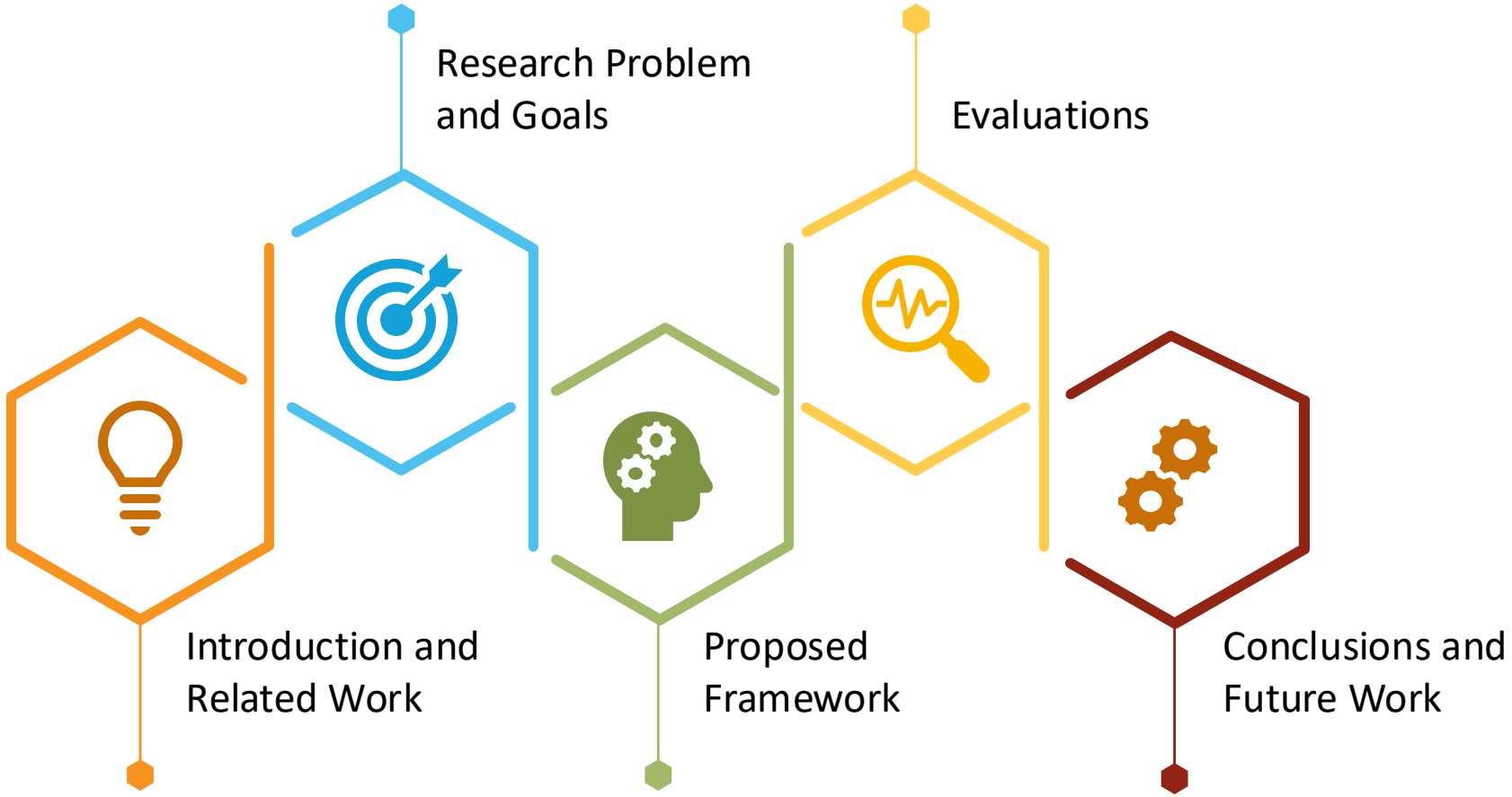
Data Cleaning and EDA Continued

- Review the Berkeley Call for Service dataset
  - There are redundant columns that are not needed
  - Can we check that the case numbers are unique?

