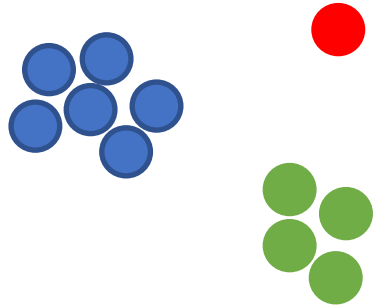
Outlier Analysis

Data Mining:
Data Mining Methods
with **Dr. Qin Lv**

Master of Science in Data Science
UNIVERSITY OF COLORADO **BOULDER**

Learning objective: Apply techniques for outlier analysis and explain how they work. Evaluate and compare methods.

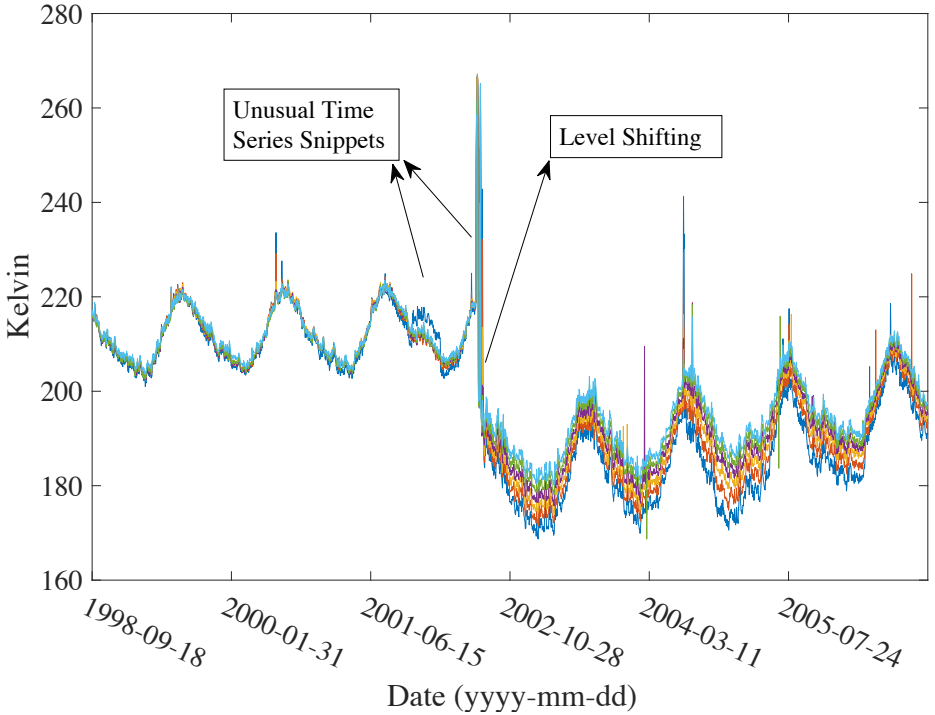
Outlier/Anomaly



- General/normal patterns
  - Frequent patterns, classification, clustering
- Abnormal: differs from the norm
- Outlier analysis/anomaly detection
  - “Outliers are errors and should be removed?”
  - Outliers may be significant events, frauds, …

Types of Outliers/Anomalies

- **Global**
  - Obj differs from the rest
- **Contextual**
  - Obj differs in a context
- **Collective**
  - Group of objs differ



Outlier/Anomaly Examples

➢ Environmental

- Temperature, commute time, wind power generation

➢ Social

- #friends, #posts, #likes, time series, network structure

➢ Financial

- Stock price, credit card transactions, spatial/temporal

Anomaly Detection: Challenges

➢ Normal vs. abnormal

- No clear definition, vary by application, noisy data

➢ Efficiency

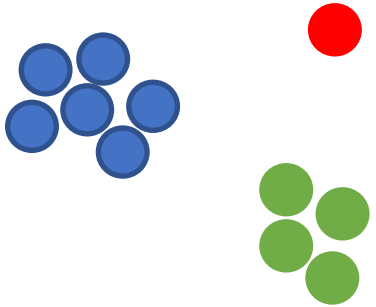
- Latency, scalability, adaptability

➢ Interpretability

- How to interpretate and explain the results?

