Data-driven Building Metadata Inference

Submitted in partial fulfillment of the requirements for

the degree of Master of Science in Building Performance and Diagnostics

June Young Park

School of Architecture
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA

Advisors:

Azizan Aziz, Assistant Research Professor

Bertrand Lasternas, Research Scientist

School of Architecture, Carnegie Mellon University

May. 2016

Abstract

Building technology has been developed due to the improvement of information technology. Specifically, a human can control and monitor the building operation by a number of sensors and actuators. The sensors and actuators are installed on every single element in a building. Thus, the large stream of building data allows us to implement both quantitative and qualitative improvements. However, there are still limitations to mapping between the physical building element and cyber system. To solve this mapping issue, last summer, a text mining methodology was developed as part of a project conducted by the Consortium for Building Energy Innovation. Building data was extracted from building 661, in Philadelphia, PA. The ground truth of the building data point with semantic information was labeled by manual inspection. And a Support Vector Machine was implemented to investigate the relationship between the data point name and the semantic information. This algorithm achieves 93% accuracy with unseen building 661 data points. Techniques and lessons were gained from this project, and this knowledge was used to develop the framework for analyzing the building data from the Gates Hillman Center (GHC) building, Pittsburgh PA. This new framework consists of two stages. In the first stage, we initially tried to cluster the data points by similar semantic information, using the hierarchical clustering method. However, the effectiveness and accuracy of the clustering method is not adequate for this framework. Thus, the filtering and classification model is developed to identify the semantic information of the data points. From the filtering and classification method, it correctly identifies the damper position and supply air duct pressure data point with 90% accuracy by daily statistical features. Having the semantic information from the first stage, the second stage figures out the relationship between Variable Air Volume (VAV) terminal units and Air Handling Units (AHU). The intuitive thermal and flow relationship between VAVs and AHUs are investigated at the beginning, and the statistical features clustering method is applied from the VAV discharge temperature data. However, the control strategy of this building makes this relationship invisible. Alternatively,

we then compared the similarity between damper position at VAVs and supply air duct pressure at AHUs by calculating the cross correlation. Finally, this similarity scoring method achieved 80% accuracy to map the relationship between VAVs and AHUs. The suggested framework will guide the user to find the desired information such as the VAVs – AHUs relationship from the problem generated by a large number of heterogeneous sensor networks by using data-driven methodology.

Table of Contents

LIST OF TABLE

LIST OF FIGURES

1. Introduction

While everyone agrees about the importance of data analysis, we have had to deal with the concern that building data does not always provide us with an ideal dataset. For example, whenever metadata is missing, the user has to manually infer the data type or ask the database manager. In the worst case, the meaning of the data is unknown. While advancements in information technology, such as the improved development of physical sensors and actuators has been beneficial, this information has not been mapped on a cyber system correctly and efficiently.

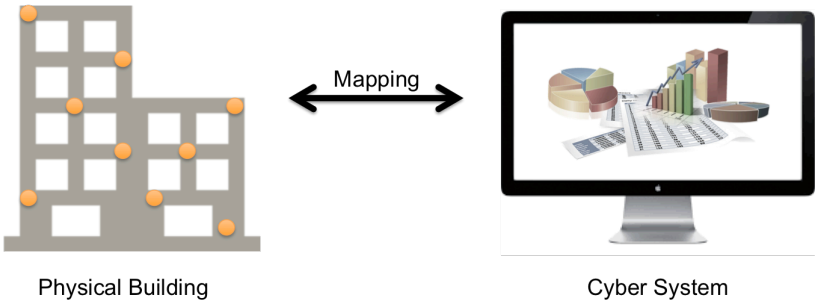


Figure 1 Cyber Physical System of building

There are three main characteristics of a building, which makes the mapping process difficult. The first reason is the huge size of the building. Normally one contains 3,000 data points for its operation. For example, building 661, Gates Hillman Center (GHC) and Mellon Institute (MI) have approximately 4,000, 15,000, and 14,000 points respectively. These numerous data points have multiple functions such as sensors, actuators, metering data or embedded virtual data points as the building control parameters. Thus, it is too large to manually handle these various data types. Secondly, the building has a relatively long life cycle, and this life cycle doesn't fit with other building equipment's life cycle (i.e. mechanical, lighting equipment, etc.). Because of this reason, different field engineers might label different point names when they reinstall or repair equipment with a certain time gap. This heterogeneity of the data point is a major impediment in finding the

generalized naming rules and the relationship among a large amount of data points. The last reason is derived from the interoperability issue from multiple Building Automation System (BAS) vendors of the building industry. For example, the National Institute of Standard and Technology found that the lack of the building interoperability standard wastes $15.8 billion annually in the US. For example, different vendors even describe a same variable air volume terminal discharge of air temperature sensor data in two different ways (i.e. vav room 3000 class room da temp or VAV-1 DAT). Due to this reason, the Department of Energy (DOE) published the Building Energy Data Exchange Specification (BEDES). It is a dictionary, which includes data format, definition and terminology about the building data stock of words. However, there are still limits in implementing this text schema in real building cases. In this regard, it is hard to find the universal rule of mapping the raw data points on the semantic information as cyber system type.

Such challenges stimulate the development of data-driven building metadata. The goal of this thesis is to develop an algorithm that mimics a well-experienced building data manager's intelligence to identify the relationship between Air Handling Units (AHU) and Variable Air Volume (VAV) terminal units. The dataset is extracted from the Building Automation and Control Network (BACnet) and Automated Logic BAS data points in a GHC building. The hierarchical clustering is firstly used to define the semantic information of the data point. This algorithm predicts the semantic information from data point names at the first stage. By changing the number of clusters, different levels of semantic information is predicted for each cluster. However, the first trial was not successful, we then proceed with new approach. The BACnet information filtering and data type classification method is developed to classify the desired data point type. Even though this filtering and classification model does not map all the data point on the cyber system, they are very effective and accurate when users know which data point type they specifically need for further analysis.

In the second stage, the algorithm also infers the mechanical functional relationship between VAVs and AHUs. Specifically, the relationship between VAVs and AHUs are studied in this thesis. This is because, most of the buildings use VAVs and AHUs as the main Heating, Ventilation, and Air Conditioning (HVAC) operation equipment, and the data points of VAVs and AHUs take up the majority of the whole building data points. Thus, this relationship is one of the most crucial parts for building data analytics. However, only a few buildings store that relationship information on BAS and people can only acquire that relationship information by interpreting the paper based mechanical drawing. The ground truth of the AHU and VAV relationship is identified through the manual inspection of BAS. This ground truth contains the physical connection between 287 VAVs and 6 AHUs in the GHC building. First and foremost, the exploratory data analysis is conducted by temperature and flow data at VAVs and AHUs. However, the strong relationship is not observed. Secondly, the VAV discharge air temperature is used as a feature variable and AHU reference number is used for prediction class values. The statistical features are extracted from the VAV discharge temperature at occupied and unoccupied time stamps. The hierarchical clustering algorithm is tested to differentiate the VAV performance to predict the assigned AHU. However, this clustering method cannot classify the AHU type due to the current control strategy of the GHC building. Changing the viewpoint to the equipment mechanical control relationship of VAVs and AHUs, the damper position data and supply air duct pressure data are queried from VAVs and AHUs respectively. By calculating the cross correlation of two signals, the similarity between those two signals is acquired. To consider the seasonal building behavior patterns, 9 month profiles are evaluated to identify the relationship between VAVs and AHUs.

The first stage essentially figures out the semantic information of the data points. For example, the user can infer the meaning of the data by BACnet information and historical database without having the informative data point names. This filtering and classification method ultimately provides the metadata for the individual building data points. The relationship information among data points also helps the

user to better understand the metadata structure. In the second stage, the algorithm specifically identifies the relationship between VAVs and AHUs. This data-driven inference approach reduces a significant amount of time for manual inspection and is a backup solution for when the user does not have the physical mechanical drawing.

2. Literature Review

2.1 Text data feature inference

To solve the heterogeneity of building data, several researchers tried to build efficient mapping methods. Bhattacharya et al. suggested automated metadata transformation by substring extraction language. The top-level expression of the language is the If-Then algorithm, and it extracts meaningful text by regular expressions. By learning from given primitive sensor metadata and common namespace sets from one building, they implemented this learned rule to 56 other buildings to evaluate the performance. 24 examples by the random generator are desired to qualify a large fraction of a commercial building. For attaining 90% accuracy, 85 examples are needed for building expert's manual inspection. Bhattacharya et al. firstly suggested that semi-automatic algorithm be used for predicting building semantic information by combining building expert's knowledge and their regular expression methodology. Schumann et al. tried to develop the text mining algorithm for predicting the format of building energy management software from primitive building data point names. They defined a building terminology dictionary in advance, which contains 504 acronyms and 125 markers, and score the similarity with 2,099 building management system raw data point names. Ultimately, the name with the highest score is labeled with raw building management system points. On average, a building expert needs to review 15 candidates of outcomes in this scoring method. Compared to the number of total labels, the user can reduce 7.5% for consideration by implementing their linguistic similarity computation method. Both literatures are the pioneers of building the metadata field. However, they only considered the text naming data features in a more linguistic approach. The general assumption for regular expression method is the text location and arrangement of data point name, and vendors have their own data point naming schema. Since Schumann et al. implemented their own building data terminology dictionary from their energy management software input formats, their text mining algorithm's performance depends on the dictionary that they used.

Thus, both still require the manual input from building experts and only are applied for a single vendor.

2.2 Numeric data feature inference

A different approach was also suggested from only considering the numeric sensor data reading from Gao et al. After cleaning the outliers and taking interpolation, the statistical features are extracted from BAS's time series data. Each data samples with statistical features are labeled by the Project Haystack metadata format. For each sample, individual tags are obtained by the binary classifiers, and the combined individual tags are considered as a complete Project Haystack tag. 8 different classification algorithms are evaluated using 20% of the training set and 80% of the testing set. Having the maximum F1 score on Random Forest classifier, the result indicates that more than 50% F1 score for half composite tags and more than 60% F1 score for half individual tags. To test out the applicability of this algorithm, Gao et al. used two different training data from December and June. As they expected, some individual tags such as 'cool', 'hot', 'pressure' show very low F1 score which means they are very sensitive to the weather. On the other hand, 'outside', 'air', 'fcu' and 'cur' seem less sensitive than the previous three tags above. Finally, they tested their algorithm by different train building. However, the result shows a very low performance and the Project Haystack convention issue is still a problem that needs to be solved. Even though the accuracy is the lowest among other metadata inference algorithms, they tried to build the Project Haystack format by multi-label classification. Having multiple prediction class values from one data point name, they built a more robust semantic information.

2.3 Text and numeric data feature inference

Several researchers consider both text naming and numeric sensor reading features to infer the metadata structure. To reduce the manual input from the domain expert and achieve high accuracy for predicting proper point type, Balaji et al. combined hierarchical clustering and a random forest classifier. Firstly they extracted unigram

features from data point names. The hierarchical clustering algorithm then calculates the Manhattan distance between building data point names and clustered them based on complete linkage of each cluster. Assuming the points within a cluster are of the same type, the building points are clustered by preprocessed string features. Then, domain experts label 10-point type on those clusters. Random forest classier is built based on these points. Random forest algorithm randomly picks several decision trees from the clustering result and evaluates the accuracy of the clustering performance. Evaluating the probability, 0.9 is the correct threshold and if the probability is less than 0.2, a domain expert is asked to label new point type. Additionally, numerical time series data is considered in the input feature. The episode of time series data consists of scaled, pattern, shape and texture based features. Those four features aid the defect of string data. Ultimately Balaji et al. achieved 98% accuracy requiring 27% fewer ground truth labels than using the regular expression method. Compared to only considering text feature methods, Balaji et al. reduced the manual input and achieved higher accuracy. However, their algorithm is still only applied to a single vendor, which is Johnsons control. Thus this algorithm has a generalization issue with other BAS vendors' naming schemata.

Hong et al. suggested using an automatic mapping algorithm, which is called transfer learning. It learns a set of statistic classifiers of the metadata from a labeled building and adaptively integrates those classifiers to a different unlabeled building from a different BAS vendor. This transfer learning algorithm takes both text data from point name and statistical numerical data from sensor reading. Both numeric and text data have their own respective implementation benefits. The numeric data is more consistent among different buildings, but it is a poor indicator of sensor type. On the other hand, the text data is a good indicator of metadata structure however it might not be consistent cross buildings. To utilize both advantages of two data types adaptively, this transfer learning algorithm assign higher weights to classifiers whose predictions on an instance's neighbors in the target building are more consistent with the text feature defined within a cluster. Three buildings' 2500 sensors are evaluated through 7 days data. The best base line classifier is selected as

Random forest, which shows 63% accuracy. On the other hand, the transfer learning algorithm achieved 85% of accuracy from the different train and test BAS vendors (i.e. Trane, Apogee, Barrington). Hong et al. tried to solve the interoperability issue by understanding the characteristics of text and numeric data.

2.4 Literature review summary

Even though building metadata inference is a new research topic in this field, several works were conducted with different feature extraction settings. Figure 2 shows the general objective of 5 different literatures. They are slightly different from the feature selection method but all of them tried to infer the metadata framework for building data. In 2014, Bhattacharya et al. and Schumann et al. started to build metadata by their own text mining approaches (i.e. regular expression, similarity scoring). However, their algorithms still require manual input labeling and have interoperability issues with different vendors' naming schemata. Different from two text mining algorithms, Gao et al. only considered statistical features by numeric sensor reading data to predict Project Haystack's format. They build the semantic information by multiple predictions. Also, they conducted learning ratio and weather sensitivity studies. On the other hand, Baljaji et al. and Hong et al. considered both text and numeric features to predict their own semantic format. Balaji et al. achieved the maximum accuracy among these algorithms and Hong et al. only tried to apply their algorithm to solve the interoperability issue among three different vendors' building data point naming schemata.

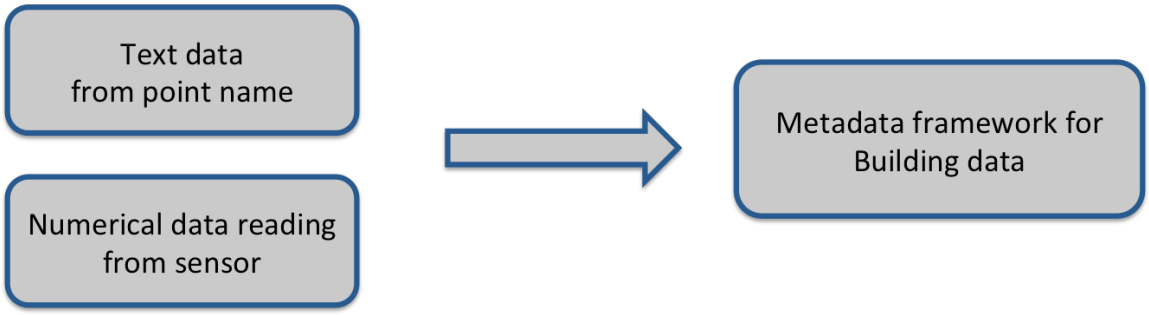


Figure 2 General objective of the literatures

Table 1 Summary table for literature review 1

Paper	Data type	BAS Vendor(s)	Algorithm
Bhattacharya et al.(2014)	Text data	Unknown	Clustering, Regex
Schumann et al. (2014)	Text data	IBM	Dictionary based similarity scoring
Gao et al. (2015)	Numeric data	Automated logic, IBM	Multi-label classification by Random forest
Balaji et al. (2015)	Text data Numeric data	Johnsons control	Clustering, Random forest
Hong et al. (2015)	Text data Numeric data	Trane, Apogee, Barrington	Adaptive transfer learning

Table 2 Summary table for literature review 2

Paper	Prediction output	Accuracy	Note
Bhattacharya et al.(2014)	Project haystack format	Semi automated	Many manual input
Schumann et al. (2014)	Own semantic format	Semi automated	Pre-defined dictionary
Gao et al. (2015)	Project haystack format	75%	Learning ratio study Statistical numeric feature
Balaji et al. (2015)	Own semantic format	98%	The highest accuracy Episodic numeric feature
Hong et al. (2015)	Own semantic format	85%	Test to other BASs adaptively learning

3. Test bed and Dataset

3.1 Gates Hillman Center

The Gates Hillman Center (GHC) is tested for the data-driven metadata inference algorithm. It is the main building of the Carnegie Mellon School of Computer Science, Pittsburgh, Pennsylvania, USA. The 217,00 sq.ft. floor area with nine floors contains 310 offices, 32 laboratories and 11 conference rooms and other minor rooms. The Automated Logic manages the building data in GHC building. Figure 4 ~ Figure 6 show the BAS webpage. Approximately, 15,000 data points with BACnet object and property information are detected in this building, and the historical data is stored from May 2015 in OSIsoft PI system. There are 287 VAV terminal units and 6 AHUs in the Gates building. And each VAV is connected to a specific AHU.

