C3889 – Capstone Project
Smart Water Quality Monitoring System

PROJECT PRESENTATION

01-AUG-2018

Presented By:

Lim Hong Loon (CET116510)

Lim Yuan Her (17060167)

Agenda

1. **Business Analysis**
   - ❑ Problem/ Solution
   - ❑ System Functional Requirements
2. **Schedule**
   - ❑ Tasks/ Milestones
3. **System Design**
   - ❑ Overall Architecture
   - ❑ Data Metrics
   - ❑ Experimental Test Setup
4. **Data Flow**
   - ❑ End-to-End Information Flow
5. **Data Visualization**
   - ❑ Dashboard
   - ❑ Notifications
6. **Data Analytics**
   - ❑ Exploratory Data Analysis
7. **Machine Learning**
   - ❑ Clustering
   - ❑ Multi-class Classification
   - ❑ Anomaly Detection
8. **Future Enhancement**
   - ❑ Exploratory Data Analysis

Business Case

- **Problem**
  - Water quality at each stage in treatment process monitored through SCADA, with analytical equipment data verified with laboratory testing once every 8 hours (Expensive and time-consuming task with quality and effectiveness dependent on laboratory technicians' experience)
  - Use of on-premise SCADA systems in traditional water treatment process requires significant upfront investments and ongoing maintenance costs with high risk of data loss from faulty equipment.

- **Solution Overview**
  - 4 major features to reduce inefficiencies in current process workflow/ infrastructue:
    - Pre-classify water quality into different grades before start of treatment process to identify treatment parameters required to optimize effectiveness and quality of treatment process
    - Employ messaging systems e.g. SMS, email etc. to notify abnormalities in water quality for pre-emptive actions to be taken
    - Use cloud-based system (Microsoft Azure) for data storage and processing
    - Use Microsoft PowerBI for dashboard visualization with multiple access modes (desktop/web/mobile)

System Functional Requirements / Implementation

- Provide real-time monitoring of water quality parameters with real-time visualization dashboard accessible via desktop, web and mobile devices

- Provide alert messaging via SMS and email notifications for abnormal water quality parameter detection and equipment malfunction

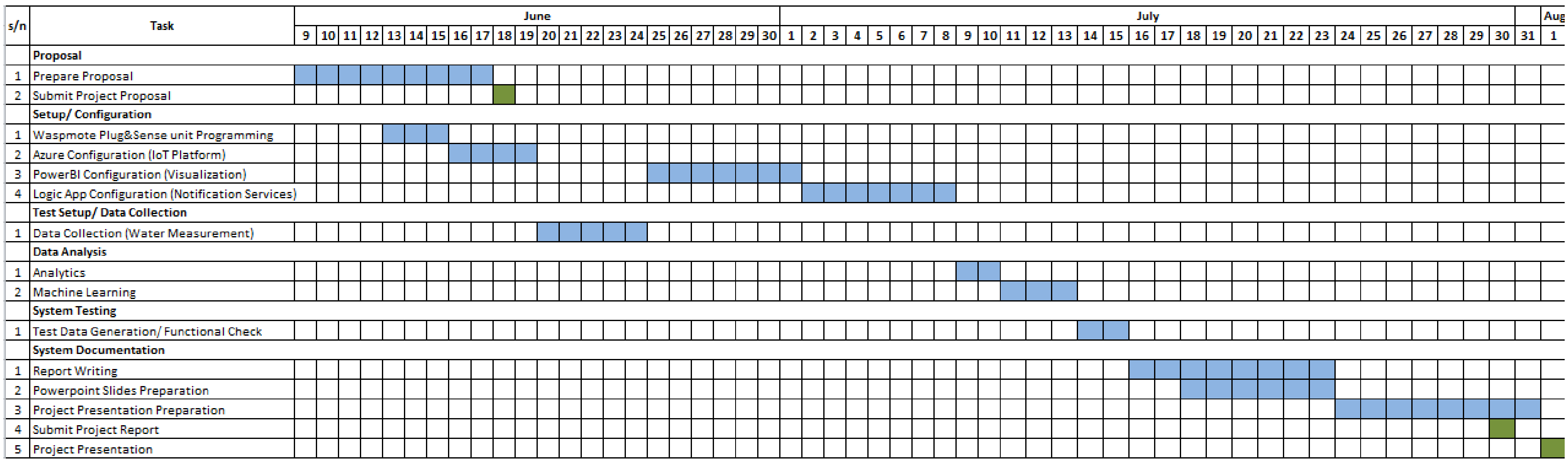
- Provide analytic capabilities e.g. using data analytics, machine learning etc. to provide the following metrics:

    - Grading classification of water quality e.g. Grade 1, Grade 2 etc.
    - Overall System Alert Status

