Maintained Individual Data Distributed Likelihood Estimation (MIDDLE)

Steven Boker (University of Virginia)
John Lach (University of Virginia)
Donald Brown (University of Virginia)

December 9, 2014

Overview

"SCH: INT: MEDIUM: Maintained Individual Data Distributed Likelihood Estimation (MIDDLE)", PI Dr. Steven M. Boker, University of Virginia, will construct and validate a revolutionary model for the accomplishing health science human-subject research with networked devices. The MIDDLE paradigm is that data can be privately maintained by participants on their personal devices and never revealed to researchers — while statistical models are fit and scientific hypotheses tested.

Intellectual Merit

The current research paradigm is that data are collected and centralized. After data collection, the central data repository is used to fit candidate statistical models by optimization to find maximum likelihood estimates of model parameters. The innovation of the MIDDLE paradigm is that data remain where they were originally collected, (e.g., on each participant's personal smartphone, computer, or home health monitor, or in a firewall protected research or clinical data repository). Statistical models are optimized by sending a candidate vector of model parameters to each individual data site, where an likelihood function is calculated locally. Only the likelihood function value (e.g., likelihood) is returned to the research study's central optimizer. The optimizer aggregates likelihood function values from responding remote data sites and chooses a new set of parameters. This process is repeated until models sufficiently converge.

The MIDDLE paradigm is not only feasible, but also solves or simplifies many problems that plague current human health science studies. A MIDDLE study provides significantly greater privacy for participants; far lower cost for the researcher and funding institute; a larger base of participants; faster determination of results and scales up to very large numbers of participants. MIDDLE facilitates the use of new mobile devices that can enable studies to be performed while participants remain in their normal living environments, thus opening novel paths in research study design.

When using the MIDDLE paradigm, data can be collected at the same time as models are optimized. This means that when sufficient data have been collected, a study can automatically terminate or switch to a cross-validation regime. Note also that individual-level variables, repeated measurements, and time series are automatically linked and estimated at the individual level first and then aggregated subsequently, consistent with a personalized medicine approach to research. Furthermore, if a participant opts into many studies simultaneously, all studies would automatically have real-time data sharing.

Broader Impacts

Improved Individual Privacy of Sensitive Health Information — Since data do not leave a participants' or patients' personal device, the risk of disclosure of sensitive information is mitigated as compared with centralized repositories. In addition, participant/patient consent management is in their own control at all times — individuals can opt in or out of specific uses of their data at any time, something that is lost when acceding control of data to a centralized repository.

Accelerated Translation of Research into Practice — As hospitals and primary care physicians install MIDDLE-compatible optimizers, a radically new approach becomes possible for translating research into medical practice. When research studies use the MIDDLE paradigm to develop predictive statistical models, models and their parameter estimates will be stored and available on the MIDDLE Host. Hospitals and primary care physicians could then have linked access to these predictive statistical models, which could be downloaded and immediately used by a healthcare provider to assist in diagnosis or to assess complex etiological risks for patients with MIDDLE-compatible home healthcare monitors.

Improved Longitudinal Data for Person-Specific Medicine — Personalized medicine requires person-specific data and models. Given participant consent, data within a participant's personal device are automatically longitudinally linked. Predictive models can be quickly translated from the MIDDLE Host and used by hospitals and primary care physicians to improve diagnoses and prescribe personalized treatment.

KEYWORDS: data privacy; electronic health records; sensor networks; statistical modeling; distributed computing

1 Project Description

A revolution is in progress in health and behavioral research. The 2013 Pew Foundation's Tracking for Health study [34] reported that 69% of Americans track some form of personal health data (PHD) and 21% of Americans do so on a personal digital device. A California Institute for Telecommunications and Information Technology study [1] reported that 75% of participants "probably" or "definitely" would be willing to share their PHD with qualified researchers, 67% felt it was either "very" or "extremely" important that their data be kept anonymous. The same study found that 54% of participants believed that they should own all their data and another 30% believed that they should share ownership.

Presently, data from participant's personal devices are centralized into a data repository. If ownership of the data remains with the participants, then it is as if centralized data are being held in trust. The data are out of the participants' control and this can be a cause for legitimate concern. For instance, in a recent study of urinary incontinence in the elderly at the University of Virginia, wireless in-home monitors were refused by some participants because participants or their caregivers do not want such data to leave the home.

Why are data needed by health scientists? The answers are, in their roots, statistical. First, data provide a foundation for statistical models to be generalized to a selected group or population. Second, individual data provide bases for statistical models of each individual's health to be used in personalized diagnosis and intervention decisions. Are the data themselves the goal? No. In each case, the goals are scientific discovery and/or decision support for healthcare professionals. Thus, the problem is fundamentally a question of statistical research methods: What information is necessary and how can this information reach the scientists/decision makers with a minimum risk of disclosure?

This leads to a radical notion — Individuals would not need to reveal their data if a method could be found that provided statistical information of equivalent or improved quality as that generated by current research practice. If such a method could be implemented, the problems of linking at the individual level would be moot: Each person would maintain sole possession of her or his own data and so an single individual could opt-in to multiple analyses. Privacy would under participant control: Each person would have responsibility for her or his own possessions.

The current project implements a method by which data can be maintained by their individual owners and not divulged, and yet statistical models can still be optimized. The MIDDLE paradigm is that data remain where they were originally collected, (e.g., on each participant's personal smartphone, computer, or home health monitor). Candidate statistical models are fit by using a secure internet connection to send a vector of model parameters to each individual's personal device. The personal device then calculates the likelihood function and only this single number is ever returned to the statistical optimizer. The optimizer aggregates likelihood function values from all data sites and chooses a new set of parameters and repeats as the models converge. **The MIDDLE project aims to develop solutions to the statistical, communications, and distributed computing issues posed by this novel paradigm and to develop open source software that implements the method to conduct health science research.**

The current project aims to provide methodology and software in order to implement a working MIDDLE system in simulation and to develop standards and software that can be used to deploy the MIDDLE architecture. Five specific aims will be accomplished by the project:

AIM-1 Develop and test novel distributed statistical estimation procedures. A sandbox simulation of the MIDDLE system will be implemented on a cluster in order to test novel statistical procedures in the context of the stochastic environment in which they will need to operate.

AIM-2 Develop and test networking protocols for statistical estimation. Bottlenecks and challenges in the distributed networking from the simulation developed in **AIM-1** will be identified and solutions implemented in order to ensure that the synchronized information needed by the optimizer and the personal devices can be handled in an asynchronous, stochastice, and fault tolerant manner.

AIM-3 Identify and provide solutions for threats to privacy and security. There are three potential points of failure resulting in loss of security, on the personal device, at the centralized optimizer, and in networking. Solutions will be developed for foreseen problems and a public testing challenge will be posed so the security community can be engaged in finding and addressing unforeseen problems.

AIM-4 Develop and implement data query software. A secure query and data management protocol and reference implementation will developed to run on a target (e.g., iOS or Android) personal device.

AIM-5 Deploy and test a prototype MIDDLE experiment. The distributed optimizer developed in **AIM-1** will be ported to run on at least one target personal device. This app and the data query software developed in **AIM-4** will be distributed to a pilot group of researchers' devices. The optimizer developed in **AIM-1** will then fit statistical models using the network protocols developed in **AIM-2** on simulated individual-level data stored on each of the personal devices while the security community (**AIM-3**) is challenged to violate the privacy of the simulated participants.

2 Significance of Proposed Work

We propose that data be retained by patients/participants in what we call *Maintained Individual Data* (MID), a software platform that would reside on a smartphone, home health monitor, computer, or cloud account possessed by the participant. *Distributed Likelihood Estimation* (DLE) refers to the process by which a central optimizer sends vectors of free parameters for candidate statistical model to each participant's personal device where a DLE remote app would query the MID, calculate the likelihood of the data returned by the query and send back to the central optimizer *only* that likelihood. The central optimizer would choose new parameters and repeat the process. Note that data could be collected simultaneously with model optimization. So, as new participants opt in, statistical power grows. When a chosen hypothesis test exceeds a chosen threshold or sufficient power has been reached, the experiment ends.

This paradigm shift in collecting and analyzing data will transform health science in the areas of data privacy, individual-level analysis, power estimation, data sharing, and the dynamic design of experiments. Currently, data are *collected*. That is to say numerical measurements from instruments, sensors, or questionnaires are centralized and stored in a data repository. When data collection is complete, this data repository is queried and analyzed. Privacy and ethical questions arise regarding to whom the data belong: the individual, the scientific project, the funding agency, or the society. We take the position that data belong to individuals and should remain in their possession [47, 60] unless they explicitly choose otherwise. This leads to new ways in which statistical analyses could proceed and scientific and individualized medical hypotheses could be tested even if individuals never divulge data. Following this approach to its logical conclusion leads to surprising efficiencies and simplifications in research methodology.

Data Privacy and Ownership — Data remain physically in each participant's possession and are not revealed during an experiment. There is no need for a central data repository. A participant can opt into (or out of) a study at any time without data management cost to the research project. Opting into a new experiment would release previously stored individual data for immediate use by the new research project even if the data were collected prior to the start of the new project, thereby automating the process of data sharing and placing the control over data sharing in the hands of the participants themselves.

Within-Person Data Linking — Longitudinal data linking in MIDDLE is automatic — all data belonging to an individual are (with consent) accessible to the MID platform, so multilevel and/or longitudinal models can be fit at the individual level without resorting to the use of personally identifying information in a central repository to link between occasions of measurement.

Data Sharing — If a participant opts into data sharing, any new experiment automatically has access to all previous data from all other experiments accessible to the participant's MID. This will accelerate scientific

discovery since new experiments will have immediate access to previously gathered individual-level data rather than waiting years for traditional data sharing. This effect will increase exponentially as individuals' MIDs become more data-rich. Automatic data sharing will dramatically increase the efficiency of NSF- and NIH-funded data collection by reducing duplication of effort.

Data Quality — Increased trust on the part of participants is likely to result in more honest responses to sensitive questions about critical variables such as drug use, HIV status or other socially sensitive behaviors. In addition, the knowledge that data are not leaving the personal device may encourage individuals to consent to use of wireless sensing devices that gather personally embarrassing data. Participant trust will depend on the trustworthiness of the source of the application download — If the participant trusts that the application will actually perform as advertised, then a significant barrier to data quality may be removed.

Optimal Power — Since data collection and data analysis happen simultaneously, experiments can be ended when either a pre-specified confidence interval for a parameter estimate is reached, pre-specified hypothesis threshold is reached, or when a pre-determined power is achieved.

3 Background

A MIDDLE experiment employs a novel method for maintaining data privacy while sharing minimum sufficient information to optimize statistical models that test scientific hypotheses. Figure 1 presents a flowchart view of a fully operational MIDDLE system connected to individuals, scientists, health care providors, hospitals, and PubMed. The current project will develop three major components of this system: i) the MID platform and DLE estimator shown in the boxes on the left; ii) the MIDDLE Optimizer shown in the box at the top; and iii) the MIDDLE Host shown in the box at the center of the figure. Beyond the current proposal, it is envisioned that data maintained within a firewall at a secure facility such as a hospital or the US Census Bureau could be able to be connected to a DLE optimization without the data crossing the firewall. Likelihoods of data given a statistical model and a vector of parameters are calculated locally on the personal devices or inside firewalls and these likelihoods are then aggregated by a research laboratory running a MIDDLE experiment. In the center of Figure 1 is the *MIDDLE Host*, which manages the communication between researchers and participants during analysis and, for many studies, also during data collection. It is imperative that the MIDDLE Host be operated by a trusted source — both in the sense of being a *trusted computing* facility as well as in the sense of being operated by an agency that has trust of the public.

