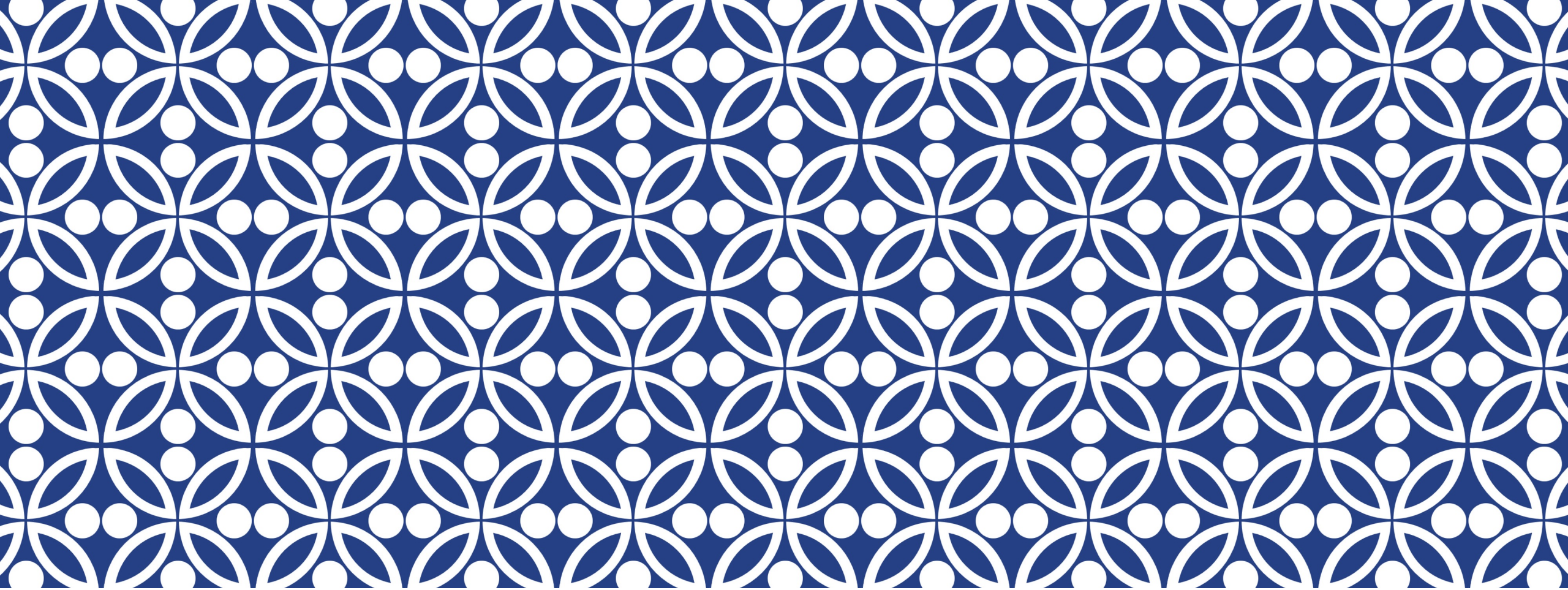


THEN AND NOW: QUANTIFYING THE LONGITUDINAL VALIDITY OF SELF-DISCLOSED DEPRESSION DIAGNOSES

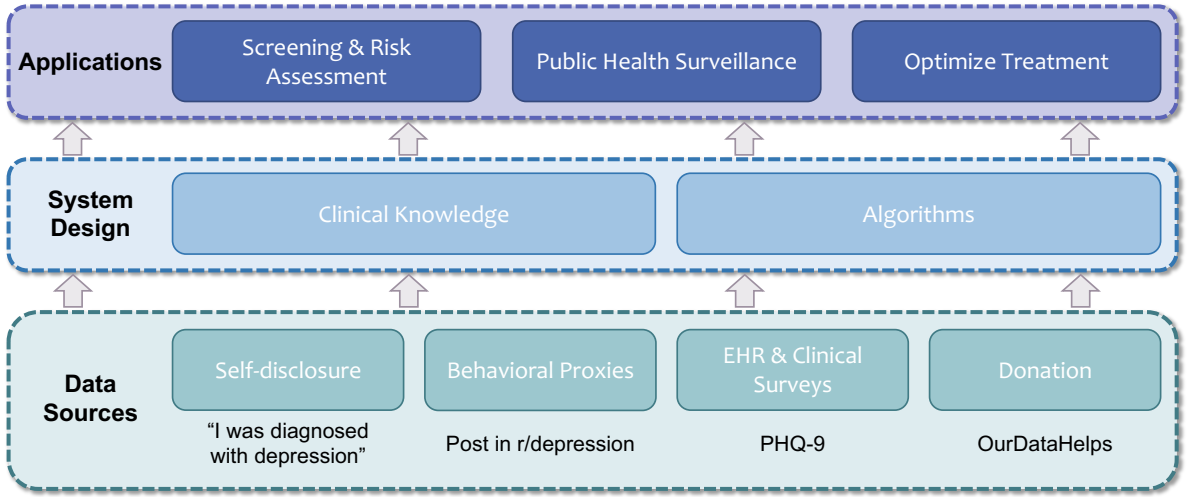
Keith Harrigian & Mark Dredze
Johns Hopkins University

