# Data Stream Mining and IT Sustainability

Philip S. Yu (psyu@cs.uic.edu)

Wexler Chair in Information Technology
University of Illinois at Chicago

# Data Stream Processing

- Data stream offers a new paradigm for
  - Managing large volume of constantly generated data
    - Potentially with multi-modalities
  - Supporting real-time response
- Most suited for monitoring or surveillance type allications

# Real-time Stream Applications

- Trade surveillance for security fraud and money laundering (also for detecting arbitrage opportunities)
- Bio-surveillance for terrorist attacks
- Sensor network for monitoring intelligent oil wells, manufacturing plants, RFID products, etc
- Network monitoring for intrusion detection
- Emergency room patient monitoring
- Web related applications
  - Click stream mining for real-time personalized recommendations
  - Text stream mining for topic detection
- System management
  - Power management for IT sustainability

# Stream Mining

- Challenges
  - Real-time: One pass
  - Resource constraints
    - Limited memory and processing power
  - Evolving stream characteristics
    - Temporal locality
    - New patterns vs outliers/anomalies
  - Noisy data
- Mining algorithms developed
  - Clustering
  - Classification
  - Frequent pattern
  - Graph/Link/Relation

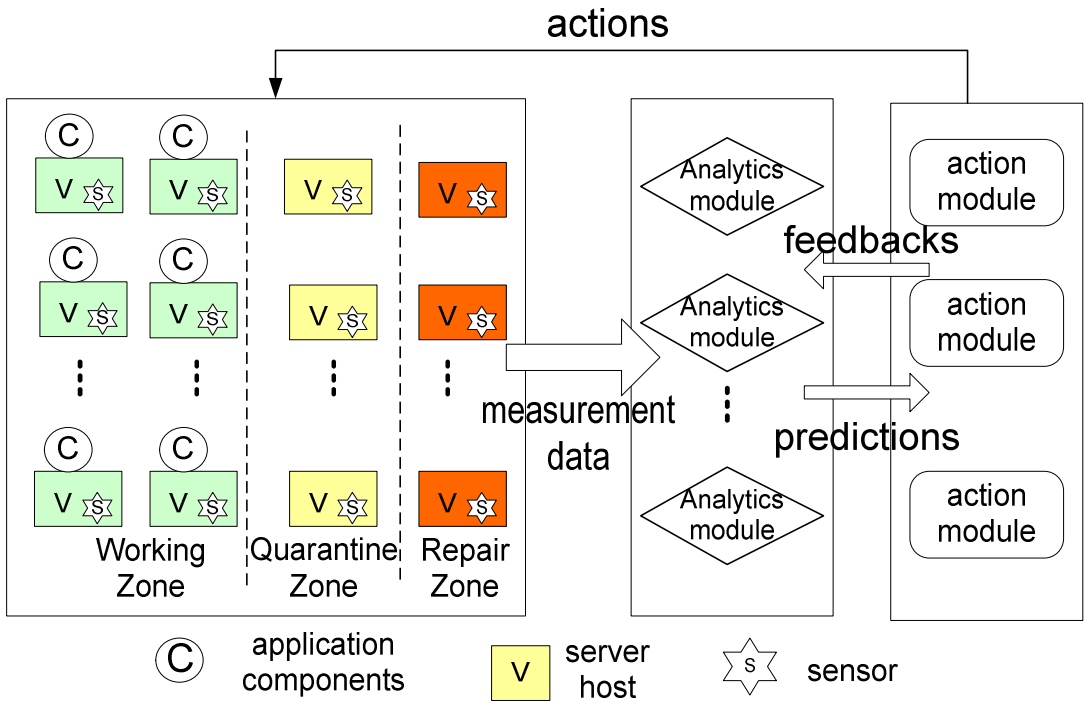
# Predictive v.s. Reactive Failure Management

- Reactive failure management
  - Action after failure
- Predictive failure management
  - Send out failure warnings in advance
    - Memory space on host B will be used up after 30 seconds with probability 98%
  - Recovery actions before failure
    - Order replacement hardware in time
    - Migrate critical application components out of problematic hosts
  - Reduce the impact of system faults

"Toward Learning-based Failure Management for Distributed Stream Processing Systems", ICDCS 2008.

