Data Science for Building Energy Management: a review

Miguel Molina-Solana[a,b], María Ros[a,*], M. Dolores Ruiz[a], Juan Gómez-Romero[a], M.J. Martin-Bautista[a]

[a] *Department of Computer Science and Artificial Intelligence, Universidad de Granada*
[b] *Data Science Institute, Imperial College London*

Abstract

The energy consumption of residential and commercial buildings has risen steadily in recent years, an increase largely due to their HVAC systems. Expected energy loads, transportation, and storage as well as user behavior influence the quantity and quality of the energy consumed daily in buildings. However, technology is now available that can accurately monitor, collect, and store the huge amount of data involved in this process. Furthermore, this technology is capable of analyzing and exploiting such data in meaningful ways. Not surprisingly, the use of data science techniques to increase energy efficiency is currently attracting a great deal of attention and interest. This paper reviews how Data Science has been applied to address the most difficult problems faced by practitioners in the field of Energy Management, especially in the building sector. The work also discusses the challenges and opportunities that will arise with the advent of fully connected devices and new computational technologies.

1. Introduction

There is a general consensus in the world today that human activities are having a negative impact on the environment and have accelerated both global warming and climate change. These environmental threats have been intensified by the emissions produced by the energy required for the lighting and HVAC (heating, ventilation and air-conditioning) systems in building constructions. According to the International Energy Agency (IEA), residential and commercial buildings are responsible for up to 32% of the total final energy consumption. In fact, in most IEA countries, they account for approximately 40% of the primary energy consumption. Similar statistics are given by the *World Business Council for Sustainable Development (WBCSD)* within the framework of its *Energy Efficiency in Buildings (EEB)* project[1]. Also provided is a comprehensive review [1] of the state of the art in building energy use (with a primary focus on energy demand).

These data indicate that inefficient energy management in aging buildings combined with rising construction activity in developed countries will cause energy consumption to soar in the near future and heighten the negative impacts associated with this consumption. Moreover, variable energy costs call for the implementation of more intelligent strategies to adapt and reduce energy consumption as well as to find alternative and sustainable energy sources. The relevance of these issues is clearly reflected in the research priorities of the European Union, as stated in its Horizon2020 Societal Challenge "Secure, Clean and Efficient Energy". This work program targets a significant reduction in energy consumption by 2020 in the transportation and building sectors, both of which have great potential for energy savings.

Increasing energy efficiency is a two-fold process. Not only does it involve the use of affordable energy sources, but also the improvement of current energy management procedures and infrastructures. The

*Corresponding author

Email addresses: miguelmolina@imperial.ac.uk (Miguel Molina-Solana), marosiz@decsai.ugr.es (María Ros), mdruiz@decsai.ugr.es (M. Dolores Ruiz), jgomez@decsai.ugr.es (Juan Gómez-Romero), mbautis@decsai.ugr.es (M.J. Martin-Bautista)

[1] http://www.wbcsd.org/web/eeb.htm

latter includes the optimization of energy generation and transportation based on user demand [2], one of the most important issues for energy companies. In this regard, computer-aided approaches have recently come into the spotlight. More specifically, increased data awareness in companies has led to the development of solutions based on Data Mining, a research area that studies how to automatically discover non-trivial knowledge from data, and Data Science, which encompasses a wide range of techniques and more complex datasets.

In the area of building energy management, Data Science is now used to address problems such as the following: (i)the prediction of energy demand in order to adapt production and distribution; (ii) the analysis of building operations as well as of equipment status and failures to optimize operation and maintenance costs; (iii) the detection of energy consumption patterns to create customized commercial offers and to detect fraud. This requires collecting data pertaining to building operation and user behavior. These data must also be interpreted to implement adapted energy management policies. The information collected may come from very heterogeneous sources ranging from in-site sensors (located in the equipment and in the immediate environment) to external parameters (e.g. weather, energy costs, etc.). These advances have also signified a shift in the perception of who owns these data and who benefits from them [3]. Customers are increasingly aware of the importance of their actions and the value of the data that they generate. In this sense, they have become actors with a key role in the energy efficiency landscape.

This paper reviews different data science techniques and explains how they have been employed to deal with the difficult challenges faced by building energy management. As reflected in recent literature on the topic, classification and clustering methods are frequently used for this purpose, but there is still room for improvement in relatively underexplored areas, such as frequent and temporal pattern discovery for load prediction. Also discussed are future trends in Data Science, which will lead to new methods and tools capable of the more intelligent processing of large amounts of data collected from multiple distributed devices. Although there are other reviews on automatic techniques for building efficiency assessment [4, 5], and on classification methods for load and energy consumption prediction [6], this work examines and discusses a broader set of data science techniques, and their applications to the different aspects of building energy management.

The paper is structured as follows. After an introduction to data science techniques (Section 2), Section 3 summarizes recent work in Energy Data Science and situates it in the context of the current requirements and needs of building energy managers. Section 4 discusses the data science techniques employed in various fields related to building energy management. Finally, Section 5 provides an overview of new approaches that are expected to lead to research advances, and concludes with recommendations and guidelines for the future.

2. Data Science

Over the years, technological tools have benefited a wide range of domains, and Energy Efficiency and Management is no exception. Developments in various areas of Information and Communications Technology (ICT), such as Control and Automation, Smart Metering, Real-time Monitoring, and Data Science, have had a tremendous impact on this field. As is well known, Data Science builds systems and algorithms to discover knowledge, detect patterns, and generate useful insights and predictions from large-scale data. It encompasses the whole data analysis process, which begins with data extraction and cleaning, and extends to data analysis, description and summarization. The results is the prediction of new values and their visualization. Data Science thus involves mathematical and statistical analysis, combined with information technology tools.

However, deriving insights from data is not only achieved by using such techniques. The expert must also manage and interpret the data in order to obtain valuable knowledge. As shown in Figure 1, the process starts with the collection of raw data. After that, it is necessary to clean the data, and select the subset that has the relevant information. For that purpose, the expert applies filters to the data or formulates queries that will eliminate irrelevant information. At this step, it is also when additional sources of information might be integrated and fused with the original data to provide further knowledge. Once the data are prepared for use, an exploratory analysis (including visualization tools) can help decide which methods or

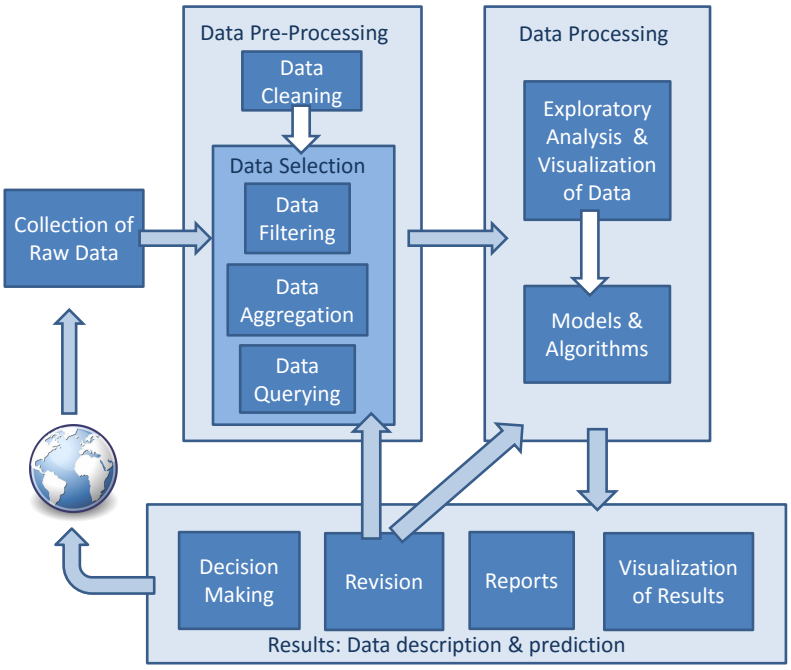

Figure 1: Data science process

algorithms are most effective to obtain the desired knowledge. The final process will lead to a set of results that guide the decision-making, which again, might rely on visualisation.

Based on the preliminary outcomes, the whole process might need to be tuned to obtain better results. This could entail setting new parameter values or adding/discarding new sets of data. Since such decisions cannot be made automatically, the participation of the expert in the analysis of the results is a crucial factor.

From a more technically perspective, Data Science comprises a set of techniques and tools which pursue different goals and depart from different situations. Some of the most popular techniques are classification, clustering, regression and association rule mining. Although these techniques have been the most frequently applied in Energy Efficiency and Management, others, which are not so well known (e.g. sequence analysis and anomaly detection), are also useful in providing solutions for building energy problems.

