Saha, K., & Reddy, M. D. Vedant Das Swain, Julie M Gregg, Ted Grover, Suwen Lin, Gonzalo J Martinez, Stephen M Mattingly, et al. 2019. Imputing Missing Social Media Data Stream in Multisensor Studies of Human Behavior. In *Proceedings of International Conference on Affective Computing and Intelligent Interaction (ACII 2019)*., http://koustuv.com/papers/ACII19_SM_Imputation.pdf

Imputing Missing Social Media Data Stream in Multisensor Studies of Human Behavior

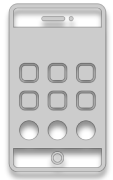
Saha, K., Reddy, M. D., Das Swain, V., Gregg, J. M., Grover, T., Lin, S., Martinez, G. J., Mattingly, S. M., Mirjafari, S., Mulukutla, R., Nies, K., Robles-Granda, P., Sirigiri, A., Yoo, D. W., Audia, P., Campbell, A. T., Chawla, N. V., D'Mello, S. K., Dey, A. K., Jiang, K., Liu, Q., Mark, G., Moskal, E., Striegel, A., & De Choudhury, M.

Koustuv Saha, Georgia Tech

Sensing Human Behavior

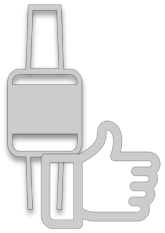
Survey Instruments

- Self-Report Questionnaires



Active Sensing

- Ecological Momentary Assessments (EMAs)



Passive Sensing

- Smartphones and Wearables
- Social Media

Social Media as a Passive Sensor

- Naturalistic setting
- Unobtrusive access
- Longitudinal and Extended Periods (beyond study period)
- Verbal and Behavioral

There are limitations associated with the social media data stream

Limitations (Social Media Data Stream)

Retrospective in nature:

So, the availability and quality of data depends on the social media use of the participant

Limitations (Social Media Data Stream)

Not everybody is on social media

Social Media population skewed towards young adults (Pew, 2018)

Limitations (Social Media Data Stream)

Data collection challenges

Changing nature of social media APIs (Facebook, Twitter, Instagram, Linkedin, etc.)

Consequences in Studies of Human Behavior

Multimodal Sensing Studies have to focus on:

- a very (social media) active participant cohort: hurts generalizability and recruitment
- disregard those with no social media data: hurts scalability
- disregard the capability of social media data stream altogether: hurts multisensor-fusion capabilities

Our work concerns…

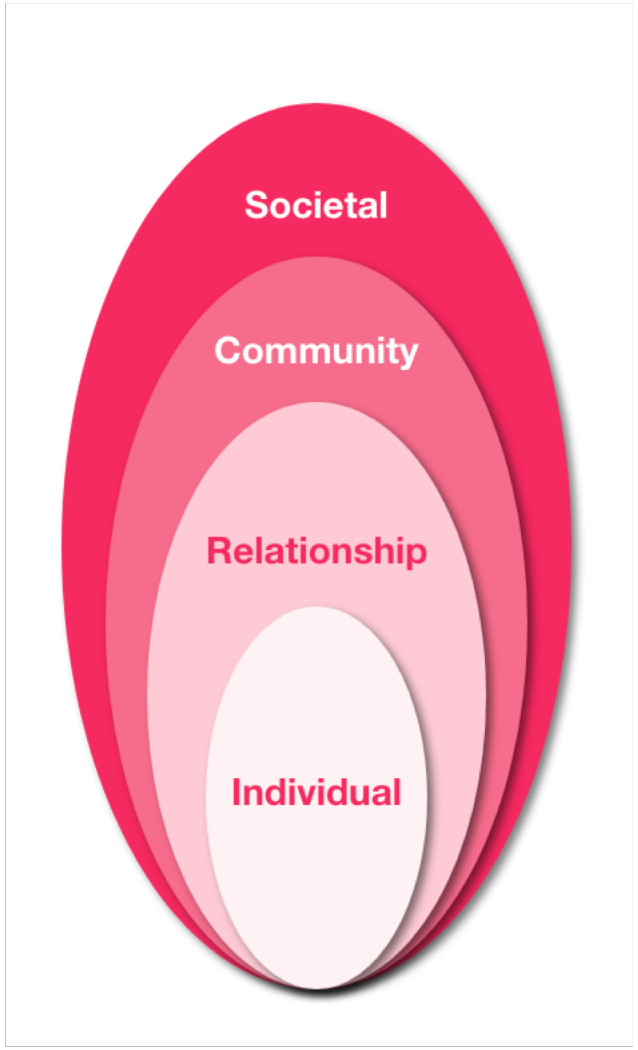
…how can we leverage the potential of social media data in multimodal sensing studies of human behavior, while navigating the challenges and limitations of acquiring social media data?

