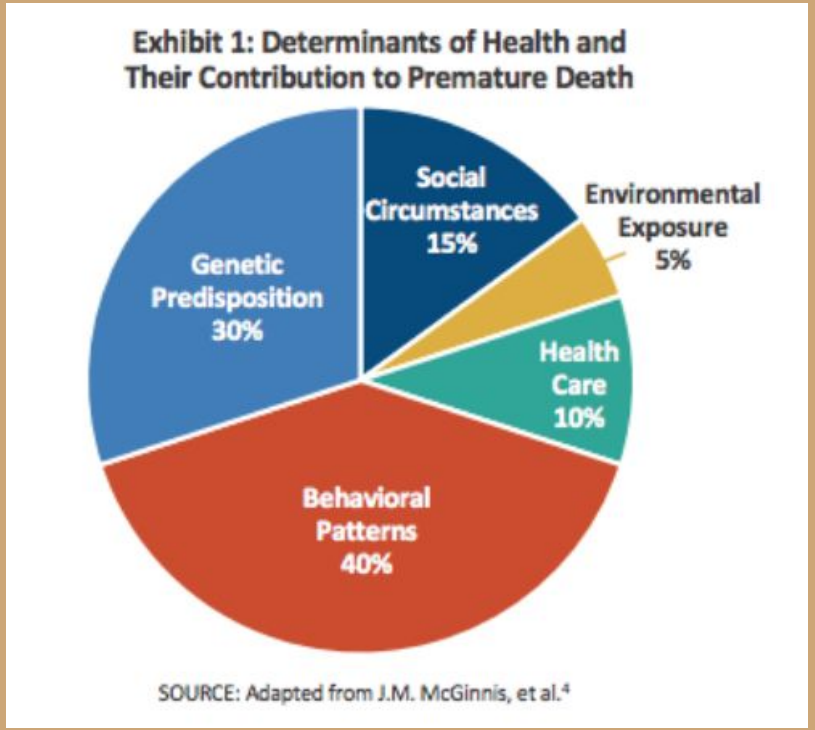
Social Media Determinants of Health

Marcus DeMaster, JingJing Rong,
Johnny Yeo

The Background

Social Factor Significantly Affect Health



The Effort

Many Big Players are Seeking to Address SDOH

- Center for Medicaid and Medicare Services (CMS)
- State Medicaid Agencies
- Managed Care Organizations (MCOs)

The Problem

It's Hard to Identify At-Risk "Social" Patients

Our Solution: The Homing Tweeter

- Utilize Tweepy to pull twitter user profiles & tweets who are
  - At-risk SDOH “social patients”
  - Self-identified as having specific medical conditions
- Create a model that can predict a user’s SDOH “riskiness”
- Provide a data tool that links the user’s SDOH scores with potential medical conditions for other users with similar SDOH profiles

What does the data look like?

Is Twitter data sufficient for this kind of work?

Rich Demographic-Inference Research on Twitter Data

- User Profile
  - Handle
  - # of Followers
  - # of Favorites
  - # of Friends
  - **Gender (demographer)**
  - **Location (carmen)**
  - **Bot Likelihood (botometer)**
- Tweets
  - Tweet text
  - Tweet Datetime
  - Recent tweets since marker tweet

Data is Self-Labeled

"I am honestly so glad I didn't go to college"

"Kinda sad but pretty glad I didn't go to college. Just counting down the days for the service "

"We need to address education costs (not subsidize them, either). Back when I went to college I only paid 7% interest as a foreigner to US."

Twitter Data is Plentiful and Expansive

- Around 6000 tweets/sec (500m tweets/day)
- There is 10+ years of data available
- As of Q3 2017, 23% of Americans as active twitter users
- The distribution of twitter users is a fairly well distributed sampling of the US. Except for age (40% 18-29 year olds use twitter), about 15-20% of the US population across these groups use twitter:
    - Men and Women
    - 30-49 years olds; 50+ year olds
    - White, Black, or Hispanic
    - High School or less; Some college; College graduate
    - Urban; suburban; rural

