CIF21 DIBBs:EI mProv: Provenance-based Data Analytics Cyberinfrastructure for High-frequency Mobile Sensor Data

PI: Santosh Kumar
Co-PIs: Zachary Ives, Ida Sim, Mani Srivastava
Chief Software Architect: Timothy Hnat

Overview of proposed mProv project

Growing Potential of High-Frequency Mobile Sensor Data

HEALTH AND WELLBEING

intelligent city

smart energy

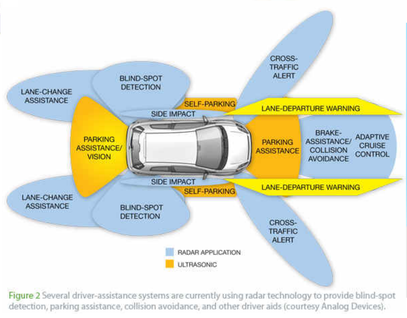


Figure 2 Several driver-assistance systems are currently using radar technology to provide blind-spot detection, parking assistance, collision avoidance, and other driver aids (courtesy Analog Devices).

Health applications are a natural focal point for research using sensor data.

Barriers in Conducting Research with Mobile Sensor Data

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

Research Funding

IRB INSTITUTIONAL REVIEW BOARD SERVICES

Research Participants Wanted

EXPECT DELAYS

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

Velocity	Variety	Volume	Variability	Veracity	Validation
Hundreds of samples/ second per sensor	Tens of sensors per device	Gygabytes/day per person	Variations in attachment, placement, signal quality	Multiple biomarkers from same sensor	Sources of valuation of specific biomarkers

mProv: Provenance Cyberinfrastructure for Mobile Sensor Data

- Builds on top of open-source MD2K cyberinfrastructure that enables collection, curation, analysis, visualization, and interpretation of high-frequency mobile sensor data
- mProv provides *data models*, *metadata standards*, *API's*, and *runtime support* for annotating sensor data streams with
  - Source – sensor type, placement, sampling rate, continuous/episodic
  - Semantics – number, probability, class/category
  - Provenance – features and rules applied to obtain a biomarker
  - Validation – specificity, sensitivity, benchmark, gold standard
  - Privacy – user controls exercised and applicable privacy policies

Proposed Works for the mProv Project

- Provenance solutions for reasoning about uncertainty and variability
  - Key construct is to create windows of data that can be treated as a unit
  - A data item can be assigned to multiple overlapping windows of different sizes
  - Provenance for each output is derived from input windows
- Create annotation model for time-varying mobile sensor data
- Develop a sensor data stream provenance annotation system
- Provide an archival service for data and metadata streams
- Manage sensor data sharing and resulting privacy risks

Validation Case Studies

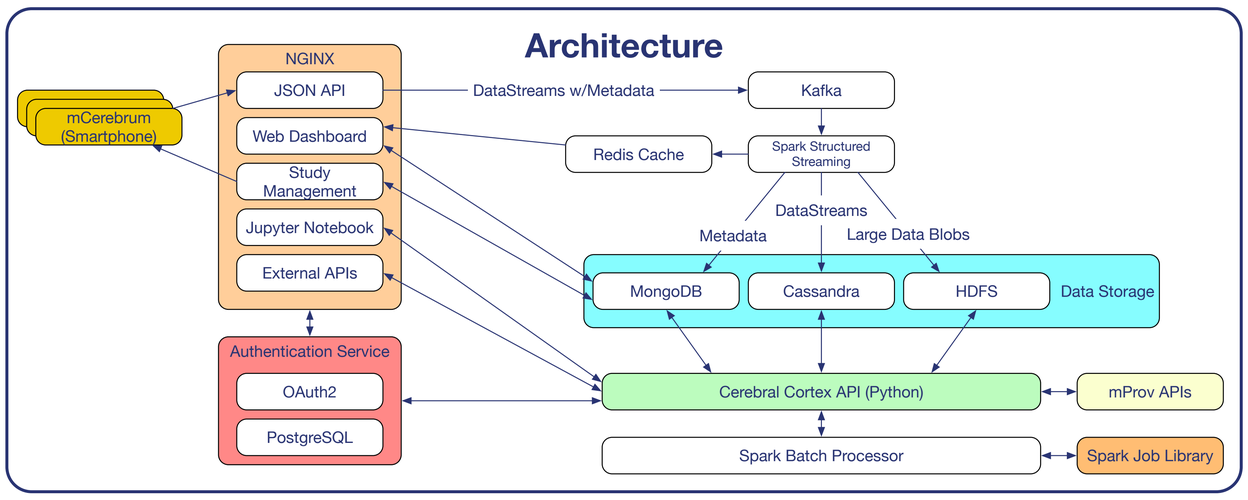
- **Goal:** Annual validation case studies will
  - Validate the effectiveness of our languages, APIs, and runtime systems
  - Evaluate how effectively "replay" experiments can be conducted using logged mobile device data annotated with provenance
  - Provide initial datasets and benchmarks for CISE researchers
- Validation studies to evaluate the adequacy of mProv in capturing the provenance reflecting the variations
  - In data source, sensor type, and attachment for stress assessment
  - In sensor location (dominant vs. non-dominant hand) on detection of hand to mouth activities such as eating and smoking
- Annual open data challenge with test problems starting in Year 3

