2016 CEGS N-GRID

Shared-Tasks and Workshop on Challenges in Natural Language Processing for Clinical Data

Michele Filannino, Amber Stubbs, Özlem Uzuner

https://www.i2b2.org/NLP/RDoCforPsychiatry/

11th November 2016
Chicago, IL

SUNY HARVARD MEDICAL SCHOOL MGH 1811 PARTNERS HEALTHCARE N-GRID

track 2: symptom severity classification

- RDoC: framework for studying mental disorders
  - integrates many levels of information (from genomics to self-report) to understand the basic dimensions of human behavior (from normal to abnormal)
  - 5 domains: **POSITIVE VALENCE***, NEGATIVE VALENCE, COGNITIVE, SOCIAL PROCESSES, AROUSAL AND REGULATORY SYSTEMS
  - how good systems are at predicting patients' symptom severity, based on initial psychiatric evaluation records?


Subject: Patient Initial Visit Note -Identifying Information Date of Service:

09/14/2067CPT Code: 90792: With medical services
Age:

43Sex: Female
Interpreter Used: None needed
Chief Complaint / HPI Chief Complaint (Patients own words)

Valentina is a 43-year old female with a past psychiatric history significant for an underlying depressive disorder, anxiety disorder, binge eating disorder, no history of prior inpatient psychiatric hospitalizations, no history of prior suicide attempts, [illegible] known history of self-injurious behavior, no history of EtOH or illicit-substance use and a past medical history significant for DVT (8-years ago while on OCP) who presents [illegible] for continued psychopharmacologic care following her transfer from Dr. Yvonne Ellison's practice. Per Valentina, "I just need to continue these medications...they have been so helpful."

History of Present Illness and Precipitating Events

Valentina describes that she first presented for binge-eating disorder treatment approximately one-year prior to current presentation, when she [illegible] that she was bingeing on a daily basis in the context of a number of stressors/[illegible], including the birth of her two children within a year of each other, marriage to her [illegible] friend at the age of 39, and ensuing conflicts with her mother as a result of the [illegible] that at that time, she binged daily, which steadily led to a marked increase in a [illegible] gain within 4-months. She states that she would binge in secret from her husband and two daughters and states that following each binge she felt "awful...ashamed." As a result, she states that her mood started to precipitously decline and states that she ultimately realized that she "could no longer fun from this...I had to seek treatment."

In an effort to seek treatment, she began seeing dieticia[illegible] in Hubert, who she states was "helpful, but not enough." She subsequently [illegible] and worked through the exercises in the book and states that that, [illegible]. She ultimately contacted one of the authors of the book and a[illegible] withher via Skype. She states that that was helpful for approxima[illegible] me her therapist suggested that she seek local ED-specific care. As such, [illegible] went an evaluation with Dr. Deon Yarbrough 6/2066 and was subsequently referred to Dr. Tonya Alston for a 20-week course of CBT-E. Valentina states that her treatment with Dr. Alston was "phenomenal...it was exactly what I needed." She concurrently began seeing Dr. Yvonne Ellison for psychopharmacologic treatment of her underlying depressive/anxiety disorder and BED. She has since responded well to her current regimen of Zoloft, Vyvanse, and Topamax.

* Responses to positive motivational situations or contexts, such as reward seeking, consummatory behavior, and reward/habit learning (alcohol, drink, abuse, repetitive, stereotypy, drug, gamble, count, craving, compulsive)

research domain criteria

RDoC
Research Domain Criteria Initiative

Construct/Subconstruct		Genes	Molecules	Cells	Circuits	Physiology	Behavior	Self-Report	Paradigms
Approach Motivation	Reward Valuation		Elements		Elements			Elements	Elements
	Effort Valuation / Willingness to Work		Elements		Elen			Elements	Elements
	Expectancy / Reward Prediction Error		Elements		Elements	Elements	Elements	Elements	Elements
	Action Selection / Preference-Based Decision Making				Elements				Elements
Initial Responsiveness to Reward Attainment		Elements	Elements		Elements		Elements	Elements	Elements
Sustained/Longer-Term Responsiveness to Reward Attainment			Elements		Elements	Elements	Elements	Elements	Elements
Reward Learning		Elements	Elements	Elements	Eleme		lements	Elements	Elements
Habit		Elements	Elements	Elements	Elements		Elements	Elements	Elements

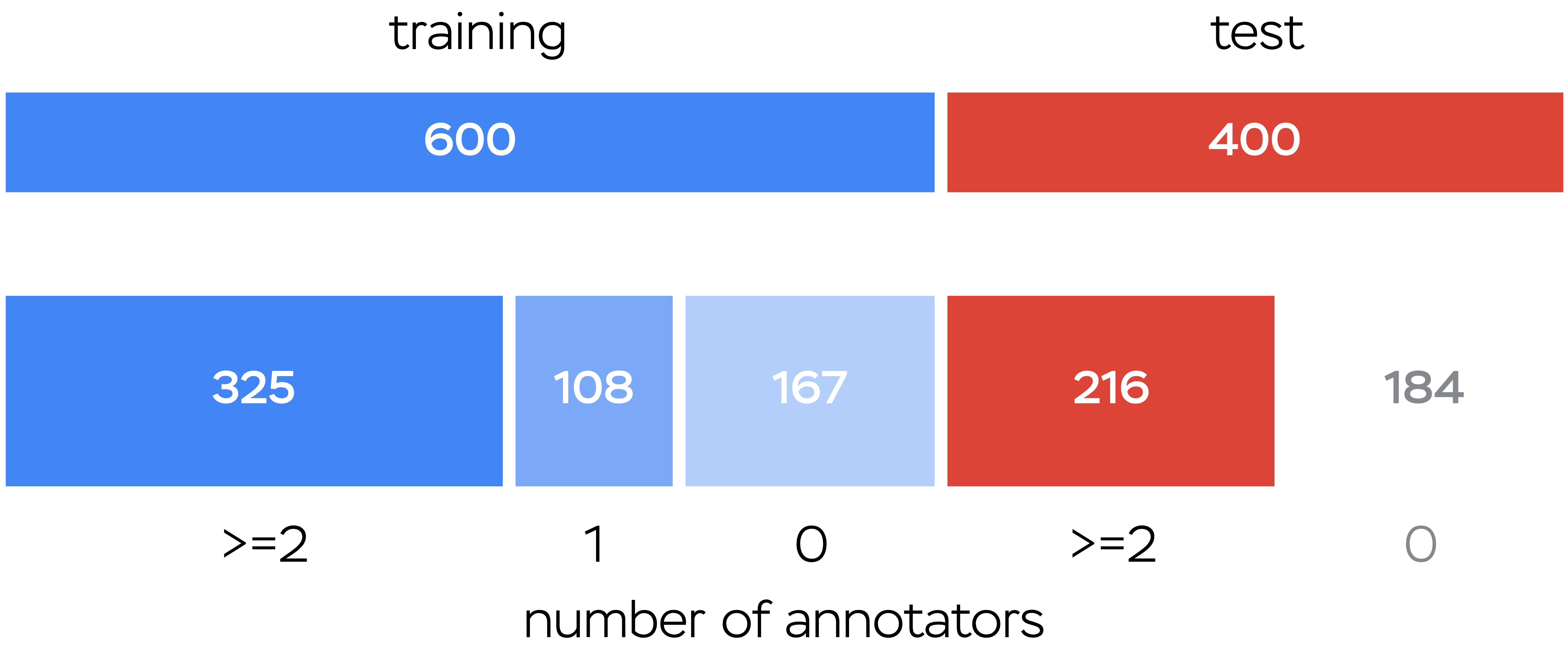
Goal tracking
Pavlovian approach
Reward-related speeding
Sign tracking

