RatSWD Working Paper Series

254

Maintained Individual Data Distributed Likelihood Estimation (MIDDLE)

Steven M. Boker, Timothy R. Brick,
Joshua N. Pritikin, Yang Wang,
Timo von Oertzen, Donald Brown,
John Lach, Ryne Estabrook,
Michael D. Hunter, Hermine H. Maes,
Michael C. Neale

January 2016

Working Paper Series of the German Data Forum (RatSWD)

The *RatSWD Working Papers* series was launched at the end of 2007. Since 2009, the series has been publishing exclusively conceptual and historical works dealing with the organization of the German statistical infrastructure and research infrastructure in the social, behavioral, and economic sciences. Papers that have appeared in the series deal primarily with the organization of Germany's official statistical system, government agency research, and academic research infrastructure, as well as directly with the work of the RatSWD. Papers addressing the aforementioned topics in other countries as well as supranational aspects are particularly welcome.

RatSWD Working Papers are non-exclusive, which means that there is nothing to prevent you from publishing your work in another venue as well: all papers can and should also appear in professionally, institutionally, and locally specialized journals. The *RatSWD Working Papers* are not available in bookstores but can be ordered online through the RatSWD.

In order to make the series more accessible to readers not fluent in German, the English section of the *RatSWD Working Papers* website presents only those papers published in English, while the German section lists the complete contents of all issues in the series in chronological order.

The views expressed in the *RatSWD Working Papers* are exclusively the opinions of their authors and not those of the RatSWD or of the Federal Ministry of Education and Research.

The RatSWD Working Paper Series is edited by:

Chair of the RatSWD
(since 2014 Regina T. Riphahn; 2009-2014 Gert G. Wagner; 2007-2008 Heike Solga)

Maintained Individual Data
Distributed Likelihood Estimation
(MIDDLE)

Steven M. Boker
University of Virginia

Timothy R. Brick
The Pennsylvania State University

Joshua N. Pritikin   Yang Wang   Timo von Oertzen
Donald Brown   John Lach
University of Virginia

Ryne Estabrook
Northwestern University

Michael D. Hunter
University of Oklahoma

Hermine H. Maes   Michael C. Neale
Virginia Commonwealth University

September 10, 2015

Abstract

Maintained Individual Data Distributed Likelihood Estimation (MIDDLE)
is a novel paradigm for research in the behavioral, social, and health sci-
ences. The MIDDLE approach is based on the seemingly-impossible idea
that data can be privately maintained by participants and never revealed
to researchers, while still enabling statistical models to be fit and scien-
tific hypotheses tested. MIDDLE rests on the assumption that participant
data should belong to, be controlled by, and remain in the possession of the
participants themselves. Distributed likelihood estimation refers to fitting
statistical models by sending an objective function and vector of parameters
to each participants' personal device (e.g., smartphone, tablet, computer),
where the likelihood of that individual's data is calculated locally. Only
the likelihood value is returned to the central optimizer. The optimizer
aggregates likelihood values from responding participants and chooses new
vectors of parameters until the model converges. A MIDDLE study provides
significantly greater privacy for participants, automatic management of opt-
in and opt-out consent, lower cost for the researcher and funding institute,
and faster determination of results. Furthermore, if a participant opts into
several studies simultaneously and opts into data sharing, these studies au-
tomatically have access to individual-level longitudinal data linked across all
studies.

Introduction

A revolution is in progress in the health, behavioral, and social sciences. Up until
recently, most research in these fields was investigator-driven and involved either randomized
controls in a laboratory setting or epidemiological/observational in natural settings. In the
1990s, small portable devices began to be used by investigators to bridge the gap between
these two paradigms in what was termed *ecological momentary assessment* (Shiffman, Stone,
& Hufford, 2008; Stone & Shiffman, 1994). This idea started slowly in part because it was
expensive to provide devices for each individual in a study. But in the 2000's smart phones
changed the landscape. As of July, 2014, comScore reported 173 million people in the U.S.
owned smartphones, of which 51.5% were running Google's Android and 42.4% were running
Apple's iOS (comScore, 2014). In July 4, 2014, the United States Census Bureau reported

the population of the U.S. to be 319 million (United States Census Bureau, 2014). Thus, more than half the U.S. population owns a smartphone and these could be used as data gathering devices for software applications running on the two most popular smartphone operating systems.

This is an incredible opportunity for investigators interested in multivariate behavioral research. But this revolution is moving faster than our sciences have reacted. Individuals are interested in tracking their own data and the software industry has responded by providing a wide array of self-monitoring applications. The 2013 Pew Foundation's Tracking for Health study (Fox & Duggan, 2013) reported that 69% of Americans track some form of personal health data and 21% of Americans do so on a personal digital device. One way of thinking about this phenomonon is as a bottom-up movement doing self-organizing person-oriented "citizen science". Not only do potential participants have personal devices that can act as a data gathering and computing platform, but many are willing to spend money in order to participate in research that can give them information about themselves.

It is possible that these individuals might not be willing to participate in organized research directed by a scientist. However, a recent California Institute for Telecommunications and Information Technology (CALIT) study (*Personal Data for the Public Good*, 2014) reported that 75% of participants "probably" or "definitely" would be willing to share their personal health data with qualified researchers. Assuming that this result generalizes and ignoring correlations between group memberships, a naive estimate of the potential participant pool with smartphones and willing to consider participation in some form of personal health research is approximately one third of the population of the U.S. — over 100 million people.

This seems as if it is a dream come true for the investigators in the health, behavioral and social sciences. However, there is a fairly serious problem that needs to be overcome before these self-organized citizen scientists can act as an engine of scientific discovery in multivariate behavioral research: participant privacy. With no concern for individual privacy, private industry has been busily gathering and mining data that is knowingly or unknowingly provided while individuals use their smartphones. But scientists must be held to a higher standard. We must protect the privacy of personally identifiable information. In fact, potential participants agree with this view. In the CALIT study cited above, 67% of respondents 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. It seems likely that cellphone owners would be angry when they find that data that they believe they own is being used by private industry without their knowledge, and there is some evidence of this in reactions to recent revelations by Facebook (Goel, 2014). Some method must be found in order address the privacy problem and give participants the ability to maintain possession and control of their data.

In current research practice, participant's data are centralized into a data repository. If ownership of the data is claimed by participants, it might seem as if centralized data are being held in trust for the participants. But legal precedent does not agree with this view. In a 2006 court case where the original investigator, Dr. William Catalona, and research participants both wanted data to be transferred from Washington University's data repository to a repository at Northwestern University, Judge Stephen Limbaugh ruled

against Northwestern University and wrote an opinion arguing that it was undisputed that Washington University was in "exclusive possession and control" of the data repository and that "control of personal property is *prima facie* evidence of ownership and anyone else claiming such property bears the burden of proof" (The Washington University, Plaintiff, v. William J. Catalona, et al., Defendants, March 31, 2006). Thus, if data are out of the participants' possession and control the court appears to argue that they are no longer owned by the participant. Thus, current research practice is at odds with participants' belief that they should maintain ownership of their data.