Figure 3 Gates Hillman Center

Figure 4 Gates building BAS 1

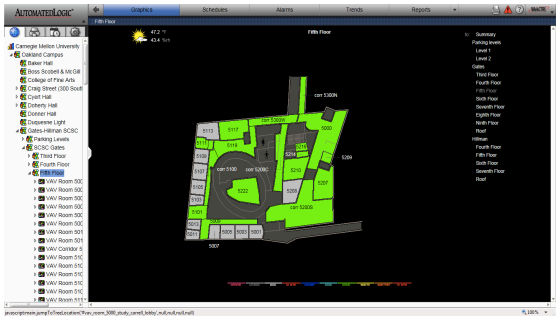


Figure 5 Gates building BAS 2

Figure 6 Gates building BAS 3

3.2 Data querying

The most important consideration for the data-driven framework development is how to acquire the dataset with ground truth. Building data flows in three steps. Firstly, based on Figure 7, building data from equipment are connected to BACNET panels. Subsequently, all the data points in these BACNET panels are transferred to the BAS, which controls and manages these series of data points. Additionally, the OSIsoft PI system is installed for data collection and storage purposes. The GHC building is controlled and managed under Automated Logic from United Technology Research Center. This BAS environment provides a clear understanding of how building data is managed and the semantic meaning of the data points. As previously explained, the historical dataset of GHC building is stored in OSIsoft PI system, thus we can efficiently access the numeric data from the GHC building system. Thus, as long as the future test building is controlled under BACnet protocol and the numerical data is stored from the BAS, we can evaluate the suggested framework to build the metadata for a different building.

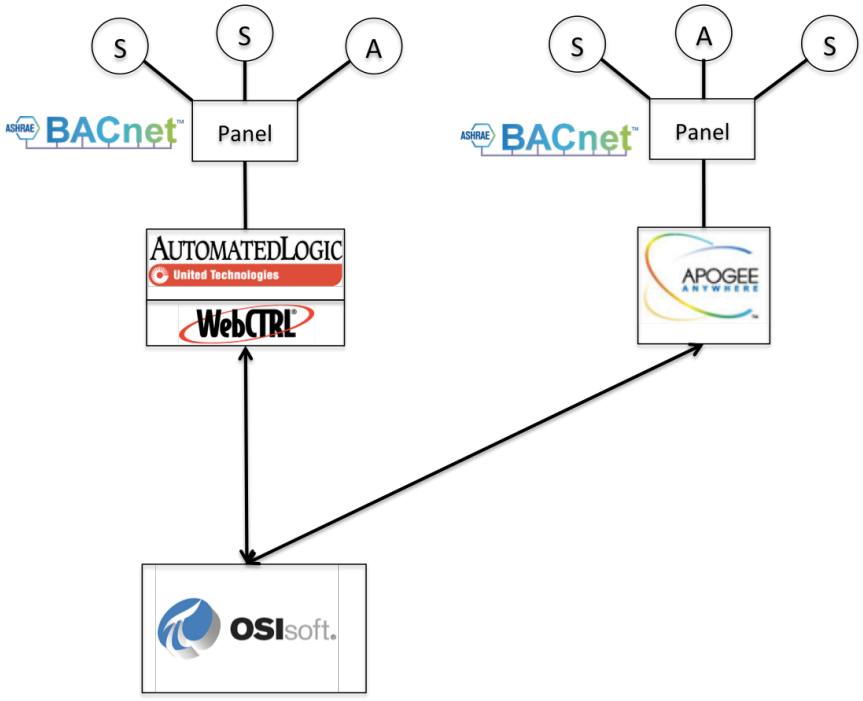


Figure 7 Data querying process

4. Methodology

4.1 Data preprocess

There are two types of data we can infer the building metadata from. The first type of building data is text data from point names. When vendors install the sensor or actuators, the field engineers named each data point by their own naming rule and each BAS vendor has its own naming schema and utilizes them for the system. Since normally field engineers name the data points names by the semantic hierarchy and application type, we can infer the hierarchical structure of metadata with a certain order (i.e. building-equipment-location-sensor type). However, once again the naming ordering rules are different by vendors and they use different terminologies. The second data type is the time-series data from sensor and actuator readings. This numeric data is more consistent even with different BAS vendor buildings. However, it is hard to indicate the sensor type by numeric data itself. Both text and numeric data types have their own advantage and disadvantage. Thus, this algorithm employs two data types, selectively. The text data feature is used for deriving the semantic information and the numeric data feature is utilized for inferring the relationship among data points.

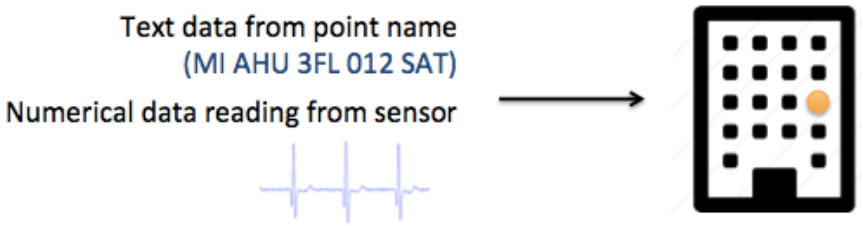


Figure 8 Two different data types acquired

Firstly, the text data is acquired from the BACnet panel and the BAS in the Gates building. There are 10 types of text data that are extracted. However, the unit of measurement, BACnet object name, and BAS vendor given name, are only selected as meaningful features. First and foremost, the unit of measurement is the obvious clue for classifying the data acquisition type by units (i.e. F, cfm, ppm, min). Secondly

the BACnet object name, which is made by field engineers also shows the data acquisition type (i.e. flow, temp, co2) in a direct way. Finally, the vendor given name contains normally 5 more words and indicates the location and equipment type in a comprehensive way. However, not all of the data points contain these three different information types. After eliminating the data points, which do not completely contain three information types, 7,434 data points are queried at the end.

Table 3 10 sample points from Gates building

	A	B	C	D	E	F	G	H
1	Location	Control Program	Name	Value	Object ID	Device ID	Object Name	Path
2	/Carnegie Mellon University/Oakland Campus/Gates-Hillman SCSC/SCSC Gates/Third Floor	Room 3000 Classroom VAV	Flow Control / Flow Input	682.0 cfm	AI:1	DEV:2401203	flow_input_1	#vav_room_3000_classroom/air_flow/flow_input
3	/Carnegie Mellon University/Oakland Campus/Gates-Hillman SCSC/SCSC Gates/Third Floor	Room 3000 Classroom VAV	Zone Temp / Zone Temp	70.1 F	AI:2	DEV:2401203	zone_temp_1	#vav_room_3000_classroom/lstat/zone_temp
4	/Carnegie Mellon University/Oakland Campus/Gates-Hillman SCSC/SCSC Gates/Third Floor	Room 3000 Classroom VAV	Zone Temp / Override Time Remaining	0.0 min	AV:1	DEV:2401203	override_time_remaining_1	#vav_room_3000_classroom/lstat/override_time_remaining
5	/Carnegie Mellon University/Oakland Campus/Gates-Hillman SCSC/SCSC Gates/Third Floor	Room 3000 Classroom VAV	Discharge Temp	56.3 F	AI:3	DEV:2401203	da_temp_1	#vav_room_3000_classroom/da_temp
6	/Carnegie Mellon University/Oakland Campus/Gates-Hillman SCSC/SCSC Gates/Third Floor	Room 3000 Classroom VAV	Zone CO2	455.0 ppm	AI:4	DEV:2401203	zone_co2_1	#vav_room_3000_classroom/zone_co2
7	/Carnegie Mellon University/Oakland Campus/Gates-Hillman SCSC/SCSC Gates/Third Floor	Room 3000 Classroom VAV	OCC Sensor	On	BI:1	DEV:2401203	occ_sensor_1	#vav_room_3000_classroom/occ_sensor
8	/Carnegie Mellon University/Oakland Campus/Gates-Hillman SCSC/SCSC Gates/Third Floor	Room 3000 Classroom VAV	HWV Command	8.3 V	AO:1	DEV:2401203	hwv_command_1	#vav_room_3000_classroom/hwv_command
9	/Carnegie Mellon University/Oakland Campus/Gates-Hillman SCSC/SCSC Gates/Third Floor	Room 3000 Classroom VAV	Airflow	685.87 cfm	AV:2	DEV:2401203	flow_1	#vav_room_3000_classroom/flow
10	/Carnegie Mellon University/Oakland Campus/Gates-Hillman SCSC/SCSC Gates/Third Floor	Room 3000 Classroom VAV	Airflow Setpoint	1200.0 cfm	AV:3	DEV:2401203	flow_sp_1	#vav_room_3000_classroom/flow_sp
11	/Carnegie Mellon University/Oakland Campus/Gates-Hillman SCSC/SCSC Gates/Third Floor	Room 3000 Classroom VAV	Cool Request	0	AV:4	DEV:2401203	cool_request_1	#vav_room_3000_classroom/cool_request

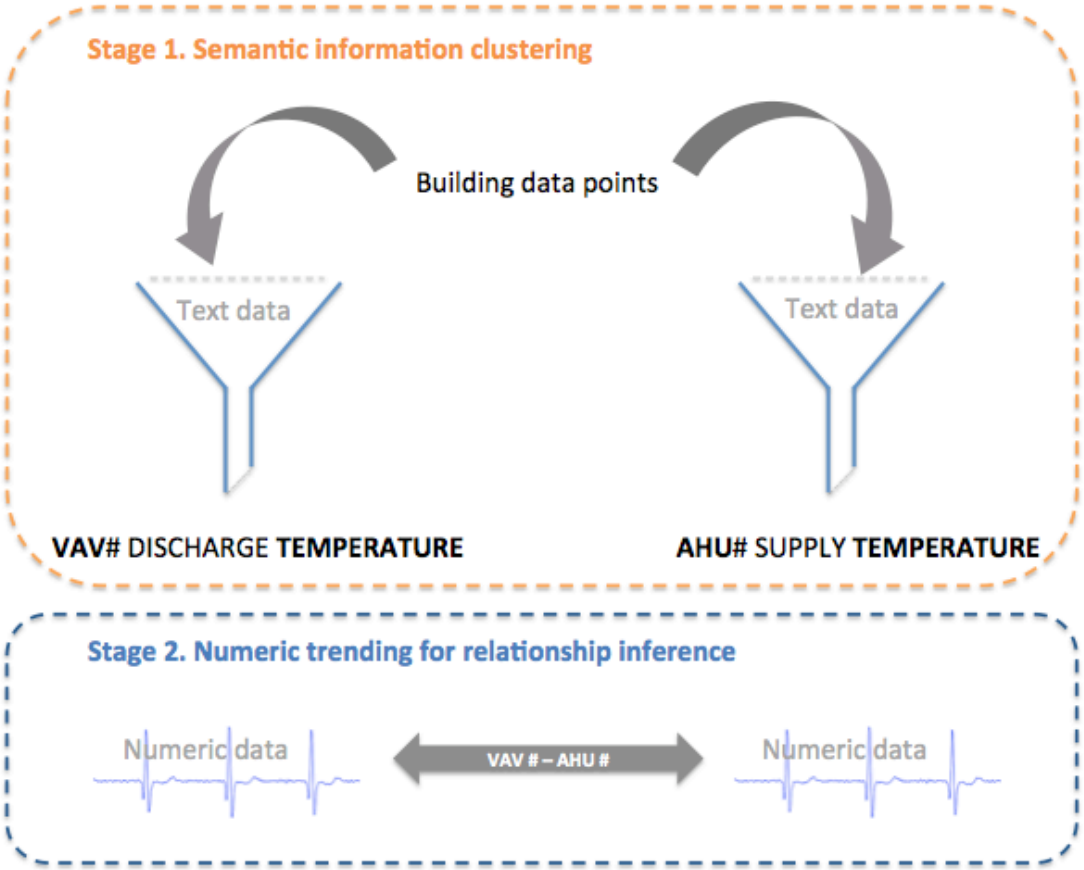


Figure 9 Two main stages of the framework

4.2 Semantic information clustering

Having the text data from data point names, we can extract the unigram features. In this feature extraction process, numerous single words from the whole data point names are extracted without the numbers and special characters. This is because, the numbers and special characters are used for equipment reference or arrangement of the point name, thus it might derive the bias error. After extracting unigram features, each data point name is converted to binary format, and this binary data basically count the occurrences of the features. Thus, we can build the binary matrix format, where each column represents the occurrences of features in binary format and, each row shows the data points. Figure 10 shows the example of the feature matrix. Having binary variables of each data point name, we can now calculate the distance among data point names. The Euclidian distance function, and complete linkage calculation method are implemented for hierarchical clustering. Since this Euclidean function is calculating the distances in the same dimension space, we can interpret the shorter distance as the higher similarity. For the distance between clusters, the complete linkage calculation method selects the furthest distance between two data points in each cluster. By doing so, we can group the data points by similarity and also differentiate each group by non-similarity.

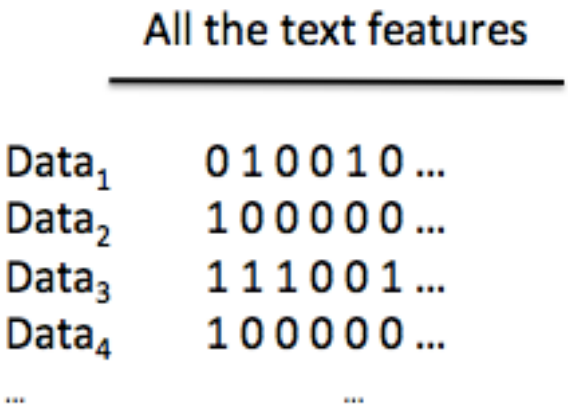


Figure 10 Feature matrix

$$\textbf{Equation 1}.\ Euclidean\ distance(data\ point1, data\ point2)$$
$$= \sqrt{(data_{X1} - data_{X1})^2 + (data_{X2} - data_{X2})^2 + \ (data_{X3} - data_{X3})^2 + \cdots}$$

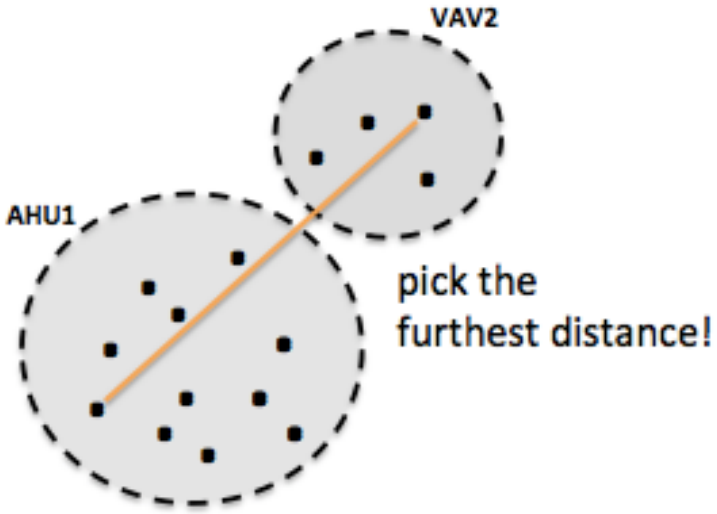


Figure 11 Complete linkage between two clusters

4.3 Data Point type classification

4.3.1 BACnet information filtering method

The Building Automation and Control Network (BACnet) is the data communication protocol, which is the most commonly used standard in the building automation and control industry. The data points in the GHC building are also managed under the BACnet protocol. As the BACnet follows the object-oriented nomenclature, the BACnet standard have 6 major objects with required properties. (Table 4 & Table 5). After querying the BACnet data points from the control panel, we can have the list of all the data point objects with required properties. The inputs are data coming into the control panel from the sensor and switch, and the outputs are data coming out from the control panel to control the actuator. Additionally, control parameters are pre-assigned or calculated within the control panel. Thus, we can filter the data points by the usage and specific property types. For example, all the temperature sensor data is acquired through filtering with Analog Input and the unit of measurement (degrees-Fahrenheit).

