Collecting **High-frequency** Mobile Sensor Data for **Long-lasting Research Utility**

**Santosh Kumar**  
*Director, MD2K Center of Excellence*  
*Professor & Lillian and Morrie Moss Chair of Excellence*  
*Department of Computer Science, University of Memphis*
Advancing biomedical discovery and improving health through mobile sensor big data

Cornell Tech ♦ Georgia Tech ♦ U. Memphis ♦ Northwestern ♦ Ohio State ♦ Open mHealth
Rice ♦ UCLA ♦ UC San Diego ♦ UC San Francisco ♦ UMass Amherst ♦ U. Michigan ♦ WVU
MD2K Multidisciplinary Team – 20 investigators

**Data Science Research**
- Santosh Kumar, *Memphis* (PI)
- Gregory Abowd, Polo Chau, and Jim Rehg, *Georgia Tech*
- Emre Ertin, *Ohio State*
- Deborah Estrin, *Cornell Tech*
- Tyson Condie, Mani Srivastava, *UCLA*
- Deepak Ganesan, Ben Marlin, *UMass*
- Susan Murphy, *Harvard*

**Health Research**
- William Abraham, *Ohio State*
- Inbal Nahum-Shani, *Michigan*
- Bonnie Spring, *Northwestern*
- Cho Lam, Dave Wetter, *Utah*
- Vivek Shetty, *UCLA*
- Ida Sim, *UC San Francisco*
- Jaqueline Kerr, *UC San Diego*
- Clay Marsh, *West Virginia*

Memphis-based headquarter hosts a team of 10 grad students, a postdoc, 3 software engineers, and 6 staff members.
Measuring Exposures, Behaviors, and Outcomes

Mobile Sensors
- Smartwatch
- Chestbands
- Smart Eyeglasses

Exposures

Behaviors

Outcomes

Advancing biomedical discovery and improving health through mobile sensor big data

Cornell Tech ♦ Georgia Tech ♦ U. Memphis ♦ Northwestern ♦ Ohio State ♦ Open mHealth Rice ♦ UCLA ♦ UC San Diego ♦ UC San Francisco ♦ UMass Amherst ♦ U. Michigan ♦ WVU
MD2K Applications – Smoking Cessation & CHF

Detect → Predict → Adapt

- MD2K Applications – Smoking Cessation & CHF

![Detect](image1)
![Predict](image2)
![Adapt](image3)

- Advancing biomedical discovery and improving health through mobile sensor big data

---

**MD2K**

Center of Excellence for Mobile Sensor Data-to-Knowledge

- Cornell Tech
- Georgia Tech
- U. Memphis
- Northwestern
- Ohio State
- Open mHealth
- Rice
- UCLA
- UC San Diego
- UC San Francisco
- UMass Amherst
- U. Michigan
- WVU
Mobile Sensor Data Sources in MD2K

AutoSense sensors: ECG, respiration, accelerometers

EasySense (contactless) sensors: heart motion, lung motion, lung fluid level

Microsoft Band: accelerometers, gyroscopes, HR

Smartphone sensors: GPS, accelerometers, self-report

MotionSense HRV: accelerometers, gyroscopes, PPG

Smart toothbrush: brushing, Pressure

Advancing biomedical discovery and improving health through mobile sensor big data

Cornell Tech  ♦  Georgia Tech  ♦  U. Memphis  ♦  Northwestern  ♦  Ohio State  ♦  Open mHealth
Rice  ♦  UCLA  ♦  UC San Diego  ♦  UC San Francisco  ♦  UMass Amherst  ♦  U. Michigan  ♦  WVU
Utility of Collecting High-frequency Sensor Data

Traditional Approach

Biomarker data

Hundreds of samples/day

Biomedical Research

Our Proposal

Raw sensor data

Biomarker Identification

Biomarker Development

Biomarker Validation

Biomarker Improvement

Biomedical Research

Advancing biomedical discovery and improving health through mobile sensor big data

