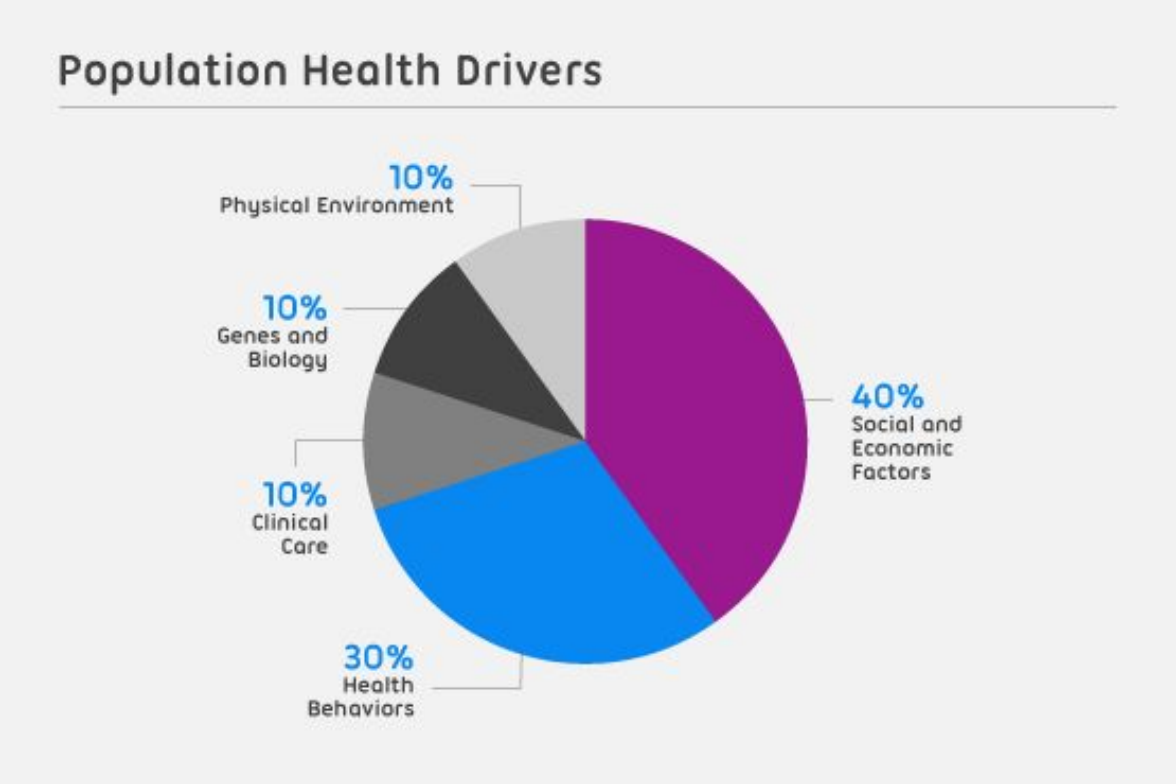
Social Media Determinants of Health

Marcus DeMaster, JingJing Rong, Johnny Yeo

Table of Contents

A Brief Run-down of what we'll cover

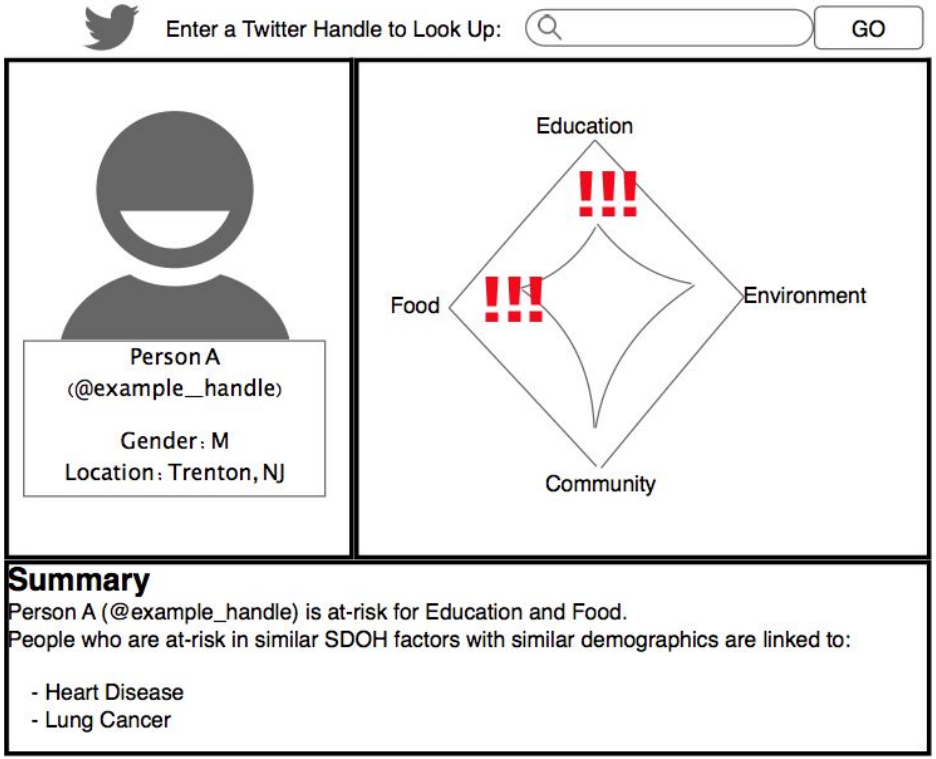
Before we start…



The Background: Social Determinants of Health Significantly Drive Population Health

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