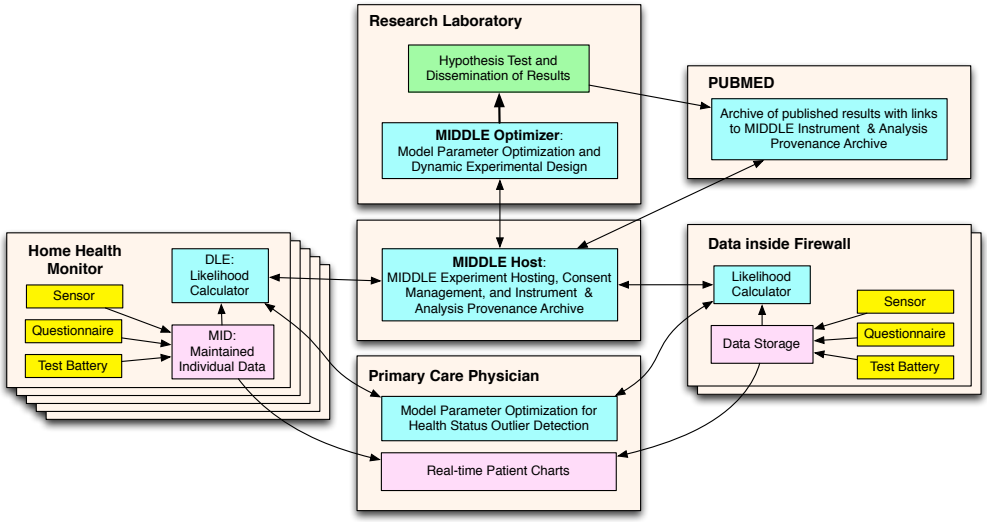

Figure 1. *Overview of information flow in MIDDLE experiments.*

There are many advantages to establishing the MIDDLE Host. Participant recruitment is centralized

across a large number of on-going and archived experiments, allowing a centralized place that prospective participants can learn about studies in which they might be interested. A rapidly growing number of health-care patients are becoming what is known as "engaged" or "informed" [44], monitoring their own health and using internet searches to learn more about their existing conditions [29]. These patients are likely to find and read advertisements for studies of interest to them and then use the MIDDLE Host as something like an "app store for science". We foresee that PUBMED articles will be linked to their instruments and analyses in the MIDDLE Host, so that researchers will be able to i) search for and find pools of prospective participants who might be willing to share their pre-existing data for a new experiment; ii) download pre-existing instruments and make modifications in order to speed their own novel instrument design; and iii) review methods sections of articles against actual analysis provenance trails.

Patients could choose to allow their primary care physicians (bottom of Figure 1) to have secure access to up-to-date longitudinal biomarkers and physiological measurements, thereby saving the physician's time. Physicians could run outlier detection models, giving real-time alerts to important changes in patients' medical status, allowing physicians access to vital alerts needed to recommend an early doctor visit rather than risk an emergency room visit at a later date.

3.1 Two Example Use Cases for a MIDDLE System

The MIDDLE project proposes to reorganize how experimental and epidemiological research are conducted. In order to give a better idea of how this reorganization might work in practice, we present a hypothetical large scale epidemiological study. We then present a design that includes random assignment of treatment and control conditions to form a planned experiment where participants must visit a laboratory for part of the study. The second study takes advantage of data sharing from the first epidemiological study. These two studies demonstrate the reasoning behind the MIDDLE computing infrastructure as well as illustrating some of the challenges that will need to be addressed.

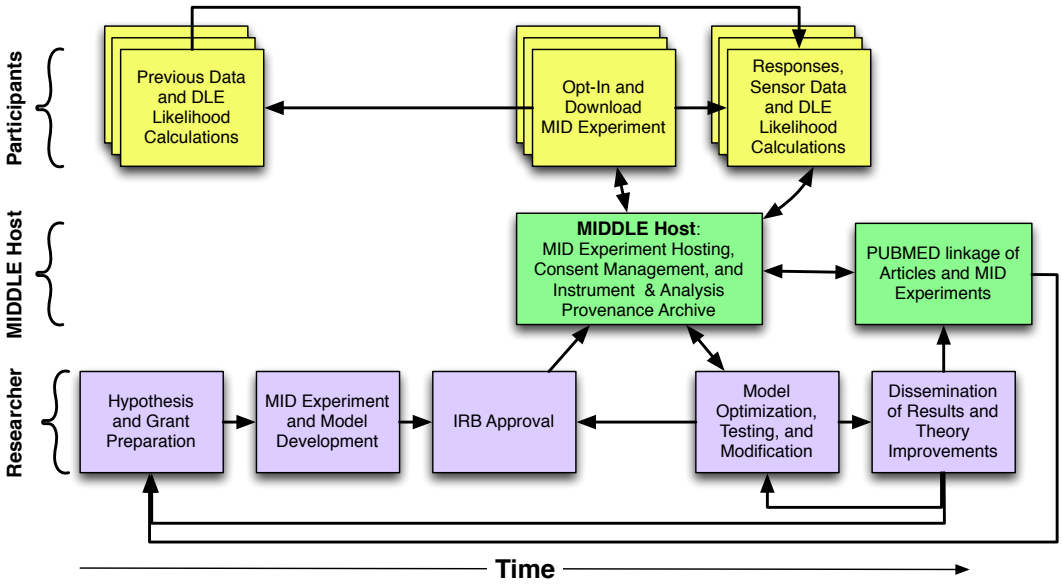

Figure 2. *Temporal flow of a hypothesis-driven MIDDLE epidemiological study.*

Figure 2 presents a timeline of how these two studies progress. The top row of the figure are the steps in which the participants engage. The center row represents the part of the process for which the MIDDLE Host is responsible. The bottom row are the steps taken by the research lab.

Epidemiological Study — A research group formulates a hypothesis about health maintenance through diet and exercise. They create a self-report questionnaire instrument, a software tie-in to accelerometer sensor

data from a smart phone, and statistical models to test the hypothesis. The research group uses a set of software tools (e.g., the MID programming interface) to create an experiment that will run on a personal device. The group submits the MIDDLE experiment and its consent instrument to an IRB. The MIDDLE experiment, DLE calculator, and consent documents are uploaded into a web-accessible central repository: the MIDDLE Host. This MIDDLE Host might be administered by a government agency (e.g., NSF, NIH, or CDC), by an academic institution, or by private industry. The MIDDLE Host then advertises the experiment to potential participants, manages opt-in (and opt-out) consent documents, generates a secure network certificate for the experiment, and allows participants to download the MIDDLE experiment software and, if necessary the MID data management app and DLE likelihood calculator app.

As participants opt into the experiment, the MIDDLE Host sends certificate information to the researcher's MIDDLE optimizer software, which then begins to optimize the pre-specified 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' DLE calculates the likelihood of the participants' data collected so far and sends that likelihood number back to the MIDDLE optimizer; (3) The optimizer chooses new parameters and repeats the process until convergence criteria are reached; (4) Either the experiment is finished (after collecting just-sufficient data) or the model or experiment is modified and the process is repeated. For follow-up experiments, IRB approved experimental modifications are uploaded to the MIDDLE Host and are re-disseminated for participant consent and opt-in.

Note that when participants give consent for 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 automatically enabled and linked at the individual level.

Once the MIDDLE experiment(s) are complete, articles are written and submitted to PUBMED. These articles are linked to the MID experiment and their statistical analysis provenance trail in the MIDDLE Host. Future researchers can thus (a) learn the exact methods and analyses that led to a published result, and (b) re-use parts of MID experiments and models in order to provide exact replications and/or maximally take advantage of the built-in data sharing enabled by the MIDDLE network of participants.

Combined In-Lab and In-Home Study with Treatment and Control — A second research group investigates running a study testing a hypothesis related to the first study. They look up the first study's results in PUBMED and follow the link to the associated models and instruments in the MIDDLE Host archive. The group downloads the MID experiment module and its statistical models, which include some of the necessary variables. However, this new hypothesis requires an experiment with an in-lab component as well as a self-report questionnaire and in-home sensor data. The group modifies the MID experiment module and statistical models and advertise their IRB-approved study on the MIDDLE Host, offering additional compensation for participants from the first study. Participants opt-in and a proportion of those from the first study opt to allow data sharing. The study quickly has a relatively large data sample. Some participants completing the in-home questionnaire are randomly selected through the MIDDLE Host for inclusion in an in-lab section and are assigned to a treatment or control condition.

For participants who opt-in, the MIDDLE Host transmits contact information to the research group, and they arrange appointments for the in-lab study. Participants bring their MID-enabled device and after the in-lab experiment is performed, data are uploaded into the MID for the participants to take home. Participants are given the choice 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' MIDs. Thus, these data are available for sharing and longitudinal linking if a participant opts into another experiment. As more data are accumulated into participants' MIDs, their data become more valuable to future experiments, and thus also of greater potential value to the participant.

3.2 Preliminary Simulation Results

Given the novelty of the MIDDLE paradigm it is reasonable to question whether a DLE algorithm can achieve convergence without a centralized and fixed data set. A significant problem is whether the constantly changing sample during estimation will prevent the estimator from converging. To answer this question, pilot simulation data were generated and two ways that DLE could be implemented were prototyped and the computational and statistical effectiveness of each were evaluated and compared. First, a population of 2000 individuals were simulated each of whom had 200 simulated observations that conformed to a latent growth curve model. This model is widely used in structural equation modeling estimation of individual-level data in the behavioral sciences. Each individual's parameters for the latent growth curve were drawn from a normal distribution with means and variances ($\mu_{intercept} = 1.5$, $\mu_{slope} = 0.6$, $\sigma^2_{intercept} = 1.2$, $\sigma^2_{slope} = 0.9$, Cov$(intercept, slope) = 0.3$) shown as the horizontal lines in Figures 3-a and 3-d.

Next, sampling schemes were simulated to represent potential ways that participants might engage with the MIDDLE estimator. Figure 3 shows results of a single run of the simulation for one choice of participant engagement parameters. A time step was defined as the interval between occasions when the MIDDLE optimizer sends out requests for likelihoods. Participant engagement was simulated as follows: At each time step: All individuals in the population who had not already opted in have a probability ($p = .03$) of opting into the experiment; All individuals who had opted into the experiment had a probability ($p = 0.005$) of opting out; All individuals in the experiment had a probability ($p = .5$) of having entered new data since the previous time step; And all individuals in the experiment had a probability ($p = 0.3$) that their device was on, able to be contacted, and had a DLE calculator running. Note that these probabilities of participant engagement make it unlikely that the same sample is used twice at any two time steps. Thus, in this sampling scheme we have ensured that the convergence can be tested when there is no "complete" data set.