Anomaly Detection: Methods

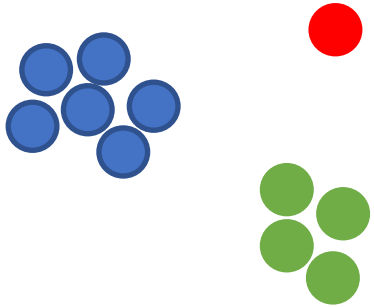
- ➢ Groundtruth label
  - Supervised, unsupervised, semi-supervised
- ➢ Assumption of normal vs. abnormal
  - Statistical model
  - Majority vs. minority
  - Proximity: distance, density



Classification-based Methods

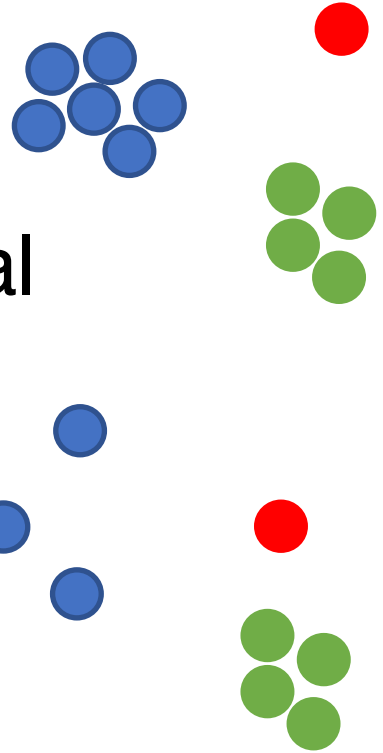
- ➢ Supervised learning
  - Normal and/or abnormal cases
  - More than one normal/abnormal classes
- ➢ Challenges
  - Class imbalance: e.g., faults are much less frequent
  - New patterns: e.g., new types of security attack

Clustering-based Methods



- ➢ Unsupervised learning
  - No pre-defined normal/abnormal cases
  - Majority vs. minority clusters
  - Multiple clusters for the normal/abnormal cases
- ➢ Generalizable to different applications
  - Clustering method, similarity measure

Proximity-based Methods



- Faraway from others => abnormal
- Distance-based
  - Absolute proximity, e.g., kNN
- Density-based
  - Relative proximity, e.g., ε-neighborhood

Semi-supervised Methods

- ➢ Partial labels: normal/abnormal cases
- ➢ Combination of clustering & classification

- Clustering
- Normal: large clusters and/or normal cases
- Abnormal: small clusters and/or abnormal cases
- Classification

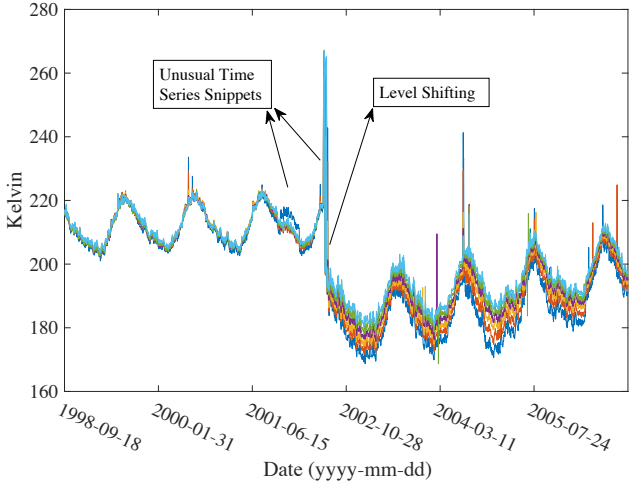
Contextual Anomaly

- ➢ Context features => behavior features
  - E.g., similar weather conditions => solar/wind power
- ➢ Identifying context
  - Frequent patterns, subspace, domain-specific
- ➢ Detecting anomaly within context
  - Transform to global outlier detection