Our Solution: The Homing Tweeter



The MVP

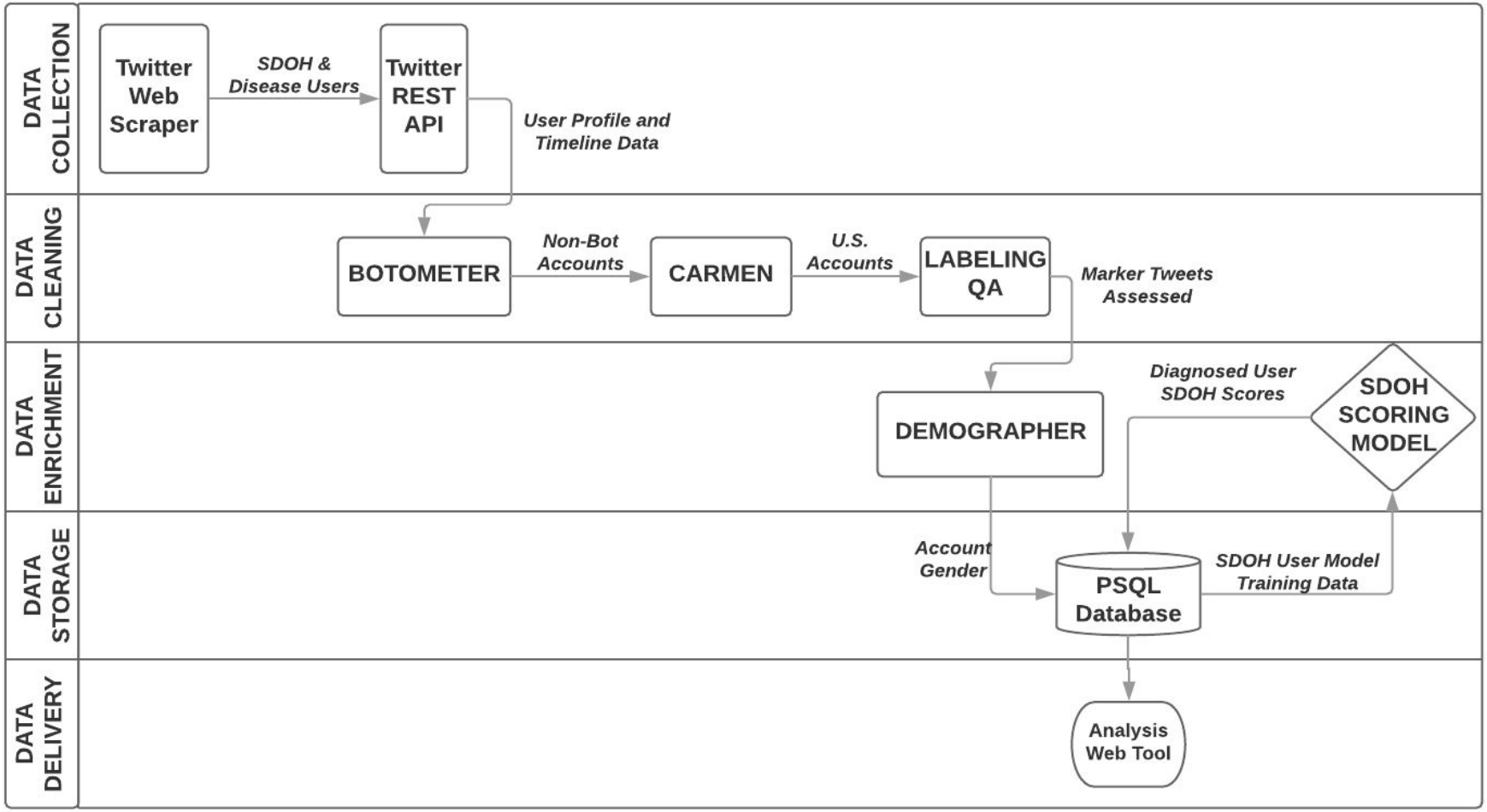
The Minimum Viable Product

- Create a data pipeline for training education model and linking to 5 diseases
- Develop a model for predicting education at-risk scores
- Create aggregate Tableau dashboard to confirm explanatory power of model