Data collection using participants' devices such as personal computers, smartphones, and tablets has become increasingly common over the past 20 years (see Dufau et al., 2011; Miller, 2012, for reviews). Methods associated with this type of data collection are often longitudinal in nature and go by names such as ecological momentary assessment (Shiffman et al., 2008; Stone & Shiffman, 1994), experience sampling (Csikszentmihalyi & Larson, 1987; Hektner, Schmidt, & Csikszentmihalyi, 2006; Larson & Csikszentmihalyi, 1983), and intensive longitudinal designs (Walls & Schafer, 2005). As technology has advanced, data from these distributed experiments has been collected by email (Boker & McArdle, 1998), web browsers (Greenwald & Nosek, 2001), social networking applications (Anderson, Fagan, Woodnutt, & Chamorro-Premuzic, 2012), and most recently by smartphone, tablet, and wearable computing applications (Benocci et al., 2010; Miller, 2012). Although the technology for data collection has improved remarkably, the basic scientific workflow has remained the same in that data are first *collected* into some central repository and only then are they analyzed.

Why are data needed for research into human health and behavior? 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-level data provide bases for statistical models of each individual's pschological or health processes to be used in personalized diagnosis and intervention decisions. Are the data themselves the goal? No. In both instances, the goals are scientific discovery and/or decision support. Thus, the problem fundamentally entails questions of statistical research methods: What information is necessary and sufficient; and how can this information reach the scientists/decision makers with a minimum risk of disclosure of the original data?

Let us assume that a researcher wishes to answer scientific questions and make diagnostic predictions, but due to privacy concerns does not desire ownership of participants' data. This way of thinking about privacy 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 any individual could opt-in to multiple analyses. Data would remain under participants' possession and control and thus participants would maintain legal ownership. Privacy maintenance would be similar to any other question of ownership: Each person would have responsibility for maintaining control of her or his own possessions.

The MIDDLE Research Paradigm

We take the position that data belong to individuals and should remain in their possession (Mandl & Kohane, 2009; Weitzman, Adida, Kelemen, & Mandl, 2011) 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.

We propose that data be retained by participants in what we call *Maintained Individual Data* (MID), a software platform that would reside on a smartphone, home health monitor, or computer in the possession and control of the participant. Data remain where they were originally collected — on each participant's personal smartphone, computer, tablet, or wearable computing device — and remain private, that is to say these data are never revealed by the participant. If an individual decides to participate in research, they consent to allow the investigator to run a likelihood calculation on their data. The only thing a participant reveals is the likelihood of their data conditional on the investigator's hypothesis.

What we call *Distributed Likelihood Estimation* (DLE) refers to the process by which a central optimizer sends vectors of free parameters for each 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.

When using the MIDDLE research paradigm, data can be collected at the same time that models are optimized. This means that when sufficient data are collected to reach a pre-selected statistical power, the study can automatically terminate or switch to a cross-validation regime. This means that individual-level variables, repeated measurements, and time series are automatically linked and model parameters are estimated at the individual level first and only then does aggregation happen. Furthermore, if a participant opts into many studies simultaneously and consents to data sharing between studies, all of the studies automatically have real-time data sharing.

Adoption of the MIDDLE approach is likely to result in a substantial increase in research efficiency and reduction in long-term cost of scientific progress. Some of the benefits to science, to individual researchers, and to research participants include:

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.

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 or modified 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 without attaining a significant result.

In the remainder of the article we first present a broad overview of traditional experimental designs and brief description of the statistical estimation technique called full information maximum likelihood (FIML). We then present a discussion of how the MIDDLE research paradigm could be implemented into a practical workflow using a communications and provenance archiving hub we call the MIDDLE Host. We next present a simulation demonstrating the feasibility of the MIDDLE optimization paradigm. A hypothetical example MIDDLE experiment is next described in order to facilitate understanding of how this workflow would be experienced by participants and investigators. Finally, we present concerns and limitations that must be kept in mind as a distributed approach such as MIDDLE is implemented. It is important to remember that not all behavioral and social science experiments are amenable to the MIDDLE approach.

Traditional Experimental Design

Traditional best-practices for behavioral, social science, and physiological research share a common sequence of events. First a hypothesis is generated. Next, an experiment that would test the hypothesis is designed and approval from the relevant Institutional Review Board (IRB) is obtained. Participants are then recruited and those that consent are enrolled to participate in the experimental protocol. Data are collected and centralized into a data repository that typically resides on a computer in a locked room in a research laboratory. These data are considered to belong to the research group that performed the experiment or collected the observational data. After the data collection phase, the centralized data are analyzed in order to estimate parameters and goodness of fit of candidate statistical models in order to compare models and test hypotheses. Results are then disseminated through journal articles and/or conference talks and posters. Finally, the experimental results are replicated and/or new hypotheses are generated. Figure 1 presents a flowchart of this process.

Consider six potential problems that accompany this approach to research: i) Who owns the data? Is it the research group who collected the data, the agency that funded the research, or do the participants retain ownership of their own data? ii) The centralized and private nature of the data repository tends to discourage open disclosure of the provenance chain of statistical analyses and thus reduces the chance that mistakes in analyses will be
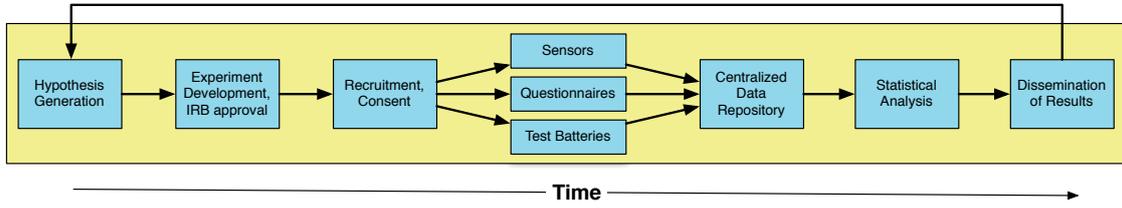
Figure 1. Traditional experimental design. In traditional designs for human subjects research, steps are performed sequentially, data are stored centrally, and statistical analysis is conducted only after the data collection portion of the experiment is concluded. This leads to long intervals between experiment replication and/or hypothesis revision as well as barriers to data sharing.

caught by reviewers. iii) During the process of designing a new experiment, power analyses (when they are conducted) are encouraged to be conservative, since one does not know the effect size in advance. iv) There can be a very long interval between the generation of a new hypothesis and the next opportunity for the hypothesis to be revised or the experiment replicated. v) Barriers to data sharing include protecting the confidentiality of the participants and a potentially long lag time during which the research group has sole right of publication using the data. vi) It is difficult, if not impossible, to perform longitudinal linking between different data repositories while maintaining participant confidentiality. Thus longitudinal studies tend to be isolated from the benefits of data sharing.