Project Schedule

Project Implementation

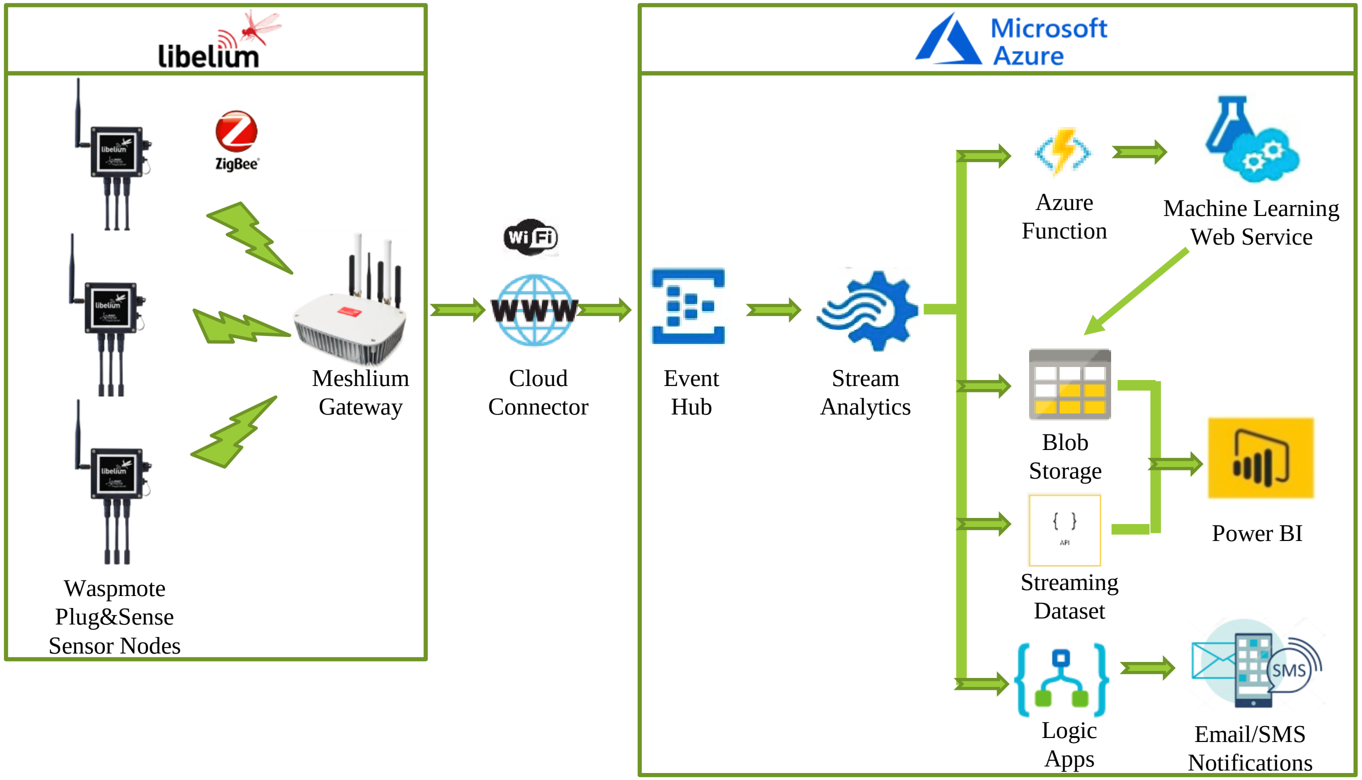
- 6 main phases
  - Proposal
  - Waspmote Programming/ Azure Setup & Configuration
  - Test Equipment Setup/ Data Collection
  - Data Analytics (Exploratory Data Analysis + Machine Learning)
  - System Testing (including functional checks)
  - Documentation (Report + Presentation)



System Design - 1

❑ Overall System Architecture

- Data Ingestion
  - Libelium Smart Water Kit (Waspmote Plug&Sense) + Meshlium Gateway
- Data Processing/Analysis/Machine Learning/Notifications
  - Microsoft Azure Platform
- Data Visualization
  - Microsoft PowerBI



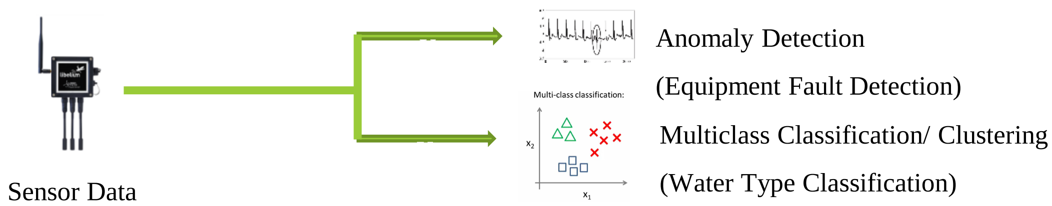
System Design - 2

❑ Data Metrics

s/n	Data	Description
1	Temperature	Measures the hotness/coldness of the water (degC)
2	pH Level	Measure of the acidity/ alkalinity of the water (scale of 0-14 with neutral point of 7)
3	Dissolved Oxygen (DO)	Measures the amount of gaseous oxygen (O2) dissolved in the water (mg/L)
4	Oxygen Reduction Potential (ORP)	Measures the water’s oxidizing power and it’s potential ability to sanitize itself (mV)
5	Conductivity	Measures the ionic strength of the water and it’s ability to conduct electricity (uS/cn)