MOTIVATION

Data Drives Mental Health Research

Clinical ground truth preferred, but not feasible to acquire at scale

Proxy- and rule-based methods sacrifice precision for increased recall



MOTIVATION

Data Drives Mental Health Research

Clinical ground truth preferred, but not feasible to acquire at scale

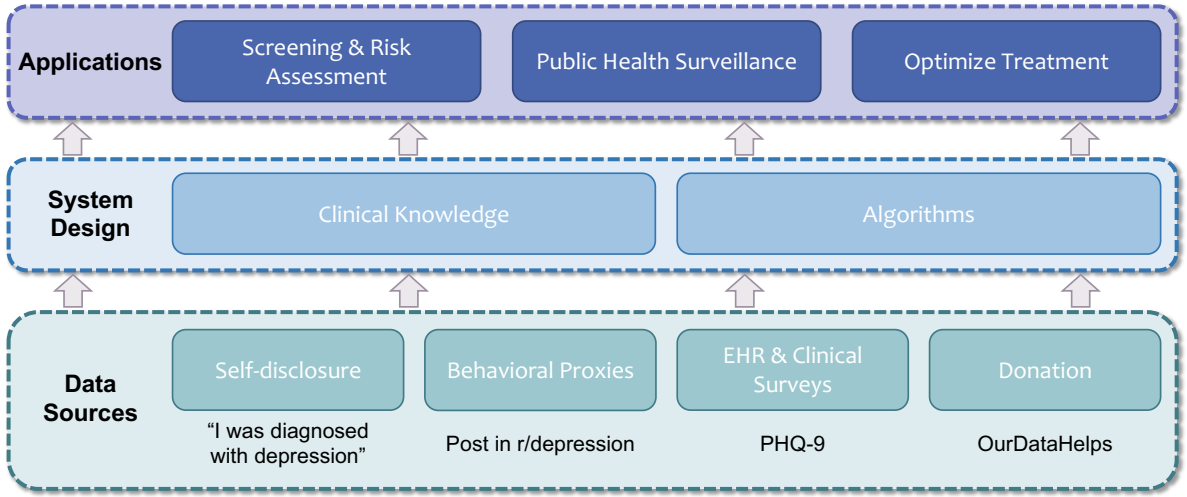
Proxy- and rule-based methods sacrifice precision for increased recall

