DATA DISCOVERY TOOLBOX

Department of Industrial Systems Engineering and Management
IE3100M Systems Design Project AY2016/2017

NUS Supervisors
Prof. Andrew Lim Leong Chye
Prof. Goh Thong Ngee

Micron Supervisors
Vincent Hong
Wu Biao

Group 21 Members
Armaan B Ashraf Ali
Lee Ming Loon
Chew Woon Sin
Lim Zhao Jun
Goh Chong Rui Gordon
Shane Leung Yu Xi
Huang Jin Jing

Project Description

Micron's Data Science Team's long term goal seeks the automation of drift detection systems for their various production processes. The team intends to create a system capable of detecting drifts, trends and anomalies without prior knowledge of dataset.

The issues that Micron's Data Science Team faces are summarized as such:

1) Overwhelming requests for data analysis by line engineers on the Team (Figure 1)

2) Limited statistical methods, functionality and flexibility of current toolbox

3) Usability of toolbox (only usable and **comprehensible** by data scientists)

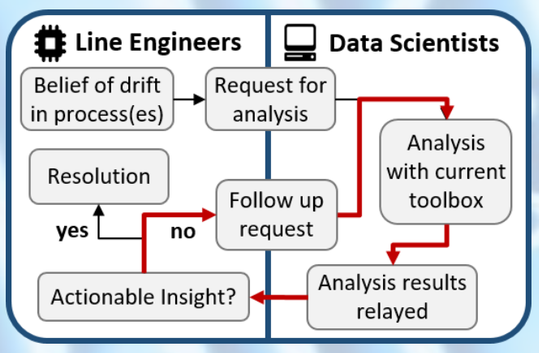

Figure 1: Process Flow for Data Analysis
(Positive feedback loop)

Request overloading (Figure 1) results from poor usability and insufficient capabilities of the toolbox (Figure 2), hampering efficiency.

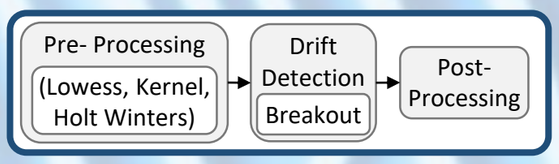

Figure 2: Simplified Current Framework

Project Objective

The Data Discovery Toolbox should address the main issues highlighted while having highly generic capabilities yet provide methods applicable to specific datasets and the following:

- A **comprehensive framework** of **non-parametric methods and algorithms** allowing **flexibility in user choice on the analysis methods** and **respective parameter customizations,** coupled with effective performance on most of the dataset types produced by Micron's fabrication plants.

- A **user-friendly interface** allowing **data scientists and line engineers alike** to **analyze, interpret and detect anomalies in data.**

Approach

The R&D process was conducted in several phases:

1) **Analysis of Current Framework**:
Identifying pitfalls and areas of improvement for development.

2) **Exploration of Datasets**:
Identifying types of datasets produced by Micron.

3) **Research on Statistical Methods & Testing**:
Based on findings, effective and efficient methods were used to analyse various dataset types.

4) **Development and Testing of the Toolbox**:
A library consisting of statistical methods and an integrated user interface synergized to achieve project objectives. Incremental, repeated usability testing conducted to provide a stable and error-free software.

Analysis

Current Framework

Pros
+ Effective drift detection algorithm
+ Noise reduction preceding analysis
+ Simple framework

Cons
- Reliance solely on default parameters prevent flexibility of analysis
- Usability of toolbox (only usable and comprehensible by data scientists)
- Little statistical analysis of dataset provided prior to drift detection
- No framework to recommend suitable methods based on dataset type
- Limited variety of methods and algorithms used, insufficient for the plethora of dataset types

Key Takeaways
- General flow of current framework may be preserved for use in development of new toolbox
- Modularity of toolbox design for developers to add new methods and built-in flexibility, are crucial to improve analysis in the long run

