The Data Holmes: Towards Automatic Metadata Inference for Sensors in Buildings

Ph.D. Dissertation Proposal

Dezhi Hong

Abstract

Ubiquitous and immersive sensing equipment and devices, e.g., the Internet of Things, are generating an explosive amount of data. To extract meaningful and actionable information out of these data inevitably requires the metadata of the generated data. However, the majority of data remain unlabeled and the generation of metadata still involves labor intensive efforts, thus fundamentally unscalable. As a solution, we propose a framework for inferring the contextual information embedded in the sensor time series, e.g., what they measure, where they locate, how they relate to each other, etc. We tease out a representative of the metadata inference problem, particularly, for commercial buildings, and demonstrate first steps towards a metadata inference solution that requires minimal human intervention. At core of our solution lies a suite of techniques that exploit both the textual and time series data of sensors in buildings. We have explored a few approaches to inferring the type and location information that show promise. Building upon the early results, we will next focus on inferring the relationship between points. This proposal provides an overview of our solution, along with some preliminary results and key challenges, proposed research and an evaluation plan.

1 Introduction

Innovations in the “Internet of Things” (IoT) devices have enabled deployment and usage at an unprecedented level, with an estimate that the number of deployed IoT devices will reach 25 billion in 2020 [22]. Technologies developed upon this infrastructure present new opportunities to better serve people in every and each aspect in life, from commute to communication, wellness monitoring to activity assistance, and so on. Distilling useful information out of the gigantic amount of data requires the metadata, or labels, of the generated data, e.g., what they measure, who and where they belong to, etc. However, the state-of-art of metadata generation still involves labor intensive efforts from personnel who often need to have domain knowledge, thus remaining highly unscalable and sometimes error-prone. To enable the analysis of IoT data at scale, we envision a tool that is capable of automatic metadata inference and generation.

We take a representative subset of the problem and focus on the metadata inference for commercial buildings. According to recent reports from the U.S. Department of Energy [48, 39], commercial and industrial buildings in the U.S. account for almost 20 percent of the country’s total energy use and a good 30 percent of that energy is used “inefficiently or unnecessarily.” Reducing this energy usage is a national grand challenge: in 2011, the U.S. government launched the Better Buildings Challenge to make these buildings at least 20 percent more efficient by 2020 [18]. To achieve this goal, many organizations are applying data analytics [19, 11] to the thousands of sensing and control points in a typical commercial building to detect wasteful, incorrect, and inefficient operations. However, these analytic engines are *tightly coupled* to the database schemata and metadata conventions in the buildings for which they were designed and thus cannot easily be applied to different buildings where the type, location, and relationships between sensors are represented differently. As a result, an analytics engine cannot be applied to a new building without first addressing the issue of *mapping*:

creating a match between the sensor streams and the inputs of a data analytic engine. In practice, the process of mapping a new building to the inputs of an analytics engine is currently a manual process that often involves a technician visiting the building to visually inspect the equipment installation. The mapping process requires significant integration effort and anecdotally can take a week or longer for each commercial building. For large organizations that are applying this approach to hundreds or more buildings, such as Microsoft's 88 Acres project [49], this process can take years. Thus, mapping is a major obstacle to applying building analytics at scale.

Even if this highly manual process is performed once, the need for additional mapping is not necessarily eliminated. New types of metadata will be required as the building is modified or renovated, as the equipment is upgraded, or as new conditions and algorithms are added to the analytics engine. Additionally, the mapping problem cannot be solved simply by investing more person hours into the problem. Even after a building is fully mapped, a new analytics engine may be developed that requires a different kind of metadata, e.g. which devices are on the northern side of the building, or which sensors are affected by a given air handler unit. These and other types of metadata may even never have been encoded in the original databases at all. Thus, as energy models and building analytics engines become more nuanced, the mapping problem will become increasingly important. As a result, we envision a system that will allow an advanced analytic engine to quickly connect to and analyze the data from a commercial building. It would extract or infer metadata values about the sensing and control points, and map them to a normalized standard metadata. The system should also enable any new building analytics engine to quickly be applied to the 10's of millions of commercial buildings across the globe. Doing so would enable a new market where boutique analytics could quickly be matched with the buildings they would benefit the most.

In this proposal, we propose *Data Holmes*, a framework for automatic metadata mapping that requires *minimal* human intervention and we demonstrate a first step towards building such a tool. The key insight that guides our solution is that a sensor typically has two attributes - its recorded name (a text string) and its numerical readings generated over time. In particular, the sensor names are likely to be good indicators of metadata structure since people would usually follow a certain naming convention, but the conventions might not be consistent across buildings. As for the sensor readings, they are more consistent across buildings. For instance, room temperature will be between 60 and 70 Fahrenheit degrees, no matter which building. Our solution will combine the complementary strengths of these two attributes to extract or infer the contextual information of points in a building, as much automated as possible. We have performed preliminary studies on inferring a key category of metadata: the *type* of sensor associated with a sensing point. Our initial results show promise in further developing techniques to infer other kinds of important metadata information, such as how are a group of sensors related to another group of sensors. We envision the main contributions in this thesis as follows:

- An algorithm that uses as source already labeled buildings to automatically infer type information for a new building. Such algorithm might not label the entire building, but those that are labeled are expected to have high accuracy.
- An algorithm to perform semi-automated type classification where manual labeling is required. Such a algorithm would compensate the first algorithm.
- An algorithm that can infer for sensors a set of important relationships that are critical to applications running in buildings.
- A tool with a graphical user interface that helps users, such as a building manager, to interactively label the sensors and equipment in a building.

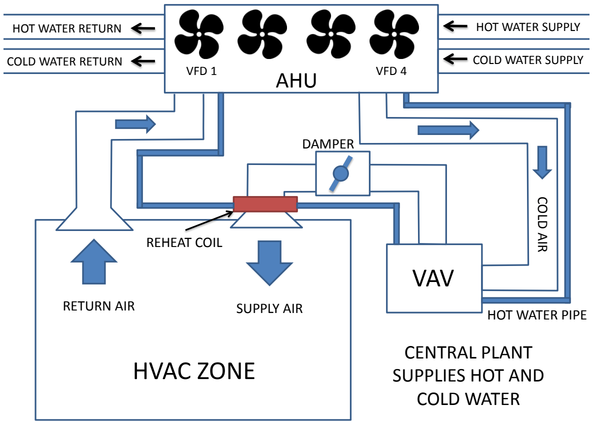


Figure 1: A typical HVAC system consisting of an air handler unit (AHU), several variable air volume boxes (VAV), water-based heating/cooling pipes and air circulation ducts. (Figure used with permission from the authors of [4].)

