Chapter 7

Data Analysis Challenges in the Future Energy Domain

Frank Eichinger

SAP Research Karlsruhe, Germany, f.eichinger@sap.com

Daniel Pathmaperuma

Karlsruhe Institute of Technology (KIT), Germany, daniel.pathmaperuma@kit.edu

Harald Vogt

acteno energy, Walldorf, Germany, harald.vogt@acteno.de

Emmanuel Müller

Karlsruhe Institute of Technology (KIT), Germany, emmanuel.mueller@kit.edu

Appears in: Ting Yu, Nitesh Chawla and Simeon Simoff (editors): Computational Intelligent Data Analysis for Sustainable Development, Chapman and Hall/CRC, 2013

7.1 Introduction

The energy system currently undergoes major changes, primarily triggered by the need for a more sustainable and secure energy supply. The traditional system relying on the combustion of fossil sources such as oil, gas and coal on the one side and nuclear technologies on the other side is not sustainable, for three main reasons: First, fossil and nuclear resources are limited and their exploitation will become more expensive (not economically sustainable). Second, the combustion of fossil sources leads to CO_2 emissions, which drive the greenhouse effect (not environmentally sustainable). Third, nuclear power plants bear certain risks in their operation and produce nuclear waste, which needs to be protected from unauthorized access. Furthermore, up to now no permanent disposal sites for nuclear waste exist, and coming generations – who will not have profited from this kind of energy source – will have to deal with it (not socially sustainable and probably not economically sustainable if all disposal costs are considered). The best way to achieve sustainability is clearly energy efficiency, i.e., all forms of saving energy. A further way are renewable energies. Luckily, they are evolving rapidly; most remarkably in the form of wind energy, photovoltaic systems, water power and biogas. In Germany, as one example, the share of renewable energies in electricity supply crossed the 20% mark in 2011 [7]. This share is planned to be quadrupled until 2050 [3]. Besides Germany, many other countries have similar plans in order to reduce greenhouse gas emissions and to fight climate change.

The rise of renewable energies comes along with a number of challenges regarding the electricity supply. Data-analysis techniques are a crucial building block which can facilitate these developments, as we will see later in this chapter. In particular, renewable energies challenge the electricity grid, which needs to be stable and should allow everybody at any point in time to consume energy. This is hard to achieve, as the electricity systems have to permanently maintain a balance between demand and supply. Storages can help achieving this, but their geographic availability is very limited, and the economics of storage do not make it feasible for large scale applications at the present time. The maintenance of the balance is getting more difficult as the share of renewable energy sources rises, due to their unsteady and fluctuating generation. This calls for so-called *smart grids*, which are a major topic of this chapter.

In particular, the future energy system – which aims to be less dependent on fossil fuels and nuclear technology and builds on more and more renewable energies – is challenged by the following four main factors:

- **Volatile Generation.** The possibly greatest issue of renewable energies is their volatile nature. It challenges the electricity system dramatically. The production of photovoltaic or wind energy does not depend on the consumers' needs, but solely on external conditions that are hard or impossible to control (e.g., general weather conditions). These do not necessarily match the energy demand patterns of consumers. In environments where renewables have a very small share of the overall production, such a natural fluctuation can be tolerated. However, the larger the share is, the more effort is needed to compensate for this effect.

Compensation for the fluctuating nature of renewables can be done, e.g., with flexible gas turbines that are permanently held in stand-by operation. They can increase production spontaneously, e.g., when clouds darken the sun and less electricity is produced. However, such stand-by operation is highly inefficient, and it is responsible for remarkable greenhouse gas emissions [136]. An alternative to extra production of energy is to shift demands, which will be targeted in more detail in the remainder of this chapter. If certain demands can be shifted to points in time where more renewable energy is available, this amount of energy can be considered as a virtual production, energy storage or buffer. Respective load-shifting programs and implementations are frequently referred to as *demand response.*[1]

In future scenarios however the situation can even be more severe. When large shares of electricity production come from renewable fluctuating sources, it might happen that the total production gets larger than the total demand. If the demand (including energy storage facilities) cannot be shifted to such periods any more, even the production of renewable energy has to be stopped, which makes its generation less efficient. This furthermore brings in the potential to substantially damage the energy generation and grid infrastructure.

