Data Council Singapre 2019

Autoencoder Forest for Anomaly Detection from IoT Time Series

Yiqun Hu, SP Group

Agenda

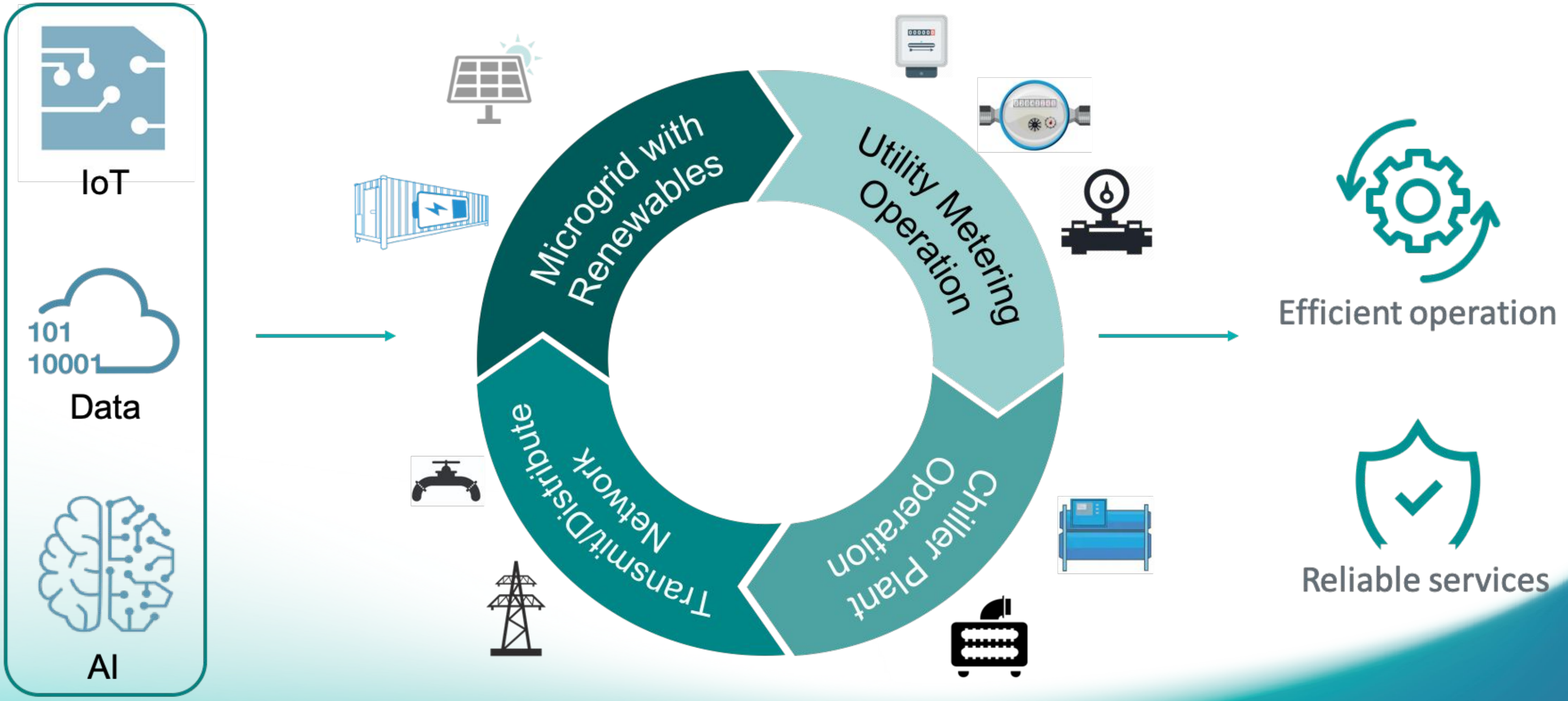
- Condition monitoring & anomaly detection
- Autoencoder for anomaly detection
- Autoencoder Forest
- End-to-end workflow
- Experiment results