Classification When classifying a set of objects, the objective is to predict the class of each one on the basis of their attributes. *Decision trees* (i.e. a kind of flowchart for the classification of new data) are a common way of performing and visualizing that classification [7]. Decision trees can be generated by many different algorithms, though the most well known are *CLS*, *ID3*, *C4.5*, *C5.0*, and *CART*. *Random Forest* is another classification technique that constructs a set of decision trees and then predicts the class by aggregating the values obtained with each tree (e.g. by using the mode or mean). This method corrects overfitting (when the models from the learning algorithm perform very well on the training set, at the cost of an increased error on the validation set), a common practical difficulty in decision trees. *Support Vector Machine (SVM)* [8] is a technique that is also used for classification. SVMs perfom classification tasks by constructing a hyperplane (or a set of hyperplanes) in a multidimensional space to separate the data (regarded as points in the space) into classes. Once the hyperplanes is constructed, it classifies the new examples according to the previously specified decision boundaries.

Bayesian classification, genetic algorithms, and neural Networks have been also employed in classification tasks. There are various approximations that use probabilistic classifiers based on the Bayes' theorem, but as a consequence, there are strong independence assumptions between the variables in-

volved [9]. Class prediction with genetic programming algorithms [10] are based on chromosome-like structures that can be combined and/or mutated with other chromosomes to create new individuals. Neural Networks (NNs) are able to predict new observations from existing ones by means of interconnected elements called neurons [11]. The main advantage of NNs is that they are robust and tolerant of errors. A self-organizing map (SOM) is a type of artificial neural network that is trained by unsupervised learning to produce low-dimensional views of high-dimensional data. Another well-known classification method is that of k-Nearest Neighbors, which classifies and object by the majority vote of its k neighbors. In other words, an object is assigned to a category based on the category of its k nearest neighbors [12].

Regression The main objective of regression analysis is to numerically estimate the relationship between variables. This involves ascertaining whether variables are independent. When they are not, it is then necessary to discover the type of dependence of their relation [13]. Regression analysis is widely used in prediction and forecasting as well as to understand how the values of dependent variables change while those of independent variables remain fixed. Linear and non-linear (polynomial, logistic, etc.) regression methods are mainly used for this purpose. In linear regression, the model assumes that variables are a linear combination of the parameters. Examples of linear regression methods are linear least squares, Bayesian linear regression, and generalized linear models (GLM). Nevertheless, linear models often do not provide a good fit to reality, and then non-linear models are required. In this case, classification-based techniques, such as support vector regression or k-Nearest Neighbors, can also be used for regression. In particular, ARMA (Autoregressive Moving Average) or ARIMA (Autoregressive Integrated Moving Average) are capable of predicting the future values of time series, based on past values. The relationship between variables can also be statistically measured by means of the standard deviation, Pearson correlation, and other correlation coefficients.

Clustering Clustering is the separation of objects into groups (clusters) based on their degree of similarity [14]. It is unsupervised, because there is no previous knowledge of the classes to which the objects can be assigned. Depending on the criterion used to measure similarity, there are different models of cluster analysis: (i) connectivity models, based on distance connectivity (e.g. hierarchical clustering); (ii) centroid models, which are constructed by assigning objects to the nearest cluster center (e.g. k-means or k-medians); (iii) distribution models using statistical distributions (e.g. expectation-maximization algorithm); (iv) density models where clusters are defined based on high-density areas in the data set; (v) graph-based models in which the data are expressed as graphs. A further distinction can be made between hierarchical and non-hierarchical models. Hierarchical models take the form of a hierarchy of clusters (e.g. hierarchical tree or agglomerative hierarchical clustering) whereas non-hierarchical models are based on a plain cluster organization without any relations between them but rather group a set of units into a pre-determined number of groups, using an iterative algorithm that optimizes a chosen criterion.

Clustering techniques are often a first step in a classification problem when there is no information about the classes. In an initial phase, clustering is used to identify groups of objects with similar features. Classification techniques are then applied to assign new objects to these groups. When there is no previous information about the objects, clustering techniques can also be used for classification purposes.

Association rules (ARs) Association rules are a useful tool for the representation of new information extracted from raw data and comprehensively expressed for decision-making in the form of implication rules of the type $A \rightarrow B$ [15]. These rules depict the frequent co-occurrence of attributes with a high reliability in a database. For example "most transactions containing beer also contain diapers" is an association rule that could be found in a supermarket database. The Apriori algorithm and its adaptations (e.g. generalized rule induction algorithm) are the most widely used, though there are others, such as the FP-Growth and ECLAT algorithms, which improve scalability in very large datasets [16, 17]. Association rules now have more sophisticated versions that not only capture correlations but other kinds of association as well. Examples include the following: (i) generalized ARs, which use

a concept hierarchy to obtain rules relating the different granularities of items; (ii) quantitative ARs, which deal with categorical and quantitative data; (iii) gradual dependence rules, which capture data tendencies by obtaining rules of the type "the more/less A → the more/less B"; (iv) sequential rules, which identify relationships between items while considering some ordering criterion (e.g. time).

Sequence discovery Sequence discovery comprises techniques that identify statistically relevant patterns in data, whose values are distributed in order [18]. Frequent problems in sequence analysis include the following: (i) the extraction of sequence information using techniques such as Motif Mining (MM); (ii) the detection of frequently occurring patterns; (iii) the search for similar sequences with a time lag by means of autocorrelation methods such as the ACF (Autocorrelation Function) and PACF (Partial Autocorrelation Function); (iv) the recovery of missing sequence members. Many of the other previously explained techniques are also capable of dealing with this kind of data.

Anomaly or outlier detection The objective of detecting anomalies is to identify items, events, or observations that deviate from expected patterns or from the usual behavior of other data items [19]. The discovery of anomalous items is crucial in the resolution of bank fraud, medical diagnoses, errors in data transmission, noise, etc. Since the previously described techniques are based on the identification/classification of similar items, most frequent patterns, etc., variations of these methods can also be employed for anomaly discovery. Methods used for this purpose are the following: density-based techniques, correlation, clustering, searching deviations from association rules, and combinations of diverse techniques using, for example, feature bagging or score normalization.

Time series analysis Time series analysis is performed on time-series data (i.e. data points that are recorded over time) in order to model data and then use the model to predict or monitor future values of the time series [20]. The most frequently used methods include the following: (i) methods for exploratory analysis (e.g. autocorrelation, trend analysis, wavelets, etc.); (ii) prediction and forecasting techniques (e.g. regression methods, signal estimation, etc.); (iii) classification methods which assign a category to patterns in the series; (iv) segmentation which aims to identify a sequence of points sharing specific properties (e.g. ARMA or ARIMA).

Most of the previously mentioned techniques have a fuzzy extension that allows them to process with imprecise and uncertain data in various domains [21]. Fuzzy logic allows a non-strict representation of object membership to a set, thus avoiding the problem of hard boundaries that are often present in basic techniques, such as clustering and classification methods. For example, fuzzy k-means is a clustering method that has proved effective in many scenarios since it permits the assignment of data elements to one or more clusters [22]. Fuzzy approaches also allow a more human-friendly representation of the extracted knowledge; since fuzzy association rules are easier to interpret than purely numerical rules [23].

3. Applications of Data Science for Building Energy Management

Data science techniques have been frequently used to support and improve basic aspects of Energy Efficiency and Management. Accordingly, this section focuses on applications of Data Science that are capable of doing the following: (1) predicting the energy demand required for the efficient operation of a building; (2) optimizing building operation; (3) enabling building retroffiting; (3) verifying the operational status and failures of building equipment and networks; (4) analyzing the economic and commercial impact of user energy consumption; (5) detecting and preventing energy fraud.

3.1. Prediction of building energy load

Energy demand, or energy load, refers to the amount of energy required at a certain time instant or interval. In particular, HVAC systems focus on thermal loads, which refer to the quantity of heating and cooling energy that must be added or removed from the building to keep its occupants comfortable. Thermal loads can be classified as *internal loads*, when heat transfer/influence is produced by elements (e.g. lightning,

equipment, or people) within the building during its operation or as *external loads* when the source of the influence is due to external (generally environmental) factors, such as, sun, air and moisture.

The detection of common patterns of building loads can be extremely complex because of the number of interrelated terms that must be dealt with. For example, Kusiak et al. [24] created a model to select the parameters required for a non-linear mapping of climate measurements. Neural networks were then used to predict the thermal loads (for heating and cooling) of a building.

Electricity loads have also been modelled with clustering methods. In line with this, Prahastono et al. [25] presented an overview of common clustering methods (e.g. hierarchical, k-means, fuzzy k-means, follow the leader and fuzzy relation) and compared their effectiveness in the classification of customers and the generation of electricity load profiles (in terms of average retail price and net generation). In contrast, Yu et al. [26] proposed the use of decision trees to develop predictive models for building energy demand since they are more easily interpreted than other classification techniques.