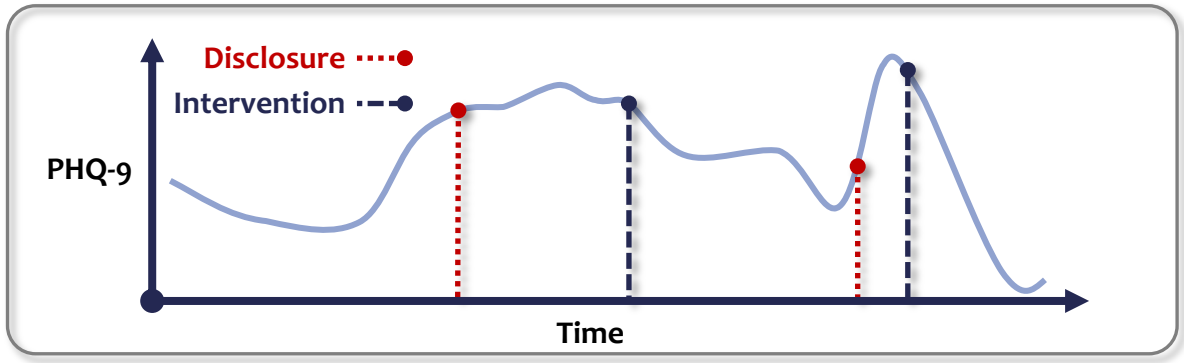
Concerns and Criticism

Construct validity

Selection bias and representation issues

Static labels for an inherently dynamical latent attribute





RELATED WORK

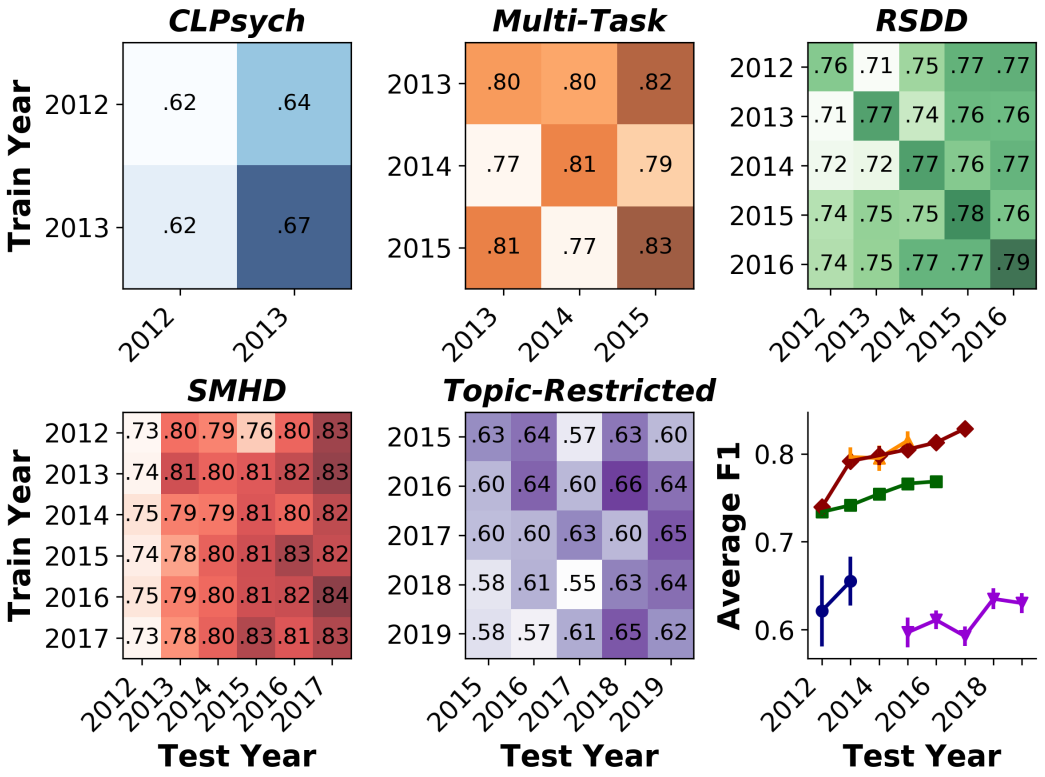
The Importance of Modeling Temporality

Time-based Priors (Wongkobap et al., 2019; Rao et al., 2020; Uban et al., 2021)

Online support groups as interventions (Chancelor et al., 2016)

Variation in predictive performance over time (Losada et al., 2017; Harrigian et al., 2020)

Clinical knowledge (Johnson and Nowak, 2002; Schoevers et al., 2005)



"Do Models of Mental Health based on Social Media Generalize?" (Harrigian et al., 2020)

SPECIFIC AIMS

To what extent do self-disclosures of a mental health diagnosis remain valid over time as proxies for mental health status?

SPECIFIC AIMS

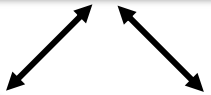
To what extent do self-disclosures of a mental health diagnosis remain valid over time as proxies for mental health status?

Quantitative Analysis

How does predictive performance change when training a classifier on new data associated with an old label?

SPECIFIC AIMS

To what extent do self-disclosures of a mental health diagnosis remain valid over time as proxies for mental health status?



Quantitative Analysis

How does predictive performance change when training a classifier on new data associated with an old label?

Qualitative Analysis

Are changes in predictive performance (or lack thereof) due to psychiatric dynamics or due to sample-related confounds?

DATA

2015 CLPsych Shared Task

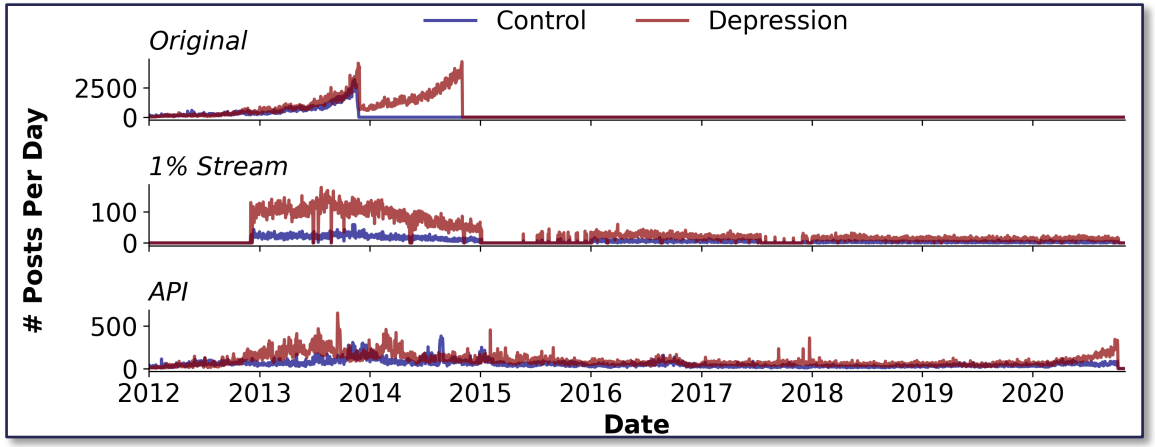
Regular-expressions and expert verification of diagnoses disclosures

Updating the Dataset

Account identifiers available w/ explicit permission from Coppersmith et al. (2015)