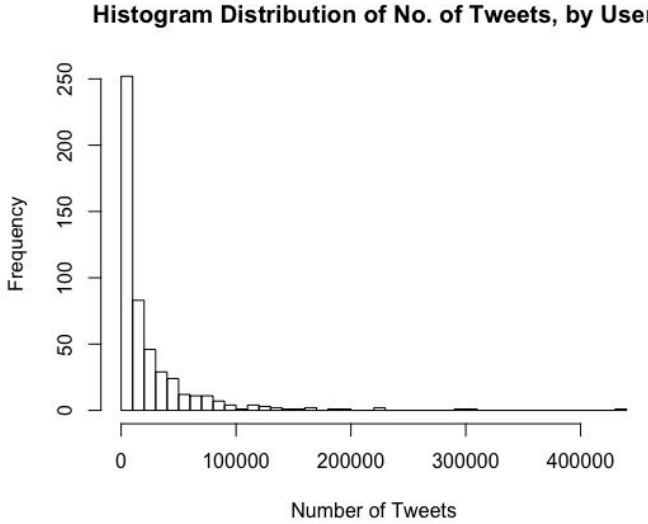
Education Risk Model EDA

"I am honestly so glad I didn't go to college"

"Kinda sad but pretty glad I didn't go to college. Just counting down the days for the service "

"We need to address education costs (not subsidize them, either). Back when I went to college I only paid 7% interest as a foreigner to US."

Average Tweets/user: **24,562**
% of US profiles (out of ~1200): **40.9%**
Gender Breakdown (M:F): **304:182**



Challenges and Considerations Overcome to Date

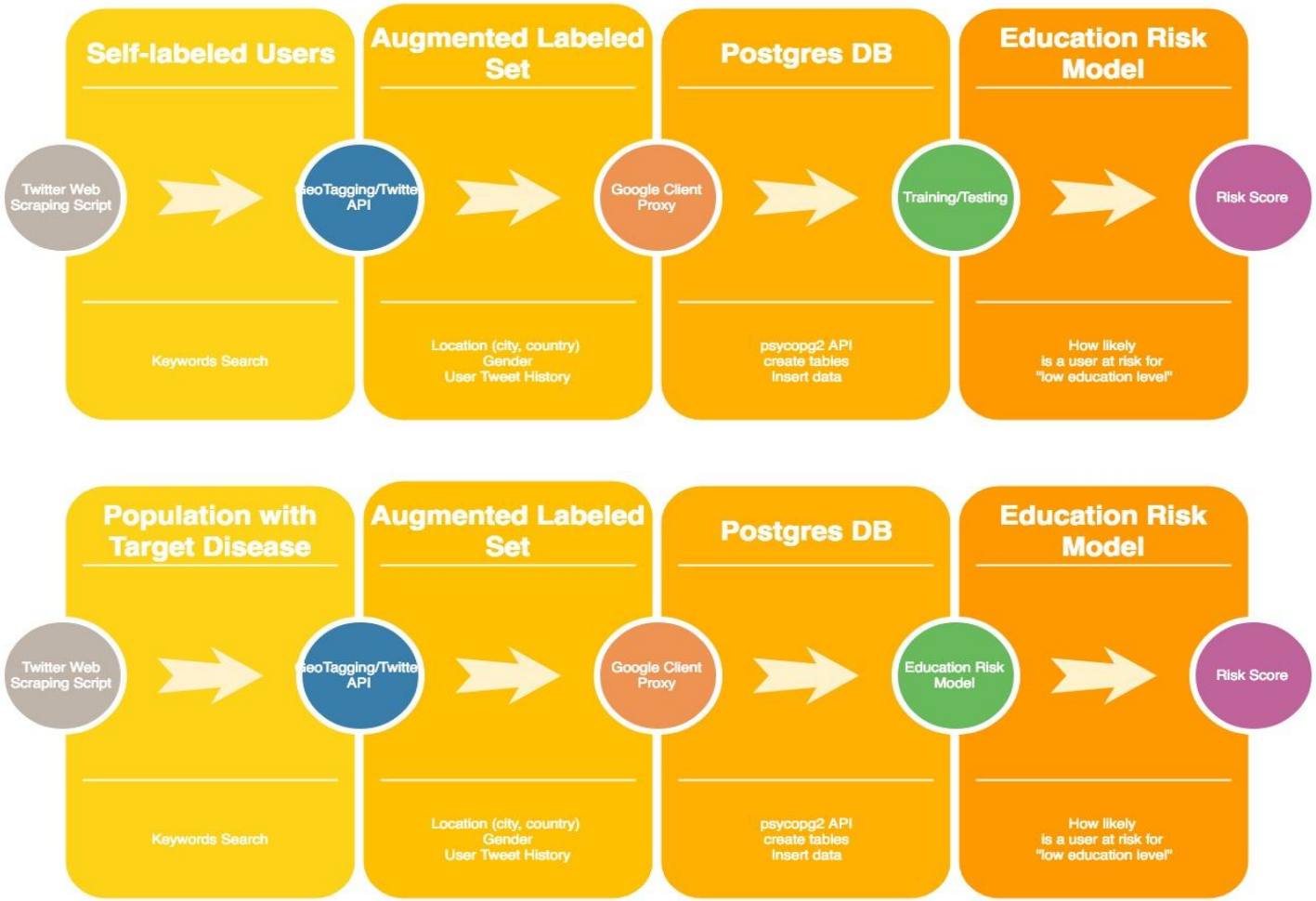
User Profile JSONs doesn't actually have much demographic data:

