

NSF CIF21 DIBBs: mProv: Provenance-Based Data Analytics Cyberinfrastructure for High-Frequency Mobile Sensor Data

PIs: Santosh Kumar, Zachary Ives, Mani Srivastava, Ida Sim

Team: John Frommeyer, Timothy Hnat, Nasir Ali, Ananda Tirtha, Sandeep Singh, Nan Zheng, Phillip Hilliard

Mobile Sensor Data Needs *Provenance* Metadata

Need for shared repositories of mobile sensor data – but data must have *provenance* to enable reproducibility, integration, and comparative analysis



Mobile sensor data introduces challenges due to volume and heterogeneity:

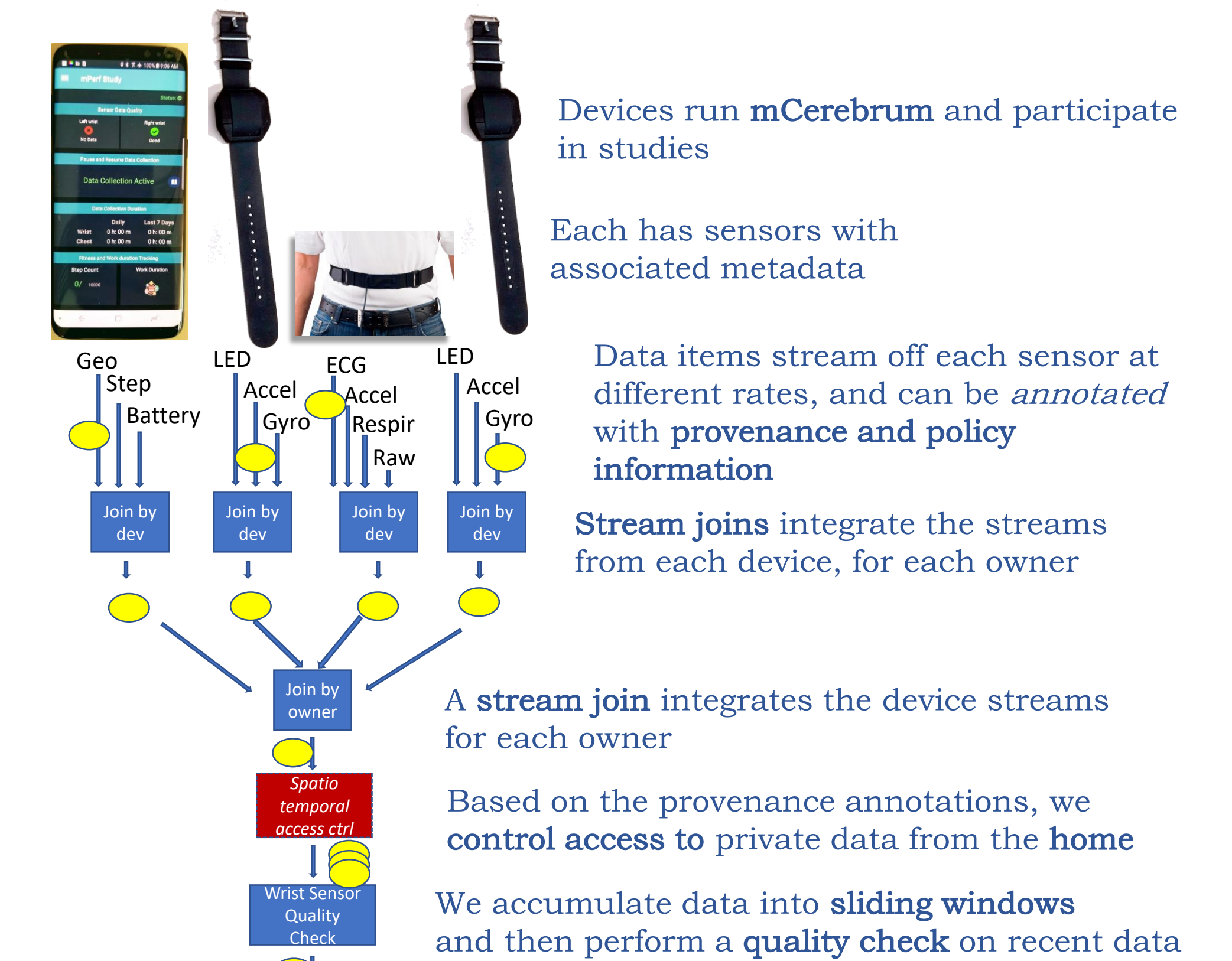
Velocity	Variety	Volume	Variability	Veracity	Validation
100s of samples / second / sensor	10s of sensors per device	GB/day per person	Variations in attachment, placement, model	Multiple biomarkers from same sensor	Sources of valuation of specific biomarkers