Traditional Likelihood Estimation

In order to provide a framework for the proposed methodology, let us first review in broad strokes one commonly used statistical estimation procedure: full information maximum likelihood (FIML). A data matrix resulting from an experiment is selected for analysis. A statistical model is built or selected which has a set of model parameters to be estimated. Starting values for the model parameters are then selected. Given starting values for the parameters, the model implies an expected covariance and means structure for the data. For each row of the data matrix the likelihood of the data is calculated given the model parameters. These individual likelihoods are logged and summed and a test is performed to see if this summed log likelihood is at a maximum. If the summed log likelihood is not at a maximum, a new set of parameters are chosen by the estimation software and new likelihoods are calculated. If the summed log likelihood is at a maximum, the software returns the summed log likelihood and the current parameter estimates. A simplified flowchart of FIML estimation is shown in Figure 2–a.

Maintained Individual Data (MID)

We propose a scientific software layer that would run on participants' personal devices (smartphones, tablets, or personal computers) and would serve to communicate between data acquisition applications and the research group sponsoring a study. One might think of this software layer as something like a personal data vault with encryption and a constrained
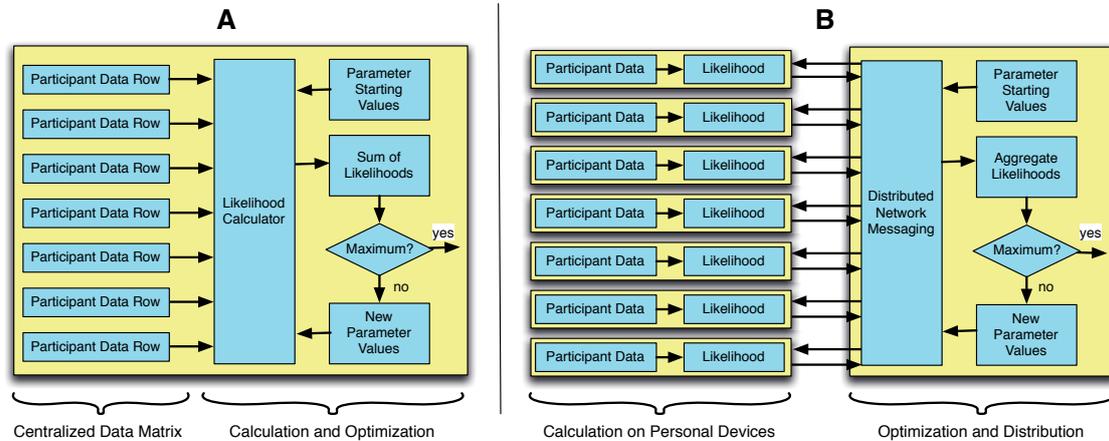
A

B

Figure 2. Simplified flowcharts for traditional full information maximum likelihood (FIML) estimation and distributed likelihood estimation (DLE). (a) In FIML, the data are centralized into a data matrix and the likelihood of each row of the data matrix contributes to the summed log likelihood. Parameters are adjusted until the summed likelihood is at a maximum. (b) In DLE, the data resides on participants' personal devices. The personal device receives model parameters from the central optimizer, calculates the likelihood of a participant's data, and passes only the likelihood back to the central optimizer. The central optimizer calculates the summed log likelihood, and if necessary, chooses new parameters to redistribute to the personal devices. We use FIML as an example although Bayesian or other optimization methods for parameter estimation can be used in DLE.

set of encrypted communication protocols. We call this software layer Maintained Individual Data (MID).

The MID software would provide several important functions. First, it would act as an intermediary that would allow research groups to advertise for participants. The MID owner could browse for studies in which she or he wanted to participate. Second, the MID would manage consent forms in a standardized way so that the potential participant could opt-in or opt-out of an experiment at any time. Third, the MID would provide a data socket that would communicate with the computer/smartphone application that presented the experimental stimulus, questionnaire, game, or other data acquisition method. Fourth, the MID would encrypt and maintain all data that was acquired by a MID-compliant application. Finally, provided appropriate participant consent, the MID would communicate with research groups that wished to test specific statistical models on the participants' data. Note that the group that tests a statistical model need not be the group that originated the experiment, testing a model would only require the MID owner's consent.

Many computer applications have been proposed to organize and store health data for large scale applications (Dolin et al., 2006; Vreeman, McDonald, & Huff, 2010) and on personal devices (IndivoHealth, 2012; Microsoft, 2012). However, these applications are designed to simply organize and protect personal health data and are not designed to participate in statistical model fitting outside of their own individual firewalls. We next propose a method that would allow large-scale networks of devices running MID software

to participate in scientific experiments without revealing the data stored by MID on the personal device.

Distributed Likelihood Estimation (DLE)

Distributed likelihood estimation (DLE) refers to a method for estimating maximum likelihood parameters and fit statistics for statistical models that is very similar to the FIML method described earlier. Figure 2–b presents a simplified flowchart of the DLE optimization procedure. As in FIML, a model and starting values are selected by the researcher. The model and parameter values imply a covariance and means structure such that the likelihood of a single participant's data can be calculated. The model and parameters are distributed from a central optimization server (the DLE server) to all of the personal devices that have consented to allow the analysis. Each participant's personal device (e.g., smartphone, laptop) uses its MID software to calculate the likelihood of the data stored on the device. The MID software then sends only the likelihood value (a single number) back to the DLE server. The DLE server then aggregates the log likelihoods from the responding MID devices and either decides it is at a log likelihood maximum or adjusts parameters and redistributes the new parameters to the MID devices.

The similarity between FIML and DLE is striking, but the two algorithms are not identical. There are two major differences: i) traditional FIML requires that all participants reveal their data to the researcher whereas DLE only requires that participants reveal the likelihood of their data given a model and a vector of parameters; ii) traditional FIML performs all of its calculations and data manipulations centrally whereas DLE performs the greatest part of its calculations in parallel on the participants' personal devices.

Several other differences may not be immediately apparent, but are nevertheless important. DLE does not require that data collection be finalized prior to the initiation of model estimation. That is to say, statistical models can begin to be tested as soon as a few participants have consented and begun the experiment. Of course, model parameters will be unreliable with only a few participants. However, as more individuals consent and participate in the experiment, the sample size contributing to each likelihood calculation will grow, and the model parameters will begin to become more and more stable. This means that a researcher can preselect a required statistical power or parameter precision for the analysis and the DLE analysis can automatically terminate or initiate a replication when the experiment reaches that criterion.