- Missing Gender
  - Used 'demographer'
  - NLP machine learning that takes in a name, and gives a gender prediction
- Missing Location data
  - Used 'carmen'
  - Based on Twitter Places, and using a small database developed by researchers

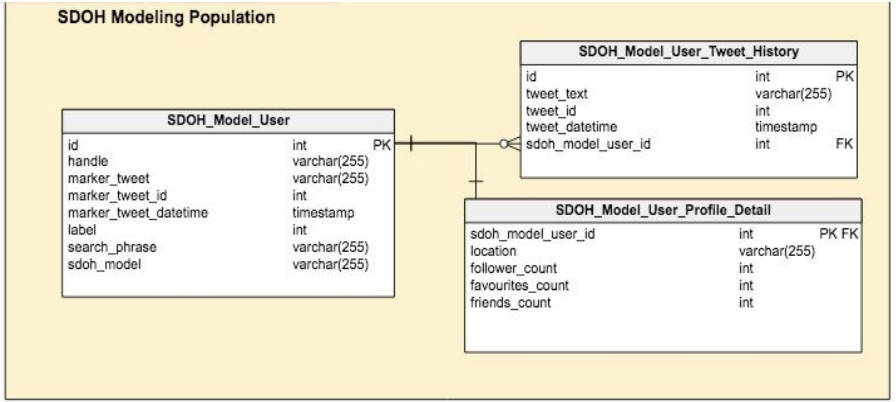
Data loading takes a long time

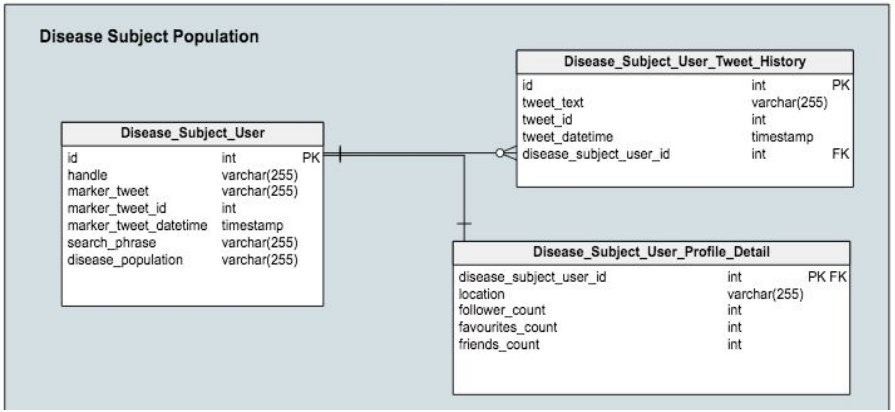
  - Tweepy has a rate limit, and forces us to wait 15 minutes (900 user/timeline get requests/15 min)
  - Proper data pipeline architecture will allow us to store data efficiently, which we'll talk about more later