The mProv Project: Enabling Provenance for Mobile Sensor Data

Standards	Capture/Storage	Privacy	Query/Replay
Create open metadata standards for provenance metadata for mobile sensor data	Define tools and services for automatically capturing provenance metadata for streaming source & derived data	Provide access control policy and enforcement mechanisms for streaming data, using metadata annotations	Develop capabilities for querying streaming data's provenance and for replaying streams from archival

mProv Use Case:

Annotations & Provenance for a Marker Stream



Marker Data Streams for Mobile Health in MD2K

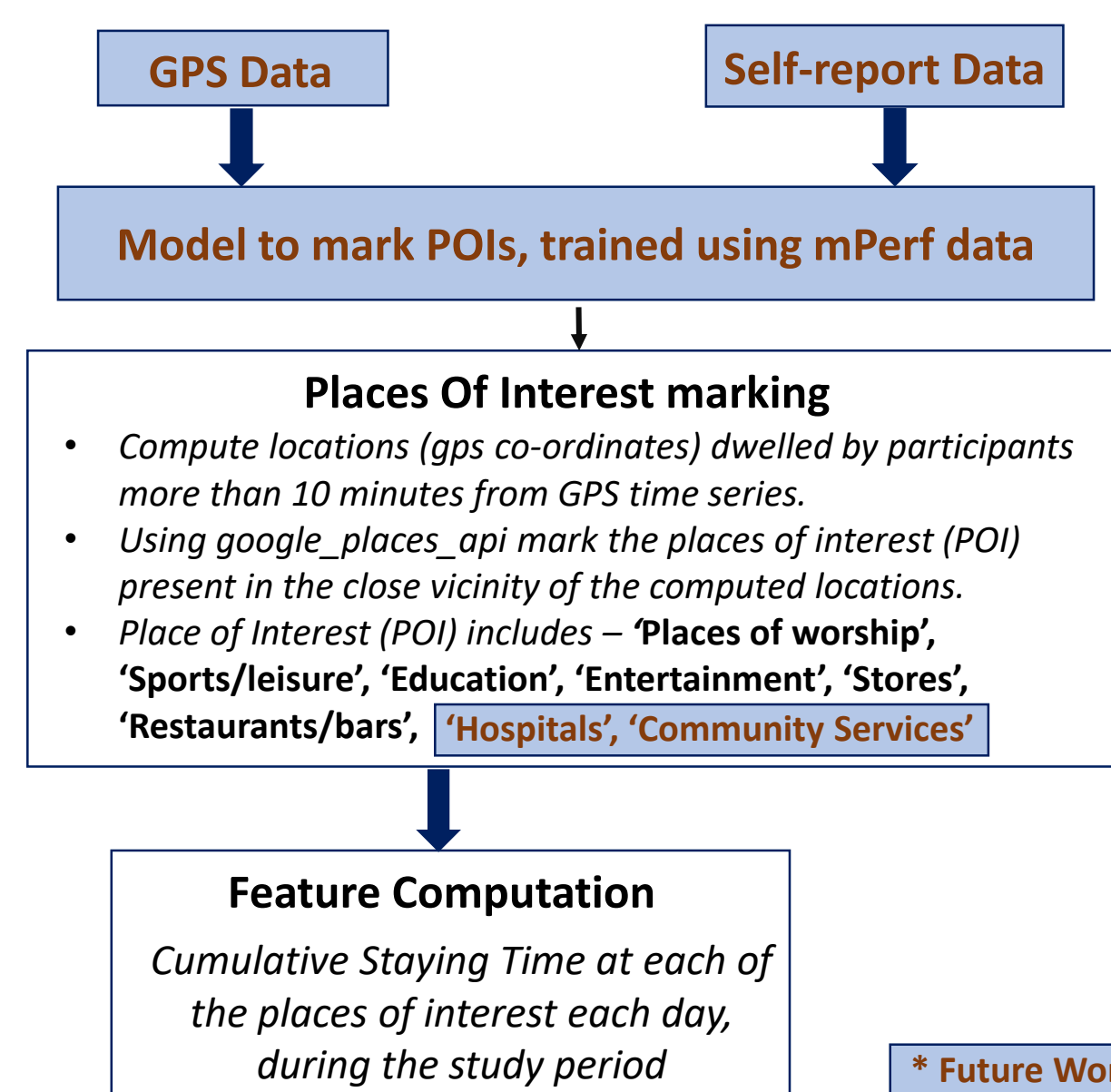


Location Data Streams

$ST_p, ET_p, 'Home'$
 $ST_p, ET_p, 'Work'$
 $ST_p, ET_p, 'Other'$
...