Another important difference between traditional FIML and DLE is that FIML always analyzes a static, centrally-stored data matrix while a DLE analysis is performed on an ever-changing subsample of the data. This is because at any particular time, some percentage of the personal devices participating in the experiment may be powered down or be off-network and unreachable, or participants may have added or withdrawn consent in the interval since the last sample was polled. Thus, a DLE analysis is akin to a naturally occurring bootstrap (Efron, 1979) analysis. As long as *holdout* (device is unreachable) is uncorrelated with the model results, bootstrapped standard errors will automatically result from a DLE analysis. A holdout likelihood calculation can be included in the DLE calculation and central aggregation of likelihoods in order to correct for sampling bias induced by holdout likelihood covarying with data likelihood.

Finally, it should be noted that since traditional FIML requires centralized data, longitudinal studies must link variables over time by some sort of identifying information so that data belonging to a single individual can be grouped. On the other hand, in a DLE analysis, all data belonging to a participant and only data that belongs to that participant is available within the MID software. As long as the participant consents to longitudinal linking of her or his data, the likelihood of the longitudinal data can be automatically calculated, even if some data resulted from a previous study. Thus, data sharing between experiments is up to the individual participant. If the participant consents to data sharing, data are automatically available and linked by participant without the researcher needing to become involved in complex data sharing agreements. This is a consequence of the data being owned by the participant, controlled by the participant, and in the possession of the participant. One possible scenario for research participant compensation when using the MIDDLE approach is by micropayments to the participant based on how informative the participants' likelihood was during model estimation. Thus, participants with better quality data would be compensated more than those whose data had low reliability. This would provide an incentive to participants to complete questionnaires and adhere to protocols during data collection.

The MIDDLE Host

Combining the Maintained Individual Data software platform with the Distributed Likelihood Estimation approach (MIDDLE) enables a novel form of experimental design for behavioral, social science, and health research. One possible parallel workflow for MIDDLE experiments is illustrated in Figure 3 where an agency (e.g., National Institutes of Health or National Science Foundation), scientific society (e.g., Association for Psychological Science, American Psychological Association, Society of Multivariate Experimental Psychology) or other organization (e.g., Institute for Social Research, or The Center for Open Science) sponsors a computing platform where MIDDLE experiments can be hosted, advertised, downloaded, and organized. While this *MIDDLE Host* is not necessary for performing an experiment using a MIDDLE approach, this central host approach offers many advantages.

From the participant's perspective, the MIDDLE Host would act as something like an "App Store for Science", where experiments could be browsed and downloaded. The MIDDLE Host would act as an intermediary between research labs and participants, so that participants could go to a single source to find experiments in which they might like to participate. Downloading an experiment and consenting to participation would initiate a communication protocol between the research lab and the participant.

From the research lab's perspective, the MIDDLE Host would fulfill a variety of tasks and services useful to rapid design and execution of behavioral and health science research. First, as a hypothesis is formulated and an experiment is designed, the MIDDLE Host would provide prototype MID software designs that could be modified to collect data that could test the hypothesis. Second, as many research labs begin to use the MIDDLE Host, a provenance trail of data that have been previously collected and models that have been previously tested would help new research cross-validate previous results and extend research into new directions while taking advantage of data available for sharing from previous participants. Third, the MIDDLE Host would act as an advertising agent for recruitment so that new participants could be quickly enrolled. Fourth, the MIDDLE Host would provide
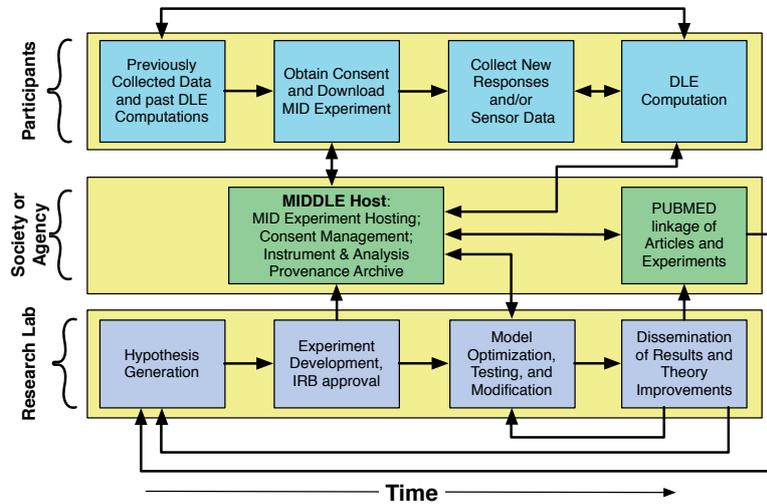
Figure 3. Flowchart of a MIDDLE experiment including a MIDDLE Host for disseminating Maintained Individual Data (MID) experiments and managing requests for Distributed Likelihood Evaluation (DLE). Participants and research labs communicate through the MIDDLE Host, which acts as something of an App Store for Science.

consent management services so that only consenting participants are connected to research labs. Fifth, the MIDDLE Host would manage privacy certificates and network communication protocols in order to help maintain data security in the participants' devices. Finally, the MIDDLE Host would provide provenance management of analyses and linkage to a publication archive (e.g., PUBMED), so that the provenance of data and analyses from articles published in PUBMED would be available to readers, increasing the reliability and reproducibility of behavioral and health science results.

The MIDDLE approach to research will accelerate the pace of discovery in the behavioral, social, and health sciences due to several efficiencies introduced by the workflow shown in Figure 3. When participants consent to data sharing, their previously collected data are automatically linked and immediately available to the new experiment. This parallel approach with automatic within-individual longitudinal data linking and sharing provides the potential for substantially reducing the time between hypothesis generation and dissemination of results while simultaneously reducing participant burden.

Efficiency of scale is available once a MIDDLE Host is up and running and many research labs are participating, but how can the process begin? Suppose one or more large scale longitudinal studies (e.g., the National Longitudinal Study of Youth, the Panel Study of Income Dynamics, or the German Socio-Economic Panel) were to implement MIDDLE data collection. This would jump-start the MIDDLE approach since data sharing and longitudinal linking would be available for new studies and thus convey immediate benefits and competitive advantage to any behavioral research that were to adopt the same MIDDLE Host as the large scale studies. Within-individual linking to large-scale studies has traditionally been impossible due to privacy risks unless the large-scale study incorporates a smaller study into the large-scale protocol. The MIDDLE approach would thus magnify

the impact of these large studies as well as opening up new within-individual data collection and analytic possibilities. Both large studies and small studies are winners in this scenario.

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 optimizer 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 ($\mu_{intercept} = 1.5$, $\mu_{slope} = 0.6$, $\sigma^2_{intercept} = 1.2$, $\sigma^2_{slope} = 0.9$, $\mathrm{Cov}(intercept, slope) = 0.3$) shown as the horizontal lines in Figures 4-a and 4-d.

Next, sampling schemes were simulated to represent potential ways that participants might engage with the MIDDLE optimizer. Figure 4 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.

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 optimizers on the data. At each time step, the first optimizer 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 optimizer to calculate the gradient and Hessian of the likelihood surface and choose a new set of model parameters. Thus the first optimizer 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 optimizer are plotted in Figure 4-a along with the number of individuals contributing likelihoods (Figure 4-b) and minus two log likelihood value divided by number of individuals (Figure 4-c) at each time step.