Table 4 The 6 major object in BACnet standard

Object	The usage
Analog Input	Sensor input
Analog Output	Control output
Analog Value	Control parameter or setpoint
Binary Input	Switch input
Binary Output	Relay output
Binary Value	Control parameter

Table 5 The required properties for BACnet object

Property	Example usage
Object Identifier	Analog Input #1
Object Name	AI 01
Object Type	Analog Input
Present Value	68
Status Flags	In Alarm
Event State	Normal
Out Of Service	False
Units	Degrees-Fahrenheit

4.3.2 Data-driven classification algorithm

Even though we can reduce the candidates for our data point semantic classification by the BACnet information filtering method, there could be some data points that have identical objects and properties. For example, there are four different pressure sensors in different locations within a single AHU. Thus, the data-driven classification algorithm is developed to differentiate the specific pressure sensor that a user needs.

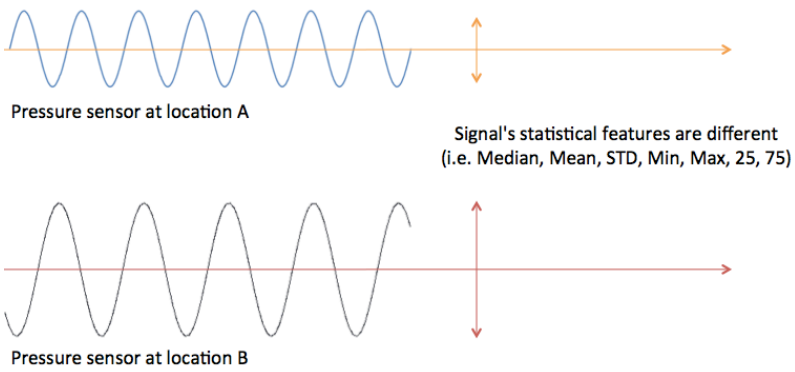


Figure 12 Hypothesis behind the classification

The data-driven classification algorithm characterizes the different signals by statistical features. Although two sensors measure the same component, the characteristics of signals are slightly different from one another based on the sensor location. As a feature selection process, 7 different statistical features (i.e. median, mean, standard deviation, minimum, maximum, 25% quartile and 75% quartile) are extracted. For model selection, the Random Forest algorithm is selected. Basically, the Random Forest algorithm builds the decision trees by random samples, and evaluates the classification probability from multiple randomly generated decision trees. Since this method classifies the sensor type based on the probability from bagging by multiple random decision trees, we can avoid the over-fitting problem. To develop the classification for the damper position, 5 days of statistical features are extracted from 297 VAVs and for the supply air duct pressure, 31 days of statistical features are extracted from 6 AHUs.

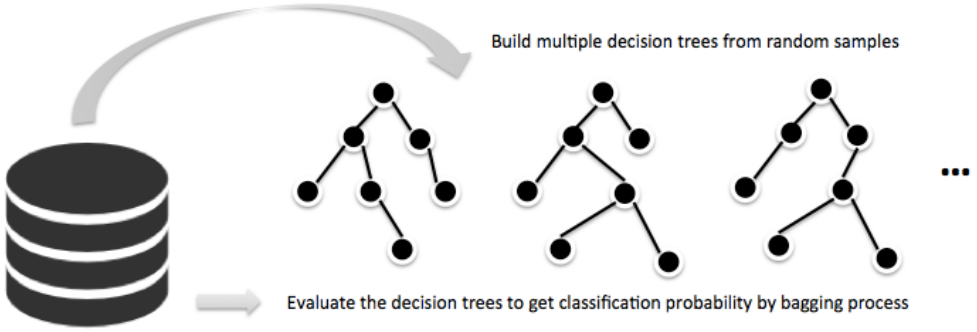


Figure 13 Random Forest algorithm

4.4 Functional relationship inference

4.4.1 Building system description

In the Gates part of the GHC building, 297 VAVs operate by 6 different AHUs. All the 297 VAV terminal units serve the air on the 3rd floor to the 9th floor, and exact VAV locations are mapped in Figure 14 ~ Figure 20. Table 6 contains the locations of AHUs and the number of VAVs that each AHU serves. Essentially, the GHC building is divided into three main programs in terms of control logic. Firstly, the public spaces (i.e. kitchen, corridor, project room, lobby) require non-stop operating everyday. Secondly, the classroom units have their own predefined schedules for operation. The last program, which comprises the largest portion of the program in this building, is the office unit. And office units in the GHC building is operated by occupancy and each office unit user can change their setpoint by their preference. Additionally, each VAV terminal unit controls both the air flow rate and temperature by damper and reheat coil respectively. The general assumption of the functional relationship inference methodology is primarily developed by considering the specific characteristics of the GHC building.

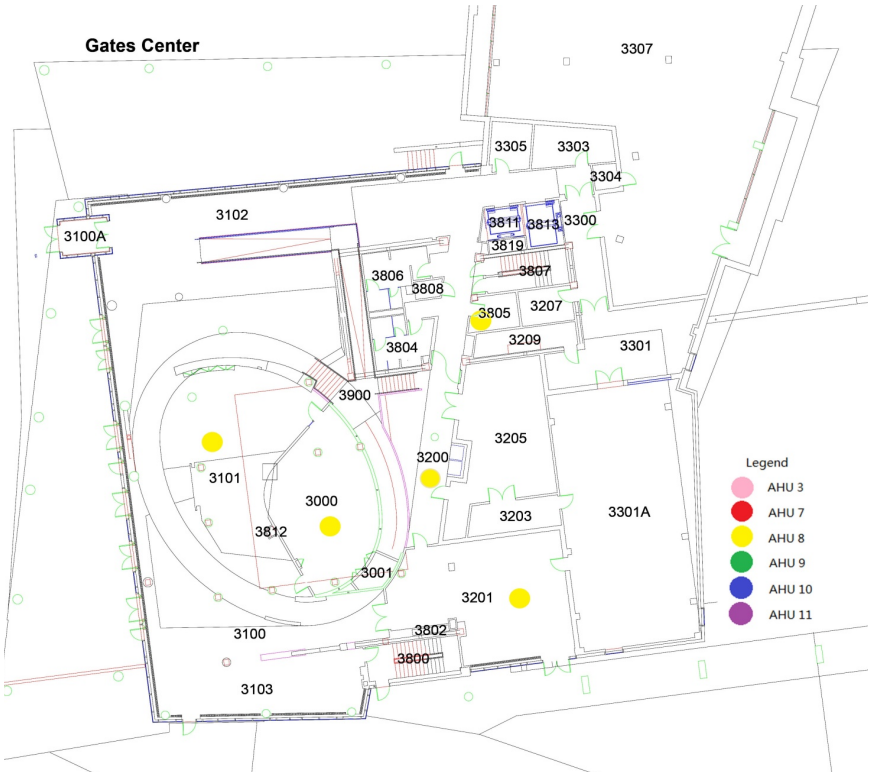


Figure 14 VAV locations on the 3rd floor

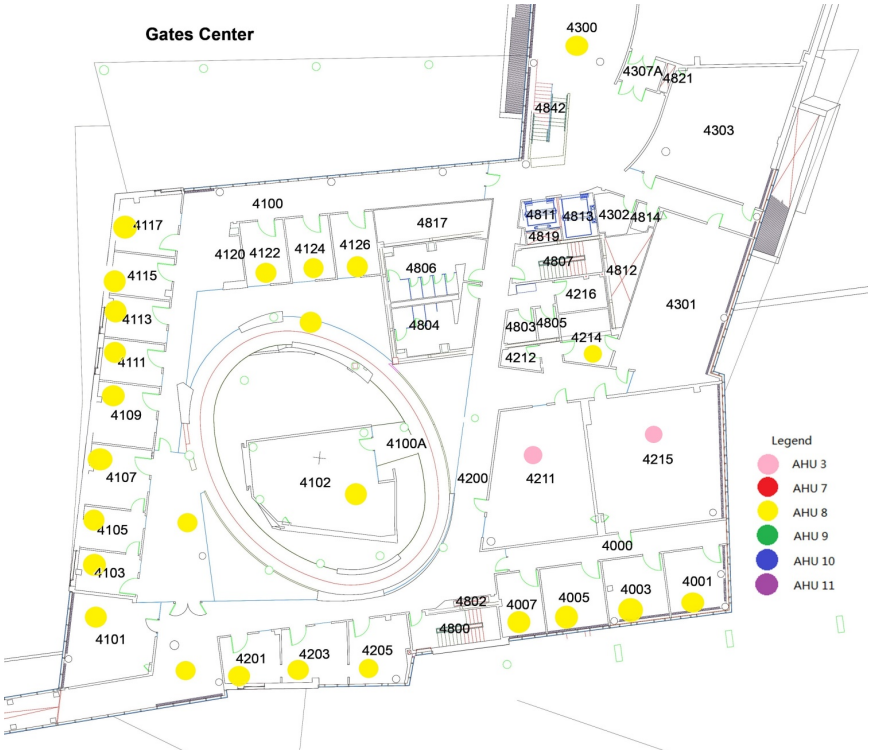


Figure 15 VAV locations on the 4th floor

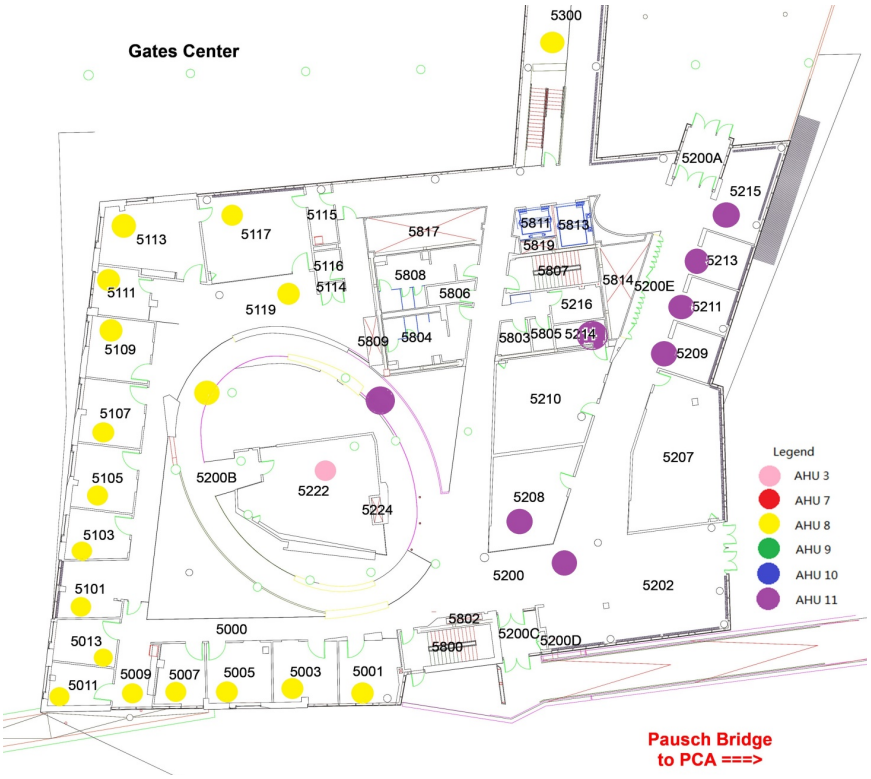


Figure 16 VAV locations on the 5th floor

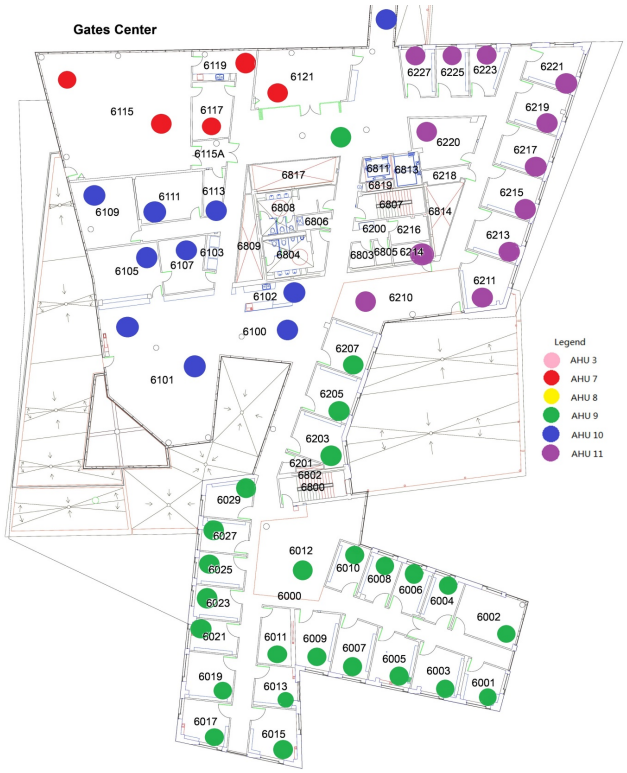


Figure 17 VAV locations on the 6th floor

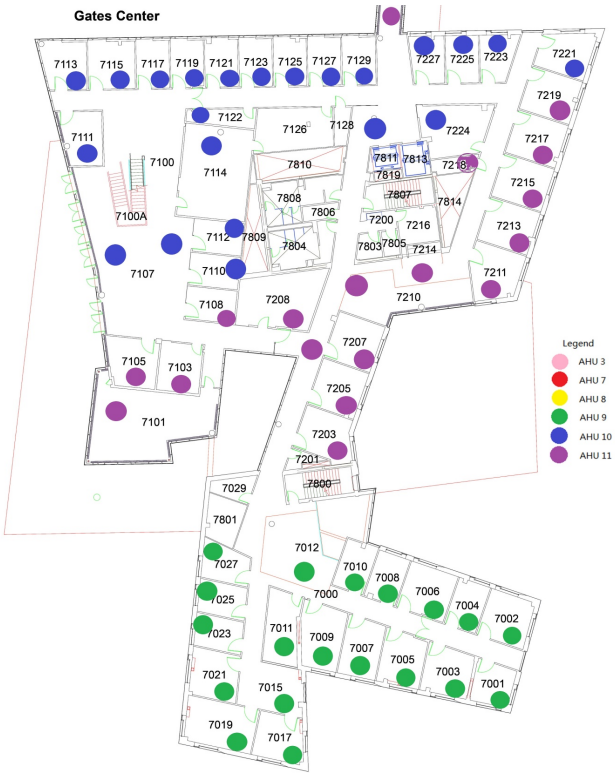


Figure 18 VAV locations on the 7th floor

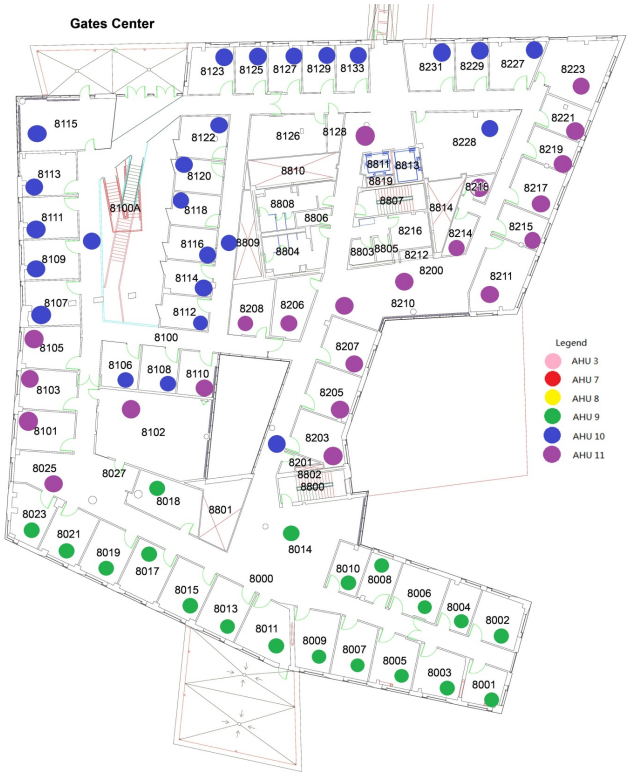


Figure 19 VAV locations on the 8th floor

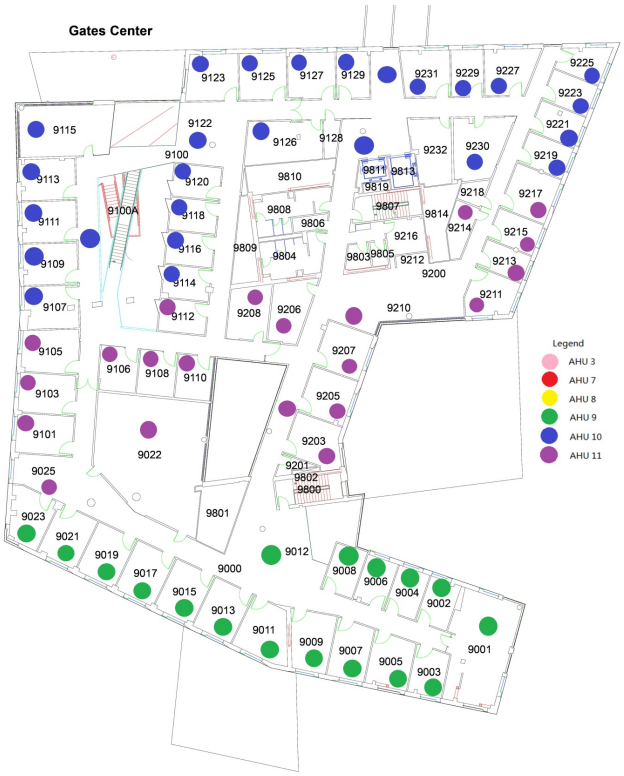


Figure 20 VAV locations on the 9th floor

Table 6 Detail information for AHUs

Air Handling Unit	Location	Number of VAVs
AHU3	3rd floor	3
AHU7	3rd floor	5
AHU8	3rd floor	53
AHU9	Rooftop floor	65
AHU10	Rooftop floor	84
AHU11	Rooftop floor	73