Query all available data from Twitter API and institution cache of 1% stream

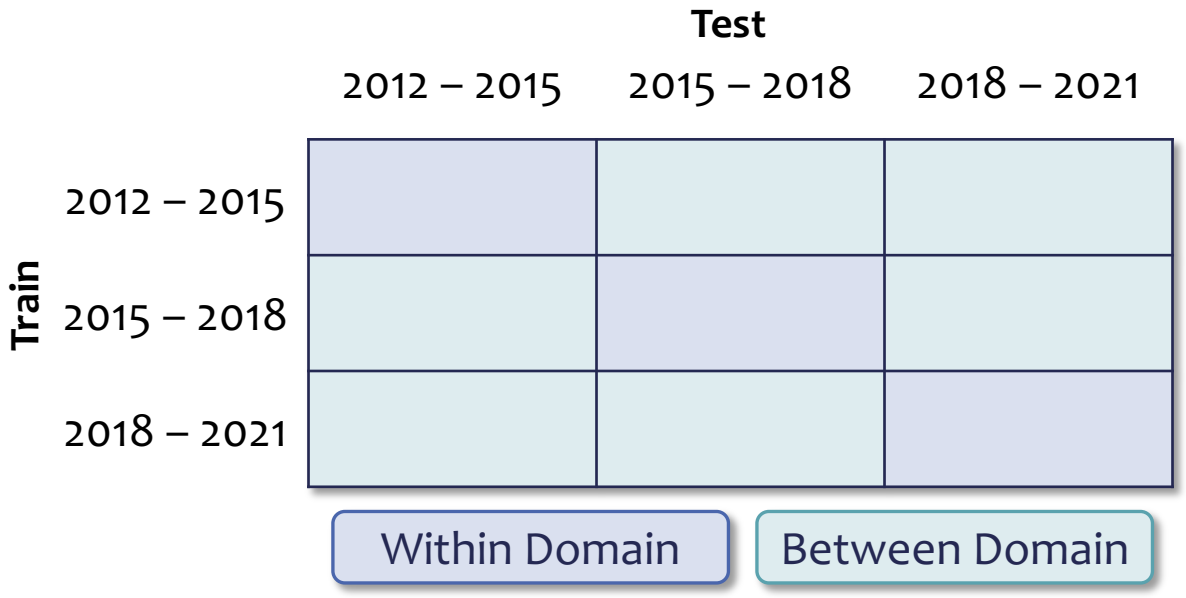
Re-establish identifier-to-label mapping by comparing timestamps and normalized text

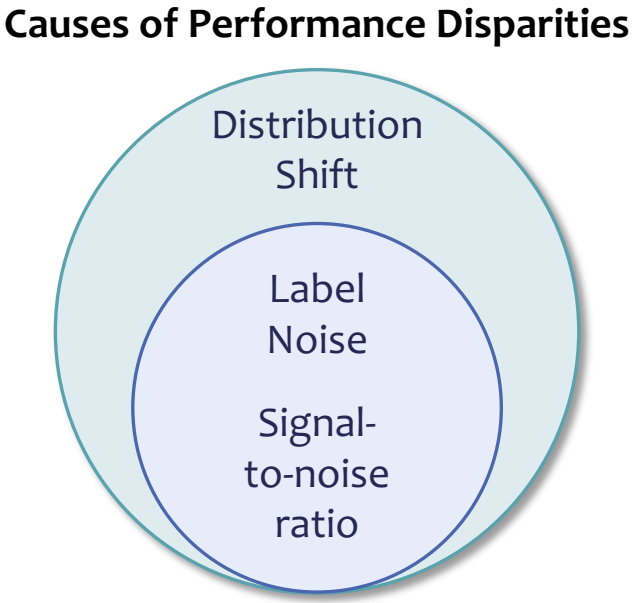


Dataset	Dates	# Individuals	# Posts
Original	2012 – 2015	D: 477 C: 872	D: 1,121,388 C: 1,907,508
Updated	2012 – 2021	D: 444 C: 172	D: 1,372,868 C: 546,826

INFERENCE UNDER LATENT DYNAMICS

How does predictive performance change when training a classifier on new data associated with an old label?





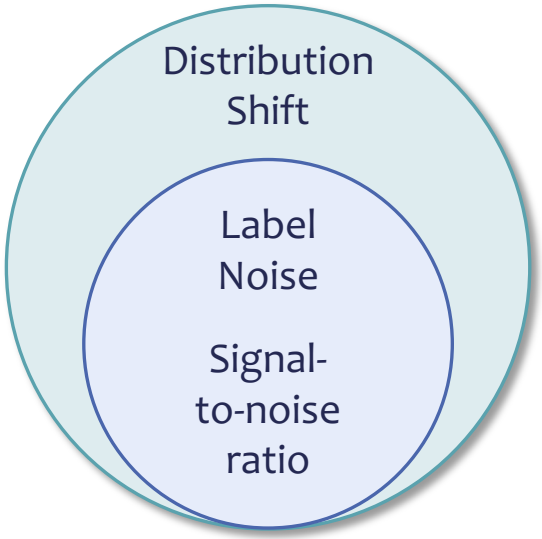
INFERENCE UNDER LATENT DYNAMICS

How does predictive performance change when training a classifier on new data associated with an old label?

Train \ Test	2012 – 2015	2015 – 2018	2018 – 2021
2012 – 2015	**0.71** (0.70, 0.72)		
2015 – 2018		0.66 (0.65, 0.66)	
2018 – 2021			0.68 (0.67, 0.69)

Within Domain | Between Domain

Causes of Performance Disparities



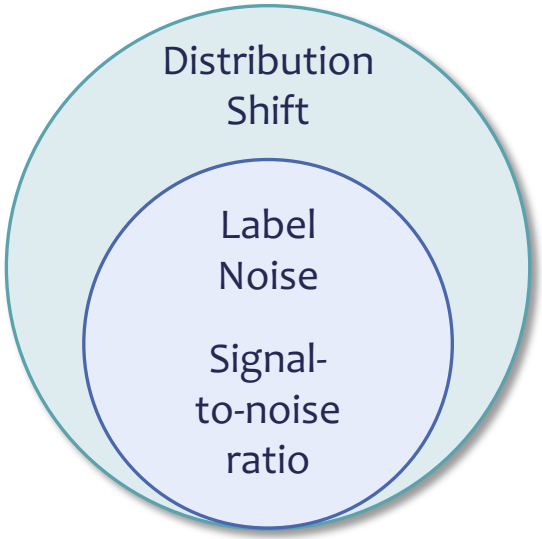
INFERENCE UNDER LATENT DYNAMICS

How does predictive performance change when training a classifier on new data associated with an old label?