Peak demand is the term used in energy demand management for a time period in which electrical power is expected to be provided at a significantly higher supply level than average. When the generation and supply levels are not able to meet the demand, this can result in power outages and load shedding, which are evident sources of customer dissatisfaction. It is thus extremely important to develop procedures that can anticipate the peak demand for any given day and model both the short-term and medium-term energy load. Traditionally, conventional analytical methods (such as regression) have been most frequently used for this purpose. However, in the last few years, other data science techniques have been applied to build predictive models from historical data. These models are extremely valuable because they are able to learn the temporal trends and extrapolate them to new scenarios.

For example, Li [27] developed a decision tree to analyze the impact of external factors on energy peaks. The objective of this research was to learn different models for building energy loads in different climates by means of two regression algorithms, and use the information from the decision trees to predict the maximum expected load for the next day. In contrast, Yang et al. [28] used C4.5 classification methodology to analyze a combination of internal and external ambient conditions, and to determine the influence of external conditions on a building's internal user comfort.

Load prediction may also help when predicting energy peaks. Fan et al. [29] proposed using a combination of models for predicting next-day energy consumption and peak power demand. They structured their proposal in three stages. The first stage involved extracting abnormal building energy consumption profiles using feature selection and clustering analysis. In the second stage, optimal inputs for eight predictive algorithms were selected by means of recursive feature elimination. Finally, in the third stage, an ensemble model was developed, whose weights were optimized by means of a genetic algorithm.

User behavior is an important factor that alters energy demand, and which has a significant impact on energy load. Various studies have used data science techniques to show that energy needs vary, depending on the activities of the building occupants [2, 3]. Consequently, a more cost-effective management of the energy load based on user behavior could result in a high percentage of savings [30].

3.2. Building operation

In the world today, building management systems generate a considerable quantity of data. These data contain information concerning temperature, humidity, flow rate, pressure, power, control signals, equipment status, etc. As such, they can be analyzed and exploited to extract operational rules to support building operation. In most cases, these rules are easily interpretable IF-THEN rules, which can help to generate recommendations for control strategies such as feedback mechanisms, and modifications of control processes [28]. For example, May-Ostendorp et al. [31] applied different classification techniques to extract rules from offline model predictive control results. The authors used those rules to achieve near-optimal supervisory control strategies for a mixed-mode building during the cooling season. Although a variety of techniques can be applied to extract IF-THEN rules, most authors use classification techniques [32, 33], especially decision tree algorithms [31, 28]. In addition, classification models are effective tools that can be used to predict building user comfort under different environmental conditions [28].

Although association rules (ARs) are less widespread in the field of Energy Efficiency and Management, they can also help to identify meaningful rules for building operation. This technique can process building-

related data and extract hidden correlations that are not so evident for experienced energy management. Yu et al. [34] used ARs to determine associations and correlations in building operation data. Other authors combine ARs with other techniques to derive interesting rules. For example, in Xiao et al. [35] ARs were used to specify the relations among the power consumptions of major components in each cluster.

Other studies focused on the improvement of specific parts or variables of the building. For instance, Kusiak et al. [33] investigated the relationships between the control settings of the air handling unit and energy usage of the HVAC system. In this case, an algorithm based on the combination of NNs was used to model the non-linear relationship between energy consumption, control settings (supply air temperature and supply air static pressure), and a set of uncontrollable parameters.

Clustering techniques have also been applied to analyze the data sets generated by building automation systems. Xaio and Fan [35] used cluster analysis to identify daily power consumption patterns, whereas Morbitzer et al. [36] applied clustering to analyze simulation results for performance predictions in order to extract predicted operation rules.

3.3. Analysis of infrastructures and retrofitting

In the last few years, new regulations have fomented the study and analysis of aspects related to energy efficiency and sustainable development. Both are now a clear priority for building designers and owners, regardless of whether the building is newly constructed or in renovation. This signifies examining the relations between energy loads, real consumption, and different building components (e.g. walls, windows, doors, lighting, heating, cooling, and ventilation). When equipment status must be verified and potential problems detected, data science techniques are a powerful tool for the extraction of meaningful patterns and correlations between those elements.

Clustering techniques are extremely effective in the analysis of correlations between building infrastructure and performance. For this purpose, Morbitzer et al. [36] applied clustering algorithms to process building monitoring data and discover non-obvious factors of energy loss in building infrastructures. However, clustering is not the only technique that has been employed. Ahmed et al. [37] used classification models to estimate indicators of building behavior, such as comfort and room usage. They concluded that their approach produced better results than traditional analytical tools. The same authors also applied classification and regression techniques couple with building indoor daylight methods to assist decision-making and optimize building design [38].

Sequence analysis techniques, such as motif mining, were used by Patnaik et al. [39] to enhance the performance of cooling infrastructure. In this same line, Shao et al. [40] extracted temporal episodic relationships to better compare systematic consumption trends in residential and commercial buildings with different electrical infrastructure.

Furthermore, Data Science can also assist building designers in the decision-making process by identifying interrelations or patterns to support the design of low-emission buildings. A case in point is Kim et al. [41], whose research focused on finding basic building elements (windows, walls, floors, etc.) that could significantly improve the energy efficiency. For this purpose, they used feature selection extraction combined with *C4.5* decision tree classification.

3.4. Fault detection and prevention

Certainly related with the previous issue, Data Science can also help in verifying the operational status and detecting faults of the building infrastructures. By continuously monitoring the building, it is possible to detect when a fault has happened (typically an anomalous event) and how it affects to other equipment (by means of correlation analysis). From the managers perspective, it is even more interesting to anticipate such faults by characterizing the situations that usually lead to them. In these regards, pattern recognition and regression techniques are very useful. They can help to implement countermeasures to increase building resilience, and to prevent costly incidents.

Capozzoli et al. [42] described a simplified approach to automatically detect faults in building energy equipment. Their methodology is based on the analysis of recorded data of active electrical power for lighting and total active electrical power by using neural networks and outlier detectors. Sedano et al. [43]

presented a similar proposal to detect thermal insulation failures in buildings also based on neural networks. The identification of the consequences of faults in critical infrastructures has been addressed by using graph analysis techniques [44]. A further step in this direction would be using Bayesian networks to model the relations between events and their likelihood. Data Science, through the identification of correlated patterns, can provide the necessary knowledge to populate such models with accurate data.

Faults due to external phenomena cannot be always anticipated by relying on past data. Weather events, such as electrical storms, can be forecast in the short term, and therefore mitigation policies can be developed. However, there are several unpredictable events, such as sabotages and random failures, that can affect the building. In these cases, performing a forensic audit can reveal as well relevant patterns leading to or increasing the impact of faults. Meléndez et al. [45] used pattern recognition techniques to identify sequences of electrical events in substations associated to major failures. This analysis process can be enhanced if sensor data is enriched with contextual information, like building semantic data [46] and cross-dependence databases [47]. In this case, data fusion techniques provide support for data alignment, event detection, and situation assessment. These approaches can be adapted to support post-crisis analysis, such as the one done by Sharma et al. [48], which studies the consequences of the Nepal earthquake in 2015, and how buildings should be improved to minimize similar catastrophes in the future.

3.5. Economic analysis of electric consumption

Numerous companies (mainly utilities) have resorted to Data Science in an effort to discover and understand how and when their customers use energy. The internal development and use of data science tools to extract such knowledge has even made certain companies more competitive than others. The techniques traditionally used for this task are classification, clustering, and pattern analysis (mostly by means of association rules).

One of the first studies in this line was by Chicco et al. [49], who grouped customers into classes, based on their electricity behavior. A modified version of the follow-the-leader algorithm and self-organizing maps were used to compare the results. The classification thus obtained was a first step towards the specification of tariff diversification options. Subsequently, Figueiro et al. [50] proposed a framework for the development and exploitation of historical data, composed of two modules: (i) a load profile module that creates a set of consumer classes by means of supervised and unsupervised clustering; (ii) a classification module that builds a model to assign consumers to these classes. Verdu et al. [51] reviewed the capacity of methods used for clustering purposes, especially self-organizing maps, to classify and filter patterns from distributor, commercial, and/or customer electrical demand databases.

In many cases, it is necessary to efficiently analyze the flow of continuous data. The ISPC algorithm (Incremental Summarization and Pattern Characterization) was used by De Silva et al. [52] to structure stream data into a data warehouse based on key dimensions for enabling a rapid interim summarization. This study enabled the creation of continuous summaries, periodically consolidating identified patterns, which thus facilitated the analysis of data and their prediction.

Models have also been created to provide insights into energy waste due to lighting. In their study of the lighting in three educational institutions, Motta-Cabrera and Zareipour [53] used association rules to discover existing relationships between time, occupancy, and lighting-related energy waste in classrooms. Santamouris et al. [54] also proposed a method for rating and classifying the energy consumption and efficiency of school buildings as compared to other similar buildings. They found that fuzzy clustering techniques could be used to obtain a more robust set of classes, thus avoiding problems stemming from unbalanced classification.