Cornell Tech ♦ Georgia Tech ♦ U. Memphis ♦ Northwestern ♦ Ohio State ♦ Open mHealth
Rice ♦ UCLA ♦ UC San Diego ♦ UC San Francisco ♦ UMass Amherst ♦ U. Michigan ♦ WVU
mHealth Biomarkers Developed in MD2K

Advancing biomedical discovery and improving health through mobile sensor big data

Cornell Tech ✦ Georgia Tech ✦ U. Memphis ✦ Northwestern ✦ Ohio State ✦ Open mHealth
Rice ✦ UCLA ✦ UC San Diego ✦ UC San Francisco ✦ UMass Amherst ✦ U. Michigan ✦ WVU
Detecting First Lapses in Smoking Cessation

Saleheen, et. al., ACM UbiComp 2015

Modeling Challenges

1. Ephemeral (very short duration)
   - 3~4 sec for each puff
   - 10,000 breaths in 10 hours
   - 2,000 hand to mouth gestures
   - But, only 6~7 positive instances
   - Need high recall & low false alarm

2. Numerous confounders
   - Eating, drinking, yawning

Main Results

• Applied on smoking cessation data from 61 smokers
• Detected 28 (out of 33) first lapses
• False alarm rate of 1/6 per day

Limitations

• Can’t detect if sensor not worn
• Can’t detect if data quality is poor
• Needs adaptation for e-cigarettes
• Difficult to validate temporal accuracy of smoking detection
Sensors-to-Markers-to-Interventions: The Case of Sensor-Triggered Stress Intervention

**SENSE**
- ECG, Respiration
- Accelerometer

**ANALYZE**
- cStress
- Activity

**ACT**
- Time series pattern mining
- Stress Episode Detection
- Intervention Trigger

- High data rate streaming
- Long battery life
- High data yield
- Real-time data quality screening
- Personalized machine learning models
- Biomarkers of health, behavior, and environment
- Validated in lab and field
- Detect trend in noisy and rapidly varying time series
- Robust to confounders and data losses
- Adapt intervention prompts to current context (e.g., driving)
1. A model to convert ECG & respiration sensor data into a continuous measure of stress
   - input: vector of all one-minute ECG + RIP statistical/aggregate features
   - output: “stress”/“non-stress” label
   - alternative output: probability of “stress”

2. Learned with Support Vector Machines (SVM)
   - Careful handling of sensor data
   - Parameters tuned for optimizing F1 score
   - Cross-subject validation for generalizability
   - Data loss (0.27 hr/day)

3. Validated on independent data sets
   - Against lab stress protocol for lab data
   - Against self-report for field data
Feature Computation from ECG

- Screen for data quality
  - Morphology, saturation, etc.
- Find R peaks
  - Pan and Tompkin’s algorithm
- Compute R-R interval
- Detect/remove noisy R-R intervals
- Normalize R-R intervals
  - Use “winsorized” (capped) mean/stdev estimates
  - Filter out activity affected data
- Compute one-minute statistics of R-R intervals

<table>
<thead>
<tr>
<th>HRV</th>
<th>variance, quartile deviation, low frequency energy (0.1–0.2Hz), medium frequency energy (0.2–0.3Hz), high frequency energy (0.3–0.4Hz), low:high frequency energy ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>non-HRV</td>
<td>mean, median, 80th percentile, 20th percentile, heart-rate</td>
</tr>
</tbody>
</table>
Feature Computation from Respiration

- Screen for data quality
  - Morphology, loosening
- Locate respiration cycles
- Detect/remove invalid cycles
  - **Amplitude**: > 20% of mean
  - **Duration**: 0.9 - 12.5 sec
- Compute base features
- Normalize base features
  - As in ECG
- Compute one-minute statistics of base-features

<table>
<thead>
<tr>
<th>Base Features</th>
<th>Aggregations</th>
</tr>
</thead>
<tbody>
<tr>
<td>inspiration duration, expiration duration, I:E ratio, stretch, respiratory sinus arrhythmia (RSA)(^1)</td>
<td>mean, median, 80th percentile, quartile deviation</td>
</tr>
<tr>
<td>breath-rate(^2), inspiration minute volume(^2)</td>
<td></td>
</tr>
</tbody>
</table>