2 Background and Related Work

2.1 An Example Building

Figure 1 illustrates a typical heating, ventilation, and air conditioning (HVAC) system deployed in modern commercial buildings. An HVAC system usually uses a combination of hot and cold water pipes in conjunction with air handler units (AHU) to maintain the appropriate thermal environment within the building. An HVAC system usually consists of several AHUs and each AHU is responsible for a physical zone in the building. An AHU consists of variable speed drives that supply cold air (cooled by the supplied cold water) using ducts to VAV boxes distributed throughout the building. The hot water loop is also connected to these VAV boxes using separate pipes. Each VAV box controls the amount of air to be let into an HVAC zone using dampers, whose opening angle can be programmed. A reheat coil, which uses supplied hot water, is used to heat the air to meet the appropriate HVAC settings for each zone.

2.2 Metadata Heterogeneity

In modern commercial buildings, a sensing or control "point" is a sensor, a controller, or a software value, e.g., a temperature sensor installed in an office room. The metadata about the point indicates the physical location, the type of sensor or controller, how the sensor or controller relates to the mechanical systems, and other important contextual information. Most of the time, the metadata is encoded as short text strings with several concatenated abbreviations in a point name. Table 1 lists a few point names of sensors in three different building management systems (Trane[1], Siemens[2] and Barrington Controls[3]). For example, the point name `SODA1R300__ART` is constructed as a concatenation of the building name (`SOD`), the air handler unit identifier (`A1`), the room number (`R300`) and the sensor type (`ART`, area room temperature). As the name indicates, this point measures the temperature in a particular room; and it also indicates the control unit that can affect the temperature in this room. Clearly different naming conventions – generally guided by the equipment, vendor, and manufacturer – are used in these buildings. For example, the notion of *room temperature* is encoded with a different

[1] `http://www.trane.com/`

[2] `http://www.siemens.com/`

[3] The company is now defunct.

abbreviation in each of the three buildings: `Temp`, `RMT` and `ART`. Such variations across different buildings impose great difficulty in quickly deploying automated analytic solutions.

Building	Point Name

A	`Zone Temp 2 RMI204`
	`spaceTemperature 1st Floor Area1`

B	`SDH_SF1_R282_RMT` `SDH_S1-01_ROOM_TEMP`
C	`SODA1R300__ART` `SODA1R410B_ART`

Table 1: Example point names of temperature sensors in three different buildings.

2.3 Related Work

Sensor Metadata Inference There have been a few recent studies addressing the sensor metadata mapping problem studied in this work. Dawson-Haggerty et al. [15] and Krioukov et al. [32] introduce a Building Operating System Service stack, whereby the underlying building sensor stock is presented to applications through a driver-based model and an application stack provides a fuzzy-query based interface to the namespace exposed through the driver interface. Although this architecture has some useful properties for easing generalizability across buildings, the driver registration process is still performed manually. Schumann et al. [43] develop a probabilistic framework to classify sensor types based on the similarity of a raw point name to the entries in a manually constructed dictionary. However, the performance of this method is limited by the coverage and diversity of entries listed in the dictionary, and the dictionary size becomes intractable when there exist a lot of variations of the same type, or conflicting definitions of a dictionary entry in different buildings. Bhattarcharya et al. [6] exploit a programming language based solution, where they derive a set of regular expressions from a handful of labeled examples to normalize the point name of sensors. This approach assumes a consistent format for all point names across buildings, which might not be true in practice (as shown in Table 1). Balaji et al. [3] propose a clustering process with active learning to infer the type of points and again they assume same type of points share a similar pattern. Gao et al. [**?**] compare the performance of a few machine learning classifiers on type classification, using standard statistical features. However, they did not leverage the fact that often there are buildings already labeled and could be used to help the classification of a new building.

There have also been efforts on identifying sensor location. Chen and Tu [9] investigate how to cluster data streams in real-time using a density-based approach with a two-tiered framework. Their approach focuses on decreasing algorithm complexity for real-time sensor stream clustering. Kapitanova et al. [30] describe a technique to monitor sensor operations in the home and identify the sensor location based on a failure or removal of of sensors. Lu et al. [34] formulate a new algorithm, particularly leveraging the semantic constraints interpreted from sensor data to determine sensor locations.

Besides, there are three major bodies of related work to ours, i.e., schema matching in database, active learning and transfer learning in machine learning.

Schema Matching of Database Automatic schema matching [41] is a classical problem in the database community where correspondences between elements of two schemas are identified as part of the data integration process. Many techniques have been proposed to achieve partial automation of the match operation for specific application domains. Doan et al. ask a user to provide semantic mappings for a small set of data sources and then train a set of learners with existing machine learning approaches to find the mappings for new data sources [17]. Dhamankar et al. extend [17] to a semi-supervised setting, where domain- specific knowledge is introduced for complex expressions learning

[16]. Madhavan et al. exploit a large collection of schemata with known mappings to learn a prior distribution of the elements and their properties [35]. The learned prior distribution is then used as constraints to help a suite of base learners to complete the matching. Though adapting learning techniques, these works mostly focus on offline supervised settings and do not emphasize the efficiency of learning methods.

Active Learning The main idea behind active learning is that a machine learning algorithm can achieve greater accuracy with fewer labeled training instances if it is allowed to choose the data from which it learns [45]. Therefore, the key research question in active learning is evaluating the informativeness of unlabeled instances for querying. Various solutions have been proposed from different perspectives, including uncertainty reduction [10], query by committee [20], and expected error reduction [5]. In particular, Nguyen and Smeulders also incorporated the idea of clustering into active learning to select the most informative instances [38]. Their proposed query strategy gives priority to instances that are close to both classification boundary and cluster centroid. Dasgupta and Hsu utilized a hierarchical clustering structure to alleviate sampling bias and improve learning efficiency [13]. They present an algorithm that is statistically consistent and guarantees to have better label complexity than supervised learning. However, in those two methods clustering is performed in an ad-hoc manner, e.g., with predefined cluster size; and neither of them considers the label proximity between adjacent instances or propagates labels to reduce the amount of labels required for model training.