Train \ Test	2012 – 2015	2015 – 2018	2018 – 2021
2012 – 2015	**0.71 (0.70, 0.72)**	0.66 (0.65, 0.66)	0.69 (0.68, 0.70)
2015 – 2018	0.66 (0.65, 0.67)	0.66 (0.65, 0.66)	0.68 (0.67, 0.69)
2018 – 2021	0.65 (0.65, 0.66)	0.67 (0.66, 0.68)	0.68 (0.67, 0.69)

Within Domain | Between Domain

Causes of Performance Disparities



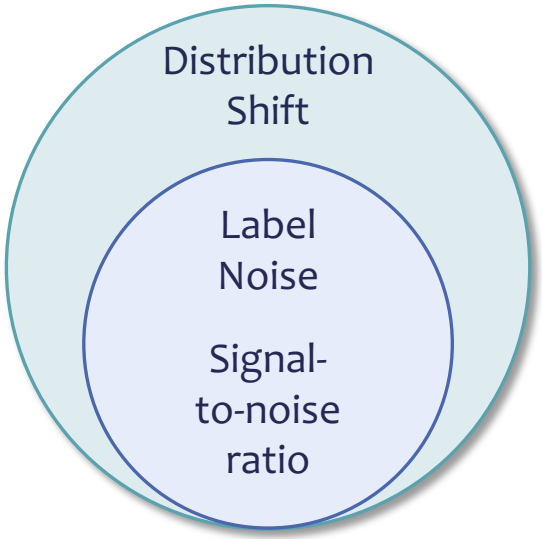
INFERENCE UNDER LATENT DYNAMICS

How does predictive performance change when training a classifier on new data associated with an old label?

Train \ Test	2012 – 2015	2015 – 2018	2018 – 2021
2012 – 2015	**0.71** (0.70, 0.72)	0.66 (0.65, 0.66)	0.69 (0.68, 0.70)
2015 – 2018	0.66 (0.65, 0.67)	0.66 (0.65, 0.66)	0.68 (0.67, 0.69)
2018 – 2021	0.65 (0.65, 0.66)	0.67 (0.66, 0.68)	0.68 (0.67, 0.69)

Within Domain | Between Domain

Causes of Performance Disparities



Has mental health status remained static? Or is there a spurious confound?

INTERPRETING MODEL PERFORMANCE

Train set debugging (Koh and Liang, 2017; Han et al., 2020)

Average influence of a tweet *x* on user-level inference

$$I(x) = \sum_{k=1}^{K} \underbrace{P_{k,\tau}(y = 1|X_\tau)}_{\text{Probability of depression given the document history } X} - \underbrace{P_{k,\tau}(y = 1|X_\tau^{\neg x})}_{\text{Probability of depression given } X \text{ without tweet } x}$$

INTERPRETING MODEL PERFORMANCE

Train set debugging (Koh and Liang, 2017; Han et al., 2020)

Average influence of a tweet *x* on user-level inference

$$I(x) = \sum_{k=1}^{K} P_{k,\tau}(y = 1|X_\tau) - P_{k,\tau}(y = 1|X_\tau^{\neg x})$$

$P_{k,\tau}(y = 1|X_\tau)$: Probability of depression given the document history *X*

$P_{k,\tau}(y = 1|X_\tau^{\neg x})$: Probability of depression given *X* without tweet *x*

An annotator is provided up to 30 tweets from each time period with highest $I(x)$

1. Indicate whether there is evidence of depression based on DSM-5 criteria and your prior knowledge regarding presentation of depression in social media
2. (If applicable) Indicate whether the depression appears to be in remission
3. Provide rational for your decision (e.g., which DSM-5 criteria, topical themes)

INTER-RATER RELIABILITY

Three Annotators (Author A_1; Non-authors B_1, B_2)

Several years experience modeling mental health within social media, but not clinical experts

Reliability Measures (Krippendorff α)

Evidence of Depression $\alpha = 0.4988$ (Fair to Moderate)