❑ Data Analysis Work Flow

- ➢ Clustering Analysis/ Multi-class Classification for Water Type Classification
- ➢ Anomaly Detection for Equipment Fault detection



Experimental Test Setup

❑ Media used

s/n	Material	Description	Remark
1	Tap Water	-	Grade 1
2	Soap		Grade 2
3	Flour		Grade 3
4	Shampoo		Grade 4
5	Sea Water		Grade 5

❑ Measurement Process

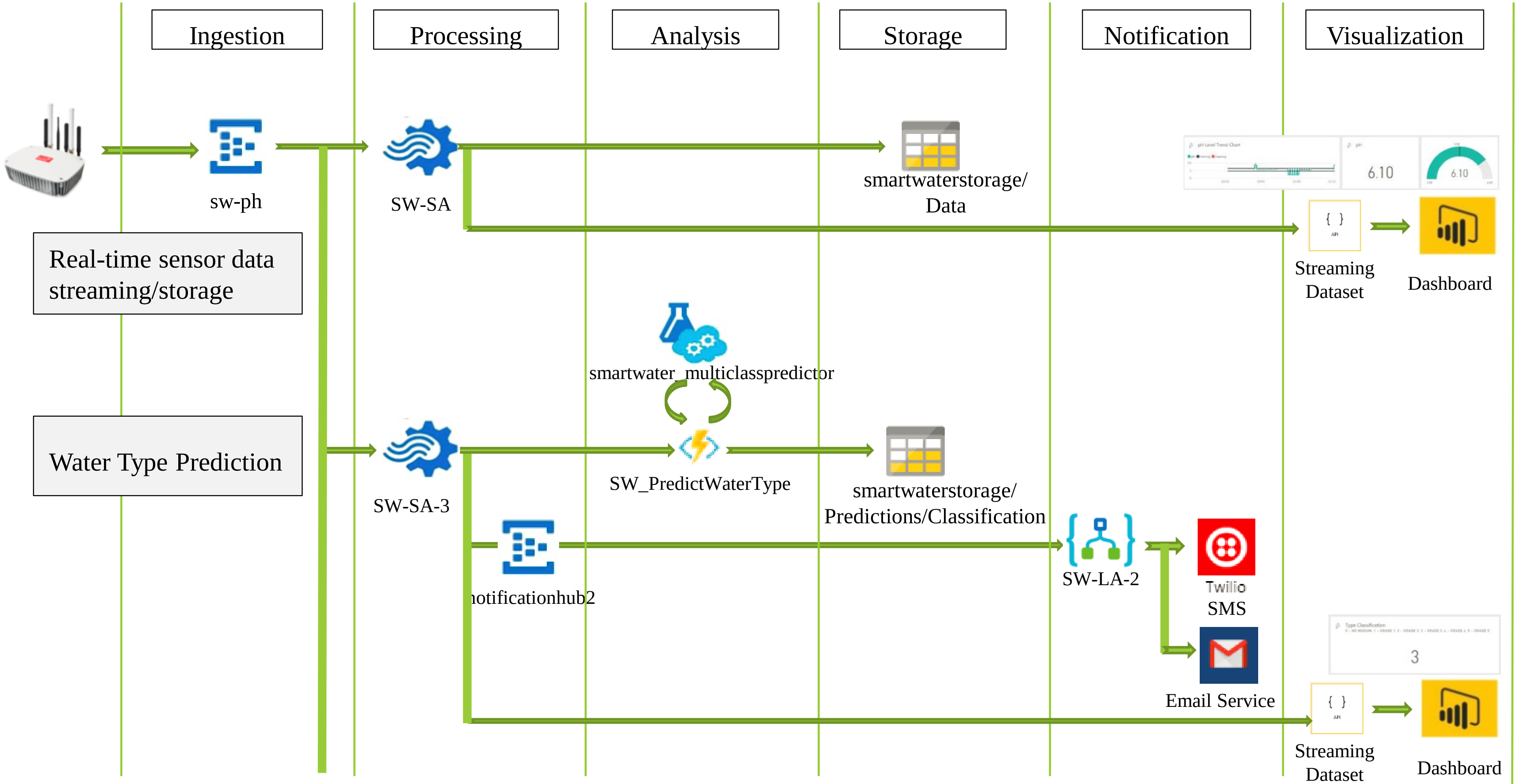
Waspmote
Plug&Sense Unit

5 Sensor probes
(Temperature, pH, Dissolved Oxygen,