The second optimizer 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 optimizer asked each MID device to perform a maximum number of likelihood calculations at each time step. The second optimizer performed calculations akin to a traditional bootstrap where each time step represented a new sample drawn from
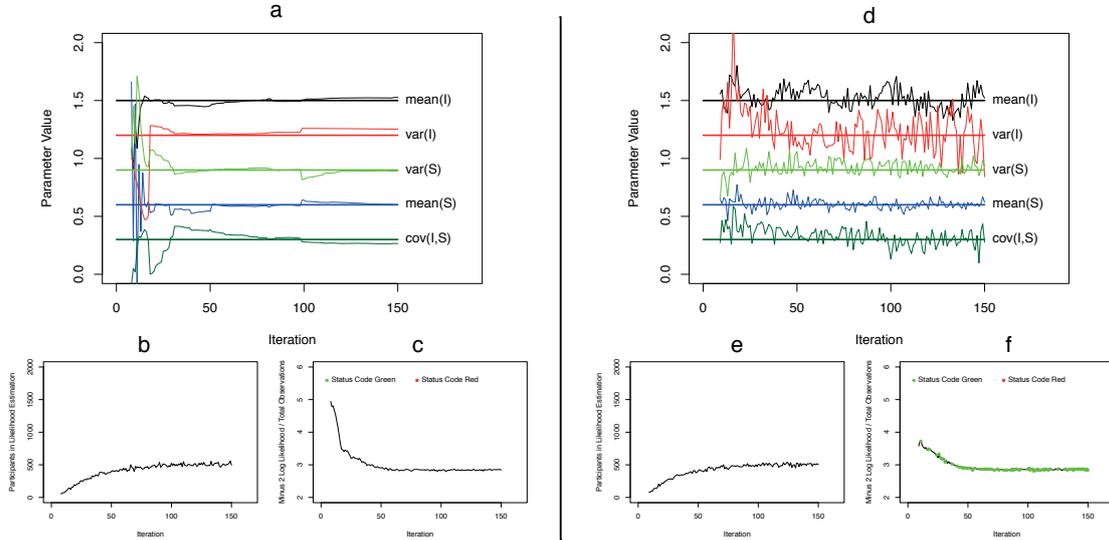
Figure 4. Results of one simulated MIDDLE experiment estimated using two different optimization criteria. For panels (a), (b), and (c), each query of the distributed devices was paired with one major iteration of the optimizer occurred. For panels (d), (e), and (f), each time the distributed devices were queried, the optimizer was allowed to come to complete maximum likelihood convergence. Note: mean(I), mean(S), var(I), var(S), and cov(I,S) refer to the simulated Latent Growth Curve mean, variance, and covariance of the latent intercept and slope.

the population. The results for the second optimizer running a single experiment are shown in Figure 4-d, 4-e, and 4-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 5 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 optimizer 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.

At first we were surprised that the single-step optimizer outperformed the full-convergence optimizer. Upon further consideration we hypothesized that the full-convergence process was overfitting each "bootstrap" sample whereas the single-step optimizer 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.
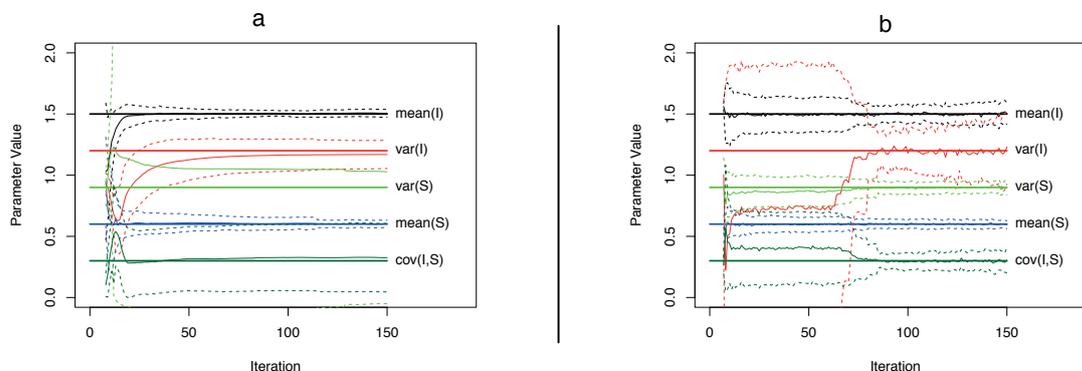
Figure 5. Means and standard deviations of 100 simulated MIDDLE experiments. Each query of the distributed devices was accompanied by (a) one major iteration of the optimizer or (b) the optimizer was allowed to reach maximum likelihood convergence. Note: mean(I), mean(S), var(I), var(S), and cov(I,S) refer to the simulated Latent Growth Curve mean, variance, and covariance of the latent intercept and slope.

Example Use Case of the MIDDLE Approach

The MIDDLE approach reorganizes how experimental and epidemiological research is 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 reorganization as well as illustrating many of the challenges that will need to be addressed.

Epidemiological Study of Diet and Exercise

As in Figure 3, a research group generates a hypothesis about diet and exercise, a self-report questionnaire instrument, a tie-in to accelerometer sensor data from a smart phone, and uses statistical models to test the hypothesis. The research group uses a set of software tools to create an experiment that will run on a MIDDLE-enabled device. The group submits the MIDDLE experiment and IRB-approved consent instrument to the MIDDLE Host. The MIDDLE Host loads the experiment, model likelihood calculator, and consent documents into a web-accessible central repository — its "app store for science". 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 app and model likelihood calculator.

As participants consent 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' personal device 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. IRB approved experimental modifications are uploaded to the NIH MIDDLE Host and are re-disseminated for participant consent and download.

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 thus automatically enabled and linked at the individual level at zero cost to the newly funded project. Participants may choose to collect personal data in their MIDDLE enabled device without previously opting into a study (e.g., using wearable activity monitors, health monitors, GPS tracking, or MIDDLE experiment questionnaires) and then track their own personal trends. If these participants then opt into an experimental analysis, they can choose to allow access of these previously collected data in the new experiment.

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

Followup In-Lab and In-Home Study with Treatment and Control

A second research group investigates running a study on a related hypothesis 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 MIDDLE 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 second research group modifies the first group's instrument and statistical models and advertises their IRB-approved study on the MIDDLE Host, offering additional compensation for participants from the first study. Participants opt-in and most 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 and consenting to in-lab followup are randomly selected through the MIDDLE Host for inclusion in an in-lab section and are assigned by the MIDDLE Host to a treatment or control condition.