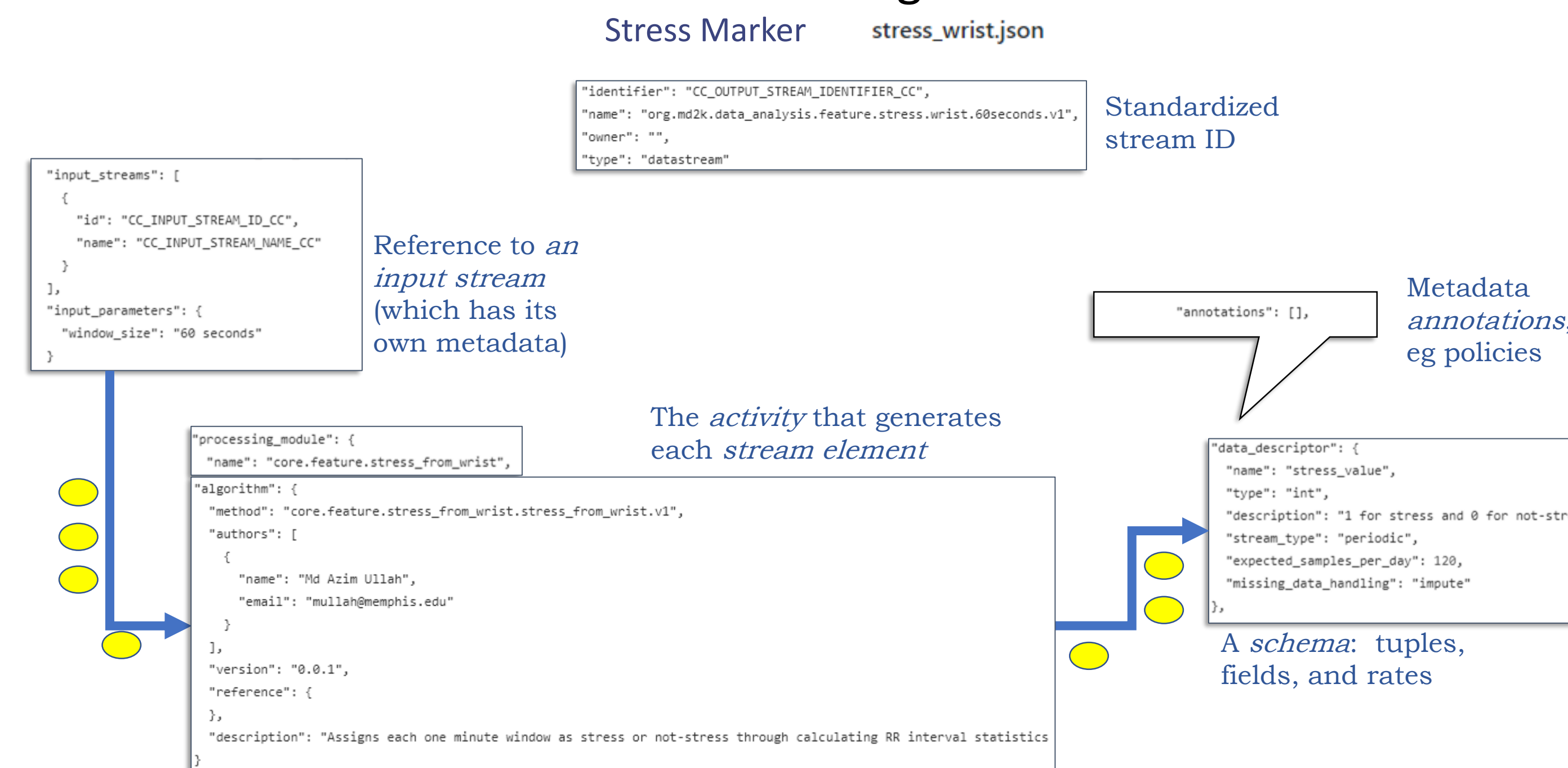
Feature Computation

- Cumulative Staying Time at a given place
- Transition Frequency between two centroids
- The total distance covered
- The maximum distance between two location
- The standard deviation of the displacement
- The maximum distance from 'home', 'work'
- The number of different significant places visited
- The number of different places visited
- The radius of gyration
- The routine index