Oxidation Reduction Potential, Conductivity)

Container with
water mixed with
media used

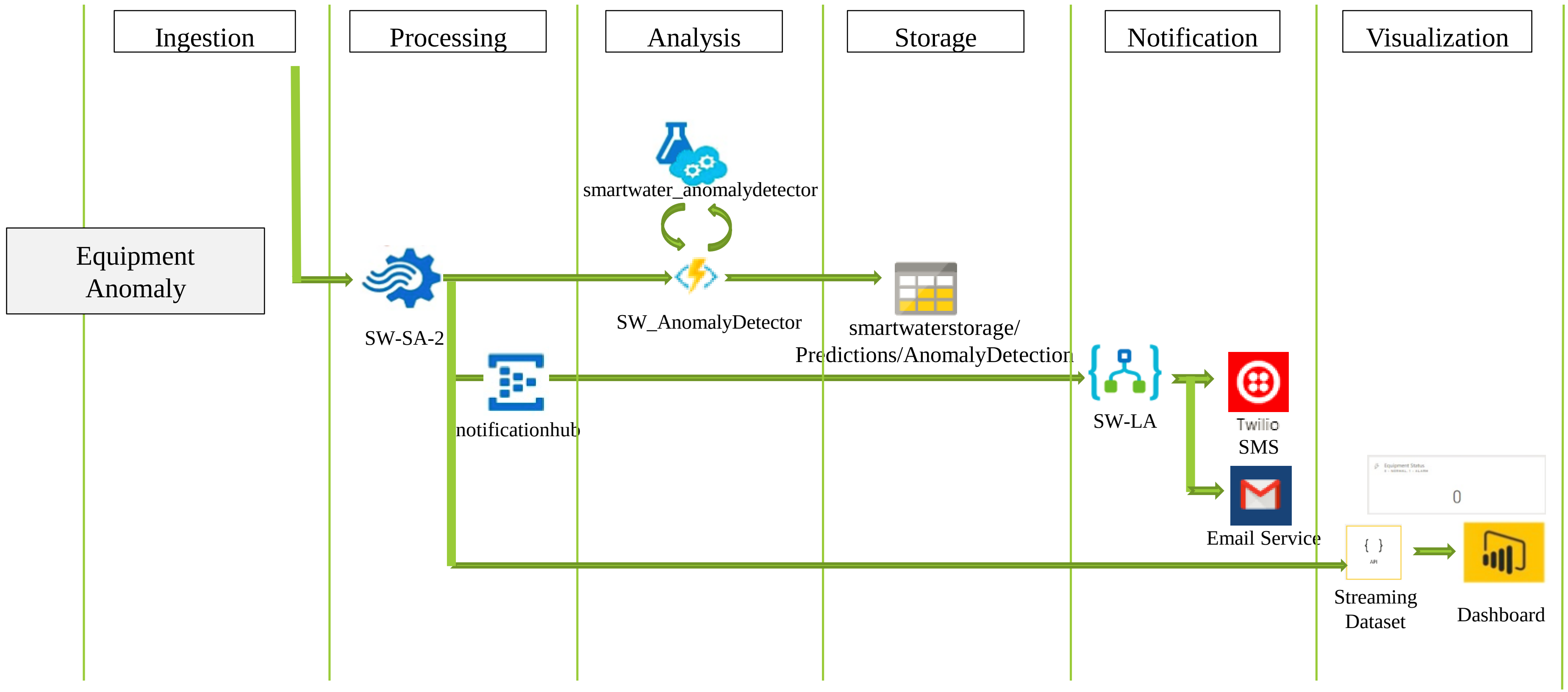
Data Flow - 1

- End-to-end sensor data flow
  - Data ingestion via Azure Event Hubs
  - Data processing via Azure Stream Analytics Jobs
  - Data Analytics via Azure Machine Learning Studio Experiments/ Notebooks
  - Data Notification via Azure Logic Apps (Twilio SMS + Google Mail built-in connectors)
  - Data Visualization via Microsoft PowerBI dashboards