Conditional monitoring & Anomaly Detection

SPgroup

Condition monitoring

SPgroup



IoT

Data

AI

Microgrid with Renewables

Utility Metering Operation

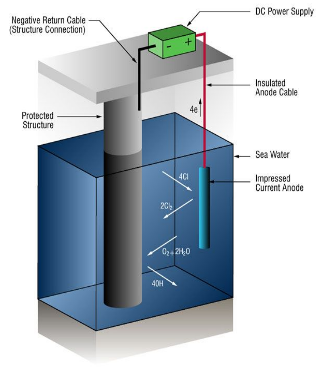
Chiller Plant Operation

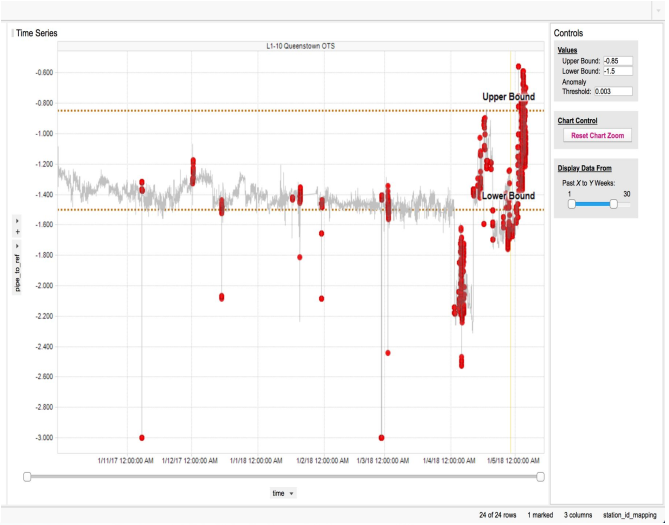
Transmit/Distribute Network

Efficient operation

Reliable services

Time-series anomaly detection





Controls

Values
Upper Bound: -0.85
Lower Bound: -1.5
Anomaly Threshold: 0.003

Chart Control
Reset Chart Zoom

Display Data From
Past X to Y Weeks:
1 30

24 of 24 rows 1 marked 3 columns station_id_mapping

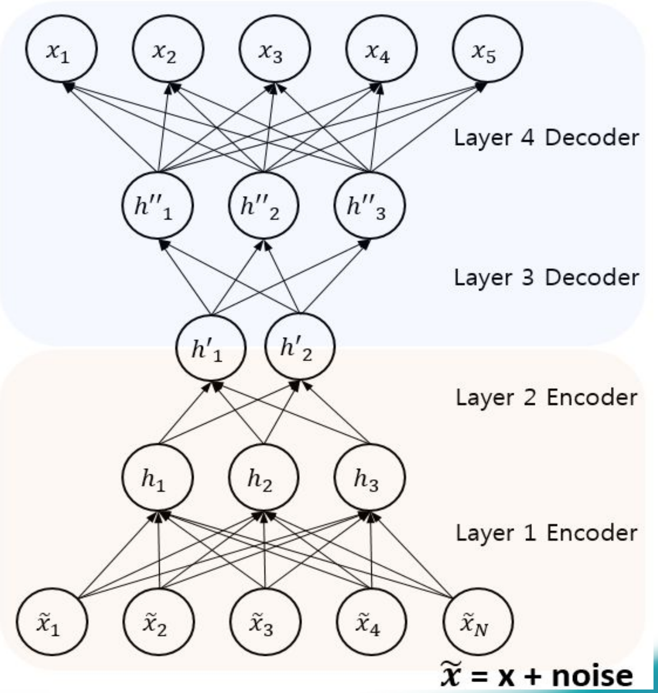
- Manual monitoring
  - Huge human effort
  - Boring task with low quality
- Rule-based method
  - Cannot differentiate different environment
  - Cannot adapt to different condition of the equipment
- Data-driven method
  - Model the common behavior of the equipment

Autoencoder for Anomaly Detection

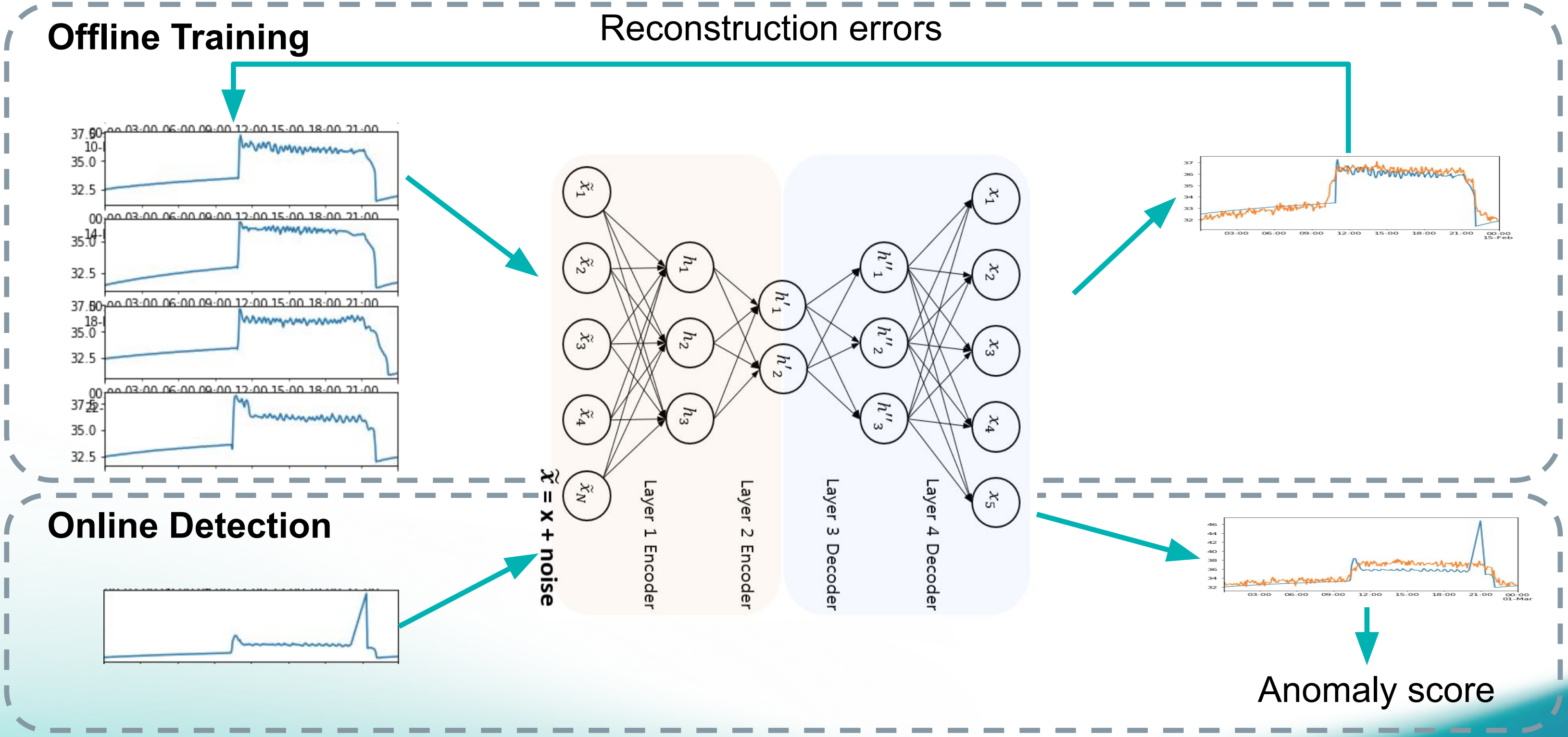
Autoencoder

- What is autoencoder
  - A encoder-decoder type of neural network architecture that is used for self-learning from unlabeled data
- The idea of autoencoder
  - Learn how to compress data into a concise representation to allow for the reconstruction with minimum error
- Different variants of autoencoder
  - Variational Autoencoder
  - LSTM Autoencoder
  - Etc.