* Future Works

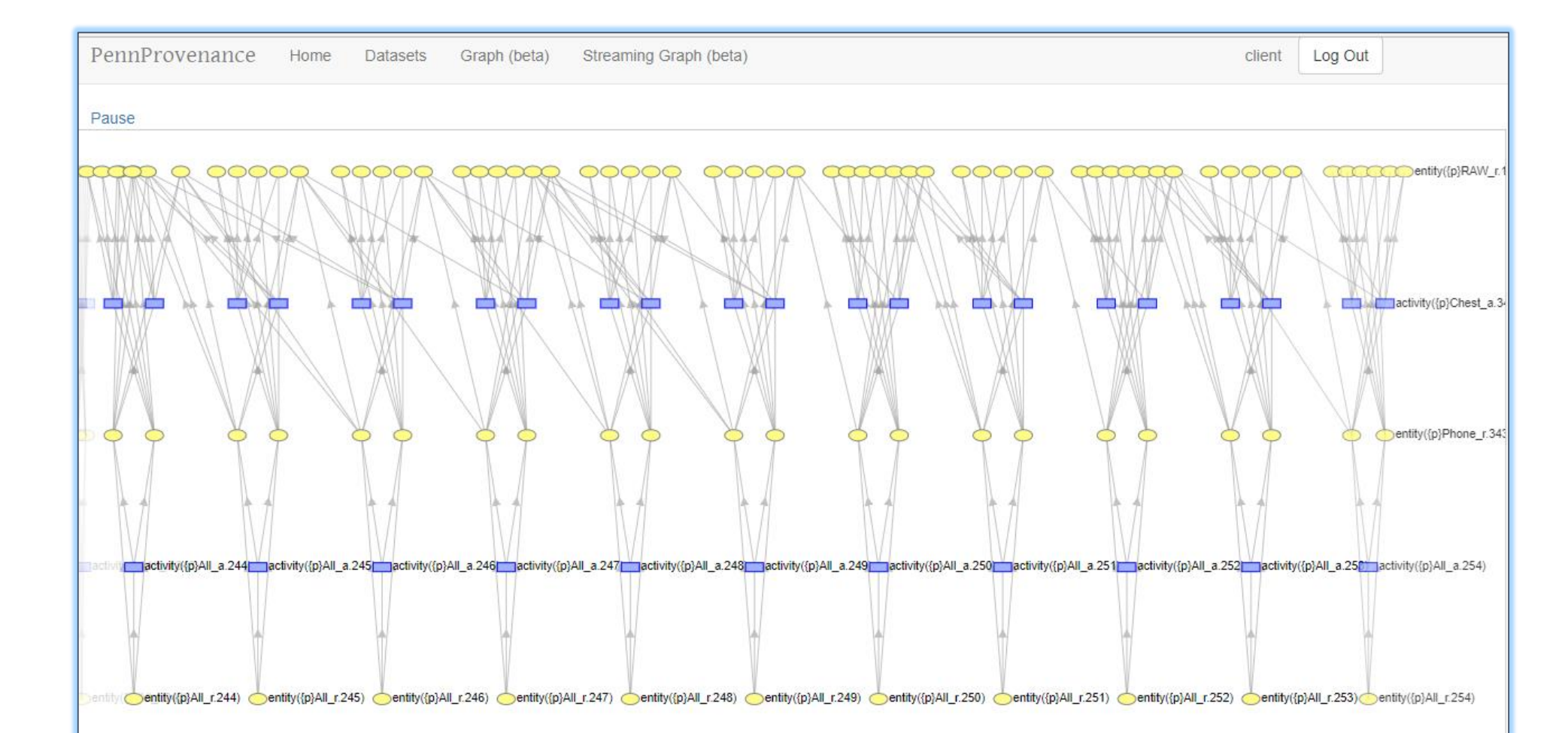
Metadata for Streaming Data in MD2K



mProv stores, streams, visualizes, & queries *provenance and annotations* for input & marker streams

