The (In)ability to Triangulate in Data Driven Healthcare Research

Philip Resnik

University of Maryland

resnik@umd.edu

SBS Decadal Survey - Workshop on Culture, Language, and Behavior
National Academies of Sciences, Engineering, and Medicine
October 11, 2017

Triangulate

Data Healthcare

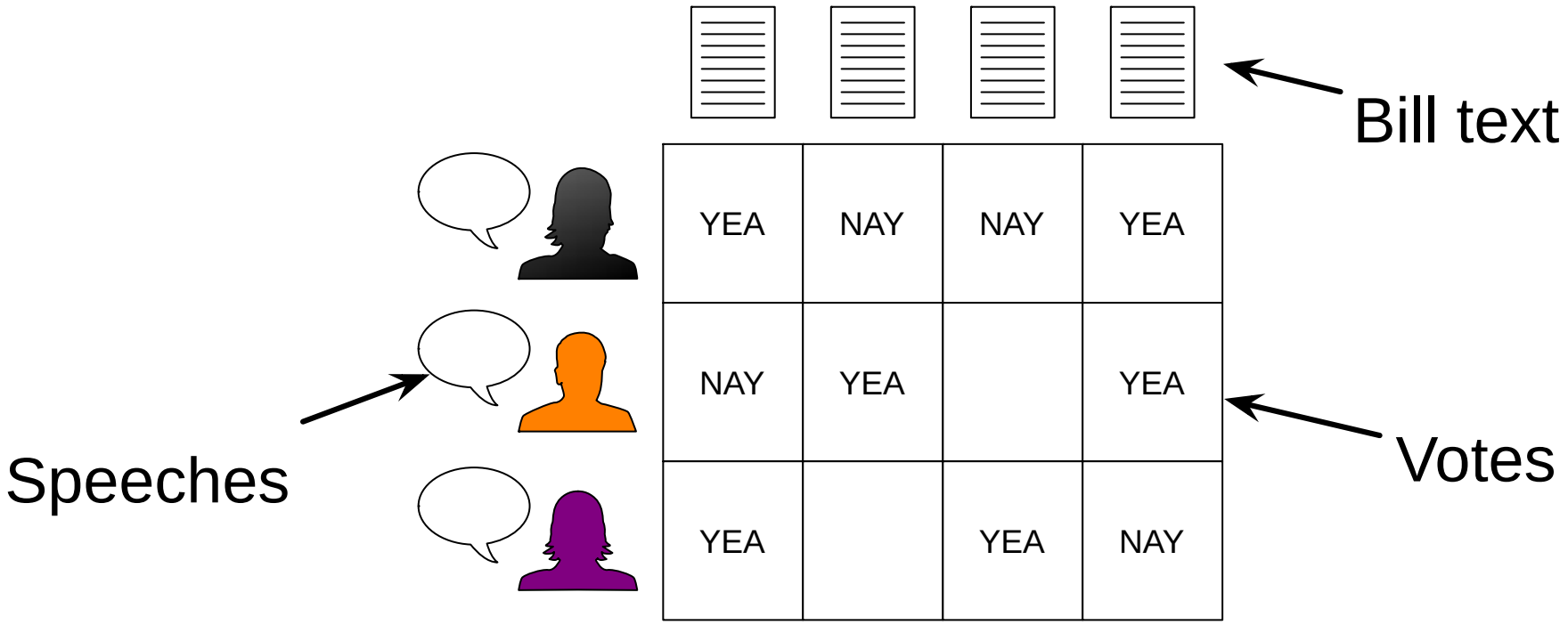
Modeling political attitudes using behavior