3.6. Energy fraud detection

Sometimes, energy consumption and services are not appropriately billed because of failures in the measurement equipment. Such failures can either be accidental or the product of fraudulent manipulation. These deviations are commonly referred to as non-technical losses (NTLs) [55], and different techniques have been successfully applied to detect them. For instance, León et al. [56] proposed a comprehensive framework to detect NTLs and recover electrical energy (lost by abnormalities or fraud). Their predictive analysis tool,

supplemented by a binary quest tree classification method, was used to discovered association rules in the data. These same authors [57] proposed an expert system for NTL detection. They used regression to study the consumption trends of customers, text mining techniques to analyze inspector commentaries, and association rules to extract additional customer information from the electric company. Nagi et al. [58], on the other hand, used Support Vector Machines to detect abnormal behavior correlated with NTL activities by analyzing customer load profile information as well as certain other attributes.

In regards to the fraudulent use of energy services by consumers and its detection, Kou et al. [59] surveyed the various techniques used for this purpose. Not surprisingly, anomaly detection is the most popular and has been frequently used to discover undesirable user behaviors such as fraud, intrusion, or account defaulting.

Cabral et al. [60] proposed a methodology for detecting fraudulent consumers, which involves the following stages: (i) generation of a database of normal and fraudulent consumers; (ii) application of diverse techniques to obtain decision rules; (iii) identification of fraudulent consumers that match the rules. In the work by Sforna [61], neural networks and fuzzy logic were combined to discover the anomalous behavior of customers. Neural networks were also employed for abnormalities and fraud detection in energy consumption in Galván et al. [62]. Filho et al. [63] described a method to fight against fraud in electricity companies, which involves a classifying algorithm, based on decision trees, to pre-select potentially fraudulent customers, who will then undergo in-site inspection for fraud or faulty measurement equipment identification. The resulting preselection increased the rate of fraud identification from 5% to 40%. Classification techniques were also used by Jiang et al. [64], who created a new automatic feature analysis method using wavelet techniques and combining multiple classifiers to identify fraud in electricity distribution networks.

4. Discussion

The research discussed in the previous sections is summarized in Table 1. The studies are classified on the basis of their field of application in Building Energy Management as well as the data science techniques employed. Table 2 shows the acronym for each technique (where appropriate).

As can be observed in Table 1, the techniques that enable an easy visualization of results, such as classification methods (e.g. decision trees) or association rules, have been widely used because they are intuitive and easy to understand. Generally speaking, the combination of various techniques seems to be the most effective way of attaining the objectives. For example, clustering techniques are often used in time series analysis. The techniques are applied to fragmented data, which have been previously extracted by means of sequence discovery. Another combination of techniques is when descriptive tools are first employed to characterize data, after which predictive techniques are applied for the purpose of detecting fraud, consumption patterns, etc. In this line, instead of using tools for outlier detection, certain fraud detection studies are based on classification into two classes (normal and fraud). Indeed, many methods for anomaly or outlier detection have been developed based on existing classification, regression, and association mining methods that compare the normal or usual trends to those that do not fit in any of the classes or clusters, or which deviate from the association rules.

Regarding the least mentioned techniques in Table 1 (e.g. time-series analysis and sequence discovery) the data collected in the majority of the studies include time-series data. This kind of data is usually analyzed with classification and clustering techniques to estimate, for instance, the future values of the series based on previous values. That is the case when predicting energy demand. Techniques such as association rules in all its variants are certainly underrepresented when modelling and predicting energy loads. In our opinion, such techniques could be used to discover relations between internal parameters or the interdependences of pieces of equipment. Moreover, fuzzy rules (which have been widely used for HVAC control) can also be used for descriptive reports of energy loads since they offer a robust representation in the context of high imprecision and uncertainty.

In other cases, the data are discretized or labeled (e.g. data are split into day/night, seasons, and more complex temporal divisions) and then tasks such as association mining are performed. One of the main challenges in this area is to appropriately handle temporal data, usually collected from sensor data or user energy consumption, even when data storage is not a viable option. Many of the techniques used need to

Fields	Techniques						
	Classification	Regression	Clustering	Association Rules	Sequence discovery	Anomaly/ outlier detection	Time Series Analysis
Load and Energy Consumption Prediction	[6] FKM [24] MLP, C1 [26] DT (C4.5) [27] DT [28] DT (C4.5) [33] MLP, C1 [65] PNN	[27] SVM [66] ARMA [67] ARMA, ARIMA	[25] HT, FLC, KM, FKM, FR [65] FKM			[29] GESD	[29] C2 [66] ARMA [67] ARMA, ARIMA [68] ARMA, MLT, DT, FL
Building Operation	[28] DT (C4.5) [31] CART [32] NB [33] MLP, C1	[31] GLM	[29] EWKM [35] AH, PAM, EWKM, KM [36]	[34] FPG [35] AP	[29] ACF, PACF	[35] AR	
Analysis/Design of Infrastructure	[37] NB, DT, SVM [38] DT [41] DT (C4.5)	[38] SVM	[36] [39] KM		[39] MM [40] MM		
Failure analysis	[42] CART [55] BN, DT	[55] PCC				[42] [43] NN	[43] NN [45]
Economic Analysis	[33] MLP, C1 [50] DT (C5.0)		[49] FLC, SOM [50] SOM, KM [51] SOM [54] FC	[53] AP			[50]
Fraud Detection	[55] BN, DT [56] QDT [58] SVM [61] NN [63] DT [64] MLP, BN	[55] PCC [57]	[51] SOM [60] SOM	[56] GRI [57] TM, AR		[56] STD [58] SVM [61] NN, FL [62] PRBFN	[55] [56] [64] WL

Table 1: Summary of data science techniques used in the field of Building Energy Management, classified by type and field of application.

Acronym	Tool
ACF	Autocorrelation function
AH	Agglomerative hierarchical clustering
AP	Apriori
AR	Association rules
ARIMA	Autoregressive Integrated Moving Average
ARMA	Autoregressive Movement Average
BN	Bayesian Networks
C1	Comparison of methods: CART, MLP, SVM, Chi-square Automatic Interaction Detector, Boosting Tree, Random Forest and k-NN
C2	Comparison of prediction methods: Multiple Linear Regression, ARIMA, Multivariate Adaptive Regression Splines, MLP, Support Vector Regression, Boosting Tree, Random Forest and k-NN
CART	Classification and Regression trees
DT	Decision Trees
EWKM	Entropy weighted k-means
FC	Fuzzy clustering
FKM	Fuzzy k-means
FL	Fuzzy logic
FLC	Follow-the-leader clustering
FPG	FP-Growth
FR	Fuzzy relation clustering
GESD	Generalized extreme studentized deviate
GLM	Generalized Linear Models
GRI	Generalized Rule Induction
HT	Hierarchical Tree clustering
KM	k-Means
k-NN	k-Nearest Neighbours
MLP	Multi-layer perceptron (Neural Networks)
MLT	McLeod.Li test
MM	Motif mining (Sequence discovery)
NB	Naive Bayes classification
NN	Neural Networks
PACF	Partial autocorrelation function
PAM	Partition around medoids clustering
PCC	Pearson Correlation Coefficient
PNN	Probability Neural Networks
PRBFN	Probabilistic Radial Basis Function Network (Neural Networks)
QDT	Quest Decision Tree
SOM	Self Organizing Maps
STD	Standard Deviation Estimation
SVM	Support Vector Machine
TM	Text Mining
WL	Wavelets

Table 2: Acronyms of the data science methods used in the research cited in Table 1.

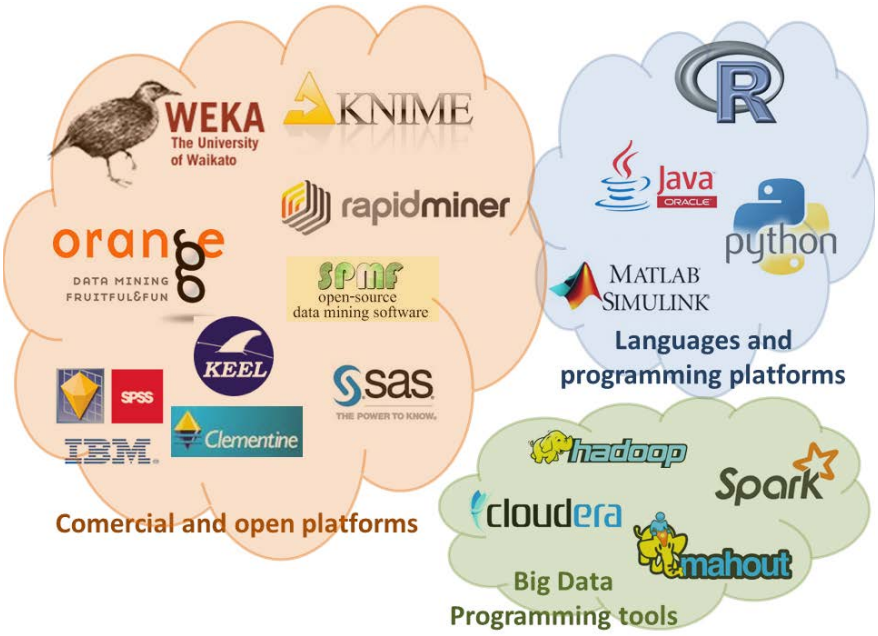

Figure 2: Selection of platforms and programming tools that implement/offer data science techniques.

be suitably adapted to obtain knowledge from this kind of data in the form of streams. In that regard, a promising research line involves the generalizations of revised techniques for the management of Big Data (see Section 5.2).

