Pervasive Health

Editors: Anind K. Dey ■ Carnegie Mellon University ■ anind@cs.cmu.edu
Jesus Favela ■ CICESE ■ favela@cicese.mx
Stephen Intille ■ Northeastern University ■ s.intille@neu.edu

Center of Excellence for Mobile Sensor Data-to-Knowledge (MD2K)

Santosh Kumar, Gregory Abowd, William T. Abraham, Mustafa al'Absi, Duen Horng (Polo) Chau, Emre Ertin, Deborah Estrin, Deepak Ganesan, Timothy Hnat, Syed Monowar Hossain, Zachary Ives, Jacqueline Kerr, Benjamin M. Marlin, Susan Murphy, James M. Rehg, Inbal Nahum-Shani, Vivek Shetty, Ida Sim, Bonnie Spring, Mani Srivastava, and Dave Wetter

The National Center of Excellence for Mobile Sensor Data-to-Knowledge (MD2K) was established in October 2014 with a grant from the National Institutes of Health under the Big Data-to-Knowledge (BD2K) program. Among the 11 centers of excellence originally funded, MD2K's unique contribution is to develop innovative tools to make it easier to gather, analyze, interpret, and capitalize on high-frequency data from mobile sensors.[1] It seeks to facilitate the monitoring of health states and quantify the temporal dynamics of key physical, biological, behavioral, psychological, social, and environmental factors that contribute to the health and disease risks of individuals. MD2K's overarching goal is to reduce the burden that complex chronic disorders place on health and healthcare by making it feasible to detect and predict person-specific disease risk factors ahead of the onset of adverse clinical events, supporting sensor-driven just-in-time interventions.

To demonstrate the utility and wide generalizability of the research and tools developed by MD2K, we initially targeted two biomedical applications—improving the success rate for smoking cessation and reducing the number of rehospitalizations in congestive heart failure (CHF). Our two chosen biomedical applications are at the opposite ends of the temporal spectrum of mortality. Smoking is the leading cause of mortality, causing 1 in 5 deaths,[2] but its mortality risk is far in the future. On the other hand, CHF, which is the leading cause of preventable rehospitalization with a readmission rate of 27 percent,[3] has an immediate mortality risk. The first (of three) iterations of these two studies with 75 participants each is currently underway. In addition, the biomedical applications addressed by MD2K have expanded to managing stress, reducing overeating, reducing cocaine use, and improving oral health.

ENABLING DIGITAL BIOBANKS FOR MHEALTH

Biomedical research studies archive biospecimens in biobanks so that such specimens can later be reprocessed to capitalize on future technological improvements, thereby supporting biomedical discoveries not possible at the time of data collection. Conversely, mobile health (mHealth) studies usually encode and retain derivative digital biomarkers (such as activity counts) that are specific to the computational models used by respective vendors at the time of data collection, because the raw sensor data processed to generate these markers is considered too voluminous to retain. This approach prevents any independent validation of these biomarkers and makes it impossible to later apply newer computational models that might more accurately measure the biomarkers.

To attain long-lasting research utility, similar to that of the biobanks, raw sensor data must be collected in a way that allows it to be reprocessed in the future to validate prior digital biomarkers, and to derive new biomarkers (see Figure 1). In addition, data science and computational research for the development and validation of new biomarkers requires the collection of raw sensor data and associated labels from observations that identify the behaviors of interest.

MD2K's work aims to enable the creation of digital biobanks in mHealth studies by developing the required computational infrastructure and disseminating that infrastructure to the broader scientific community through open source software and tools.

OVERVIEW OF MD2K'S RESEARCH

MD2K brings together investigators, postdocs, and students in computing, engineering, medicine, behavioral science,