- **Distributed Generation.** The future energy generation will be more distributed. Today, it builds on a comparatively small number of large, central power plants. These power plants feed-in energy from a higher voltage level to a lower voltage level, resulting in an unidirectional power flow. In the future, in particular due to the rise of renewable energies, the number of power-generating units is likely to rise dramatically and to be a lot more distributed. On the one side, this is due to the increasing number of photovoltaic systems installed on private and industrial buildings and due to wind turbines which are not always part of larger wind parks. On the other side, small biogas-based power plants and *cogeneration units (combined heat and power units, CHP units)* become more and more popular. CHP units can be a contribution to a more sustainable energy generation, too. This is as they are more efficient compared to pure fossil-based generation of electricity which does not make use of the waste heat. In addition, they can potentially be used to generate energy when production is low. In particular, so-called *micro CHP units* are increasingly installed in private houses, partly triggered by incentives

[1] *Demand response* is considered to be an element of the broader field of *demand-side management* which also includes energy efficiency measures [110].

from government programs. These developments turn many *consumers* into so-called *prosumers*. Prosumers are consumers owning, for example, photovoltaic systems or micro CHP units generating energy which is fed into the electricity grid if it exceeds their own demand.

Distributed generation challenges the electricity system, which has been designed for a more central generation with a comparably small number of large power plants. Much of today's grid infrastructure is built on the understanding and technology that power flows from higher to lower voltage levels. Distributed generation units however will operate at different voltage levels, which can potentially result in an unidirectional power flow within the grid. Issues arise, for instance, when photovoltaic panels installed on many roofs in a certain neighborhood feed their electricity into the local low-voltage distribution grid. These grids were originally designed solely to distribute energy from higher to lower voltage levels. Massively distributed generation can therefore lead to severe grid conditions (i.e., voltage and frequency fluctuation) and in the worst case to power outages [81]. This happens for instance when transmission-system capacities are temporarily not high enough or if there is a surplus of energy in a certain grid segment which cannot be transferred to higher-level parts of the grid. Another issue can be, for example, that wind parks in remote areas generate lots of energy which cannot be consumed locally. In such cases, the electricity grid might not have enough capacity to transport the electricity to areas where it is needed. Thus, wind turbines might have to be stopped temporarily. Solving such issues by means of increased grid capacities is surely possible, but very expensive in cases where such peak situations occur very rarely. Additionally, the construction of new power lines is very often opposed by local residents. Besides enhanced infrastructure, information technology will therefore play a more important role in the future.

- **New Loads.** Not only the supply side undergoes major changes, also on the demand side new loads are arising. In particular, *air conditioning*, *heat pumps* and *(hybrid) electric vehicles* become more and more popular. The latter is caused by the aim to make mobility more sustainable by driving vehicles with electric energy generated from renewable sources. It also makes countries more independent from fossil fuels. Several countries have programs to support such electric mobility. As one example, the German government has released the goal to have one million electric vehicles in Germany by 2020 and six million by 2030 [3]. This is ambitious, as by the end of 2010, only 40 thousand out of 42 million registered cars in Germany were (hybrid) electric vehicles [5].

  Electric mobility can only be sustainable if the consumed energy is sustainable, too. As electric mobility will lead to an increased consumption of energy, even more renewable sources are needed to satisfy this new demand. Furthermore, the demand for charging electric vehicles is highly volatile, and peak demands can hardly be supplied if they are not aligned with the production and distribution of energy. In addition, charging too many electric vehicles at the same time in the same segment of the electricity grid might lead to overloads and in the worst case to power outages. At the same time, the typical average load of the electricity grid leaves by far enough capacity to charge electric vehicles. Again, tackling such peak demands with increased grid capacities is expensive, and it does not solve the issue with insufficient production of renewable electricity at certain points in time. Therefore, an intelligent control of charging is necessary to integrate electric vehicles with the smart grid.

Electric mobility should not only be seen as a challenge, but as a chance to realize the smart grid. For example, intelligent techniques could schedule the charging processes of electric vehicles in order to avoid electricity network issues, to realize demand response (e.g., charge when renewable production is high and further demand is low) and to fulfill user needs (e.g., have the vehicle charged to a certain level when the user needs it). In addition to smart charging, the batteries of electric vehicles might also be used as a buffer in the electricity grid. This is, energy from the vehicle might be fed into the electricity network *(vehicle to grid, V2G)* when production is low or if there are demand peaks [118]. However, as electric mobility is still in its infancy, many challenges – including data analysis – need to be tackled to integrate it with the smart grid. As an example, to achieve user acceptance, intelligent systems need to capture and predict user behavior in order to have a vehicle sufficiently charged when the user needs it.