Table 7 Control strategy in GHC building

Zone program	Control strategy
Public area (kitchen, corridor, project room)	Operating everyday
Class room	Unoccupied mode (12:00 am ~ 7:00 am) Occupied mode (7:00 am ~ 10:00 pm) Unoccupied mode (10:00 pm ~ 12:00 am)
Office unit	Occupancy sensor based control Occupants can change setpoint

4.4.2 Exploratory numeric trending method

An exploratory method is implemented for investigating the intuitive relationship between VAVs and AHUs. Three different types of data are collected from VAVs and AHUs. From VAVs, discharge air tempereature, flow and damper position are collected in every 5 minute interval. Also, supply air temperatuer, flow and static pressure are gathered from AHUs. The detail locations of the data point at VAVs and AHUs are highlighted in the orange box in Figure 21 and Figure 22. The numeric trending data is sampled on a typical summer and winter day in terms of building occupancy and outdoor weather conditions. To consider fully occupied building operation, the first days of Fall and Spring sememster are selected (i.e. Fall semester - August 31st 2015, Spring semester - January 11th 2016). The air temperture and flow are the most visible air distribution relationship of VAVs and AHUs, because the higher temperature or flow rate from the AHU is distributed into the higher temperature or flow rate at the VAV terminal unit as well. Additionally, damper position from VAVs and static pressure from AHUs are selected because they have a mechanical control relationship. For example, if the VAV increases the opening of the damper, the static pressure, with a short time lag, at AHU is going to decrease. The damper opening would be the trigger for increasing fan flow rate.

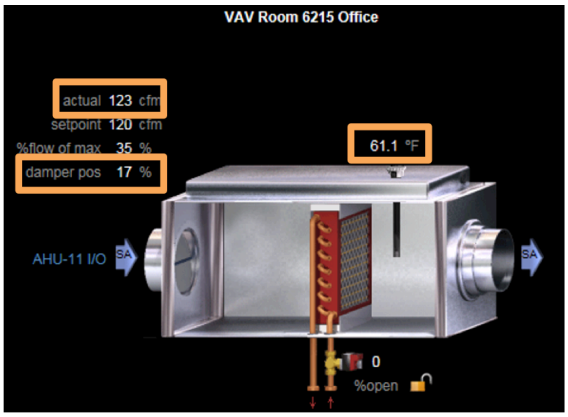


Figure 21 Data point locations at VAV unit

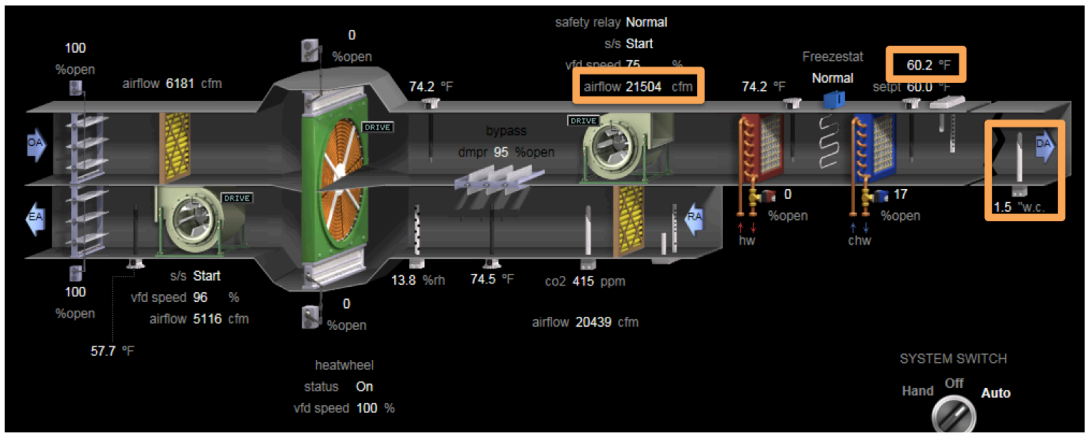


Figure 22 Data point locations at AHU

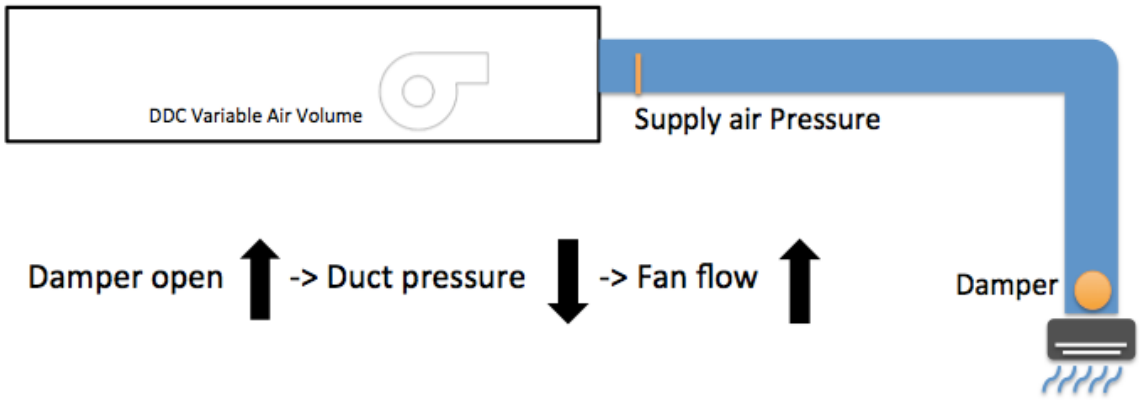


Figure 23 Relationship between damper position and duct pressure

4.4.3 Statistical features clustering method

To identify the relationship between VAVs and AHUs, the statistical features (i.e. mean, median, standard deviation, quartiles, min, max) are extracted from the VAV discharge temperature sensor data. The functional relationship inference is conducted with an unsupervised clustering algorithm in this pilot study. We group the VAVs based on their temperature trend, and compare how many VAVs are correctly assigned to each AHU. Even though this pilot study only aims to count the number of VAVs on each AHU, it would be a great stepping-stone for feature extraction for future study. We can simply count how many AHUs are installed in the buildings, and this number is used as the clustering number. Similar to the semantic information clustering algorithm, we can thus calculate the distances from the statistical features of the VAV discharge temperature sensor data. Therefore, we can group the VAVs by the assigned AHUs. For the statistical features clustering method, the historical data for discharge air temperature (November 18th, 2015) is queried by a 6-minute interval from 287 different VAVs. Since the temperature trends are different in occupied (9am ~ 5pm) and unoccupied time stamps (12am ~ 9am and 5pm ~ 12pm) by control logic, the queried time-series data is divided into two-time stamps.

4.4.4 Similarity scoring by cross correlation

To compute the similarity between two signals, we implemented a cross correlation method. Basically cross correlation is the dot product of two vectors, and two signals are represented by two vectors for this calculation. The calculation result of cross correlation indicates similarity for some time lags. The range for this similarity score can be from the minimum value -1.0 (negative correlation) to maximum value 1.0 (positive correlation), and the similarity score 0 means that it is impossible to find the correlation between two signals. The damper position from 297 VAVs and supply air side static pressure at 6 AHUs are selected for this similarity scoring method because, as we previously discussed in the prior section, the damper position and the static pressure have a negative correlation. Six different profiles are

extracted from those two signals, and the detail sampling date is tabulated on Table 8.

Table 8 Six profiles for cross correlation

	Summer	**Winter**
Daily (5 minute interval)	Aug.31 – Sep.1 2015 (288 data points)	Jan.11 – Jan.12 2016 (288 data points)
Weekly (5 minute interval)	Aug.31 – Sep.7 2015 (2,016 data points)	Jan.11 – Jan.18 2016 (2,016 data points)
Monthly (5 minute interval)	Aug.31 – Sep.28 2015 (8,064 data points)	Jan.11 – Feb.8 2016 (8,064 data points)

$$Equation 2.\ Cross\ correlation\ function$$

$$(f \star g)[n] = \sum_{m=-\infty}^{\infty} f[m]\ g[m+n]$$

By computing the similarity scores between the single VAV damper position signal and 6 different AHUs static pressure signals, a specific VAV can have 6 similarity scores from the calculation. The algorithm predicts the AHU of the highest absolute scoring value. The reason why we take the absolute value of the score is that the relationship between damper position and static pressure has a negative correlation and we want to compare the similarity even though they have the negative relationship.

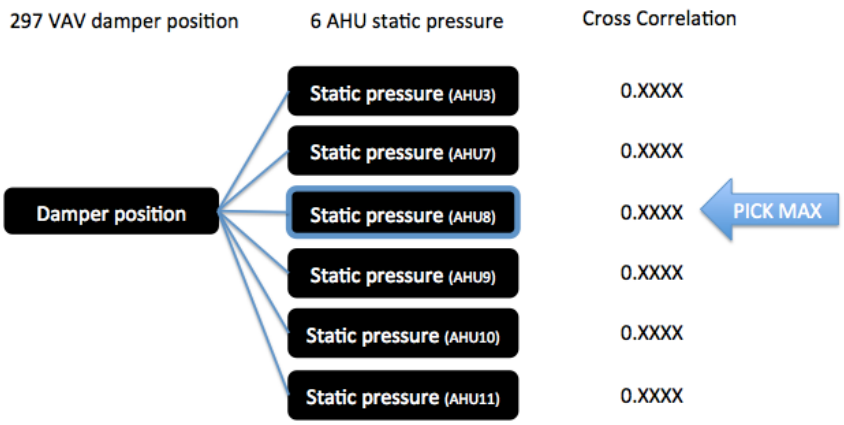
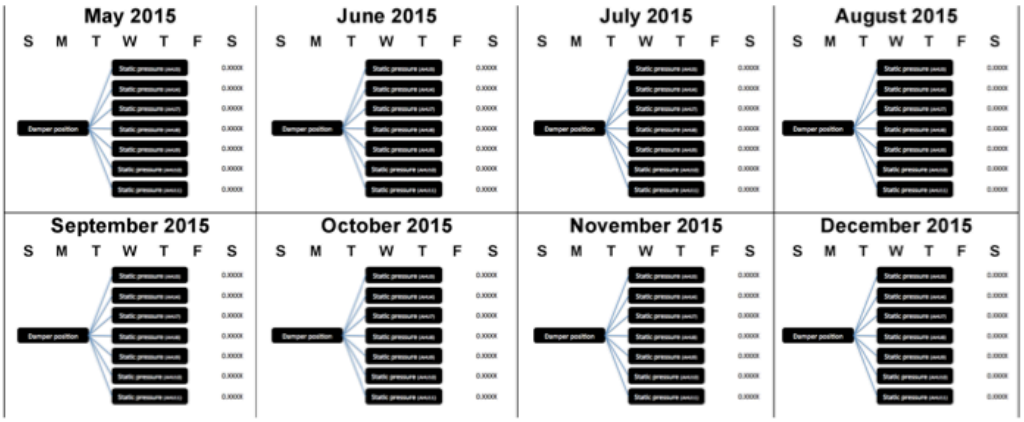
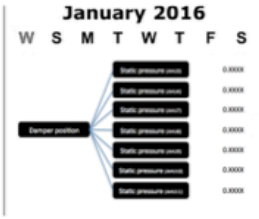


Figure 24 Similarity scoring process

4.4.5 Multiple scores prediction method

The variation of the signal profile is the important factor for investigating the cross correlation function of two signals. Even though the control schedule is repetitive throughout a year, the outdoor weather conditions would be the primary trigger for the variation of signal profiles. Thus, the experiment is extended for 9 month profiles to cover different weather conditions (i.e. 3 months – swing season, 3 months – cooling season and 3 months – heating season). By extending the range of experiments for May 2015 to January 2016, the algorithm can cover four different seasonal factors (i.e. Spring, Summer, Fall and Winter). Firstly, 9 different monthly profiles are collected with 5 minute intervals. By computing the similarity scores with 9 months, a single VAV damper position signal can have 9 different predicted AHUs for each month. The most frequently predicted AHU from 9 months is selected as the final prediction of this algorithm. If the multiple AHUs are tied in the prediction result, the AHU with the higher score is the prediction.





	May	June	July	Aug	Sep	Oct	Nov	Dec	Jan	Prediction
VAV	AHU3	AHU7	AHU8	AHU8	AHU8	AHU7	AHU3	AHU9	AHU8	**AHU8**

Figure 25 Multiple scores prediction method

5. Evaluation

5.1 Semantic information clustering

Having 7,435 data points, we can extract 238 unique unigram features (i.e. cfm, flow, input, vav). Each data point is evaluated by the occurrences of 238 text features, and the result is represented in binary format. Thus, we can build a binary feature matrix (7435 data points by 238 text features). The 2,000 data points are evaluated for this experiment. The detailed experiment procedure is explained in Figure 26.

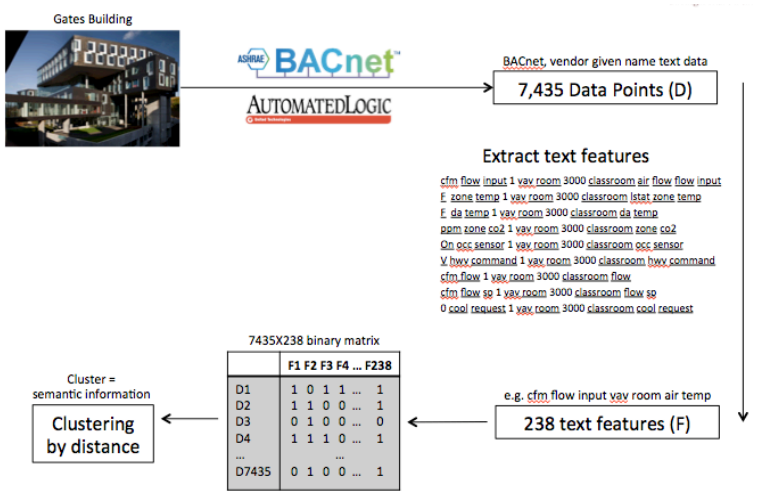


Figure 26 Text feature extraction procedure

Based on the calculation method described in Equation 1 and Figure 11, the 2,000 data points are clustered together. In Figure 27, the x-axis represents the individual data points, and they are clustered by the distance index. The highest clustering number is one, which is the Gates building, and the lowest clustering number is 2,000, which is the individual data points. The clustering numbers of 10, 50 and 249 were evaluated as the parametric study to investigate the relationship between clustering number and semantic information type.