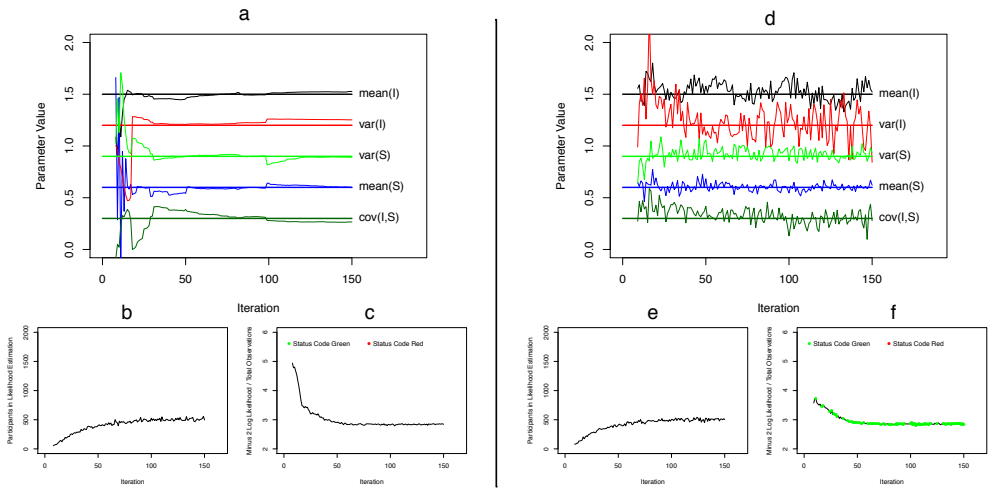

Figure 3. *Results of one simulated MIDDLE experiment estimated using two different optimization criteria. (a,b,c) Each query of the distributed devices was paired with one major iteration of the optimizer occurred. (d,e,f) Each time the distributed devices were queried, the optimizer was allowed to come to complete maximum likelihood convergence.*

In order to understand the difference between convergence using a fixed data set and convergence when the data are in continuous flux, we ran two separate estimators on the data. At each time step, the first estimator asked all available MID devices to return the likelihood of their data for a target set of parameters and for the associated minor steps that would allow the estimator to calculate the gradient and Hessian of the likelihood surface and choose a new set of model parameters. Thus the first estimator asked each MID device to perform the minimum likelihood calculations needed to generate improved model parameter

estimates at the next time step. The model parameter point estimates for the first estimator are plotted in Figure 3-a along with the number of individuals contributing likelihoods (Figure 3-b) and minus two log likelihood value divided by number of individuals (Figure 3-c) at each time step.

The second estimator asked all available MID devices to return the likelihood of their data and repeated that request until convergence prior to moving to the next time step. Thus, the second estimator asked each MID device to perform a maximum number of likelihood calculations at each time step. The second estimator performed calculations akin to a traditional bootstrap where each time step represented a new sample drawn from the population. The results for the second optimizer running a single experiment are shown in Figure 3-d, 3-e, and 3-f. While these plots are only one simulated experiment out of the more than 500 we ran in the pilot, the results are representative of the full simulation.

Figure 4 plots the means and standard deviations of 100 runs of the simulation described above. Here it becomes quite obvious that while the ever-changing sample due to the participant engagement probabilities poses a significant challenge for the standard likelihood estimation procedure. On the other hand, the single-step estimator appears resistant to that challenge and produces smaller standard deviations of estimates for every parameter except the lowest line, the covariance between intercept and slope.

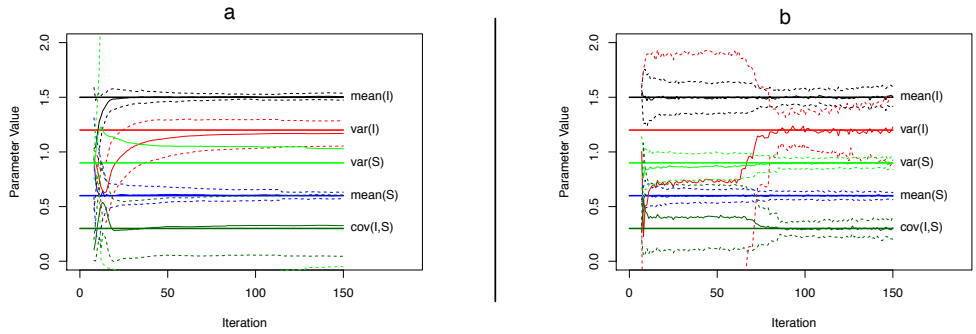

Figure 4. *Means and standard deviations of 100 simulated MIDDLE experiments. Each query of the distributed devices was accompanied by (a) one major iteration of the optimizer or (b) the optimizer was allowed to come to maximum likelihood convergence.*

At first we were surprised that the single-step estimator outperformed the full-convergence estimator. Upon further consideration we hypothesized that the full-convergence process was overfitting each "bootstrap" sample whereas the single-step estimator incorporated the bootstrap sampling into the likelihood convergence process and thus became more resistant to overfitting. **While much more needs to be accomplished before DLE becomes a well-known statistical technique with a comprehensive list of advantages and disadvantages, the results of this pilot simulation provide evidence that DLE is likely to be an efficient and unbiased estimator when used in the context of a MIDDLE experment.**

4 Proposed Work

The overall aim of the project is to prove that the MIDDLE approach is feasible by innovating solutions to statistical, privacy, distributed computing, and data management problems that stand in the way of an implementation of the MIDDLE system. The best proof of feasibility is a demonstration, so we first intend to write and deploy proof-of-concept implementations in *sandbox environments*, safe testbed environments in which no actual participants' data are at risk of exposure or loss. As outlined in the five specific aims of the project, these testbeds will involve two stages:

1. A simulation environment will be developed to run on a cluster (**AIM-1**) so that the problems and approaches discussed in Sections 4.1 through 4.9 below can be implemented and tested efficiently.

2. A pilot real-world sandbox (**AIM-5**) will be developed by porting the MID data query and management system (**AIM-4**) and DLE likelihood calculator (**AIM-5**) to a selected personal device platform (e.g., iPhone or Android device). These apps will be distributed to UVA researchers' personal devices along with simulated data. Then a modified version of the central optimizer (**AIM-1**) will communicate with the personal devices using the network protocol developed in **AIM-2** to fit candidate statistical models to simulated individual-level data stored on the researchers' devices. In this way, no actual participant data will be used and so if privacy breaches occur during security community stress testing (**AIM-3**), no participant data will be at risk.

The next sections break MIDDLE implementation into categories that can be implemented in parallel and discuss alternative solution paths to be followed if the main solution path is shown to be impossible or infeasible. In this way, we mitigate risk of a single point of failure in the overall design.

4.1 Distributed Likelihood Estimation

Estimating parameters of a statistical model using the MIDDLE system will require i) optimizer software running centrally; ii) DLE and MID software running on many personal devices; and iii) a trusted and secure communication protocol running on a wide area network linking the central optimizer and the personal devices. The optimizer will be pre-supplied with a set of starting values for a given model. The DLE and MID software will be pre-supplied with an likelihood function definition that includes tokenized matrix algebra and a data linkage definition. Through the MIDDLE Host, the optimizer will send several vectors of parameters representing *minor steps* in optimization to all MIDDLE devices that have opted into the experiment. DLE calculators will use MID to query the data on the personal device and will calculate the value of the likelihood function for each vector of parameters given their available data and return the calculated function values to the MIDDLE Host which will pass them back to the optimizer. MIDs with no available data will return a missing value token. The optimizer will aggregate the likelihood function values for each of the minor steps and calculate a gradient and Hessian at the starting values. The optimizer will then make a major step and broadcast a new set of minor step parameter vectors.

This method mimics the method of Full Information Maximum Likelihood (FIML) for obtaining parameter estimates. The properties of this algorithm are well-known when the data are constant throughout optimization [61, 33]. However, convergence properties are not well-known when the data are dynamically changing. In the case of a MIDDLE optimization, data collection and optimization will co-occur. Participants may opt-in or opt-out at any time. MIDs may be online or offline at any time. Algorithms such as Kalman Filtering [43] and Expectation Maximization [31, 6], or Bayesian methods [35] such as Gibbs Sampling [36, 26], and other Markov Chain Monte Carlo [9] methods for optimization may provide bases for solutions to this problem. While solutions to distributed likelihood optimization are challenging, based on the preliminary simulation, we expect to find at least one straightforward solution. The project will discover, report, and provide open source code implementations of solutions to these problems, testing to find optimal solutions within a variety of distributed modeling contexts.

It should be noted that when the data environment is in flux, models chosen for comparison are likely to need to be optimized simultaneously in order to be able to compare relative model fit between two or more candidate models. This means that model comparisons can occur online while data are being collected.

4.2 Parameter Stability and Power Estimation

Since the data are in flux during optimization, standard calculations for parameter stability and power estimation will need to be revised. These problems bear some resemblance to those posed by resampling and/or permutation testing. However, the MIDDLE situation continuously brings new information into the optimization, whereas both resampling and permutation testing use randomization to reduce effects of new

information due to the method by which samples are selected for estimation. This new information is both a help and a hindrance. We believe that we can capitalize on the information generated when participants opt in and opt out to provide a better estimate of generalizability. A naive solution would be to collect data and optimize until a moving window of iteration-to-iteration fluctuations in classically calculated standard errors fall below some chosen epsilon for some chosen number of iterations. Thus, there is at least one solution to this problem. However, it is likely that other solutions exist that can be shown to outperform the naive solution, thus reducing the required sample size for a chosen power. The project will find, report, and implement alternative solutions in open source software.

4.3 Identity / Group Linking

Some studies will require group membership information. For instance, an analysis of social networks or family relationships will require individual participants to be identified with a group. This group membership information is data and as such should be treated as belonging to the individual. A method must be implemented where an individual can opt-into a group membership. One solution would be for group members to choose a common pass phrase and give that to the MIDDLE Host consent controller when they opt-in. Multigroup model membership could then be incorporated into the MIDDLE experiment software uploaded to each group member's MID. Once the group membership data is stored in the MID, it is straightforward to calculate likelihood functions conditional on group membership. However, optimization will require aggregating likelihood function values conditional on group membership. In order for group membership to not be revealed to the central optimizer, one solution could be to allow peer-to-peer aggregation of likelihood functions prior to transmission to the MIDDLE optimizer.

Other studies will require (or could benefit from) linkage of in-home collected MID data with other data collected in a laboratory setting or pre-existing physician or hospital records. One solution to this problem is for participants to request that data from physicians or hospitals be uploaded into their MID. This is a sensible solution, although there are numerous practical roadblocks inherent in moving data between any two devices in a form that becomes immediately usable on the receiving device.

4.4 Identity / Data / Likelihood Function Decoupling

Encryption is not the same as privacy. Of course, data on the MID and transmissions among MIDs and MIDDLE servers will need to be encrypted. However, no encryption is totally secure, it is merely expensive and difficult to break. Increased focus on privacy reduces the payoff to a potential attacker who manages to intercept and decrypt transmissions on the MIDDLE network, or from malicious actors within the MIDDLE system. While the data on a given MID would still be susceptible to a determined decryption attack, our approach to data ownership improves data privacy by decentralizing participant data. The potential reward for decrypting a single MID is much lower than the reward for a successful attack on a centralized data repository. Any given MID is therefore a much less attractive target for identity thieves.

The problem of privacy is not solved by encryption, since a malicious user within the system might still be able to access participant data [21, 42]. To illustrate this problem, consider that some statistical models may reveal private information simply by calculating a likelihood function — similar to a game of "twenty questions". For instance, if a model only estimates the mean of a binary variable, say EyesAreBlue, a likelihood function could be used to determine all of the individuals in a sample who have blue eyes. One solution to this problem would be to randomly assign artificial groups for peer-to-peer aggregation of likelihood functions so that no single persons' likelihood function is ever known to the central optimizer. This is an example of *decoupling identity and likelihood functions*. If we can assure that the more the MIDDLE system knows about the likelihood function, the less it knows to whom the likelihood function belongs, then we have mitigated the risk of abuse by malicious users of the system.

Some applications, e.g., individual-level data plots, require revealed data. In order to preserve privacy, revealing raw data would need to be accompanied by hiding the identification of the nodes revealing data, both from the server collecting the data and from other nodes revealing data. One possible solution would be for a research group to generate a set of randomized "noise" data values and pass this to the first participant node for data collection. The first node would then, with some probability, replace a randomly-chosen row of the table with its own data values. The first node would randomly select another MIDDLE device in the set, and pass the newly modified table to that device. The new MIDDLE device would then repeat the process until a certain number of nodes had been able to modify the table. No node along the way would be able to identify the source of any row of data, or even if those values were real data. The original research group would receive the final values, remove any rows that remained unchanged, and have a subset of the data from some of the nodes on the path for plotting, but would be unable to determine which nodes contributed which data, or even which nodes were represented in the data set.

