Data Science and Applied Technology Core RC4

Investigators: Todd Manini, Parisa Rashidi, Jiang Bian, Mamoun Al Mardini, Sanjay Ranka

Professional staff: Tonatiuh Viramontes Mendoza, Matthew J McConnell

Current trainees: Scott Vouri, Erta Cenko, Yashaswi Karnati

Past trainees: Dottington Fullwood, Anis Davoudi, Subhash Nerella, , Aseem Thakkar, Rohith Kessi Reddy, Duane Corbett, Josh Brown, Matin Kheirkhahan, Sanjay Nair (graduated), Manoj Battula

Overall description: Data Science and Applied Technology (DSAT) Core (RC4) provides many unique attributes, such as: developing software for interactive mobile technology (e.g., wearable sensors that are programmable in real time); validating new sensing technology; warehousing data; repurposing data; and applying machine learning techniques to domain problems. DSAT provides a central hub of expertise in computer science, biomedical engineering, biomedical informatics, data science, applied technology, epidemiology, and content expertise in the assessment of mobility

The core has 800 sqft of dedicated space to conduct research. This space is outfitted with 8 workstations with large monitors (>20 inches). It also has an 8X4 ft conference table that serves as collaboration space. with office is adjacent to the space where clinical assessments are performed. The space is designed for programmers and trainees to download, store and analyze data. The space is situated immediately office/work areas of clinic research staff that include study coordinators, student assistants, and collaborating investigators. The core also has full access to The Department of Computer and Information Sciences and Engineering (CISE) computer cluster space consisting of a head node with dual Opterons, 16GB of memory and 3.5TB of storage with 20 worker nodes with dual Opterons and 32GB of memory running Linux (Ubuntu Server 10.04). These will be used for prototype software development. All graduate students have access to a workstation that can be used to access this cluster. All faculty offices are equipped with a Windows or Linux workstation with standard software installations. Wireless access is available throughout the CSE Building and all of campus.

Real-time Online Assessment and Mobility Monitor app

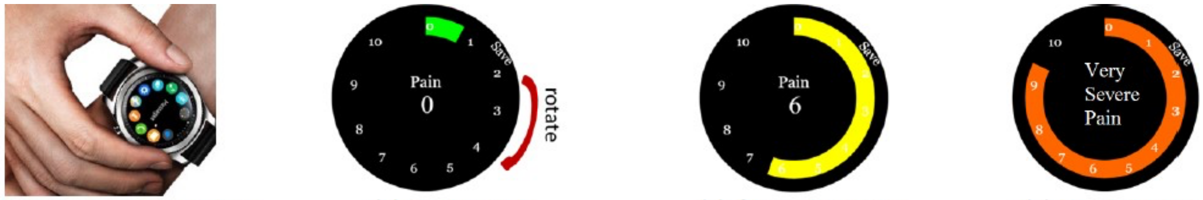
ROAMM (Real-time Online Assessment and Mobility Monitor) is a customizable platform, developed at the University of Florida, designed specifically for smartwatches. It offers long-term and continuous connectivity, bidirectional interactivity, and remote programmability through a smartwatch. The goal of this platform is to allow for remote monitoring of patients using smartwatches. This approach reduces and potentially removes in-person repeated follow-up visits while also providing detailed objective information surrounding the onset and recovery from a fall or other episodic health events (e.g. hospitalizations). Moreover, because of their immense popularity, mobile devices overcome racial, geographic and educational divides and has potential of digitally engaging and retaining diverse populations in healthcare research.

ROAMM spans four main categories of measurement: i) mobility, ii) ecological momentary assessment/patient reported outcomes; iii) cognition; and iv) intervening health events monitoring using a smart watch platform. The components include: 1) a secure smart watch application (app), which collects and summarizes sensor monitored data (e.g. tri-axial accelerometer, GPS location); 2) a graphical interface allowing ecological momentary assessments; 3) interface and/or voice recording for cognitive assessment; and 4) secure servers for configuring the application parameters, data storage, advanced analytics, web-based data visualization and remote management of sensors.