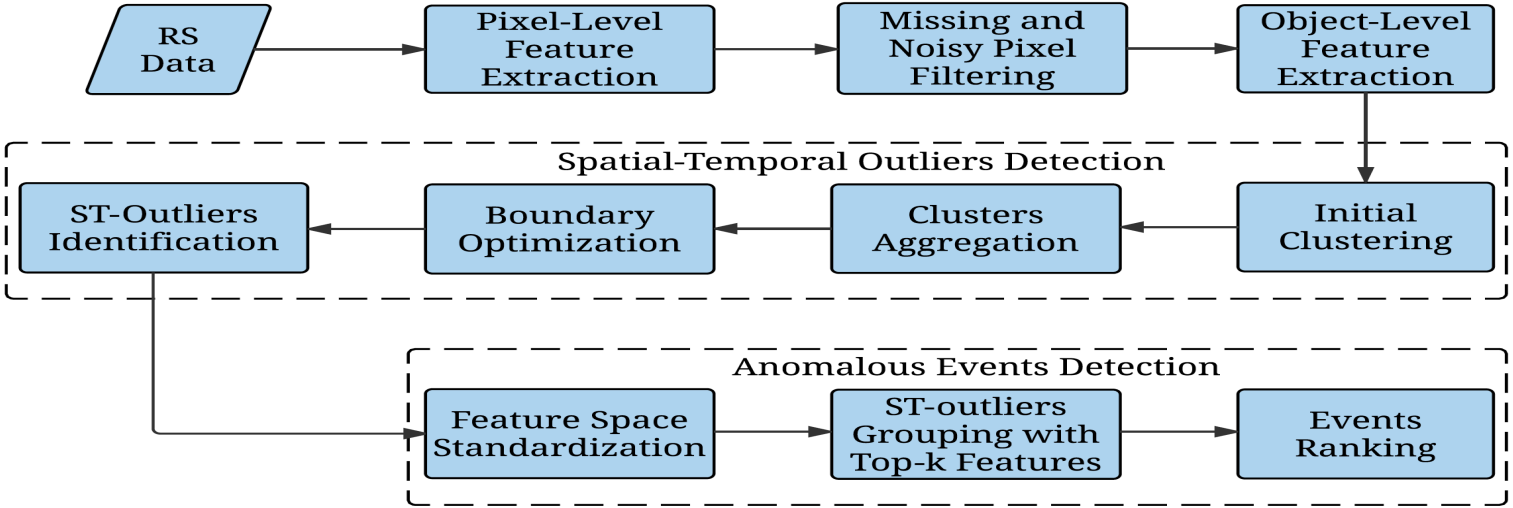
Collective Anomaly

- A group of objects deviate from the norm
  - E.g., Distributed Denial-of-Service (DDoS) attack
- Structural relationship among objects
  - E.g., purchases of specific items at certain time/location
- Super object: a group of related objects
  - Transform to global outlier detection