Fundamentally, this problem reduces to a modified version of the problem of anonymous peer-to-peer file sharing. The proposed solution is similar to a probabilistic anonymous bus system [7]. Other options, such as the P^5 protocol [54] and onion routing [32] will also be examined to choose a method that will preserve privacy while enabling applications such as individual-level data plots.

4.5 Likelihood Function Information and Jittering

Since the calculated value of the likelihood function is the primary information leaving the MIDDLE-enabled personal device, it is also the primary possibility for revealing data from the individual. A participant may be especially concerned about revealing sensitive data, e.g., genetic information, HIV status, or responses concerning illegal behavior. The structure of the system must not only protect privacy by technical means, but also reassure the participant in a believable manner. If this can be achieved, participants are more likely to answer truthfully [39]. One possible solution to this problem is to calculate how much information is being revealed by an likelihood function calculation conditional on the MID data while not exceeding a specified revelation threshold. If the threshold is exceeded, the likelihood function value could be *jittered* by adding a number drawn from a normal distribution with zero mean. The information threshold and standard deviation of the normal distribution could be controlled by a user interface function: essentially a "privacy dial" so that if participants were worried about their data being revealed, this setting could be increased. The result of likelihood function jittering is decreased power in an experiment, but not increased bias. This is a vast improvement over the current situation where participant dissembling or nonrandom nonresponse can introduce substantial bias into results.

The project will develop methods for calculating revealed information and implement MIDDLE user interface elements that allow for user control of information jittering. Power calculations that can account for function value jittering will be designed, reported, and implemented in open source by the MIDDLE project team.

4.6 MIDDLE Reference Implementation, Application Programmer Interface (API) and Software Developer's Kit (SDK)

The MID and DLE apps operating on a personal device comprise a software protocol existing within one of many possible operating systems and hardware platforms. Necessary characteristics include i) the ability to establish secure communications with a MIDDLE Host; ii) sufficient matrix numeric functions to be able to calculate a wide range of likelihood functions; iii) encryption and decryption algorithms; iv) data query protocols; and v) a well-defined applications programmer interface and software developer's kit that allows user I/O for experimental procedures. We will follow open source standards (e.g., Direct Project, [53], CONNECT, [57], Indivo [41]), and take into consideration ease of porting to commercial platforms.

The MIDDLE platform must simultaneously be flexible, extensible and secure [45, 48]. While these design criteria are not mutually exclusive (note e.g., Linux), they represent a substantial challenge for implementation on personal devices. The user interface for the MID must be so simple that at least 80% of adults can be expected to be able to successfully use it. In addition, the software development kit for new experiments must be so intuitive that those without expertise in computer science can build and deploy experiments and statistical models.

An API for experiment modules will be developed as part of the project. While the interface is, as yet, unspecified, it will be standards based, most likely around HTML and Java so that a modified open source browser or plug-in can present experimental paradigms and record responses and sensor data [59]. A Request for Comments will be issued to both the programming community and the community of NSF researchers in order to improve initial design decisions made by the MIDDLE programming team. The API for data manipulation will be based on a lightweight version of the R language, allowing pre-processing of data, e.g., upsampling or downsampling for sensor synchronization, pre-whitening for time series analysis, and time-delay embedding for derivative calculations [17]. The API for likelihood functions will be based on the specification of likelihood functions in the OpenMx Structural Equation Modeling software [11, 10]. OpenMx is developed and maintained by members of the MIDDLE project programming team and is open source Apache 2.0 licensed, so can be used without license fees by the MIDDLE project or other subsequent projects from other teams. The API for secure connection to external sensor devices will need to be specified so that manufacturers of sensor devices can write appropriate drivers with a minimum of effort and maximum flexibility [49].

4.7 MIDDLE Host Implementation

The MIDDLE Host will be required to i) manage participant recruiting and consent; ii) provide download capabilities for experiment and model likelihood function download to MIDs; iii) manage optimization communication between MIDDLE optimizers and MIDs; iv) archive experiments, models, and statistical optimization provenance trails; v) provide links to and from PUBMED. The MIDDLE project will write a prototype of the MIDDLE Host as part of the cluster simulation in **AIM-1** using standards-based database management tools, network communications protocols.

Much of the MIDDLE Host software requires design and implementation specifics, but does not require novel computer science. However, managing communication, calculation, and storage on personal devices (e.g., smartphones or tablets) will require novel work in distributed parallel computing. Devices could be polled by the MIDDLE Host or could send requests for parameter vectors to the MIDDLE Host. Devices will only have a probability of being available at any one time, and so their computation and data become a probability function. Optimization of design and implementation of the MIDDLE Host will require attention to factors such as communication bandwidth and latency during statistical optimization, management of provenance, and security of consent documents.

4.8 Network Identity Obfuscation and Certificate Management

Our prototype will be based on the evolving standards and reference software implementations of the Nationwide Health Information Network [58] and the CONNECT software toolkit [57]. The best platform currently known to the investigators is that of the Direct Project [53] which will the basis for our model of study participant identifiers, consent documents, anonymization/de-identification, and metadata harmonization. The threat model, against which the Direct communication model is designed, is compliant with that of HL7 and CDA. To the extent possible, we will maintain and demonstrate compliance with applicable standards for health data management and interchange (e.g., HL7 , CDA , and CCD). Where existing and evolving standards are too complex for use within the resources available to this project, we will implement

simplified "stubs" as demonstration placeholders for more robust and standards-compliant solutions that can be plugged in to future implementations based on our reference sandbox source code.

4.9 Consent Options and Risk Management

The user interface for obtaining consent from a MIDDLE participant will need to include a variety of options that are not contained in standard consent documents. For instance, sharing across experiments and risk management options could be included in consent documents. Some individuals might be willing to donate data to a research laboratory and their primary care physicians while others not wish to reveal any data. Some individuals may be willing to allow plots of selected raw data to be generated locally, anonymized, and transmitted for research purposes. Others may wish to only reveal likelihood values, in which case their data could be used by models that calculate the necessary information to create aggregated plots showing group means and confidence intervals. Still others might be willing to reveal historical data but not be willing to participate in new data collection paradigms. The project will design and implement a variety of sample consent documents and ask them to be reviewed by several IRBs for approval. These template consent documents will be reported and incorporated into the reference implementation of sample MIDDLE experiment modules. MIDDLE-style research presents novel opportunities for improved participant consent.

4.10 Documentation for Researchers

Documentation appropriate for NSF and NIH researchers will be written and tested in focus groups and in online discussion forums. The documentation will cover reference implementation details such as building MIDDLE experiment modules, building consent modules with common options, creating and running statistical models, and linking with firewalled facilities using a MIDDLE Host. At regular intervals this documentation project will release Request for Comments (RFCs) so that the community of health science researchers can participate in building understandable and immediately useful documentation. The documentation will be released online on the MIDDLE Project's web site in tutorial, reference, and Wiki format. Open-access discussion forums will be implemented on the web site in order to facilitate continuing discussion outside the context of formal RFCs.

4.11 Outreach to Researchers, Physicians, Industry, and Participants

Outreach is essential if the MIDDLE approach is to be adopted as a research paradigm. This phase of the project will begin with scholarly articles, editorials in the mainstream press, blog postings, radio and TV interviews, and press releases. This outreach will be intended to educate the audience about the paradigm shift inherent in the MIDDLE approach and its benefits for the medical and health science community. One way to amplify outreach is to focus attention on companies in the medical devices industry who might have financial interest in the success of the MIDDLE approach and who, in turn, might engage in public relations efforts to educate the community and publicize the benefits of the approach. We intend to ensure that as the MIDDLE project matures, there will be growing interest in seeing it implemented by many software and hardware manufacturers. The MIDDLE project will provide information and standards that will allow for these implementations.

As the MIDDLE project comes to fruition, the team will focus on educating the public about the approach. Materials with crisp, clear language will be developed and tested in focus groups of the target audience of participants. Online material, mainstream media interviews, and blog posts will be released as soon as sufficient progress has been made and sufficient industry interest has been generated. We expect this phase to begin during the third or fourth year of the project.

5 Timeline

Figure 5 illustrates a timeline for accomplishing the tasks and problems outlined in Section 4. The work can be categorized into six phases: 1) **Research** into existing literature and design of analytic, statistical, or computer science solutions; 2) Development of **Prototype** software or mathematical algorithms; 3) Alpha- or beta- **Testing** of software or algorithms; 4) Production **Coding** of reference software implementations; 5) **Documentation** of the software, interfaces, algorithms, statistical methods, or data management techniques; and 6) Request for Comments (**RFC Period**) when the community is asked to give feedback in web-based forum groups. Between these phases are **Milestone** points when evaluation of algorithms, software, interfaces, and comments from the community are integrated.

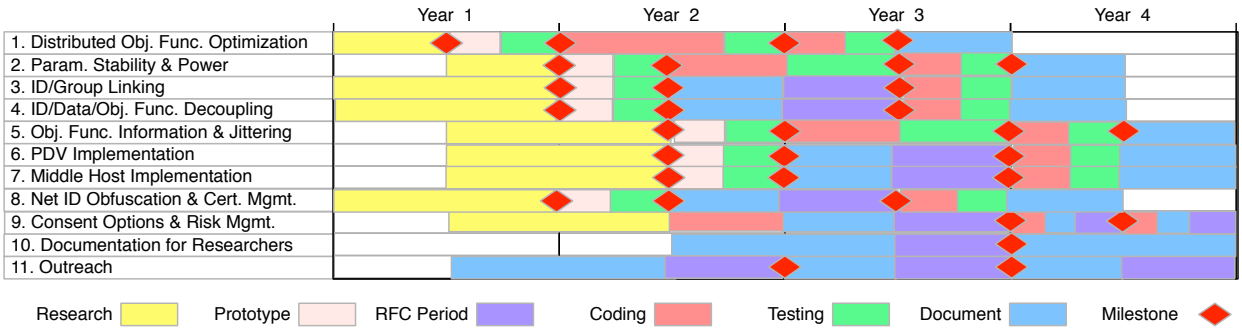

Figure 5. *Timeline for the development phases and milestone/evaluation points.*

Note that milestone points are aligned at 6 month intervals in order to allow for potential interdependency of decisions. All major design and approach decisions will have been made by the middle of year 3. The remaining milestones will concern implementation, user interface, and results from beta-testing. Community discussion will be solicited throughout the project using a dedicated web-based open-access forum and wiki. RFCs will be published and widely advertised on the web. Prior to decision points, even when there is no formal RFC, forum posts will solicit community discussion. Final decisions will be made by consensus of the PIs, CoIs, and consultants. If group consensus cannot be reached, a decision will be made by vote of the three PIs. When an unforeseen problem arises that requires additional expertise, consultant(s) with the requisite expertise will be identified and invited to join the team for the duration of that problem.

6 Broader Impacts

6.1 Benefits To Health Science Research

Accelerated Translation of Research into Practice — As hospitals and primary care physicians install MIDDLE-compatible optimizers, a radically new approach to research translation becomes possible. NSF-funded studies will use the MIDDLE system to develop predictive statistical models. These predictive models and their parameter estimates will be stored and available on the MIDDLE Host. Hospitals and primary care physicians can have immediate access not only to the published articles on PUBMED, but also direct access to the predictive models. These models can be downloaded and immediately used by the healthcare provider to assist in diagnosis or assess complex etiological risks in patients with MIDDLE-compatible home healthcare monitors. We expect that a subscription "push" service (possibly AMA-sponsored) will eventually be established, further reducing the healthcare providers' time and effort for translation of NSF-funded research into medical practice.