Data Flow - 2

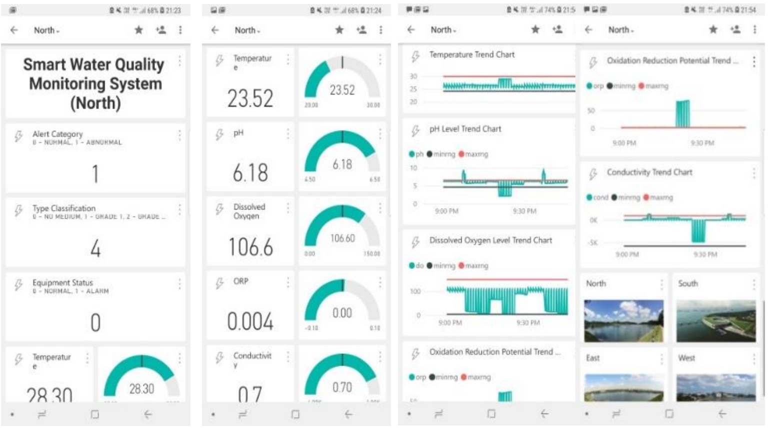
- Data Information Flow



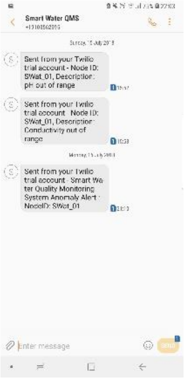
Data Visualization (Dashboard/ Notifications)

- Real-time Visualization Dashboard
  - Displays sensor data and analytics prediction output (water type classification/ equipment anomaly detection)
  - Accessible from web browser and from mobile devices (using PowerBI App)
  - Notification via SMS/ email

Web Browser View



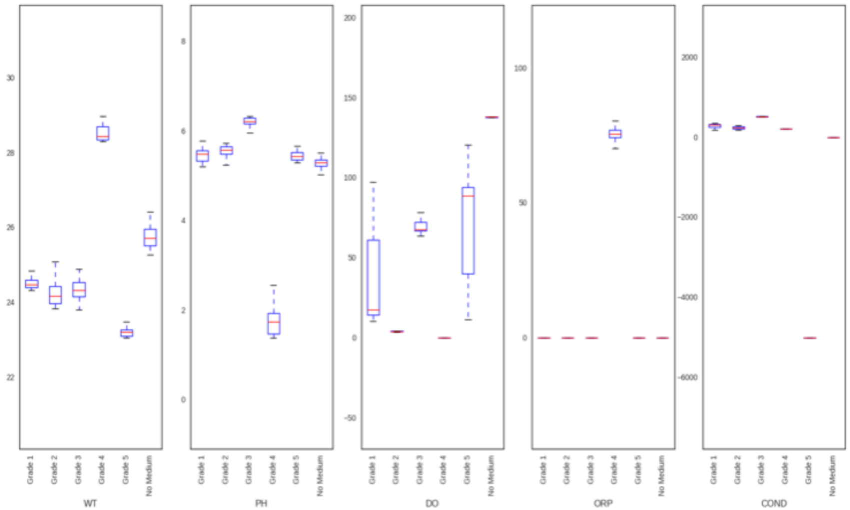
Mobile Device View

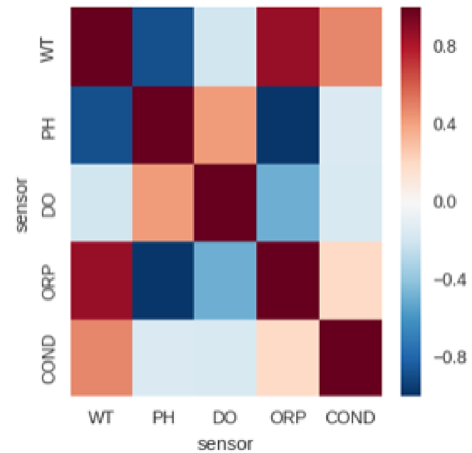


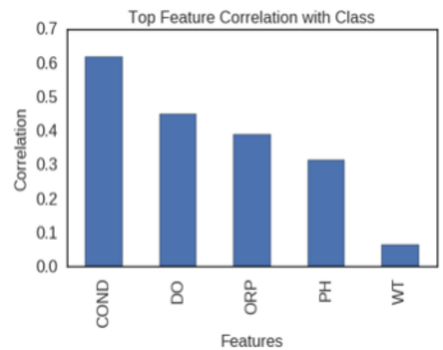
SMS/ Email Notifications

Data Analytics

- Feature Distribution/ Correlation
  - Clear distinguishing features among different water types
  - Strong correlation between pH/ORP, pH/WT features
  - Conductivity displays strongest correlation with class attribute



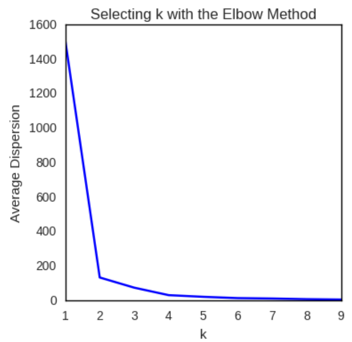




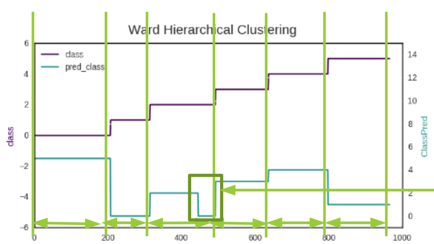

```
Top Absolute Correlations

sensor  sensor
PH      ORP       0.973068
WT      PH        0.877850
        ORP       0.873558
        COND      0.490239
DO      ORP       0.488710
PH      DO        0.425376
ORP     COND      0.201313
WT      DO        0.195034
DO      COND      0.158004
PH      COND      0.147031
dtype: float64
```