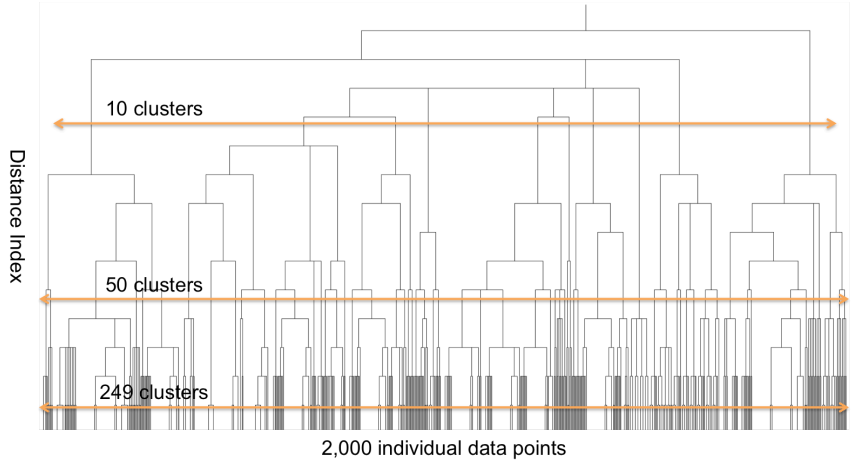


Figure 27 Dendogram result of hierarchical clustering

From Figure 28 and Table 9, if the 2,000 data points are grouped by 10 clusters, each cluster represents the major system categories (i.e. AHU, VAV, Metering). By the manual verification of the clustering result, 4 different unique major system types are clustered but assigned to multiple clusters. For example, cluster number 0, 1, 2, 4, 5 and 9 shows all VAV semantic information. Assuming the probability of correct clustering at 70%, we can achieve 80% accuracy for a 10 clustering result. And the accuracy for each cluster is manually inspected by a simple if-then code (i.e. if the text data contains VAV, then it is correctly clustered for the VAV cluster).

Table 9 Ten clustering result 1

Clustering Number = 10	
Category of semantic	Major system (AHU, VAV, Electrical Metering, Water)
Number of unique semantic	4
Accuracy of total clustering	8/10 = 80% ($P_{correct}$ = 0.7 for individual cluster)

Table 10 Ten clustering result 2

Cluster	Data points (EA)	Semantic Information	Accuracy (%)
0	957	VAV	95.8
1	141	VAV	87.2
2	73	VAV(override)	84.9
3	**437**	**AHU**	**69.6**
4	160	VAV(error)	77.5
5	**87**	**VAV(alarm)**	**67.8**
6	68	Metering	100
7	34	Water system	100
8	20	Metering	100
9	23	VAV(lobby)	100

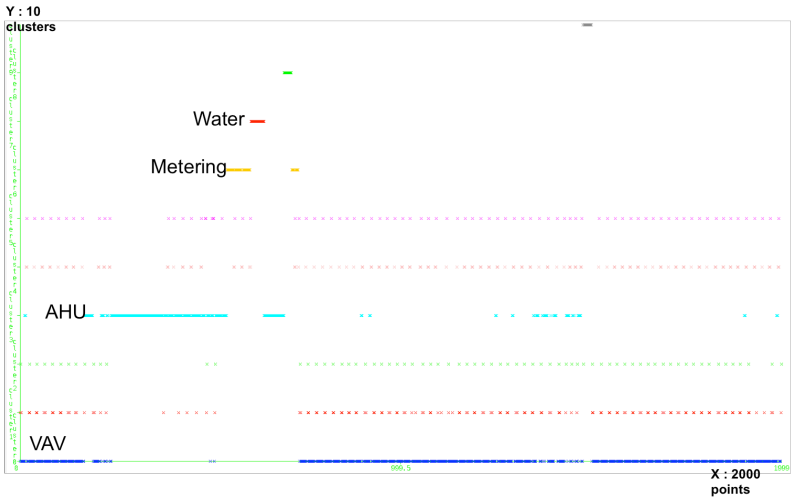


Figure 28 Clustering result (10)

From Figure 29 and Table 11, if the 2,000 data points are grouped by 50 clusters, each cluster represents the acquisition type categories (i.e. temperature, flow, CO2). By the manual verification of the clustering result, 18 different unique major system types are clustered, but assigned to multiple clusters. Assuming the probability of correct clustering at 70%, we can achieve 60% accuracy for a 50 clustering result. And the accuracy for each cluster is manually inspected by a simple if-then code (i.e.

if the text data contains temp, then it is correctly clustered for the temperature cluster).

Table 11 Fifty clustering result 1

Clustering Number = 50	
Category of semantic	Data acquired type (temperature, flow, RH, CO2, …)
Number of unique semantic	18
Accuracy of total clustering	30/50 = 60%

Table 12 Fifty clustering result 2

Cluster	Data points (EA)	Semantic Information	Accuracy (%)
0	166	flow	100
1	58	zone temp	100
2	73	override	100
3	55	da temp	100
4	106	co2, occ, run, schedule	<50
5	81	command	100
6	149	run	100
7	65	damper position	100
8	58	run, schedule, alarm	<50
9	130	error	100

Cluster	Data points (EA)	Semantic Information	Accuracy (%)
10	56	alarm	100
11	121	zone temp	>50
12	10	zone temp	100
13	68	run	100
14	32	schedule	100
15	26	RH	100
16	83	flow	>50
17	18	zone temp	100
18	24	alarm	100
19	30	zone temp	100

Cluster	Data points (EA)	Semantic Information	Accuracy (%)
20	8	valve, alarm...	<50
21	5	error, alarm...	<50
22	25	alarm, hand...	<50
23	27	alarm, run...	<50
24	7	valve	100
25	5	valve, damper...	<50
26	22	start, error...	<50
27	30	error	100
28	40	temp, run...	<50
29	12	fault	100

Cluster	Data points (EA)	Semantic Information	Accuracy (%)
30	25	speed, failure	<50
31	4	request	100
32	6	occupancy	100
33	46	alarm, temp...	<50
34	59	alarm, flow...	<50
35	17	request	100
36	12	status	100
37	7	alarm, error	<50
38	6	temp, velocity	<50
39	54	power	>50

Cluster	Data points (EA)	Semantic Information	Accuracy (%)
40	4	energy, error...	<50
41	13	power	100
42	12	energy	100
43	5	alarm, error...	<50
44	53	energy, power	<50
45	20	energy, power	<50
46	14	power	100
47	24	temp	100
48	23	flow, temp...	<50
49	6	flow, request	<50

Figure 29 Clustering result (50)

The last parametric study case is the 249 clustering experiment. When Automated Logic installed the BAS for GHC building, they predefined the 249 unique semantic information types for each data points. Based on Figure 30 and Table 13, the hierarchical clustering algorithm doesn't predict the given semantic information well. The accuracy is only 22%. This low accuracy is driven by the calculation method of complete linkage. Since the complete linkage method calculates the distance between two clusters by the furthest distance, this method is not applicable for the extremely high number of clustering. Another interesting finding is detected from Figure 31, where both given semantic information and the algorithm prediction shows the Pareto distribution. The Pareto distribution means 20% of the semantic information takes 80% of the total data points. This distribution result indicates that after the certain maximum limit of the clustering number, ultimately doesn't guarantee a good accuracy rate because of the skewed data distribution. For example, the majority of data shows VAV in terms of major system type and temperature or flow in terms of acquisition type. Thus, it is very important to choose a correct number of clustering before conducting the hierarchical clustering procedure.

Table 13 249 clustering result

Clustering Number = 249	
Category of semantic	Vendor given semantic information
Number of unique semantic	249
Accuracy of total clustering	55/249 = 22%

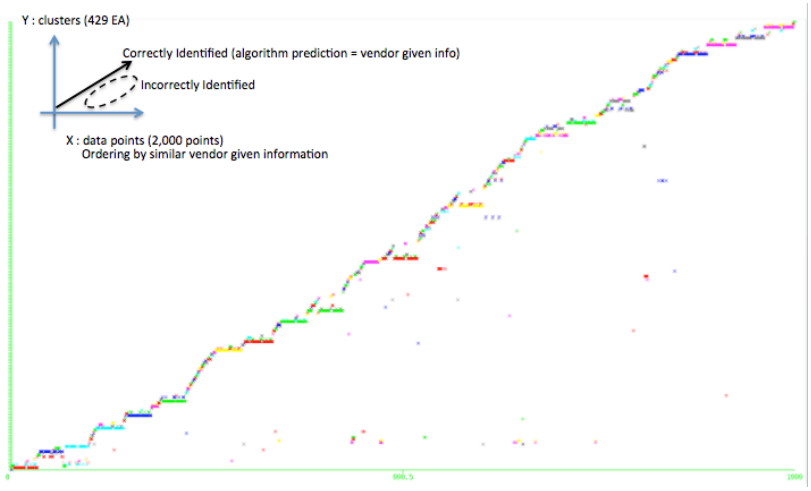


Figure 30 Clustering result (249)

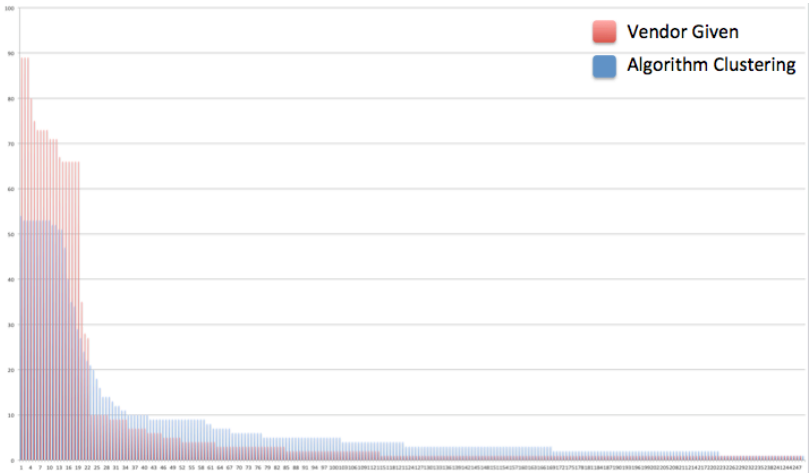


Figure 31 Clustering result comparison with ground truth

5.2 Data point type classification

5.2.1 Classification for the damper position

To identify the relationship between VAVs and AHUs, the damper position data is required from the VAV unit. Normally, a single VAV unit contains around 20 BACnet data points, and 4 of them are Analog Values (i.e. Airflow, Airflow setpoint, Damper position and Reheat Valve) as the control parameters. Additionally, we can filter the data points by unit of measurement (%). After filtering the data points by Analog Value and unit (%), we can only have two candidates (damper position or reheat valve) for classifying the damper position data type.

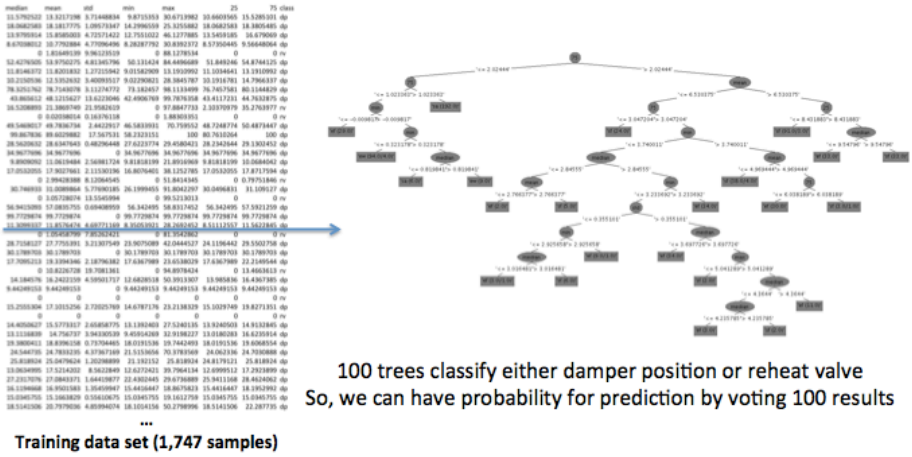


Figure 32 Random decision tree by a statistical feature sample

Calculating statistical features from randomly chosen 1,747 samples from previously filtered data point, we can build the feature table (7 features by 1,747 samples) with the manually labeled ground truth class values. From this feature table and ground truth labels, 100 decision trees are randomly generated. Having the classification result from 100 random decision trees, we can build the probability distribution for each class value. To evaluate the model by a new data set, 569 samples from March 2016 are collected. The model for classifying the damper position data classifies 569 samples correctly out of 594 samples (=95.79%) with low out-of-bag error (0.0567) and high kappa statistics (0.8197). To interpret the confusion matrix, the horizontal and vertical axis represent the actual and prediction data point types, and the diagonal values are correctly classified data

point types from this model. The model classifies the damper position data type from the actual damper position data type accurately. However, it misclassifies 25 data points as damper position instead of the reheat valve, thus, this model does not guarantee the higher accuracy for differentiating the reheat valve data type. Since the aim of the suggested classification algorithm is to identify the damper position data type from all the data points in the VAV unit, the skewness of the classification result is not a major problem. For the training and testing process for the Random Forest method, we used Weka, which is a Java based machine learning tool developed from the University of Waikato, New Zealand.

Table 14 Confusion matrix from the classification result 1

		(Prediction)	
		Damper position	Reheat valve
(Actual)	Damper position	**502**	0
	Reheat valve	25	**67**

```
inst#,     actual, predicted, error, probability distribution
     1        1:dp       1:dp          *0.952  0.048
     2        1:dp       1:dp          *0.941  0.059
     3        1:dp       1:dp          *0.932  0.068
     4        1:dp       1:dp          *0.94   0.06
     5        2:rv       2:rv           0.152 *0.848
     6        1:dp       1:dp          *0.94   0.06
     7        1:dp       1:dp          *0.955  0.045
     8        1:dp       1:dp          *0.938  0.062
     9        1:dp       1:dp          *0.939  0.061
    10        1:dp       1:dp          *0.937  0.063
    11        2:rv       2:rv           0.32  *0.68
    12        2:rv       2:rv           0.161 *0.839
    13        1:dp       1:dp          *0.93   0.07
    14        1:dp       1:dp          *0.941  0.059
    15        1:dp       1:dp          *0.932  0.068
    16        1:dp       1:dp          *0.999  0.001
    17        1:dp       1:dp          *0.949  0.051
```

Figure 33 Example result for the Random Forest model 1

5.2.2 Classification for the supply air duct pressure

To identify the relationship between VAVs and AHUs, the supply air duct pressure data is required from the AHU unit. Normally, a single AHU unit contains over 100 BACnet data points, and around 10 different types of sensors are installed (i.e. temperature, pressure, humidity, airflow, etc.). To acquire the supply air duct

pressure data point, we filtered the data point by the object (Analog Input) and the unit of measurement (in H2O). After filtering, we can have four candidates (supply air duct pressure, exhaust fan pressure, supply fan pressure and enthalpy wheel pressure) for classifying the supply air duct pressure.

To build the model for classifying the pressure sensor type, we calculated statistical features from 700 pressure sensor data samples from 6 AHUs with manually labeled ground truth class values. From this data collection, we generate 100 random decision trees to build the classification model. By having the classification result from the 100 random decision trees, we can gain a classification probability result for the test dataset from March 2016 (236 samples). The Random Forest model for classifying the damper position data classifies 222 samples correctly out of 536 samples (=94.06%) with low out-of-bag error (0.0375) and high kappa statistics (0.9182). Referring to Table 15, the model classifies the supply air and enthalpy wheel pressure sensor correctly, but it misclassifies some of the exhaust fan and supply fan pressure sensor data. Once again, the aim of this classification model is identifying the supply air duct pressure sensor, therefore, we can utilize this model for further analysis.

Table 15 Confusion matrix from the classification result 2

		(Prediction)			
		EF	SF	SA	EW
(Actual)	EF	**49**	6	0	4
	SF	4	**75**	0	0
	SA	0	0	**70**	0
	EW	0	0	0	**28**

Legend : EF(Exhaust Fan), SF(Supply Fan), SA(Supply Air), EW (Enthalpy Wheel)

inst#	actual	predicted	error	probability distribution			
1	3:sa	3:sa		0	0	*1	0
2	2:sf	2:sf		0.06	*0.94	0	0
3	2:sf	2:sf		0.02	*0.98	0	0
4	3:sa	3:sa		0	0	*1	0
5	4:ew	4:ew		0	0	0	*1
6	4:ew	4:ew		0	0	0.01	*0.99
7	3:sa	3:sa		0	0	*1	0
8	2:sf	2:sf		0	*1	0	0
9	1:ef	4:ew	+	0.18	0	0	*0.82
10	2:sf	2:sf		0.03	*0.97	0	0
11	3:sa	3:sa		0	0	*1	0
12	2:sf	2:sf		0	*1	0	0
13	3:sa	3:sa		0	0	*1	0
14	1:ef	1:ef		*0.74	0.26	0	0
15	3:sa	3:sa		0	0	*1	0

Figure 34 Example result for the Random Forest model 2

5.3 Functional relationship inference

5.3.1 Exploratory numeric trending result

As Figure 35 ~ Figure 40 show, it is hard to capture the relationship between discharge air temperature and flow at VAVs, and supply air temperature and flow at AHUs. The reason why both temperature and flow do not have a strong relationship is due to the control logic of the GHC building. As previously explained, AHUs serve three different program zones through VAVs and the majority of VAVs are controlled based on occupancy variable in the office program units. Both program variety and uncertain occupancy behavior create the complex temperature and flow relationship between VAVs and AHUs. Thus, it is hard to extract the VAVs – AHUs relationship information from the direct numeric trending of temperature and flow sensor data.