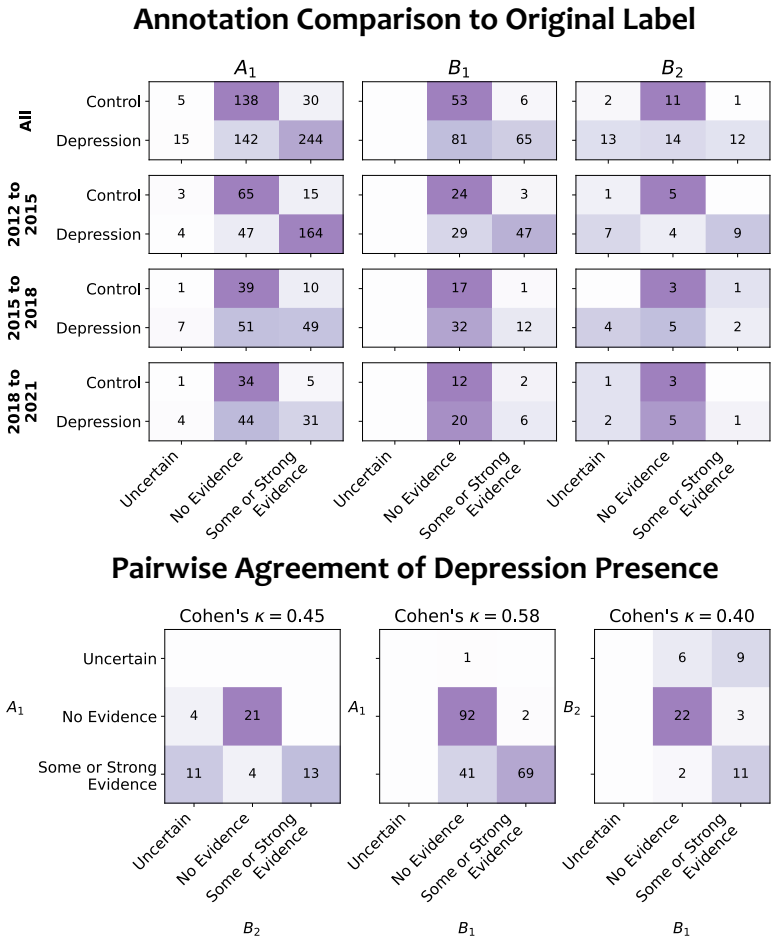
Remission Status $\alpha = 0.3561$ (Poor to Fair)

Causes of Disagreement

Prevalence to indicate uncertainty about label

Exposure bias

Sensitivity to depressed mood and/or negative emotion



QUANTITATIVE RESULTS

Decrease in evidence of depression

76% of individuals in original depression group during 2012-2015

45% and 39% of individuals during latter two time periods

	Dates	Total	Some Evidence	Strong Evidence	Remission
Control	2012 – 2015	83	15	3	1
	2015 – 2018	50	10	2	0
	2018 – 2021	40	5	0	0
Depression	2012 – 2015	215	164	136	10
	2015 – 2018	107	49	28	2
	2018 – 2021	79	31	16	1

Distribution of Annotations (A_1 only)

QUANTITATIVE RESULTS

Decrease in evidence of depression

76% of individuals in original depression group during 2012-2015

45% and 39% of individuals during latter two time periods

Label Noise

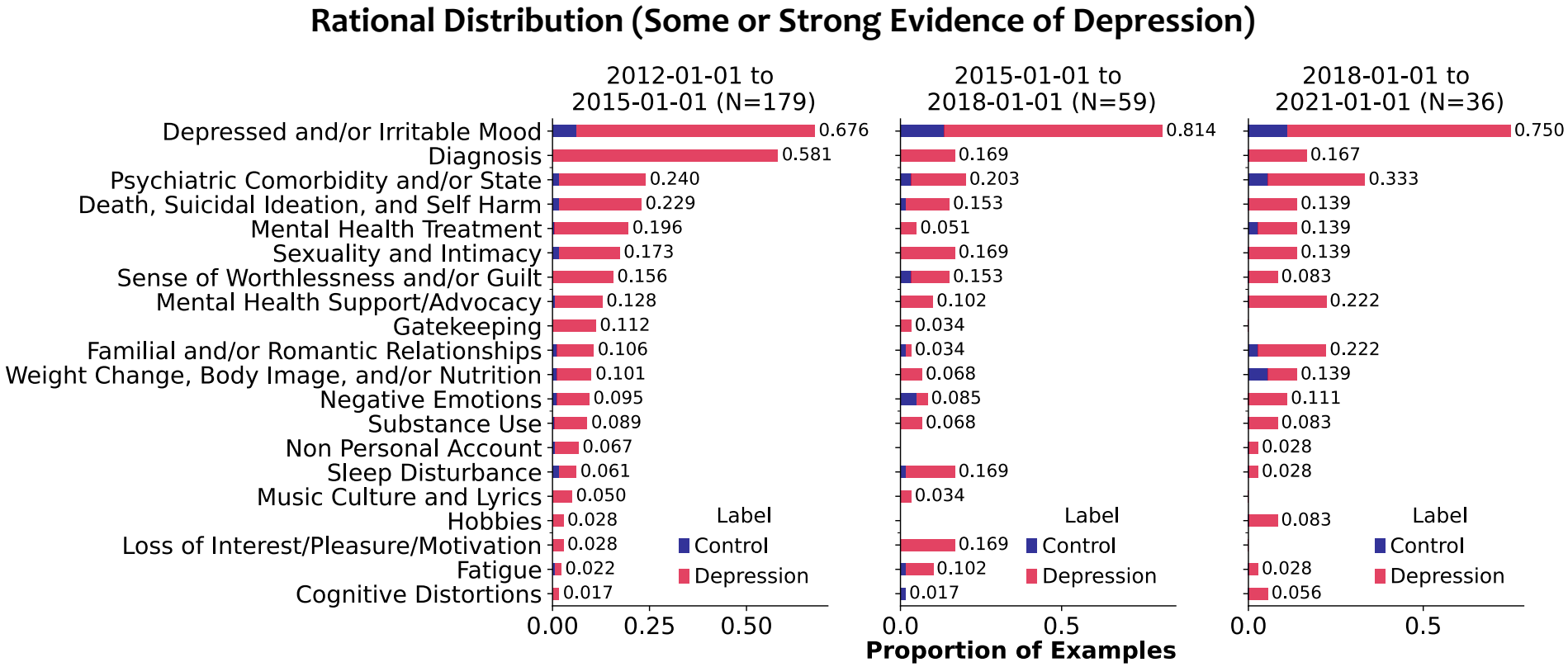
4% of Control group shows strong evidence of depression (Wolohan et al., 2018)