Autoencoder Neural Network



Autoencoder for anomaly detection



Offline Training

Reconstruction errors

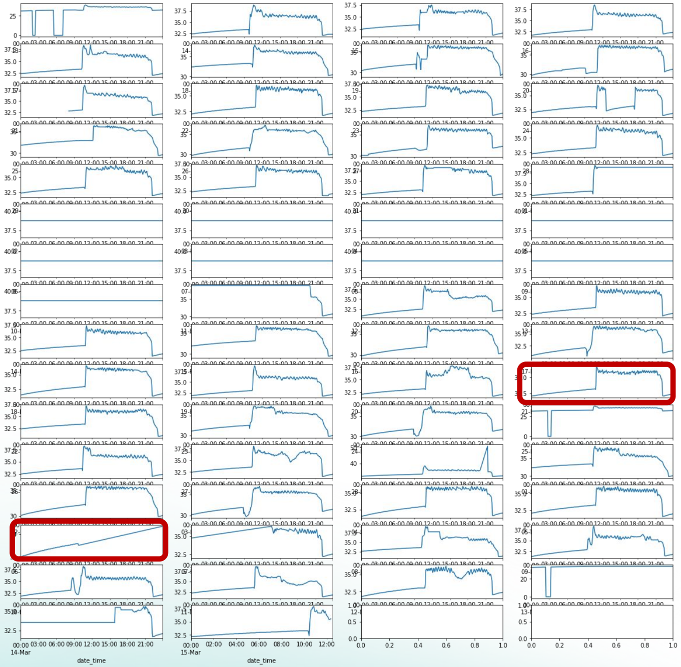
Online Detection

Anomaly score

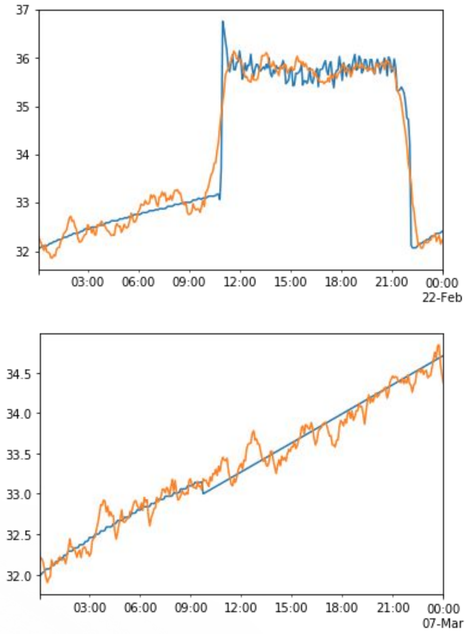
Autoencoder Forest

SPgroup

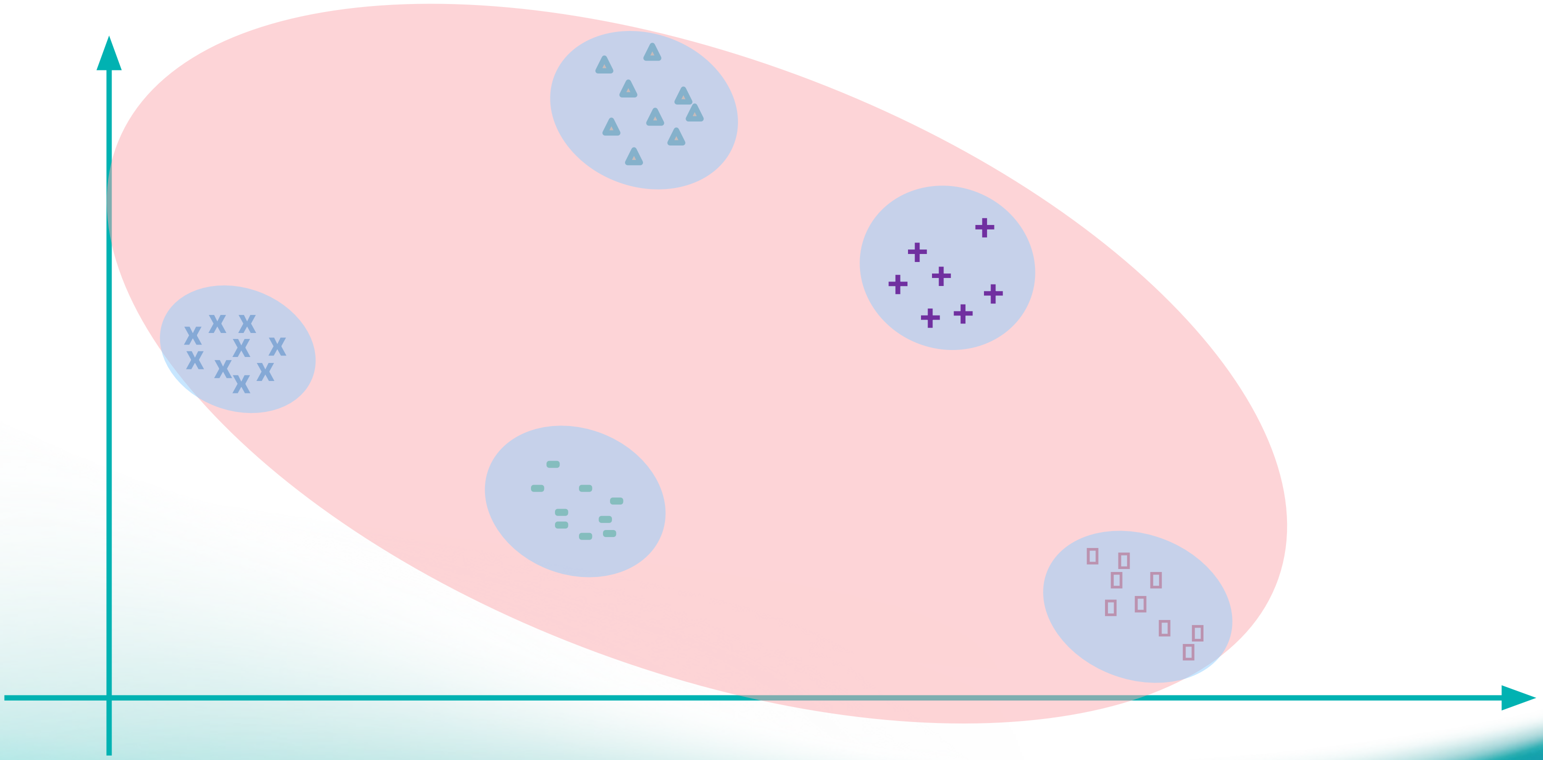
A key challenge of autoencoder



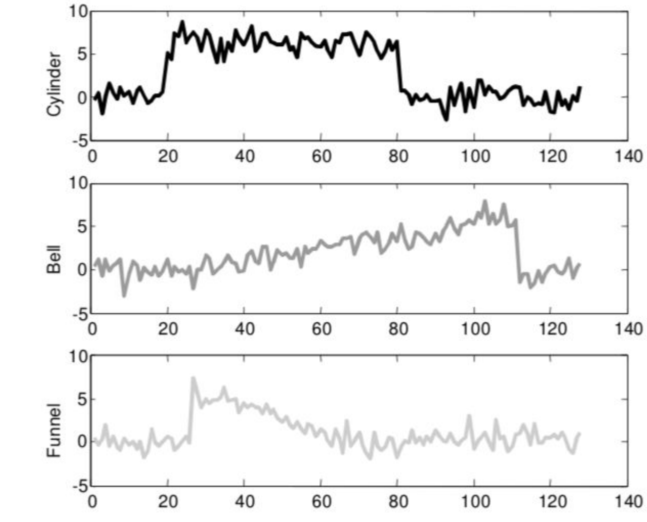
Single Autoencoder



The idea of autoencoder forest



Clustering subsequence is meaningless