YEA	NAY	NAY	YEA
NAY	YEA		YEA
YEA		YEA	NAY

Martin and Quinn, 2002; Bafumi et al., 2005; Gerrish and Blei, 2011

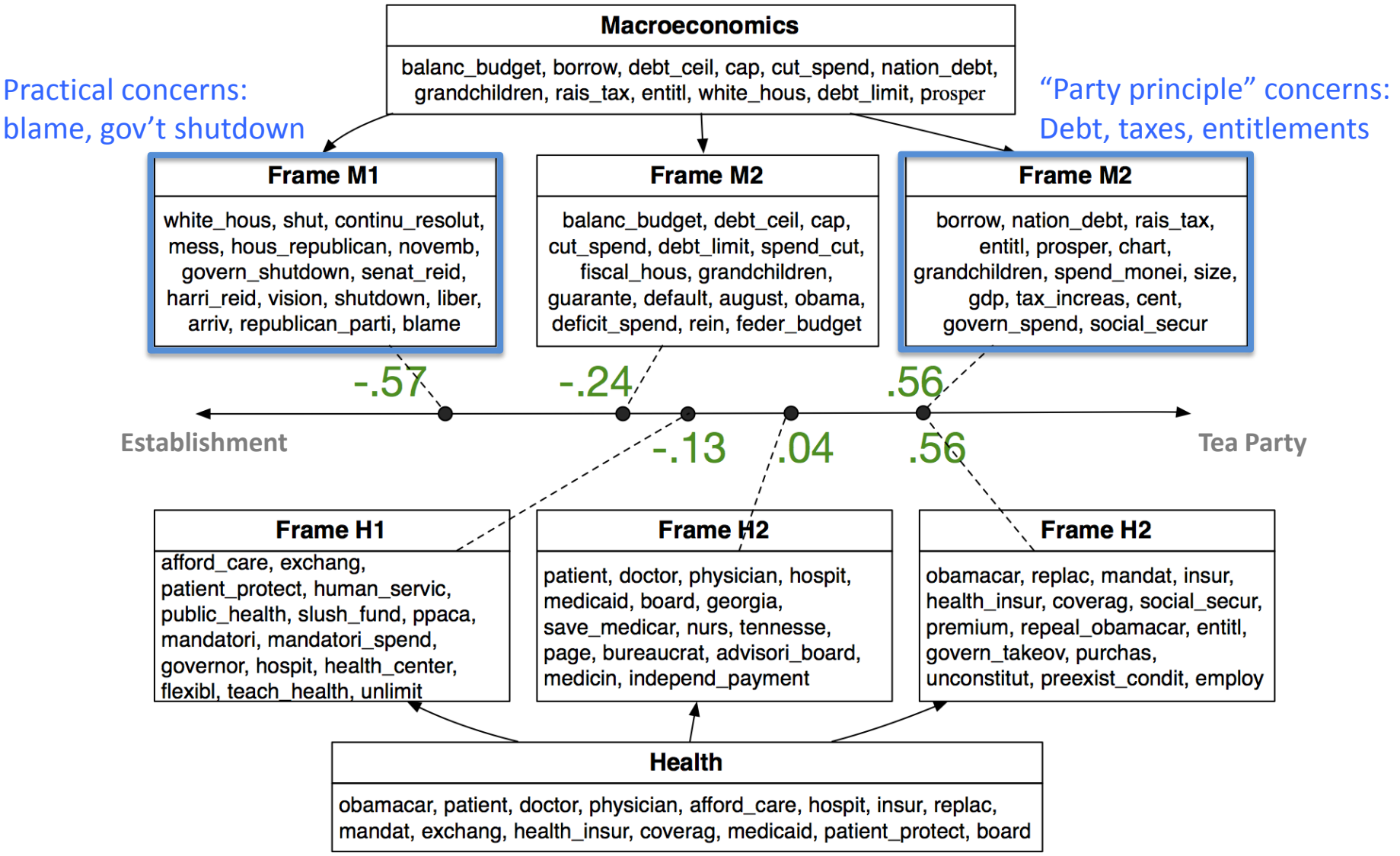
Figure adapted from Viet-An Nguyen

Triangulating: behavior and language



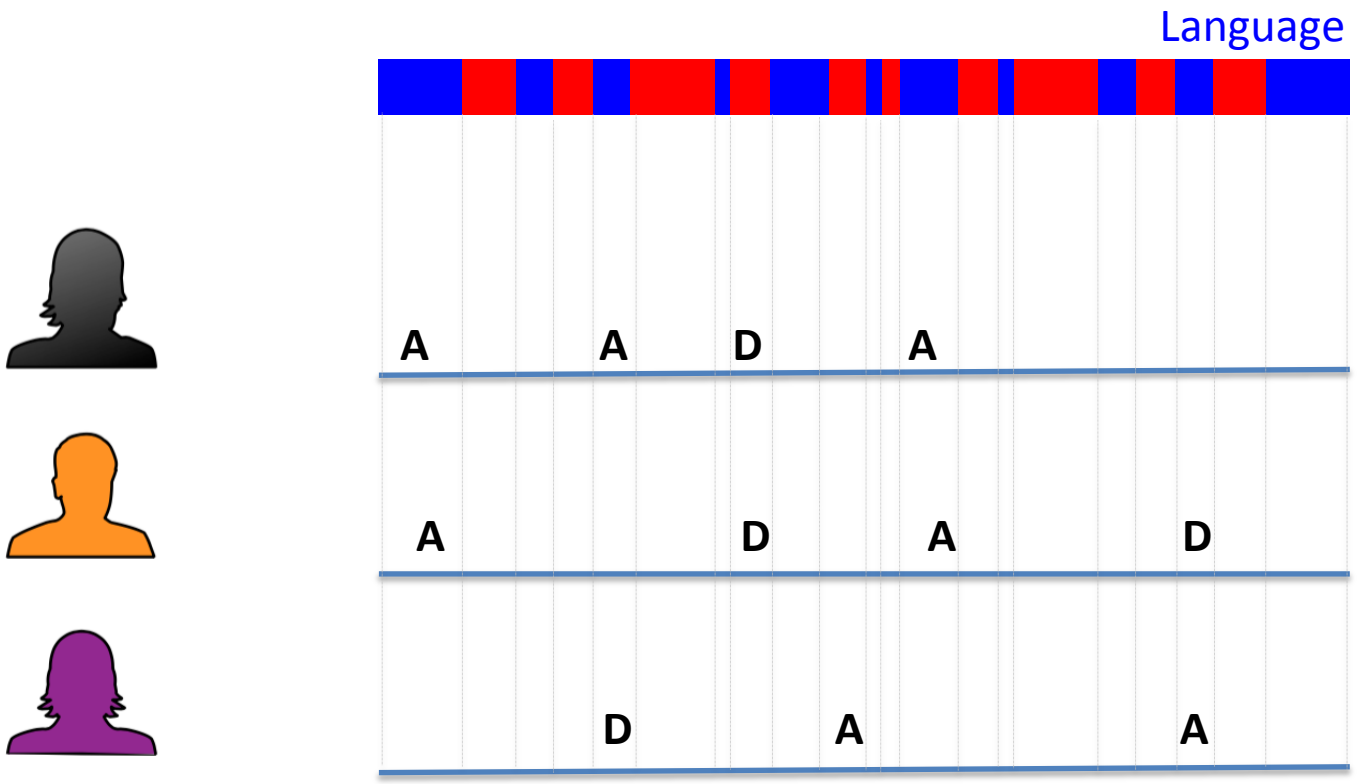
$$p(v_{a,b} \mid \boldsymbol{u}_a, x_b, y_b, \hat{\vartheta}_b) = \Phi\left(x_b \sum_k \hat{\vartheta}_{b,k} u_{a,k} + y_b\right)$$

Nguyen et al., 2015
Extending Gerrish and Blei 2012, Lauderdale and Clark 2014
Figure adapted from Viet-An Nguyen



Nguyen et al., Tea Party in the House: A Hierarchical Ideal Point Topic Model and Its Application to Republican Legislators in the 112th Congress. Association for Computational Linguistics, Beijing, July 2015.