Besides electric mobility, *heat pumps* become more popular in certain regions. Such devices draw thermal energy from the environment for heating or cooling. They however are a non-negligible load in the electricity system, and might be used to shift demand to a certain extend.

- **Liberalization.** Besides the aim for sustainability and besides technological developments, the energy system is also challenged by legislation [104]. While traditionally electricity generation, transition, distribution and retail has been done by regional monopolies, the electricity market is now liberalized in many countries. Since the end of the 1990s, the mentioned tasks are separated and competition is introduced for generation and retail of electrical energy. Besides generation and retail companies, liberalization leads to a number of further actors, such as transition and distribution-network operators, balance-responsible parties, metering operators, value-added-service providers etc. Particularly actors related to the operation of the grid are typically regulated by governmental authorities. From a more technical perspective, a larger number of actors is involved in the electricity system. They all generate data, and this distributed data needs to be exchanged with other partners in order to fulfill their respective tasks. This opens-up new interesting possibilities for data analyses as well as the need for ways of treating data in a privacy-preserving way.

To wrap-up the challenges, sustainable energy systems will undergo major changes in the future. Generation and production will be more volatile and the landscape becomes more fragmented, both from a technical (distributed generation, volatile generation and consumption) and an organizational perspective (new and more specialized actors). Most important is the paradigm shift from *demand-driven generation* in the past to *generation-driven demand* in the future, triggered by renewable energy generation and new loads. Generation-driven management of energy consumption in the smart grid is a complex optimization problem [138]. It involves the operation of certain distributed energy sources and the control of energy consumption, for example via market-based mechanisms. In the liberalized electricity market, different actors will probably contribute to the solution of the optimization problem in a distributed manner. Further, the solution will likely be hierarchical: Certain optimizations will be done on the transition grid, further optimizations at the distribution grids, the next ones might be more fine-grained consumption and generation shifting between neighbor consumers and small energy generators. At these different levels, consumption, generation and grid-usage data will arise at different levels of aggregation and induce new challenges for data analysis.

Another rather technical development which affects current and future energy systems is *smart metering*. Smart electricity meters facilitate fine-grained measurements of energy consumption, production and quality and the communication of respective measurements. They are one of the technological foundations of many future energy scenarios described in this chapter. Also from a legislative point of view, their introduction is promoted, and many countries have respective programs. As one example, the European Union wants to achieve an 80% share of smart meters by 2020 [1].

This chapter will review selected scenarios in the field of the smart grid in more detail. Most of them contribute to the realization of the paradigm shift from demand-driven generation to generation-driven demand or to achieving energy efficiency. These elaborations will reveal a number of data-analysis challenges, which have to be tackled to implement the scenarios. These challenges will be discussed in more detail along with possible solutions.

The remainder of this chapter is organized as follows: Section 7.2 describes the current status of the electricity market. Section 7.3 reviews selected future-energy scenarios which are building blocks of the smart grid. Based on Sections 7.2 and 7.3, Section 7.4 describes the resulting data-analysis challenges and highlights first solutions. Section 7.5 concludes.

7.2 The Energy Market Today

In this section, we give a short overview about the current market for electric energy, with Europe as an example. This section is not only a basis to understand the future-energy scenarios described in the following section, it already bears some data-analysis challenges. In particular, prediction of consumption and generation plays already an important role in today's energy market. In this section, we will first introduce the different roles in the energy market (Section 7.2.1). Then, we describe energy trading (Section 7.2.2) and energy balancing in the electricity grid (Section 7.2.3).