Pipeline


Self-labeled Users
Twitter Web Scraping Script
Keywords Search
Augmented Labeled Set
GeoTagging/Twitter API
Location (city, country)
Gender
User Tweet History
Postgres DB
Google Client Proxy
psycopg2 API
create tables
insert data
Education Risk Model
Training/Testing
How likely
is a user at risk for
"low education level"
Risk Score
Population with Target Disease
Twitter Web Scraping Script
Keywords Search
Augmented Labeled Set
GeoTagging/Twitter API
Location (city, country)
Gender
User Tweet History
Postgres DB
Google Client Proxy
psycopg2 API
create tables
insert data
Education Risk Model
Education Risk Model
How likely
is a user at risk for
"low education level"
Risk Score

Database Schema





Algorithms, Techniques

- **Feature Extraction**
  - Text mining, timestamp, LabMT sentiment analysis, LIWC category feature, WordNet
  - Interactions and friendships
  - User profile/ demographic information
- **Predictive Model Construction**
  - SVM
  - Regression
  - Decision trees/random forest
  - Neural network
- **Model Verification**
  - n-fold cross validation
  - separate self-labeled data into training vs testing

Questions?

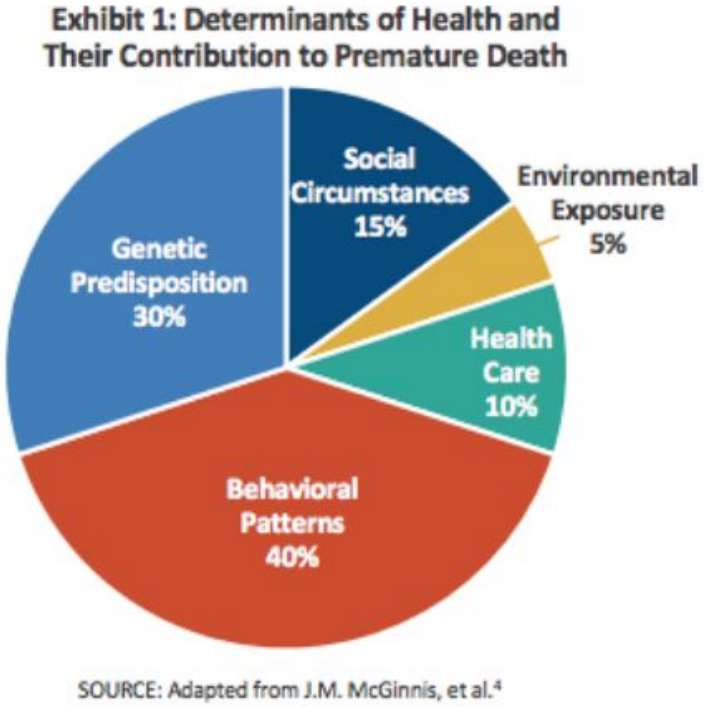
What are SDOH?

“Social Determinants of Health (SDOH) are conditions in the environment in which people are born, live, learn, work, play, worship, and age that affect a wide range of health, functioning, and quality-of-life outcomes and risks.”

-healthypeople.gov (HHS Office of Disease Prevention and Health Promotion)

https://www.healthypeople.gov/2020/topics-objectives/topic/social-determinants-of-health

Social Factors Significantly Affect Mortality Risk



https://www.chcs.org/media/CHCS-SDOH-Measures-Brief_120716_FINAL.pdf

Players Seeking to Address SDOH

- Center for Medicaid and Medicare Services (CMS)

- State Medicaid Agencies

- Managed Care Organizations (MCOs)

Efforts at the Federal Level

CMS Accountable Health Communities Model

- Launched May 1st, 2017
- 5 year study
- $120M grant for 32 “Bridge” Organizations
  - Oregon Health & Science University, Hackensack University, Baltimore Health Dept.
  - Screen Medicare/Medicaid beneficiaries for social needs
  - Help high-risk individuals navigate community services

Efforts at the State Level

Exhibit 2: Current Data Collection on Common SDOH Domains in Select States*†

SDOH Domains	KS	MA	MI	NY	OR	TN	VT	WA
Housing	✔	✔	✔	✔	✔	✔	✔	✔
Family and Social Support	✔	✔	✔	✔		✔	✔	✔
Education and/or Literacy	✔	✔	✔		✔		✔	✔
Food Security		✔	✔		✔	✔	✔	✔
Employment	✔	✔	✔	✔	✔	✔	✔	✔
Transportation		✔	✔				✔	✔
Criminal Justice Involvement	✔	✔		✔	✔		✔	✔
Intimate Partner Violence		✔		✔	✔			

*Data collected from Medicaid beneficiaries at the individual and/or population level.
† Data not systematically collected on the entire Medicaid population.