mProv captures (in a stream and in storage) the output of the pipeline, but also the inputs and steps, and any annotations on them!

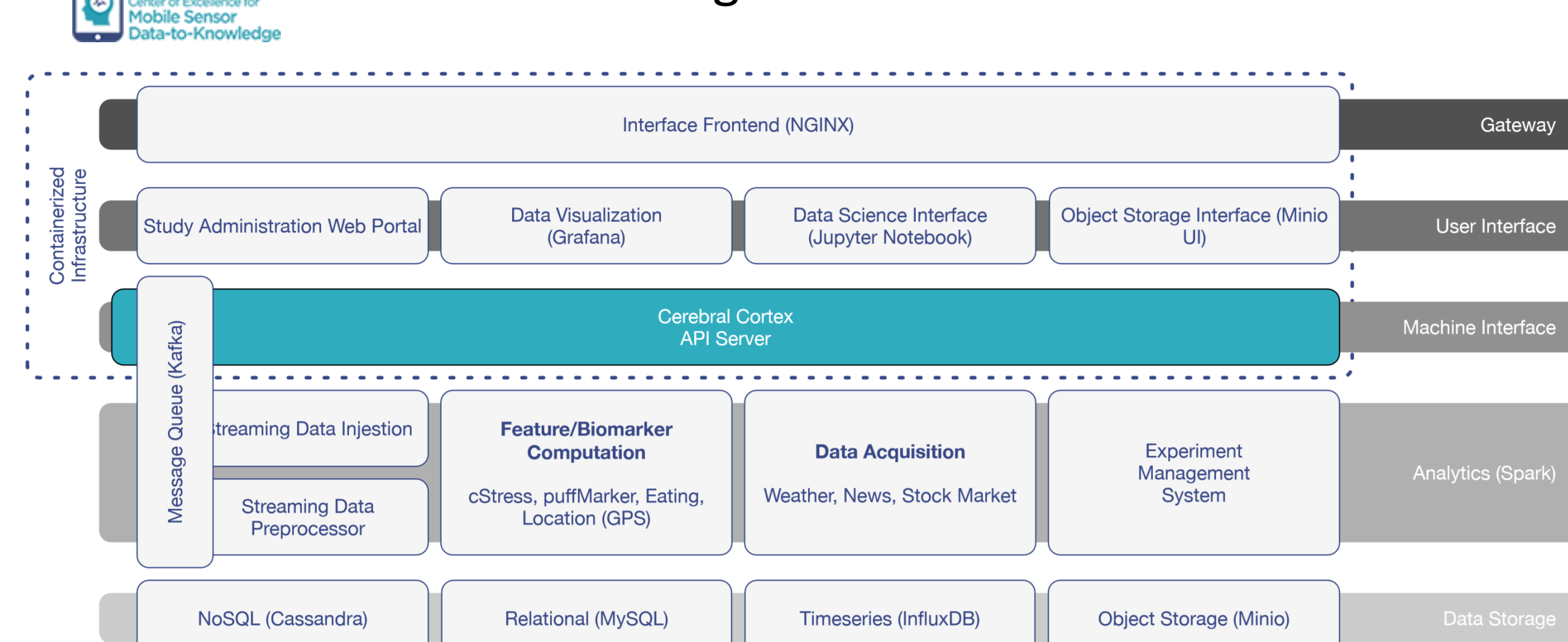
entity (data) collection ("window") activity (processing step)