7.2.1 Actors in the Energy Market

As mentioned in the introduction, liberalization leads to a number of new actors/roles in the energy market. Caused by regulation, this can vary in different countries, and some of the roles might be taken on by the same entity (e.g., a generator might also act as a retailer who sells energy to consumers). This number of (new) actors in the energy market is also interesting from a data-analysis point of view: Most of these actors have access to potentially interesting data or could profit from data provided by different actors. Investigating the actors and their data leads to interesting opportunities for data analysis and maybe even to new roles such as analytic service providers. In the following, we introduce the most common actors/roles that are relevant for the remainder of this chapter (see [2]):

- **Generator.** A company which produces electrical energy and feeds it into the transportation or distribution grid. Generators might use conventional power plants or generate renewable energy.
- **Consumer.** An industrial or private entity or person who consumes electrical energy.
- **Prosumer.** A consumer who also produces electrical energy. A difference to the generator is that the prosumer might entirely consume its own generation. The generation is typically done by renewable sources or micro CHP (combined heat and power) units.
- **Distribution System Operator (DSO).** Operates regional electricity distribution grids (low and medium voltage) which provide grid access to consumers, prosumers and small generators. Historically, such grids were intended to distribute centrally produced energy to the consumers. Today, it can happen as well that temporarily more energy is fed into a grid than energy is taken from the grid. The DSO also plans, builds and maintains the grid infrastructure.
- **Transmission System Operator (TSO).** Operates a transmission grid (high voltage) and transmits electrical energy in large quantities

over large distances. This includes providing grid access to large generators and consumers and to the DSOs. The TSO is responsible for the overall stability of its parts of the transmission grid and for providing balancing power (see Section 7.2.3). The TSO also plans, builds and maintains the grid infrastructure.

- **Balance-Responsible Party (BRP).** Is responsible that the scheduled supply of energy corresponds to the expected consumption of energy in its balance area. To achieve this, the BRP eliminates upcoming imbalances using balancing power with the help of the TSOs. The BRP financially regulates for any imbalances that arise.
- **Retailer.** A company which buys electrical energy from generators and sells it to consumers. The retailer has also to interact with DSOs and possibly metering operators to provide the grid access to the consumers.
- **Metering Operator.** Provides, installs and maintains metering equipment and measures the consumption and/or generation of electrical energy. The readings are then made accessible (possibly in an aggregated manner) to the retailer, to the consumer/prosumer and/or to other actors. Frequently, the role of the metering operator is taken on by DSOs.
- **Energy Market Operator.** An energy market may be operated for different kinds of energy to facilitate the efficient exchange of energy or related products such as load-shifting volumes (demand response). Typical markets may involve generators selling energy on a wholesale market and retailers buying energy. Energy market operators may employ different market mechanisms (e.g., auctions, reverse auctions) to support the trade of energy in a given legal framework.
- **Value-Added Service Providers.** Such providers can offer various services to different actors. One example could be to provide *analytic services* to final customers, based on the data from a metering operator.

7.2.2 Energy Trading

In order to supply their customers (consumers) with electrical energy, the retailers have to buy energy from generators. In the following, we will not consider how prosumers sell their energy, as this varies in different countries. While consumers traditionally pay a fixed rate per consumed kilowatt hour (kWh) of energy to the retailers (typically in addition to a fixed monthly fee), the retailers can negotiate prices with the generators (directly or at an energy exchange). A procurement strategy for a retailer may be to procure the larger and predictable amount of their energy need in advance on a long-term basis. This requires analytic services and sufficiently large collections of consumption data. The remaining energy demand which is hard to predict on the long run is procured on a short-term basis for typically higher prices. Similarly, generators of electricity need to predict in advance which amounts of energy they can sell.

7.2.2.1 Long-Term Trading

While electric energy has traditionally been traded by means of bilateral *over-the-counter (OTC)* contracts, an increasing amount of energy is nowadays traded at power exchanges. Such exchanges trade standardized products, which makes trading easier and increases the comparability of prices. While there are different ways of trading energy, double auctions as known from game theory [54] are the dominant means for finding the price [137].

As one example, the *European Energy Exchange AG (EEX)* in Leipzig, Germany trades different kinds of standardized energy futures and options. These products describe the potential delivery of a certain amount of energy at certain times in the future. The delivery has to be within one of the transportation grids. Besides the traded products, they also provide clearinghouse services for OTC contracts. The volume traded at the EEX derivate market for Germany amounts to 313 terawatt hours (TWh) in 2010, 1,208 TWh including OTC transactions [6]. The latter number corresponds roughly to two times of the energy consumed in Germany in the same time.

7.2.2.2 Short-Term Trading

Short-term trading becomes necessary as both consumption and production cannot be predicted with 100% accuracy. Therefore, not all needs for energy can be covered by long-term trading. In particular, fluctuating renewable energies make correct long-term predictions of production virtually impossible. As one example, wind energy production can only be predicted with sufficient accuracy for a few hours in advance. Therefore, energy exchanges are used for short-term trading, again making use of different kinds of auctions. At such exchanges, retailers can buy electrical energy for physical delivery in case their demand is not covered by futures. Again, the delivery has to be within one of the transportation grids.