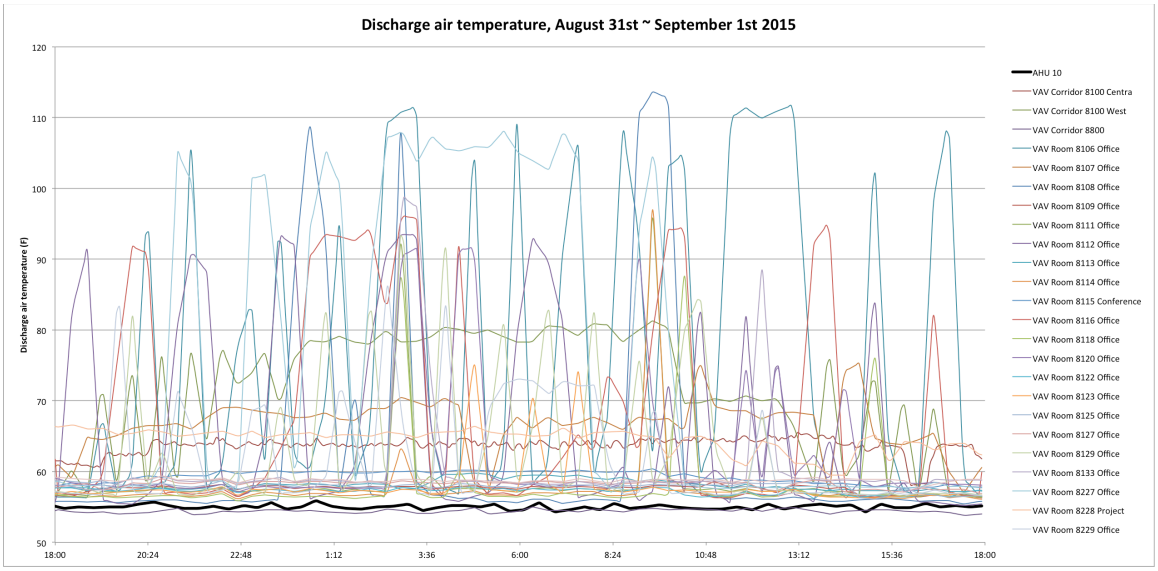


Figure 35 AHU - VAV temperature distribution on Summer day

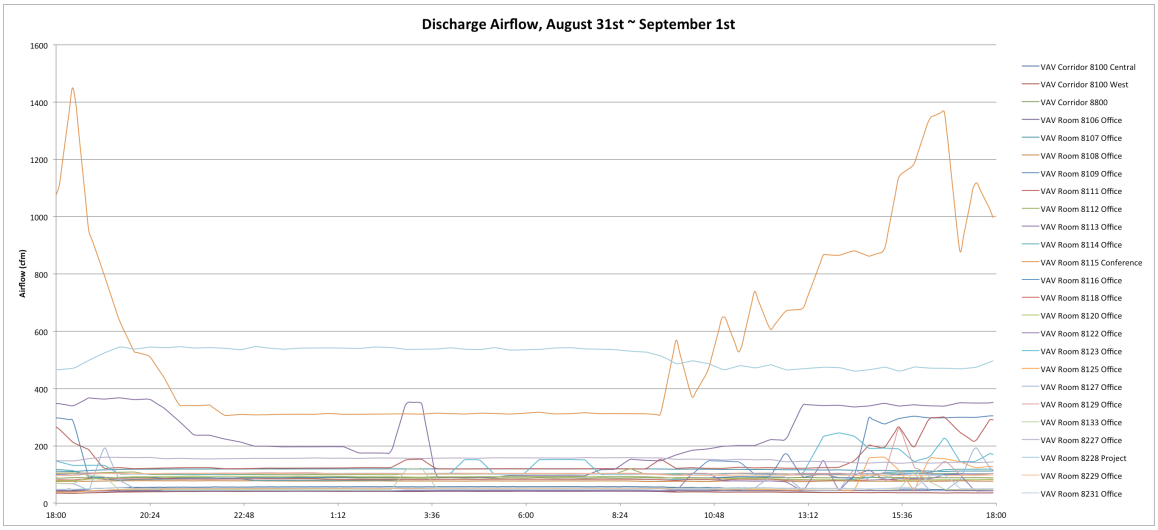


Figure 36 VAV airflow distribution on Summer day

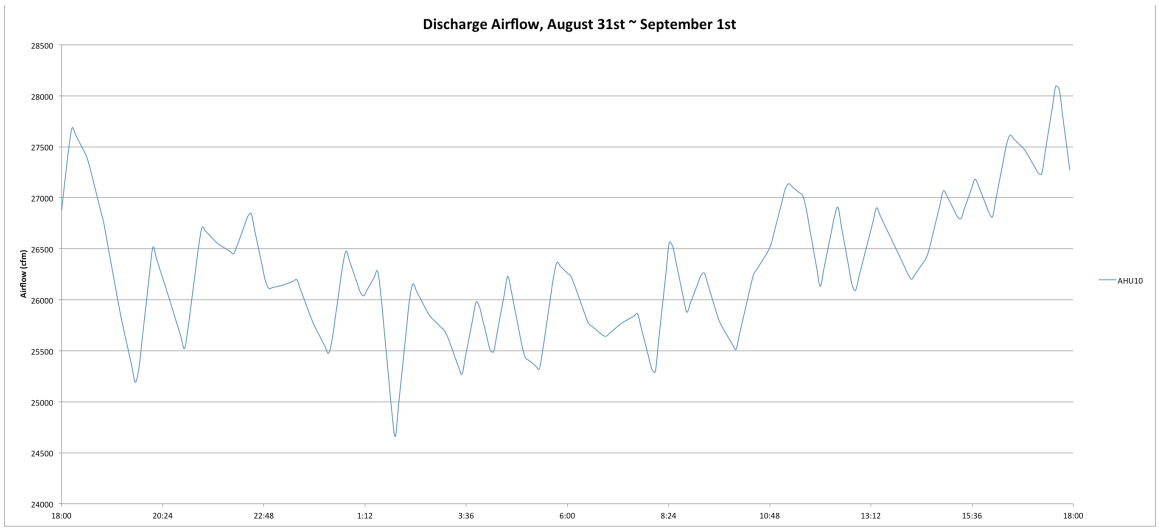


Figure 37 AHU airflow distribution on Summer day

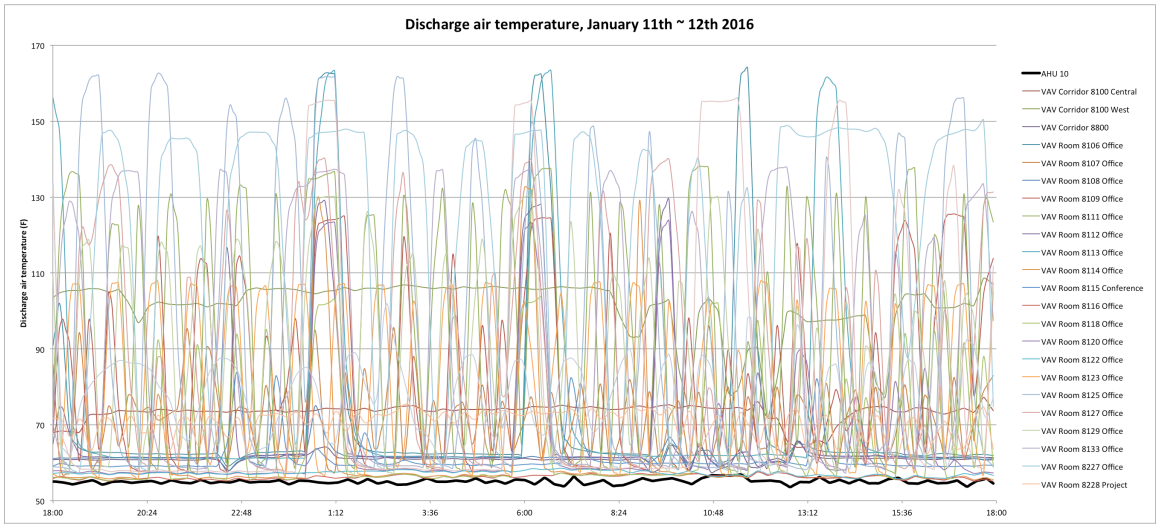


Figure 38 AHU - VAV temperature distribution on Winter day

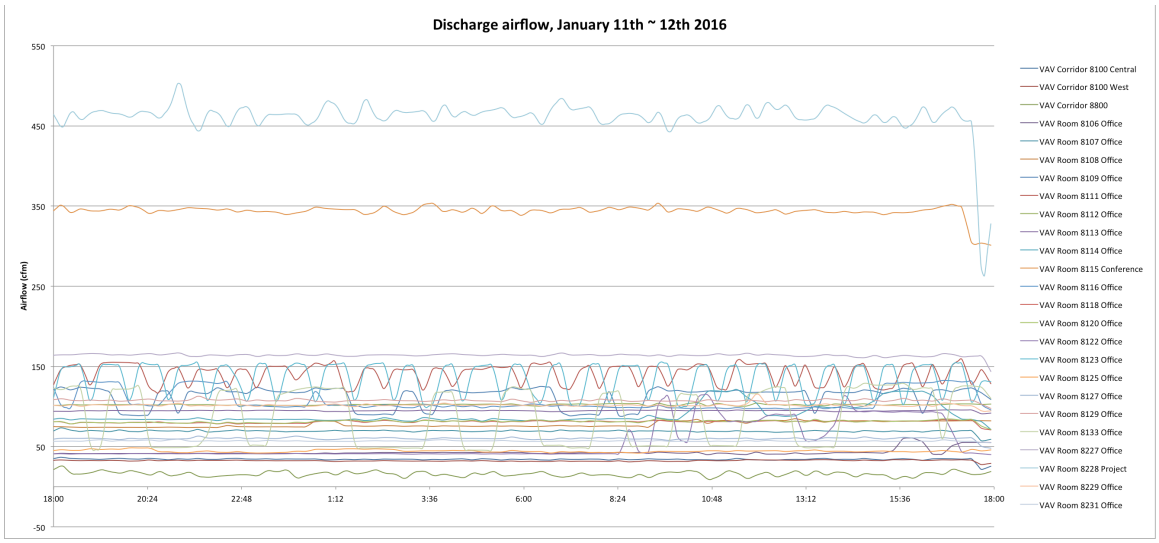


Figure 39 VAV airflow distribution on Winter day

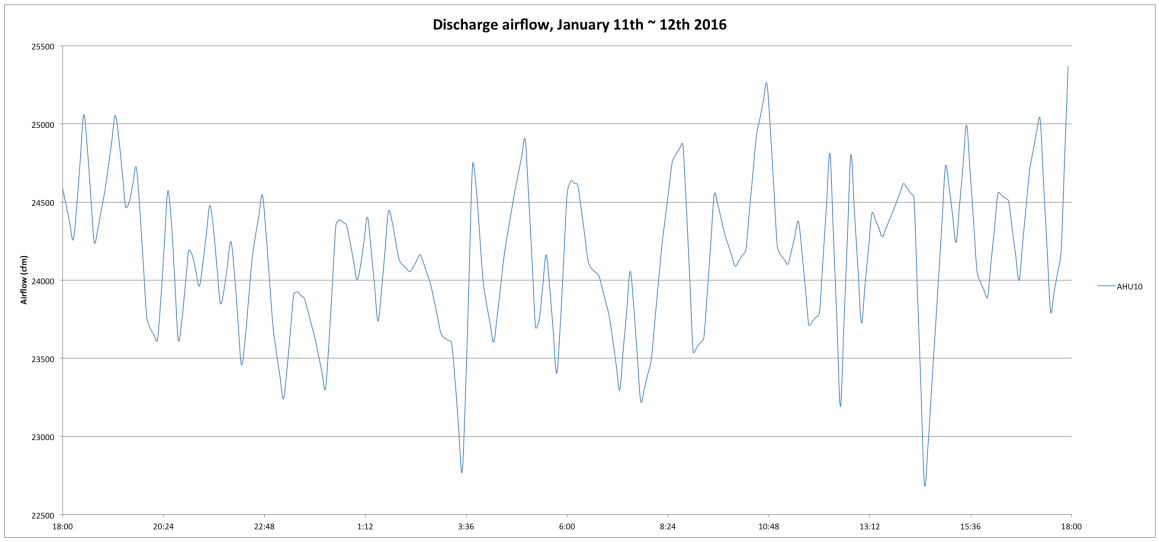


Figure 40 AHU airflow distribution on Winter day

5.3.2 Statistical features clustering result

The numeric trending for AHUs – VAVs relationship inference is investigated by not just direct comparison of numeric trending, but trying to cluster the VAVs by assigned AHUs with statistical features. 14 statistical features are extracted for clustering VAVs by assigned AHUs. The clustering algorithm is evaluated by implementing a similar setting of the semantic information clustering algorithm. Figure 41 shows the result of the algorithm and it conveys a low performance. The clustering pattern indicates poor performance. For example, AHU 11 actually should have 83 VAVs but it predicts 180 VAVs. To verify whether this feature extraction method is correct or not, each statistical feature distribution is plotted in Figure 42 ~ Figure 45. Intuitively, their distributions are similar even though they are assigned by 6 different AHUs. This is because, as long as their set point temperature is the same for all the VAVs, it is then difficult to differentiate the VAV numeric trending by only statistical features of the temperature data.