Exploration of Datasets

From Data Science Team's findings as well as statistical analyses, the **time series** datasets generally fall under categories below:

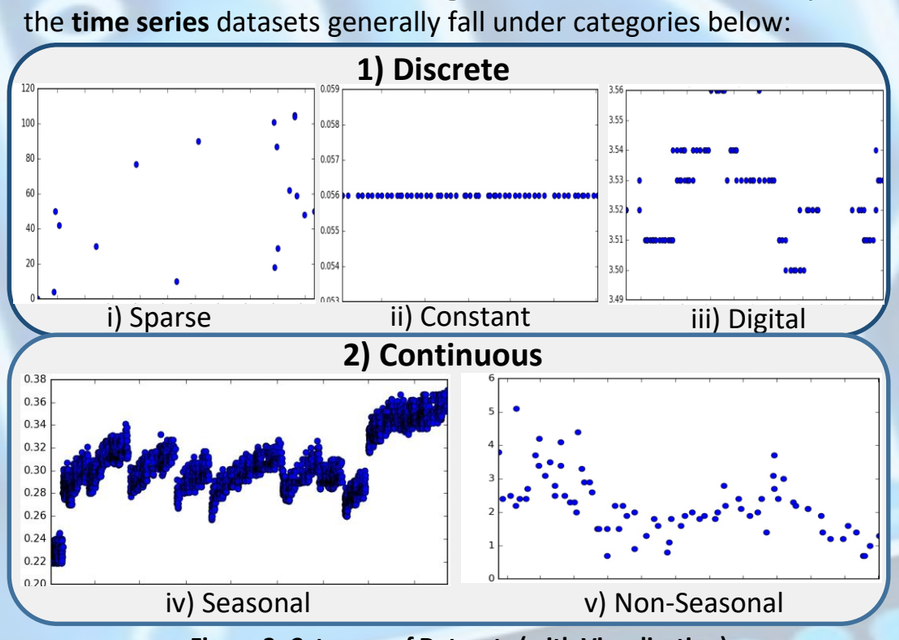

Figure 3: Category of Datasets (with Visualization)

Figure 3 as above demonstrates the various expected distributions associated with the nature of data (discrete, continuous).

Results

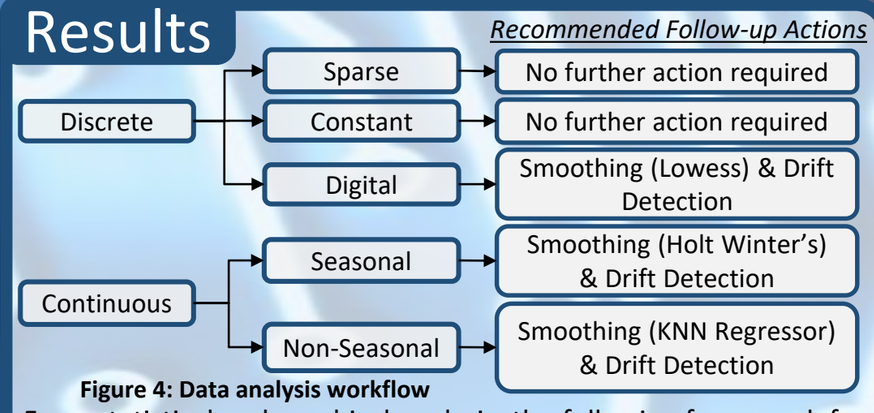

Figure 4: Data analysis workflow

From statistical and graphical analysis, the following framework for classification and recommendation of methods was formulated to improve efficiency of default methods and parameter selection. Dataset type was identified using the methods in the table below.

Dataset Classification

Discrete: The heuristics developed, considers the time proximity of data points, and number of points that constitute a discrete y-value or an outlier. Their frequency determines the respective discrete sub-categorizations seen in Figure 4 above.