Creating a Data Pipeline

Tweepy + Demographic Feature Additions

Pipeline Overview



Data QA Process

		Bot Likelihood<0.4		Tweet Age < year		Manual Review
Tweets	680k	634k	485k	485k	468k	461k
Users	7800	6618	6612	6611	6195	6096
	Initial Counts		**Remove RTs**		**Tweet Counts between 30 & 100**	

10

Developing the Model

Tweepy + Demographic Feature Additions

LR Model V0

Accuracy:

0.6 (partial dataset) => 0.93 (partial dataset) => 0.59 (full dataset)

- Accuracy is up with the partial dataset
- Does not provide the same insights as the existing researches on disease populations
- Accuracy drops back to 0.59 when training on 600k data

=> Data Examination & Data Cleaning

LR with Re-labeled Data

0.59 => 0.65

Relevant Features:

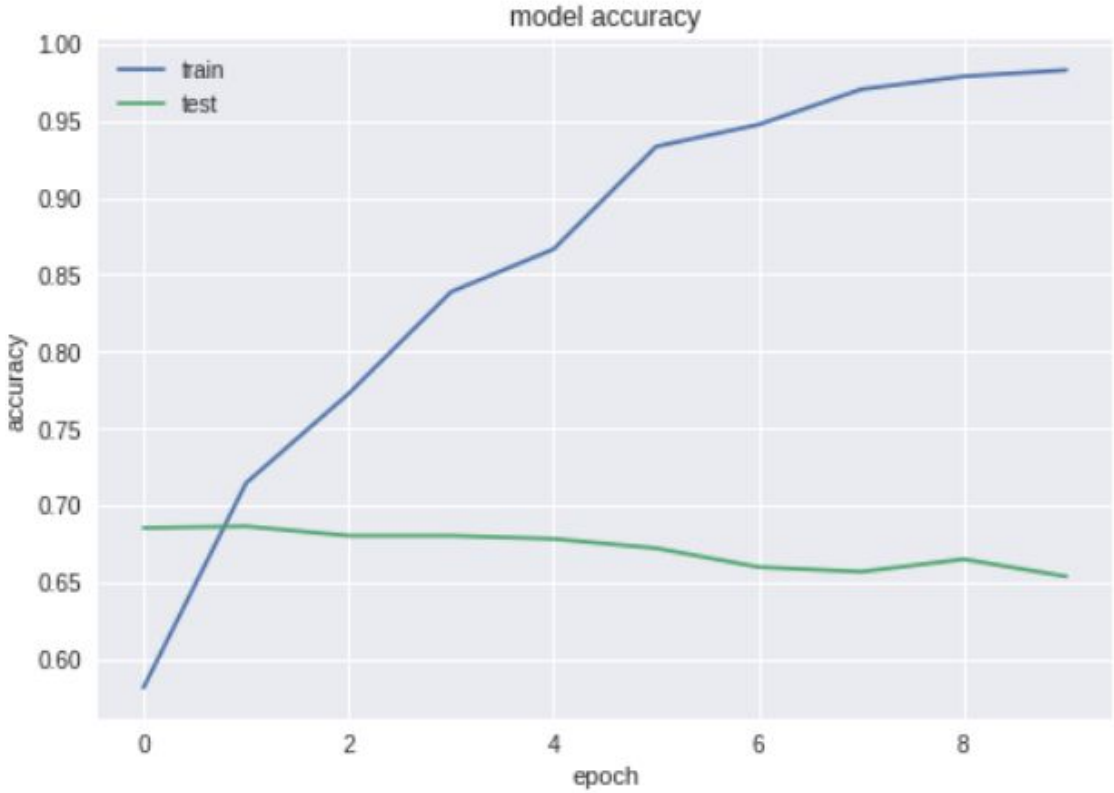
1. Polarity
2. Subjectivity
3. Gender

LR Model performance < NN Model

```
Optimization terminated successfully.
         Current function value: 0.654931
         Iterations 5
                           Logit Regression Results
==============================================================================
Dep. Variable:                      y   No. Observations:                 6096
Model:                          Logit   Df Residuals:                     6088
Method:                           MLE   Df Model:                            7
Date:                Fri, 16 Mar 2018   Pseudo R-squ.:                 0.03556
Time:                        06:35:38   Log-Likelihood:                -3992.5
converged:                       True   LL-Null:                       -4139.7