The EPEX Spot SE (European Power Exchange) in Paris, France trades energy for several European markets. There, trading is divided as follows:

- In **day-ahead auctions**, electrical energy is traded for delivery on the following day in 1-hour intervals [6]. These auctions take place at noon on every single day. Bids can be done for individual hours or blocks of several hours, and the price can be positive or negative. Negative prices may occur, for instance, when the predicted regenerative production is very high and the demand is low, possibly at a public holiday.
- In **intra-day trading**, electrical energy is traded for delivery on the same or following day, again in 1-hour intervals [6]. Each hour can be traded until 45 minutes before delivery; starting at 3:00 pm, all hours of the following day can be traded. Bids can as well be done for individual hours or blocks of hours, and prices can again be positive or negative.

At the EPEX, 267 TWh were traded in 2010 by means of day-ahead auctions and 11 TWh during intra-day trading (German, French and Austrian market) [6]. The sum of these trades roughly corresponds to half of the electrical energy consumed in Germany in the same time (most of this energy is traded for the German market).

7.2.3 Energy Balancing

To assure a reliable and secure provision of electrical energy without any power outages, the energy grids have to be stable at any point in time. In particular, it needs to be ensured that the production always equals the consumption of energy. In practice, avoiding imbalances between generation and demand is challenging due to stochastic consumption behavior, unpredictable power-plant outages and fluctuating renewable production [137].

On a very coarse temporal granularity, a balance is achieved by means of energy trading (see Section 7.2.2) and data-analysis mechanisms, in particular prediction and forecast: Retailers buy the predicted demand of their customers, and generators sell their predicted generation. As mentioned in Section 7.2.1, the balance-responsible parties (BRPs) make sure that the scheduled supply of energy corresponds to the expected consumption of energy. This expected consumption is also derived using data-analysis techniques. The transmission system operators (TSOs) are responsible for the stability of the grids. In the following, we describe how they do so.

From a technical point of view, a decrease in demand leads to an increase of frequency, while a decrease in production leads to a decrease of frequency (and vice versa for increases). Deviations from the fixed frequency of 50 Hz in electricity grids need to be avoided in real time, as this might lead to damage of the devices attached to the grid.

Typically, frequency control is realized in a three-stage process: *primary*, *secondary* and *tertiary control*. The primary control is responsible for very short deviations (15–30 seconds), the secondary control for short deviations (max. 5 minutes) and the tertiary control for longer deviations (max. 15 minutes) [137]. The control process can be realized by various means, e.g., standby generators, load variation of power plants or load shifting. Different measures involving different types of power plants have different reaction times and are therefore used for the different control levels.

The capacities needed for balancing the grids are again frequently traded by means of auctions: Primary and secondary capacities are traded biannually, while tertiary capacity is traded on a daily basis [137]. Usually, only pre-qualified actors are allowed in such balancing markets, as it needs to be ensured that the requirements at the respective level can be fulfilled technically. Primary-control bids consist of the offered capacity and the price for actual delivery of balancing power. In contrast, bids for secondary and tertiary control consist of one price for making available a certain capacity and a price for consumed energy [137]. Thus, capacities play an important role in the balancing market and actors are partly paid for a *potential supply*, which hinders a generator from selling this energy at the regular market.

For costs that arise from the actual energy delivered for secondary and tertiary control, the generators who caused deviations or retailers who did not procure the correct amount of energy for their customers are charged [81, 137]. Costs for primary control and for the capacities of secondary and tertiary control are paid by the corresponding grid operators [137], who earn grid-usage fees. Since costs at the balancing market are typically higher than on the normal energy market, generators and retailers are stimulated to make best-possible predictions of energy production/consumption [81].

7.3 Future Energy Scenarios

In this section, we describe a number of visionary energy scenarios for the future. They represent building blocks of a possible *smart grid*. We describe the scenarios from a researcher's point of view – they assume certain technical developments and may require legislative changes. We have selected scenarios which are commonly discussed in the scientific and industrial communities, either in the described form or in some variation.