Faster and Easier Data Sharing — Data sharing currently requires a five-year wait period. In the MIDDLE system, data sharing with automatic longitudinal linking can occur for on-going studies. Data belong to participants, and so participants can decide to allow data sharing for as many studies as they wish. However,

this means that the MIDDLE Host must be able to detect and estimate the effect of *influential observations*, i.e., individuals who participate in many studies and who might also be overweighted by a representativeness model. This problem currently exists in published results, but until now there has been no way to estimate influence across multiple studies since within-person linkage across studies is currently extremely difficult, when not impossible. Current approaches to inter-study statistics are primarily *meta-analysis* based, but aggregation of aggregates obscures individual etiologies. The MIDDLE approach provides immediate and automatic *mega-analysis*: raw individual-level data from many studies contribute to statistical analyses.

Improved Longitudinal Data for Person-Specific Medicine — Personalized medicine requires person-specific data and models. Data within a MID are automatically longitudinally linked for any study to which the participant consents. Predictive models can be quickly translated from the MIDDLE Host and used by hospitals and primary care physicians to improve diagnoses and prescribe personalized treatment, not only for NSF participants but also for any patient with a MIDDLE-compatible home health monitor.

Reduced Burden and Mitigated Risk for Participants — Since fewer new data are required, either within-person or in terms of sample size, participant burden is decreased per experiment. The accelerated pace of research may absorb this reduction in burden as an individual may choose to participate in more studies. Risk of data exposure is reduced since data are always within the participant's control and not solely stored in centralized location. A participant can opt-out of an experiment at any time and their data do not need to be found and deleted.

More Reliable Methods, Instruments, and Statistical Tests — Methods, instruments and the provenance of the statistical tests will be archived for any article in PUBMED using the MIDDLE system. This will improve quality control of instruments, methods and statistical modeling. Open source sharing of the MIDDLE Host archive contents will maximize researchers' access to these tools and methods. As more researchers use the MIDDLE Host, these improvements will further accelerate.

More Consistent Standards for Data Access and Analysis — The MIDDLE project will define an open source Applications Programmer Interface (API) standard for use with MIDDLE-enabled personal devices. It is widely recognized by manufacturers and researchers that personal health monitors with wireless sensors are one of the main growth markets in personal devices [2, 50]. By setting an open source API standard early, NSF and NIH can influence the intercompatibility of these devices. There are many reasons why device manufacturers may find it profitable to advertise their products as "NSF/NIH Compatible", causing a coalescence in data access and analysis standards.

6.2 Other Benefits

Until now we have focused on how health and behavioral research will benefit. However, the broader impacts of the MIDDLE project are also remarkable. Since the MID and DLE apps, MIDDLE Host system, and MIDDLE reference optimizer will be open source, Apache 2.0 licensed, and standards-based, both commercial and academic uses will be allowed without license fees. Thus it is likely that MIDDLE-compatible systems will be developed by multiple software developers for personal computers, smart phones, tablets, and purpose built wireless sensor devices. There will likely be multiple implementations of the MIDDLE Optimizer, both from commercial software publishers and open source teams that produce systems such as R [40] and OpenMx [20].

Driver Health Monitoring — Any networked sensor system that needs to maintain privacy while monitoring statistical information could use the MIDDLE paradigm. For instance, a MIDDLE system could detect health anomalies for vehicle operators without needing to transmit private health data to a central repository.

Primary Care — Medical technology manufacturers are likely to see the MIDDLE system as an opportunity to manufacture new devices and/or provide new services to physicians. Physicians would buy these products because they would improve patient care while at the same time reducing time spent per case. Patients would approve of these services due to more timely care, improved decision aids for informed treatment choice

[51], and the fact that a smart phone loaded with a standards-based up-to-the-minute medical history would improve the chance of survival in a life-threatening medical emergency [8]: EMS technicians could begin diagnostic analysis while en route to the scene.

Hospitals — Hospitals, in particular Veterans Administration Hospitals, are likely to implement MIDDLE Host and MIDDLE-compatible data repositories because they would be able to automatically take advantage of state-of-the-art statistical models for outlier detection developed by health science researchers. The NSF MIDDLE Host would vastly accelerate the translation of research into medical practice, thereby saving lives.

Individual Ownership of Personal Data — In all of the cases outlined above, personal data remains the property of the individual, thereby revolutionizing the basic economic model of large-scale research both public and private. We believe that this philosophic shift is as economically disruptive as when ownership of private property becomes newly allowed in a formerly command-driven economic system. We predict that a market-driven *personal data economy* will arise as individuals realize that their data is property with worth. It is difficult to predict what innovations will arise from this market-driven personal data economy, but we are confident that the pace of innovation and discovery will be vastly accelerated while, at the same time, risks to individual privacy will be mitigated relative to the current "collectivist" data economy.

7 Results from Prior NSF-Sponsored Research

PI Boker has been funded by several research grants including BCS-0527485/BCS-0742705 ("Collaborative Research DHB: Coordinated motion and facial expression in dyadic conversation," PI S. Boker, $402,618, 1/1/06–12/31/09). *Intellectual Merit:* Boker's group developed novel statistical methods, developed computer software, and performed experiments using facial tracking avatars in videoconference experiments and in "thin slice" emotion rating experiments. We disseminated these results in journal articles, edited chapters, proceedings papers, conference presentations, invited talks, and on the web. Publications resulting from this work include [18, 13, 16, 15, 12, 14, 19, 24, 30, 52, 5, 56, 25]. *Broader Impacts:* The software developed under NSF BCS–0527485 is open source and licensed under Apache 2.0. It uses a software library (DeMOLib) that is licensed under the BSD license. This means that the full software package can be distributed free for use in research institutions. The software is in beta test at the Max Planck Institute for Human Development, the University of Michigan, and Pennsylvania State University.

Lach has been supported by several NSF grants. One representative project is IIS-1065262 ($1,224,000, 8/1/11-7/31/15, SHB: Medium: Non-Intrusive Multi-Patient Fall-Risk Monitoring in Health Care Facilities), the goal of which is to design and deploy body sensor technology that gives health care workers the ability to identify fall prone individuals as soon as mobility impairments arise so that targeted interventions can be made before fall events occur. *Intellectual Merit:* This project has explored on-node signal processing and low power RF transceivers to balance energy consumption and mobility analysis fidelity in wireless on-body inertial sensing, as well as determined the appropriate algorithms to identify fall prone individuals in controlled in- and out-of-lab data collections. Published papers to date include [3, 55, 46, 37, 38, 28, 27, 23, 22, 4]. *Broader Impact:* This project is working toward a new paradigm in elderly patient care to target fall prevention interventions to high fall risk individuals before fall incidents occur, preserving patients health and quality of life while reducing health care costs. Early stages of this capability are being explored with the Northern Virginia Fall Prevention Coalition (NVFPC) and Inova Loudoun Hospital.

8 Literature Cited

[1] Personal data for the public good. Technical report, California Institute for Telecommunications and Information Technology, 2014. http://hdexplore.calit2.net/wp/project/personal-data-for-the-public-good-report/.

[2] H. Almedar and C. Ersoy. Wireless sensor networks for healthcare: A survey. *Computer Networks*, 54:2688–2710, 2010.

[3] I. Armenti, P. Asare, J. Su, and J. Lach. A methodology for developing quality of information metrics for body sensor design, wireless health. *Wireless Health*, 2:1–8, 2012.

[4] P. Asare, R.F. Dickerson, X. Wu, J. Lach, and J.A. Stankovic. Bodysim: A multi-domain modeling and simulation framework for body sensor networks research and design. In *International Conference on Body Area Networks*, volume 72, pages 1–2, 2013.

[5] K. T. Ashenfelter, S. M. Boker, J. R. Waddell, and N. Vitanov. Spatiotemporal symmetry and multifractal structure of head movements during dyadic conversation. *Journal of Experimental Psychology: Human Perception and Performance*, 34(4):1072–1091, 2009. PMID: PMC19653750.

[6] L. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics*, 41:164–171, 1970.

[7] Amos Beimel and Shlomi Dolev. Buses for anonymous message delivery. *Journal of Cryptology*, 16:25–39, 2008.

[8] M. Benocci, C. Tacconi, E. Farella, L. Benini, L. Chiari, and L. Vanzago. Accelerometer–based fall detection using optimized ZigBee data streaming. *Microelectronics Journal*, 41(11):703–710, 2010.

[9] B. A. Berg. *Markov Chain Monte Carlo Simulations and Their Statistical Analysis*. World Scientific, Singapore, 2004.

[10] S. Boker, M. Neale, H. Maes, M. Wilde, M. Spiegel, T. Brick, J. Spies, R. Estabrook, T. Bates, P. Mehta, T. von Oertzen, R. Gore, M. Hunter, D. Hackett, J. Karch, A. Brandmaier, J. Pritikin, M. Zahery, and R. Kirkpatrick. OpenMx 2.0 User Guide, 2014. University of Virginia, Department of Psychology, Box 400400, Charlottesville, VA 22904. http://openmx.psyc.virginia.edu.

[11] S. Boker, M. Neale, H. Maes, M. Wilde, M. Spiegel, T. Brick, J. Spies, R. Estabrook, S. Kenny, T. Bates, P. Mehta, and J. Fox. OpenMx: Multipurpose software for statistical modeling, version 1.2, 2012. http://openmx.psyc.virginia.edu.

[12] S. M. Boker. Dynamical systems and differential equation models of change. In H. Cooper, A. Panter, P. Camic, R. Gonzalez, D. Long, and K. Sher, editors, *APA Handbook of Research Methods in Psychology*, pages 323–333. American Psychological Association, Washington, DC, 2012.

[13] S. M. Boker and J. F. Cohn. Real time dissociation of facial appearance and dynamics during natural conversation. In M. Giese, C. Curio, and H. Bültoff, editors, *Dynamic Faces: Insights from Experiments and Computation*, pages 239–254. MIT Press, Cambridge, MA, 2010.

[14] S. M. Boker and J. F. Cohn. Real time dissociation of facial appearance and dynamics during natural conversation. In C. Curio, H. H. H. Bültoff, and M. Giese, editors, *Dynamic Faces: Insights from Experiments and Computation*, pages 239–254. MIT Press, Cambridge, MA, in press.

[15] S. M. Boker, J. F. Cohn, B.-J. Theobald, I. Matthews, T. R. Brick, and J. R. Spies. Effects of damping head movement and facial expression in dyadic conversation using real–time facial expression tracking and synthesized avatars. *Philosophical Transactions of the Royal Society*, 364:3485–3495, 2009. PMCID: PMC2781890.

[16] S. M. Boker, J. F. Cohn, B.-J. Theobald, I. Matthews, M. Mangini, J. R. Spies, Z. Ambadar, and T. R. Brick. Something in the way we move: Motion dynamics, not perceived sex, influence head movements in conversation. *Journal of Experimental Psychology: Human Perception and Performance*, 37(2):631–640, 2011. PMID: PMC21463081.

[17] S. M. Boker, P. R. Deboeck, C. Edler, and P. K. Keel. Generalized local linear approximation of derivatives from time series. In S.-M. Chow, E. Ferrer, and F. Hsieh, editors, *Statistical Methods for Modeling Human Dynamics: An Interdisciplinary Dialogue*, pages 161–178. Taylor & Francis, Boca Raton, FL, 2010.

[18] S. M. Boker and M. Martin. On the equilibrium dynamics of meaning. In M. Edwards and R. MacCallum, editors, *Current Issues in the Theory and Application of Latent Variable Models*, pages 240–252. Taylor & Francis, New York, 2013.

[19] S. M. Boker, P. C. M. Molenaar, and J. R. Nesselroade. Issues in intraindividual variability: Individual differences in equilibria and dynamics over multiple time scales. *Psychology & Aging*, 24(4):858–862, 2009.