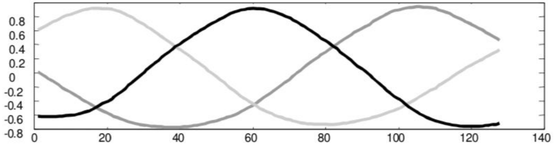
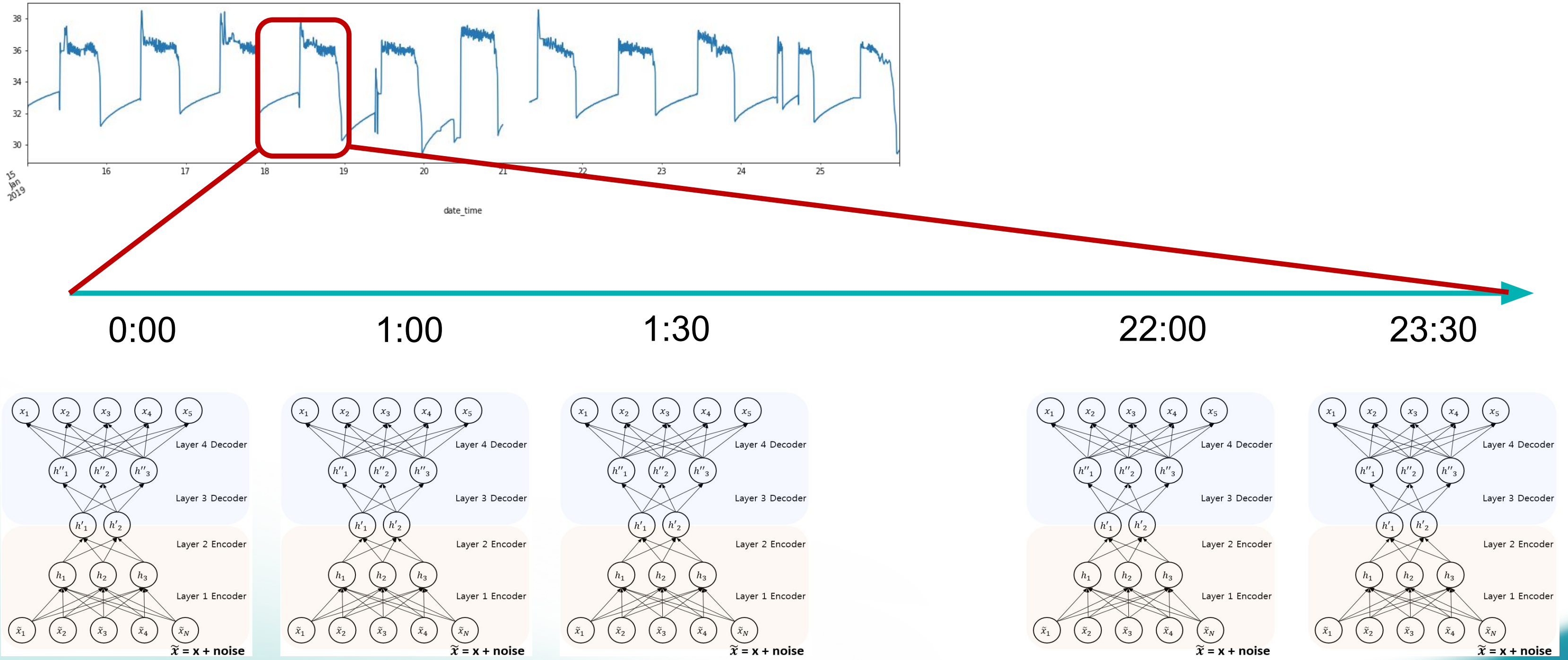


Figure 9. The three final centers found by subsequence clustering using the sliding window approach. The cluster centers appear to be sine waves, even though the data itself is not particularly spectral in nature. Note that with each random restart of the clustering algorithm, the phase of the resulting "sine waves" changes in an arbitrary and unpredictable way.