ROAMM watch app. The watch application is built on the Samsung Galaxy and Apple watch. The application collects self-reported as well as sensor- monitored data and processes them into informative features. The mobile application has the following unique features and advantages:

1. Data is transferred using 4G and LTE GSM allowing for automatic real-time data uploads in a minimally obtrusive manner
2. It is flexible enough to accommodate different types of studies with different targets and outcomes while allowing for constructions of variables instantly from the raw data on the device.

3. It supports interactive interfaces (e.g., prompting for reporting symptoms or asking to recharge), as seen in Figure 2.
4. Configuration can occur remotely (e.g. stopping GPS to save battery life).



(a) Bezel or button rotation *(b) Rating pain* *(c) Changing color* *(d) Final design*

Figure 2. ROAMM app interface showing rating of a patient reported outcome (e.g., pain)

ROAMM server. The server configures the application parameters for data collection and provides a means for data storage, analysis, and visualization services to the researcher/administrator. The server software consists of the following features and advantages:

1. It provides a user-friendly platform to configure and review data from the watches' applications remotely. Such configuration includes identifying the sensors to use, their sampling rate, and setting the parameters necessary for aggregating the collected raw data. It provides a user interface that summarizes statistics about the status of the watches and the underlying data.
2. For better control, user management and security modules are housed on the server. New roles (e.g., researcher and administrator) with the desired access and privileges are defined and assigned.

The server can receive data from hundreds of smart watches operating simultaneously in the field and store the data in a high-performance and fault-tolerant database. Data security is maintained by granting the access to the database only through the web interface, which itself retrieves data via a database view with least privileges. The server is founded on Amazon Web Services (AWS) which allows flexible options for ROAMM to grow. AWS give us the tools to provide a user-friendly platform interface for visualization, campaign management (activating the watches, participant management, creating new campaigns, downloading the data) and to modify configurations remotely. Configuration includes identifying the sensors to use, constructing questions, timing, sampling rate, and aggregating raw data. AWS is used to develop highly scalable, secure serverless architecture to enable remote data collection and real time transmission. For security, including Identity and Access Management, only admins are permitted to manage user access to different resources (software-and hardware-based encryption) that protects against distributed denial-of-service attacks (DDoS attacks). We currently use the following services to enable the ROAMM platform:

Elastic Computer Cloud (EC2), Lambda, API Gateway, Dynamo Database, Simple Storage Service (S3), Transcribe, relational database services (RDS).

a. Lambda is used to define micro services that support real time campaign management, data collection, transmission, and user management.
b. API gateway is used as a management tool that sits between the watches and Lambda.
c. RDS, S3 buckets are used to store the data collected and support the visualization dashboards hosted on an EC2 instance.

ROAMM campaigns. Study "campaigns" allow investigators to customize their watch graphical interface while also adjusting sensor settings. Campaigns are built through a tabbed workflow via a user-friendly web-portal. The portal also serves as a hub for initializing smartwatches and aggregating information from sensors and EMA in a uniform way that can be visualized in real-time. First, users are able to control how the sensors operate on the smartwatch (sampling rate, activation status) and what aggregate features to compute and store in the database (total activity, raw accelerometer). Second, users are able to select and control ecological assessments and patient reported outcome prompts via the graphical interface - when they are delivered, what questions to ask, scaling choices (e.g. 0-10, 1-5), frequency and color scheme. During this customization, certain limits are placed on users— e.g. amount of text and scaling placed on the screen. Importantly, software developers or computer programing expertise is NOT required to customize these elements— modifiable elements are presented in a simple "user friendly" interface and work flows through the ROAMM web portal.

The *ROAMM web portal* section provides details about the components of this customizable portal. Following is a summary of the flexible features ROAMM campaigns provide:

- Allows a researcher to execute a study that is customized to their needs.
- Allows to configure both user-reported out comes and sensor-based data
- Supports the collection and analysis of this data in a robust and scalable fashion, we developed an event driven, serverless computing platform using AWS cloud services
- Allows multiple campaigns to run concurrently each under the auspices of a different researcher

ROAMM data security. Data security during transmission and storage are of utmost importance for our future goals. We have cleared our existing approaches through the UF security protocols. Data are transmitted to a cloud-based storage and database management environment on Amazon Web Services (AWS). AWS security standards for data storage and HIPAA risk management program aligns with FedRAMP and NIST 800-53, which are higher standards than the HIPAA Security Rule. Following are the AWS security features provided by the ROAMM platform.