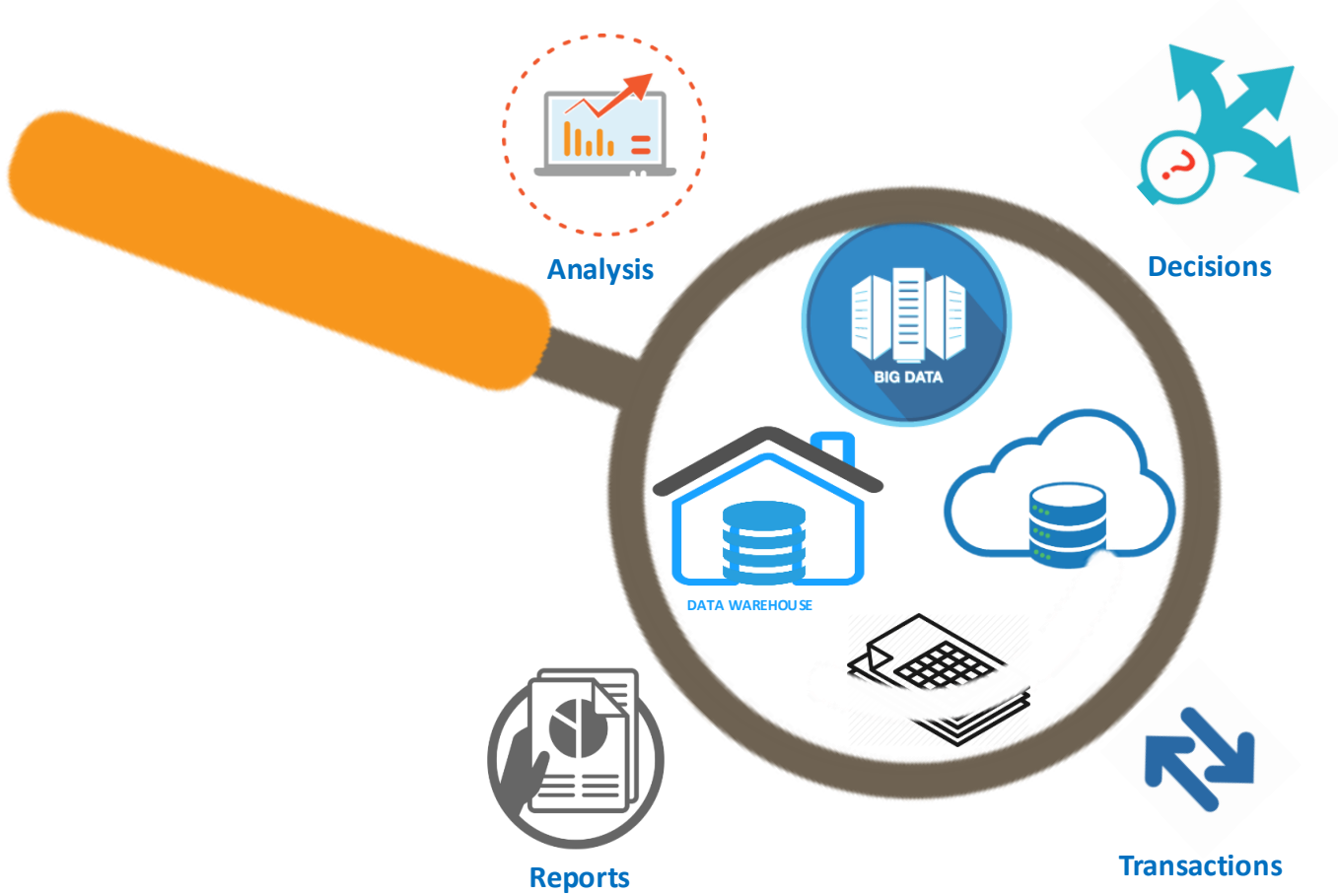
Anomaly Detection and Explanation in Big Data

Sean Kang

Outline


Research Problem and Goals
Evaluations
Introduction and Related Work
Proposed Framework
Conclusions and Future Work

Data Quality is Critical to Every Enterprise



Non-Sequence vs. Sequence Data

A *Non-sequence* dataset D is a set of d-dimensional records $D = \{R_0, \dots, R_{n-1}\}$

A breast cancer dataset that contains values of *tumor size* for different patients, Or
A set of purchase orders from an online e-commerce system

A *Sequence* (time series) dataset T is a sequence of d-dimensional records $T =< R_0, \dots, R_{n-1} >$

A climate dataset that contains *wind speed*, *snow depth*, and *temperature* over time, Or
A dataset set of movement and position data for military or commercial aircrafts, over time

where

- $R_i = (a_i^0, \dots, a_i^{d-1})$ is a record $(0 \leq i \leq n-1)$
- a_i^j is j^{th} attribute of i^{th} record

$d = 1$ for univariate time series
$d > 1$ for multivariate time series

Data Quality Tests

Validate data in a data store to detect violations of constraints that are imposed by application domain experts and data model

- Constraints over single attributes

wind_speed must be positive, vehicle speed must be positive

- Constraints over multiple attributes

If *vehicle_speed* is greater than zero, then altitude must be non-zero , if it is an aircraft

- Constraints over single attributes in multiple records

patient_weight growth rate over time must be positive and in the range [4, 22] lb for every infant, Or
Fuel amount must be zero or higher over time, for a jet plane

- Constraints over multiple attributes in multiple records

Mean value of *daily_delivered* electricity to *premise_classification*="Residential" must be in the range [0-20] kWh

Limitations of Existing Approaches

Pre-specified constraints

- Domain-independent approaches
  - Limited to trivial constraints
- Domain-specific approaches
  - Specified set of constraints is incomplete

Discovered constraints

- AI-based and statistical-based approaches (anomaly detection):
  - Lack of explanation – one of the limitation of such system
  - Typically require labeled data (supervised ML approach for training datasets)
    - Very hard to train with large data sets
  - Prone to generating false alarms (especially with unsupervised approach)