[1]. Eamonn Keogh, Jessica Lin, Clustering of Time Series Subsequences is Meaningless: Implications for Previous and Future Research

Autoencoder forest based on time



0:00

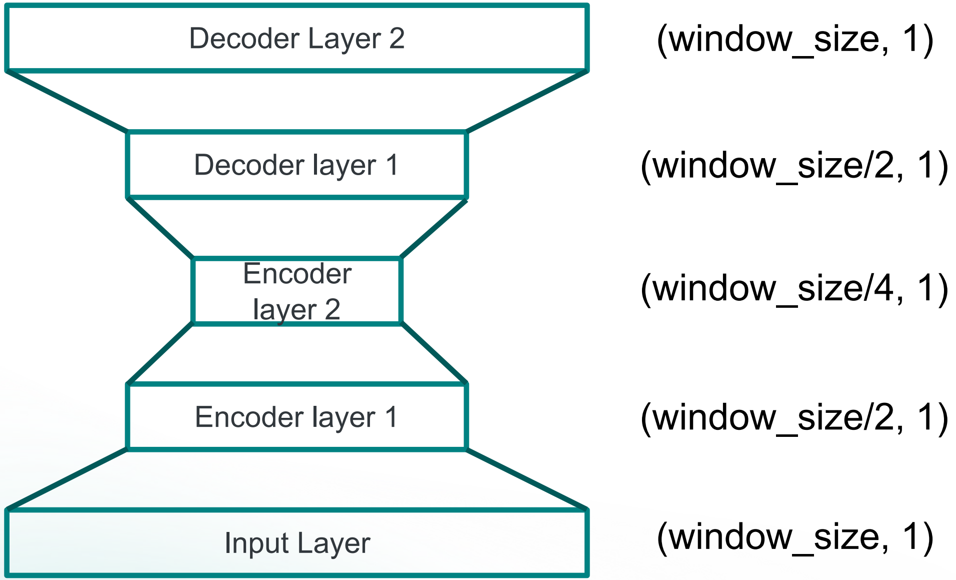
1:00

1:30

22:00

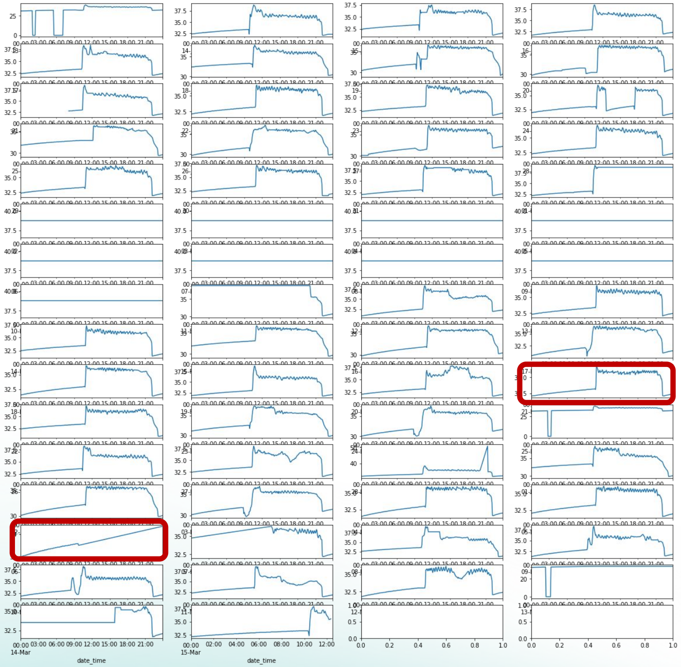
23:30

Training autoencoder forest

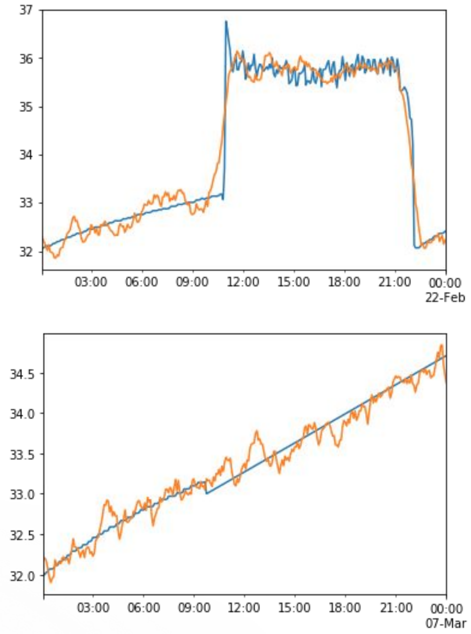


- Structure is fixed for every autoencoder. (try to make it as generic as possible)
- Each autoencoder within forest is independent. So the training is naturally parallelizable
- Using early stopping mechanism, the training of individual autoencoder can be stopped at similar accuracy.