Published by the IEEE CS ■ 1536-1268/17/$33.00 © 2017 IEEE

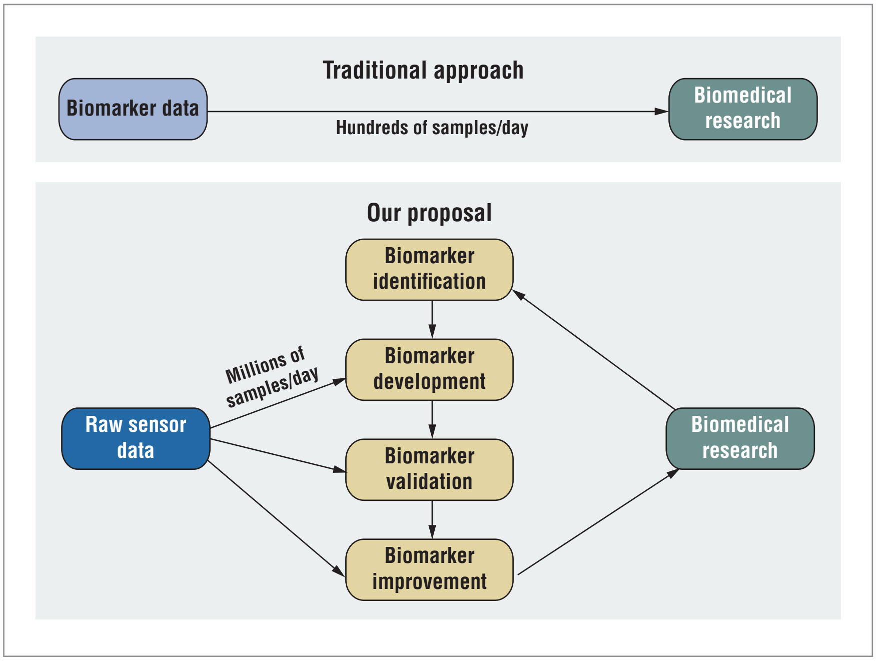

Figure 1. Benefits of collecting raw sensor data (instead of biomarkers only) in mobile health (mHealth) studies.

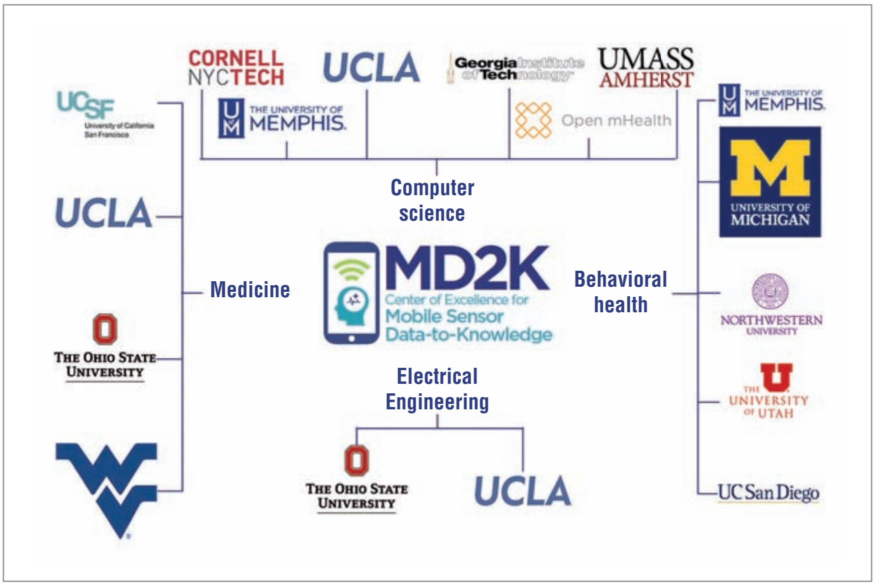
CORNELL NYCTECH
UCLA
Georgia Institute of Technology
UMASS AMHERST
UCSF University of California San Francisco
THE UNIVERSITY OF MEMPHIS
Open mHealth
THE UNIVERSITY OF MEMPHIS
Computer science
UCLA
UNIVERSITY OF MICHIGAN
Medicine
MD2K
Center of Excellence for Mobile Sensor Data-to-Knowledge
Behavioral health
NORTHWESTERN UNIVERSITY
THE OHIO STATE UNIVERSITY
THE UNIVERSITY OF UTAH
Electrical Engineering
THE OHIO STATE UNIVERSITY
UCLA
UC San Diego

Figure 2. Organizations and disciplines comprising the MD2K Center of Excellence.

and statistics from 12 universities, as shown in Figure 2. It takes a comprehensive approach to advancing the science of mHealth and generating resources to enhance the research community's capacity to conduct data science and biomedical research with mobile sensor data.

MD2K is developing general principles and computational methods to infer markers of patient health and behavioral, physical, social, and environmental risk factors. It's also developing data-analytic tools to mine biomarker time series to discover indicators and predictors of vulnerable states. Furthermore, it's developing an extensible computational architecture for mobile sensor big data computing platforms—encompassing sensors, smartphones, and the cloud—to collect, analyze, manage, and store mobile sensor data. Finally, MD2K is conducting field studies on newly abstinent smokers and CHF patients to inform the design of MD2K technology and evaluate its utility in biomedical knowledge discovery.

RESOURCES FOR THE COMMUNITY

The work of MD2K has resulted in four types of resources for both the data science and biomedical research communities: sensors, digital biomarkers, software, and training.

