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

Award: ACI-1640813; PI: Santosh Kumar; Co-PIs: Zachary Ives, Ida Sim, Mani Srivastava

New wearable and implantable sensors can continuously capture human health, behavior, and environmental risk factors. The need for computational models of human health and behavior, combined with the uncertainty and variable quality of sensor data collected in the mobile setting, motivate our project. We address the need to *provide access* to live and re-playable datasets and processing pipelines for such data; enforce *regulatory obligations* in using mobile sensor data collected from humans; and perform *metadata capture and reasoning* to enable applications and developers to assess the provenance, quality, and integrity of data and inferences made from it. We build upon devices, ongoing user studies, and tools from the NIH-funded MD2K Center of Excellence, which records, detects, and predicts user behavior based on wearable sensor data, and upon the Open mHealth data standard. The mProv team currently targets challenges revolving around *integrating metadata capture, annotation, and derivation* within its current data stream analytics pipeline.

In the MD2K sensor-triggered stress intervention study, data is collected from wearable sensors worn on chest and wrists: each device has an accelerometer, and the chest belt has an electrocardiogram (ECG) and respiration. Sensor data is processed to infer whether sensors are properly attached, and whether user behavior indicates stress. A data analysis and diagnosis pipeline in mCerebrum aggregates and streams the data back to a cloud-based data processing platform, Cerebral Cortex, built over open-source storage and data processing tools (MongoDB, Cassandra, Spark). A "pipeline" of code modules operates over streaming sensor data, infers stress levels based on the properties of the data in stream windows, and triggers interventions. The current application does not address the issue that different active sensors, sensor combinations, and associated models have different fidelity. As an initial mProv use case, we address the need to quantify output quality based on input data: (1) As raw sensor data is acquired, we *annotate it with provenance metadata* fields as defined by the Open mHealth project, and any sensor parameters. (2) As data streams are processed by each individual Spark module, we automatically *annotate each output data sample with derived provenance information*, connecting the result back to the source samples from which it was derived, as well as the computational steps involved in creating it. (3) Finally, we take the output predictions of the analysis pipeline, and *use data provenance to determine a confidence level* in the data – based on whether the original sensor data was acquired by the ECG and respiration sensors worn on chest, the wrist-worn devices, or some combination thereof.

Currently, we are augmenting the Apache Spark-based mCerebrum platform to associate provenance metadata with the sensors, "attaching" user and device metadata from a remote database. This requires us to carefully define a general set of metadata fields that are appropriate for the sensor, environment, and patient. We will also develop a set of metadata annotations for data sharing and privacy policies that will be carried with a sensor data stream, enriched as the stream is transformed and aggregated by entities along the end-to-end path, and enforced by mProv system services prior to transfer across trust domains. An implementation challenge is how to optimize network traffic and data processing throughput.

A second research focus is to *propagate* provenance annotations from input data, through algorithmic modules, to the individual outputs of these modules. While this might be achieved by manually logging the provenance, a better approach is to let the runtime system *automatically* capture relationships between each output data item, and the computation and source data items from which it was created. Leveraging "fine-grained" provenance techniques developed for database query operators, we will supply a set of basic data manipulation primitives that can be composed to perform data processing, event detection, filtering, data splitting, and data merging. As necessary, our framework will also incorporate user-defined operations – at some cost to the ability to capture fine granularity of provenance metadata. We plan to rewrite MD2K modules to take advantage of these primitives, and to validate (1) that they can capture the common stream processing operations required in the setting, (2) they are natural for the programmer to use, (3) they preserve a full re-playable record of the data relationships.