Transfer Learning Applying transfer learning to cross building sensor type classification saves extra effort in manual annotation by exploiting the labels in the already well annotated buildings. There are several categories of transfer learning, e.g., inductive, transductive, and multi-task transfer learning as comprehensively surveyed in [40]. Inductive transfer learning [12] assumes the set of class labels in the target domain is different from those in the source domain, and aims at achieving high classification performance in the target domain by transferring knowledge from the source domain. Multi-task transfer learning [8] has a similar setting, but tries to learn from the target and source domains simultaneously. Transductive transfer learning [14] assumes the source and target domains have the same set of labels, but different marginal distribution of features or conditional distribution of labels. This breaks the basic identical and independent assumption in classical supervised learning models and makes them inept. Typical solutions in transductive transfer learning reweigh the source domain trained classifiers' predictions in target domain, e.g., instance-based local weighting [7, 26, 46]. But these solutions usually assume that only the marginal distribution of features differ in the source and target domain. Ensemble methods are therefore explored to assign different weights to a set of classifiers to accommodate the varying conditional probabilities of labels in the target domain [1, 27]. Our problem setting falls into this category: we assume we have well-labeled instances in a source building, but do not have any labeled instances in the target building. We exploit different properties of a sensor point to perform the transfer learning: sensor's timeseries data is utilized to estimate a diverse set of classifiers to transfer knowledge from the source building to the target building; sensor names in the target building are used to compute the ensemble weight of classifiers during knowledge transfer.

3 Preliminary Results

The goal of this research is to provide a solution to inferring metadata for sensor time series, as much automated as possible. We have developed a suite of different techniques to recognize sensor type and co-location information, spanning from supervised to unsupervised approaches. For the supervised case, we propose two different approaches - fully automated inference and semi-automated inference. In the fully automated case, we transfer the information from other well-mapped buildings to automate the inference for a new building, while in the semi-automated case, we formulate an active learning algorithm that involves manual labeling, to relieve the inference process for a single building. The semi-automated algorithm is used when the first automated solution is inadequate, because it does not label all points. As for the unsupervised approach, we take the perspective from how much sensors are predictive of each other and formulate a sparse regression-based solution to cluster the points of the same type.

3.1 Fully Automated Inference

The Fully automated technique will leverage buildings that have already been manually mapped to an analytics engine and will attempt to learn how to map sensing and control points to a given analytics engine. The key challenge will be that no two buildings are identical twins: the types of sensors and controller and the format of the point names will vary greatly based on the building's equipment, its vendor and manufacturer, and the contractor who originally installed and configured it. Nonetheless, many buildings do have commonalities between them. We will use and develop techniques to find and exploit these commonalities whenever possible. The insight that guides our solution is that a sensor typically has two attributes - its recorded name (a text string) and its numerical readings generated over time. The sensor names are likely to be good indicators of metadata structure but are not likely to be consistent across buildings, while the sensor readings are more consistent across buildings but are less likely to be indicators of metadata structure. We will combine the complementary strengths of these two attributes to automatically recognize metadata structure in one building based on another building.

Specifically, we construct classifiers based on the data features since they are more likely to be consistent across buildings. However, a single supervised classifier might not perform well on all instances in the new building due to the inductive bias inherent in classifier training. Hence, we employ an ensemble of classifiers, where each classifier captures a different "perspective" in predicting the sensor type. When being applied to a new building, since different classifiers might be effective in predicting different instances, we appeal to the instance-specific local weighting method proposed in [21] to weight different classifiers while ensemble. In our solution, the weight is derived based on the consistency between a classifier's predictions and the instance's local clustering structure, which is estimated by the sensor names in the target building.

We have conducted experiments on a dataset with 7 days of data from over 2,500 sensors located in 3 commercial buildings across 2 different college campuses. Ground truth labels were created manually for all buildings. The types covered and number of streams per type is illustrated in Table 2. We then applied our techniques on each building to automatically infer the type information of the other 2 buildings. The results are illustrated in Figure 2: it indicates that this approach can automatically label at least 36% of the points with more than 85% accuracy, and in some cases labels up to 81% of the points with 96% accuracy. In contrast, simply training classifiers on one building and applying them directly to another building without reweighting achieves only 63% label accuracy on average. Our technique does not label all points, but those that are labeled have high label accuracy. Thus, it will only support a fraction of analytics algorithms. This limitation is to be addressed by our semi-automated algorithm.

3.2 Semi-Automated Inference

As demonstrated, the automated technique does not label everything; we need to address the unlabeled points. We develop a novel clustering-based active learning algorithm to perform semi-automated sensor type classification within a single commercial building. Semi-automated inference will require a technician to manually map points to the inputs of the analytics engine, and will try to minimize the number of points that must be manually mapped. The key insight is that points of the same type of input often have similarities in their string-based point names. This is often true because vendors will use certain naming conventions for the same type of points.

	Building		
Type	A	B	C
CO_2	16	52	0
Humidity	54	52	0
Air Pressure	142	216	215
Room Temp	159	231	208
Facility Operation Status	59	72	41
Facility Control	0	138	403
Setpoint	140	486	229
Air Flow Volume	14	172	9
Damper Position	0	290	10
Fan Speed	0	25	15
HW Supply Temp	27	1	0
HW Return Temp	15	1	0
CW Supply Temp	18	2	11
CW Return Temp	15	3	10
Supply Air Temp	20	17	3
Return Air Temp	6	2	4
Mixed Air Temp	5	2	3
Ice Tank Entering Temp	1	2	0
Ice Tank Leaving Temp	1	4	0
Occupancy	25	52	0
Timer	0	0	15
Sum	575	1124	1166

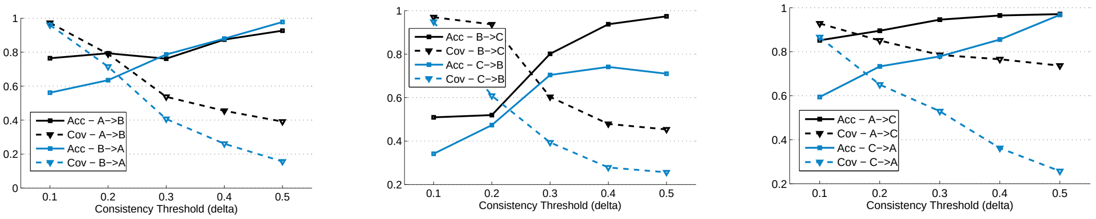
Table 2: Number of points by type for the 3 test buildings. “Temp” stands for “temperature”, “HW” for “hot water” and “CW” for “cold water”.