For streaming data – we can query about its provenance and metadata to, e.g., filter or rank the result

For stored data – we can merge streams and replay precisely, differentiate among elements, and see if they were produced similarly

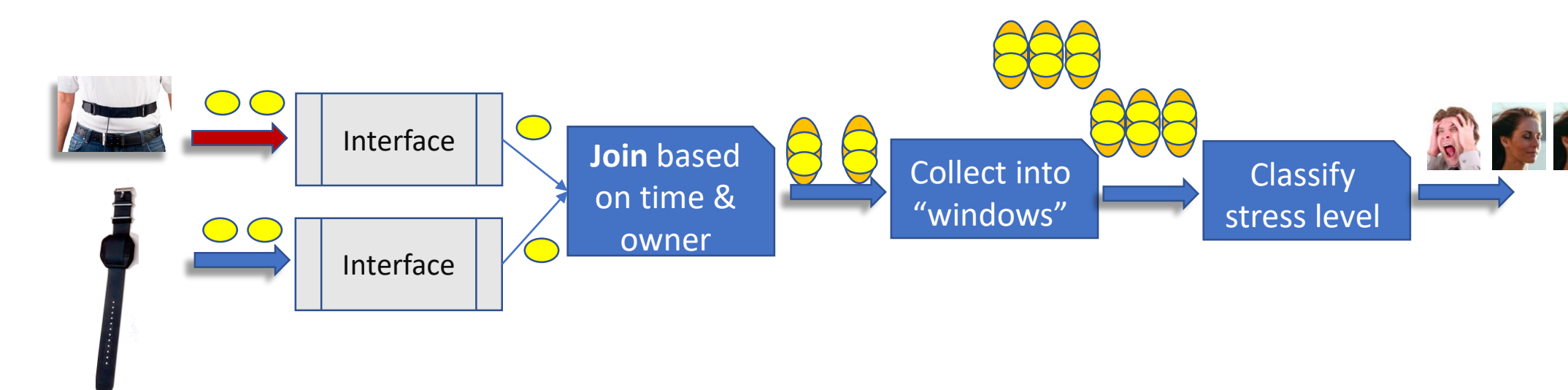
MD2K Mobile Sensor Big Data Platform: Cerebral Cortex



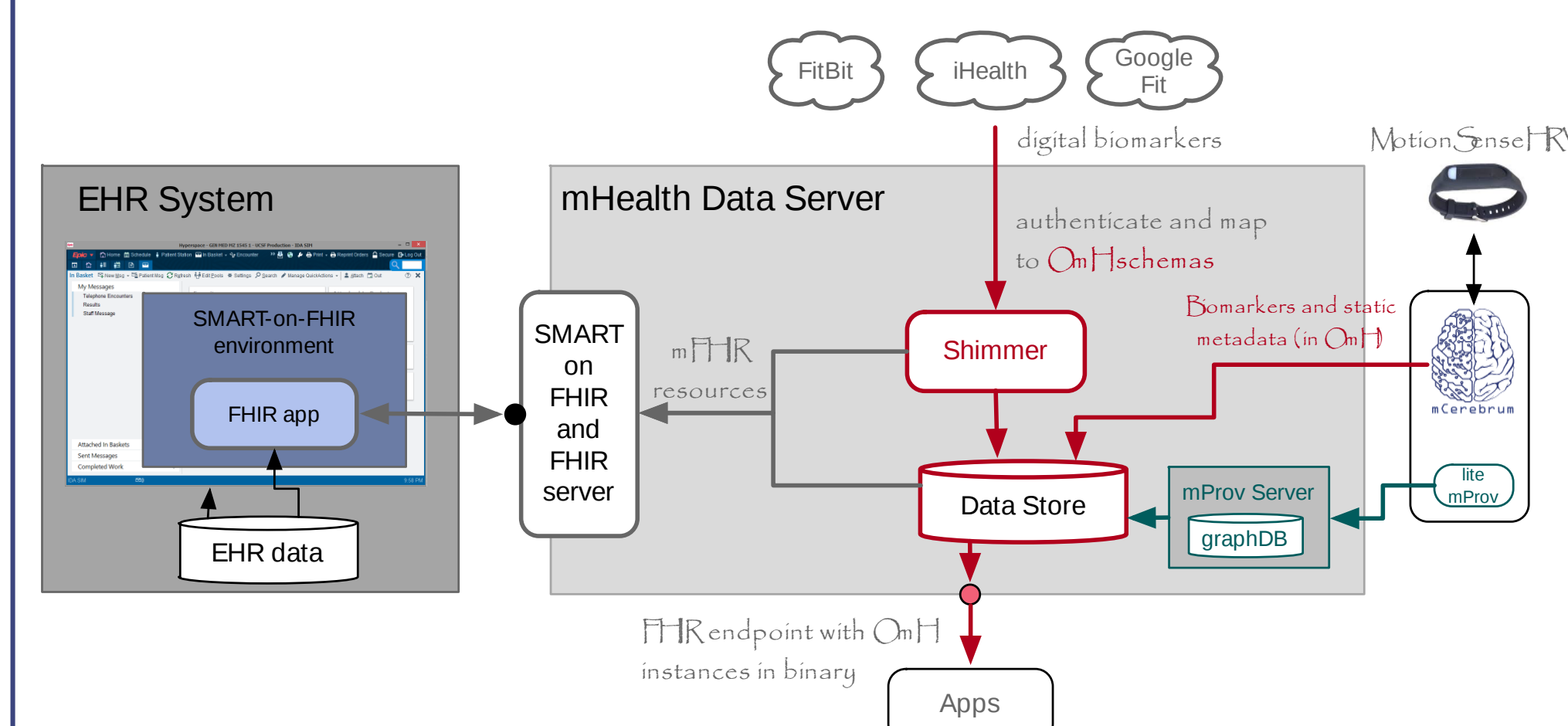
Stream Processing in Cerebral Cortex: Raw Streams → Marker Data Streams