- **HIPPA compliant**: AWS enables covered entities and their business associates subject to HIPAA to securely process, store, and transmit PHI
- **Identify & access management**: securely manage identities, resources, and permissions among team members and customers.
- **Data protection**: AWS provides services that help you protect your data, accounts, and workloads from unauthorized access. AWS data protection services provide encryption and key management and threat detection that continuously monitors and protects your accounts and workloads.
- **Infrastructure protection**: AWS protects web applications by filtering traffic based on rules that you create. For example, you can filter web requests based on IP addresses, HTTP headers, HTTP body, or URI strings, which allows you to block common attack patterns, such as SQL injection or cross-site scripting.

Wearable sensor hardware

- The core possesses 60 Actigraph GT1M, GT3X and LINK accelerometer models (The Actigraph Inc. Pensacola, FL). The monitors are small (3.8 x 3.7 x 1.8 cm), lightweight (27 g) and include a uniaxial and triaxial accelerometers. The accelerometers measure accelerations in the range of 0.05-2 G with a band-limited frequency of 0.25-2.5 Hz. The monitors are initialized, and data downloaded with the ActiLife software (Version 3.3.0).
- The core owns 35 Samsung Gear S smartwatches that possess customized software to program “apps”. Programming is done in TIZEN and Android operating systems. The applications loaded on the Gear S device run in a webkit based browser environment. Tizen provides API libraries to interface with its sensors as well as other system-level functionality and notifications.
- The core owns and distributes 25 Apple watch 3's and SE series smartwatches.
- The core owns and operates 4 Android mobile smartphones (2 Nexis and 2 Galaxy's) and 4 Apple iPhones.
- The core possesses multi-sensor technology through a portable armband (HealthWear Bodymedia, Pittsburgh, PA). The Sensewear armband uses a dual-axis accelerometer, a heat flux sensor, a galvanic skin response sensor, a skin temperature sensor, and a near-body ambient temperature sensor to capture data. Data from multi-sensor technologies are comparable to energy expenditure measured with doubly-labeled water.
- The core possesses 4 Empatica E4 wristband wrist worn wearable multi-sensor. The E4 measures blood volume pulse through a photoplethysmography Sensor - from which heart rate, heart rate variability (HRV), and other cardiovascular features are derived. An electrodermal Activity Sensor measures sympathetic nervous arousal and derives features related to stress, engagement and excitement. It also has a tri-axial accelerometer, event mark button and infrared thermopile for peripheral skin temperature.

Wearable technology validation and implementation services

- Validation against “gold-standard” measures of energy expenditure via indirect calorimetry and visual observation
- The core conducts focus groups and key informant interviewing to evaluate the acceptance of new technology. These groups help to optimize the adherence and retention in future studies utilizing this technology.
- The core possesses two k5 Cosmed Indirect calorimeters for validation of wearable technology

Mobility and activity measures using wearable technology

- Summary measures of physical activity include:
    - Total physical activity time (any type of activity at any intensity)
    - Time spent at specific intensities of physical activities
        - Sedentary
        - Light
        - Light-moderate
        - Moderate
        - Moderate-vigorous
        - Vigorous
- Mobility characterization
    - Total steps per day
    - Cadence (steps/min during active bout)
    - Step bouts – steps taken at a specific pre-determined cadence
- Basic GPS monitoring and tracking
    - Excursion size - average of maximum distance from the home for each excursion away from home
    - Excursion span - average daily maximum distance between all recorded locations away from home. Measures travel clusters, independent of maximal distance traveled.
- Geographical information systems for combining mobility patterns with the contextual environment
    - Geocoding and mapping according to CDC tracts
    - Adjacency/distance measures – for measuring distances between places and mobility patterns
    - Overlays – Points of interest (crime locations, walkways, sidewalks, parks, transportation services)

Wearable technology device administration, initialization and cloud computing

Databases available (either stored publicly or in a local repository) for secondary data analyses supported by RC4