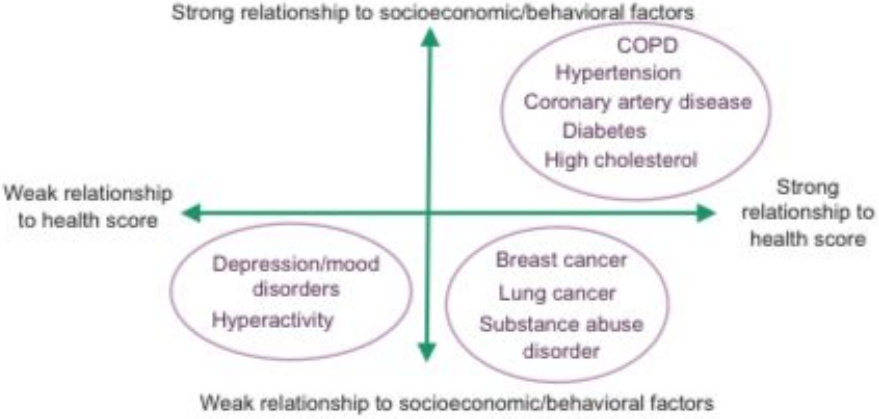
https://www.chcs.org/media/CHCS-SDOH-Measures-Brief_120716_FINAL.pdf

Efforts by Managed Care Organizations

➔ **December 2017 Moody's Analytics study of 24M BCBS members under age 65.**

Chart E1: Taxonomy of Index Health Conditions

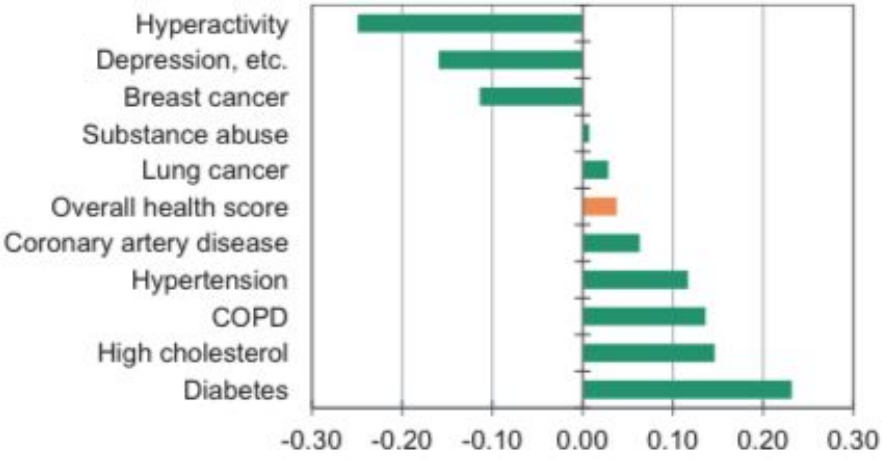
Conditions vary in important ways in terms of what drives them and how they relate to overall health