Triangulating: behavior and language



Non-political actors

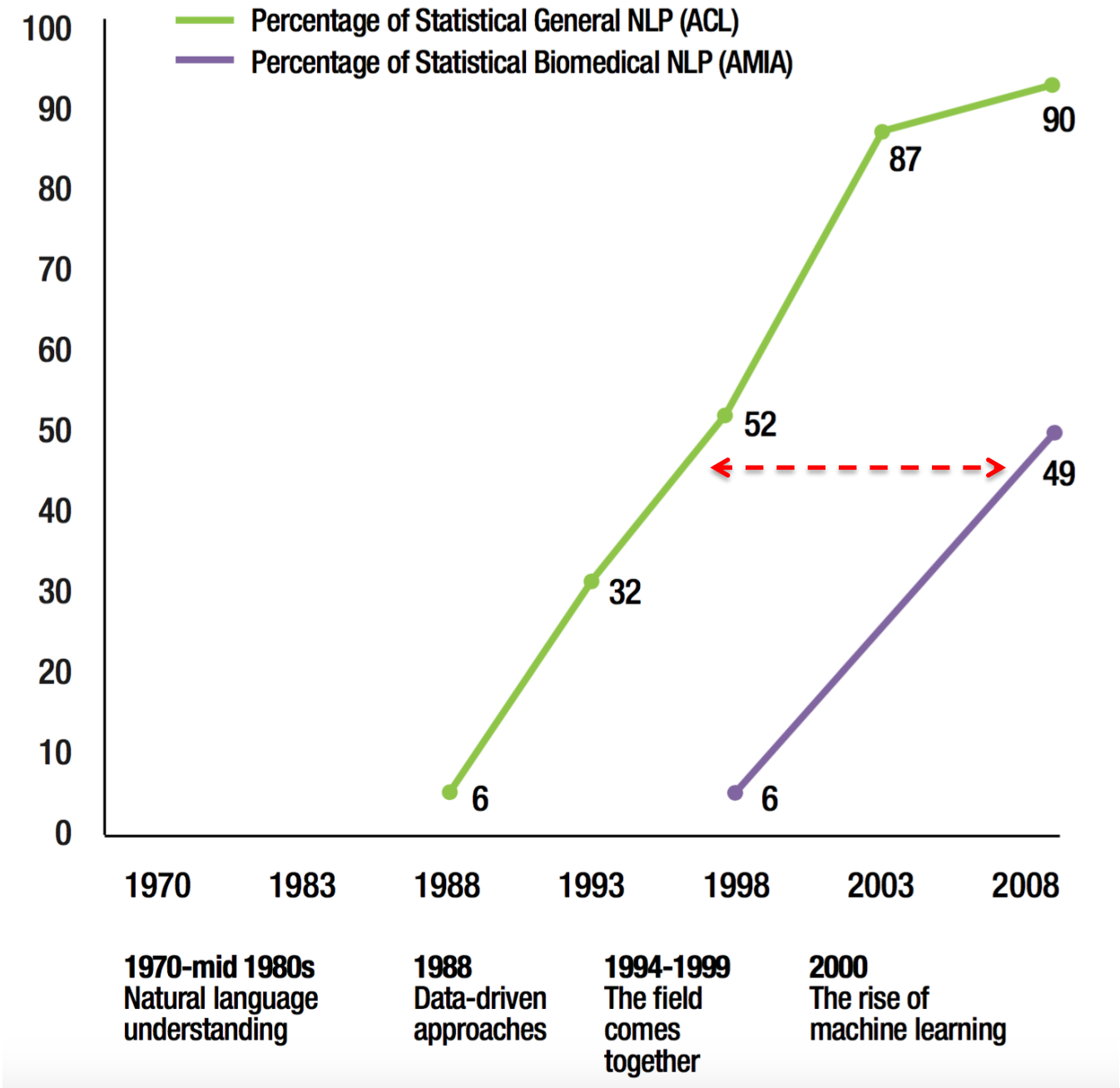
"Votes"

Real-time responses

Social media sentiment

Daniel Argyle, Philip Resnik, and Vlad Eidelman, Using Ideal Point Models to Characterize Political Reactions in Non-Political Actors, Seventh Annual Conference on New Directions in Analyzing Text as Data, Boston, Oct 14-15 2016

Data Healthcare



Adapted from http://multimedia.3m.com/mws/media/988566O/paths-to-success-cac-nlp-white-paper.pdf.