[20] S. M. Boker, M. Neale, H. Maes, M. Wilde, M. Spiegel, T. Brick, J. Spies, R. Estabrook, S. Kenny, T. Bates, P. Mehta, and J. Fox. OpenMx: An open source extended structural equation modeling framework. *Psychometrika*, 76(2):306–317, 2011. PMCID: PMC3525063.

[21] A. Boldyreva, M. Fischlin, A. Palacio, and B. Warinschi. A closer look at PKI: Security and efficiency. In Tatsuaki Okamoto and Xiaoyun Wang, editors, *Public Key Cryptography*, volume 4450 of *Lecture Notes in Computer Science*, pages 458–475. Springer, Berlin, 2007.

[22] B. Boudaoud, H.C. Powell Jr., and J. Lach. Application-informed platform evaluation for commercial-off-the-shelf dynamic voltage scaling. In *International Conference on Electronics Circuits and Systems*, in press.

[23] J.S. Brantley, A.T. Barth, and J. Lach. Optimizing battery lifetime-fidelity tradeoffs in bsns using personal activity profiles. In *International Conference on Body Area Networks*, pages 106–112, 2012.

[24] T. R. Brick and S. M. Boker. Correlational methods for analysis of dance movements. *Dance Research Electronic*, in press.

[25] T. R. Brick, J. R. Spies, B. Theobald, I. Matthews, and S. M. Boker. High–presence, low–bandwidth, apparent 3–d video–conferencing with a single camera. In *Proceedings of the 2009 International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*. IEEE, 2009.

[26] G. Casella and E. George. Explaining the Gibbs sampler. *American Statistician*, 46:167–174, 1992.

[27] S. Chen, J.S. Brantley, T. Kim, S.A. Ridenour, and J. Lach. Characterising and minimising sources of error in inertial body sensor networks. *International Journal of Autonomous and Adaptive Communications Systems*, 6:253–271, 2013.

[28] S. Chen and J. Lach. Nonlinear feature for gait speed estimation using inertial sensors, international conference on body area networks. In *International Conference on Body Area Networks*, pages 185–188, 2013.

[29] K. Davis, S. C. Schoenbaum, and A.-M. Audet. A 2020 vision of patient–centered primary care. *Journal of General Internal Medicine*, 10(10):953–957, 2005.

[30] P. R. Deboeck, S. M. Boker, and C. S. Bergeman. Modeling individual damped linear oscillator processes with differential equations: Using surrogate data analysis to estimate the smoothing parameter. *Multivariate Behavioral Research*, 43(4):497–523, 2008.

[31] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

[32] Roger Dingledine, Nick Mathewson, and Paul Syverson. Tor: the second-generation onion router. In *Proceedings of the 13th conference on USENIX Security Symposium - Volume 13*, SSYM'04, pages 21–21, Berkeley, CA, USA, 2004. USENIX Association.

[33] Ronald A. Fisher. On the mathematical foundations of theoretical statistics. *The Philosophical Transactions of the Royal Society*, 222:309–368, 1922.

[34] S. Fox and M. Duggan. Tracking for health. Technical report, Pew Research Center, 2013. http://www.pewinternet.org/files/old-media//Files/Reports/2013/PIP_TrackingforHealth%20with%20appendix.pdf.

[35] A. Gelman, J. Carlin, H. Stern, and D. Rubin. *Bayesian Data Analysis*. Chapman and Hall, London, 1995.

[36] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741, 1984.

[37] J. Gong, P. Asare, J. Lach, and Y. Qi. Piecewise linear dynamical model for human actions clustering from inertial body sensors with considerations of human factors. In *International Conference on Body Area Networks*, in press.

[38] J. Gong and J. Lach. Reconfigurable differential accelerometer platform for inertial body sensor networks. *IEEE SENSORS*, pages 795–805, 2013.

[39] M. Goodstadt and V. Gruson. The randomized response technique: a test on drug use. *Journal of the American Statistical Association*, 70(352):814–818, Dec 1975.

[40] Ross Ihaka and Robert Gentleman. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5(3):299–314, 1996.

[41] IndivoHealth. Indivo, the personally controlled health record, 2012. http://indivohealth.org. Accessed January 2, 2012.

[42] R. Kainda, I. Flechais, and A. Roscoe. Security and usability: Analysis and evaluation. In *International Conference on Availability, Reliability and Security*, pages 275–282. IEEE, 2010.

[43] R. E. Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME Journal of Basic Engineering*, 82:35–45, 1960.

[44] J. Kivits. Researching the "informed patient". *Information, Communication, & Society*, 7(4):510–530, 2004.

[45] H. Lee, K. Park, B. Lee, J. Choi, and R. Elmasri. Issues in data fusion for healthcare monitoring. In *Proceedings of the 1st international conference on Pervasive Technologies Related to Assistive Environments*. ACM, 2008. DOI: 10.1145/1389586.1389590.

[46] T.E. Lockhart, C. Frames, R. Soangra, and J. Lach. Identifying fall risk in the obese elderly using a wireless sensor and movement complexity. *Wireless Health*, in press.

[47] K. D. Mandl and I. S. Kohane. No small change for the health information economy. *New England Journal of Medicine*, 360(13):1278–1281, 2009.

[48] K. D. Mandl, P. Szolovits, and I. S. Kohane. Public standards and patients' control: How to keep electronic medical records accessible but private. *British Medical Journal*, 322:283–287, 2001.

[49] O. G. Morchon and H. Baldus. Efficient distributed security for wireless medical sensor networks. In *International Conference on Intelligent Sensors, Sensor Networks and Information Processing, 2008*, pages 249–254. IEEE, 2008.

[50] L. Mottola and G. P. Picco. Programming wireless sensor networks: Fundamental concepts and state of the art. *ACM Computing Surveys*, 43(3):19:1–51, 2011.

[51] A. O'Connor, J Wennberg, F. Legare, H. Llewellyn-Thomas, B. Moulton, K. Sepucha, A. Sodano, and J. King. Toward the "tipping point": Decision aids and informed patient choice. *Health Affairs*, 26(3):716–725, 2011.

[52] T. v. Oertzen and S. M. Boker. Time delay embedding increases estimation precision of models of intraindividual variability. *Psychometrika*, 75(1):158–175, 2010.

[53] Direct Project. The Direct Project overview, 2010. http://wiki.directproject.org/file/view/DirectProjectOverview.pdf. Accessed January 2, 2012.

[54] Rob Sherwood, Bobby Bhattacharjee, and Aravind Srinivasan. p^5: A protocol for scalable anonymous communication. *Proceedings of the 2002 IEEE Symposium on Security and Privacy (S&P'02)*, pages 1–13, 2002.

[55] R. Soangra, T.E. Lockhart, J. Lach, and E.M. Abdel-Rahman. Effects of hemodialysis therapy on sit-to-walk characteristics and dynamic stability in esrd patients. *Annals of Biomedical Engineering*, 41:795–805, 2013.

[56] B. Theobald, I. Matthews, M. Mangini, J. Spies, T. Brick, J. F. Cohn, and S. M. Boker. Mapping and manipulating visual prosody. *Journal of Language and Speech*, 52(2):369–386, 2009. PMCID: PMC2716035.

[57] U.S. Department of Health & Human Services. CONNECT community portal, 2011. http://www.connectopensource.org/. Accessed January 2, 2012.

[58] U.S. Department of Health & Human Services. Nationwide health information network, 2011. http://healthit.hhs.gov/portal/server.pt/community/healthit_hhs_gov__nationwide_health_information_network/1142. Accessed January 2, 2012.

[59] G. Virone, A. Wood, L. Selavo, Q. Cao, L. Fang, T. Doan, Z. He, R. Stoleru, S. Lin, and J. A. Stankovic. An assisted living oriented information system based on a residential wireless sensor network. In *1st Transdisciplinary Conference on Distributed Diagnosis and Home Healthcare*, pages 95–100. IEEE, 2006. DOI: 10.1109/DDHH.2006.1624806.

[60] E. R. Weitzman, B. Adida, S. Kelemen, and K. D. Mandl. Sharing data for public health research by members of an international online diabetes social network. *PLoS One*, 6(4):1–8, 2011.

[61] Halbert White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25, 1982.

Data Management

Data Description
The data in this project will consist of the open-source software code written as part of the project and simulated data used to verify statistical routines. All software will be released under the Apache 2.0 license, encouraging others to freely use the code for any project asking in return that the code be properly cited and attributed to its authors.

Format
The software will be kept on a publicly readable revision control system housed on computers in PI Boker's lab.

Sustainability and Dissemination
The computers in PI Boker's lab will perform regular off-line backups. These backups will adequately support all data formats. The software revision control system will also be made available on UVa maintained servers for sustained availability to the public.

Security
Not applicable.

Intellectual Property
UVa has well defined policies with regard to IP that will be strictly followed. All project members will belong to the UVa community, and as such if UVa develops any IP created by this project, it will own that technology. Since the fundamental objectives of the project are basic research, a significant amount of patentable IP is not anticipated. Software developed as part of the project will be open-source and released under the Apache 2.0 License. Standard header files that assigns proper attributions to authors of code will be utilized. In addition, students will be taught to give proper attribution for ideas in papers and presentations.

List of Project Personnel and Partner Institutions

1. **Steven Boker; University of Virginia; PI.**
2. **Donald Brown; University of Virginia; Co-I.**
3. **John Lach; University of Virginia; Co-I.**
4. **Timothy Brick; Pennsylvania State University; Consultant.**
5. **Michael Neale; Virginia Commonwealth University; Consultant.**

List of Project Personnel Collaborations

- Collaborators for Steven Boker; University of Virginia; PI

1. Allen, J. (University of Virginia);
2. Amabadar, Z. (University of Pittsburgh);
3. Ashenfelter, K. (U.S. Bureau of the Census);
4. Bates, T. (University of Edinburgh);
5. Bergeman, C. S. (University of Notre Dame);
6. Borkowski, J. G. (University of Notre Dame);
7. Brick, T. (Pennsylvania State University);
8. Brown, D. (University of Virginia);
9. Burt, S. A. (Michigan State University);
10. Clark, L. A. (University of Notre Dame);
11. Cohn, J. (University of Pittsburgh);
12. Cummings, J. R. (Michigan State University);
13. Deboeck, P. (University of Kansas);
14. Diamond, L. (University of Utah);
15. Dodson, G. (University of Virginia);
16. Edler, C. (University of Iowa);
17. Erbacher, M. K. (James Madison University);
18. Estabrook, R. (Northwestern University);
19. Farr, R. H. (University of Massachusetts)
20. Farris, J. R. (Pennsylvania State University);
21. Fujita, F. (Indiana University, South Bend);
22. Fox, J. (McMaster University);
23. Gasimova, F. (Max Planck Institute for Human Development);
24. Haedt–Matt, A. A. (Michigan State University);
25. Hopwood, C. J. (Michigan State University);
26. Hueluer, G. (Max Planck Institute for Human Development);
27. Hunter, M. D. (University of Oklahoma);
28. Hu, Y. (Texas State University);
29. Keel, P. (Florida State University);
30. Kenny, S. (University of Chicago);
31. Klump, K. (Michigan State University);
32. Lach, J. (University of Virginia);
33. Maes, H. H. (Virginia Commonwealth University);
34. Martin, M. (University of Zurich);
35. Mangini, M. (Concordia College);

