Why Focus on Cities?
- 82% of the US population lives in urban areas.
- Cities provide essential infrastructure and services.
- Huge amount of real-time data is being generated by each sector.
- We need an End-to-End Knowledge Discovery Cyberinfrastructure for effective analysis and policy support.

Design Goals
- Real time interface
- Online data analytics
- Usable discovery mechanism
- Secured access control

Citadel Metadata/Data model
Graph-structured Metadata, RDF
- Numerous different hierarchical information.
  - Types of sensors: Temperature? Humidity? Traffic?
  - Target entity of sensors: Outside air? Indoor? Car?
  - License of the data: Open-source? Closed-source?
- Many components are interrelated. It can be expressed in RDF easily. (nodes and directed edges)

Spatio-Temporal Data
- Both time and space indices are important
  - How fast is this bus moving? How long/large is this protest?
- Data Model:
  - A data point associated with spatio-temporal indices and values.
  - Spatial index can be a point, polygon, line.
  - Values are numbers currently.
  - A cell is (Data UUID, Timestamp, GeometryType, List of Lat/Lng, Value)
- Geomesa: GeoHash and timestamp for index + load balancing

Architecture
- Spatio-Temporal DB: Geomesa on top of HBase
- Graph-structured Metadata DB: Virtuoso with SPARQL.
- REST Framework: Vert.x for Microservice Framework.
- Message Bus across different Components: Kafka
- Analytics: Spark

Provenance
- Users need provenance of analytics.
  - How is this analytics generated?
  - What data does this analytics refer?
  - How much credible is this result?
- Proposed Provenance Model (using PROV-O and VOAG):
  - Enables feature inheritance such as permission levels.

Access Control
- Many datasets are confidential and are shared with authorized entity.
- Traditional Access control mechanism like Role-Based Access Control and Attribute-Based Access Control exists but are not exploited in city data hub scenario.
- General access control framework over complex data structure of heterogeneous datasets is complicated.
- In order to allow collaboration for confidential datasets, we propose a framework where access control is based on;
  - Content of data streams.
  - Attributes of data streams with focus on space, time and values.
  - Allow controlling resolution of data stream.
  - Expressing data sharing policies and exploring policy enforcement.

Preliminary Evaluation
Metadata Management Performance
Current metadata types per point: name, point type, unit
Configuration: In Azure cloud, a host machine (8 cores, 32GB), 8 clients in parallel. Averaged over 10,000 requests.

Spatio-Temporal DB Performance
Current data types per point: latitude, longitude, timestamp, num value
Configuration: In Azure cloud, a host machine (8 cores, 32GB), single client averaged over 1000 requests.

Acknowledgements
This research is funded in part by the National Science Foundation under awards IIS-1636916, IIS-1638879, IIS-1636936, OAC-1640813, CI-1331615, and CSR-1526841, and by the King Abdullah University of Science and Technology under KAUST Sensor Initiative.