Our work contributes…

… a statistical framework to impute missing social media features by learning individuals' observed behaviors from other passive sensing streams (Bluetooth beacons, wearables, and smartphone sensors).

Social Ecological Framework

Human behaviors and attributes can be considered to be deeply embedded in the complex interplay between an individual, their relationships, the communities they belong to, and societal factors[+].



[+]Ralph Catalano. 1979. Health, behavior and the community: An ecological perspective. Pergamon Press New York.

The Tesserae Project

757

Participants

By leveraging passive sensors, this study aims at proactively identifying changes in an individual that may impact their wellbeing and job performance

Wearable

Smartphone

BT Beacon

Social Media

Surveys

Data and Problem (Predicting Psychological Attributes)

- 603 participants with physical sensor (Bluetooth, Smartphone, and Wearable) data
- 496 participants with social media (Facebook) data (~82% of the dataset)

Therefore,

- to include all the participants, we could only incorporate the physical sensor features,

Or,

- To include all the sensor modalities, we can only include a subset of participants.

Imputing missing social media features help us use **all sensors** and **all participants' data**

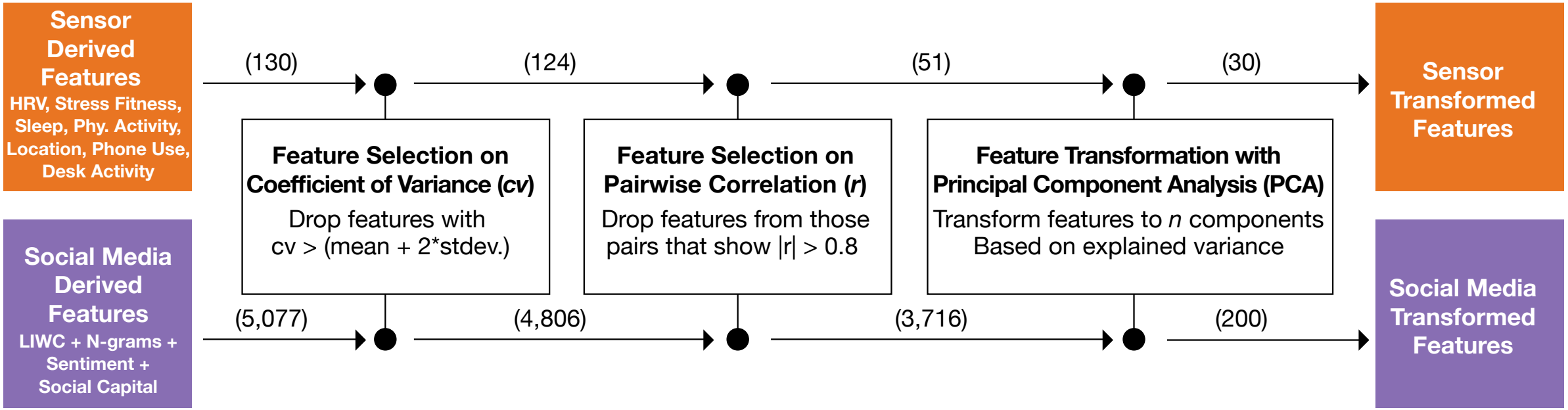
Feature Engineering

- Features known in theory to be predictive of psychological constructs (personality traits, affect)

- **Physical Sensor Features**: heart rate, heart rate variability, sleep, stress, step count, physical activity, mobility, phone use, call use, work duration (130 raw features)