36. Matthews, I. (Disney);
37. McArdle, J. J. (University of Southern California);
38. Mehta, P. (University of Houston);
39. Monpetite, M. (Illinois Wesleyan University);
40. Neale, M. C. (Virginia Commonwealth University);
41. Nesselroade, J. R. (University of Virginia);
42. Nicholson, J. S. (St. Jude Children's Research Hospital);
43. Paskus, T. S. (NCAA Research)
44. O'Connor, S. (Michigan State University);
45. Oertzen, T. v. (University of Virginia);
46. Ong, A. (Cornell University);
47. Pettersson, E. (Karolinska Institute);
48. Pritikin, J. N. (University of Virginia);
49. Racine, S. E. (Michigan State University);
50. Robitzsch, A. (Max Planck Institute for Human Development);
51. Rojas, E. E. (Michigan State University)
52. Schmidt, K. M. (University of Virginia);
53. Sisk, C. L. (Michigan State University);
54. Sperry, S. (Michigan State University);
55. Spiegel, M. (Renaissance Computing Center);
56. Spies, J. (University of Virginia);
57. Staples, A. (Indiana University);
58. Sturge-Apple, M. (University of Rochester);
59. Suisman, J. (Michigan State University);
60. Theobald, B.-J. (University of East Anglia);
61. Thompson, J. (Michigan State University);
62. Tellegen, A. (University of Notre Dame);
63. Tiberio, S. (Oregon Social Learning Center);
64. Watson, D. (University of Notre Dame);
65. Wilde, M. (University of Chicago);
66. Wilhelm, O. (Max Planck Institute for Human Development);

- Collaborators for Donald Brown; University of Virginia; Co-I

1. Bass, E. (Drexell University);
2. Boker, S. (University of Virginia);
3. Borne, P. (Ecole Centrale Lille);
4. Breton, M. (UVa);
5. Gorr, W. (CMU);

6. Haimes, Y. (UVa);
7. Kovatchev, B. (UVa);
8. Lach, J. (University of Virginia);
9. Liebowitz, J. (University of Maryland);
10. Lin, S. (AT&T);
11. Llinas, J. (U. Buffalo);
12. Marin, J. (BBN);
13. McGinnis, M. (ODU);
14. Tien, J. (University of Miami);
15. Trainer, T. (U.S. Military Academy)
16. White, C. (GeorgiaTech);

- Collaborators for John Lach; University of Virginia; Co-I

1. Abdel-Rahman, E. (University of Virginia);
2. Acton, S. (University of Virginia);
3. Amy LaViers, A. (University of Virginia);
4. Anderson, M. (Carilion Clinic Center for Healthy Aging);
5. Bankole, A. (Carilion Clinic Center for Healthy Aging);
6. Barnes, B. (University of Virginia);
7. Barth, J. (University of Virginia);
8. Bennett, B. (Northwestern Health Sciences University);
9. Bhansali, S.(Florida International University);
10. Boker, S. (University of Virginia);
11. Brandt-Pearce, M. (University of Virginia);
12. Broshek, D (University of Virginia);
13. Calhoun, B. (University of Virginia);
14. Datta, S. (Pennsylvania State University);
15. Evans, D. (University of Virginia);
16. Freeman, F. (University of Virginia);
17. Ha, D. (Virginia Polytechnic University);
18. Jafari, R. (University of Texas-Dallas);
19. Jur, J. (North Carolina State University);
20. Kerrigan, C. (OESH);
21. Lockhart, T. (Virginia Polytechnic University);
22. Misra, V. (North Carolina State University);
23. Muth, J. (North Carolina State University);
24. Patek, S. (University of Virginia);
25. Peden, D. (University of North Carolina);

26. Qi, J. (University of Virginia);
27. Roberto, K. (Virginia Polytechnic University);
28. Robins, G. (University of Virginia);
29. Rose, K. (University of Virginia);
30. Russell, S. (University of Virginia);
31. Skadron, K. (University of Virginia);
32. Smith-Jackson, T (North Carolina A&T);
33. Specht, J. (University of Iowa);
34. Spruijt-Metz, D. (University of Southern California);
35. Stankovic, J. (University of Virginia);
36. Trolier-McKinstry, S. (Pennsylvania State University);
37. Vallas, C. (University of Virginia);
38. Wentzloff, D. (University of Michigan);
39. Whitehouse, K. (University of Virginia);

Coordination Plan

Personnel Roles

PI. Boker's work has focused on statistical methods for the analysis of within person data and statistical software for the estimation of structural equation models. CoI Brown's work has focused on fusion of networked data. CoI Lach's work has focused on data recording from wireless sensor networks. The current project will merge these three teams for the first time, resulting in a team with unique expertise available to accomplish the aims of the MIDDLE project.

Steven Boker; University of Virginia; PI. Dr. Boker, Ph.D is Professor of Psychology at UVA and directs the Quantitative Psychology program. Dr. Boker will be the Program Director. He will be responsible for: i) personnel decisions; ii) decisions requiring reallocation of resources; iii) final decisions on overall software design when the development team is not in consensus. Dr. Boker will direct the project and facilitate interactions between the team members. He will be responsible for proposing and adjudicating statistical software design in collaboration with the rest of the development team. He will be responsible for coordinating with the work being performed in Dr. Brown's and Dr. Lach's labs. At scientific meetings and workshops he will educate the research community on the implementation of the statistical methods and their implications for data privacy. He will be available for public relations outreach to health sciences communities and to the public through mass media.

Donald Brown; University of Virginia; CoI. Dr. Brown is William Stansfield Calcott Professor of Engineering and Applied Science and the inaugural Director of the University of Virginia Data Science Institute. Dr. Brown will assist in the design of required data fusion and network protocols as well as statistical optimization. At scientific meetings and workshops he will educate the research community on the implementation of the software and statistical methods. He will be available for public relations outreach to the medical community and to the public through mass media.

John Lach; University of Virginia; Co-I Dr. Lach is Professor and Chair of Electrical and Computer Engineering at UVA. Dr. Lach will assist in the integration of the DLE and MID modules into wireless devices as well as coordinating the MIDDLE network protocols with existing networking standards. At scientific meetings and workshops he will educate the research community on the implementation of the software and statistical methods on wireless devices. He will be available for public relations outreach to the medical community and to the public through mass media.

Timothy Brick; Pennsylvania State University; Consultant. Dr. Brick is an assistant professor at Pennsylvania State University. Dr. Brick is a lead software developer for the OpenMx calculation engine and optimizer. His extensive experience with this open source code is invaluable. He will assist in coding the MIDDLE software and be author/co-author of articles resulting from the project.

Michael Neale; Virginia Commonwealth University; Consultant. Dr. Neale is professor of Psychiatry and Human Genetics at Virginia Commonwealth University. He is the original author of the Mx Structural Equation Modeling software and with over 44,000 citations and an H-index over 100 is among the most influential methodologists in the world. His expertise in statistical estimation techniques and in big data problems resulting from whole genome analyses will be invaluable. He will assist in the design of MIDDLE estimation procedures and be author/co-author of articles resulting from the project.

Coordination Mechanisms

Progress will be evaluated at monthly meetings of both site teams to assess milestone achievements, approve any changes to experimental or analytic designs, and/or to reassign responsibilities and resources. Decisions regarding software design and interfacing and/or reallocation of responsibility will be made by consensus. If consensus cannot be reached, these decisions will be adjudicated by PI Boker.

The site at which proposed research is to be conducted is separated from that of the consultants. However, all sites have access to video–conferencing abilities within their labs using computers connected to the internet. Monthly video–conference meetings between the UVA team members and the consultants will be scheduled. Funds are provided in the budget justification for continuing licenses for the videoconferencing software. In addition, there will be face–to–face meetings on–site at UVA attended by all team members. Funds are provided in the budget for the consultants to travel to these meetings. Additional meetings will be arranged in connection with conference travel.

Fiscal and Management Coordination

Dr. Boker will serve as contact PI and will assume overall responsibility for fiscal and administrative management. Dr. Boker will be responsible for communication with NSF and submission of annual reports. PI Boker, CoI Brown, and CoI Lach will discuss all fiscal and administrative management issues and fully engage in collaborative problem solving efforts should issues arise.

Procedures for Resolving Conflicts

If a conflict develops, PI Boker, CoI Brown, and CoI Lach will meet and attempt to resolve the dispute. If they fail to resolve the dispute, the appropriate departmental administrators representing the PIs will meet and attempt in good faith to settle the dispute, claim or controversy arising out of or relating to the interpretation, performance or breach of the leadership or research plan. If the departmental administrators fail to resolve the disagreement within 30 business days, the disagreement will be referred to Thomas Skalak, PhD, Vice President for Research, University of Virginia.

Communication Among Investigators and Data Sharing

Planning and development issues will be discussed at monthly joint lab meetings. In addition there will be regular weekly software development meetings where interface and design issues will be discussed and resolved.

Publication and Intellectual Property Policies

Publication authorship will be based on the relative scientific contributions of PI Boker, CoI Brown, CoI Lach, and other Senior/Key personnel. Authorship will be discussed by all key personnel and determined by consensus of the team members. PIs Boker, CoI Brown, and CoI Lach will be included on all publications resulting from the project.

Change in PI Location

If Dr. Boker moves to a new institution, every attempt will be made to transfer the relevant portion of the grant to the new institution. In the event that PI Boker cannot carry out his duties, either CoI Brown or CoI Lach will be recruited as a replacement.

Human Subjects Protection

Not applicable.

Postdoc Mentoring Plan

This mentoring plan establishes a framework for the postdoctoral researchers who will support this project. One postdoctoral researcher will be based at the University of Virginia and will work to connect the multiple threads of research contained in this proposal. The postdoctoral researcher will be mentored by PI Boker, CoI Brown, and CoI Lach.

Orientation and Core Research Agenda

The postdoctoral researchers' tenure will begin with an in-depth conversation with their respective PIs about their individual career goals and existing research strengths. From this, a tailored plan will be developed for the postdoctoral researcher that aligns with the goals of the project and her or his individual interests. The postdoctoral researcher will be mentored in technical areas, as appropriate to her or his experience and interests. These areas will include design, performance, analysis and evaluation of distributed statistical analysis, estimation of models when data are in flux, and mathematical analysis of performance of these algorithms. The postdoctoral researcher will be mentored in designing research sub-projects and developing and carrying out novel, cross-disciplinary approaches to solving these problems. The postdoctoral researcher will be paired with a graduate student so that she or he will have the opportunity to mentor and to learn how to incorporate a student effectively in research. In concert with her or his research, the postdoctoral researcher will learn about conceptualizing research projects, including writing grant proposals.

The postdoctoral researcher will have the opportunity to work with the mentors' labs, lead important segments of work, and gain experience in both experimental and analytical research. As Boker, Brown, and Lach have experience in cross-disciplinary collaborations, they will serve as appropriate mentors for the cross-disciplinary work of the post doctoral researcher. They will mentor him or her by having him or her participate in a combination of research, teaching, and outreach activities that will lead to a successful portfolio to use in applying for jobs and will serve them well throughout their career. The National Academies of Science and Engineering guidance on how to enhance the postdoctoral experience will be followed. The postdoctoral researchers will also complete training in the conductance of responsible research.

Professional Development and Career Launching

Boker, Brown, and Lach will work with the postdoctoral researcher on communication skills, including writing technical papers, presenting to technical audiences, presenting to broad audiences and teaching. The postdoctoral researcher will have opportunities to present their research regularly at weekly project and lab meetings. They will also be expected to present at conferences and workshops, and will have the opportunity to give lectures as part of the courses taught by the PIs. The postdoctoral researcher will also have the opportunity to participate in outreach programs; this will give additional experience in communicating to diverse audiences, teaching and mentoring.

Facilities, Equipment, and Other Resources

8.1 Human Dynamics Lab (HDL)

Dr. Boker's HDL is located in Gilmer Hall on campus at the University of Virginia and is comprised of 4 rooms: (1) a 4.5m^2 machine room with soundproofing, separate air conditioning, and 4800 watt uninterruptible power supply, (2) a 30.5m^2 motion capture room, (3) a second 30.3m^2 motion capture room with a conference seating area, and (4) a sound isolated 17.8m^2 control room with sound isolated windows looking into each of the motion capture rooms. The two motion capture rooms are separated by a sound isolated non–ferrous wall with no electrical or plumbing in it so that a single magnetic field can span the two motion capture rooms. Each motion capture room has an 8.9m^2 non–ferrous stage composed of six reconfigurable $1.2\text{m} \times 1.2\text{m} \times 0.3\text{m}$ sections. Desks, tables, and chairs are available to seat 5 research assistants.

