# **Twitter Project notes**

## Research Question:

* Does the rate of Tweets about HIV match the prevalence of disease in American big cities?

OR

* How does knowledge of HIV, as measured by tweets about the disease, relate to disease prevalence in American big cities?
  + Later we could factor in Tweets that include inaccurate information. Are those positively correlated with disease prevalence?

Methods:

* **Questions:**
  + Cities with over X population?
  + How are big cities defined?
  + Top 50 cities in terms of population?
  + How are we processing tweets?
  + What are the exclusion/inclusion criteria?
  + What are the search terms?
  + How long/often are we running this search?
  + Demon: Tweets per time period, in each city?
  + Number: Topic specific tweets per time period, in each city?
  + Search terms:
    - HIV
      * >>HIV, HIV/AIDS, HIV testing, HIV medication, AIDS test, HIV test, HIV+, HIV(+), rapid-HIV test, rapid HIV test, ora-sure, orasure, Acquired Immune Deficiency Syndrome, Acyclovir, ADAP, Kaposi's Sarcoma, Thrush
        + [AIDS Institute](https://www.theaidsinstitute.org/education/aids-101/glossary-hivaids-related-terms)
* Ideas for mapping tweets:
  + “We used the set of rules described in Schwartz et al.53 to map location fields to counties. The locations fields were broken up into sequences of words (tokenized) and then matched to country names. Out of those messages either mentioning the country as the United States or not mentioning a country, we used the words preceding the country and attempted to match city and state names. City population information was used when the user provided a city without a state in order to determine whether the city was 90% likely to be in any state; if so, we paired the city with its most likely state. Otherwise, the tweet was discarded. For example, if Springfield, Illinois has a population of approximately 117,000 and the sum of populations across all cities named Springfield is 187,000, then we would calculate the likelihood that “Springfield” is referring to Springfield, Illinois as 117,000 / 187,000 = 62.6%; thus, Springfield would not be mapped.”
    - (Ireland, Chen, Schwartz, Ungar, & Albarracin, 2016)

# **Other Twitter Related Ideas**

* Could you evaluate the impact of tweets aimed at countering misinformation
  + How often are they retweeted?
  + How many followers do the people who post those tweets have?
  + Figure out the same thing for people who post misinformation
  + Questions
    - Who is winning the Twitter information war when it comes to HIV?
    - What areas post the most mis-information/good info
      * Does this relate to HIV rates in those areas?
* What factors impact how often a tweet is retweeted
  + *SEE* (Lohmann et al., 2018)
  + Are funny tweets about HIV better than serious tweets?
  + Are Tweets from a health agency more effective than a tweet from a private account?
  + Problems:
    - Do we need IRB approval?
    - What ethical considerations are there?
* Do tweets about disease match up with incidence?
* Impact of news on awareness of diseases
  + Mine News RSS feeds
* Does the rate of Tweets about disease risk behaviors correlate with the actual rate of infectious disease in big US cities.
  + Need
    - Disease rates in big cities
    - Tweets about the disease
    - Tweets about risk factors
    - Total Tweets in a big city per (day, hour)
  + Start simple
    - Basic prevalence of tweets about disease
  + Get complex
    - Do Tweets about disease risk factors correlate to the prevalence of disease
    - Do monthly Tweets about disease risk factors correlate with future disease rates
* Standing Questions
  + What searches do we need to run?
    - What are our search terms?
  + How long do we need to run the program for
    - Should we run it for 24hrs or 8 hrs for multi days…?

# Topic of interest:

# **TO DO:**

* ~~Modify the code so that time is more of a factor~~
  + ~~So that we can run the code for set amount of time and so that~~
  + ~~We can schedule a time to run the~~ 
    - ~~Use a while loop~~
      * ~~Add vars~~
        + ~~Start time~~
        + ~~Run length/end time~~
      * ~~If /while start LT current time then do nothing~~
        + ~~Else run code for run length~~
* Find out if you can pull old twitter data?
  + Is there code to pull retrospective data from Twitter
  + We could validate a monthly predictive model using this
* Research Carmen
  + See healthtweets.org presentation
* ~~See if I can pull the Florida groceries from google map mining code (other)~~
  + Yes, I can and have
* ~~Change delimiter to something used less frequently (~)~~
* ~~Only print tweets with geocode, if geo codes re needed~~
  + ~~Get clean Tweets about HIV~~
* ~~Review research on uses of twitter in PH research~~
* IF twitter does not work
  + Look for machine learning dataset
    - Find big publicly available longitudinal data sets
  + Can we do a survival analysis with machine learning
    - Framingham data?

Ireland, M. E., Chen, Q., Schwartz, H. A., Ungar, L. H., & Albarracin, D. (2016). Action Tweets Linked to Reduced County-Level HIV Prevalence in the United States: Online Messages and Structural Determinants. *AIDS Behav, 20*(6), 1256-1264. Retrieved from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4867271/pdf/nihms743860.pdf>. doi:10.1007/s10461-015-1252-2

Lohmann, S., White, B. X., Zuo, Z., Chan, M. S., Morales, A., Li, B., . . . Albarracin, D. (2018). HIV messaging on Twitter: an analysis of current practice and data-driven recommendations. *Aids, 32*(18), 2799-2805. Retrieved from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6615455/pdf/nihms-1527627.pdf>. doi:10.1097/qad.0000000000002018