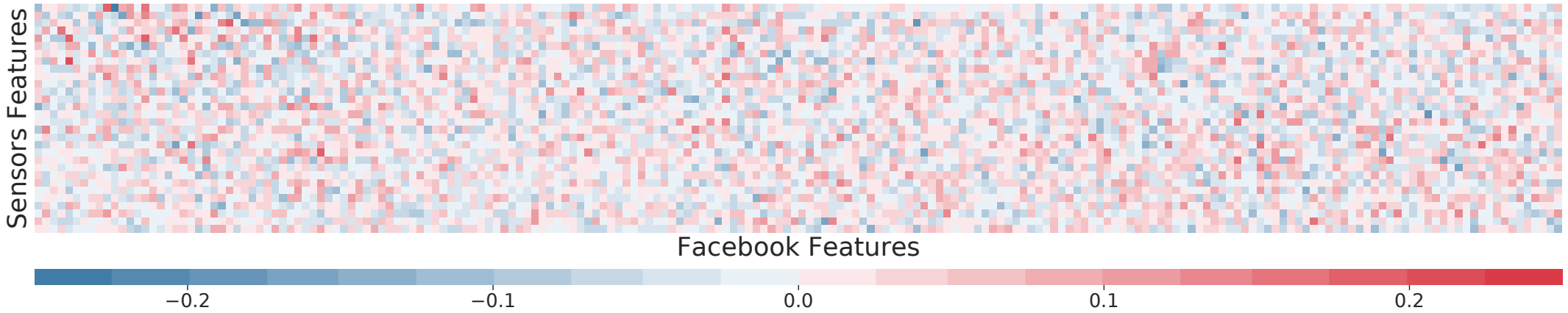
- **Social Media Features**: psycholinguistic attributes (LIWC), top n-grams, sentiment, social capital (number of check-ins, engagement, activity with friends, etc.) (5,077 raw features)

Feature Engineering (Selection & Transformation)



Imputing Social Media Features

Can sensor features predict social media features?



Pearson's correlation (*r) ranges between -0.21 and 0.22 showing the likelihood of weak correlation*

Imputing Social Media Features: Methods

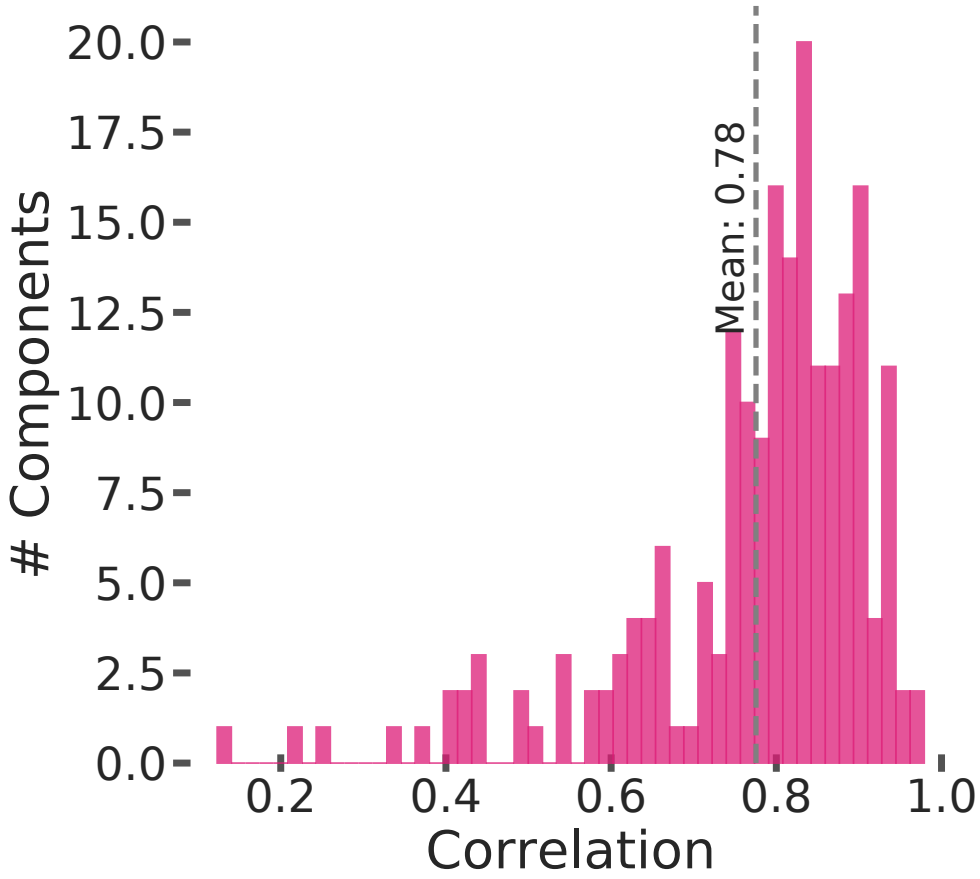
For each of the 200 social media transformed features, we build a separate model that:

- uses the sensor transformed features as the independent variables and
- predicts the corresponding social media transformed feature as the dependent variable.

Imputing Social Media Features: Results

k-fold cross-validation and pooled accuracy (Pearson's correlation (*r*) between actual and predicted features

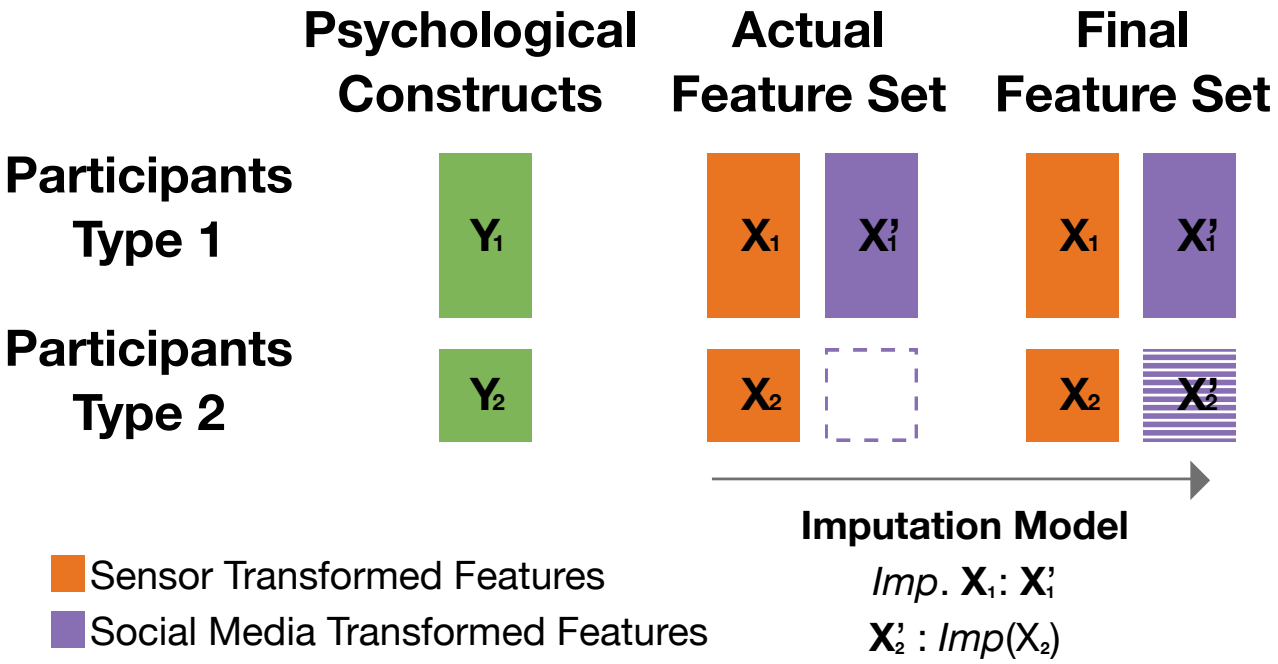
GBR (Gradient Boosted Random Forest Regression) performs the best: mean r = 0.78



Is Imputation Effective?