Machine Learning - Clustering

- Selection of optimal K
  - Uses Elbow Method
  - Number of Clusters = 6

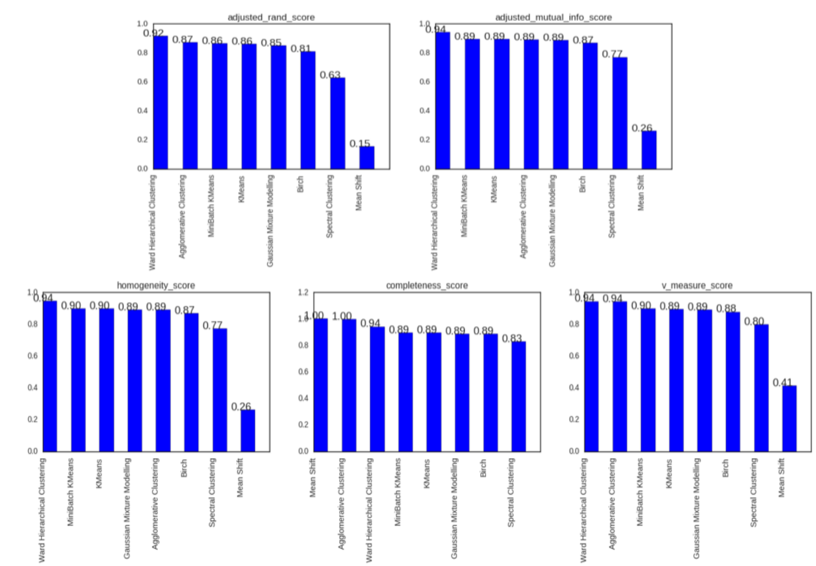


- Clustering labelling vs class information



Partial Class 2 data misclassified as Class 1

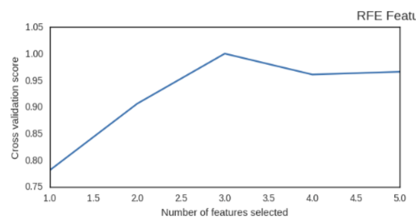
- Algorithm Evaluation
  - 8 algorithms tested
    - K-Means, Gaussian Mixture Modelling, Spectral Clustering, Ward Hierarchical Clustering, Mean Shift, Birch, MiiniBatch K-Means, Agglomerative Clustering
  - Criteria used (closer to 1 -> better clustering result)
    - Adjusted Rand Score
    - Adjusted Mutual Information Score
    - Homogenity/Completeness/V-Measure scores
  - **<u>Ward Hierarchical Clustering</u>** selected

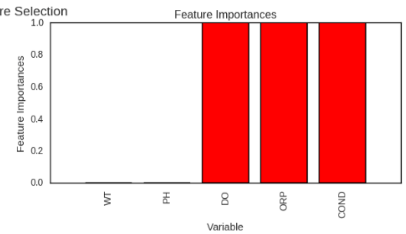


Machine Learning – Multiclass Classification (1)

- Feature Selection
  - Used to select features that contribute most to the prediction variable to improve accuracy and reduce overfitting
  - 3 selection methods used - Reduced Feature Elimination (RFE), Model Feature Importance (MFI), Factor Analysis (FA)
  - RFE shows 3 most importance features are DO (Dissolved Oxygen), ORP (Oxidation Reduction Potential) and Conductivity (COND)
  - MFI shows all 5 features collectively contribute to >95% importance with top 3 features identified as DO, ORP and COND
  - Factor Analysis shows all 5 features account for the variability in the dataset

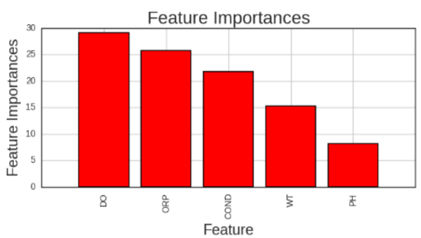
```
Optimal number of features : 3
```

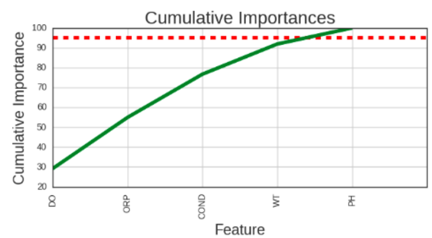




	WT	PH	DO	ORP	COND
0	-0.8498	0.2151	8.3927	-5.6838	-1995.6224
1	-0.4055	0.7676	51.8879	-16.3169	0.2669
2	1.4047	-1.1852	7.2845	22.2898	-0.0383
3	0.1875	-0.1530	-0.0159	-0.1494	-0.0003
4	0.0000	0.0000	-0.0000	0.0000	-0.0000

```
Number of important features:  5
```