Seasonal: A periodogram (which utilizes Fast Fourier Transform) is used to identify the average periodicity in each dataset.

Methods

The following methods used for processing datasets are located in the source code of the Data Discovery Toolbox as shown below:

Pre-Processing

Lowess (Locally Weighted Smoothing)

It is a non-parametric regression technique which uses local polynomial regression and thus resilient against outliers.

Holt Winters (Exponential Smoothing)

This method uses triple exponential smoothing that accounts for non-stationarity and seasonality. The coefficients for each component are obtained by minimising Mean Square Error (MSE).

K-Nearest Neighbours Regressor

This is a non-parametric method takes K nearest neighbours and outputs a weighted average value for the object based on a weight function.

Local Polynomial Kernel Regressor

This is a non-parametric regression technique that estimates the conditional expectation of a random variable, which is calculated by for each point by assigning weights to neighbouring points (based on a kernel function) and finding the weighted average.

Exponential Weighted Moving Average (EWMA)

EWMA is an infinite impulse response filter that applies weighting factors to decrease exponentiation and thus smooth the dataset.

Drift Detection

Change Point

Change point algorithm uses a hypothesis test of 'no change' vs 'change', whereby 'no change' assumes the data follows the same distribution, and 'change' assumes multiple distributions within the dataset.

Breakout

Breakout Algorithm uses E-Divisive with Medians (EDM) to detect divergence in mean. EDM is non-parametric which is capable to detect multiple breakouts in a given time series.

Toolbox

Integrated User Interface (Software)