A sampling of NLP research datasets

Blog Authorship Corpus: consists of the collected posts of 19,320 bloggers gathered from blogger.com in August 2004. 681,288 posts and over 140 million words. (298 MB)

Cornell Movie Dialog Corpus: contains a large metadata-rich collection of fictional conversations extracted from raw movie scripts: 220,579 conversational exchanges between 10,292 pairs of movie characters, 617 movies (9.5 MB)

Enron Email Data: consists of 1,227,255 emails with 493,384 attachments covering 151 custodians (210 GB)

Hansards text chunks of Canadian Parliament: 1.3 million pairs of aligned text chunks (sentences or smaller fragments) from the official records (Hansards) of the 36th Canadian Parliament. (82 MB)

Reddit Submission Corpus: all publicly available Reddit submissions from January 2006 - August 31, 2015). (42 GB)

Twitter Sentiment140: 1.6 million Tweets related to brands/keywords. (77 MB)

Yahoo! Answers Comprehensive Questions and Answers: Yahoo! Answers corpus as of 10/25/2007. Contains 4,483,032 questions and their answers. (3.6 GB)

https://github.com/niderhoff/nlp-datasets

A sampling of *healthcare* NLP research datasets

SemEval-2017: Clinical TempEval. 400 manually de-identified **clinical notes and pathology reports from cancer patients at the Mayo Clinic.**

CLEF eHealth 2016. Suominen H, Zhou L, Hanlen L, Ferraro G. Benchmarking Clinical Speech Recognition and Information Extraction: New Data, Methods, and Evaluations. JMIR Med Inform 2015;3(2):e19 **Synthetic dataset** of 101 handover records.

MIMIC-III, a freely accessible critical care database. Johnson AEW, Pollard TJ, Shen L, Lehman L, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, and Mark RG. Scientific Data (**2016**). ~2M free text notes from ~40K **critical care patients at Beth Israel Deaconess Medical Center**.

CLPsych 2015. Triage of posts from a **mental health forum**; 65K posts.

Choudhury, Munmun De et al. "Discovering Shifts to Suicidal Ideation from Mental Health Content in Social Media." CHI (**2016**). ~80K posts from **mental health related forums** on Reddit.

CLPsych 2016. Triage of posts from a mental health peer support forum; 65K posts.

***Not* clinical ground truth**

What's the problem?

- HIPAA balkanizes research
- Language data is hard to fully de-identify
- EHRs create pressure to avoid language
- **It's easy to just work on something else**

How NLP can help cure cancer?

Regina Barzilay
CSAIL, MIT

Interpretable Neural Models

Goal: Generate rationale behind the predictions

this beer pours ridiculously clear with tons of carbonation that forms a rather impressive rocky head that settles slowly into a fairly dense layer of foam. this is a real good lookin' beer, unfortunately it gets worse from here ... first, the aroma is kind of bubblegum-like and grainy. next, the taste is sweet and grainy with an unpleasant bitterness in the finish. ... ... overall, the fat weasel is good for a fairly cheap buzz, but only if you like your beer grainy and bitter .