Sensors

Several sensors have been developed and deployed by MD2K in mHealth field studies:

- *EasySense*—a contactless microradar sensor that can detect heart and lung motion and assess change in the lung fluid level;
- *MotionSenseHRV*—a wrist-worn sensor that can measure hand gestures via accelerometers and gyroscopes and interbeat intervals via optical sensors for computing heart rate variability indices; and
- *AutoSense*—a chest-worn sensor suite that can measure cardiorespiratory parameters via ECG and respiration, and movement of the torso via accelerometers.

Each of these sensors collects high-quality raw sensor data in the field to facilitate the development and validation of biomarkers. In addition, an energy-efficient design ensures that sensors last the entire day on a single battery charge while streaming raw sensor data in real time to a smartphone.

Digital Biomarkers

MD2K has developed several digital biomarkers. The cStress model measures stress likelihood for each minute, not confounded by physical activity, from interbeat intervals (of heart) and respiration. Time series pattern mining methods have also been developed that identify major daily stress episodes from the cStress time series.

The puffMarker model identifies smoking episodes from hand gestures and breathing pattern, while the mCrave model can estimate craving for cigarettes in newly abstinent smokers using physiological sensor data. A computational model has also been developed to detect cocaine use from interbeat intervals. Finally, models have also been developed to detect eating from hand gestures.

Several other biomarkers are currently under development. One uses EasySense to detect lung fluid congestion, and two others exploit eyeglasses—one uses smart eyeglasses that capture blink patterns to detect fatigue, and another uses camera-equipped eyeglasses to detect TV viewing. Oral health biomarkers use hand gestures to detect brushing and flossing.

Software

Open source software for mobile phones and the cloud have been developed with accompanying user manuals so they can be used in all MD2K user studies and can be made available widely for the research community to collect high-frequency raw sensor data in their field studies.

mCerebrum is a configurable software platform that supports reliable high-frequency data collection from mobile and wearable sensors and supports real-time processing of this data for sensor-triggered, just-in-time adaptive interventions. It provides the ability to collect raw sensor data from Easy-Sense, AutoSense, MotionSense, Microsoft Band, Oral-B smart toothbrushes, and Omron weighing scales and blood pressure monitors. In the future, it will likely also support Android Wear smartwatches.

In addition to collecting streaming raw sensor data concurrently from multiple wearable sensors, mCerebrum also includes the ability to assess the quality of incoming data in real time and enable data quality correction via sensor reattachment or rearrangement. mCerebrum also includes apps to compute, in real-time, stress, smoking, driving, and activity biomarkers, each of which can be used to trigger notifications or launch health interventions. For example, in the smoking cessation field study at Northwestern University, stress episodes detected from cStress are being used to trigger a just-in-time stress reduction intervention.

Finally, mCerebrum includes native support for privacy control as well as prompting for momentary self-reports, called Ecological Momentary Assessment (EMA). EMAs can be triggered at random times (for baseline data collection), based on time of day, or in response to self-reported events and sensor-detected events. The timing of EMAs across all trigger sources are carefully coordinated to ensure limiting of user burden. mCerebrum's modular design currently includes more than 20 apps to allow the mixing and matching of desired functionality for specific studies.

Cerebral Cortex is the cloud companion of mCerebrum. It includes support for receiving and remotely monitoring data collection and for study-wide data analysis, visualization, model development, and intervention design for mobile sensor data. It's designed to support thousands of concurrent instances of mCerebrum and provides the ability to develop machine-learning models on study-wide datasets, and its interoperable interfaces allow for the aggregation of other data sources, including non-sensor data.

The MD2K software suite currently supports seven different field studies at seven sites across the US for improving smoking cessation, reducing overeating, reducing illicit drug use, reducing stress, reducing readmission in CHF patients, and improving oral health. These studies are expected to collectively generate over half a million hours of raw sensor data (150 terabytes) from more than 1,000 participants.

All MD2K software is freely available from GitHub (https://github.com/MD2Korg) and can be configured to support high-rate data collection for specific studies via simple changes in configurations. Open source licensing also allows for others to incorporate new apps for supporting other sensors or new computational models for existing or new biomarkers.

MD2K welcomes the research community to both use and contribute to the software platforms, to expand the data sources, improve the computational models for computing existing biomarkers, and introduce models for computing new mHealth biomarkers.

Training

To increase the capacity of the data science and biomedical research communities to collect, analyze, interpret, and capitalize on high-frequency mobile sensor data in real-world environments, MD2K provides several training resources.