Dopamine,
Serotonine

DAT, DR2,
TREK1

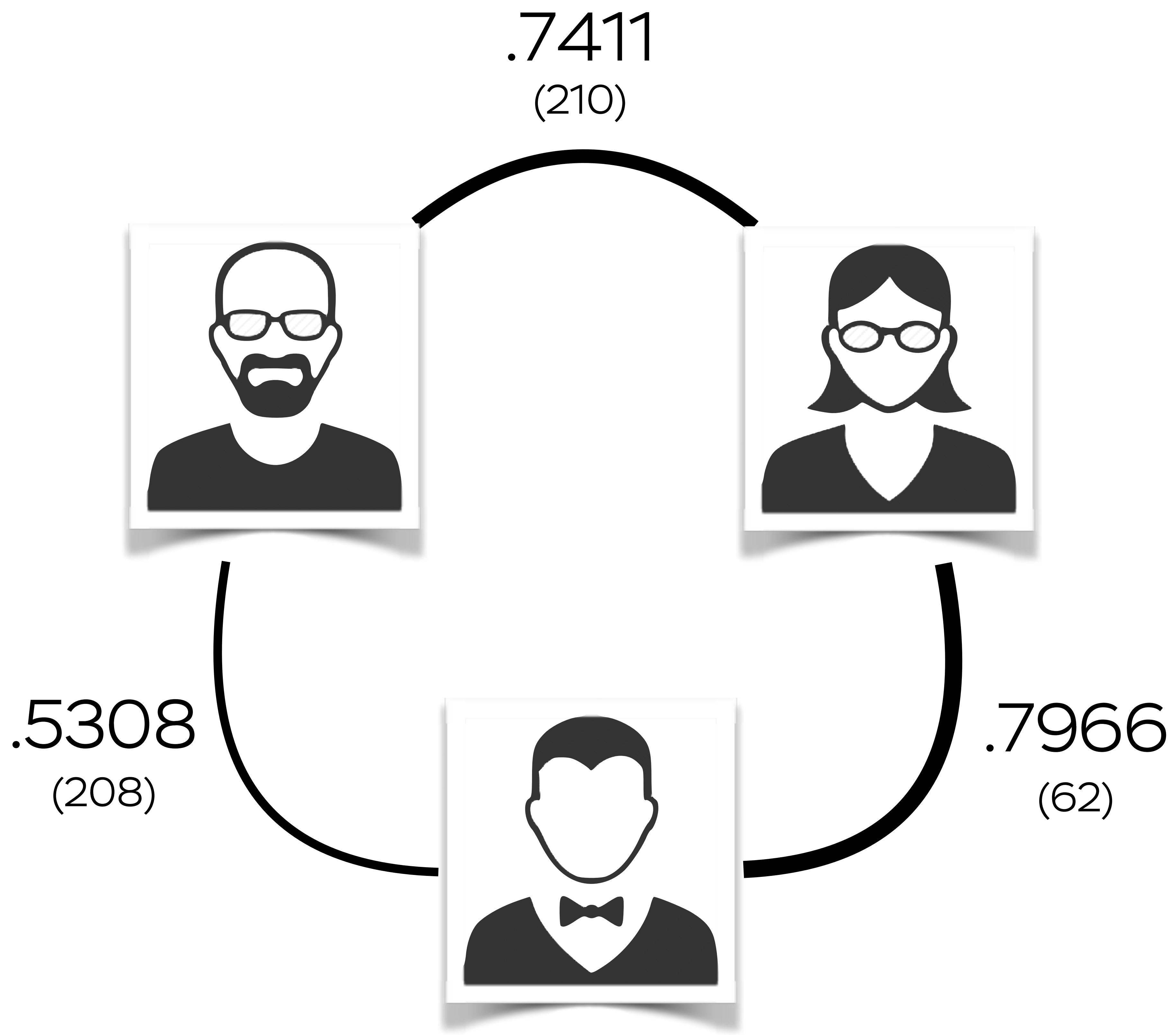
Compulsive behaviors
Repetitive behaviors
Stereotypic behaviors