Sources: BCBS, Moody's Analytics

Chart E2: Education Has Mixed Effects

Effect of % population with college degree on condition z-score

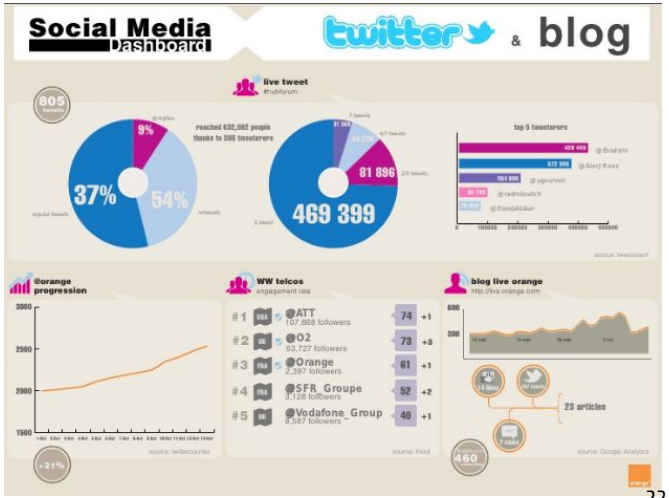


Sources: BCBS, Moody's Analytics

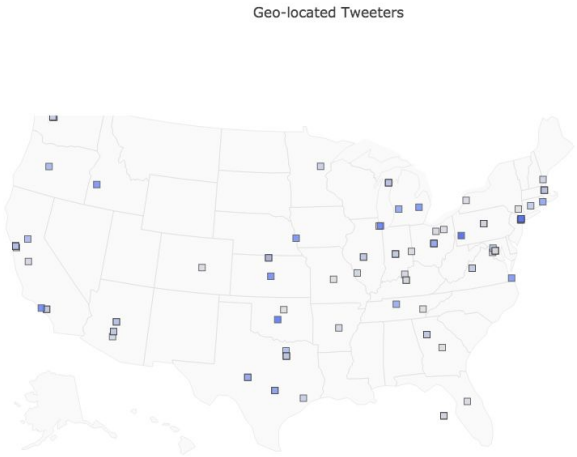
https://www.bcbs.com/sites/default/files/file-attachments/health-of-america-report/BCBS.HealthOfAmericaReport.Moodys_02.pdf

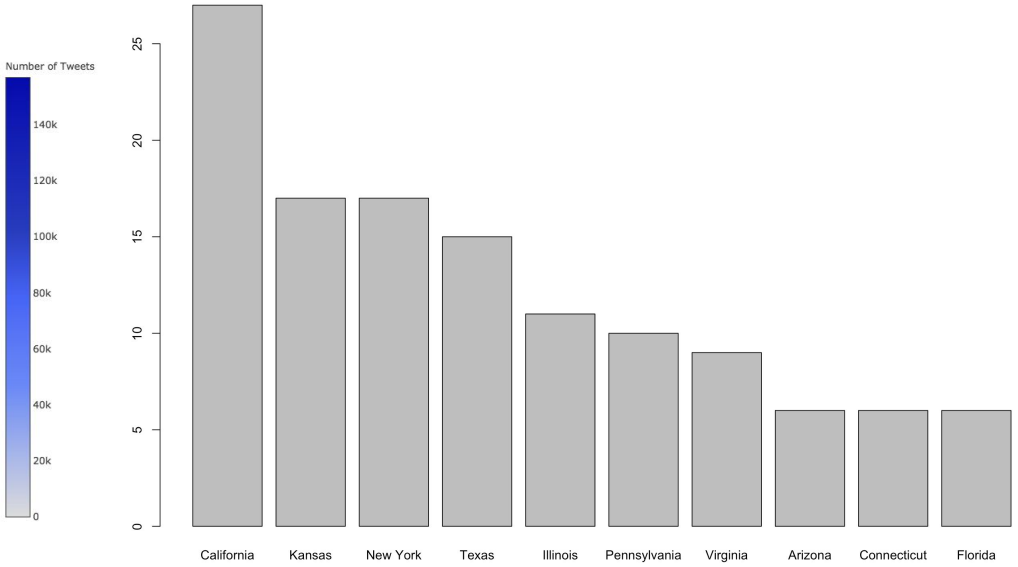
Overcoming Current Pain Points

- Healthcare orgs screen SDOH for those already receiving care
- Collecting data through questionnaires is highly manual and incomplete
- A social media SDOH monitoring tool...
  - SDOH NLP predictive models
  - Broader analysis
  - Automated
  - Longitudinal
- Minimum Viable Product
  - Education Risk Model → 5-6 Chronic Diseases



Geo-Located Tweeters & Tweet Distribution by State (sampled)





Disease Subjects

“I dead forget I have COPD until Im having trouble breathing”

"Thanks Clive, this pollution is is having a negative affect on my COPD."

"Got my physical results back....apparently I have hypertension and a possible anxiety issue. Well duh! I work retail lol"

"I have type 2 diabetes but my life will go on just got to make some changes for the better"

Type 2 Diabetes

Average Tweets/user: **20,640**
% of US profiles (out of ~280): **35.6%**
Gender Breakdown (M:F): **63:35**