The literature also mentions various data science methods with better features than those in Table 1. Examples include methods with more accurate results, methods capable of handling temporal data or data streams, etc., which could feasibly be applied to Building Energy Management. However, most of them are not available for direct use and require extensive training to set the parameters for an optimal result. Our revision found that the most frequently used data science algorithms in the energy field are those found in commercial toolboxes and open platforms (see Figure 2).

Commercial toolboxes such as *Intelligent Miner*, *SPSS*, *Clementine*, *Oracle Data mining*, *RapidMiner*, and *SAS* incorporate a diverse palette of tools with good documentation that allow non-programming users to quickly learn how to apply them. However, the disadvantage of these toolboxes is that they obviously cannot include cutting-edge techniques. In contrast, open platforms such as *WEKA*, which is more oriented towards data mining and machine learning research, or *Matlab* and *R*, which are more programming-oriented, are not as popular in this field because they can be quite challenging for non-technical users, and they require specific background knowledge of the methods. Nevertheless, there are less well-known platforms that have been used in other fields, such as financial analysis. Besides being a good source of data analysis and visualization methods, these platforms are generally updated with the latest improvements in the scientific community. Examples include KNIME (Konstanz Information Miner), KEEL (Knowledge Extraction based on Evolutionary Learning), SPMF (open-source data mining software) and Orange, an open-source data mining platform. For more programming-oriented users, there is also Python scripting and Java Data Mining modules (JDM). It is also necessary to consider new tendencies and methods for managing big data, which may be a good option for managing temporal and stream data. Promising tools are those provided by Apache Spark, Hadoop, and Mahout.

An interesting issue, which is rarely addressed when selecting an appropriate method, is the interpretability of the results. Since the knowledge extracted will eventually be used by humans, the results should be presented in an understandable way so that the maximum benefit can be derived from the analysis. In this respect, new techniques with suitable knowledge and data representations are in the process of be-

ing developed. Their representations approximate human thinking/language in order to incorporate expert knowledge in the analysis process. Also interesting are techniques that can handle imprecise data, which is inherent from data compiled from sensors, and/or uncertainty of data. This and other challenges in Energy Efficiency and Management are discussed in the following section.

5. Current trends and challenges

This work is not complete without a reflection on the issues that now affect and will affect the evolution of Energy Efficiency and Management and its relation with ICT. These trends will strengthen the progressively expanding role of Data Science in this field, and will lead to pioneering applications. At present, the tendency is towards more specialized solutions. Suitable synergies between energy companies and technological enterprises can lead to innovative business ideas linked to cheap monitoring devices, cloud computing, open data, etc.

In today's world, one of the best strategies for a company is to specialize in a frequently occurring problem in the market and to solve it more effectively than the competition. In this context Big Data tools applied to energy problems are very much in the spotlight since they promise highly specialized solutions for the extremely competitive utility market. In this sense, they are capable of outperforming other more traditional tools. Even though they are only applicable to small areas, the results can lead to huge financial savings and benefits.

Apart from Big Data, other technologies that are expected to have a significant impact on Energy Efficiency and Management include Smart metering, the Internet of Things and Cloud computing. Nevertheless, these technologies bring to the forefront crucial issues that must be addressed, namely, privacy along with uncertainty and imprecision.

5.1. Smart metering

Smart metering is the continuous monitoring of energy consumption with a view to gaining a better understanding of the energy consumption, generation, and transportation stages. In the future, it will significantly improve the decision-making phase, and also positively influence the energy behavior of final users and managers (energy awareness). Smart metering signifies a vast change in electricity or gas metering since it measures the precise amount of consumed energy and records the exact time when consumption occurs. As a result, bills are no longer an aggregated value of energy consumption (or of estimated consumption) over a long period of time. Rather, they are real measurements in very short time intervals. Since smart metering permits the real-time visualization of consumed energy, users are able to better understand their own energy behavior. This has the advantage of reducing energy waste, and encourages users to modify underperforming habits.

Although smart metering was first proposed in the 1970s, such meters have only been widely deployed in the last ten years. Commercial interests and technical advances have now reached the point at which present and future investments in smart metering in the European Union have been estimated at €51 billion, with potential financial benefits ranging from €14 billion to €67 billion [69].

Nevertheless, smart meters bring some privacy concerns. For instance, from smart metering information it might be possible to get private details of a home, such as the number of occupants, their daily routines, their electronic devices, etc. Therefore, despite the fact that smart metering has enormous social and technical benefits, it also brings with it security and privacy concerns since users are obliged to share usage data [70]. Additional care should be taken so as not to identify individual users during the analysis process. The most frequent solution for this problem is data anonymization [71]. However, such issues must be specifically addressed in order to guarantee user privacy and to comply with government regulations (see Section 5.5.

The norm is that users must agree on what they share with utility companies (e.g. frequency of data collection, type of data sent, data granularity, etc.). Nonetheless, gathering, storing, and analyzing these data are not the whole story. The big challenge in the Energy Management, and especially for Energy Big Data, is to understand which part of the data is useful for answering the questions that facilitate decision-making.

5.2. Big Data

The huge amount of data generated by sensors can be exploited to increase energy efficiency. However, the sheer quantity of the data poses a challenge at various levels for traditional data analysis approaches. New infrastructure and tools need to be developed to deal with what is now known as *Big Data*. The broader definition of Big Data is data that is too large and complex to be handled by traditional databases. According to Laney, Big Data is best described by three Vs [72]: volume, variety, and velocity. *Volume* refers to the huge amount of data already produced by organizations; *variety* refers to its diversity as the data come from different sources (structured, unstructured, multimedia, textual, etc.); and *velocity* refers to the speed at which these data are obtained and accessed.

Big Data has gained the interest of governments and companies. For instance, the European Commission and Europe's data industry (including companies such as ATOS, Orange, SAP, and Siemens) have agreed to invest €2.5 billion in a public-private partnership (PPP) that aims to strengthen the data sector and put Europe at the forefront of the global data race [73]. There are many industries, such as Health, Retail and Education, in which Big Data solutions are flourishing and are now taking the lead [74]. Those solutions are providing companies with a competitive advantage, and the energy industry is no exception. Companies in the energy sector are not indifferent to this trend and are taking a keen interest in Big Data collection and analysis tools. Accordingly, they are improving their networks with the use of sensing devices that monitor energy status. For instance, according to Lesser [75], when smart meters are fully deployed, this will potentially generate up to 1000 petabytes of data annually. This huge flow of information requires summarization processes and smarter analyses, and Big Data has a lot to say in that regard. In fact, IBM [76] specifically states the challenges and opportunities of Big Data as applied to energy management, and lists the actors involved.

It is thus not surprising that large technology corporations are taking an interest in the field of Energy Management. For example, Siemens and Teradata have begun to work together on data science solutions to control and monitor their energy infrastructure [77]. Furthermore, many startups have a brighter future since Google's acquisition of Nest Labs, a small company which was adding sensors, computing, and communications technology to make everyday objects more useful. Nevertheless, this acquisition has raised concerns about privacy and the use that corporations, such as Google, might make of the data reflecting user energy behavior. This issue is a sensitive one that should be carefully addressed by governments and policymakers. There is currently a great deal of movement in this area. In fact, apart from the number of government-funded projects, there are also hundreds of companies that have entered the arena and are currently proposing innovative solutions and business opportunities.

5.3. Internet of Things

The Internet of Things (IoT) is the term for a world in which objects, and not only people, will be permanently connected and able to interact through the Internet. The IoT is expected to foster a huge number of new applications such as environmental monitoring, healthcare, and efficient energy management in smart homes. According to Rifkin [78], the IoT can potentially disrupt the economy, as we know it, by reducing the dependency on central entities and promoting a more collaborative economy with reduced costs and automated processes.

In regards to energy efficiency and management, it is expected that smart meters and appliances will soon be connected to the Internet, and also equipped with some kind of intelligence. The result will be a fully connected and sensorized environment. For instance, by moving to phones and other wearables, these devices will generate a huge quantity of signals and data. This will lead to the refinement and enhanced accuracy of contexts and models, thus enabling new and fascinating applications. Moreover, it will produce extremely accurate monitoring of energy flows, based on real-time data and situational awareness, as opposed to historical data patterns. In this way, energy distribution can be easily predicted and corrected in almost real-time, which will avoid errors and prevent overloads.

