**Using Federated Learning to Address User Heterogeneity, Missingness and Privacy**

The ubiquitous nature of smartphones and wearables, such as Fitbits, generates a lot of potential for data collection and analysis, but comes with the issue of protecting user privacy. People carry these items often, so we can collect high-frequency data by sampling at multiple time points throughout the day. When collecting mobile health data, we want to ensure that users’ confidential health information is not leaked to others. Privacy breaches often happen when data servers are compromised. To guarantee privacy, we propose using federated learning. In federated learning, there is no global server that contains all users’ information. In fact, as we will explain, model training occurs on each individual’s smartphone, and raw personal data across users is not stored anywhere. Federated learning also addresses some of the other typical challenges in mobile health data: accounting for user heterogeneity since each user has unique traits, and missingness not at random, since users can become disengaged with a smartphone app over time.

One method to address user privacy while taking advantage of sharing data across users is federated learning, a technique recently published.[[1]](#endnote-1) The process is as such: we train one model a little at a time on each user’s data, and get updates from each user. We send the updates, but not the individual data, to the central server. We choose a method to aggregate the updates, and produce a ‘global’ model server-side. The global model state is transmitted to each user, and we train on each user’s data again. This process iterates. The global model learns from individual models, but individual data stays on each user’s wearable, and cannot be reused for other purposes without user consent. By training a global model over all users, federated learning has more training data, compared to if we only trained an individual model on each user, in which case we lack data for users who are less engaged. For the rest of the paper, we use ‘wearables’ to refer to activity trackers as well as smartphones, both of which are carried on the user for most of the day and have the potential to collect user data.

Federated learning is generally done using neural networks, and neural networks are an appropriate choice to handle user heterogeneity and data missing not at random. Neural networks can be trained to recognize which users are similar to each other and make similar predictions, and pull information from similar users when a user is missing data. As neural networks are quite sensitive to hyperparameters, one of the areas I will focus on is developing methods to optimize hyperparameters for time series prediction in mobile health. This will also help inform hyperparameter tuning for federated learning, as currently there is no standard method for it. In addition, since neural networks cannot take missing values, I will explore and evaluate different imputation methods.

### Case Study: Intern Health Study

We turn to a case study of mobile health data from the Intern Health Study, a year-long study conducted by The Sen Lab at the University of Michigan that tracks physical activity, sleep duration and quality, heart rate, and daily user-inputted ordinal mood from over 500 users’ Fitbits and smartphones.[[2]](#endnote-2)

This data has one strong advantage because it was not collected using federated learning. As such, we know the results of the global model, and we can compare a federated learning model against the true global model. We find that the federated model does almost as well as the global model, and that the federated model does much better than individual models. This is important because individual and global models do not necessarily protect user privacy, as we put all users’ raw data together to train our model. By finding that federated learning does much better than individual models, and nearly as well as global models, we are not losing much prediction power while guaranteeing user privacy through federated learning, and we are gaining the significant benefit of guaranteed user privacy.

While many people are taking advantage of the opportunity to collect data from wearables, wearables often contain confidential data and privacy is important. The innovations I develop will bring further advances in applications such as mobile health and the social sciences, both of which often use individual data and require user privacy. Knowledge gained from this project will be disseminated through publications and open-source software so that other researchers can apply rigorous federated learning and neural networks methods to time series. This research will provide new methods for addressing the challenges that come with data from wearables, and facilitate continued learning and growth in the mobile health field.

1. **References**  Konecny, J., et. al. (2016). Federated Learning: Strategies for Improving Communication Efficiency. In *NIPS Workshop on Private Multi-Party Machine Learning*. [↑](#endnote-ref-1)
2. The members of The Sen Lab are here: https://www.srijan-sen-lab.com/people. The co-PIs are Dr. Srijan Sen and Dr. Constance Guille, and Dr. Elena Frank is the Study Director. [↑](#endnote-ref-2)