For participants who opt-in, the MIDDLE Host transmits contact information to the research group, which arranges appointments for the in-lab study. Participants bring their personal device to the lab and the in-lab data are uploaded into the personal device for the participants to take home. Participants can choose whether or not the lab will be allowed to archive a copy of their data. The analysis and write up proceed in the same manner as in the epidemiological experiment. Note that the in-lab data are always uploaded to the participants' devices. Thus, these data are available for sharing and longitudinal linking

in other experiments. As more data are accumulated into participants' personal devices, their data become more and more valuable to future researchers, and thus of greater market value to the participant.

Benefits of the MIDDLE Approach

Accelerated Pace of Discovery

First, as the MIDDLE approach is adopted, previously collected data will become available to new experiments. Second, data collection in any single experiment can be stopped when minimally sufficient power or statistical precision is achieved. Third, analysis happens simultaneously with data collection. These three conditions mean that the data collection and analysis phases of a funded project will take less time, and therefore a smaller proportion of a given project's budget. A primary rate-limiting factor for discovery is total grant funding available. If each project costs less, more proposals can be funded. And since each project takes less time, the total rate of discovery is improved. The rate of discovery will continue to improve since each year more data will be preexisting on participants' devices.

Reduced Burden and Mitigated Risk for Participants

Since fewer new data are required, either within-individual or in terms of sample size, participant burden per experiment is decreased. 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 stored in a centralized location. Reduced risk of data disclosure is likely to improve the chance that participants will answer sensitive questions such as drug use, sexual history, and/or HIV status. Any participant can opt out of an experiment at any time and their data do not need to be found and deleted in the central database.

Accelerated Translation of Research into Practice

As hospitals and clinicians install MIDDLE-compatible optimizers, a radically new approach to research translation becomes possible (see Figure 6). If a patient opts to allow a clinician secure access to her or his data, the clinician can have automatic online access to up-to-date longitudinal biomarkers and physiological measurements, thereby saving clinicians' time. The clinician could run an outlier detection model, giving real-time alerts to important changes in a patient's medical status, allowing the clinician to recommend a clinic visit early rather than risk an emergency room visit later. As research studies use the MIDDLE system to develop predictive statistical models, hospitals and clinicians connected to the MIDDLE Host can have direct access to these predictive models. These models can be downloaded and used by the health care provider to assist in diagnosis or to assess complex etiological risks in patients using MIDDLE-compatible home health care monitors, thereby reducing time and effort required to translate research into clinical practice.
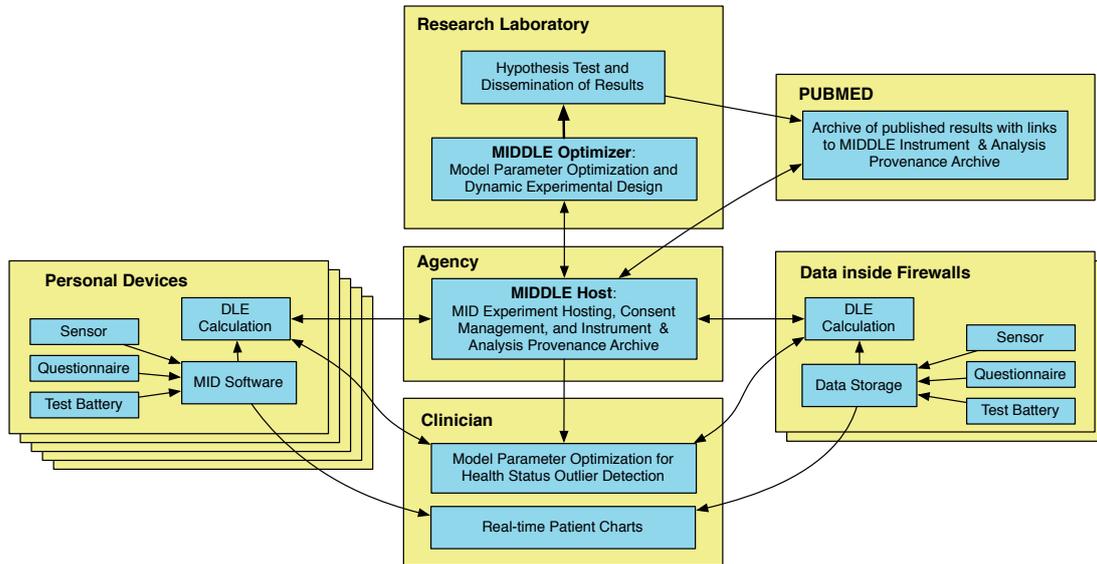
Figure 6. Clinicians and hospitals could use the MIDDLE approach to calculate likelihood of patient health problems from daily observations recorded by home health monitors. These predictive models could be directly accessed from a MIDDLE Host accelerating the translation of research into clinical practice.

Automatic Data Sharing

Data sharing often requires a multi-year wait period. In the MIDDLE approach, data sharing with automatic longitudinal linking can occur for on-going studies. Since data are analyzed as they are being collected, an original study automatically has first access to new data. However, the authors of the original study are under time pressure to publish their results quickly, since other research studies may recruit participants from the same pool as the original study, triggering data sharing. While this is an additional pressure for the original study authors, the effect on science as a whole is positive, since there is an additional incentive to keep the time between data collection and publication short. 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 unduly weighted 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-individual linkage across studies is currently difficult or impossible to implement. 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*: the 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 participant's MIDDLE-compatible device 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 NIH participants but also for any patient with a MIDDLE-compatible home health monitor.

Inter-Site Linking for Data inside Firewalls

Statistical models can be fit when data come from two or more facilities that each require that sensitive data not leave their respective facility. Each facility running a MIDDLE objective function calculator can participate in fitting a statistical model in the same manner that individual personal devices participate. When individual-level data are linked between institutions and participant consent is given, individual-level models can be run using data from personal devices and multiple institutions. For instance, one facility might house individual genome data, while another facility has phenotypic clinical test results, while an individual's device might contain experience-sampling, accelerometer, heart rate, and blood pressure time series. Statistical models could then be fitted that link all of these variables and time series to study gene by environment interaction.

More Reliable Methods, Instruments, and Statistical Tests

Methods, instruments and the provenance of the statistical tests will be archived for articles 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.

Expanded Participant Pool with Better Generalizability Estimates

Linkage to US Census and other very large scale data allows estimation of representativeness of any particular sample. All samples are non-representative to some degree. State-of-the-art multivariate weighting models can then be developed that could be applied to any study run with the MIDDLE system, thereby improving generalizability for any given sample size.

More Consistent Standards for Data Access and Analysis

Implementing the MIDDLE approach will require defining an Applications Programmer Interface (API) standard for use with MIDDLE-compatible devices. It is widely recognized by manufacturers and researchers that personal health monitors with wireless sensors are one of the main growth markets in small devices (Almedar & Ersoy, 2010; Mottola & Picco, 2011). By defining an open source API standard early, an agency e.g., NIH or society e.g., APS could have an influence on the intercompatibility of these devices. There are many reasons why device manufacturers may find it profitable to advertise their products

as "NIH Compatible" or "APS Compatible", encouraging a coalescence in data access and analysis standards.