corpus

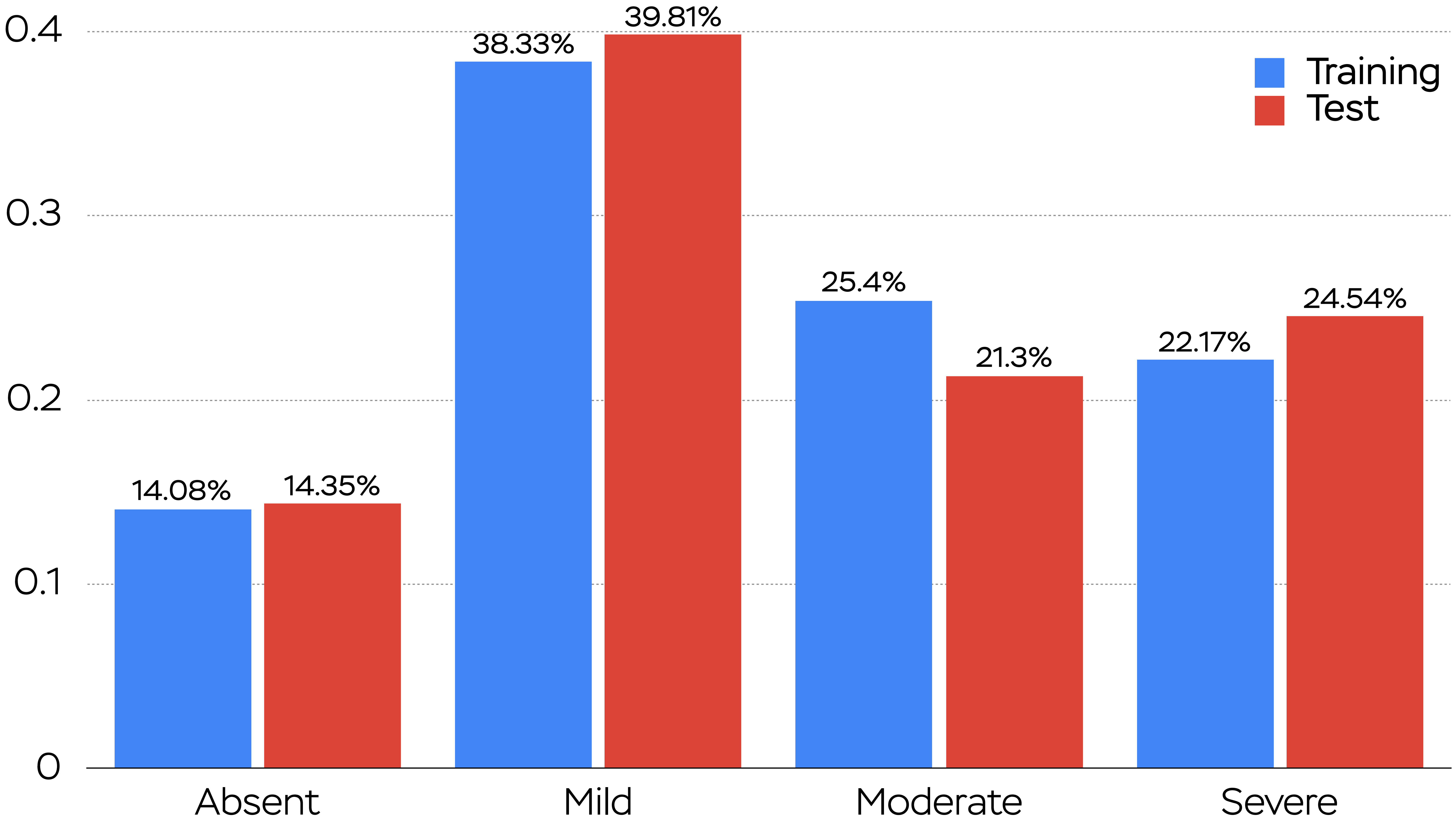

training
test
600
400
325
108
167
216
184
>=2
1
0
>=2
0
number of annotators

annotation process

- 3 expert psychiatrists with several years of experience
- MGH and Harvard Med. Sc.
- 2 annotation:
  - tie-broken by the 3rd
    - adjudicated by the most experienced
- 1 annotation:
  - the most experienced



distribution of classes



track 2: performance measures

- Nominal Classification measures:
  - Precision, Recall, F1-measure, Accuracy, Cohen's Kappa coefficient, Scott's Pi coefficient
- Ordinal/Interval Classification measures:
  - Median Absolute Error, Mean Absolute Error, Mean Squared Error
- Continuous Regression measures:
  - R^2 coefficient, Pearson's correlation coefficient

- **ranking**
  - Macro-averaged Mean Absolute Error

MAEμ

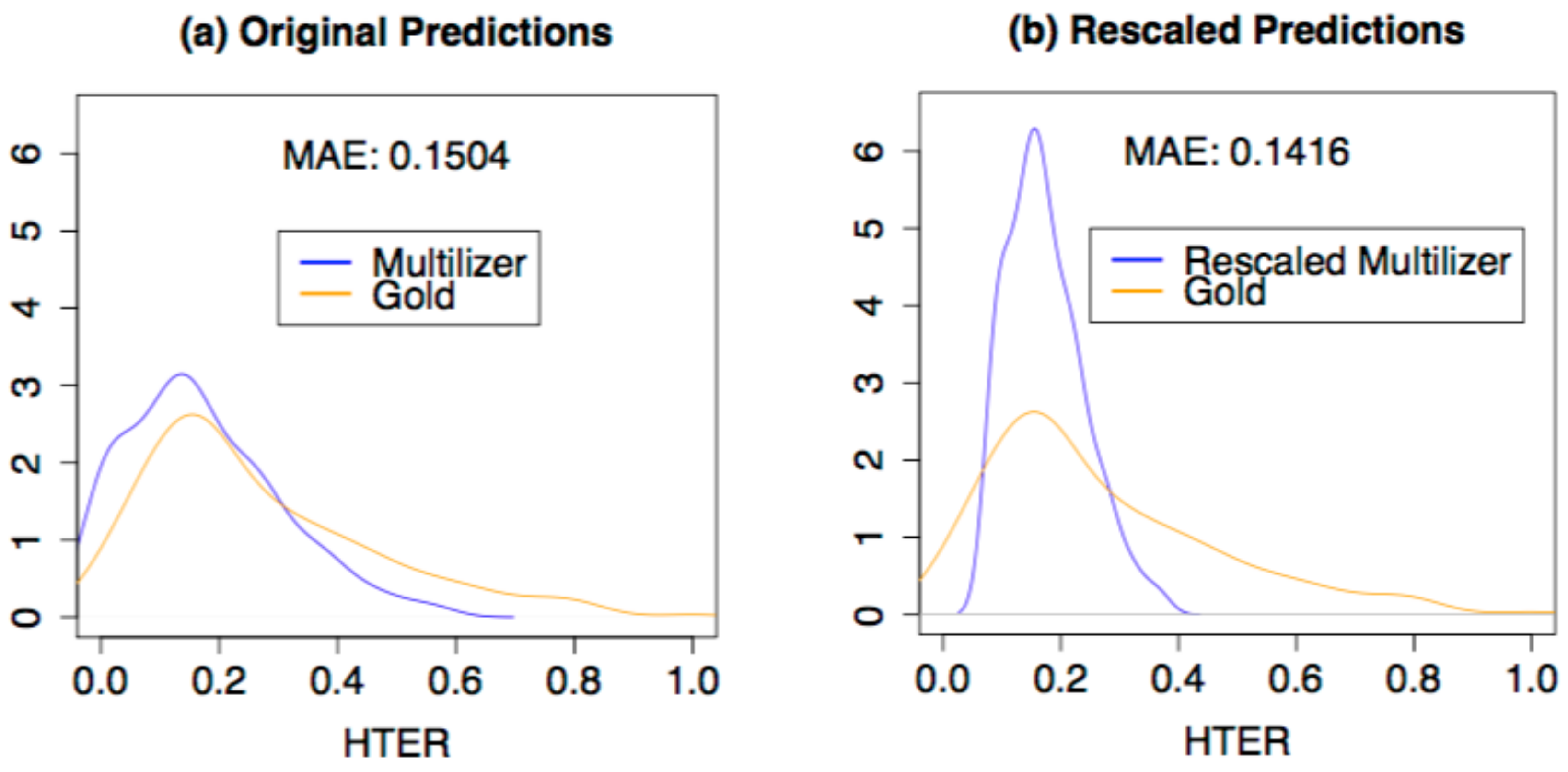
test docs

prediction

gold

$$MAE^{\mu}(h,Te)=\frac{1}{|Te|}\sum_{\mathbf{x}_i\in Te}|h(\mathbf{x}_i)-y_i|$$

predictions can be **adjusted** by guessing the central tendency!



1. Baccianella, S., Esuli, A., & Sebastiani, F. (2009). Evaluation Measures for Ordinal Regression. In 2009 Ninth International Conference on Intelligent Systems Design and Applications (pp. 283–287). IEEE. http://doi.org/10.1109/ISDA.2009.230
2. Graham, Y., 2015. Improving evaluation of machine translation quality estimation. In 53rd Annual Meeting of the Association for Computational Linguistics and Seventh International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (pp. 1804-1813).
3. Moreau, E. and Vogel, C., 2014, August. Limitations of MT Quality Estimation Supervised Systems: The Tails Prediction Problem. In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics (pp. 2205-2216). Dublin City University and Association for Computational Linguistics.
4. Celli, F., Pianesi, F., Stillwell, D. and Kosinski, M., 2013, June. Workshop on computational personality recognition (shared task). In Proceedings of the Workshop on Computational Personality Recognition.
5. Machácek, M. and Bojar, O., 2014, June. Results of the WMT14 metrics shared task. In Proceedings of the Ninth Workshop on Statistical Machine Translation (pp. 293-301).
6. P. Nakov and A. Ritter and S. Rosenthal and V. Stoyanov and F. Sebastiani (2015). SemEval-2016 Task 4: Sentiment Analysis in Twitter. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016).