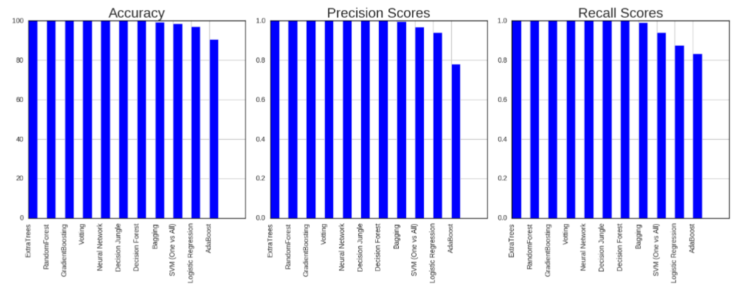





```
Number of features for 95% importance: 5
```

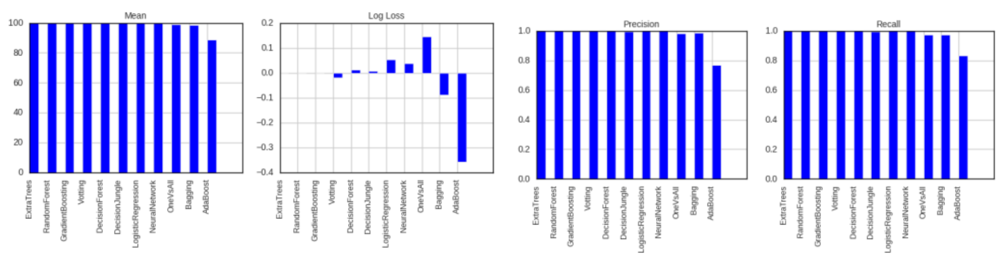
Machine Learning - Multiclass Classification (2)

- Multi-Class Classification
  - 11 classifier algorithms
    - 6 ensemble (Extrra Trees, Random Forest, Gradient Boosting, Bagging, Votting, Adaboost)
    - 5 Azure (Neural Network, Decision Jungle, Decision Forest, SVM (One Vs All), Logistic Regression)
  - Decision Forest / Random Forest classifiers found suitable for performing water type classification.



Name	Accuracy	Precision	Recall
ExtraTrees	100.0000	1.0000	1.0000
RandomForest	100.0000	1.0000	1.0000
GradientBoosting	100.0000	1.0000	1.0000
Votting	100.0000	1.0000	1.0000
Neural Network	100.0000	1.0000	1.0000
Decision Jungle	100.0000	1.0000	1.0000
Decision Forest	100.0000	1.0000	1.0000
Bagging	99.4845	0.9958	0.9907
SVM (One vs All)	98.6183	0.9691	0.9420
Logistic Regression	97.0639	0.9408	0.8768
AdaBoost	90.7216	0.7807	0.8333

Classifier Algorithm Evaluation Results



Name	Accuracy	Logloss	Precision	Recall
ExtraTrees	100.0000	-0.0014	1.0000	1.0000
RandomForest	100.0000	-0.0004	1.0000	1.0000
GradientBoosting	100.0000	-0.0007	1.0000	1.0000
Votting	100.0000	-0.0201	1.0000	1.0000
DecisionForest	100.0000	0.0125	0.9988	0.9988
DecisionJungle	100.0000	0.0061	0.9966	0.9962
LogisticRegression	100.0000	0.0515	1.0000	1.0000
NeuralNetwork	100.0000	0.0361	1.0000	1.0000
OneVsAll	98.9659	0.1447	0.9826	0.9733
Bagging	98.4515	-0.0879	0.9891	0.9763
AdaBoost	88.4882	-0.3575	0.7693	0.8333

Classifier Algorithm Cross-Validation Evaluation Results

Accuracy Estimation of RandomForest on validation dataset

```
Standardized: 0.90206185567

Normalized: 1.0

Rescaled: 1.0

Binarized: 0.536082474227
```