                                        LLR p-value:                 9.237e-60
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
x1            -2.4060      0.300     -8.017      0.000      -2.994      -1.818
x2             3.5565      0.473      7.514      0.000       2.629       4.484
x3             0.6231      0.055     11.404      0.000       0.516       0.730
x4             0.0017      0.001      2.663      0.008       0.000       0.003
x5            -0.0012      0.000     -3.261      0.001      -0.002      -0.000
x6            -0.0034      0.001     -4.283      0.000      -0.005      -0.002
x7         -8.214e-09   2.43e-09     -3.380      0.001    -1.3e-08   -3.45e-09
x8         -5.228e-09   1.54e-09     -3.395      0.001   -8.25e-09   -2.21e-09
==============================================================================
```

Deep Learning Model

- Keras and Tensorflow
- Simple Sequential Model
- Word embeddings only
- Val. Acc.: 69%

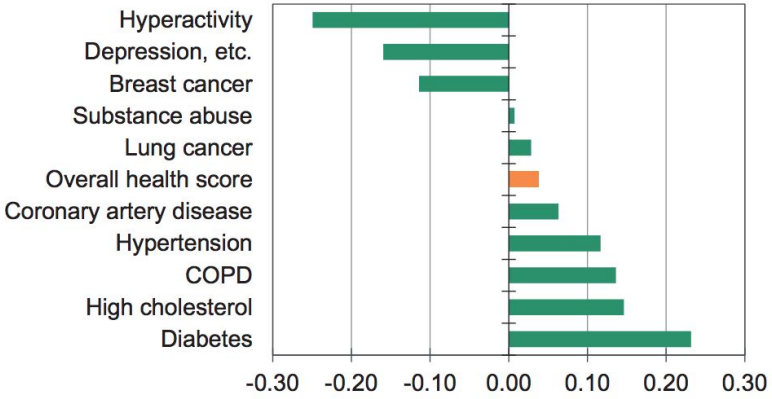


Aggregate Tableau Dashboard

Exploring Findings

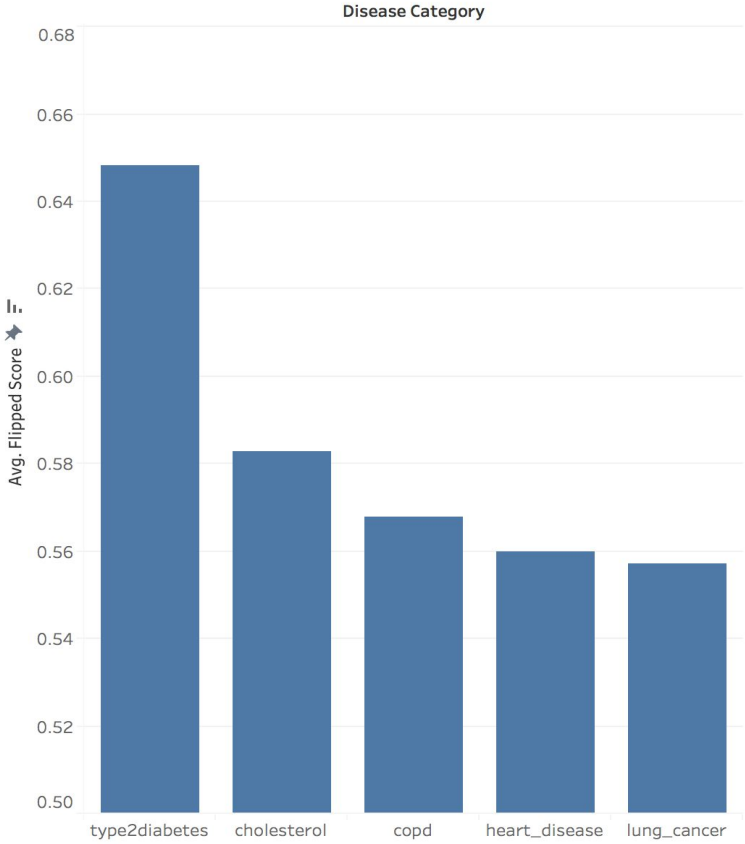
Chart E2: Education Has Mixed Effects

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



Sources: BCBS, Moody's Analytics

BarPlot of Avg Scores, by Disease



Roadmap Ahead

The path to our final MVP

- **Week 11:** Further work on model implementation, web tool development
- **Week 12:** Final analytics front-end built and incorporated into web tool
- **Week 13:** Implement 1 or 2 more SDOH models (housing, employment)
- **Week 14:** Final Presentation


ration of naphthalene.
(N)
p-BrC6H4