Example: Remote Sensing Data

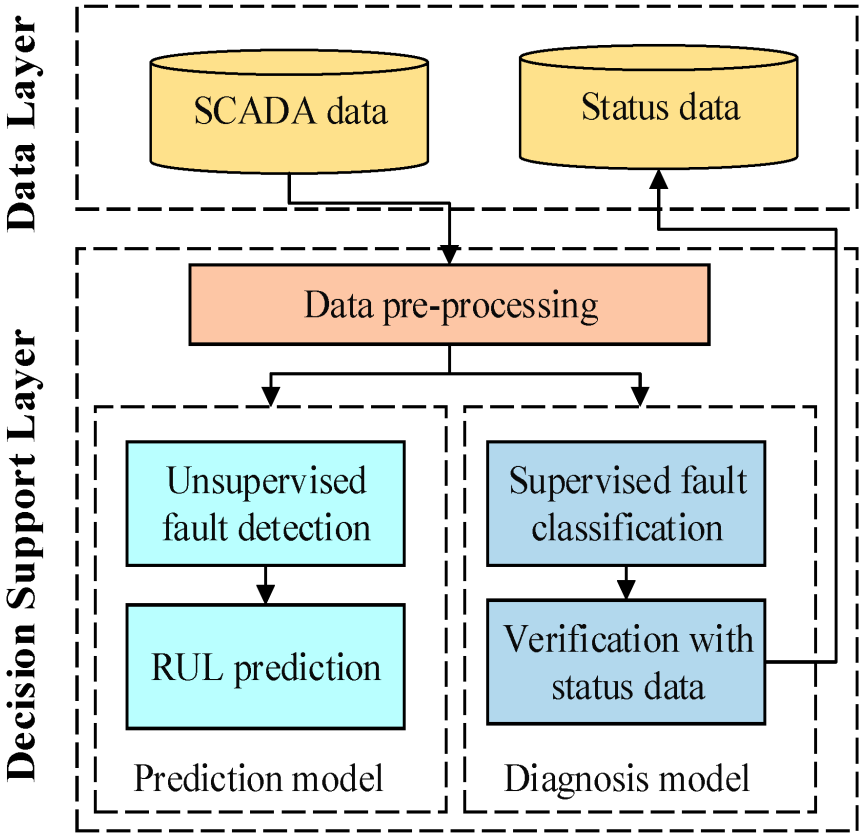


- Spatial-temporal anomaly

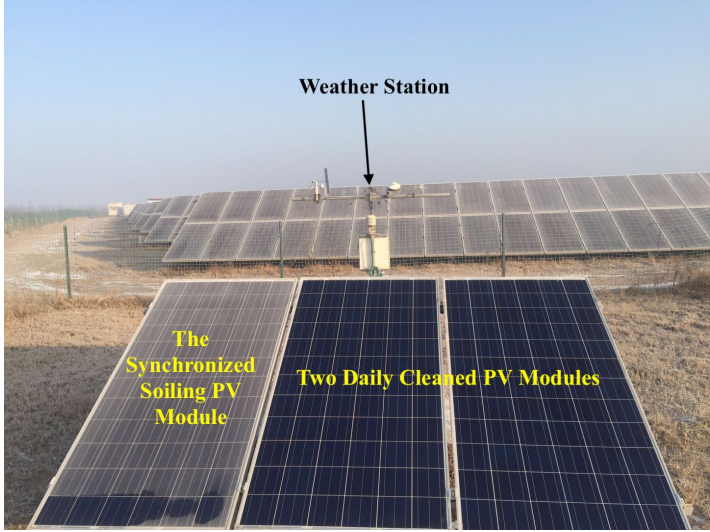


Example: Wind Farm

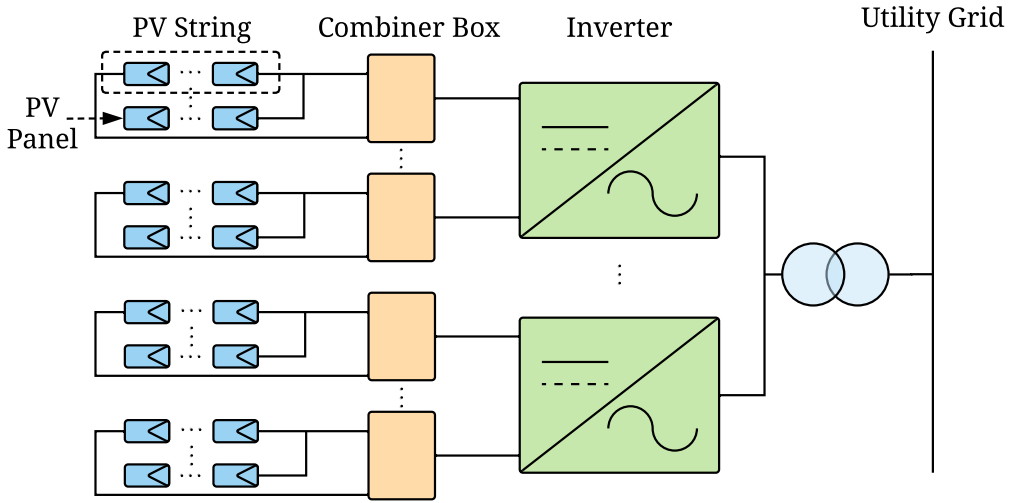
- Fault prediction
- Fault diagnosis



Example: Solar Farm (1)

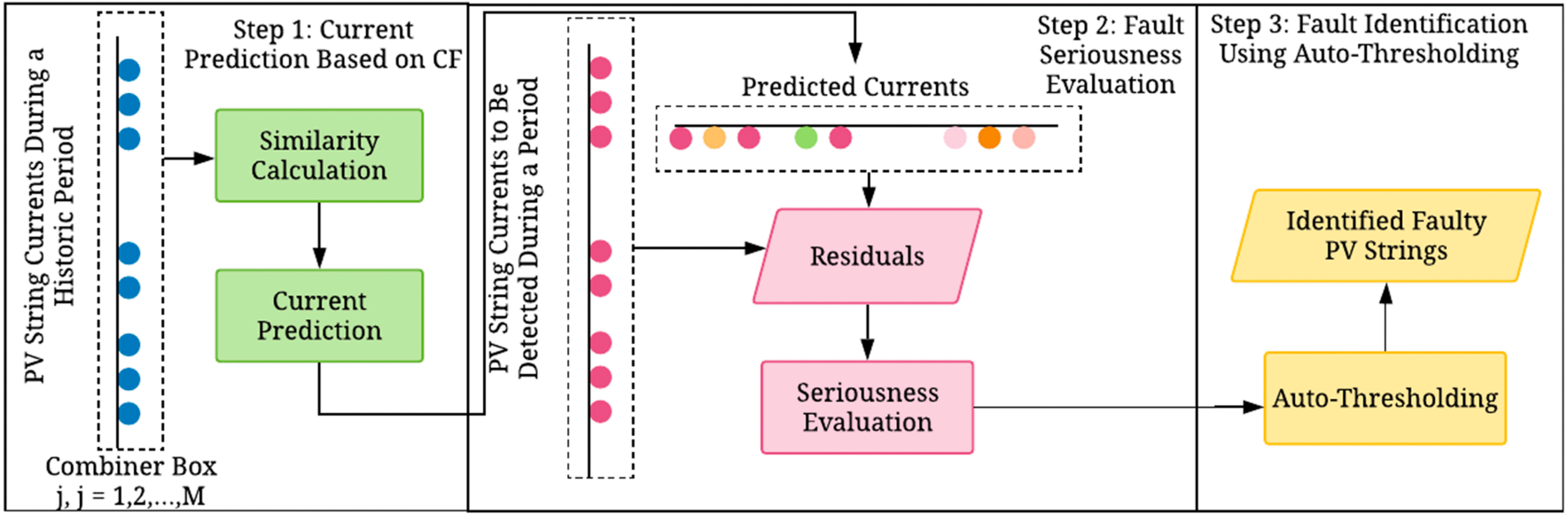


- Hierarchical anomaly



Example: Solar Farm (2)

- Collaborative fault detection



Summary: Outlier Analysis

➢ Types of outliers/anomalies

- Global, contextual, collective

➢ Methods

- Unsupervised, (semi-)supervised, context, structure

➢ Interpretation

- Errors or significant events