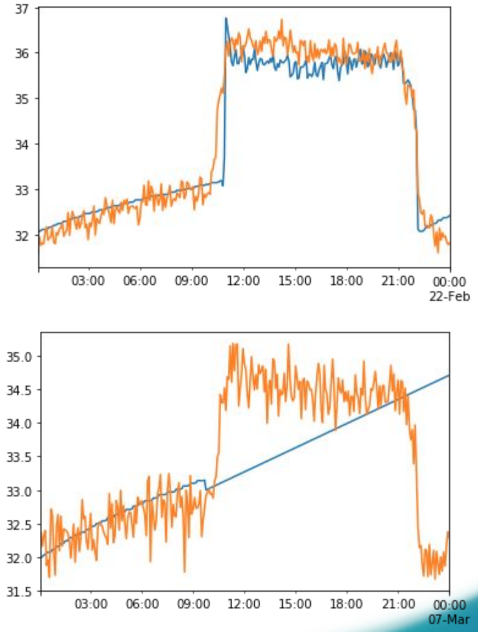
Autoencoder Forest



Single Autoencoder

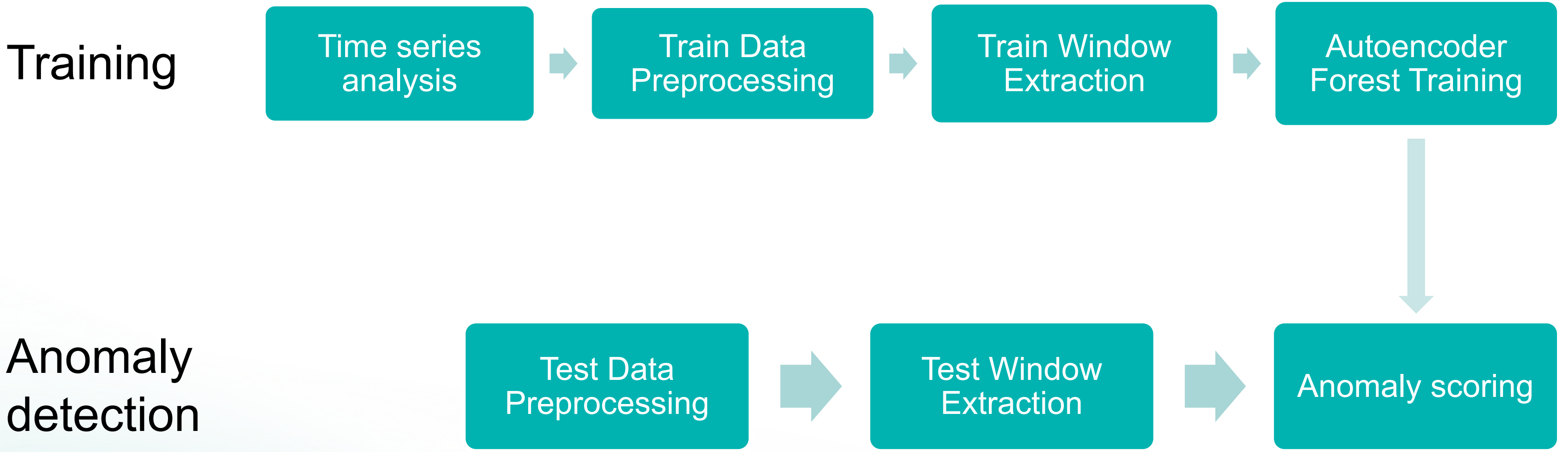


Autoencoder Forest



End-to-end Workflow

Automatic end-to-end workflow



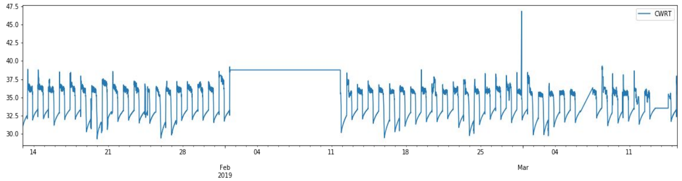
Training

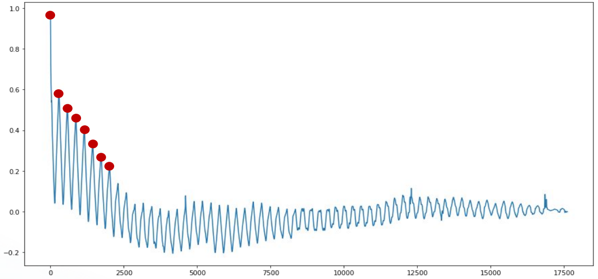
Time series analysis → Train Data Preprocessing → Train Window Extraction → Autoencoder Forest Training

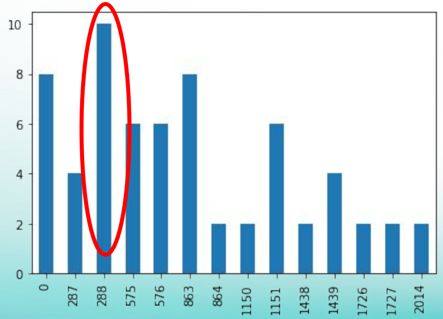
Anomaly detection

Test Data Preprocessing → Test Window Extraction → Anomaly scoring

Periodic pattern analysis



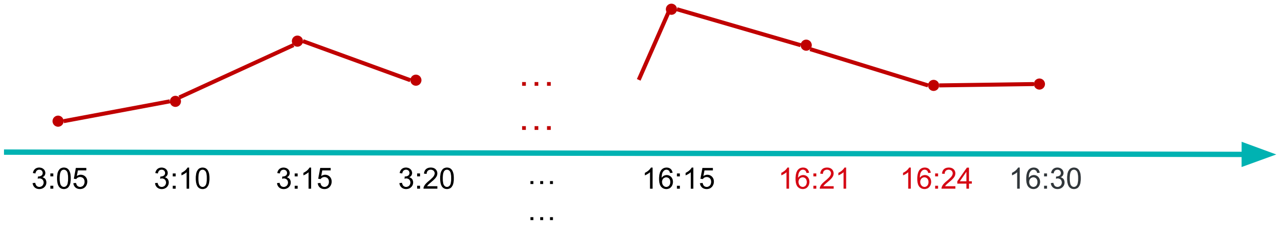




- Automatic determine the repeating period in time series
  - Calculate autocorrelations of different lags
  - Find the strong local maximum of autocorrelation
  - Calculate the interval of any two local maximum
  - Find the mode of intervala

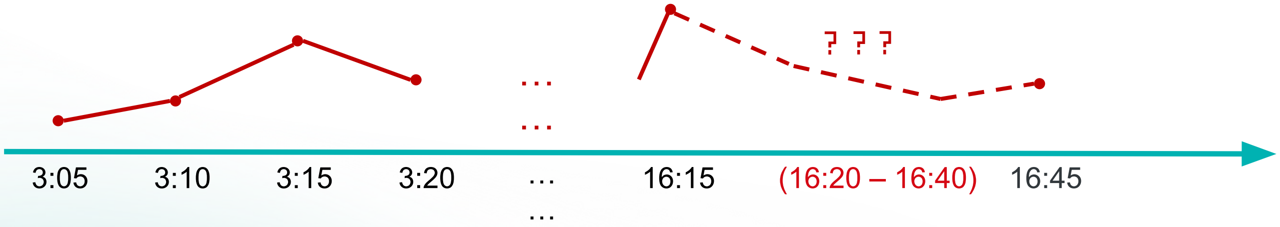
Missing data handling

Misalignment



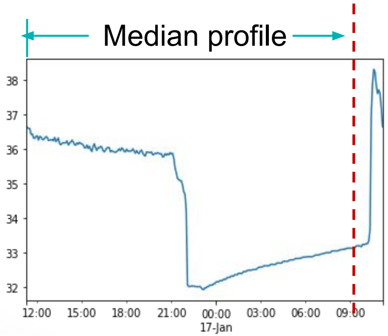
- No need to impute

Missing

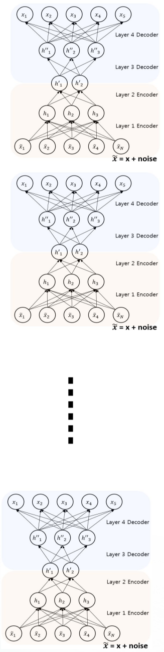


- If missing gap is small, impute with neighbouring points;
- If missing gap is large, impute with the same time of other periods;