Dr. Boker's office and graduate students offices are located off–campus at the 1023 Millmont Building which houses the quantitative psychology area. Dr. Boker's 18.5m^2 office and 19.6m^2 conference room and three graduate student offices at 1023 Millmont are available to the project.

Major Equipment (HDL): Computer equipment in the lab and available to the project include: 1 Apple 4-core 3.0 GHz Xserve with 8 GB RAM and 250 GB of storage; 3 2007-vintage Apple MacPros with 8 GB of RAM and 250 GB disk. There is an internal SAN linking all of the computers in the control room and machine closet to the XRAID via dual 4Gb fiberchannel links through a fiberchannel switch and SAN software. In order to isolate and manage network traffic within the lab, there is are two separate internal gigE ethernet networks (one for the SAN metadata and one for other network traffic) tying together all the machines with CAT7 shielded network wiring and two managed switches, one with two gigE uplinks to the campus fiber optic backbone.

Video and audio equipment available to the project include 2 Panasonic IK-M44H genlockable "lipstick" color video cameras, three MOTU V3HD video capture converters, two 3–panel Numark LCD video monitors, two Earthworks directional microphones, a Yamaha 01V96 multichannel digital audio mixer, a Kramer 8×8 video switcher, a Z-Systems 8×8 ADAT format optical audio switcher, 6 1×6 video distribution amps, a Hottronic video frame delay, an AV Toolbox genlockable XVGA scan converter, a Burst Electronics blackburst generator, a Presonus 6 channel headphone amp, and a Canon GL-2 mini DV camcorder. All video timestamping comes from an ESE 185-U GPS synchronized master clock which outputs a genlocked SMPTE time code.

A 47m^2 AccessGrid video conferencing facility with 4 cameras and a 4 projector wall is installed in the 1023 Millmont Building and will be available to the project for multiway video conference meetings with the Co-PIs and consultant.

8.2 Computer Engineering

The computer engineering group in the UVA ECE Department has approximately 2,500 square feet of research laboratory space containing more than $25 million in hardware and software. The facilities are partitioned into several laboratories including the Integrated Circuits Laboratory (ICL), the Embedded Systems Laboratory (ESL), the Systems Integration Laboratory (SIL), and the High-Performance Low-Power (HPLP) Laboratory. As the names imply, the facilities provide a complete environment for the design, fabrication, and testing of prototype hardware/software systems from initial concept to final implementation.

The Integrated Circuits Laboratory (ICL) has a comprehensive infrastructure of software and computing hardware to support the design and test of digital electronic systems. These tools include numerous state-of-the-art design capture and simulation systems operating on high-performance workstations. The available EDA software from Mentor Graphics and Cadence provides automation at any point in the four major phases of the design process (design capture, simulation, physical layout, and test). The tools support several

different design entry mechanisms including schematic capture and hardware description languages (VHDL and Verilog). Multiple levels of simulation are available including system, gate, and circuit levels. These different levels can be intermixed within a single design to perform more efficient simulations of complex systems. Also included in the tool suite are advanced integrated circuit development tools to create complex custom or semi-custom parts. Additional CAD software is provided by Vantage Analysis Systems, Cascade Design Automation, Summit Design, Tanner Tools, Actel, Meta-Software, MicroSim, Lucent Technologies, and ICED.

The Embedded Systems Laboratory (ESL) is a combination research and educational lab with a local network of high-performance PCs and a Windows server. The lab includes many prototyping boards, and the PCs all have data acquisition cards for the many external systems tests that are performed. Two modern Tektronix Logic Analyzers support the hardware design and debugging efforts, and a BP Universal Device Programmer enables the rapid prototyping of systems on a wide variety of programmable logic devices.

The Systems Integration Laboratory (SIL) supports printed circuit board (PCB)-level prototype development. The lab contains several digital storage oscilloscopes and one high-speed analog oscilloscope. Other general-purpose test equipment includes digital multimeters, power supplies, and function generators. Finally, the lab also contains an SRT Sierra series surface mount machine that allows for the assembly and population of single or double-sided surface mount PCBs.

The High-Performance Low-Power (HPLP) laboratory performs research in the area of low power design and design automation for high performance systems. In addition to the department computing facilities, the HPLP lab has exclusive access to a Sun Ultra and two IBM AIX workstations as well as eight Pentium-III systems with workstation-grade monitors for running IC design software. For IC and FPGA testing, the labs operate a HP 82000-D100 tester which can present the test vectors used in simulations to the physical device. Also available for use with the HP82000 is a Tektronix HFS 9003 stimulus system capable of generating stimuli in the GHz range and an HP54120B digitizing oscilloscope mainframe system.

Current and Pending Support

Steven Boker

Current

- **Adolescent Peer and Family Relationship Predictors of Adult Health** (PI: Allen; role: Co-I). Source: National Institute on Child Health and Human Development (2R01HD058305-17A1). 07/10/2008 – 6/30/2018. Amount: $2,756,492 Total Costs.

Pending

- **Psychometric and Genetic Assessments of Substance Use** (PI: Neale; role: Co-I, Subcontract Consortium PI). Source: National Institute on Drug Abuse. 02/01/2015 – 1/31/2020. Amount: $555,550 Subcontract Total Costs.
- **Methodology for tracking and measuring interpersonal affective communication processes** (PI: Boker). Source: National Cancer Institute. 07/01/2015 – 6/30/2020. Amount: $2,273,139 Total Costs.
- **CompCog: RI: Medium: Spatiotemporal Dynamics of Affective Communication in Dyadic Conversations via Data-driven Style-based Robotic Confederates** (PI: LaViers; role: Co-I). Source: National Science Foundation. 08/01/2015 – 07/30/2019. Amount: $xxx Total Costs.
- **SCH: INT: Maintained Individual Data Distributed Likelihood Estimation (MIDDLE)** (PI: Boker). Source: National Science Foundation. 08/01/2015 – 07/30/2019. Amount: $1,526,366 Total Costs. (This proposal)

John Lach

Current

- **SCH: INT: Collaborative Research: BESI: Behavioral and Environmental Sensing and Intervention for Dementia Caregiver Empowerment** (PI)
  Source of Support: National Science Foundation Total Award Amount: $574,951 (UVa portion) Total Award Period Covered: 09/01/14 - 08/31/18 Location of Project: University of Virginia Person-Months Per Year Committed to the Project: 0.75 Calendar Months

- **NSF Nanosystems Engineering Research Center for Advanced Self-Powered Systems of Integrated Sensors and Technologies (ASSIST)** (co-PI) Source of Support: National Science Foundation Total Award Amount: $18,494,327 Total Award Period Covered: 09/01/12 - 08/31/17 Location of Project: University of Virginia Person-Months Per Year Committed to the Project: 1.0 Calendar Months

- **SHB: Type I (EXP): Personalized Signal Processing for Early Diagnosis of Mobility Impairment** (co-PI) Source of Support: National Science Foundation Total Award Amount: $360,128 Total Award Period Covered: 09/01/12 - 08/31/15 Location of Project: University of Virginia Person-Months Per Year Committed to the Project: 0.25 Calendar Months

- **SHB: Medium: Non-Intrusive Multi-Patient Fall-Risk Monitoring in Health Care Facilities** (co-PI) Source of Support: National Science Foundation Total Award Amount: $1,200,000 Total Award Period Covered: 08/01/11 - 07/31/15 Location of Project: Virginia Tech, University of Virginia Person-Months Per Year Committed to the Project: 0.5 Calendar Months

- **Ultra Low Power Processing in Wireless Sensor Nodes** (PI) Source of Support: US Army Night Vision Laboratory Total Award Amount: $355,000 Total Award Period Covered: 01/01/09 - 12/31/14 Location of Project: University of Virginia Person-Months Per Year Committed to the Project: 0.5 Calendar Months

- **CDI-Type I: Accelerating Simulations Using CPU+FPGA Heterogeneous Processing** (co-PI) Source of Support: National Science Foundation Total Award Amount: $595,518 Total Award Period Covered: 09/15/11 - 08/31/15 Location of Project: University of Virginia Person-Months Per Year Committed to the Project: 1.0 Calendar Months

- **Correlates Among Nocturnal Agitation, Sleep, and Urinary Incontinence in Dementia** (co-PI) Source of Support: National Institutes of Health Total Award Amount: $428,269 Total Award Period Covered: 04/01/12 - 03/31/15 Location of Project: University of Virginia, University of Iowa Person-Months Per Year Committed to the Project: 0.6 Calendar Months

Pending

- **STTR Phase I:Gait Tracker Shoe for Long Term Accurate Determination of Gait Parameters and Activity** (co-PI) Source of Support: National Science Foundation Total Award Amount: $125,662 (UVa portion) Total Award Period Covered: 01/01/15 - 12/31/15 Location of Project: OESH Shoes, University of Virginia Person-Months Per Year Committed to the Project: 0.5 Calendar Months

- **ViCTER: Phase II Studies of Gamma Tocopherol as an Intervention for Environmental Asthma** (co-PI) Source of Support: National Institutes of Health Total Award Amount: $334,152 (UVa portion) Total Award Period Covered: 05/01/14 - 04/30/17 Location of Project: University of North Carolina - Chapel Hill, North Carolina State University, University of Virginia Person-Months Per Year Committed to the Project: 1.2 Calendar Months

- **THz Resonance Spectroscopy of HPV-Related Molecular Components for Cervical Cancer Early Detection and Prognosis** (co-PI) Source of Support: National Institutes of Health Total Award Amount: $3,480,959 Total Award Period Covered: 07/01/15 - 06/30/20 Location of Project: University of Virginia Person-Months Per Year Committed to the Project: 0.5 Calendar Months

- **REU Site: Wireless Health** (PI) Source of Support: National Science Foundation Total Award Amount: $299,077 Total Award Period Covered: 09/01/15 - 08/31/18 Location of Project: University of Virginia Person-Months Per Year Committed to the Project: 0.12 Calendar Months

- **Using Collaborative Teaching and Learning to Revolutionize an Electrical and Computer Engineering Department** (PI) Source of Support: National Science Foundation Total Award Amount: $1,998,281 Total Award Period Covered: 09/01/15 - 08/31/20 Location of Project: University of Virginia Person-Months Per Year Committed to the Project: 0.25 Calendar Months

- **SCH: INT: Maintained Individual Data Distributed Likelihood Estimation (MIDDLE)** (PI: Boker). Source: National Science Foundation. 08/01/2015 – 07/30/2019. Amount: $1,526,366 Total Costs. 0.5 Calendar Months (This proposal)

Donald Brown

Current

- **Asymmetric Threat Tracking for Expeditionary and Counterinsurgency Basing** (PI: Brown) Dates: 1/13-12/15 Source: Technology Service Corporation (U.S. Army Rapid Innovation Fund) Co-PIs: Matthew S. Gerber Amount: $600k
- **Asymmetric Threat Tracking for Force Operating Base Sustainment Planning** (PI: Brown) Dates: 1/13-12/15 Source: Technology Service Corporation (U.S. Army Rapid Innovation Fund) Co-PIs: Matthew S. Gerber Amount: $600k

Pending

- **SCH: INT: Maintained Individual Data Distributed Likelihood Estimation (MIDDLE)** (PI: Boker). Source: National Science Foundation. 08/01/2015 – 07/30/2019. Amount: $1,526,366 Total Costs. 0.5 Calendar Months (This proposal)