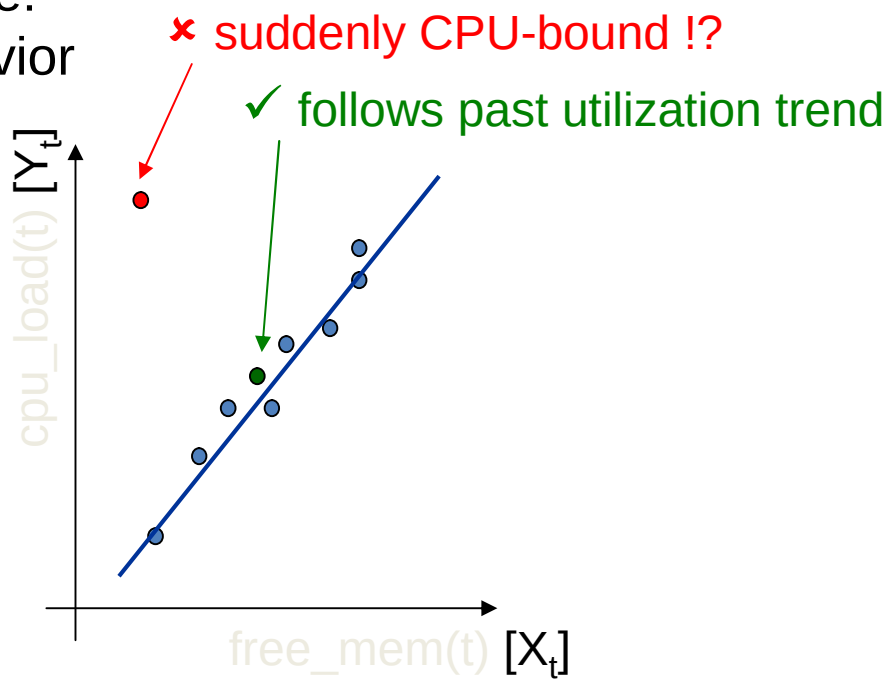
# Framework

- **A predictive failure management framework**
  - Utilize system analytic model for failure prediction
  - Perform proper fault handling based on predictions



# SAAO System Analytics Intuition – Step 1

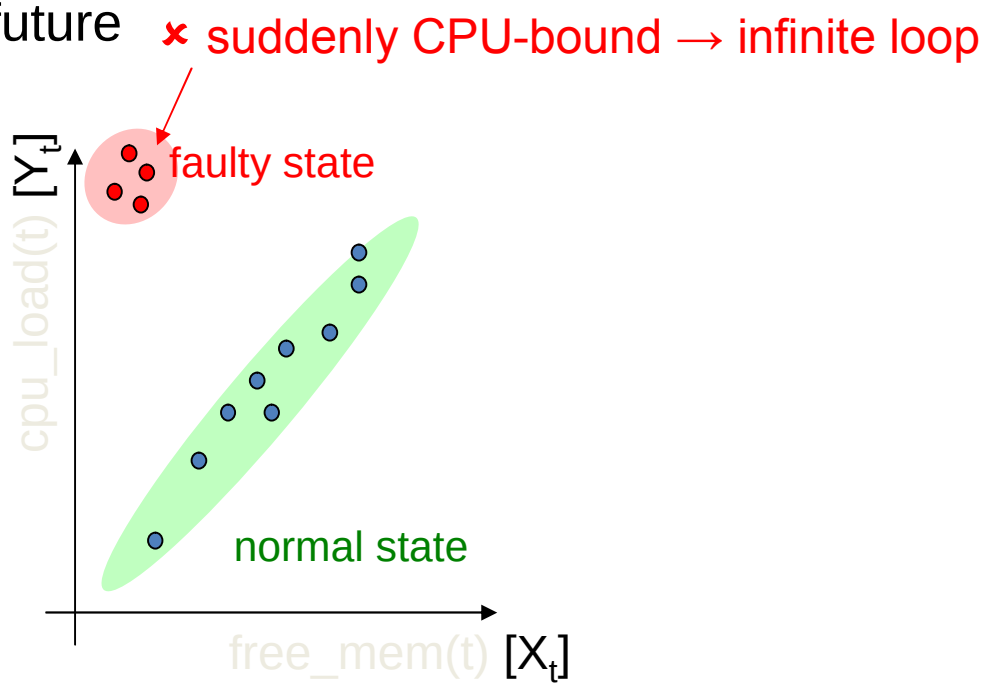
Potential unknown failure:
previously unseen behavior




Outlier detection
“abnormal vs. normal”

# SAAO System Analytics Intuition – Step 2a

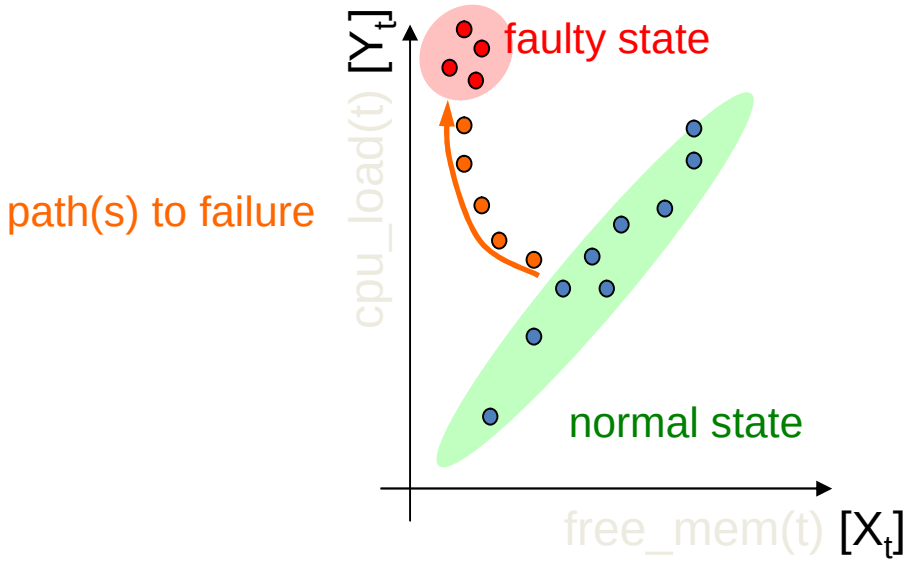
Failure identified:
Characterizing failure for future automatic identification