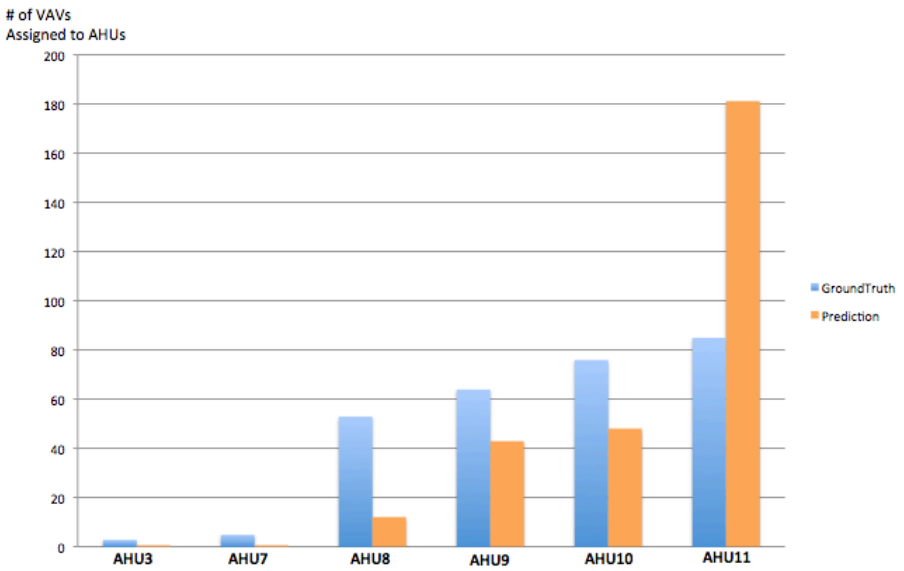


Figure 41 VAV - AHU mapping inference result

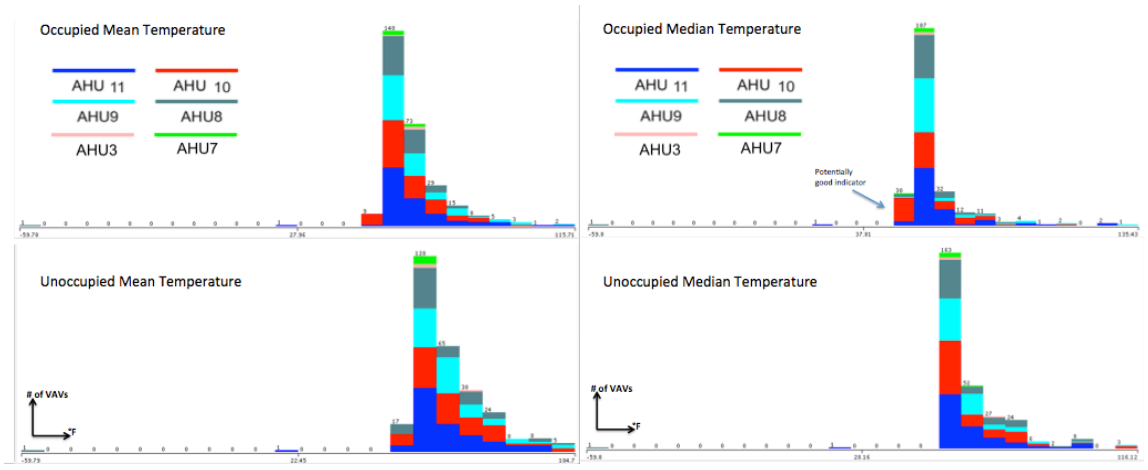


Figure 42 Mean and Median distribution

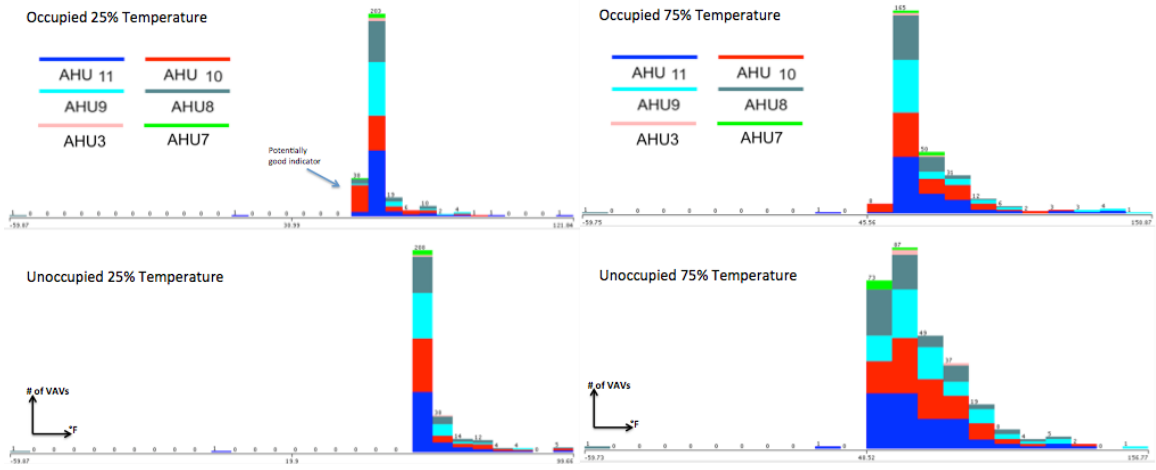


Figure 43 Quartiles distribution

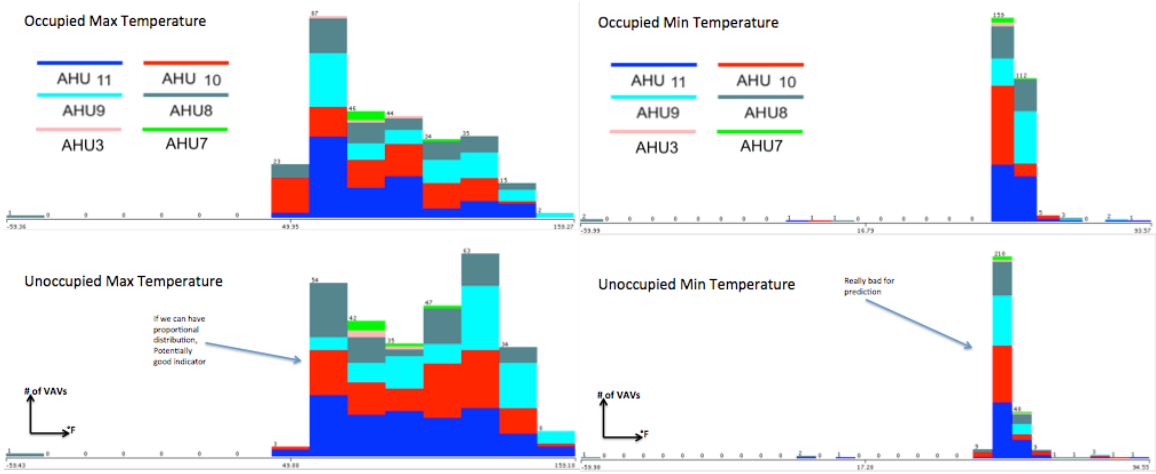


Figure 44 Max and Min distribution

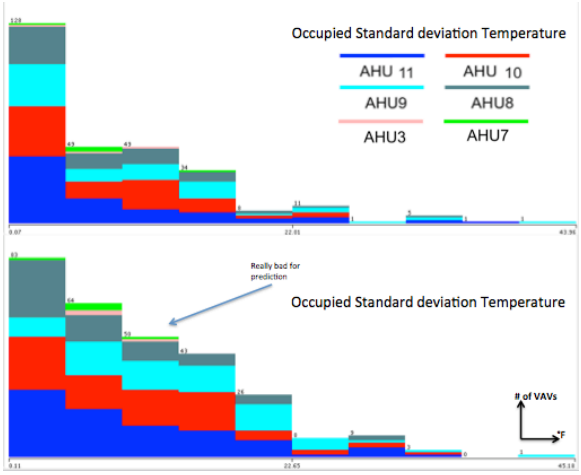


Figure 45 Standard deviation distribution

5.3.3 Similarity scoring result

Similarity scoring by cross correlation is implemented for damper position at VAVs and static pressure at AHUs. Figure 46 shows a complete cross correlation result of a typical summer day. The result indicates they have a negative correlation with a short time lag. The absolute number of the maximum cross correlation value is 0.9143, and it is the maximum score among other AHUs scores. Thus, we can infer that VAV_room4215 is connected with AHU3. Additionally, Figure 47 represents the cross correlation result of a typical winter month. Their relationship is also a negative correlation with 0.8432 maximum score, thus the prediction of VAV_room6119 is AHU7.

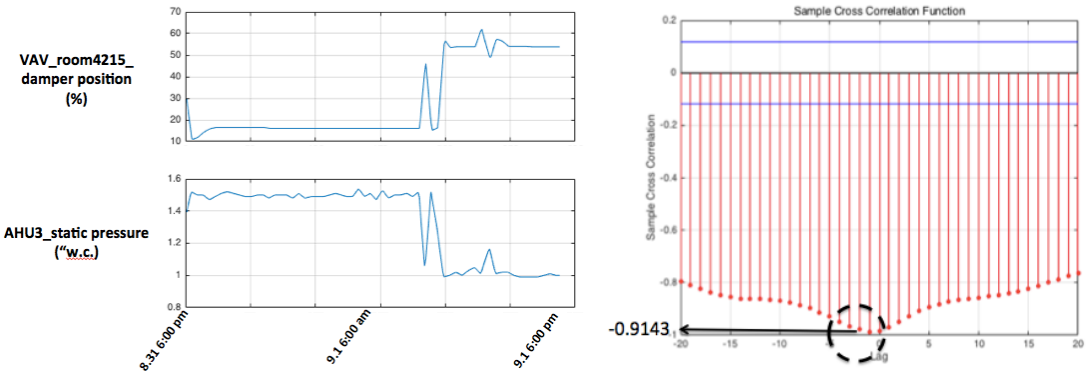


Figure 46 Cross correlation result on Summer day

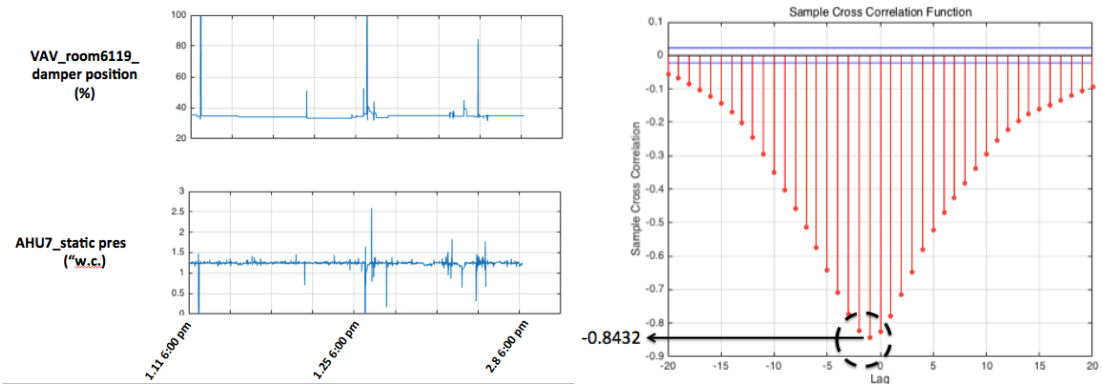


Figure 47 Cross correlation result on Winter month

After calculating all the similarity scores for the combination of VAVs and AHUs, we can know the inference accuracy result on Table 16. Among 6 profiles, the monthly profile has the maximum accuracy (40.50%). However, those six profiles do not have enough variation in terms of seasonal factors. To cover seasonal building operation behavior, the dataset must be extended to multiple months.

Table 16 Inference accuracy result of 6 profiles

	Summer	Winter
Daily	58/297 = 20.78%	21/297 = 7.52%
Weekly	81/297 = 29.03%	69/297 = 24.73%
Monthly	70/297 = 25.08%	113/297 = 40.50%

(correctly matched VAV / total number of VAV)

5.3.4 Multiple scores prediction result

Extending the dataset from 1 month to 9 month profiles, the cross correlation function can consider various seasonal factors. The 9 month profiles are the minimum number covering three different seasonal behaviors. Since the 14 damper positions at VAVs are not collected properly, the total number of VAVs is reduced to 283. The seasonal characteristics of the profiles are selectively beneficial for predicting the AHU. For example, the summer peak profile from August is a good indicator for AHU9, on the other hand, the winter peak profile from January provides a higher accuracy for AHU11. To adaptively predict the correct AHU from

the cross correlation result, the most frequently predicted AHU type from 9 months is predicted at the final stage of this algorithm.

Table 17 Multiple scores prediction method result

Assigned AHU	Correctly matched	Accuracy
AHU3	3/3	100%
AHU7	4/5	80%
AHU8	45/53	85%
AHU9	56/65	86%
AHU10	72/84	86%
AHU11	26/73	63%
Total	226/283	79.85%

Table 17 indicates the relationship inference result, the overall accuracy is 80%, but the relationship between VAVs and AHU11 does not show the strong relationship compared to other AHUs. The VAVs served from AHU11 are located in various thermal characteristic zones. Essentially AHU11 serves the second largest number of VAVs (73ea), and the AHU11 distributes the air into 5 different floors (from the 5^{th} floor to the 9^{th} floor). It means the vertical distance between VAVs and AHU11 is the maximum distance among other AHUs. In addition, as Figure 14 ~ Figure 20demonstrate, the VAVs, which connect to AHU11, are facing various orientations compared to other VAVs.

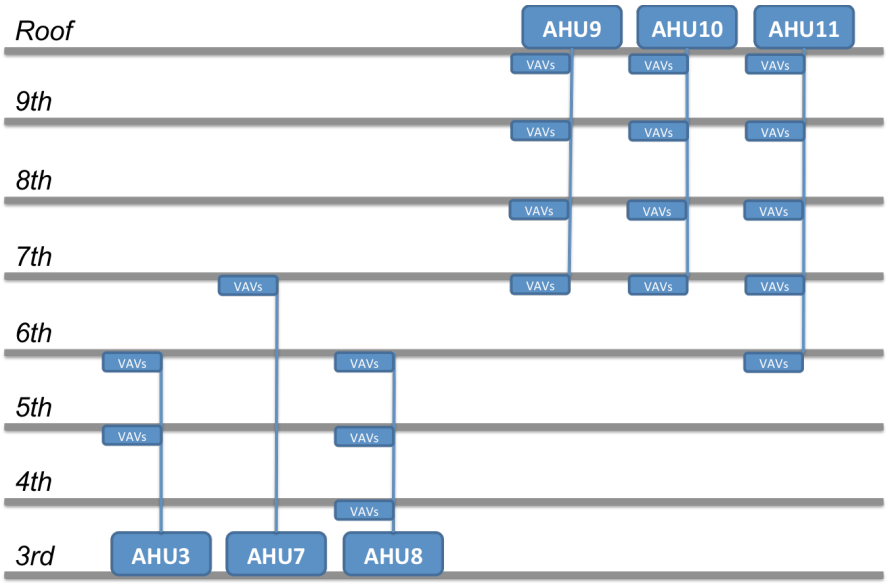


Figure 48 Vertical distances between VAV - AHU

6. Conclusion

6.1 Framework deployment

The final outcome of this paper provides the framework for future users to utilize the building data in more efficient ways. The primary objective of this framework is discerning the relationship between VAVs and AHUs from the heterogeneity of the building data points. The framework mainly consists of two stages. The first is filtering and classifying the data point by BACnet information and daily statistical features for acquiring damper position and supply air duct pressure data. The second stage is figuring out the relationship between VAVs and AHUs by the data that we investigated in the first stage.

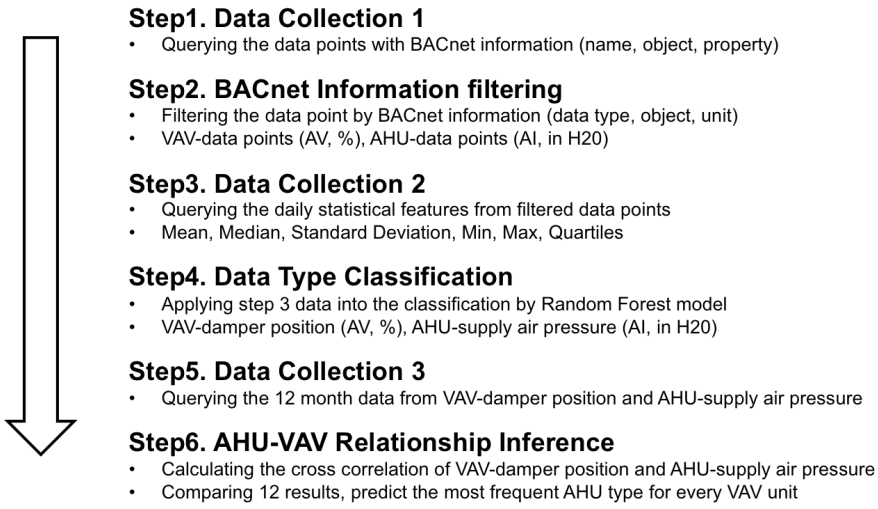


Figure 49 Detail procedure of the framework

The detail procedure of the framework has 6 steps and is described in Figure 49. Firstly, the user can collect the BACnet information (i.e. names, objects and properties) with the data point, and filter the data point list by what the user needs. For example, if the user wants to find the supply air duct pressure sensor data, the user can filter the data points by object type (analog input) and unit of measurement (in H2O).

Data points collection with BACnet information

Object name	Object	Unit	Value	...
#scsc_ahu-8_io/m005	BI	on/off	on	...
#scsc_ahu-8_io/o901	AI	F	58	...
#scsc_ahu-8_io/q292	AI	in H20	1.25	...
#scsc_ahu-8_io/q305	AO	%	93	...
#scsc_ahu-8_io/m039	BO	Normal/alarm	Normal	...
#scsc_ahu-8_io/da_setpt	AV	F	60	...
...	...	...	...	...

Filtering data points by desired object and unit

Object name	Object	Unit	Value	...
~~#scsc_ahu-8_io/m005~~	~~BI~~	~~on/off~~	~~on~~	...
~~#scsc_ahu-8_io/o901~~	~~AI~~	~~F~~	~~58~~	...
#scsc_ahu-8_io/q292	**AI**	**in H20**	**1.25**	**...**
#scsc_ahu-8_io/q305	**AO**	**%**	**93**	**...**
~~#scsc_ahu-8_io/m039~~	~~BO~~	~~Normal/alarm~~	~~Normal~~	...
~~#scsc_ahu-8_io/da_setpt~~	~~AV~~	~~F~~	~~60~~	...
...	...	...	...	...

Figure 50 Example of step 1 and 2

After having all the candidates for the damper position data points in VAVs and the supply air duct pressure data points in AHUs, the daily statistical features (i.e. mean, median, standard deviation, minimum, maximum and quartiles) are extracted from those candidate data points. These daily statistical features are applied to the classification model for labeling their semantic information. The classification model is developed by Random Forest algorithm and the historical random samples. Essentially, the classification model identifies the semantic information (damper position, supply air duct pressure) by the probability calculated through the results of the 100 decision trees.

