# Maintained Individual Data Distributed Likelihood Estimation (MIDDLE)

Joshua N. Pritikin, Department of Psychology

Yang Wang, Department of Systems and Information Engineering

The current research paradigm for social, behavioral, and epidemiological sciences is that data are collected into a centralized repository. After data collection, this central data repository is used to fit candidate statistical models, commonly obtaining maximum likelihood or Bayesian estimates of model parameters and fit statistics. Boker et al. (2013) proposed a new paradigm wherein data remain where they were originally collected—on each participant's personal smartphone, computer, tablet, or wearable computing device—and remain private, that is to say these data are never revealed by the participant. This new paradigm would ease a variety of new study designs that involve bigger data than previously practical. For example, it would become feasible to study patient history across hospitals even when data is legally prohibited from leaving the hospital. MIDDLE would facilitate longitudinal linking of within participant data and sharing of data between studies. In general, there is potential for substantially reducing the time between hypothesis generation and dissemination of results while simultaneously reducing participant burden.

Imagine we are 10 years in the future. The MIDDLE system is fully implemented and smoothly coordinating research. Suppose your team has a hypothesis about diet and exercise that involves a self-report questionnaire instrument and smart phone accelerometer sensor data. Using the MIDDLE experiment creation web page, your team creates a detailed description of the experiment and draws the statistical models that represent your hypotheses.

After Institutional Review Board (IRB) approval, the MIDDLE service publishes the experiment in a web-accessible central repository—its "app store for science." The MIDDLE service manages opt-in (and opt-out) consent documents and facilitates the download of the MIDDLE experiment app (including the model likelihood calculator) to each participants' smart phone. As participants consent into the experiment, the researcher's MIDDLE optimizer software begins the optimization of statistical models.

The optimization process proceeds as follows: (1) The optimizer chooses starting values for all parameters and sends these to all current participants; (2) Each participants' personal device calculates the likelihood of the participants' data collected so far and sends only that likelihood number back to the MIDDLE optimizer; and (3) The optimizer chooses new parameters and repeats the process until convergence criteria are reached. Privacy is maintained. At no time do participant data leave the participant's device! IRB-approved experimental modifications can be re-disseminated for participant consent and update.

When participants consent to the use of previously-collected data, each new experiment starts optimization with a large set of data automatically shared from previously-run MIDDLE experiments. Longitudinal data collection is thus automatically enabled and linked at the individual level at zero cost to a new project. Participants could even choose to collect personal data outside the context of an experiment using wearable activity monitors, health monitors, GPS tracking, or questionnaires. If these participants then opt into an experimental analysis, they can choose to allow access of these previously collected data in the new experiment.

A second research group runs a study on a hypothesis related to the first study. They look up the first study's results in a publication archive and follow the link to the associated models and instruments in the MIDDLE archive. However, their new hypothesis requires an experiment with an in-lab component as well as a self-report questionnaire and in-home sensor data. The second research group modifies the first group's instrument and statistical models and advertises their IRB-approved study, offering additional compensation to participants from the first study. Participants opt-in and most of those from the first study opt to allow data sharing. Within a week, the study has a relatively large data sample. Participants who consent to the in-lab followup are randomly selected to a treatment or control condition.

For participants who opt-in, the MIDDLE service transmits contact information to the research group, which arranges appointments for the in-lab study. Participants bring their personal device to the lab and the in-lab data are uploaded into the personal device for the participants to take home. Participants are given the choice of whether to allow the lab to archive a copy of their data. The analysis and write up proceed in the same manner as in the epidemiological experiment. Note that the in-lab data are always uploaded to the participants devices. Thus, these data are available, with participant consent, for sharing and longitudinal linking in other experiments. As more data are accumulated into participants' personal devices, their data become more and more valuable to future researchers, and thus of greater value to the participant.

In summary, MIDDLE permits statistical models to be fit and scientific hypotheses tested while maintaining the privacy of participants' data. Participants' data are never revealed to researchers. MIDDLE will provide automatic management of opt-in and opt-out consent; lower cost for the researcher and funding institute; and faster determination of results. Robust privacy protection will reassure participants involved in sensitive research and permit research on topics that are otherwise too sensitive or involve data that is legally required to be kept private.

## Proposed Work

There are software, privacy, and statistical challenges involved in building the MIDDLE system. To mitigate the risk for a full-scale, multi-year implementation project, we proposed to build an open-source prototype. A prototype of this system will allow us to explore privacy and statistical challenges in more detail than is presently possible. Example questions that can be explored in a prototype include:

- What convergence criteria are most suitable for a distributed likelihood calculation wherein participants may choose at any time to enter or leave the system?

- What choices are available for standard errors and how do these choices influence interpretation?
- Can we compensate for the bias that will be introduced when participation is correlated with measured or missing data?

We cannot address all of these questions but our work will be published with an open-source license. We will endeavor to make it convenient for other research groups to install and work with our prototype to investigate issues of interest.

In addition to statistical questions, a prototype will facilitate the investigation of privacy concerns. We will work to identify weak points in the system most vulnerable to surreptitious monitoring or malicious actors within the system. We will identify areas where cryptography may contribute to ensuring privacy. One promising refinement to enhance individual privacy is to perform peer-to-peer aggregation of objective functions prior to transmission to the MIDDLE optimizer. Peer-to-peer aggregation can help protect the privacy of the connection between participants and the likelihood of the model given their data.

## Approach

A prototype MIDDLE service is too much work to build from scratch. We will endeavor to reuse existing software components wherever possible to speed us to our goal. For example, `OpenMx` is free and open source software for use with R that allows estimation of a wide variety of advanced multivariate statistical models (Boker et al., 2011). `OpenMx` consists of a library of functions and optimizers that allow workers to quickly and flexibly define a structural equation model (SEM) and estimate parameters given observed data. We will improve the modularity of `OpenMx` so it can adapt to the MIDDLE architecture so that the optimizer can be separated from the evaluation of individual data rows. Specifically, we propose to build a prototype MIDDLE system with a plug-in style architecture that includes the following modules:

- A distributed full information maximum likelihood optimizer as a central service
- A personal device emulation module (PDEM) with a plug-in API
- An individual record likelihood estimator that runs on the PDEM
- A peer-to-peer and peer-to-server communications module
- A personal data storage and retrieval module that runs on the PDEM

With these components, we will simulate a complete experiment on the University of Virginia Linux cluster from start to finish.

Our proposed project is to build a prototype research platform that embodies the essential characteristics we anticipate in a full-scale MIDDLE system. This will lay the groundwork to explore a broad variety of exciting research questions. In addition, Joshua will focus on longitudinal multilevel modeling applications and Yang will focus on distributed and peer-to-peer likelihood estimation algorithms. These are the two main research areas where filling gaps in our knowledge will remove formidable barriers to a full scale MIDDLE implementation.