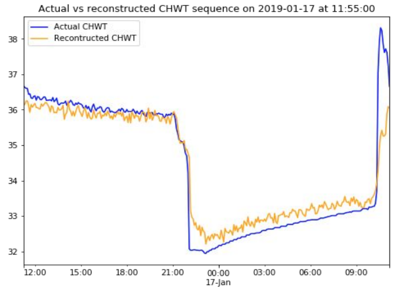
Anomaly scoring



Extract the sequence window end at time *t*



Learned autoencoder forest



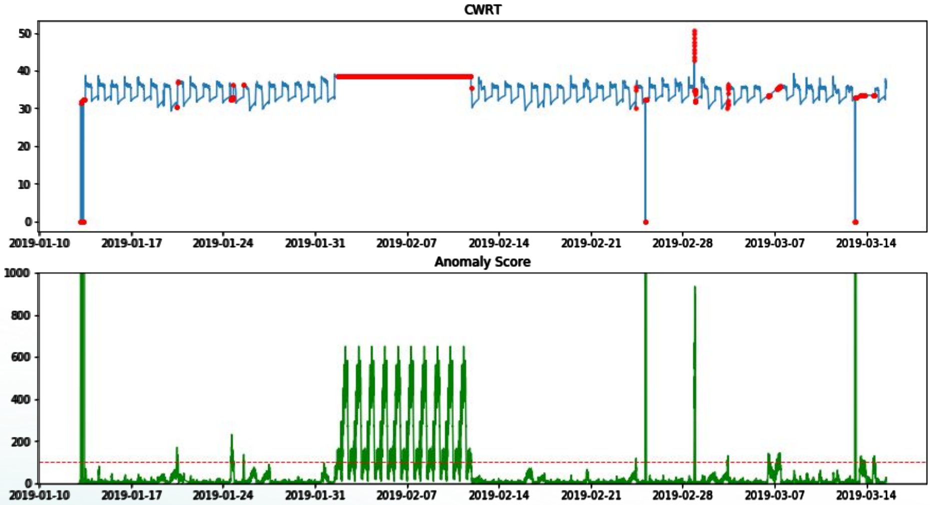
Corresponding autoencoder reconstruct the sequence window at time *t*

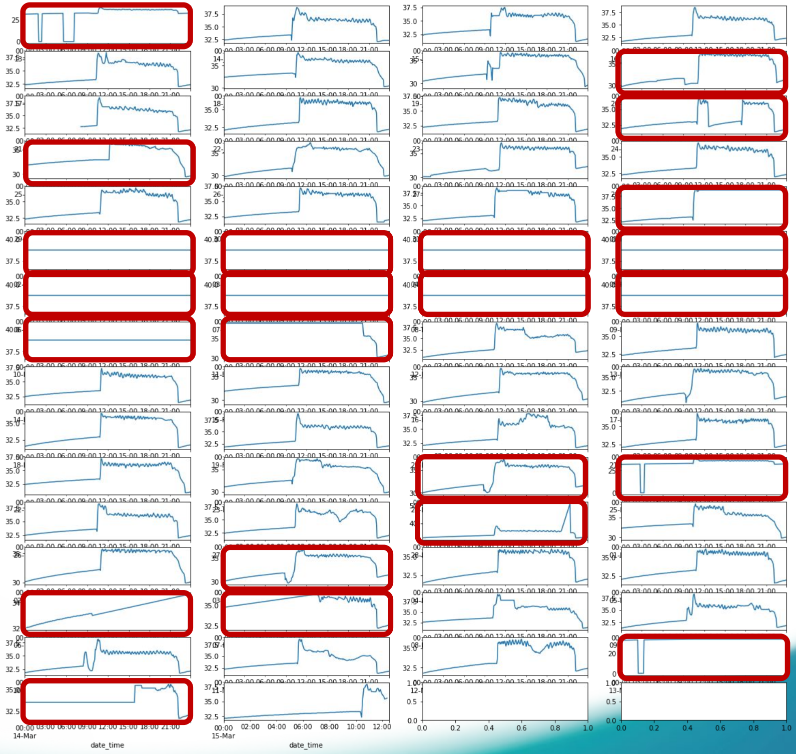
Score= $\sum_{k=1}^{K}(Y_k - \check{Y}_k)^2$