The key research question in active learning is evaluating the informativeness of unlabeled instances for querying. In our solution, we first cluster the unlabeled points based on their point names, and those in the same cluster are more likely to share the same label. Hence, the acquired labels can be propagated to the unlabeled neighbors in the same cluster to expedite classifier training. As such, we try to choose examples from different clusters, or from different points within a given cluster. The examples selected for labeling are chosen based on both their representativeness in the cluster and the informativeness of the cluster itself. We also define an adaptive approach to propagate labels based on the connectivity of the labeled examples’ neighborhoods.

To investigate the effectiveness of the proposed semi-automated solution for sensor type classification, we performed experimental comparisons against the state-of-the-art active learning algorithms on a large collection of real sensor stream data, which includes over 20 different sensor types and 2,500 sensors in three different commercial buildings. As shown in Figure 3, our method achieved increased classification performance with reduced amount of manual labels. Our solution achieved more than 92% accuracy with at least 16% fewer manual labels than the state-of-the-art active learning algorithms.

3.3 Unsupervised Clustering

We also approach the type inference problem from an unsupervised perspective: without any manual labels, we propose to cluster sensor time series based on the relationship among each other. The key insight is that one time series is likely to be predictive of other time series of the same type. We model the data with an autoregressive vector model and hypothesize that the sensor reading at each time point is more correlated with the readings from previous time points of the same type, rather than from

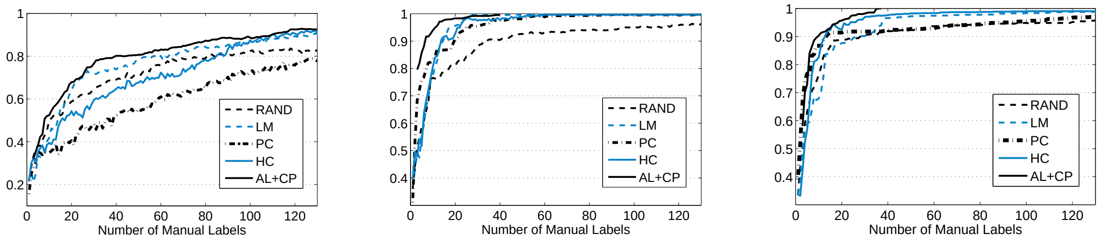


(a) A and B

(b) B and C

(c) A and C

Figure 2: Type classification accuracy (Acc) against labeled percentage (Cov) with transfer learning between different pairs of buildings (denoted as X->Y). As we increase the threshold, the coverage drops while the overall accuracy increases.



(a) Building A

(b) Building B

(c) Building C

Figure 3: Classification accuracy on three different buildings: comparing to the baselines, our method (AL+CP) is able to achieve better accuracy on all buildings with less labeled examples.

a different type, which would naturally impose sparsity on the relationships among the time series. Following the intuition, from a high level, the algorithm starts from estimating the causal relationship between each pair of time series by sparse regression. Based on the estimated relationships, we then apply standard spectral clustering algorithm [37] to identify the underlying type clusters. In addition, we focus on the high-dimensional regime, where the number of time series d to cluster can possibly exceed the length of each time series T, which casts a challenging unconstrained problem to solve where we have more variables to estimate than equations available. As the IoT market soars, it is likely to see millions of sensors deployed, but not millions of data readings from each of them. We also conducted experiments with a dataset comprising 204 streams of four different types from 51 rooms in an office building on campus. The proposed algorithm can cluster the points into type clusters with more than 90% accuracy (measured by Rand Index), compared to the accuracy of around 55% by baselines.

4 Proposed Research

Our preliminary results of the proposed techniques demonstrate promise of recognizing the type information of sensor time series. However, we argue that these solutions are designed for general-purpose classification and clustering, and therefore are expected to work for learning other kind of metadata information with proper modifications.

Building upon the results of type inference, we will refine our techniques for a more scalable roll out of building analytics engines. We will focus more on the inference of relationships among point, e.g., how a point or a group of points are related to another point, in terms of location and functionality. The key insight is that groups of points usually have similarities to other groups of points. For example, room temperature sensors are not only similar to other room temperature sensors, but they all have a distinct relationship to setpoint values, air handler control parameters, and so on. Following such insight, once we identify certain type of points, we can exploit such physical influences between points to infer the relationship.

We will explore and develop techniques to minimize the manual efforts required for inferring such relationships: first, we will establish new datasets containing the relational information of points and identify a set of important relationships with the help from a domain expert . Second, we will develop new features for representing the data and new distance metrics for clustering, in order to improve the performance of both the fully and semi-automated techniques. Third, we will explore the possibility of combing the fully and semi-automated techniques to further boost the performance, since the two are complementary in usage. Fourth, we will devise new algorithms for inferring the relationships among data. The approaches outlined previously make it possible to achieve the proposed objectives. More specifically, the following tasks define the scope of work and are ultimately the method for achieving the goals of this proposal.

4.1 Design Algorithms for Inferring Relationships

It is worth noting that the aforementioned methods in the preliminary study are not restricted to type classification only, whereas they ought to work for other kind of contextual information, e.g., for clustering or classification to recognize the relationships between points. The focus of the proposal is on relationship inference for sensor time series - how they are related to each other, for example, with regard to the HVAC zone, VAV/AHU they belong to. Table 3 summarizes the type of sensors and equipment as well as the relationships among them required by some common applications running in commercial buildings.