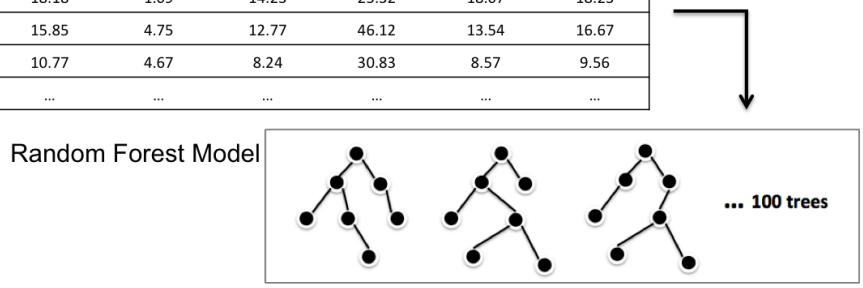
Object name	Object	Unit	Value	...
~~#scsc_ahu-8_io/m005~~	~~BI~~	~~on/off~~	~~on~~	...
~~#scsc_ahu-8_io/o901~~	~~AI~~	~~F~~	~~58~~	...
#scsc_ahu-8_io/q292	**AI**	**in H20**	**1.25**	**...**
#scsc_ahu-8_io/p815	**AI**	**in H20**	**6.21**	**...**
~~#scsc_ahu-8_io/m039~~	~~BO~~	~~Normal/alarm~~	~~Normal~~	...
~~#scsc_ahu-8_io/da_setpt~~	~~AV~~	~~F~~	~~60~~	...
...	...	...	...	...

Calculating daily statistical features from filtered data points

Name	Mean	Median	STD	Min	MAX	25%	75%
#scsc_ahu-8_io/q292	11.57	13.23	3.74	9.87	30.67	10.66	15.21
#scsc_ahu-8_io/p815	18.06	18.18	1.09	14.23	25.32	18.07	18.23
#scsc_ahu-8_io/s422	13.97	15.85	4.75	12.77	46.12	13.54	16.67
#scsc_ahu-8_io/y212	8.67	10.77	4.67	8.24	30.83	8.57	9.56
...	...	...	...	...	...	...	...

Figure 51 Example of step 3

Name	Mean	Med	STD	Min	MAX	25%	75%
#scsc_ahu-8_io/q292	11.57	13.23	3.74	9.87	30.67	10.66	15.21
#scsc_ahu-8_io/p815	18.06	18.18	1.09	14.23	25.32	18.07	18.23
#scsc_ahu-8_io/s422	13.97	15.85	4.75	12.77	46.12	13.54	16.67
#scsc_ahu-8_io/y212	8.67	10.77	4.67	8.24	30.83	8.57	9.56
...	...	...	...	...	...	...	...



Name	Mean	Med	STD	Min	MAX	25%	75%	Data type
#scsc_ahu-8_io/q292	11.57	13.23	3.74	9.87	30.67	10.66	15.21	Supply Air Pressure
#scsc_ahu-8_io/p815	18.06	18.18	1.09	14.23	25.32	18.07	18.23	Exhaust Fan Pressure
#scsc_ahu-8_io/s422	13.97	15.85	4.75	12.77	46.12	13.54	16.67	Supply Fan Pressure
#scsc_ahu-8_io/y212	8.67	10.77	4.67	8.24	30.83	8.57	9.56	Enthalpy Wheel Pressure
...	...	...	...	...	...	...	...	...

Figure 52 Example of step 4

Since the user acquired the damper position and supply air duct pressure data point from the previous steps, the user can investigate the VAVs – AHUs relationship by the acquired data points. Querying the 12 months data by 5 minute intervals from both damper position and supply air duct pressure, the user needs to calculate the similarity scores (absolute number of cross correlation) of 12 monthly profiles. Therefore, each VAV unit has 6 similarity scores from 6 different AHUs per month. Considering the maximum similarity scores as the prediction of the specific month, each VAV unit has 12 predictions from 12 month profiles. Ultimately, the framework returns the most frequently predicted AHU type as the final outcome of the relationship.

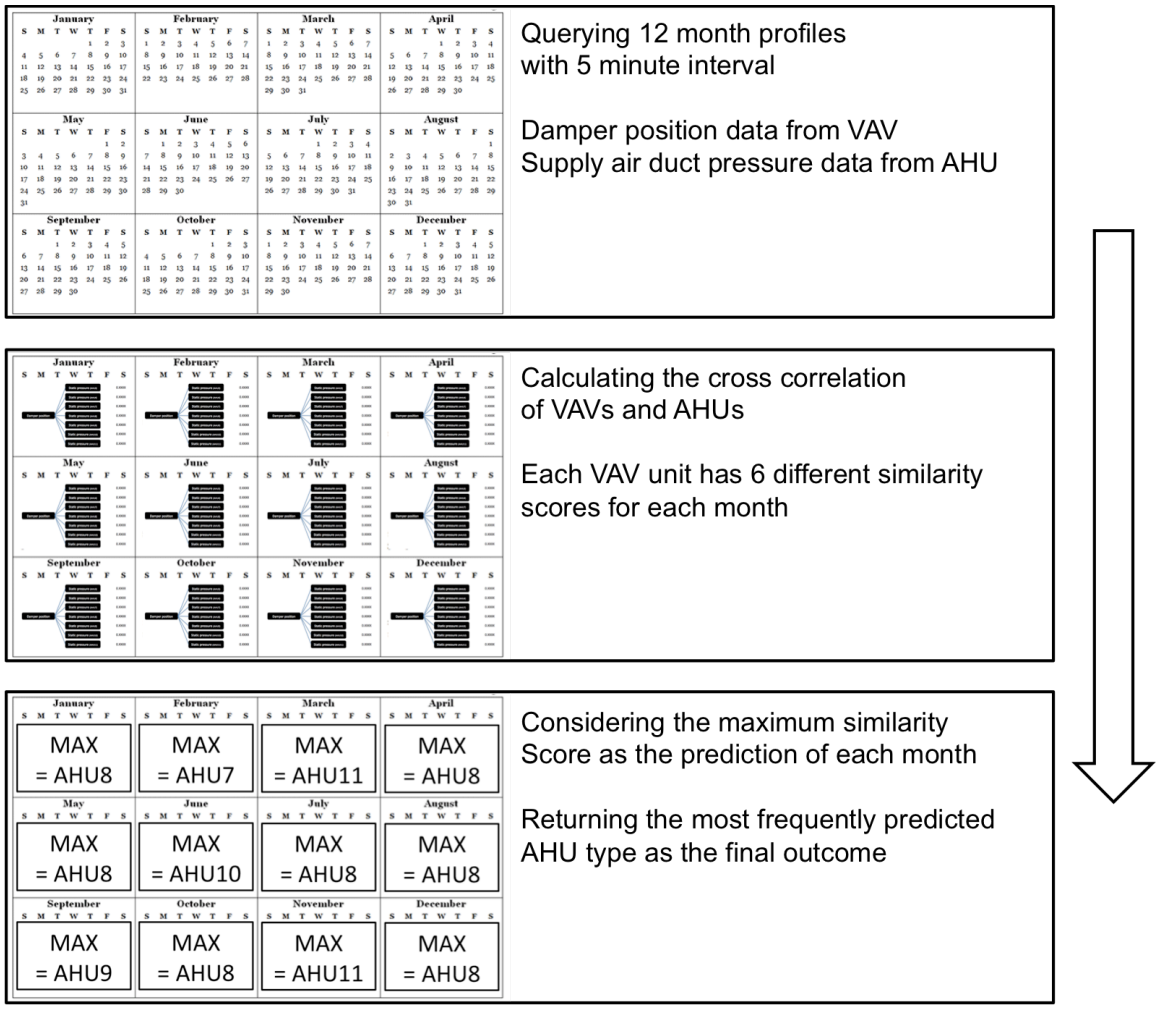


Figure 53 Example of step 5 and 6

The overall methodology for this framework is driven by the physical cyber system of the building. To discern the physical relationship, conventional practice dictated that engineers checked the equipment and drawing manually. However, since information technology has dramatically improved, where the building industry actively utilizes the technology into the building system, the building industry is now ready to derive the useful information from the building database. For example, the GHC building's data points are managed under BACnet protocol and their historical database is stored. Thus, the physical relationship is indirectly inferred through the data mining of the cyber system database. Assuming that a building is managed under BACnet protocol and at least 9 months of building data is stored,

this framework is very useful when the user has no information on the relationship between VAVs and AHUs. Even though the mechanical drawing is missing or the user cannot read it, they can diagnose the mechanical system with this framework.

VAV unit	AHU_A	AHU_B	AHU_C	AHU_D	Prediction
VAV_1	3	6	0	0	AHU_B
VAV_2	1	3	4	1	AHU_C
VAV_3	1	2	5	1	AHU_C
VAV_4	5	1	0	3	AHU_A
VAV_5	2	4	2	1	AHU_B
...	...	...	...	...	...

Figure 54 Final framework outcome

6.2 Finding & Limitation

This project has two primary goals. The first goal is mapping the semantic information on the data points and the second one is investigating the VAVs – AHUs relationship by the semantic information that is acquired in stage 1. To accomplish the first objective, initially we try to cluster the data point by the similar semantic information. However, the accuracy for clustering the acquisition type (i.e. temperature, pressure, etc.) shows less than 20%. This is because when we have only the text features for clustering, we are then limited in being able to cluster the semantic information. In the GHC building data points, the data point name includes the equipment type (i.e. VAV, AHU, etc.), however the detail sensor information (i.e. supply air duct pressure, mixed air temperature, etc.) is not represented as the text data. In some data points, we can find the acronyms and synonyms on the point names, but the majority of data point names contains the company's naming

ordering rule (combination of one character and 3 digit numbers). Moreover, there are too many data points in the building database, and some of the data points are just constant numbers or binaries as the control parameters. In this initial trial to find the semantic information, we realize that considering only text features is not enough to infer the semantic information, therefore, we need to reduce the data points based on our needs.

The second method is developed by the two lessons that we learned during the first trial. Essentially the filtering stage utilizes the BACnet information for reducing the useless data points of our framework. The BACnet information is very useful to understand how the building data points are managed and, as long as the building data points are managed under the BACnet protocol, we can implement this filtering method easily for other buildings. However, even though this filtering method reduces the data points, it is not enough to find the desired data type. For example, reheat valve and damper position data points are both analog values with percentage as the unit of measurement. To solve this type of problem, we consider the numerical features to classify the desired data type. The daily statistical features of numerical signals are very informative in classifying the data type. This filtering and classification method has implementation limits when the user knows which data type the user needs. On the other hand, the first clustering method is mapping all the data points with similar semantic information. Additionally, the current filtering and classification model is only tailored to find the damper position and supply air duct pressure data in VAV and AHU respectively. Once again, however, if the user knows what they need for further analysis, this method is very informative in finding the desired data point type.

To accomplish the second goal of this project, which is an investigation of the VAVs – AHUs relationship, the exploratory data analysis is conducted with temperature and flow data. Since a single AHU serves an average of 46 VAV units, and a building contains various program zones, it is hard to find the intuitive relationship in medium sized commercial buildings like the GHC building. To learn more

information about the temperature data, the statistical features are extracted from VAV units. However, this statistical feature method is not informative in mapping those VAV units on the AHU type, because the control set-points are almost identical for 6 different AHUs. Therefore, we invert our viewpoint to the mechanical relationship. Typically, the damper position at VAV unit is calculated based on the sensor reading at supply air duct pressure, therefore they have a negative correlation. By calculating the cross correlation between damper position from VAVs and supply air duct pressure from AHUs, we can score the correlations between VAVs and AHUs. Based on the building behavior in heating, cooling or swing seasons, different profiles are collected. Therefore, we consider 9 month profiles to calculate the cross correlations. Since we need to query and calculate 9 ~ 12 month profiles, this step requires high computational demands. To achieve the 80% accuracy, at least 9 month profiles are required.

6.3 Future work

The next step for this project involves the evaluation of a new building where this framework is applied. We developed the framework for future users to overcome interoperability issues with other buildings. However, the Random Forest classification model in the first stage is driven by the historical data of the GHC building. Thus, the Random Forest classification model should be evaluated with other building data for the active usage of this framework.

This framework is especially developed for inferring the VAVs – AHUs relationship by damper position and supply air duct pressure data. It is important to note though, that there are a variety of mechanical relationships in the building system. The future user can develop their methodology within this framework structure. For example, if they train the classification model based on the different location temperature sensor data, the new framework will define the locations of different temperature sensors. Furthermore, if we can collect the entire possible train model

for data type classification, this framework will ultimately map all the data points with the semantic information.

Bibliography

Automated Logic //. (n.d.). Retrieved April 27, 2016, from http://www.automatedlogic.com/

BACnet Website. (n.d.). Retrieved April 27, 2016, from http://www.bacnet.org/

Balaji, B., Verma, C., Narayanaswamy, B., & Agarwal, Y. (2015, November). Zodiac: Organizing Large Deployment of Sensors to Create Reusable Applications for Buildings. In *Proceedings of the 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments*(pp. 13-22). ACM.

Bernaden III, A., Decious, G. M., Seem, J. E., Drees, K. H., West, J. D., & Kuckuk, W. R. (2001). *U.S. Patent No. 6,219,590*. Washington, DC: U.S. Patent and Trademark Office.

Bhattacharya, A., Culler, D. E., Ortiz, J., Hong, D., & Whitehouse, K. (2014).*Enabling portable building applications through automated metadata transformation*. Technical Report UCB/EECS-2014-159, EECS Department, University of California, Berkeley.

Bhattacharya, A. A., Hong, D., Culler, D., Ortiz, J., Whitehouse, K., & Wu, E. (2015, November). Automated metadata construction to support portable building applications. In *Proceedings of the 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments*(pp. 3-12). ACM.

Bouckaert, R. R., Frank, E., Hall, M., Kirkby, R., Reutemann, P., Seewald, A., & Scuse, D. (2013). Waikato Environment for Knowledge Analysis (WEKA) Manual for Version 3-7-8. *The University of Waikato, Hamilton, New Zealand*.

Bushby, S. T. (1997). BACnet TM: a standard communication infrastructure for intelligent buildings. *Automation in Construction*, *6*(5), 529-540.

Carnegie Mellon School of Computer Science. (n.d.). Retrieved April 27, 2016, from http://www.cs.cmu.edu/

Chen, S., & Demster, S. J. (1996). *Variable air volume systems for environmental quality*. McGraw Hill Professional.

Department of Energy. (n.d.). Retrieved April 27, 2016, from http://energy.gov/

Gao, J., Ploennigs, J., & Berges, M. (2015, November). A Data-driven Meta-data Inference Framework for Building Automation Systems. In *Proceedings of the 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments* (pp. 23-32). ACM.

Gut, E. B. (1988). *U.S. Patent No. 4,751,501*. Washington, DC: U.S. Patent and Trademark Office.

Hong, D., Wang, H., Ortiz, J., & Whitehouse, K. (2015, November). The Building Adapter: Towards Quickly Applying Building Analytics at Scale. In*Proceedings of the 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments* (pp. 123-132). ACM.

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*,*32*(3), 241-254.

Koc, M., Akinci, B., & Bergés, M. (2014, November). Comparison of linear correlation and a statistical dependency measure for inferring spatial relation of temperature sensors in buildings. In *Proceedings of the 1st ACM Conference on Embedded Systems for Energy-Efficient Buildings* (pp. 152-155). ACM.

Lewis, J. P. (1995, May). Fast normalized cross-correlation. In *Vision interface* (Vol. 10, No. 1, pp. 120-123).

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news*, *2*(3), 18-22.

Liu, M. (2003). Variable speed drive volumetric tracking for airflow control in variable air volume systems. *Journal of Solar Energy Engineering*, *125*(3), 318-323.

Mayfield, E., & Rosé, C. (2012). LightSIDE: text mining and machine learning user's manual.

National Institute of Standards and Technology. (2004). Retrieved April 27, 2016, from http://www.nist.gov/index.html

OSIsoft: Global Leader in Operational Intelligence. (n.d.). Retrieved April 27, 2016, from http://www.osisoft.com/

Pritoni, M., Bhattacharya, A. A., Culler, D., & Modera, M. (2015). Short Paper. *Proceedings of the 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments - BuildSys '15*. doi:10.1145/2821650.2821677

Project Haystack. (n.d.). Retrieved April 27, 2016, from http://project-haystack.org/

Schumann, A., Ploennigs, J., & Gorman, B. (2014, November). Towards automating the deployment of energy saving approaches in buildings. In*Proceedings of the 1st ACM Conference on Embedded Systems for Energy-Efficient Buildings* (pp. 164-167). ACM.