Compute reconstruction error as anomaly score

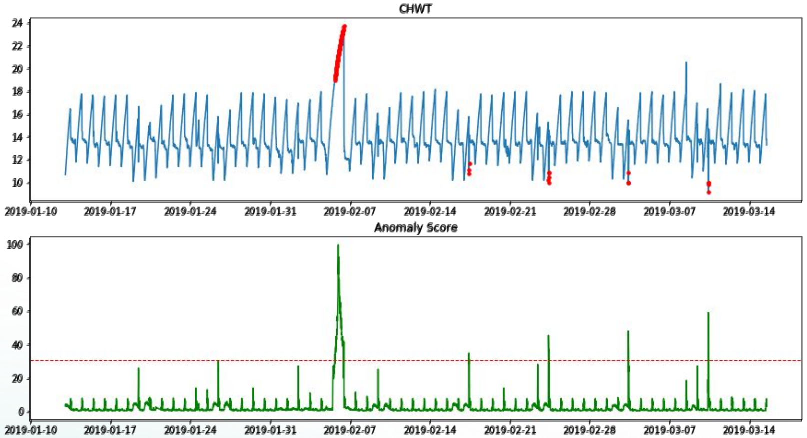
Experiment Results

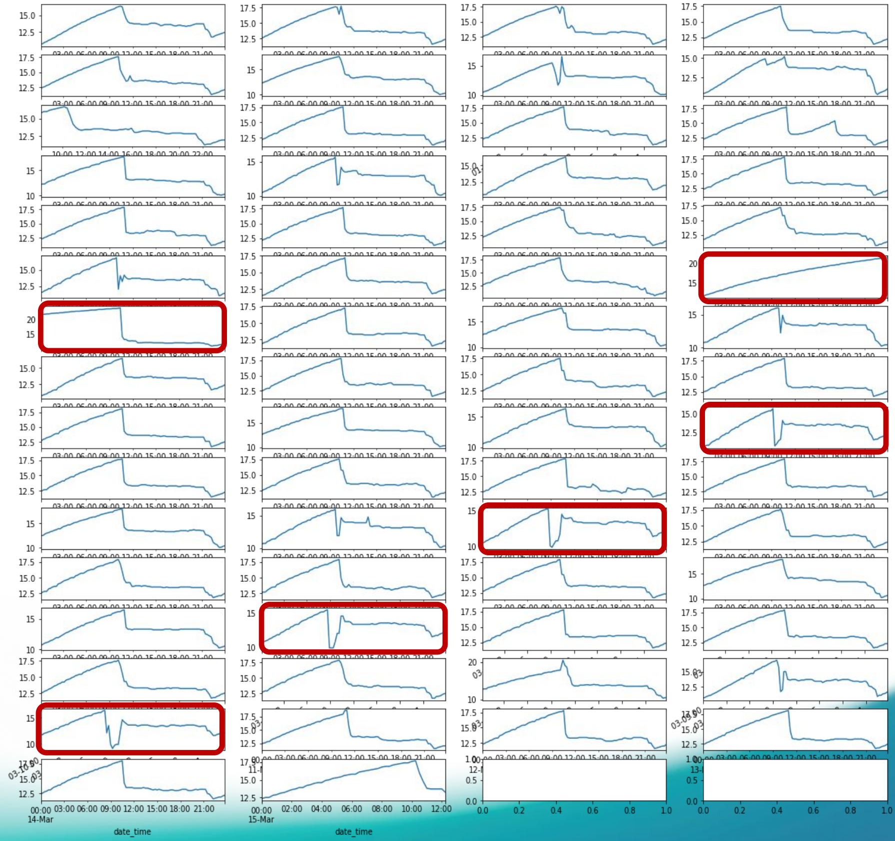
Cooling tower – return water temperature





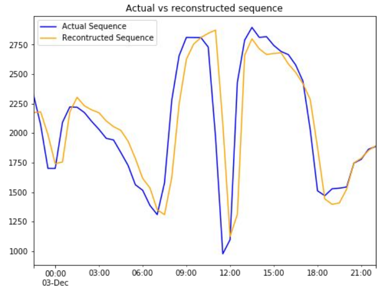
Chiller – chilled water return temperature





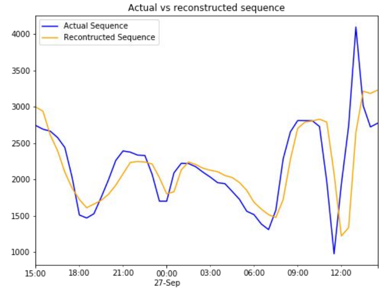
Smart meter – half hour consumption

Normal data

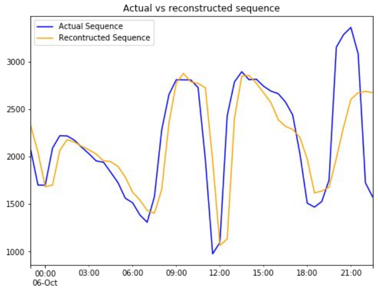


2018-12-03 22:00:00

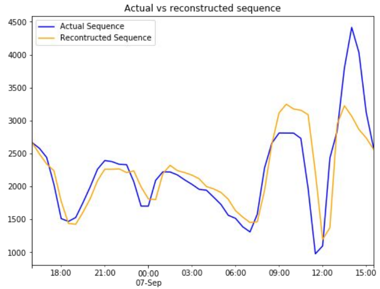
Top 3 Detected Anomaly



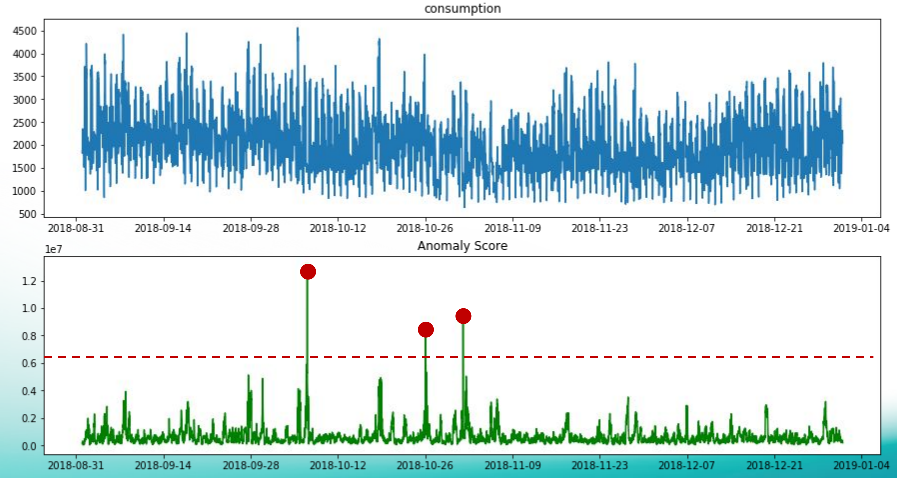
2018-09-27 14:30:00



2018-10-06 22:30:00



2018-09-07 15:30:00



We have built systems with lots of time series data

Chronos+

A common platform for time series data, with built-in AI capabilities

SPgroup

powering the nation