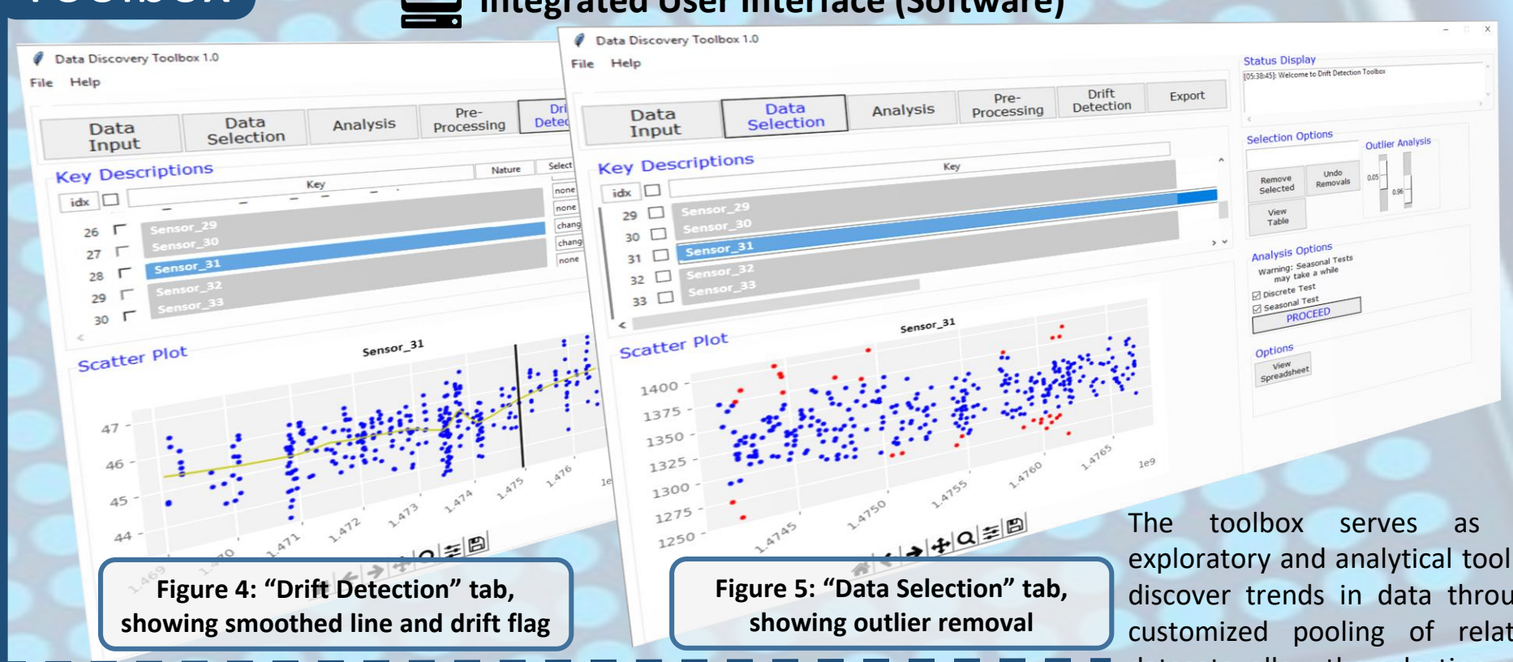

Figure 4: "Drift Detection" tab, showing smoothed line and drift flag

Figure 5: "Data Selection" tab, showing outlier removal

The toolbox serves as an exploratory and analytical tool to discover trends in data through customized pooling of related datasets, allow the selection of a wide array of pre-processing and processing methods, as well as the customization of the parameters used in these methods.

The toolbox consists of 2 layers of interaction; the source code and the integrated user interface. The source code contains all the relevant methods and functions used in the integrated systems.

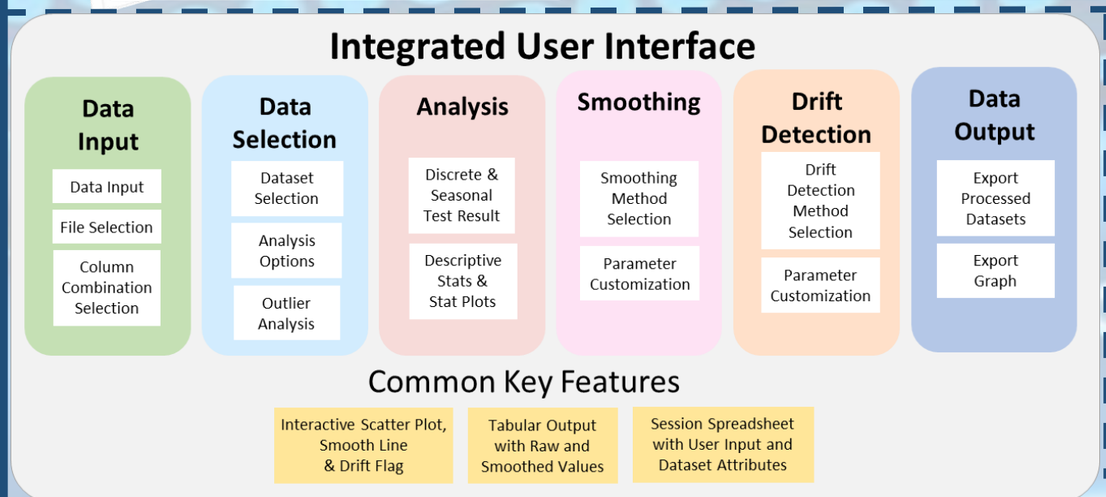

Figure 6: Integrated User Interface

The system flow as seen above in Figure 6 of the interface is described as follows:

(1) Data Input: Input choice and selective grouping of data by data descriptors.
(2) Data Selection: Select relevant datasets for analysis and perform outlier removal.
(3) Analysis: Performs a dataset analysis to identify distribution discreteness as well as seasonality in a dataset. From these attributes, methods and parameters will be recommended for subsequent steps.
(4) Smoothing: Select relevant methods and adjust parameters accordingly to smooth datasets for improved efficiency of drift detection.
(5) Drift Detection: Selection of relevant methods, set parameters to detect drift.
(6) Data Output: Exporting of tabular and graphical outputs of processed data.