Connected devices can also perform power management tasks with greater precision and faster response times than manual or human-dependent systems, thus saving energy, prioritizing usage, and setting policies for response to outages. The main drawback of the Internet of Things is that it raises new concerns about data privacy, data sovereignty, and security (see Section 5.5).

5.4. Cloud computing

Cloud computing [79] has its origins in the concept of grid computing, whose goal is to reduce computation costs and increase the flexibility and reliability of systems by aggregating computing resources. According to the NIST [80], cloud computing is "a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g. networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction".

Thus, Cloud Computing can be regarded as a distributed system that provides computing services via a computer communication network such as the Internet. Within this context, a service is regarded as a task that has been encapsulated in such a way that it can be automated and supplied to the clients through a consistent and constant procedure. In addition, resources in the cloud are transparent to the final users, who do not need to know their exact location or be aware of the underlying architecture [79]. This trend to relocate the computing capabilities to the premises of a third party has grown in recent years. Specialized companies can thus provide dedicated premises to host the computing necessities of other companies, delegating to them the responsibility for the integrity of loaded data (without data losses), secure access, and the correct performance of services. This allows companies to improve their performance and increase their flexibility while reducing costs.

For most companies, cloud computing seems a plausible choice since they can avoid scalability problems, and reduce deployment costs and time. Without making an up-front investment in computing capabilities, companies are able to scale their needs in a quick and easy way. Additionally, Cloud computing enables continuous and transparent updates and improvements, which are readily available to customers. However, these advantages come at a cost. Because of security constraints and privacy concerns, some industries are still reluctant to embrace cloud computing and cloud technologies in general.

Companies such as Amazon, Google, and Microsoft offer commercial products that provide an infrastructure for cloud computing. At a relatively low cost, developers can transparently allocate computational resources to run their applications, and subsequently increase or reduce these resources to adapt to processing requirements. In addition to their cloud computing platforms, these companies have developed flagship applications such as Google Gmail, Microsoft Office 365, Adobe Creative Cloud, etc. Other well-known companies, such as Dropbox and Netflix, extensively use cloud platforms adapted to a variable number of users in their normal operation. In all likelihood, future energy services that must process data generated by a large number of users will also rely on cloud computing.

Once again, when using Cloud computing, the problem of data privacy and security looms on the horizon [81]. Since company data and applications are stored in a third-party infrastructure, which potentially shares hardware with other businesses, this could put user privacy at risk. Moreover, handing over confidential information to another company makes most people feel insecure. In addition, governments have been accused of participating, either legally or illegally, in sniffing data stored in shared data centers.

It is also true that because of the inherent distribution of resources in cloud computing, latency and connectivity issues can also be a problem for certain users. Therefore, fault tolerance and recovery mechanisms should be implemented by the service provider to ensure access to resources without losses. Finally, there are major challenges in the domain of energy efficiency, which involve the reduction of the fingerprint of supporting big cloud data centers. In fact, many companies have moved their infrastructures to northern countries where they can cool the equipment more easily and cheaply. They can also update their infrastructures with low-power microprocessors that reduce energy consumption.

5.5. Privacy and security issues

As explained in the previous sections, *Smart metering*, *Big Data*, *IoT*, and *Cloud computing* naturally lead to security and privacy concerns [82]. This means that it is first necessary to develop mechanisms that ensure that the information remains secure. With new cloud services available, many data analyses of the future will be performed in remote data centers, which are either owned by energy companies or third parties. The case of third-party ownership adds still another actor to the process, which may raise concerns for both the users and utility company. Furthermore, there is the question of which data can be collected

as well as its ownership. This is particularly relevant to personal data. The legal frameworks in different countries, as well as the ethical aspects not addressed in the law, worsen this problem [83]. Therefore, it is necessary to establish a common reference that clearly defines what private data are, who can use them, and for what purpose.

To handle these privacy issues, security mechanisms and privacy schemes must be enforced. A first step has already been taken by the European Commission regarding smart grids and smart metering. Data Protection Directive 95/46/EC proposes a "Recommendation on the Data Protection Impact Assessment Template for smart grid and smart metering systems" [84] which provides guidelines on how to support the security of the implementation of smart grids and smart metering by data controllers. The objective of this recommendation is to ensure progress towards the full harmonized protection of personal data as well as to enhance security in smart grids and metering throughout the European Union.

5.6. Uncertainty and imprecision

As already stated, energy systems are complex dynamic systems, which are not susceptible to precise modelization. To further complicate the situation, data are often gathered from a myriad of heterogeneous sensors, which are not fully reliable. In this situation, techniques that can deal with uncertainty and imprecision in models and data seem like a sensible choice. Soft computing techniques in particular offer an effective solution for studying and modeling the stochastic behavior of renewable energy generation, operation of grid-connected renewable energy systems, and sustainable decision-making, *inter alia*. In fact, their tolerance of imprecision, uncertainty, partial truth, and approximation makes them useful alternatives to more conventional techniques [85].

Moreover, the incorporation of uncertainty in certain energy consumption processes provides a more realistic description of them, thanks to better simulation models of building performance. A good example is the analysis performed by de Wit and Augenbroe [86], in which the uncertainties identified in building design simulations had a direct effect on performance assessment and quantification of design performance. These authors focused on the study of uncertainty propagation in decision-making, where inputs sustaining rational decisions with respect to energy use, thermal comfort, etc. come from domain experts in building physics. More recently, Hopfe and Hensen [87] analyzed energy performance (energy consumption and thermal comfort) and considered three different groups of uncertainties for building design: (i) physical uncertainties inherent in physical properties, which appear in quantified measurements; (ii) design uncertainties, such as changes in the room geometry or the window size; (iii) scenario uncertainties that are linked to building usage. The results of their study confirmed the added value and usefulness of integrating uncertainty analysis in building performance simulations [87].

Furthermore, one of the features of data within the context of Big Data is their heterogeneity and, more often than not, their lack of structure. This lack of precision highlights the advisability of using techniques such as fuzzy logic, which can model the information more satisfactorily than conventional tools that can only process structured data. An interesting proposal in that regard is that of the fuzzy extension of Building Information Models [46]. Another characteristic in the context of Big Data is that data are mostly irrelevant in isolation. What really matters is their overall tendency or texture. Appropriately modelling these flows of imprecise and uncertain information was the objective of research such as Saleh and Masseglia [88] and Delgado et al. [89].

Finally, fuzzy logic can also benefit Energy Big Data, insofar as the representation of conclusions and interpretability of the data. In this sense, linguistic summarization techniques can enhance the legibility of results obtained with data science tools and make them more understandable for human users. For example, Chen [90] revised a set of data science techniques in which the use of fuzzy logic led to more user-friendly results. Interpretability is crucial because in most cases, the report will be sent to a manager or operator, who must be able to understand it and act on it.

6. Conclusions

This paper has reviewed recent developments in information technologies and their influence on Building Energy Management. We examined the usefulness of various data science techniques that have been applied

or could be applied to solve energy problems. Given the current challenges that must be addressed in energy management, it is evident that data science techniques will be widely applied in the near future.

In all areas, the discovery and exploitation of the information hidden within collected data is extremely useful. However, in the case of energy consumption, this is even more so because of the economic and environmental implications. In Building Energy Management, the identification of equipment and user consumption patterns will doubtlessly save money, improve comfort, and reduce contaminant emissions. The economic impact of this sector is reflected in the number of new companies that are currently applying Data Science to the analysis of energy consumption data, user habits, and building infrastructure. This is leading to synergies between energy companies and information technology enterprises who are beginning to work together towards more efficient energy management.

This new context actively challenges researchers to develop solutions for the management of huge amounts of heterogeneous data in real time, as well as to find ways to deal with its associated uncertainty. Data science techniques have shown themselves to be valuable tools capable of extracting and exploiting the knowledge and information inherent in user data. In the near future, Big Data techniques will expand these possibilities and democratize them. This will enhance energy awareness, since users will have access to more data and be able to understand their own energy consumption habits. In this regard, companies have begun to realize that energy savings are not only a question of optimizing components, but also of understanding and acting on user behaviors.

Acknowledgements

This research was partially funded by the Spanish Government (TIN2012-30939 and TIN2015-64776-C3-1-R projects), the Andalusian Regional Government (P11-TIC7460 project), the European Commission (FP7 *Energy IN TIME* project, grant agreement no. 608981). Juan Gómez is supported by University of Granada under the Programme 'Proyectos para la incorporación de jóvenes doctores a nuevas líneas de investigación'.

[1] WBCSD. Energy efficiency in buildings: Business realities and opportunities. Technical report, The World Business Council for Sustainable Development, 2007.

[2] H. Allcott and S. Mullainathan. Behavior and energy policy. *Science*, 327(5970):1204–1205, 2010.

[3] T.A. Nguyen and M. Aiello. Energy intelligent buildings based on user activity: A survey. *Energy and Buildings*, 56:244–257, 2013.