Streams of data from sensors are being processed via **pipelines of modules** in Apache Spark on the cloud, with modules written in Python (Pandas + decorators)

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



Open Metadata Standards for Health Device Data

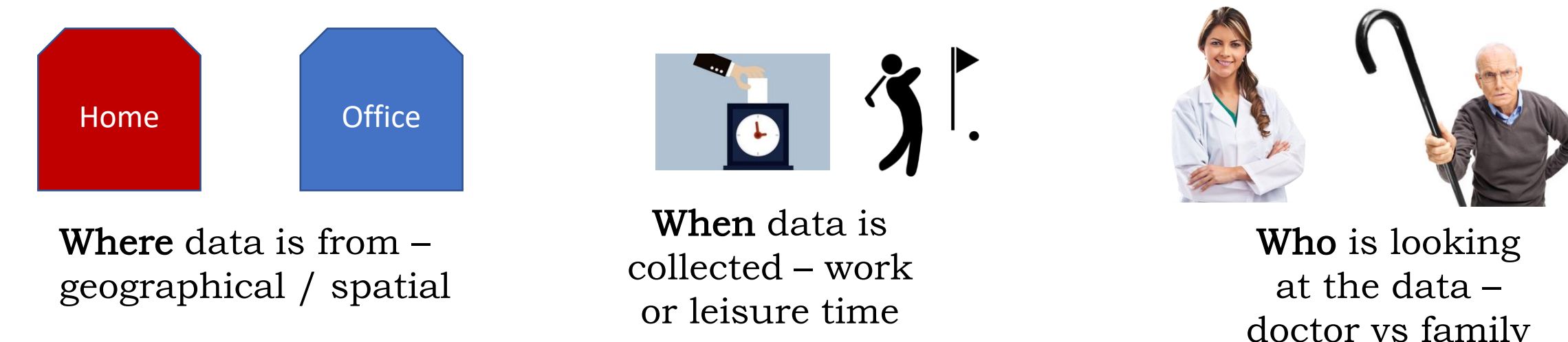


- Open mHealth** standardized schemas for 91 health data elements
- Shimmer** tool maps biomarkers to OmH schemas
- mFHIR Implementation** Guide standardizes OmH schemas in FHIR, for integration with EHR

Access Control Based on Data, Metadata, and Provenance

Data stream elements are associated with *policy* and *annotation* information and provenance

Access control policies restrict stream visibility, using metadata about:



mProv Provenance Platform