MAE^M

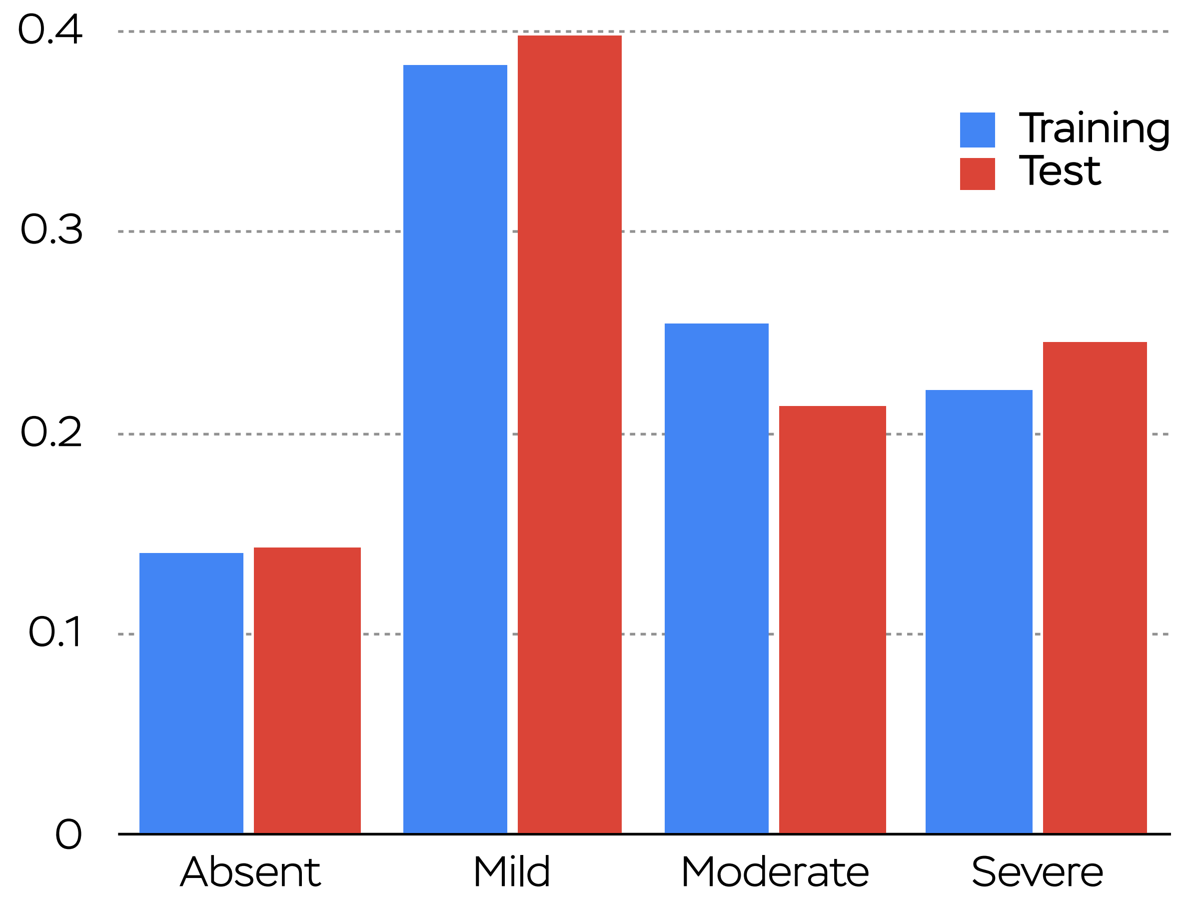
test docs with class j

class

prediction

gold

$$MAE^M(h, Te) = \frac{1}{|\mathcal{C}|} \sum_{j=1}^{|\mathcal{C}|} \frac{1}{|Te_j|} \sum_{\mathbf{x}_i \in Te_j} |h(\mathbf{x}_i) - y_i|$$



- copes with imbalanced data
- the under-represented classes counts as any other class, rather than proportionally to their frequencies
- normalized by maximum error

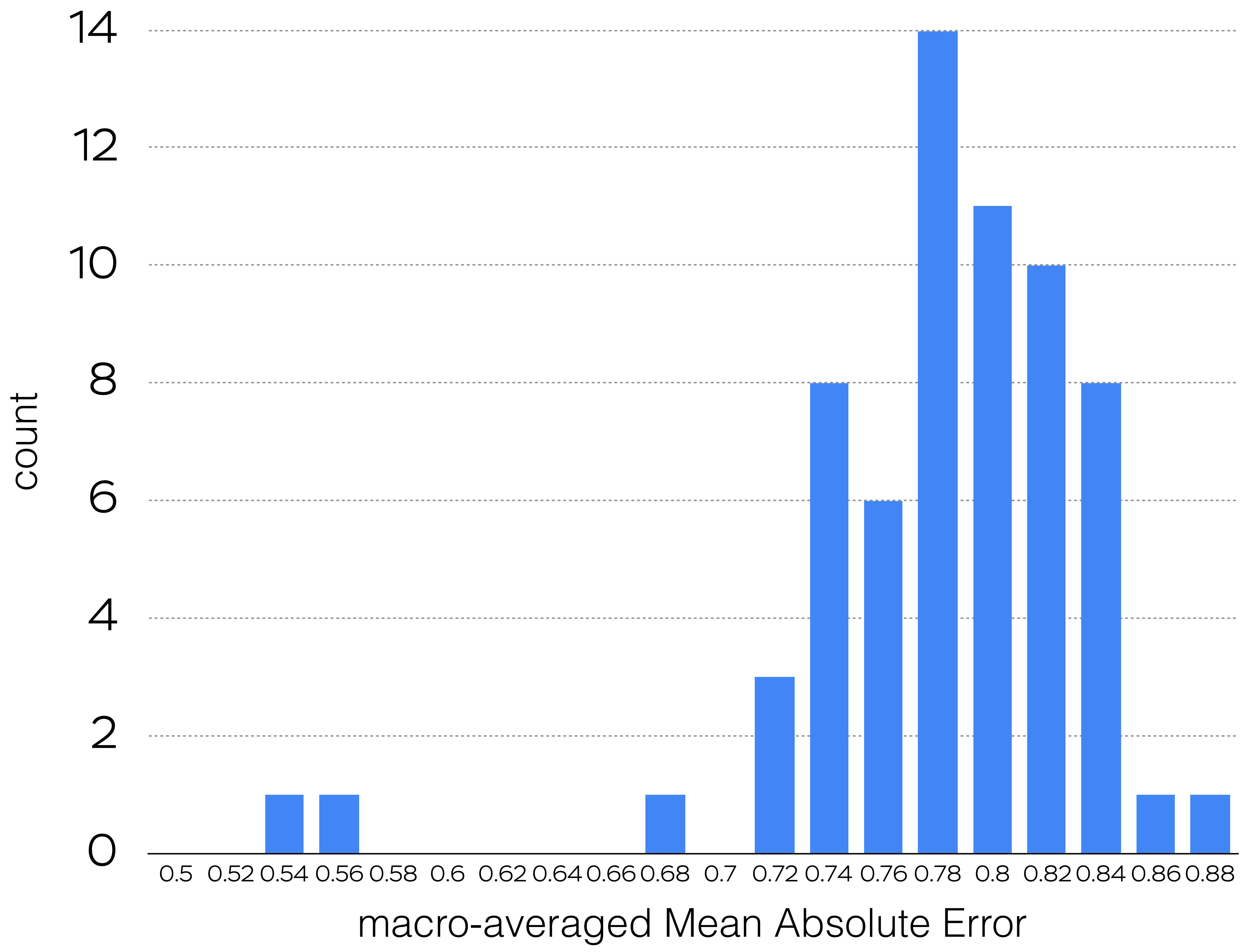
participation

- 15 countries
- 50 teams
- 65 institutions
- 154 researchers

track 2

- 11 countries
- 24 teams
- 42 institutions
- 110 researchers
- 65 submitted runs

general results (all runs)