Ratings

Look: 5 stars

Aroma: 2 stars

Key properties of rationales:

- short and coherent pieces of text from the original input
- sufficient for prediction as substitution of the original input

Rationals are not provided during training

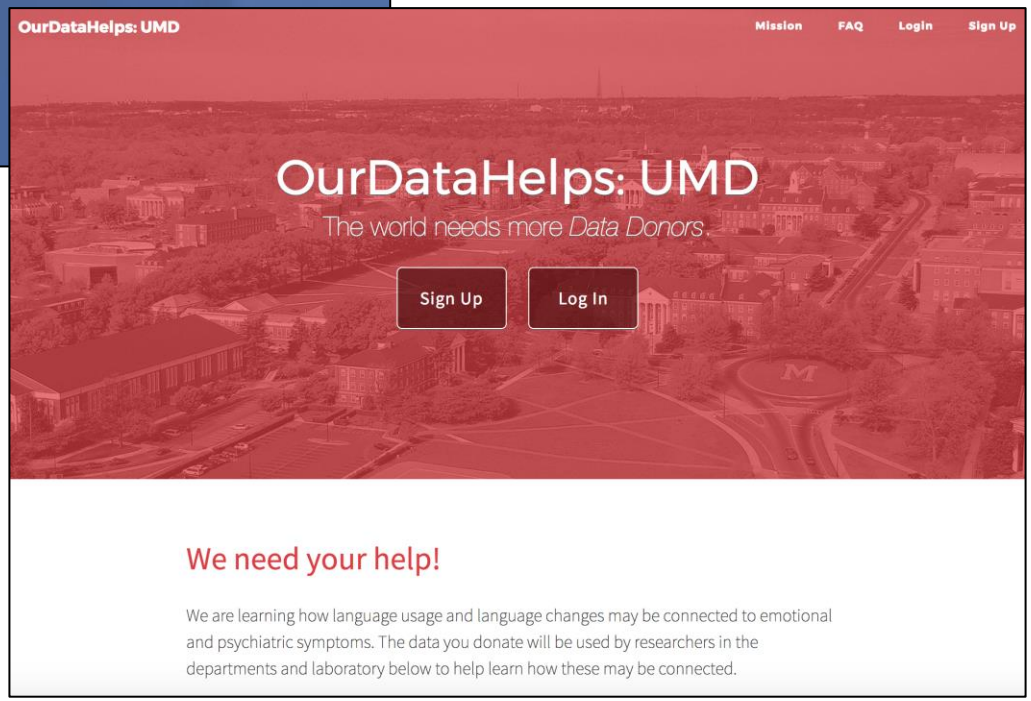
NAACL 2016 keynote
Adapted from https://people.csail.mit.edu/regina/talks/CNLP.pdf

What's the problem?

- HIPAA balkanizes research
  - *Researchers can't fix HIPAA*
- Language data is hard to fully de-identify
  - *High accuracy automation isn't enough*
- EHRs create pressure to avoid language
  - *NLP is helping, but not fast enough*
- **It's easy to just work on something else**
  - ***We need to find a different way***

ntfy

ourdatahelps.org



umd.ourdatahelps.org

Philip Resnik and Deanna Kelly, Development of Computational Modeling to Identify Symptom Changes in Schizophrenia and Depression, UMB/UMCP MPower seed grant

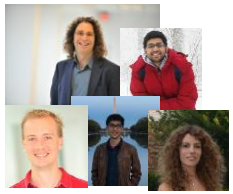
Recruiting

Consent infrastructure

Non-PHI clinical data

Social media

Collection/anonymization infrastructure



Computing environment

Progress

- UPenn Linguistic Data Consortium (LDC)
  - Framingham Heart Study project
- Health Natural Language Processing Center (hNLP)
  - LDC-like repository/dissemination of healthcare data
- NIH "All of Us" (Precision Medicine Initiative)
  - "EHR data may be sent directly by the participant's health care provider or sent by the participant to the program through Sync for Science … The initial data types to be included are demographics, visits, diagnoses, procedures, medications, laboratory tests, and vital signs, but may be expanded to all parts of the EHR, including health care provider notes."
  - A chicken-egg problem:
    - The unstructured data problem requires more investment
    - We need progress with unstructured data to justify it

Take-aways

- Healthcare is a national security issue
- Language data is a hugely valuable resource for triangulation
- We have a *lot* of catching up to do
- More needs to be done faster