---

1. RSA
2. breath-rate
Minimizing the Data Loss due to Physical Activity

Sarker, et. al., ACM CHI 2016

Data Loss due to Physical Activity

Recovered: 27.6%

Heart-rate exponential recovery

\[ HR_{\text{RR}} = HR_{\text{Rest}} + (HR_{\text{Peak}} - HR_{\text{Rest}}) e^{-\frac{t-t_0}{\tau}} \]  
[Freeman, PCD 2006]

<table>
<thead>
<tr>
<th></th>
<th>Discard 2 minutes</th>
<th>Proposed Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Activity</td>
<td>22.7%</td>
<td>22.7%</td>
</tr>
<tr>
<td>Residue</td>
<td><strong>35.0%</strong></td>
<td><strong>7.4%</strong></td>
</tr>
<tr>
<td>Total</td>
<td>57.7%</td>
<td>30.1%</td>
</tr>
</tbody>
</table>
Training & Validation Methods

- **Lab**: Model trained using lap protocol
  - Public speaking, mental arithmetic, and cold pressor sessions

- **Field**: A Bayesian Network model to map minute-level outputs from cStress to self-reports
  - Random prompts (5-15 per day)
Validation in Independent Lab and Field Data

- **Lab Validation**: Cross-subject validation with $n=21$, 1600 minutes of lab data
- Stress sessions consist of public speaking, mental arithmetic, cold pressor

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>F1</th>
<th>AUC</th>
<th>Hit-rate</th>
<th>TPR</th>
<th>FPR</th>
<th>C. Kappa</th>
<th>Optimal hyper-parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>0.81</td>
<td>0.96</td>
<td>0.93</td>
<td>0.84</td>
<td>0.05</td>
<td>0.77</td>
<td>$C=90.5097$</td>
</tr>
<tr>
<td>ECG</td>
<td>0.78</td>
<td>0.95</td>
<td>0.92</td>
<td>0.72</td>
<td>0.05</td>
<td>0.73</td>
<td>$\gamma=0.000345267$</td>
</tr>
<tr>
<td>HRV</td>
<td>0.56</td>
<td>0.78</td>
<td>0.84</td>
<td>0.55</td>
<td>0.1</td>
<td>0.46</td>
<td>$bias=0.339329$</td>
</tr>
<tr>
<td>RIP</td>
<td>0.75</td>
<td>0.93</td>
<td>0.90</td>
<td>0.83</td>
<td>0.09</td>
<td>0.69</td>
<td>$C=2$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$\gamma=0.00552427$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$bias=0.340407$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$C=724.077$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$\gamma=0.0220971$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$bias=0.250926$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$C=1448.15$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$\gamma=0.000488281$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$bias=0.308312$</td>
</tr>
</tbody>
</table>

- **Field Validation**: 1601 self-report EMA from $n=23$ over 7 days in the field
- Bayesian network model to map cStress onto self-reported stress data

<table>
<thead>
<tr>
<th></th>
<th>train</th>
<th>field</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median F1</td>
<td>0.75</td>
<td>0.71</td>
</tr>
<tr>
<td>Median AUC</td>
<td>0.85</td>
<td>0.60</td>
</tr>
<tr>
<td>Median Accuracy</td>
<td>0.9</td>
<td>0.72</td>
</tr>
</tbody>
</table>
Validation of cStress on 38 Drug Users Dataset

4 weeks of sensor wearing by polydrug users at NIDA IRP (PI: Dr. Kenzie Preston)

Overall Cronbach’s Alpha of the Study = 0.843

Cronbach’s Alpha of last five participants = 0.335

[Sarker, et. al., ACH CHI’16]
Validation of cStress on Smoking Data

1 day pre-quit and 3 days post-quit sensor wearing by 61 newly abstinent smokers at UMN (PI: Dr. Mustafa al’Absi)

- Overall Cronbach’s Alpha of the Study = 0.76
- Median F1 score = 0.646

- Lower F1 score than other datasets
  - Imputed missing data, but using simple carry-forward
  - Lower consistency of self-reports (0.76 vs. 0.843)
Stress Likelihood Timeseries