mProv Impact of Science and Society

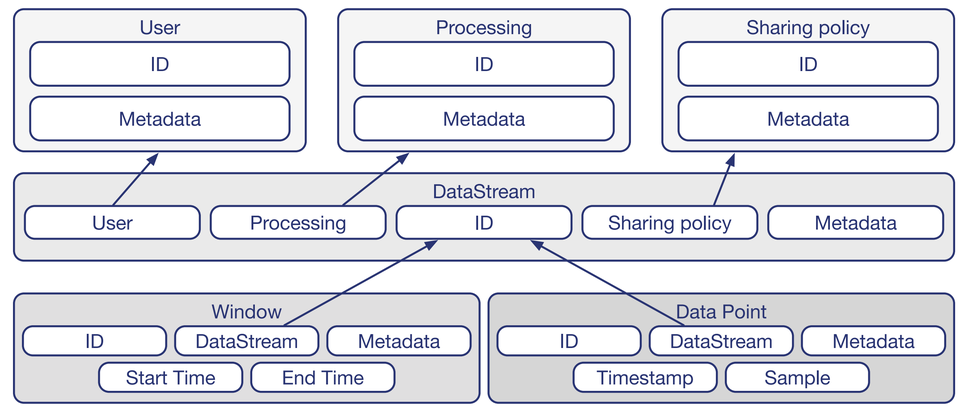
- MD2K platform enables collection and analysis of mobile sensor data in the natural environment
- mProv provides a companion cyberinfrastructure to annotate data streams with source, placement, semantics, provenance, validation, and privacy to facilitate data sharing
- MD2K and mProv together deliver an end-to-end cyberinfrastructure to catalyze research with mobile sensors
- Accelerating research with mobile sensor data brings tangible benefits to the society, e.g., improving health & wellness

Progress since launch (September 2016)

Architecture

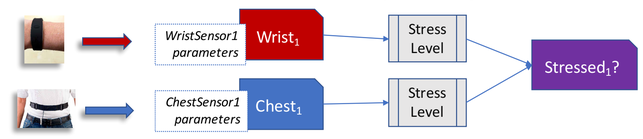


Data and Metadata Representations

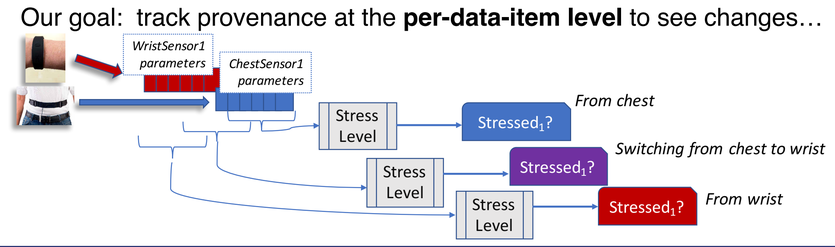


Fine-Grained Provenance Over Data Stream Computations

Provenance is often captured at the "streams & modules" level.
But if 2 sensors' output is "merged" together, we can't differentiate…



Our goal: track provenance at the **per-data-item level** to see changes…



Mechanisms for Automatic Provenance Generation

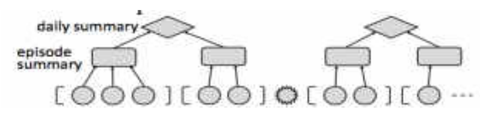
Streams from sensors are being processed via **pipelines of modules**

Can we **automatically record** provenance for data computed by modules?

➢Idea: Compose modules from smaller **primitive operations**

e.g., *map + reduce, filter chains, SparkSQL relational algebra*

1. Operations from our library: **automatically track provenance derivations**
2. User-defined operations: **all inputs linked to output** or **programmer calls an API**



We can compose hourly "summary" streams over individual stress event episodes, which are at the bottom level

- **StreamQRE** library defined in [*Mamouras et al, submitted for publication*] and being integrated into mCerebrum

Privacy Aware Data Sharing

- **Context-sensitive Control over Sharing**
  - Easy-to-manage fine-grained user control over sensor data collection process while balancing privacy and data quality
  - Platform mechanisms to prevent leakage of sensor data to other applications
- **Metadata for Sharing and Privacy**
  - Schema to carry sharing policy ("Terms of Service") with the sensor data
  - Compute sharing policy for outgoing sensor data by fusing sharing policies of incoming data as well as sharing policy of processing module
  - Cryptographically bind a recipient to the sharing policy
  - Changes in sharing policy triggered by new knowledge or change in context

Standards

- Builds upon Open mHealth data standard, which includes
  - 91 schemas for common mHealth data elements
  - standardized representation of time and units
  - annotation to standard medical vocabularies
- mProv will revise and extend Open mHealth's current basic metadata schema

Website: http://mprov.md2k.org/

mProv

Giving mHealth researchers the means to manage metadata for streaming sensor data

University of Memphis • University of Pennsylvania • University of California - San Francisco • University of California, Los Angeles • Open mHealth

mProv is supported by a National Science Foundation grant (ACI-1640813) under the CIF21 DIBBs:EI program | mProv@md2k.org | mprov.md2k.org