Concerns, Limitations, and Opportunities

A number of problems must be solved in order to implement a MIDDLE system. While some of these are complicated, there is no reason that a MIDDLE system could not be implemented using currently available technologies. We next present a number of issues that should be kept in mind if one were to develop a successful MIDDLE system. While the following list is not exhaustive, it provides highlights of areas where more research is needed. The MIDDLE paradigm provides a very different way of thinking about statistical methods and as such poses a wide range of questions that we believe will become an active new area of methodological inquiry.

Security

Security and privacy of the network must be excellent. It should be noted that encryption is not the same as privacy. Of course, data on the personal device and transmissions among personal devices and MIDDLE servers will need to be encrypted. But, no encryption is totally secure, it is merely expensive and difficult to break. The MIDDLE approach focuses on privacy, which 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. This is due to the fact that MIDDLE communications are information-impoverished; only models are sent upstream and only likelihood values are returned. While the data on a given personal device 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 personal device is much lower than the reward for a successful attack on a current centralized repository. Any given personal device is therefore a much less attractive target for identity thieves. While no system will ever be entirely secure or private, risks to privacy will be substantially mitigated by implementing the MIDDLE system.

Estimation

Estimation of statistical models from a dynamically changing sample requires new algorithms and convergence criteria. Since the data are in flux during DLE optimization, standard calculations for parameter stability and power estimation will need revision. These problems bear resemblance to those posed by resampling or permutation testing. However, the MIDDLE approach brings novel information into the estimation problem, whereas resampling and permutation testing use explicit randomization applied to existing information. Bayesian methods may provide insight into how to address this problem of expanding information resampling. For instance, it may be possible to capitalize on the information generated when participants opt in and opt out to provide a better estimate of generalizability. A naive solution to the estimation problem would be to collect data and optimize until 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 very likely that other solutions exist that outperform the naive solution, thus reducing the required sample size for a chosen power.

Power

Power estimated from a fixed data set can be misleading and can lead to surprising uncertainty in estimating proportions of studies that fail to replicate (Maxwell, Lau, & Howard, in press). Current practice assumes that the data are known and fixed, but the MIDDLE paradigm requires one to consider statistical estimation in the context of uncertainty in the data. While this poses problems for estimation as discussed in the previous paragraph, it also provides an opportunity to rethink what is meant by "replication" and/or "cross-validation". Every time a MIDDLE optimizer requests the likelihood of a new set of parameter point estimates it is querying a data set that may have changed in some unknown way. This forces us to confront the possibility that using a fixed data set typically leads to overfitting: current practice may lead us to be more confident in our results than we should be. Since the MIDDLE paradigm combines the data gathering and estimation processes, there may be a way to improve confidence estimates if we consider a something like "confidence in incremental replication" as part of the estimation stopping rule. At the very least, the relationship between power, confidence, and replication in the context of MIDDLE estimation require further study.

Inference

One of the primary goals of psychological and medical research is to improve the ability to predict the effects of some intervention. In current practice, this is taken to mean that random assignment of treatment and control are used in a controlled setting to provide an estimation of confidence in a statement of causal inference. Random assignment of treatment and laboratory controlled experimental settings in a MIDDLE experiment need not differ from a typical protocol. However, there is a weakness in the logic of the randomized control paradigm: if only a proportion of the population is susceptible to the treatment, the estimated effect size generalized to the population at large is attenuated. The MIDDLE paradigm provides the possibility of experiments that are dynamically modified on a person-specific basis based on changing parameter estimates. It may be that mixture distribution models could be fit while simultaneously altering individual treatment conditions according to Bayesian posteriors. This could lead to inference at both population-level and person-specific effects of the treatment. Possible applications could include adaptive training programs, cognitive behavior therapies, or pharmaceutical dosages.

Group Membership

Some analyses 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 membership in a group. 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 personal device. Once the group membership data are stored in the personal device, it is straightforward to calculate objective functions conditional on group membership. However, optimization will require aggregating objective function values conditional on group membership. In order for group membership to not be revealed to the central optimizer, one solution would allow peer-to-peer aggregation of objective functions prior to being transmitted to the MIDDLE optimizer. Again, this is an issue that requires substantial study.

Informed Consent and Institutional Review Boards

A modification of current best practices in informed consent will need to be developed. 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, longitudinal linking and sharing across experiments can be separate consent items. Risk management options can be included in consent documents. One participant might be willing to donate his/her data to a research laboratory and their primary care physician while another participant may wish to not reveal any data. Some individuals may be willing to allow plots of selected raw data to be generated locally and transmitted to the research study. Others may wish to only reveal function 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. Others might be willing to reveal historical data but not be willing to participate in new data collection paradigms.

The changes in informed consent described above will need to be codified and incorporated into Institution Review Board (IRB) training. This process will require extensive discussion in the research ethics community in order to provide clear guidance to IRBs. One possible benefit of this discussion is that since consent can be isolated to instances of analysis (rather than solely at the time of data gathering), the MIDDLE paradigm may provide a way to resolve the current impasse between NIH guidelines requiring data sharing and IRB guidelines on data privacy that can preclude data sharing.

Backup and Archiving

Data privacy requires that data not be disclosed. But also, data must not be lost. Mechanisms for secure data backup must be available for individual personal devices. Individuals must be given a choice of backup mechanisms. One reasonable choice would be encrypted backup onto a cloud facility (Bhadauria, Chaki, Chaki, & Sanyal, 2013; Jansen & Grance, 2011). Some participants may wish to only maintain a private MID backup in their homes, although this is vulnerable to permanent loss from fire or theft. Hospitals and/or primary care physicians may choose to offer access to encrypted cloud-based MID backup as part of their health care services. While it is clear that backup to cloud storage does expose the individual's data to a risk of disclosure, in the MIDDLE paradigm, the choice of how much risk is acceptable remains with the individual whereas in the current paradigm that choice is made by the research lab.

Since data are maintained by individual participants, what happens in cases of participant mortality? One solution to this problem is inherent in the fact that an individual's data are private property. Thus, a participant's data could be part of an estate. Data could

be willed to science, in which case an archive repository for data of deceased participants would need to be maintained.

Conclusions

This article only covers a few of the major discussion points that have been raised as we consider the implications of the paradigm shift in data collection and analysis implied by the MIDDLE approach. Suffice it to say that while there are complex problems remaining to be solved, the benefits of the MIDDLE approach are so great that we foresee a system akin to MIDDLE being an inevitable component of the future of behavioral, social, and health science research.

The basic premise of the MIDDLE approach is that data remain in participants' possession and control and thus remain the personal property of each participant. This will transform the economic model of large-scale research both public and private. We believe that this philosophic shift is as revolutionary as when ownership of private property becomes allowed in a formerly command-driven economic system. We predict that a market-driven personal data economy will arise as individuals realize that personal data are personal property, must be kept in their possession and control, and have accumulating worth directly related to the data's quality and scarcity. Unforeseen innovations will surely arise from this new market-driven personal data economy. We are confident that as the MIDDLE research paradigm becomes widespread, the pace of innovation and discovery in the behavioral, social and health sciences will be vastly accelerated while risks to individual privacy will be considerably mitigated relative to current research practice.