RandomForest Classifier Performance on Transformed dataset

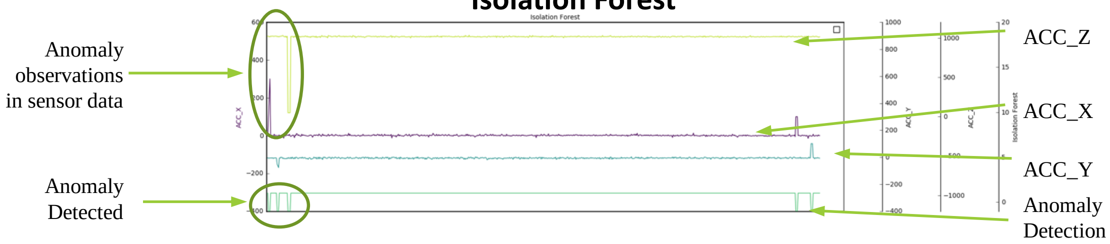
```
Best Params: {'min_samples_split': 2, 'bootstrap': True, 'min_samples_leaf': 1, 'n_estimators': 100, 'max_features': 4, 'max_de
pth': 10}
Best Score: 1.0
Best Estimator: RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
            max_depth=10, max_features=4, max_leaf_nodes=None,
            min_impurity_split=1e-07, min_samples_leaf=1,
            min_samples_split=2, min_weight_fraction_leaf=0.0,
            n_estimators=100, n_jobs=1, oob_score=False, random_state=None,
            verbose=0, warm_start=False)
```

RandomForest Algorithm Hyperparameter tuning results

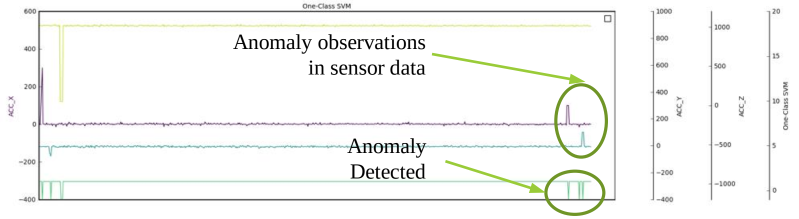
Machine Learning – Anomaly Detection

- Anomaly Detection
  - 5 algorithms evaluated
    - 3 Scikit-Learn (OneClassSVM, Isolation Forest, Robust Covariance)
    - 2 Azure machine learning specific (One Class Support Vector Machine, PCA Based Anomaly Detection)
  - One Class SVM found suitable for performing equipment fault detection

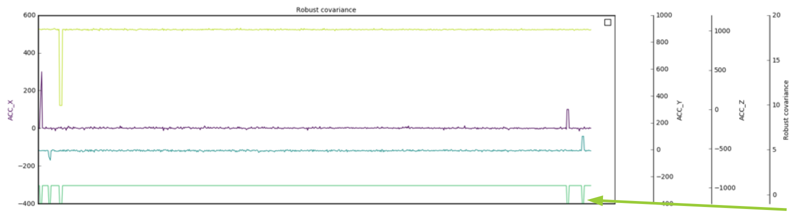
Isolation Forest

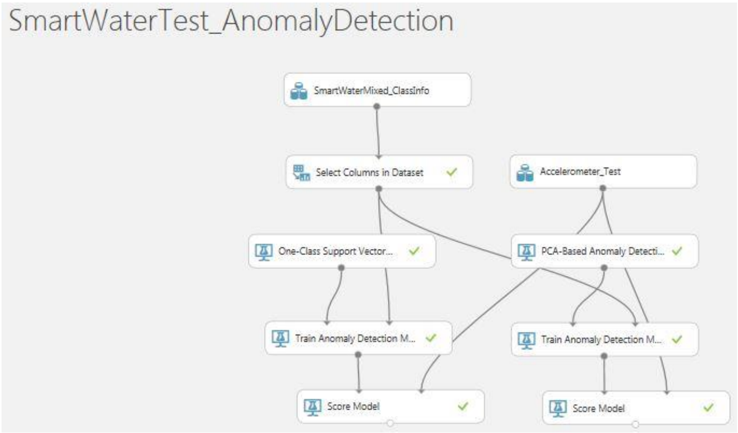


One-Class SVM



Robust Covariance





Further Enhancement

- Use Azure IoT Hub Connector to connect Meshlium Gateway to Microsoft Azure to support cloud-to-device messaging and per device connection authentication

- Extend visualization capabilities using chatbots to provide a natural language interface for querying sensor data information on mobile devices/ social media platforms

- Install sensors to measure rainfall in each reservoir to design an integrated system to channel water to different parts of Singapore based on min holding capacity based on the amount of rainwater collected

- Customize automatic trigger Alert for individual Water Catchment(Reservoir) e.g. Heavy Rainfall, monitor Algae and Bacteria Growth activity in reservoir

- Use solar panel installed on Reservoir water surface to power the Libelium Sensor and use remote CCTV to detect illegal fishing and water sports activities around reservoirs and facial recognition to trigger alert on illegal activities

- Implement Libelium Water Sensor System for HDB Rooftop Water Tanks to monitor water quality and trigger alert to town council Ops rooms to send cleaners to wash the water tank should the condition of drinking water falls below an unacceptable standards

Q&A