From the Table 3, we see that a few important relationships are required by most applications, `isLocIn` which captures the location information of a certain sensor or piece of equipment, `hasPoint` that describes what sensing and control points are associated with a piece of equipment (e.g., an AHU), and `hasPart` for representing hierarchical location information. As a result, our focus will be on inferring relationships as identified above, e.g., which points are co-located, and which set of sensors are associated with the same equipment.

Structurally, one building is divided into several smaller HVAC zones and conditioned in separate. Each HVAC zone usually consists of multiple rooms that are physically close to each other, and is conditioned by one or more VAVs. Intuitively, we would expect similar, if not simultaneous, responses in physical property (e.g., changes in temperature, or humidity level) of rooms and offices in the same zone. However, identifying such changes specific to a zone or room is challenging in that the actual responses in physical measurements comprise influences from a few different factors. For instance, room temperatures on the same side of a building is affected by sun activities, reflected as the diurnal patterns observed in Figure 4a and 4c. In addition, based on the findings from the clustering analysis in § 3.3, we see that the type of a point is another driven factor - numerical readings from the same type of sensors would show similar trends.

Mixture of Latent Factors We hypothesize that each sensor time series is generated from a mixture of latent driven factors, e.g., diurnal pattern, type factor and other local influences. We will explore techniques, such as Hidden Markov Models or Principal Component Analysis, to model the underlying factors affecting the sensor time series. Once we are able to decompose the signal into different driven factors, we can subtract out the dominant ones, e.g., the type factor. We will next investigate the effectiveness in identifying relational clusters with the features and distance metric discovered from early steps.

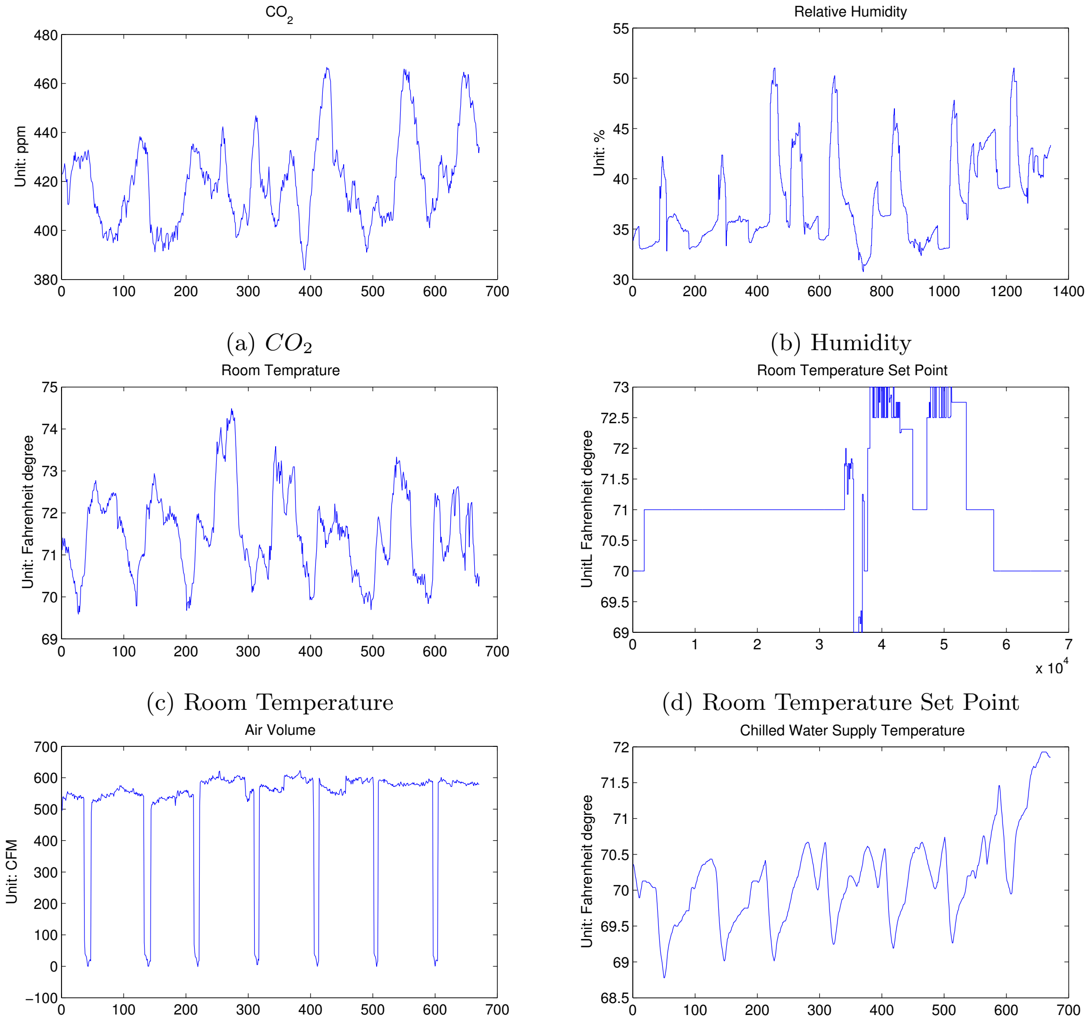
4.2 Establish New Dataset and Evaluation Metrics

Another key piece in the proposed work will be to establish a new dataset containing relational information of points that are critical to applications, e.g., which VAV or AHU each point belongs to. We will work closely with Trane[4] to establish such a new data sets that are representative of the problems faced by Trane. The data sets will include a minimum of 20 manually mapped buildings, i.e. buildings that were originally configured by another company such as Johnson Controls and that were manually mapped by Trane personnel to a Trane analytics engine. Thus, these buildings will have both the original point names and any other metadata, and will also have the mapping from the original configuration to the inputs of the analytics engine by Trane, along with the numerical readings. This dataset will implicitly define an evaluation metric: how many of and with what accuracy the original configuration can be automatically mapped to the correct inputs. However, we will also seek guidance from Trane personnel to understand the relative importance of each kind of metadata information, e.g., location vs sensor type, in order to create a better weighted evaluation metric for the metadata inference.