The authors intend to actively pursue inquiry into problems that need to be solved prior to implementing the MIDDLE paradigm. However, the number of open questions raised in the current article are larger than we can reasonably address. We firmly believe that something akin to what is described here will be part of the future of psychological and medical science. If the reader has considered this new research paradigm and in the process formed new questions or solutions, then the authors' intentions will have been fullfilled.

References

Almedar, H., & Ersoy, C. (2010). Wireless sensor networks for healthcare: A survey. *Computer Networks*, *54*, 2688–2710. (doi: 10.1016/j.comnet.2010.05.003)

Anderson, B., Fagan, P., Woodnutt, T., & Chamorro-Premuzic, T. (2012). Facebook psychology. *Psychology of Popular Media Culture*, *1*(1), 23–37. (doi: http://dx.doi.org/10.1037/a0026452)

Benocci, M., Tacconi, C., Farella, E., Benini, L., Chiari, L., & Vanzago, L. (2010). Accelerometer–based fall detection using optimized ZigBee data streaming. *Microelectronics Journal*, *41*(11), 703–710. (doi:10.1016/j.mejo.2010.06.014)

Bhadauria, R., Chaki, R., Chaki, N., & Sanyal, S. (2013). A survey on security issues in cloud computing. *IEEE Communications Surveys and Tutorials*, *arXiv:1109.5388v1*. (doi: http://arxiv.org/abs/1109.5388)

Boker, S. M., & McArdle, J. J. (1998). A psychotelemetry experiment in fluid intelligence. In J. J. McArdle (Ed.), *Human cognitive abilities* (pp. 215–229). Hillsdale, NJ: Lawrence Erlbaum Associates.

comScore. (2014). *comScore Reports July 2014 U.S. smartphone subscriber market share.* `http://www.comscore.com/Insights/Market-Rankings/comScore-Reports-July-2014 -US-Smartphone-Subscriber-Market-Share`. (Accessed 2015-02-25)

Csikszentmihalyi, M., & Larson, R. (1987). Validity and reliability of the experience-sampling method. *The Journal of nervous and mental disease*, *175*(9), 526.

Dolin, R. H., Alschuler, L., Boyer, S., Beebe, C., Behlen, F. M., Biron, P. V., et al. (2006). HL7 clinical document architecture, release 2. *Journal of the American Medical Informatics Association*, *13*(1), 30–39. (doi: http://dx.doi.org/10.1197/jamia.M1888)

Dufau, S., Duñabeitia, J. A., Moret-Tatay, C., McGonigal, A., Peeters, D., Alario, F.-X., et al. (2011). Smart phone, smart science: How the use of smartphones can revolutionize research in cognitive science. *PloS one*, *6*(9), e24974. (doi: 10.1371/journal.pone.0024974)

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, *7*, 1–26.

Fox, S., & Duggan, M. (2013). *Tracking for health* (Tech. Rep.). Pew Research Center. `http://www.pewinternet.org/files/old-media//Files/Reports/2013/ PIP_TrackingforHealth%20with%20appendix.pdf`. (Accessed 2015-02-25)

Goel, V. (2014). Facebook tinkers with users' emotions in news feed experiment, stirring outcry. *The New York Times*. `http://www.nytimes.com/2014/06/30/technology/facebook-tinkers-with-users- emotions-in-news-feed-experiment-stirring-outcry.html?_r=0`. (Accessed 2015-02-25)

Greenwald, A. G., & Nosek, B. A. (2001). Health of the implicit association test at age 3. *Experimental Psychology*, *48*(2), 85–93.

Hektner, J. M., Schmidt, J. A., & Csikszentmihalyi, M. (2006). *Experience sampling method: Measuring the quality of everyday life.* SAGE Publications, Incorporated.

IndivoHealth. (2012). *Indivo, the personally controlled health record.* (http://indivohealth.org. Accessed January 2, 2012)

Jansen, W., & Grance, T. (2011). *Guidelines on security and privacy in public cloud computing* (Tech. Rep.). Gaithersburg, MD: National Institute of Standards and Technology.

Larson, R., & Csikszentmihalyi, M. (1983). The experience sampling method. In H. T. Reis (Ed.), *Naturalistic approaches to studying social interaction: New directions for methodology of social and behavioral sciences* (pp. 41–56). San Francisco: Jossey-Bass Publishers, Inc.

Mandl, K. D., & Kohane, I. S. (2009). No small change for the health information economy. *New England Journal of Medicine*, *360*(13), 1278–1281. (doi: 10.1056/NEJMp0900411)

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (in press). Is psychology suffering from a replication crisis? what does "failure to replicate" really mean? *American Psychologist*, *??*(??), ??–??

Microsoft. (2012). *Microsoft healthvault.* (http://www.microsoft.com/en-us/healthvault/. Accessed January 2, 2012)

Miller, G. (2012). The smartphone psychology manifesto. *Perspectives on Psychological Science*, *7*(3), 221–237. (doi: 10.1177/1745691612441215)

Mottola, L., & Picco, G. P. (2011). Programming wireless sensor networks: Fundamen-

tal concepts and state of the art. *ACM Computing Surveys*, *43*(3), 19:1–51. (doi: 10.1145/1922649.1922656)

Personal data for the public good (Tech. Rep.). (2014). California Institute for Telecommunications and Information Technology. `http://hdexplore.calit2.net/wp/project/personal-data-for-the-public-good-report/`. (Accessed 2015-02-25)

Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological momentary assessment. *Annual Review of Clinical Psychology*, *4*, 1–32. (doi: 10.1146/annurev.clinpsy.3.022806.091415)

Stone, A. A., & Shiffman, S. (1994). Ecological momentary assessment (EMA) in behavorial medicine. *Annals of Behavioral Medicine*, *16*, 199–202.

United States Census Bureau. (2014). *U.S. and world population clock.* `http://www.census.gov/popclock/`. (Accessed 2015-02-25)

Vreeman, D. J., McDonald, C. J., & Huff, S. M. (2010). Loinc(r): a universal catalogue of individual clinical observations and uniform representation of enumerated collections. *International Journal of Functional Informatics and Personalised Medicine*, *3*(4), 273–291. (doi: 10.1504/IJFIPM.2010.040211)

Walls, T. A., & Schafer, J. L. (Eds.). (2005). *Models for intensive longitudinal data.* Oxford: Oxford University Press.

Weitzman, E. R., Adida, B., Kelemen, S., & Mandl, K. D. (2011). Sharing data for public health research by members of an international online diabetes social network. *PLoS One*, *6*(4), 1–8. (doi:10.1371/journal.pone.0019256)