Study name	N	Age	Design
ADAPT[1]	316	65+	RCT
ENRGISE	300	70+	RCT
COVID-19 survey*	1392	20+	OLC
GEM[2]	3069	75+	RCT
LEAPS[3]	408	62	RCT
LIFE[4] ^	1,635	70-89	RCT
LIFE-Pilot[5] ^	424	70-89	RCT
PEAKS	100	70+	OCS
LOOK-AHEAD[6, 7]	5,145	45-75	RCT
SHEP[8-10]	4,736	60+	RCT
TRAIN[11-13]	290	50+	RCT
TTrial[14, 15]	790	65+	RCT
ChoresXL	280	20+	OCS
Baltimore Longitudinal Study of Aging ^	1581	20+	OLC
Sacopenia Definitions and Outcomes Consortium *	18,767	70+	OLC
EPESE[16, 17] ^	14,000	65+	OLC
OneFlorida Data Trust* (see citation)	15,700,000	20+	OLC/EHR
Health ABC[18, 19] ^	3,075	70-80	OLC
InChianti[20, 21]	1,020	65+	OLC
mtDNA genomic sequence*	3,499	70-89	OLC+RCT
NHANES[22] ^	1,286	65+	OCS
UDS-NACC Alzheimer's[23] ^	32,364	50+	OLC
Aging EHR hospital repository*	21,615	65+	OCS/EHR
UK Biobank* ^	502,625	20+	OLC
PECAN - preoperative physical & cognitive impairment	14,000	65+	OLC/EHR
UF OAIC pilot studies[24]	519	60+	RCT, OLC
UF INFORM**[25]	9,000	50+	OLC
WHAS[26, 27]	1,002	65+	OLC
WHI[28-30]	161,808	50-79	RCT, OLC
WHIMS[31]	4532	50-79	RCT, OLC
Total	**16,509,578**		

*=new study; ** Movement disorders database, see letter of Dr. Okun; RCT=Randomized Clinical Trial; OLC= Observational Longitudinal Cohort; OCS= Observational Cross-Sectional; EHR=Electronic Health Records. ^ data used by OAIC, but is publicly available

- A Map-Reduce framework (Apache Spark), as well as pre-defined scripts that leverage machine learning methods scaled for big data (SparkML), are included in the server software to be able to retrieve and analyze large amounts of data in real-time.
- Amazon web-services capability
- The server provides a platform to register participants, personalize the application based on their preferences, and configure data col- lection settings according to study requirements.
- Data collection configuration includes identifying active sensors, specifying their sampling rates, and defining the parameters used to aggregate the raw data into study-required variables.
- All of the configuration steps on the server are done remotely and without the requirement that the watches be collected from and returned to the participants.

Data repository

A data repository of de-identified data has been created for investigators to address age-related questions. A description of data available to investigators are listed in the adjacent table.

Big data analyses

- Time series data analyses and processing
- Supervised, semi-supervised and unsupervised machine learning
  - decision trees (random forest, classification of regression trees)
  - support vector machines, bag of words
  - deep learning
  - feature extraction
  - feature validation
  - pattern discovery
  - cluster analysis
  - models (support vector machines, artificial neural networks, deep learning)
- Epidemiological analyses [regression, random effects modeling, event time analyses (Poisson, Cox Hazard)].
- A suite of software to conduct a variety of analyses and visualization. These include STATA, SAS, SPSS, StatTransfer, MatLAB, Labview 16.0, R v3.4, JAVA and Enthought Canopy (Python).

Web portal development

- The portal provides information from all actively deployed watches and the collected data for each separate device.
- It presents summary statistics of activities, the current status of the watches, and detailed visualizations for the activity data.
- The data can be accessed through the web portal for data exploration and analysis

Data visualization

- Data are Visualized using GraphPad Prism 5.0, Tableau, Visual Studio, Adobe Photoshop, Adobe Illustrator and custom visualization packages in R.

Fast Healthcare Interoperability Resources (FHIR) SMART FHIR

- Fast Healthcare Interoperability Resources (FHIR) is a way for merging data formats for sharing electronic health records. It serves as a way to facilitate interoperation between mobile technology and electronic health systems. The core contains expertise for integrating mobile technology with FHIR exchange patient-generated data with the Epic® health record system and vice-versa.
- App Orchard for developing new EPIC interface apps

Consulting services. These services are provided due to the expertise as of the current investigators