[4] `http://www.trane.com/`

Entities	Occupancy Modeling [29]	Energy Apportionment [28]	Web Displays [2]	Model-Predictive Control [47]	Participatory Feedback [31]	Fault Detection and Diagnosis [42]	NILM [36]	Demand-Response [50]
Sensors								
Temp Sensor	X					X		
CO2 Sensor	X							
Occ Sensor	X	X			X			
Lux Sensor		X			X			
Power Meter	X	X	X		X		X	X
Airflow Sensor			X					
Equipment								
Lighting	X	X			X			
Reheat Valve			X			X		
VAV			X	X				
AHU				X		X		
Chilled Water			X			X		
Hot Water			X			X		
Locations								
Building				X		X		
Floor				X	X		X	
Room	X	X	X	X	X		X	
HVAC Zone	X		X	X				
Lighting Zone	X				X			
Relationships								
Sensor `isLocIn` Loc.	X	X			X		X	
Equip `isLocIn` Loc.		X			X		X	X
Equip `hasPoint` Sensor	X		X			X	X	X
Zone `hasPart` Room	X			X	X			

Table 3: This table shows at a high level which entities and relationships are required by some representative applications running in commercial buildings.



(a) CO_2

(b) Humidity

(c) Room Temperature

(d) Room Temperature Set Point

(e) VAV Airflow Volume

(f) Chilled Water Supply Temperature

Figure 4: Different type of sensors show different characteristics, but can have similar diurnal patterns.

4.3 Create New Features and Distance Metrics

Once the data set and evaluation metrics are defined, we will first formulate new features to be used for classification - our current work uses several features of both the data and the name associated with a particular point. The data features are summarized in Table 4 and include things like the maximum, minimum, the quartiles, and the moments. The name features are the k-mers of length 3 and 4 for all point names: all the possible substrings of length k contained in a string. For example, {`zone, temp, rmi`} will yield a set {`zon, one, tem, emp, rmi`} with k=3.

Since we will study other kind of metadata, we will explore a broader set of features to better represent the raw sensor time series, including “shapelets” [52, 53], or repetitive shapes in the data stream, and correlations with already-labeled data streams. “Shapelets” are subsequences of time series that are representative of a class, and have proven empirically effective in time series classification. Such an example could be the signature response of CO2 sensors when people walk into an unoccupied room. Another possible direction is to leverage the observed correlations with already-labeled data streams. For instance, if we already label a point A as “VAV temperature setpoint” and always

Category	Statistical Function	Acronym
Extrema	Minimum	min
	Maximum	max
Average	Median	med
	Root Mean Square	rms
Quartiles	1st, 3rd Quartile	q1, q3
	Inter-quartile range	iqr
Moments	Variance	var
	Skewness	skew
	Kurtosis	kurt
Shape	Linear Regression Slope	slope

Table 4: Features extracted at window level from each sensor time series.

observe corresponding changes in another point B whenever there is a change in A, then B is likely to be a temeprature measurement controlled by A. We expect feature exploration as such to improve the performance of both the fully automated and semi-automated techniques, which fundamentally forms the bottleneck of metadata inference.

With the refined feature set, we will set out to devise new distance metrics for clustering - our preliminary work clusters points simply based on the Euclidean distance between the vectors of *k*-mers. In the proposed work, we will explore customized distance metrics that help better capture the similarity/dissimilarity among time series data. We will apply techniques such as distance metric learning [51] to the data represented with new features, and perform statistical test, such as the permutation test, to decide the usefulness of the new distance metric. Once the new distance metric passes the significance test, we will proceed to develop new techniques for inferring the relationships between points.

4.4 Combine the Semi- and Fully Automated Techniques

Our current work uses transfer learning for fully-automated mapping and active learning for semi-automated inference. In the proposed work, we will design new techniques to combine these two for cases where we have both manually mapped buildings and manually labeled examples. The simplest such approach would be to apply transfer learning first and then active learning to cover any remaining, un-mapped points. The results are illustrated in Figure 5 where we do not see much improvement from the simple combination over simply active learning. This approach does not use the manually labeled examples to improve transfer learning, or the other way around.

There are two possible directions to better engage the two techniques: on the one hand, if we can accurately estimate the quality of each already manually labeled building, we can query buildings with high confidence for automatic labels with *no cost* during active learning; on the other hand, when acquiring a manual label of an example, which we assume is always correct, the information could be used to help reweight the manually labeled buildings. For the first scenario, graphical models, such as Conditional Random Field [33] (CRF), are a good candidate framework for jointly estimating the confidence of each labeled building and measuring the informativeness of examples for active learning. The key question is how to design proper features for such models. As for the second direction, we will develop new algorithms that use manually labeled examples to help reweight both the learned classifiers and the distance metrics, in order to improve the effectiveness of transfer learning.

4.5 Incorporate With Existing Tools and Documentation

After creating new automatic mapping techniques, we will incorporate them with existing tools, including the techniques described earlier to parse point names [6] as well as those to match point

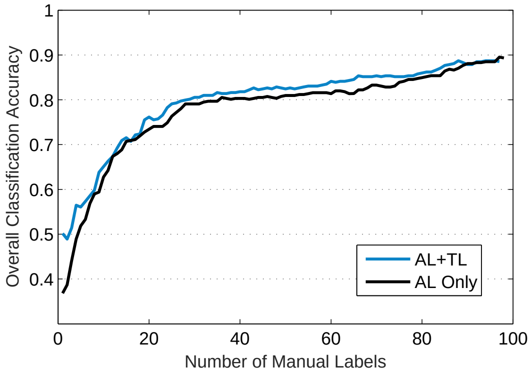


Figure 5: A simple marriage of transfer learning and the active learning-based approaches (AL+TL) does not significantly reduce the amount of manual labeling.

names to a manually constructed dictionary of building terms [44]. We will also formulate common point name structures based on the data sets provided by Trane. These techniques will then be combined with the automatic mapping techniques produced by this project, and the improvement in mapping accuracy will be evaluated. The software produced is expected to be a stand-alone tool that can be used by any analytics engine based on manually mapped buildings and manually labeled examples. In addition to this software artifact, we will produce documentation on the automatic mapping techniques, the parsing techniques, and the dictionary of building terms. The final outcome of this project will be documentation of techniques developed such that they can be integrated into other software products.