First, it co-sponsors the annual Office of Behavioral and Social Sciences Research/National Institute on Drug Abuse funded mHealth Training Institutes (mHTI) held at UCLA each summer. The immersive week-long program is designed to train a diverse group of participants from academia in core mHealth perspectives and methodologies in a boot camp setting. The selected participants are grouped into teams that each work on an mHealth problem under the guidance of mHealth and big data thought leaders. To date, over 70 mHealth scholars from more than 45 institutions have been trained through the mHTI. Applications for mHTI are usually due by the end of January (see https://mhealth.md2k.org/mhealth-training-institute).

Second, MD2K has created a virtual collaboratory called mHealthHUB (https://mhealth.md2k.org), a dedicated website that serves as an organizing hub and online repository of mHealth tools, technologies, and educational materials, as well as forums for the rapidly growing community of mHealth researchers across the globe to connect and collaborate. Recorded videos of all the mHTI lectures and regular MD2K webinars are curated on this site for broad dissemination. MD2K welcomes the community to contribute news articles, research, and projects and other training resources to this website for broader dissemination.

Third, MD2K develops and releases training manuals and videos to accompany MD2K's open source software (mCerebrum and Cerebral Cortex). These materials help users understand how to use the MD2K software for collecting, curating, analyzing, and interpreting high-frequency mobile sensor data.

Introductory Book on mHealth. A new book, *Mobile Health: Sensors, Analytic Methods, and Applications*, edited by James M. Rehg, Susan Murphy, and Santosh Kumar of MD2K and soon to be published by Springer, provides an introduction to the field of mHealth from a sensor perspective. The book, currently in production, will provide an in-depth look at the three key elements of mHealth technology:

- the development of sensors that identify key health-related behaviors,
- the use of analytic methods to predict current and future states of health and disease, and
- the development of mobile interventions to improve health outcomes.

The book is meant to serve as a readily accessible and comprehensive introduction to current and future research directions in mHealth, especially for those entering this field.

Institutional Review Board Language. During the course of designing the institutional review board (IRB) applications and consent forms for MD2K studies, IRB language has emerged for sharing high-frequency mobile sensor data among researchers spread across a dozen institutions. The raw sensor data might reveal privacy-sensitive information (such as location, stress level, smoking behavior, and cocaine use), so appropriate care must be taken to protect participant privacy while allowing responsible scientific investigations. Examples of such appropriate IRB language used in MD2K field studies can be accessed from mHealthHUB.

FUTURE OUTLOOK

In the near future, the work of MD2K is expected to generate a variety of resources for the mHealth community. First, the entire software and hardware ecosystem for collecting raw sensor data to develop and validate biomarkers (including energy-efficient sensors and servers for setting up a locally situated private MD2K cloud) is expected to be available for the community in the fall of 2017. Second, MD2K field studies will generate unique data that, with appropriate IRB approvals, can be made available for third-party research by requesting investigators. Third, computational models for several new biomarkers are expected to be published and implemented in MD2K software repositories. Fourth, both science and associated software support will become available for designing, implementing, evaluating, and refining sensor-triggered just-in-time interventions. Methodological advances in both computational modeling as well as biomedical research with sensor-derived biomarkers will continue to be published in research articles, all of which can be accessed from the MD2K website (https://md2k.org).

In addition, a new project recently funded under the Data Infrastructure Building Blocks program of the National Science Foundation is helping introduce provenance capabilities to the MD2K cyberinfrastructure to facilitate the sharing of high-frequency raw sensor data for third-party research. The mProv project complements MD2K by developing data models, metadata standards, APIs, and runtime support for annotating sensor data streams with information about

- the source (sensor type, placement, and sampling rate—continuous or episodic),
- semantics (such as number, probability, and class/category);
- provenance (the features and rules applied to obtain a biomarker),
- validation (specificity, sensitivity, benchmark, or gold standard used for validation), and
- privacy (the user controls exercised and applicable privacy policies).

A key aspect of this will be semi-automated generation of provenance metadata from computational pipelines. We expect mProv to enable replay, interpretability, comparative analysis, and reproducibility.

Mobile sensor data presents tremendous potential for advancing and improving human life. Collecting and sharing raw sensor data is, however, essential to bring about scientific rigor through reproducibility and comparative analysis. Data sharing can also significantly expand the scientific field by enabling scientists to perform analyses without acquiring the data themselves. The MD2K Center of Excellence is contributing to this rapidly growing field by providing an ecosystem of tools for collecting, analyzing, and sharing raw sensor data obtained from the natural field environment of individuals.