Classification
“fault identification”

# SAAO System Analytics Intuition – Step 2b

Warning lead-time:
Characterizing state
*right before* failure




Classification w/ time
"failure lead-time"

# Power Management

- A potential new stream processing/mining application on system management
  - Need to monitor power or temperature conditions
- Challenge
  - How to schedule tasks under power constraints or objectives

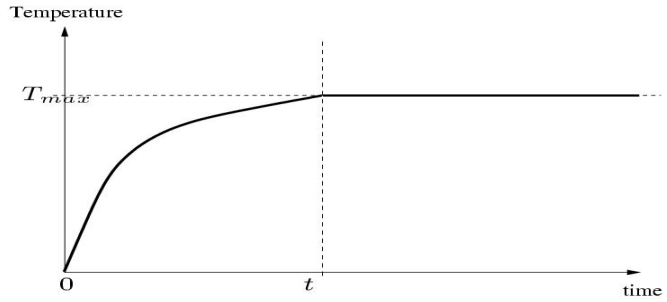
# Temperature Aware Task Scheduling

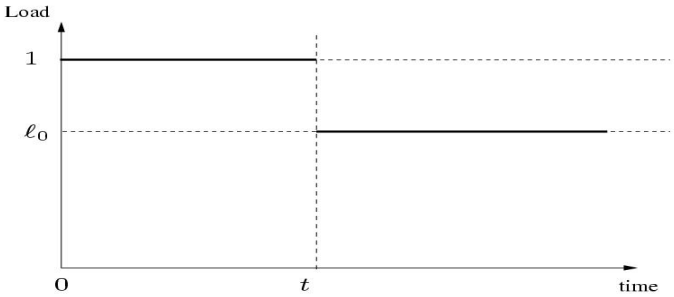
- Processor can change speed (dynamic voltage scaling)
- Scheduling policy needs to consider both
  - Placement of tasks
  - Processing speed at each node
- Question
  - Is constant policy optimum  from the power consumption view point?
  - What is the alternative?
    - Zig-zag?

"On Temperature-aware Scheduling for Single-processor Systems", IEEE Intl. Conf. on High Performance Computing, 2007.

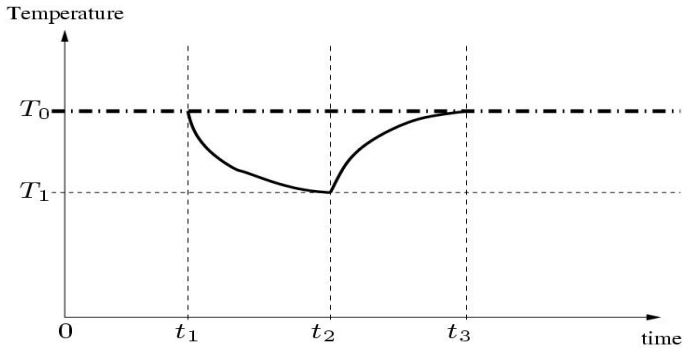
# Different Scheduling Policies

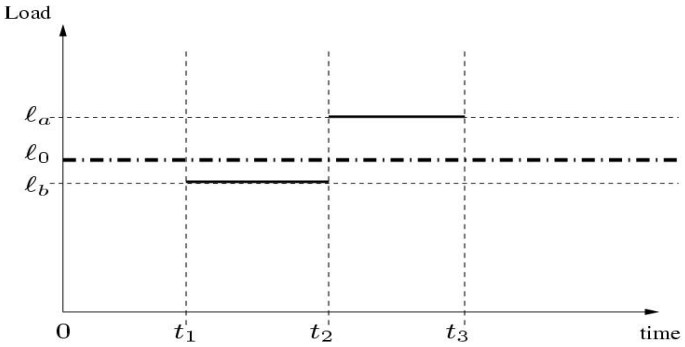
**Constant:** Simplest policy for ensuring that temperature threshold is not exceeded (operate at load $\ell_0$).




**Zig-Zag:** Policies that alternate between stages of cooling and heating.



# Summary

- Data stream offers a new paradigm for handling monitoring/surveillance type applications
- Provide predictive/proactive management instead of reactive management
- Offer new opportunities by coupling stream mining with power management