5 Evaluation Plan

Datasets Currently we have a dataset consisting of more than 4,000 points from 4 office buildings. Besides, we will have a new dataset of more than 20 buildings with ground truth labels collected with the help from Trane, as described in § 4.2. These two datasets will be the basis of our evaluation.

Effectiveness of Features and Distance Metrics To evaluate the effectiveness of new features for type classification, we will use the new set of features for the fully automated technique. Similar to the setup in the preliminary study, we will use one building for training and another different building for testing. We will measure how many and with what accuracy the points in the testing building can be automatically labeled, with regard to their type. As a baseline, performance of the original set of features will be measured with the same metrics.

As for the distance metric for generating clusters, we will quantify the quality of clustering with standard metrics such as rand index. We will also use the new distance metric in both the fully and semi-automated techniques. For the former, we will measure the performance with the same metrics as above, while for the latter we will measure the labeling accuracy versus the number of manual labels required. In both cases, we will use as a baseline the current clustering algorithm simply based on Euclidean distance.

Advantage of Combining the Fully and Semi-Automated Inference Techniques Once the exploratory analysis and the algorithmic design is complete, the combined inference technique will be evaluated in terms of the accuracy with which it labels new buildings and the amount of manual labels required. This will be evaluated by training the algorithm on a subset of the example buildings and testing on the others, by training on a subset of points and testing on the others, or by combining the two approaches. The combined algorithm will be compared with two different baselines: the original fully automated and semi-automated techniques. Since the new technique will combine the strengths of both techniques, it is expected to label points with higher accuracy and fewer manual labels.

Accuracy of Algorithms for Relationship Inference The mixture model is expected to characterize the underlying driven factors, and we will be able to subtract out the dominant factors. On the residue, we will extract the set of features identified in § 4.3 and apply classification and clustering techniques used in § 3.1 and § 3.2 with cross-validation, to measure the accuracy of inference of different set of important relationships. Additionally, we will revisit the evaluation metrics applied to the approach by communicating results with Trane personnel and asking for feedback.

6 Expected Publications

1. "Towards Automatic Spatial Verification of Sensor Placement in Building." [23] - Published
2. "Clustering-based Active Learning on Sensor Type Classification in Buildings." [25] - Published
3. "The Building Adapter: Towards Quickly Applying Building Analytics at Scale." [24] - Published
4. "Automated Metadata Construction to Support Portable Building Applications." [6] - Published
5. "High-dimensional Time Series Clustering with Provable Guarantee." - In Submission
6. "Brick v1.0 - Towards a Unified Metadata Schema for Buildings." - In Submission
7. One paper on *relationship inference of sensor time series.* - Ongoing

7 Timeline

Date	**Task**
July - September	Data collection and ground truth labeling
September - January	Data analysis and algorithm development
January - March	Evaluation and paper submission
April - June	Thesis writing and dissertation defense

References

[1] C. G. Atkeson, A. W. Moore, and S. Schaal. Locally weighted learning. *Artif. Intell. Rev.*, 11(1-5), Feb. 1997.

[2] B. Balaji, H. Teraoka, R. Gupta, and Y. Agarwal. Zonepac: Zonal power estimation and control via hvac metering and occupant feedback. In *Proceedings of the 5th ACM Workshop on Embedded Systems For Energy-Efficient Buildings*, pages 1–8. ACM, 2013.

[3] B. Balaji, C. Verma, B. Narayanaswamy, and Y. Agarwal. 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*, pages 13–22. ACM, 2015.

[4] B. Balaji, J. Xu, A. Nwokafor, R. Gupta, and Y. Agarwal. Sentinel: Occupancy based hvac actuation using existing wifi infrastructure within commercial buildings. In *Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems*, SenSys '13, 2013.

[5] M.-F. Balcan, A. Broder, and T. Zhang. Margin based active learning. In *Proceedings of the 20th Annual Conference on Learning Theory*, COLT'07, Berlin, Heidelberg, 2007. Springer-Verlag.

[6] A. A. Bhattacharya, D. Hong, D. Culler, J. Ortiz, K. Whitehouse, and E. Wu. Automated metadata construction to support portable building applications. In *Proceedings of the 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments*, pages 3–12. ACM, 2015.

[7] S. Bickel, M. Brückner, and T. Scheffer. Discriminative learning for differing training and test distributions. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, 2007.

[8] R. Caruana. Multitask learning. *Machine Learning*, 28(1), 1997.

[9] Y. Chen and L. Tu. Density-based clustering for real-time stream data. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '07, 2007.

[10] D. A. Cohn, Z. Ghahramani, and M. I. Jordan. Active learning with statistical models. *J. Artif. Int. Res.*, 4(1), Mar. 1996.

[11] Comfy. `https://gocomfy.com/`.

[12] W. Dai, Q. Yang, G.-R. Xue, and Y. Yu. Boosting for transfer learning. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, New York, NY, USA, 2007.

[13] S. Dasgupta and D. Hsu. Hierarchical sampling for active learning. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, New York, NY, USA, 2008.

[14] H. Daumé, III and D. Marcu. Domain adaptation for statistical classifiers. *J. Artif. Int. Res.*, 26(1), May 2006.

[15] S. Dawson-Haggerty, A. Krioukov, J. Taneja, S. Karandikar, G. Fierro, N. Kitaev, and D. Culler. Boss: Building operating system services. In *Proceedings of the 10th USENIX Conference on Networked Systems Design and Implementation*, NSDI'13, 2013.

[16] R. Dhamankar, Y. Lee, A. Doan, A. Halevy, and P. Domingos. imap: Discovering complex semantic matches between database schemas. In *Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data*, SIGMOD '04, pages 383–394, New York, NY, USA, 2004. ACM.

[17] A. Doan, P. Domingos, and A. Y. Halevy. Reconciling schemas of disparate data sources: A machine-learning approach. In *Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data*, SIGMOD '01, pages 509–520, New York, NY, USA, 2001. ACM.

[18] U. DOE. Better buildings challenge. *http://www4.eere.energy.gov/challenge/sites/default /files/uploaded-files/may-recognition-fs-052013.pdf (Feb. 26, 2014)*, 2013.