- Stress
- Not Stress
- Binary Classification Threshold

Participants

Stress Likelihood
Mining Stress Episodes in cStress Time Series

Stress Likelihood \rightarrow \text{Stress Density}

Options for Intervention Timing

Stress Density = \text{Area/Time}
Generating Intervention Triggers - Model Training

Moving from a Single Threshold to Dual Thresholds To Optimize Confidence

![Stress Density of an Episode](image)

- **Stressed**
- **Unsure**
- **Not Stressed**

**Advancing biomedical discovery and improving health through mobile sensor big data**

Cornell Tech ♦ Georgia Tech ♦ U. Memphis ♦ Northwestern ♦ Ohio State ♦ Open mHealth Rice ♦ UCLA ♦ UC San Diego ♦ UC San Francisco ♦ UMass Amherst ♦ U. Michigan ♦ WVU
Thresholds for Reactive Stress Intervention

<table>
<thead>
<tr>
<th>Lab Study (Stress Density)</th>
<th>Precision and Recall</th>
<th>Field Study (per day)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>95%</td>
<td>90%</td>
</tr>
<tr>
<td>Threshold 1</td>
<td>0.29</td>
<td>0.29</td>
</tr>
<tr>
<td>Threshold 2</td>
<td>0.44</td>
<td>0.42</td>
</tr>
<tr>
<td>Field Study (per day)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Precision and Recall

<table>
<thead>
<tr>
<th></th>
<th>90%</th>
<th>85%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Threshold 1</td>
<td>0.36</td>
<td>0.36</td>
</tr>
<tr>
<td>Threshold 2</td>
<td>0.33</td>
<td>0.36</td>
</tr>
</tbody>
</table>
Utility of Collecting High-frequency Sensor Data

Traditional Approach

- Biomarker data
  - Hundreds of samples/day
  - Biomedical Research

Our Proposal

- Raw sensor data
  - Millions of samples/day
  - Biomarker Identification
  - Biomarker Development
  - Biomarker Validation
  - Biomarker Improvement
  - Biomedical Research

MD2K
Advancing biomedical discovery and improving health through mobile sensor big data
Cornell Tech • Georgia Tech • U. Memphis • Northwestern • Ohio State • Open mHealth
Rice • UCLA • UC San Diego • UC San Francisco • UMass Amherst • U. Michigan • WVU
MD2K Mobile Software Platform (open-source)

- **High Data Rate Sensors**
  - Chestband (ECG, RIP, Accel)
  - Smart Watches
  - EasySense Chest Sensor
  - 70+ million samples/day
  - 13 million
  - 10+ million
  - 38+ million

- **Low Data Rate Sensors**
  - Oral-B Toothbrush
  - iCO Smokerlyzer
  - Omron BP/Weight

- **Smartphone**
  - Wireless Radios: Bluetooth, ANT+
  - Machine Learning Models

- **Biomarkers**
  - Stress
  - Smoking
  - Eating
  - Lung Congestion
  - Heart Motion
  - Location
  - Activity
  - Driving
  - Drug use

- **Privacy Controller**
  - Encryption-based Privacy Filter

- **Data Quality**
  - Smartphone Sensors (GPS, Accel, Gyro)
  - 1500
  - 4
  - 6

- **Secure Local Data Storage**

- **Secure Cloud Data Storage: Cerebral Cortex**

- **Sensitive Marker Obfuscation**

---

**MD2K**

Advancing biomedical discovery and improving health through mobile sensor big data

- Cornell Tech
- Georgia Tech
- U. Memphis
- Northwestern
- Ohio State
- Open mHealth
- Rice
- UCLA
- UC San Diego
- UC San Francisco
- UMass Amherst
- U. Michigan
- WVU
Key Capabilities of mCerebrum