With participation of and contribution from the community, mobile sensor research can soon become a mainstream scientific discipline with wide-ranging applications in health, transportation, energy, home, and workplace. P

ACKNOWLEDGEMENTS

We thank Barbara Burch Kuhn from the University of Memphis for her editorial help. The authors acknowledge support by the National Institutes of Health (NIH) under grant U54EB020404 (by NIBIB) through funds provided by the trans-NIH Big Data-to-Knowledge initiative and support by the National Science Foundation under grant ACI-1640813.

Lab data collected by Mustafa al'Absi at the University of Minnesota was used to develop and validate the cStress model. Lab data collected by Annie Umbricht at Johns Hopkins University and field data collected by Kenzie Preston at NIDA Intramural Research Program (IRP) was used to develop and validate the cocaine detection model. The lab study conducted by Santosh Kumar at the University of Memphis was used to develop the puffMarker model. The field studies conducted at Memphis, Minnesota, and NIDA IRP were used to validate the cStress model, and the field study at Minnesota was used additionally to validate the puffMarker and mCrave models. Finally, the lab and field studies conducted by Gregory Abowd at the Georgia Institute of Technology and by Deepak Ganesan at the University of Massachusetts, Amherst were used to develop and validate eating detection models.

REFERENCES

1. S. Kumar et al., "Center of Excellence for Mobile Sensor Data-to-Knowledge (MD2K)," *J. Am. Medical Informatics Assoc.*, vol. 22, no. 6, 2015, pp. 1137–1142; doi:10.1093/jamia/ocv056.
2. *The Health Consequences of Smoking—50 Years of Progress: A Report of the Surgeon General*, US Dept. Health and Human Services, 2014.
3. R.M. Coffey et al., "Congestive Heart Failure: Who Is Likely to Be Readmitted?" *Medical Care Research and Rev.*, vol. 69, no. 5, 2012, pp. 602–616; doi: 10.1177/1077558712448467.

Santosh Kumar is a professor and Moss Chair of Excellence in computer science at the University of Memphis and director of the MD2K Center of Excellence. He is the corresponding author for this article. Contact him at skumar4@memphis.edu.

Gregory Abowd is a Regents' Professor and J.Z. Liang Chair in the School of Interactive Computing at the Georgia Institute of Technology.

William T. Abraham is Chair of Excellence in Cardiovascular Medicine and Chief of Cardiovascular Medicine at the Ohio State University College of Medicine. He leads MD2K's congestive heart failure studies.

Mustafa al'Absi is a Max & Mary La Due Pickworth Chair and professor of behavioral medicine and the founding director of Duluth Medical Research Institute at the University of Minnesota Medical School.

Duen Horng (Polo) Chau is an assistant professor at the Georgia Institute of Technology's School of Computational Science and Engineering.

Emre Ertin is a research associate professor with the Department of Electrical and Computer Engineering at the Ohio State University. He is the sensors lead for MD2K.

Deborah Estrin is a professor of computer science at Cornell Tech in New York City, professor of public health at Weill Cornell Medical College, and co-founder of Open mHealth.

Deepak Ganesan is a professor in the Department of Computer Science at the University of Massachusetts, Amherst.

Timothy Hnat is the chief software architect for MD2K at the University of Memphis.

Syed Monowar Hossain is the lead software engineer for MD2K at the University of Memphis.

Zachary Ives is a professor of computer and information science and associate dean of the School of Engineering and Applied Science, University of Pennsylvania.

Jacqueline Kerr is an associate professor of family and preventive medicine at the University of California, San Diego.

Benjamin M. Marlin is an assistant professor at the University of Massachusetts, Amherst.

Susan Murphy is the H.E. Robbins Distinguished University Professor of Statistics, professor of psychiatry, and research professor at the Institute of Social Research at the University of Michigan.

James M. Rehg is a professor in the School of Interactive Computing at the Georgia Institute of Technology. He is deputy director of MD2K and leads MD2K's data science research.

Inbal Nahum-Shani is an assistant professor at the Institute for Social Research at the University of Michigan.

Vivek Shetty is a professor of oral and maxillofacial surgery and an assistant vice-chancellor for research at the University of California, Los Angeles. He heads the MD2K Training Core.

Ida Sim is a professor of medicine and co-director of biomedical informatics at the University of California, San Francisco. She leads MD2K's consortium activities.

Bonnie Spring is a professor of preventive medicine, psychology, psychiatry, and public health at Northwestern University. She leads MD2K's smoking cessation studies.

Mani Srivastava is a professor of electrical engineering and computer science at the University of California, Los Angeles.

Dave Wetter is the Jon M. and Karen Huntsman Presidential Professor in Population Health Sciences at the Huntsman Cancer Institute at the University of Utah.