[4] Z. Li, Y. Han, and P. Xu. Methods for benchmarking building energy consumption against its past or intended performance: An overview. *Applied Energy*, 124:325–334, 2014.

[5] T. Nikolaou, D. Kolokotsa, and G. Stavrakakis. Review on methodologies for energy benchmarking, rating and classification of buildings. *Advances in Building Energy Research*, 5(1):53–70, 2011.

[6] K. Zhou, S. Yang, and C. Shen. A review of electric load classification in smart grid environment. *Renewable and Sustainable Energy Reviews*, 24:103–110, 2013.

[7] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G.J. McLachlan, A. Ng, B. Liu, P.S. Yu, Z.-H. Zhou, M. Steinbach, D.J. Hand, and D. Steinberg. Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14(1):1–37, 2008.

[8] K.P. Bennett and C. Campbell. Support vector machines: Hype or hallelujah? *ACM SIGKDD Explorations Newsletter*, 2(2):1–13, 2000.

[9] Y.N. Andrew and I.J. Michael. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In T.G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 841–848. MIT Press, 2002.

[10] W. Banzhaf, F.D. Francone, R.E. Keller, and P. Nordin. *Genetic Programming: An Introduction: on the Automatic Evolution of Computer Programs and its Applications*. Morgan Kaufmann Publishers Inc., 1998.

[11] D. Kriesel. *A Brief introduction to neural networks*. Available at http://www.dkriesel.com, 2007.

[12] V. Nitin Bhatia. Survey of nearest neighbor techniques. *International Journal of Computer Science and Information Security*, 8(2):302–305, 2010.

[13] S. Chatterjee and A.S. Hadi. *Regression Analysis by Example*. John Wiley & Sons, 2013.

[14] A.K. Jain, M.N. Murty, and P.J. Flynn. Data clustering: A review. *ACM Computing Surveys*, 31(3):264–323, 1999.

[15] R. Agrawal, T. Imielinski, and A. Swami. Mining associations between sets of items in massive databases. In *ACM-SIGMOD International Conference on Data*, pages 207–216, 1993.

[16] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining, (First Edition)*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2005.

[17] M. J. Zaki. Scalable algorithms for association mining. *IEEE Transactions on Knowledge and Data Engineering*, 12(3):372–390, 2000.

[18] N.R. Mabroukeh and C.I. Ezeife. A taxonomy of sequential pattern mining algorithms. *ACM Computing Surveys*, 43(1):1–41, 2010.
[19] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Computing Surveys*, 41(3):15:1–15:58, 2009.
[20] J.D. Hamilton. *Time Series Analysis*. Princeton University Press, 1994.
[21] E. Hüllermeier. Fuzzy methods in machine learning and data mining: Status and prospects. *Fuzzy Sets and Systems*, 156(3):387–406, 2005. 40th Anniversary of Fuzzy Sets.
[22] J.C. Bezdek, R. Ehrlich, and W. Full. FCM: The fuzzy c-means clustering algorithm. *Computers & Geosciences*, 10(2-3):191–203, 1984.
[23] M. Delgado, M.D. Ruiz, D. Sánchez, and J.M. Serrano. A formal model for mining fuzzy rules using the {RL} representation theory. *Information Sciences*, 181(23):5194–5213, 2011.
[24] A. Kusiak, M. Li, and Z. Zhang. A data-driven approach for steam load prediction in buildings. *Applied Energy*, 87(3):925–933, 2010.
[25] I. Prahastono, D. King, and C.S. Ozveren. A review of electricity load profile classification methods. In *Universities Power Engineering Conference, 2007. UPEC 2007. 42nd International*, pages 1187–1191, Sept 2007.
[26] Z. Yu, F. Haghighat, B.C.M. Fung, and H. Yoshino. A decision tree method for building energy demand modeling. *Energy and Buildings*, 42(10):1637–1646, 2010.
[27] Z.Y. Li. An empirical study of knowledge discovery on daily electrical peak load using decision tree. *Advanced Materials Research*, 433–440:4898–4902, 2012.
[28] G. Yang, E. Tumwesigye, B. Cahill, and K. Menzel. Using data mining in optimisation of building energy consumption and thermal comfort management. In *Software Engineering and Data Mining (SEDM), 2010 2nd International Conference on*, pages 434–439, June 2010.
[29] C. Fan, F. Xiao, and S. Wang. Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques. *Applied Energy*, 127:1–10, 2014.
[30] M.V. Moreno, M.A. Zamora, and A.F. Skarmeta. User-centric smart buildings for energy sustainable smart cities. *Transactions on Emerging Telecommunications Technologies*, 25(1):41–55, 2014.
[31] P.T. May-Ostendorp, G.P. Henze, B. Rajagopalan, and C.D. Corbin. Extraction of supervisory building control rules from model predictive control of windows in a mixed mode building. *Journal of Building Performance Simulation*, 6(3):199–219, 2013.
[32] A. Ahmed, J. Ploennigs, Y. Gao, and K. Menzel. Analyse building performance data for energy-efficient building operation. In *26th International Conference on Managing IT in Construction*, Istanbul, Turkey, 2009.
[33] A. Kusiak, M. Li, and F. Tang. Modeling and optimization of HVAC energy consumption. *Applied Energy*, 87(10):3092–3102, 2010.
[34] Z. Yu, F. Haghighat, B.C.M. Fung, and L. Zhou. A novel methodology for knowledge discovery through mining associations between building operational data. *Energy and Buildings*, 47:430–440, 2012.
[35] F. Xiao and C. Fan. Data mining in building automation system for improving building operational performance. *Energy and Buildings*, 75:109–118, 2014.
[36] C. Morbitzer, P. Strachan, and C. Simpson. Data mining analysis of building simulation performance data. *Building Services Engineering Research & Technology*, 25(3):253–267, 2004.
[37] A. Ahmed, N.E. Korres, J. Ploennigs, H. Elhadi, and K. Menzel. Mining building performance data for energy-efficient operation. *Advanced Engineering Informatics*, 25(2):341–354, 2011.
[38] A. Ahmed, M. Otreba, N.E. Korres, H. Elhadi, and K. Menzel. Assessing the performance of naturally day-lit buildings using data mining. *Advanced Engineering Informatics*, 25:364–379, 2011.
[39] D. Patnaik, M. Marwah, R. Sharma, and N. Ramakrishnan. Sustainable operation and management of data center chillers using temporal data mining. In *Procs. 15th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, pages 1305–1314, 2009.
[40] H. Shao, M. Marwah, and N. Ramakrishnan. A temporal motif mining approach to unsupervised energy disaggregation: Applications to residential and commercial buildings. In *1st International Non-Intrusive Load Monitoring Workshop*, 2012.
[41] H. Kim, A. Stumpf, and W. Kim. Analysis of an energy efficient building design through data mining approach. *Automation in Construction*, 20(1):37–43, 2011.
[42] A. Capozzoli, F. Lauro, and I. Khan. Fault detection analysis using data mining techniques for a cluster of smart office buildings. *Expert Systems with Applications*, 42(9):4324–4338, 2015.
[43] Javier Sedano, Leticia Curiel, Emilio Corchado, Enrique de la Cal, and José R. Villar. A soft computing method for detecting lifetime building thermal insulation failures. *Integrated Computer-Aided Engineering*, 17(2):103–115, 2010.
[44] George Stergiopoulos, Panayiotis Kotzanikolaou, Marianthi Theocharidou, and Dimitris Gritzalis. Risk mitigation strategies for critical infrastructures based on graph centrality analysis. *International Journal of Critical Infrastructure Protection*, 10:34–44, 2015.
[45] J. Meléndez, O. Quiroga, and S. Herraiz. Analysis of sequences of events for the characterization of faults in poser systems. *Electric Power Systems Research*, 87:22–30, 2012.
[46] Juan Gmez-Romero, Fernando Bobillo, Maria Ros, Miguel Molina-Solana, M. Dolores Ruiz, and M.J. Martn-Bautista. A fuzzy extension of the semantic building information model. *Automation in Construction*, 57:202–212, 2015.
[47] R. Zimmerman. Decision-making and the vulnerability of interdependent critical infrastructure. In *Systems, Man and Cybernetics, 2004 IEEE International Conference on*, volume 5, pages 4059–4063, 2004.
[48] K. Sharma, L. Deng, and C.C. Noguez. Field investigation on the performance of building structures during the april 25,