PREDICTING PSYCHOLOGICAL CONSTRUCTS WITH MULTIMODAL SENSING DATA

Predicting Psychological Constructs with Multisensor Data



Base Models (Who have social media data)

$S_1.X_1 : Y_1$

$SS_1.X_1 + X_1' : Y_1$

Models (Who do not have social media data)

$S_2.X_2 : Y_2$

$SS_2.X_2 + X_2' : Y_2$

Final Models (All participants)

$S_3.\ (X_1 + X_2) : (Y_1 + Y_2)$

$SS_3.\ (X_1 + X_2) + (X_1 + X_1') : (Y_1 + Y_2)$

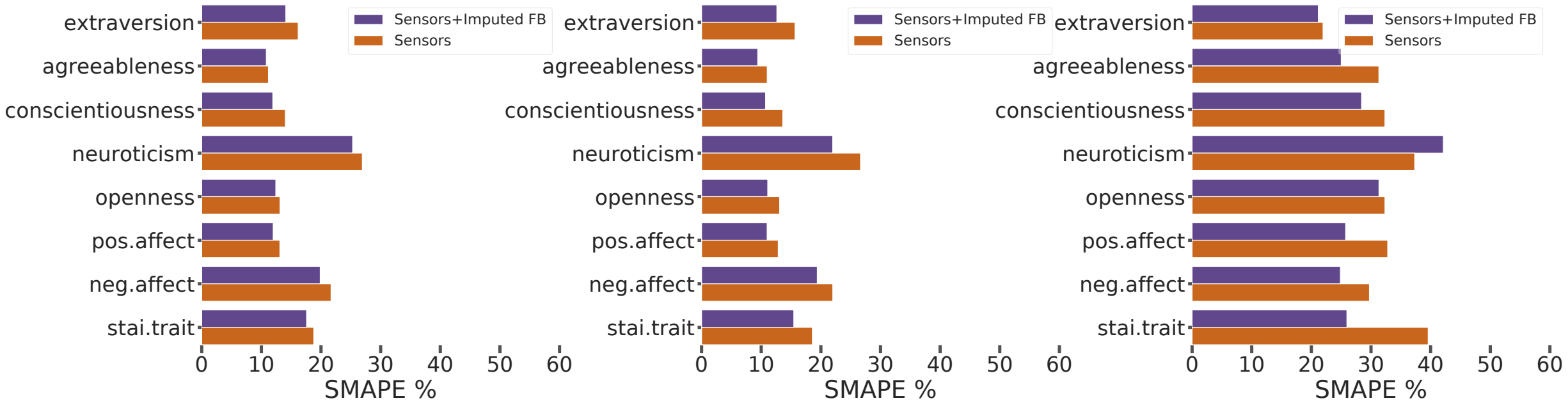
Predicting Psychological Constructs with Multisensor Data

We evaluate all our prediction models using three kinds of algorithms:

- Linear Regression
- Gradient Boosted Regression
- Neural Network Regression

The above algorithms cover a broad spectrum of algorithm families

Effectiveness of Social Media Feature Imputation

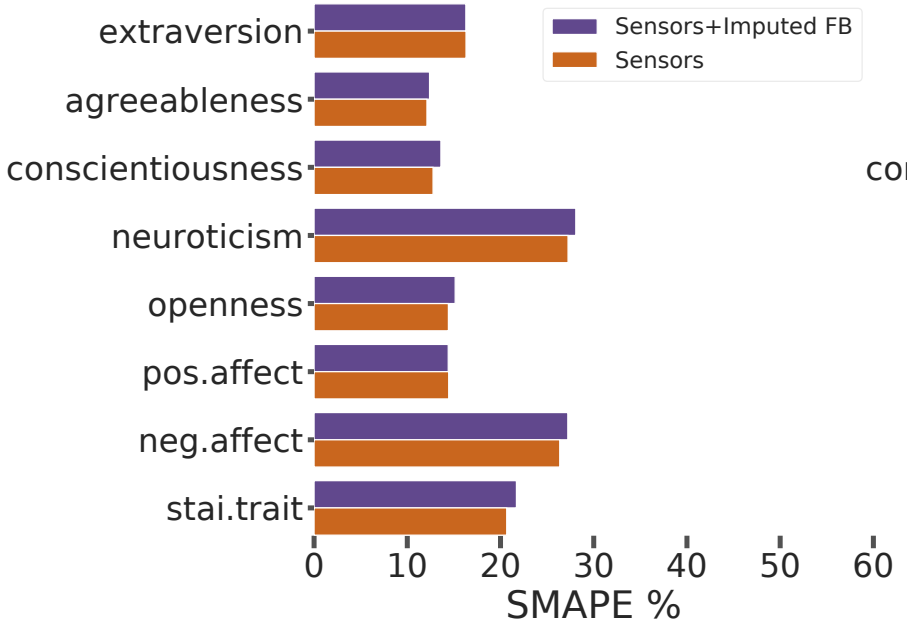


Linear Regression | Gradient Boosted Regression | Multilayer Perceptron

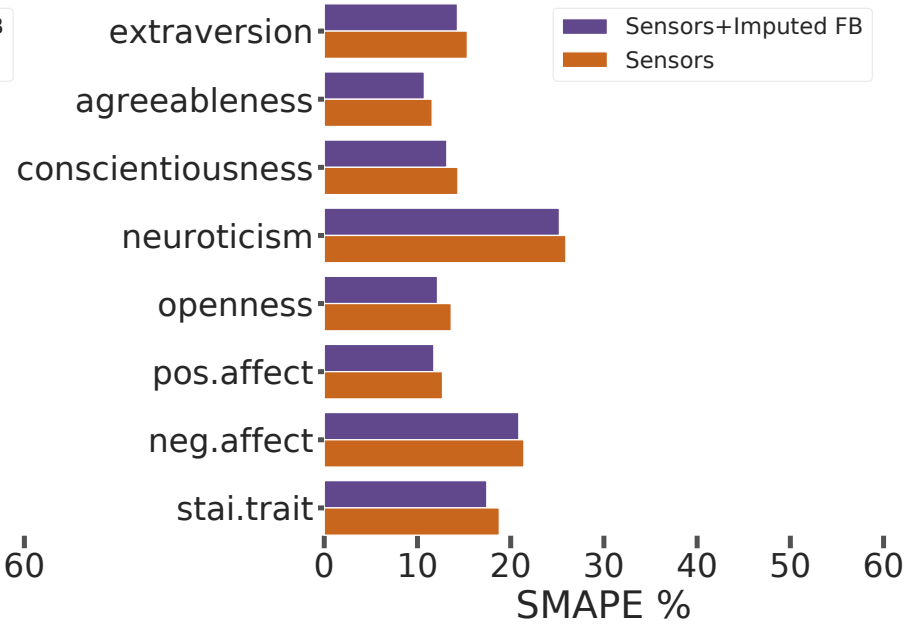
SMAPE comparing three models that use physical sensor features vs. those that use sensor and imputed features to predict psychological constructs on the entire dataset

Outcome: Imputed Social Media Features Improve Predictions

Robustness Against Other Imputation Approaches



Mean Imputation



Random Imputation

SMAPE comparing prediction models that use sensor features vs. those that use sensor and mean- / random- imputed features

Outcome: Mean / Random Imputation does not improve (or even depletes) predictions

Discussion

- Contribution: A framework to impute social media features in longitudinal and large-scale multimodal sensing studies of human behavior
- Theoretically situated in the Social Ecological Model
- Similar approach can be applied for other sensors
- Ethics: Should imputation be done on those individuals who *do not want to share* their social media data?

Ethics

- Latent dimensions do not necessarily translate to social media activity or behavior
- Caution against the use as a means to surveil
- Should imputation be done on those individuals who *do not want to share* their social media data?

Saha, K., & Reddy, M. D. Vedant Das Swain, Julie M Gregg, Ted Grover, Suwen Lin, Gonzalo J Martinez, Stephen M Mattingly, et al. 2019. Imputing Missing Social Media Data Stream in Multisensor Studies of Human Behavior. In *Proceedings of International Conference on Affective Computing and Intelligent Interaction (ACII 2019)*., http://koustuv.com/papers/ACII19_SM_Imputation.pdf

Imputing Missing Social Media Data Stream in Multisensor Studies of Human Behavior

This research is supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA Contract No. 2017-17042800007.

Thank You

@kous2v| koustuv.saha@gatech.edu | koustuv.com