Min:
0.524597

Max:
0.863019

Average:
0.771492

Median:
0.775882

Standard Deviation:
0.056080

TOP 10 (best runs only)

Rank	Institutions	Score
1	SentiMetrix Inc.	0.863019
2	The University of Texas at Dallas	0.840963
3	University of Kentucky	0.838615
4	University of Pittsburgh	0.825594
5	Med Data Quest Inc.	0.817474
6	Harbin Institute of Technology Shenzhen Graduate	0.816844
7	University of Minnesota	0.814971
8	Antwerp University Hospital	0.806356
9	LIMSI-CNRS	0.801738
10	The University of Manchester	0.801143

24 teams -> 24 best runs

TOP 10 (all runs)

Rank	Institutions (# run)	Score
1	SentiMetrix Inc. (#3)	0.863019
2	The University of Texas at Dallas (#3)	0.840963
3	University of Kentucky (#3)	0.838615
4	University of Kentucky (#1)	0.837284
5	SentiMetrix Inc. (#1)	0.836503
6	University of Kentucky (#2)	0.835138
7	SentiMetrix Inc. (#2)	0.833281
8	University of Pittsburgh (#3)	0.825594
9	The University of Texas at Dallas (#2)	0.824262
10	University of Pittsburgh (#2)	0.821807

Organizing committee:
Ozlem Uzuner, co-chair, SUNY at Albany
Amber Stubbs, co-chair, Simmons College
Michele Filannino, co-chair, SUNY at Albany

Tianxi Cai, Harvard School of Public Health
Susanne Churchill, Harvard Medical School
Isaac Kohane, Harvard Medical School
Thomas H. McCoy, MGH, Harvard
Roy H. Perlis, MGH, Harvard
Peter Szolovits, MIT
Uma Vaidyanathan, NIMH
Philip Wang, American Psychiatric Association

thanks!

SUNY HARVARD MEDICAL SCHOOL MGH 1811 PARTNERS HEALTHCARE N-GRID

2016 CEGS N-GRID Discussion

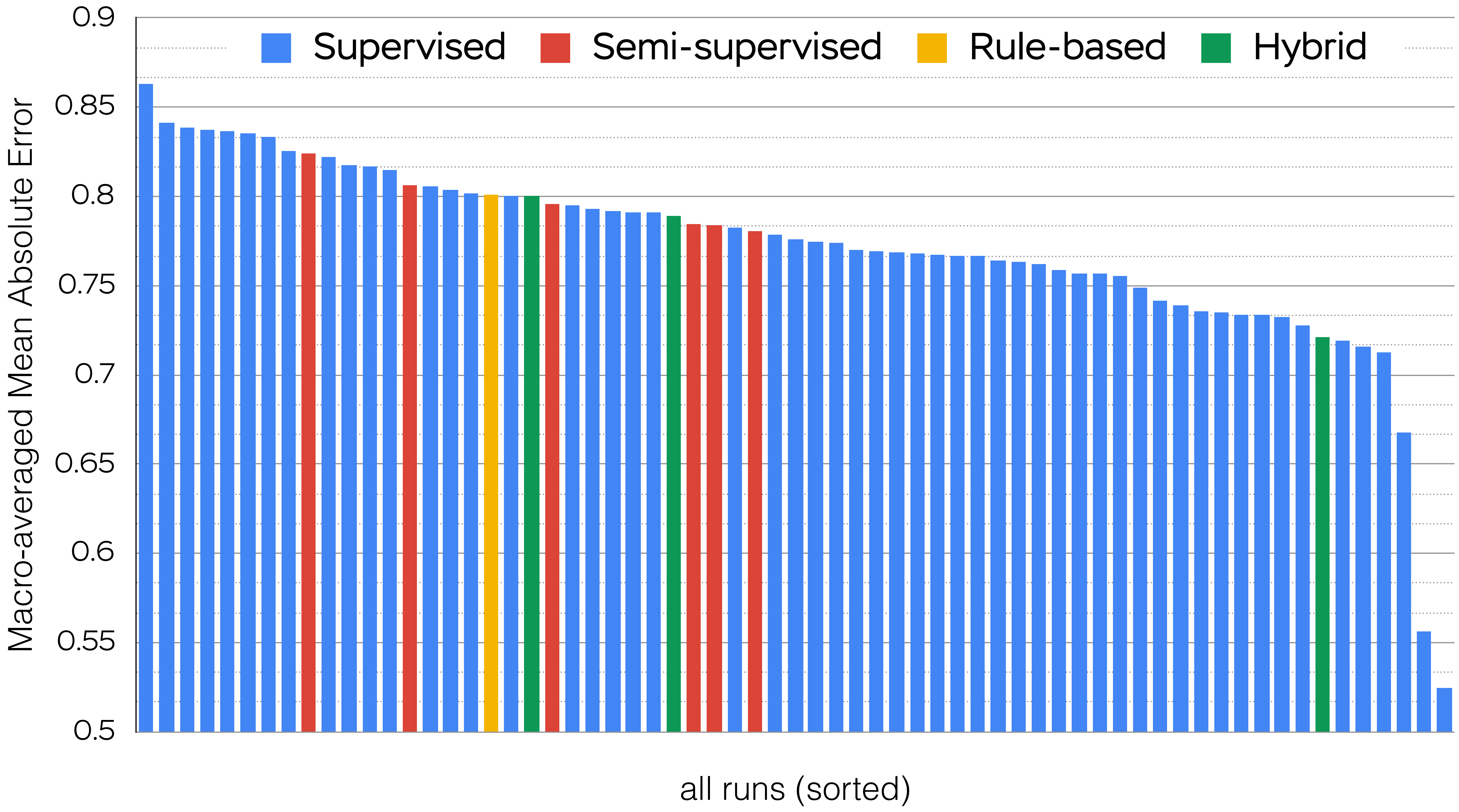
Michele Filannino, Amber Stubbs, Özlem Uzuner

https://www.i2b2.org/NLP/RDoCforPsychiatry/

11th November 2016
Chicago, IL

SUNY · Simmons College · MIT · Harvard Medical School · MGH · Partners HealthCare · N-GRID