[19] B. Dong and K. Lam. A real-time model predictive control for building heating and cooling systems based on the occupancy behavior pattern detection and local weather forecasting. *Building Simulation*, 7(1), 2014.

[20] Y. Freund, H. Seung, E. Shamir, and N. Tishby. Selective sampling using the query by committee algorithm. *Machine Learning*, 28(2-3), 1997.

[21] J. Gao, W. Fan, J. Jiang, and J. Han. Knowledge transfer via multiple model local structure mapping. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, 2008.

[22] Gartner Inc. `http://www.gartner.com/newsroom/id/2636073`.

[23] D. Hong, J. Ortiz, K. Whitehouse, and D. Culler. Towards automatic spatial verification of sensor placement in buildings. In *Proceedings of the 5th ACM Workshop on Embedded Systems For Energy-Efficient Buildings*, pages 1–8. ACM, 2013.

[24] D. Hong, H. Wang, J. Ortiz, and K. Whitehouse. 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*, pages 123–132. ACM, 2015.

[25] D. Hong, H. Wang, and K. Whitehouse. Clustering-based active learning on sensor type classification in buildings. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 363–372. ACM, 2015.

[26] J. Huang, A. Smola, A. Gretton, K. Borgwardt, and B. Schölkopf. Correcting sample selection bias by unlabeled data. In *Advances in Neural Information Processing Systems*, volume 19. The MIT Press, Cambridge, MA, 2007. Pre-proceedings version.

[27] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. Adaptive mixtures of local experts. *Neural Comput.*, 3(1), Mar. 1991.

[28] M. Jahn, T. Schwartz, J. Simon, and M. Jentsch. Energypulse: tracking sustainable behavior in office environments. In *Proceedings of the 2nd International Conference on Energy-Efficient Computing and Networking*, pages 87–96. ACM, 2011.

[29] D. Jung, V. B. Krishna, N. Q. M. Khiem, H. H. Nguyen, and D. K. Yau. Energytrack: Sensor-driven energy use analysis system. In *Proceedings of the 5th ACM Workshop on Embedded Systems For Energy-Efficient Buildings*, pages 1–8. ACM, 2013.

[30] K. Kapitanova, E. Hoque, J. A. Stankovic, K. Whitehouse, and S. H. Son. Being SMART about failures: assessing repairs in SMART homes. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, UbiComp '12, 2012.

[31] A. Krioukov, S. Dawson-Haggerty, L. Lee, O. Rehmane, and D. Culler. A living laboratory study in personalized automated lighting controls. In *Proceedings of the third ACM workshop on embedded sensing systems for energy-efficiency in buildings*, pages 1–6. ACM, 2011.

[32] A. Krioukov, G. Fierro, N. Kitaev, and D. Culler. Building application stack (bas). In *Proceedings of the Fourth ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings*, BuildSys '12, 2012.

[33] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the eighteenth international conference on machine learning, ICML*, volume 1, pages 282–289.

[34] J. Lu and K. Whitehouse. Smart blueprints: automatically generated maps of homes and the devices within them. In *Proceedings of the 10th international conference on Pervasive Computing*, Pervasive'12, 2012.

[35] J. Madhavan, P. A. Bernstein, A. Doan, and A. Halevy. Corpus-based schema matching. In *Proceedings of the 21st International Conference on Data Engineering*, ICDE '05, pages 57–68, Washington, DC, USA, 2005. IEEE Computer Society.

[36] A. Marchiori and Q. Han. Using circuit-level power measurements in household energy management systems. In *Proceedings of the First ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings*, pages 7–12. ACM, 2009.

[37] A. Y. Ng et al. On spectral clustering: Analysis and an algorithm. 2002.

[38] H. T. Nguyen and A. Smeulders. Active learning using pre-clustering. In *Proceedings of the Twenty-first International Conference on Machine Learning*, ICML '04, New York, NY, USA, 2004.

[39] U. D. of Energy Better Buildings program. Total annual cost of energy in the commercial and industrial sector. Mar. 2015.

[40] S. J. Pan and Q. Yang. A survey on transfer learning. *Knowledge and Data Engineering, IEEE Transactions on*, 22, Oct 2010.

[41] E. Rahm and P. A. Bernstein. A survey of approaches to automatic schema matching. *the VLDB Journal*, 10(4):334–350, 2001.

[42] J. Schein, S. T. Bushby, N. S. Castro, and J. M. House. A rule-based fault detection method for air handling units. *Energy and Buildings*, 38(12):1485–1492, 2006.

[43] A. Schumann, J. Ploennigs, and B. Gorman. Towards automating the deployment of energy saving approaches in buildings. In *Proceedings of the 1st ACM Conference on Embedded Systems for Energy-Efficient Buildings*, BuildSys '14, 2014.

[44] A. Schumann, J. Ploennigs, and B. Gorman. Towards automating the deployment of energy saving approaches in buildings. In *Proceedings of the 1st ACM Conference on Embedded Systems for Energy-Efficient Buildings*, pages 164–167. ACM, 2014.

[45] B. Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.

[46] H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2), 2000.

[47] D. Sturzenegger, D. Gyalistras, M. Morari, and R. S. Smith. Semi-automated modular modeling of buildings for model predictive control. In *Proceedings of the Fourth ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings*, pages 99–106. ACM, 2012.

[48] E. S. p. U.S. Environmental Protection Agency. Useful facts and figures. June 2007.

[49] J. Warnick. 88 acres: How microsoft quietly built the city of the future. *http://www.microsoft.com/en-us/stories/88acres/88-acres-how-microsoft-quietly-built-the-city-of-the-future-chapter-1.aspx (May 8, 2015)*, 2012.

[50] T. Weng, B. Balaji, S. Dutta, R. Gupta, and Y. Agarwal. Managing plug-loads for demand response within buildings. In *Proceedings of the Third ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings*, pages 13–18. ACM, 2011.

[51] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell. Distance metric learning with application to clustering with side-information.

[52] L. Ye and E. Keogh. Time series shapelets: A new primitive for data mining. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, 2009.

[53] J. Zakaria, A. Mueen, and E. Keogh. Clustering time series using unsupervised-shapelets. In *Proceedings of the 2012 IEEE 12th International Conference on Data Mining*, ICDM '12, 2012.