Non-zero proportion of individuals discussing *prior experience* with depression (remission)

	Dates	Total	Some Evidence	Strong Evidence	Remission
Control	2012 – 2015	83	15	3	1
	2015 – 2018	50	10	2	0
	2018 – 2021	40	5	0	0
Depression	2012 – 2015	215	164	136	10
	2015 – 2018	107	49	28	2
	2018 – 2021	79	31	16	1

Distribution of Annotations (A_1 only)

QUALITATIVE RESULTS



Themes: personality (e.g., elevated neuroticism), comorbid conditions, and a propensity for oversharing (e.g., taboo topics)

RECOMMENDATIONS

Individuals who disclose a mental health diagnosis systematically differ from the larger population of individuals living with that condition (Ernala et al., 2019)

We need to differentiate between:

1. Classifiers that estimate whether someone has a mental health condition
2. Classifiers that estimate whether someone has a mental health condition *AND* discloses their condition online

RECOMMENDATIONS

Individuals who disclose a mental health diagnosis systematically differ from the larger population of individuals living with that condition (Ernala et al., 2019)

We need to differentiate between:

1. Classifiers that estimate whether someone has a mental health condition
2. Classifiers that estimate whether someone has a mental health condition *AND* discloses their condition online

Annotate date of diagnosis and comorbidities

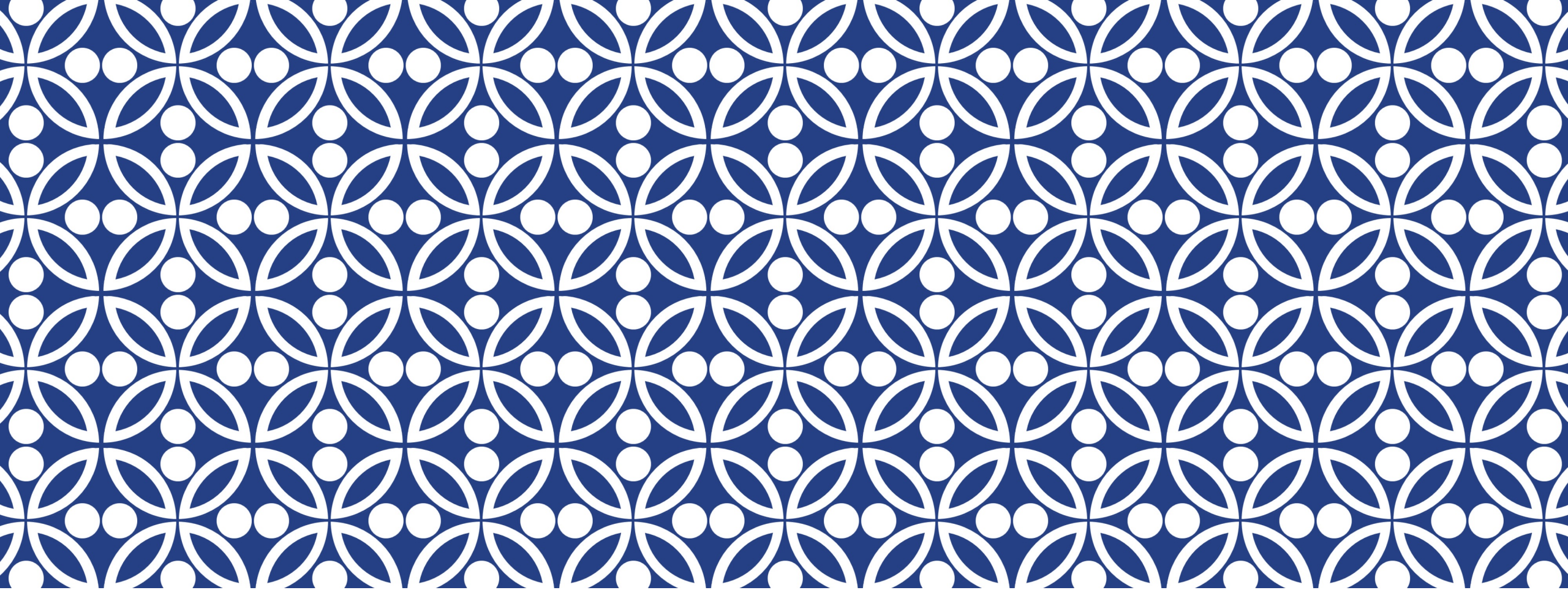
"1 year ago", "Depression and Eating Disorder"