methods

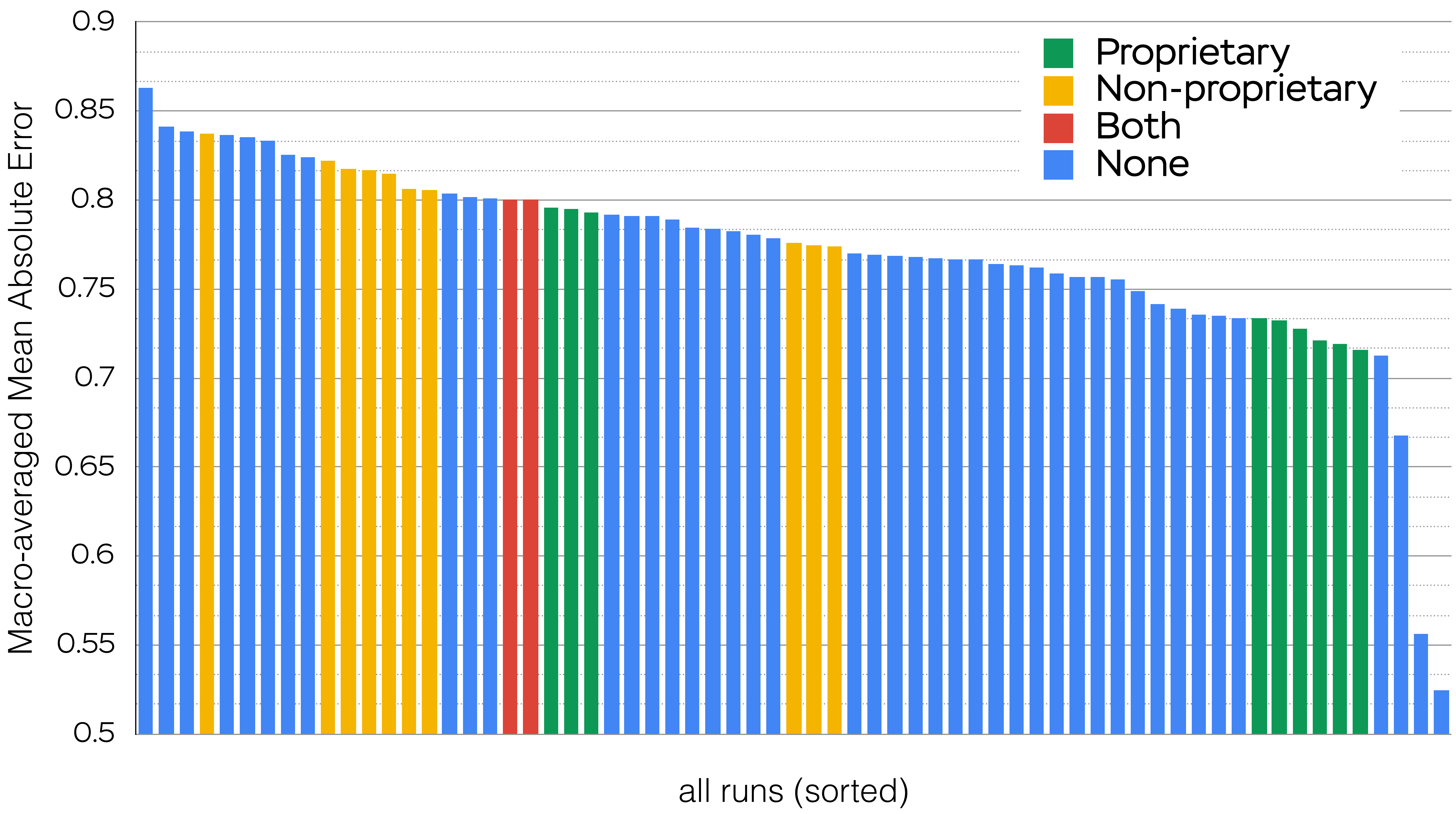


Supervised
Semi-supervised
Rule-based
Hybrid
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
Macro-averaged Mean Absolute Error
all runs (sorted)

learning techniques

- Supervised:
  - SVM, neural networks, conv. NN, Random Forest, Logistic Regression, SVR, Naïve Bayes, Bayesian nets, Gradient Tree Boosting, Log. Ordinal Regression
- Semi-supervised:
  - self-training
- Unsupervised:
  - embeddings, brown clustering, skip-grams
- Rule-based:
  - hand-crafted rules, association rules
- Hybrid:
  - hand-crafted rules + neural network

resources

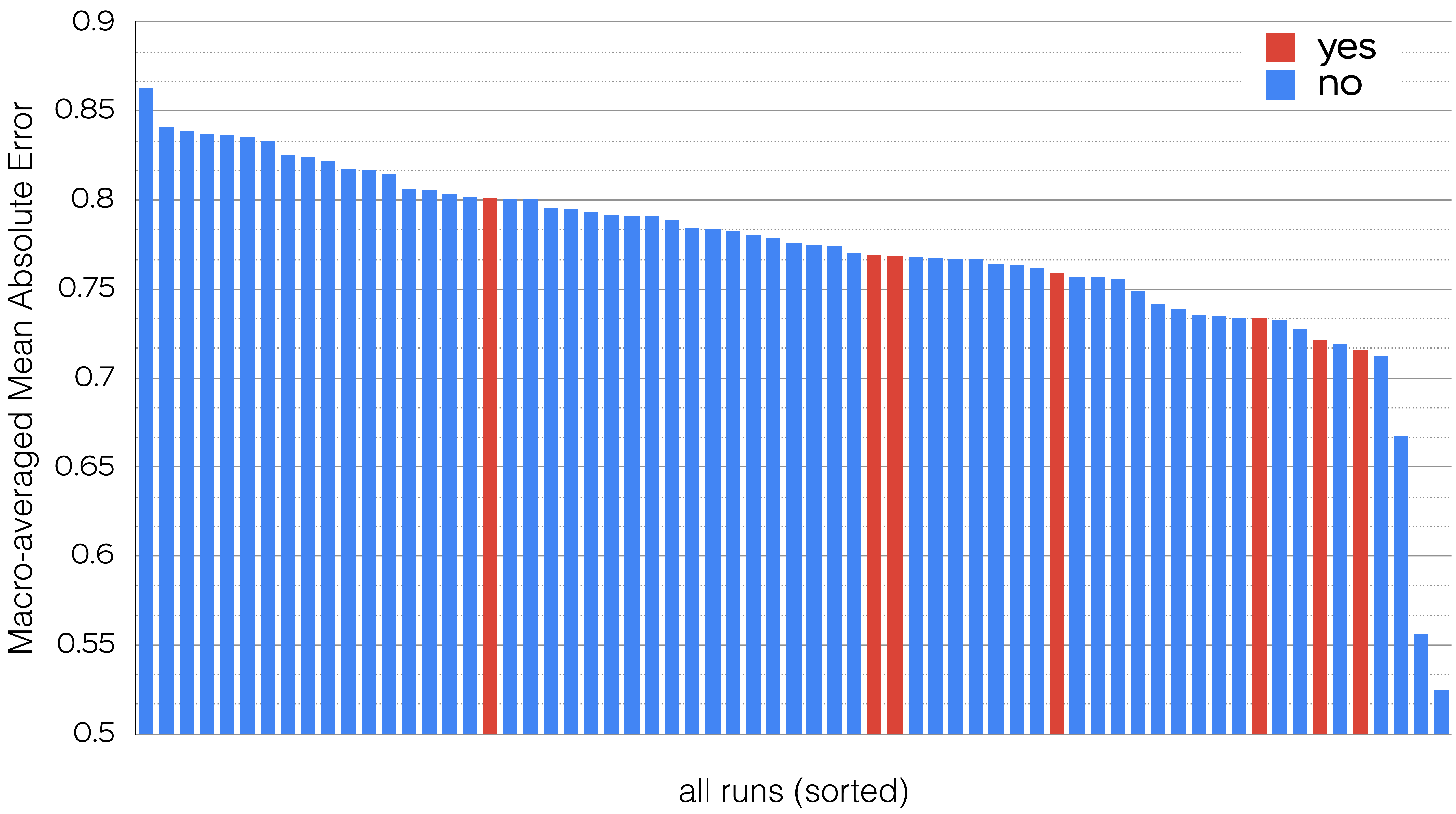


Macro-averaged Mean Absolute Error
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
Proprietary
Non-proprietary
Both
None
all runs (sorted)