2015, Gorkha earthquake in Nepal. *Engineering Structures*, 121:61–74, 2016.
[49] G. Chicco, R. Napoli, F. Piglione, P. Postolache, M. Scutariu, and C. Toader. Load pattern-based classification of electricity customers. *Power Systems, IEEE Transactions on*, 19(2):1232–1239, May 2004.
[50] V. Figueiredo, F. Rodrigues, Z. Vale, and J.B. Gouveia. An electric energy consumer characterization framework based on data mining techniques. *IEEE Trans. on Power Systems*, 20(2):596–602, 2005.
[51] S.V. Verdú, M.O. García, C. Senabre, A.G. Marín, and F.J.G. Franco. Classification, filtering, and identification of electrical customer load patterns through the use of self-organizing maps. *IEEE Transactions on Power Systems*, 21(4):1672–1682, 2006.
[52] D. De Silva, X. Yu, D. Alahakoon, and G. Holmes. A data mining framework for electricity consumption analysis from meter data. *Industrial Informatics, IEEE Transactions on*, 7(3):399–407, 2011.
[53] D.F. Motta-Cabrera and H. Zareipour. Data association mining for identifying lighting energy waste patterns in educational institutes. *Energy and Buildings*, 62:210–216, 2013.
[54] M. Santamouris, G. Mihalakakou, P. Patargias, N. Gaitani, K. Sfakianaki, M. Papaglastra, C. Pavlou, P. Doukas, E. Primikiri, V. Geros, M.N. Assimakopoulos, R. Mitoula, and S. Zerefos. Using intelligent clustering techniques to classify the energy performance of school buildings. *Energy and Buildings*, 39(1):45–51, 2007.
[55] I. Monedero, F. Biscarri, C. León, J.I. Guerrero, J. Biscarri, and R. Millán. Detection of frauds and other non-technical losses in a power utility using Pearson coefficient, Bayesian networks and decision trees. *International Journal of Electrical Power & Energy Systems*, 34(1):90–98, 2012.
[56] C. León, F. Biscarri, I. Monedero, J.I. Guerrero, J. Biscarri, and R. Millán. Variability and trend-based generalized rule induction model to NTL detection in power companies. *Power Systems, IEEE Transactions on*, 26(4):1798–1807, 2011.
[57] C. León, F. Biscarri, I. Monedero, J.I. Guerrero, J. Biscarri, and R. Millán. Integrated expert system applied to the analysis of non-technical losses in power utilities. *Expert Systems with Applications*, 38(8):10274–10285, 2011.
[58] J. Nagi, K.S. Yap, S.K. Tiong, S.K. Ahmed, and M. Mohamad. Nontechnical loss detection for metered customers in power utility using support vector machines. *Power Delivery, IEEE Transactions on*, 25(2):1162–1171, April 2010.
[59] Y. Kou, C.-T. Lu, S. Sirwongwattana, and Y.-P. Huang. Survey of fraud detection techniques. In *2004 IEEE International Conference on Networking, Sensing and Control*, volume 2, pages 749–754, 2004.
[60] J.E. Cabral, J.O.P. Pinto, E.M. Martins, and A.M.A.C. Pinto. Fraud detection in high voltage electricity consumers using data mining. In *Transmission and Distribution Conference and Exposition, 2008. TD. IEEE/PES*, pages 1–5, April 2008.
[61] M. Sforna. Data mining in a power company customer database. *Electric Power Systems Research*, 55(3):201–209, 2000.
[62] J. Galván, E. Elices, A.M. Noz, T. Czernichow, and M. Sanz-Bobi. System for detection of abnormalities and fraud in customer consumption. In *12th IEEE/PES Conf. Electric Power Supply Industry*, 1998.
[63] J.R. Filho, E.M. Gontijo, A.C. Delaiba, E. Mazina, J.E. Cabral, and J.O.P. Pinto. Fraud identification in electricity company customers using decision tree. In *2004 IEEE International Conference on Systems, Man and Cybernetics*, volume 4, pages 3730–3734, Oct 2004.
[64] R. Jiang, H. Tagaris, A. Lachsz, and M. Jeffrey. Wavelet based feature extraction and multiple classifiers for electricity fraud detection. In *Transmission and Distribution Conference and Exhibition 2002: Asia Pacific. IEEE/PES*, volume 3, pages 2251–2256, Oct 2002.
[65] N. Anuar and Z. Zakaria. Electricity load profile determination by using fuzzy {CMeans} and probability neural network. *Energy Procedia*, 14:1861–1869, 2012. 2011 2nd International Conference on Advances in Energy Engineering (ICAEE).
[66] A. Vaghefi, M.A. Jafari, E. Bisse, Y. Lu, and J. Brouwer. Modeling and forecasting of cooling and electricity load demand. *Applied Energy*, 136:186–196, 2014.
[67] P. Chujai, N. Kerdparasop, and K. Kerdprasop. Time series analysis of household electric consumption with ARIMA and ARMA models. In *Proceedings of the International MultiConference of Engineers and Computer Scientists*, volume I, pages 295–300, 2013.
[68] A. Azadeh, M. Saberi, S.F. Ghaderi, A. Gitiforouz, and V. Ebrahimipour. Improved estimation of electricity demand function by integration of fuzzy system and data mining approach. *Energy Conversion and Management*, 49(8):2165–2177, 2008.
[69] A. Faruqui, D. Harris, and R. Hledik. Unlocking the € 53 billion savings from smart meters in the EU: How increasing the adoption of dynamic tariffs could make or break the EU's smart grid investment. *Energy Policy*, 38(10):6222–6231, 2010.
[70] P. McDaniel and S. McLaughlin. Security and privacy challenges in the smart grid. *IEEE Security Privacy*, 7(3):75–77, 2009.
[71] C. Efthymiou and G. Kalogridis. Smart grid privacy via anonymization of smart metering data. In *2010 First IEEE International Conference on Smart Grid Communications (SmartGridComm)*, pages 238–243, 2010.
[72] D. Laney. 3-D data management: Controlling data volume, velocity and variety. Technical report, Gartner, 2001.
[73] European Commission Press Release. European commission and data industry launch 2.5 billion partnership to master big data. [Online; accessed: 2014-12-15] `http://europa.eu/rapid/press-release_IP-14-1129_en.htm`.
[74] B. Schmarzo. *Big Data: Understanding How Data Powers Big Business*. Wiley, 2013.
[75] A. Lesser. How energy data will impact the smart grid, 2013. [Online; accessed: 2015-04-13] `http://research.gigaom.com/report/how-energy-data-will-impact-the-smart-grid/`.
[76] IBM. Managing big data for smart grids and smart meters. White paper, IBM Software, 2012.
[77] Teradata. Siemens and Teradata form global strategic partnership for big data in the utility sector. [Online; accessed: 2015-03-30] `http://www.teradata.com/News-Releases/2013/Siemens-and-Teradata-form-global-strategic-partnership-for-big-data-in-the-utility-sector/`.
[78] J. Rifkin. *The Zero Marginal Cost Society: The Internet of Things, the Collaborative Commons, and the Eclipse of*

Capitalism. St. Martin's Press, 2014.
[79] R. Buyya, J. Broberg, and A. Goscinski, editors. *Cloud Computing: Principles and Paradigms*. Wiley, 2011.
[80] National Institute of Standards and Technology (NIST). The NIST definition of Cloud Computing. `http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf`. Accessed: 2014-12-07.
[81] M.N.O. Sadiku, S.M. Musa, and O.D. Momoh. Cloud computing: Opportunities and challenges. *IEEE Potentials*, 33(1):34–36, 2014.
[82] D. Zissis and D. Lekkas. Addressing cloud computing security issues. *Future Generation Computer Systems*, 28(3):583–592, 2012.
[83] T. Craig and M.E. Ludloff. *Privacy and Big Data*. O'Reilly Media, 2011.
[84] European Commission. Commission Recommendation on Data Protection Impact Assessment Template for Smart Grid and Smart Metering systems (2014/724/EU). `http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv:OJ.L_.2014.300.01.0063.01.ENG`. Accessed: 2015-03-20.
[85] K. Gopalakrishnan, S.K. Khaitan, and S. Kalogirou, editors. *Soft Computing in Green and Renewable Energy Systems*, volume 269 of *Studies in Fuzziness and Soft Computing*. Springer Berlin Heidelberg, 2011.
[86] S. de Wit and G. Augenbroe. Analysis of uncertainty in building design evaluations and its implications. *Energy and Buildings*, 34(9):951–958, 2002. A View of Energy and Bilding Performance Simulation at the start of the third millennium.
[87] C.J. Hopfe and J.L.M. Hensen. Uncertainty analysis in building performance simulation for design support. *Energy and Buildings*, 43(10):2798–2805, 2011.
[88] B. Saleh and F. Masseglia. Discovering frequent behaviors: time is an essential element of the context. *Knowledge and Information Systems*, 28(2):311–331, 2010.
[89] M. Delgado, W. Fajardo, and M. Molina-Solana. Representation model and learning algorithm for uncertain and imprecise multivariate behaviors, based on correlated trends. *Applied Soft Computing*, 36:589–598, 2015.
[90] H. Chen. Applications of fuzzy logic in data mining process. In Y. Bai, H. Zhuang, and D. Wang, editors, *Applications of Fuzzy Logic in Data Mining Process*, pages 249–260. Springer London, 2006.