Sample control groups using propensity score matching

Distribution of interests, personality, temporal activity

Identify and filter sample selection biases

Fan accounts, non-personal accounts



CONTACT

Keith Harrigian
https://kharrigian.github.io

Mark Dredze
https://www.cs.jhu.edu/~mdredze/

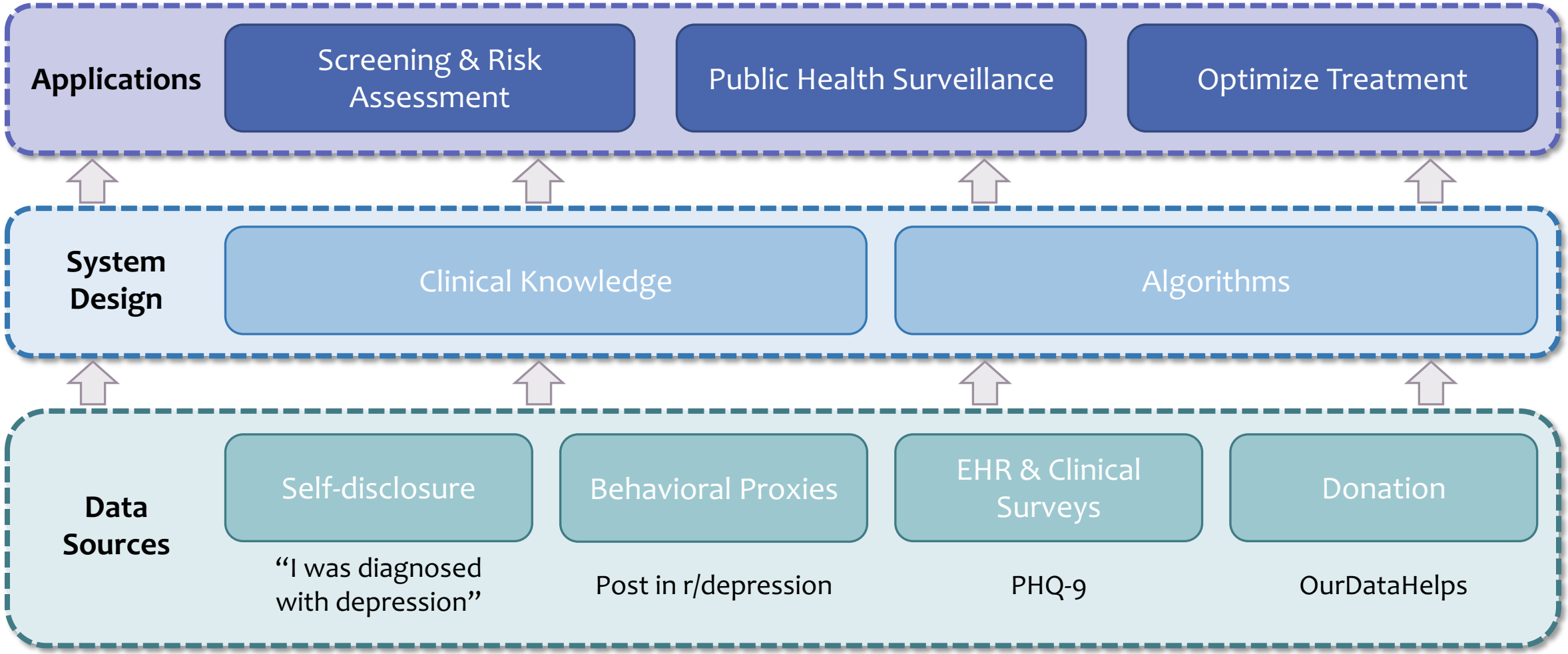
Exemplary tweets for each rational category identified during train-set debugging procedure

Evidence (Rational)	Exemplary Tweet
Diagnosis Disclosure	Bipolar disorder and depression. My doctor finally agrees.
Depressed & Irritable Mood	No one ever asks if I'm doing fine.
Loss of Interest/Pleasure/Motivation	…realizing you don't care about the things you used to enjoy
Weight, Body Image, & Nutrition	Not that anyone cares, but I'm almost at my goal weight.
Sleep Disturbance	I CANT SLEEP. PAIN. JUST LIKE ALWAYS.
Fatigue	mentally drained from this pandemic.
Sense of Worthlessness & Guilt	When you let someone do anything to you…
Impaired Thought	I'm failing my classes because I'm depressed.
Death & Self Harm	My scars are faded… unless you care to look close.
Cognitive Distortions	I always think my bf is going to leave me
Treatment	Scared to tell a women that I'm in therapy.
Gatekeeping	depression isn't just a bad day. fuck you all.
Sexuality & Intimacy	Who wants to come take some pics of me for only fans?
Negative Emotions	I feel like no one cares even though I know they do
Coping Strategies	Art is always the easiest way to distract me from my anxiety
Psychiatric Comorbidity & State	I am anorexic and I cut myself
Non-psychiatric Comorbidity	Could use a little bit of aid #DisabilityAid
Substance Use	Weed makes the dreams go away and that's a good thing
Support & Advocacy	RIP Chester. If you're going through pain, please reach out to me.
Personality & Identity	Lol grandma still think I'm bringing a boy home
Music Culture & Lyrics	#FallingInReverse :D
Familial/Romantic Relationships	Mom: You'll never lose weight. Me: is that why dad left?
Hobbies	Missin the old days when everyone played Pokemon yellow
Non-personal Accounts	My life was about to fall apart until I found the Calm app…

Diagnosis Disclosure

DSM-5 Criteria

Empirical Themes


Applications
Screening & Risk Assessment
Public Health Surveillance
Optimize Treatment
System Design
Clinical Knowledge
Algorithms
Data Sources
Self-disclosure
Behavioral Proxies
EHR & Clinical Surveys
Donation
"I was diagnosed with depression"
Post in r/depression
PHQ-9
OurDataHelps