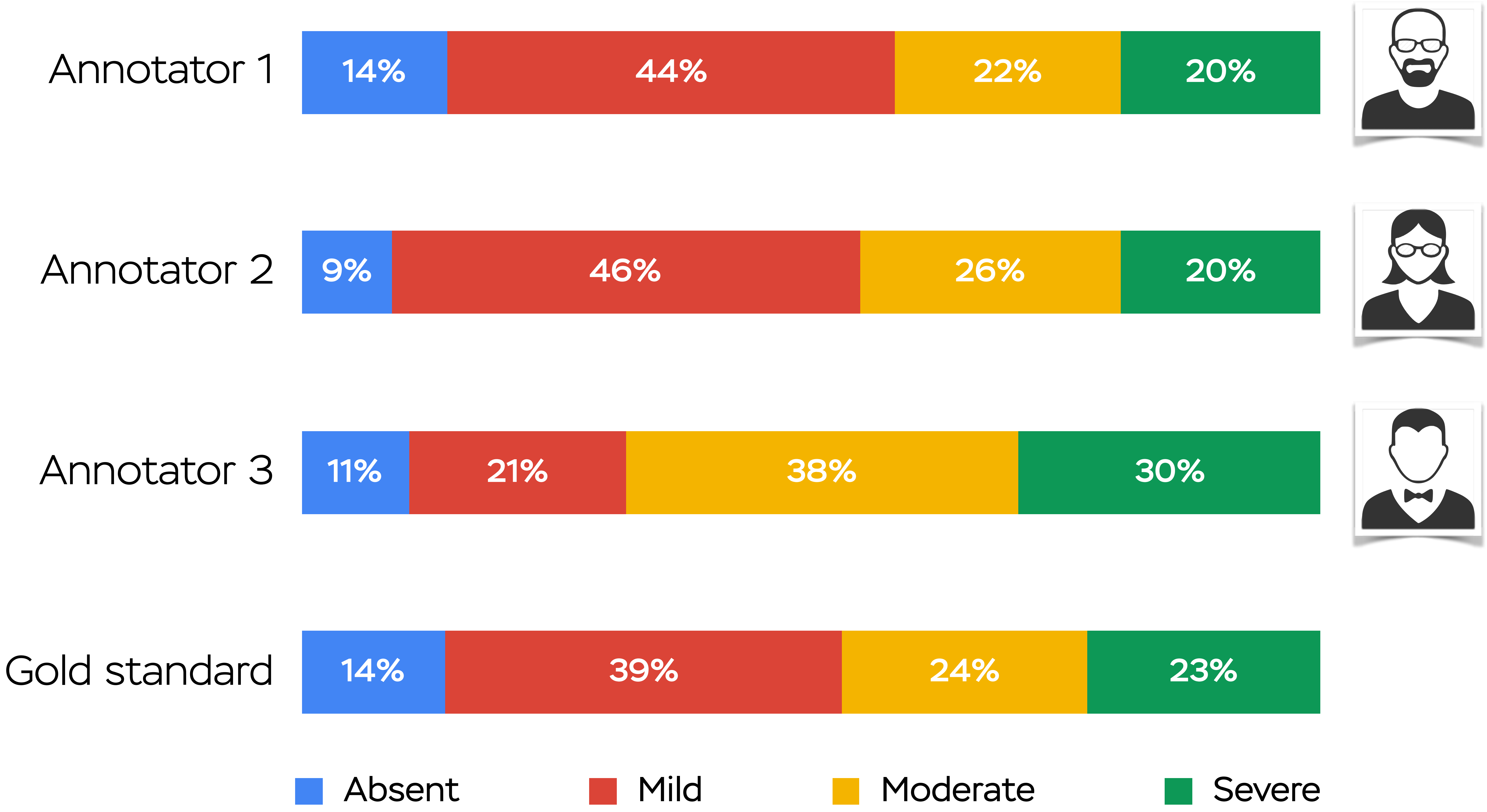
resources

- Pre-processing & Feature extraction:
  - OpenNLP, MetaMap, NLTK, cTAKES, MedLEE, LingScope, GENIA, DSM Ontology, SentEmotion
- Corpora:
  - Gigaword, PubMed Central
- Machine learning:
  - scikit-learn, scipy, Weka, XGBoost, Mandolin, liblinear, word2vec, libSVM

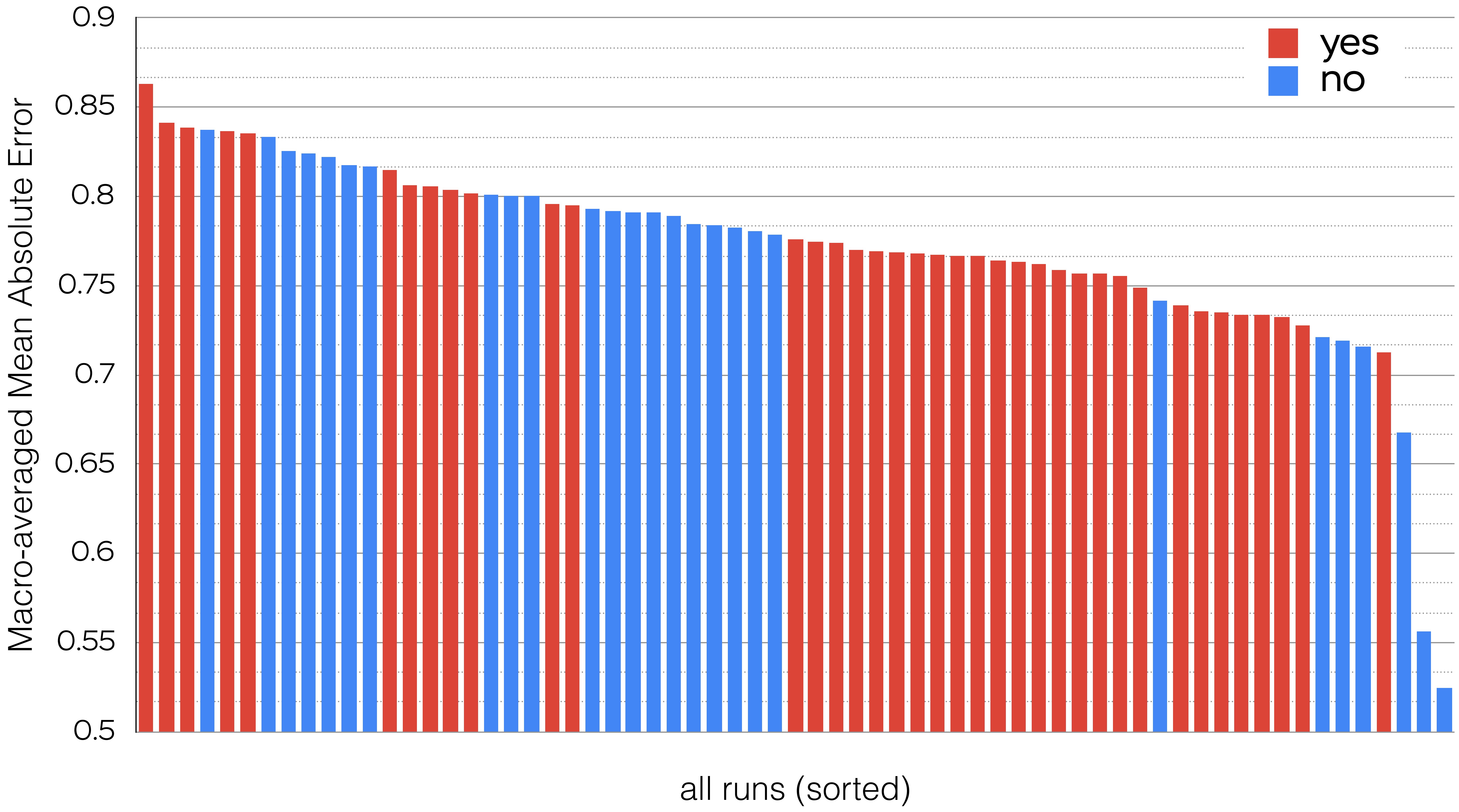
medical experts involvement



annotators vs gold



used data: 1 annotator



errors (TOP 5 best runs)