- The UF Health Integrated Data Repository (IDR) and I2B2.
    - The core provides consulting services with regard to the IDR when used for aging-specific studies that meet the theme of the OAIC. The IDR enables new research discoveries as well as patient care quality and safety improvements through a continuous cycle of information flows between our clinical enterprise and research community. In its simplest form, a data repository is a collection of disparate data organized in a manner that lends itself to understanding relationships between data elements to answer questions.
    - The UF Health IDR currently consists of a Clinical Data Warehouse (CDW) that aggregates data from the various clinical and administrative information systems, including the Epic EMR. The CDW contains demographics, inpatient and outpatient clinical encounter data, diagnoses, procedures, lab results, medications, select nursing assessments, co-morbidity measures and select perioperative anesthesia information system data.
    - The CDW data contains “Fully Identified Data” and is fundamental to institutional business processes and secured per UF&Shands policies.
    - Access to IDR data is provided through the NIH-funded i2b2 tool, which provides researchers access to a HIPAA-compliant and IRB-approved “Limited Data Set.” Faculty researchers can query the i2b2 Limited Data Set to identify cohort counts as they prepare grant proposals, plan clinical trials, and write IRB protocols.

Partnerships established with the DSAT core.

- **E-health core.** Dr. Bian leads the e-Health core as part of the The University of Florida Health Cancer Center (UFHCC) Cancer Informatics Shared Resources program. This partnership, along with Dr. Bian investigator status on DSAT, allows for the two resources leverage knowledge gained in each program separately. The Shared Resources’ functions include:
    - Design, develop, and implement novel informatics methods, tools, and systems for capturing and synthesizing data to support clinical activities and clinical research;
    - Develop tools and methods to transform data collected by eHealth technologies, integrating with other relevant data sources, into actionable knowledge.
    - Develop and implement eHealth interventions and make eHealth tools freely available;
    - Support investigators to engage communities and key stakeholders in the development of eHealth tools and other patient- or clinician-facing technologies that are relevant to addressing the needs of patients, especially those in the UF Health catchment area, in collaboration with the UFHCC Community Outreach and Engagement (COE) Program;
    - Facilitate integration of tools into the electronic health record (EHR) systems and liaise with the EPIC/MyChart team at UF Health;
    - Liaise with the IDR/i2b2 team at UF Health when new data elements are being considered for i2b2; and
    - Liaise with UF Health IT security to ensure that tools being developed meet the security and privacy standards necessary.

- **OneFlorida Clinical Data Research Network.** RC4 actively utilizes the resources provided by the OneFLorida Data Trust. Following an investment of $100 million, in 2011 UF Health opened a new electronic medical record system and a clinical data warehouse that was the foundation for the development of an integrated data repository. Over the past 4 years, the IDR system expanded to the OneFlorida Network— a statewide Clinical Data Research Network (CDRN) that will join the PCORnet to optimize opportunities for conducting comparative effectiveness research (CER).

- In 2012, One Florida cared for over 15 million unique patients, or about 70% of all Floridians, through a network of 22 hospitals, 1240 clinics/practices and 4100 providers.
- The centerpiece of the One Florida CDRN is the OneFlorida Data Trust, a secure, de-identified data repository in which UF Health, Orlando Health, Florida Medicaid/CHIP, and the Florida Department of Health currently participate.
- To date, the OneFlorida Data Trust houses data on 15 million patients, including demographic information, diagnoses, procedures, lab results, health care visit details, nurse assessments, bio-specimen availability, and vital statistics records.

- ***intelligent HEAlth Lab* (i-HEAL).** Dr. Rashidi’s laboratory is a close partner of DSAT. The lab focuses on: (1) transforming patient care in the Intensive Care Unit by developing autonomous monitoring tools using advanced machine learning techniques and (2) Developing intelligent tools for monitoring activity, cognitive and mental conditions of community-dwelling patients.
- Resources and personnel are often shared between the two entities. The *intelligent HEAlth Lab* (i-HEAL) is located at the New Engineering Building (NEB). It includes desk space for up to 10 students. It contains a Dell Precision T5610 server with 64GB of memory and Dual Intel® Xeon processor and five networked workstations, each being equipped with four microprocessors. The software licensed to Dr. Rashidi’s lab for advance data analysis and programming include MTLAB, Visual Studio, Enthought Canopy (Python), WinEdt for LaTeX editing, and Microsoft suite for text and graphics processing.