1. Support for high-frequency streaming data
   – 800+ Hz for 70 million samples per day

2. Connectivity to diverse sensors and radio
   – ANT, Bluetooth, Bluetooth Low Energy (BLE), etc.

3. Continuous data collection and real-time data quality monitoring

4. Real-time computation of biomarkers
   – Stress, smoking, driving, activity, etc.

5. Biomarker-triggered notification/intervention
Field Studies Using MD2K Software

<table>
<thead>
<tr>
<th>Study</th>
<th>Users</th>
<th>Person-Days</th>
<th>Samples (Billions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Northwestern (Smoking and Eating)</td>
<td>225</td>
<td>3,150</td>
<td>136</td>
</tr>
<tr>
<td>Rice (Smoking)</td>
<td>300</td>
<td>4,200</td>
<td>182</td>
</tr>
<tr>
<td>Utah (Smoking)</td>
<td>300</td>
<td>4,200</td>
<td>182</td>
</tr>
<tr>
<td>Vermont (Smoking and fMRI)</td>
<td>90</td>
<td>1,260</td>
<td>55</td>
</tr>
<tr>
<td>Moffitt (Smoking and Stress)</td>
<td>24</td>
<td>336</td>
<td>15</td>
</tr>
<tr>
<td>Ohio State (Heart Failure)</td>
<td>225</td>
<td>6,750</td>
<td>224</td>
</tr>
<tr>
<td>UCLA (Oral Health)</td>
<td>162</td>
<td>29,160</td>
<td>968</td>
</tr>
<tr>
<td>Johns Hopkins (Cocaine Use)</td>
<td>25</td>
<td>350</td>
<td>18</td>
</tr>
<tr>
<td>Dartmouth (Behavior Change)</td>
<td>100</td>
<td>1400</td>
<td>58</td>
</tr>
<tr>
<td>Minnesota (Workplace Performance)</td>
<td>800</td>
<td>56,000</td>
<td>2,891</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>2,251</td>
<td>106,806</td>
<td>4,729</td>
</tr>
</tbody>
</table>

Entire ecosystem (sensors, software, cloud) to be available end of 2017
mProv: Provenance Cyberinfrastructure for Mobile Sensor Data

Due to lack of data sharing, everyone needs to collect their own data.

Sharing of raw mobile sensor data can accelerate research, but provenance infrastructure is needed to enable reproducibility and comparative analysis.

<table>
<thead>
<tr>
<th>Velocity</th>
<th>Variety</th>
<th>Volume</th>
<th>Variability</th>
<th>Veracity</th>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hundreds of samples/sec per sensor</td>
<td>Tens of sensors per sensor</td>
<td>Gigabytes per day per person</td>
<td>Variations in attachment, placement, signal quality</td>
<td>Multiple biomarkers from same sensor</td>
<td>Sources of validation for specific biomarkers</td>
</tr>
</tbody>
</table>

Advancing biomedical discovery and improving health through mobile sensor big data

Cornell Tech ♦ Georgia Tech ♦ U. Memphis ♦ Northwestern ♦ Ohio State ♦ Open mHealth ♦ Rice ♦ UCLA ♦ UC San Diego ♦ UC San Francisco ♦ UMass Amherst ♦ U. Michigan
Indicators of Everyday Job Performance

Big Five Personality Traits

- Agreeable
- Open
- Extraverted
- Conscientious
- Stable

work hard now. it'll pay off later.

Counterproductive Workplace Behaviour

GOOD JOB

CORPORATE CITIZENSHIP
Healthier, Wealthier, and Happier You

Detect → Predict → Adapt

healthy me

Success
Growth
Career

Advancing biomedical discovery and improving health through mobile sensor big data

Cornell Tech ● Georgia Tech ● U. Memphis ● Northwestern ● Ohio State ● Open mHealth
Rice ● UCLA ● UC San Diego ● UC San Francisco ● UMass Amherst ● U. Michigan
For More Information

• MD2K website: md2k.org; Email: info@md2k.org
• mProv Website: mprov.md2k.org/
• Software Overview: md2k.org/software/platform
• Software Download: GitHub: github.com/MD2Korg
  • 20+ mCerebrum Android applications
• Software Documentation: docs.md2k.org
• Questions and Answers: discuss.md2k.org
• mHealthHUB: mhealth.md2k.org/
• mHTI: mhealth.md2k.org/mhealth-training-institute