gold labels ↓	1st best run				2nd best run				3rd best run				4th best run				5th best run			
	abs	mild	mod	sevr	abs	mild	mod	sevr	abs	mild	mod	sevr	abs	mild	mod	sevr	abs	mild	mod	sevr
Absent	**21**	10	0	0	**17**	10	4	0	**23**	8	0	0	**23**	6	1	1	**20**	9	2	0
Mild	7	64	13	2	3	60	20	3	8	72	6	0	9	52	23	2	5	65	12	4
Moderate	2	12	29	3	1	9	27	9	4	12	28	2	2	12	29	3	4	12	27	3
Severe	0	3	9	41	0	1	7	45	0	7	24	22	0	4	18	31	0	5	21	27

errors$^{\text{summed-up}}$ (TOP 5 best runs)

gold labels ↓	absent	mild	moder	severe
Absent	104	43	7	1
Mild	32	313	74	11
Moderate	13	57	140	20
Severe	0	20	79	166

scores' meaning

- ⋆ **None**: no evidence of these symptoms at any time
- ⋆ **Mild**: some evidence of these symptoms but never the focus of treatment
- ⋆ **Moderate**: symptoms sufficient to be focus of outpatient treatment (intensive outpatient program, maintenance, operating under the influence of substances)
- ⋆ **Severe**: symptoms sufficient to warrant inpatient treatment/hospitalization now or at some point (symptoms of withdrawal, blackouts from alcohol)

depression

- ★ A patient who is depressed can have a null severity level (0).
- ★ A patient who is depressed and has decreased interest scores at least mild (1).
- ★ Examples: people with major depressive disorder
- ★ A patient who is depressed and has decreased interest and that is the main focus of treatment scores at least moderate (2).
- ★ Examples: people who need an intervention to get out of bed
- ★ A patient who is depressed and needs hospitalization/ electroconvulsive therapy scores severe (3)

addiction

- A patient who is violent and has a history of substances abuse scores at least moderate (2).
- A patient who smokes cigarettes scores at least mild (1).
- A patient who uses alcohol (EtOH) or street drugs scores at least moderate (2).
- A patient who uses marijuana (MJ):
  - occasional use not focus of a treatment = mild (1)
  - specific focus of a treatment = moderate (2)
  - inpatient (or could have been) = severe (3)
- A patient who has legal troubles (driving under the influence, arrest)/ intensive outpatient program because of substance abuse scores at least moderate (2).
- A patient who has blackouts/detoxification/withdrawal symptoms because of substance abuse scores severe (3).
- A patient who participate to Alcoholics Anonymous scores at least moderate (2)

motivation (decrease)

- A psychotic patient who is amotivated (anhedonia):
  - not focus of treatment = mild (1)
  - needs treatments = moderate (2)
  - requires ER or hospitalization = severe (3)
- A patient who has little interest or pleasure scores AND this is not a focus of treatment, scores mild (1)

motivation (increase)

- A patient who has mania:
  - few hypomanic symptoms scores mild (1)
  - elevation/abnormal drive scores moderate (2)
  - dangerous behavior that needs hospitalization scores severe (3)
  - is already hospitalized scores severe (3)
- A patient who has obsessive compulsive disorder:
  - present AND no under prescription scores mild (1)
  - present AND under prescription scores at least moderate (2)
  - present AND hospitalized scores severe (3)
- A patient who has bipolar disorder:
  - has symptoms scores mild (1)
  - present AND under prescription scores at least moderate (2)
  - present AND hospitalized scores severe (3)

what we learned about psychiatry

- a disturb is in the positive valence if there's a rewarding component (overdose intentional vs. unintentional)
- main sources of positive valence are abnormal changes in drive/approach/motivation: increase (mania, substance) and decrease (symptoms of depression)
- depression, anxiety, anhedonia, lack of motivation are correlated
- violence not because of substances usage doesn't score in positive valence
- cutting is an abnormally rewarding behavior
- PSTD has a negative valence component (arousal, social withdrawal, cognition, intrusive thoughts)
- any mood/anxiety problem falls in the negative valence

some fact about the data

- ★ all the notes contain a review of systems which is standardized, although some categories come and go (PTSD)
- ★ most notes contain useful alcohol screen
- ★ some words like panic, abuse are hard because they appear negated and not
- ★ Important dimensions:
- ★ the frequency in history
- ★ is or is not the focus of a treatment (not focus: mild, focus: moderate, inpatient: severe)
- ★ the gravity of the specific event

challenges!

- ★ patients describing someone else's symptoms (spouse worried about her husband, parent worried about one child)
- ★ vague prior event reference
- ★ clinician diagnosis not consistent with symptoms (patient describes panic attacks, diagnosed with general anxiety)
- ★ extremely complex patients (PTSD + borderline + substance +/- psychosis)
- ★ extremely ill patients where history is not helpful (but mental status and impression may be)
- ★ many syndromes are reflected in multiple domains. (depressed mood (negative), loss of interest (positive), insomnia (arousal), OCD, PTSD, panic)
- ★ many domains reflect multiple syndromes (cognition = ADHD, depressive symptoms, psychosis, PTSD)

keywords

depression, anxiety, anhedonia, bipolar disorder, smoking, smoking history, cigarette, EtHO (alcohol), MJ (marijuana), street drugs, illegality, manic, mood, decreased interest, decreased pleasure, violence, blackout, amotivation, cutting, bingering (addiction to crack/cocaine), purging (to cause intestinal evacuation), IOP (intensive outpatient program), PHP (partial hospitalization program), DUI (driving under the influence), OUI (operating under the influence), legal trouble, arrest, outpatient treatment, CBT (cognitive behavioral therapy), AA (Alcoholics Anonymous), hypomania, hypomanic, hyperactivity, increased libido, elevation, abnormal drive, ECT (Electroconvulsive therapy), MDD (Major depressive disorder), OCD (obsessive compulsive disorder), BPAD (bipolar disorder), ER (emergency room), ROS (review of systems), PTSD (post-traumatic stress disorder), ADHD (attention deficit hyperactivity disorder)

Organizing committee:
Ozlem Uzuner, co-chair, SUNY at Albany
Amber Stubbs, co-chair, Simmons College
Michele Filannino, co-chair, SUNY at Albany

Tianxi Cai, Harvard School of Public Health
Susanne Churchill, Harvard Medical School
Isaac Kohane, Harvard Medical School
Thomas H. McCoy, MGH, Harvard
Roy H. Perlis, MGH, Harvard
Peter Szolovits, MIT
Uma Vaidyanathan, NIMH
Philip Wang, American Psychiatric Association

thanks!