Scenario 1 (Access to Smart-Meter Data) Smart metering *is a key building block of the* smart grid*, as many further scenarios rely on the availability of fine-grained information of energy consumption and production. For instance, smart metering can enhance load forecasting and enable demand response, dynamic pricing etc. Some of these scenarios are described in the following. However, smart metering is not only an enabler for other scenarios. Giving users access to their energy consumption profiles can make them more aware of their consumption and improve energy efficiency. This is important as many consumers have little knowledge about their energy consumption.*

For purposes of billing, smart-meter data is typically generated in 15-minute intervals. This is, the meter transfers the accumulated consumption of the consumer every 15 minutes to the metering operator. Technically, smart meters can increase the temporal resolution of consumption data, e.g., measure the consumption within every second or minute. This allows to get a detailed picture of the energy consumption, down to the identification of individual devices (e.g., coffee machines), as each device has its typical load curve. Such fine-grained data could also be transferred to a metering operator. In addition, metering data at any granularity can be made available within a home network, e.g., to be accessed via visualization tools on a tablet computer.

Access to consumption profiles for energy consumers can be more than pure numbers or simple plots. Respective visualization tools can easily show comparisons to previous days, weeks etc. In the case of service providers, they can also provide comparisons to peer groups of consumers having similar households (in terms of size and appliances used). Furthermore, devices can be identified by *their load profile [113], which can be visualized additionally. This leads to an increased awareness of the energy consumption of each single device.*

A number of studies have investigated the effects of giving users access to smart-metering data. The authors of [126] have carried out controlled field experiments with 2,000 consumers and came to the conclusion that giving access to detailed consumption data may lower the total energy consumption moderately by about 4%. Other (meta) studies suggest that savings can be even a little higher [39, 45, 94].

Scenario 2 (Demand Response with Dynamic Prices) *Energy retailers procure parts of their energy needs at the spot market, where the prices reflect the actual availability of (renewable) energy (see Section 7.2.2.2). At the same time, energy consumers typically procure energy for a fixed price per kilowatt hour (kWh), additionally to a consumption-independent monthly fee. Although this is comfortable for the consumers and the sum of the monthly or annual bill can be easily calculated if the consumption is known, there are no incentives to consume energy when large amounts are available (e.g., when wind-energy production is high) or to save energy when production is low. As discussed in the introduction, such a paradigm shift from a demand-driven generation to a generation-driven demand is necessary to facilitate the integration of renewable energies. One approach to realize this is* dynamic pricing. *In a nutshell, the retailer communicates the actual or future prices of energy per kWh to their consumers, and they can decide if they react to such incentives or if they prefer to consume energy when they want it while tolerating possibly higher prices.*

The predecessors of dynamic prices are tariff schemes where electricity is cheaper during the night, reflecting a lower average demand at this time. However, this does not consider the fluctuating production of renewable energy. Dynamic price mechanisms can take these fluctuations into account and incorporate them in the prediction of renewable generation and demand of the consumers. Dynamic price schemes mainly differ in the range of possible prices, in the resolution of time ranges in which prices are valid and in the time spans in which the prices are communicated in advance. Very short time spans enable highly dynamic demand response, but also make it hard to plan energy consumption. Furthermore, legislation may demand certain minimum time spans. A more detailed review of dynamic prices can be found in [70].

The German research project MeRegio *investigates the user acceptance of so-called 'energy traffic lights' (see [4] for more information). These are small devices receiving the energy tariffs for the forthcoming 24 hours via radio. The granularity of prices comprises three discrete levels, i.e., 'low', 'normal' and 'high'. Consumers can see these levels and plan accordingly. In addition, the device visualizes the three levels in different colors (e.g., red stands for 'high') which makes the consumers more aware of the current price of electricity. Of course, dynamic prices are not only intended for human interpretation. Intelligent devices can receive such price signals and decide when a certain process should be started (a dishwasher can for instance be* *started automatically when the lowest tariff starts). This automation is done in more extend in so-called* smart homes *which are described in Scenario 4.*

Besides the promising approach to realize demand response, dynamic prices also bear risks. As one example, it might happen that many devices start operation when a low-price time span starts. This may challenge the distribution grid as sudden significant increases in demand can hardly be handled by energy balancing mechanisms and should therefore be avoided. This might be realized by using individual dynamic prices for the different consumers (slight differences in order to smoothen demand curves), or by having a much finer granularity of prices. If granularity is fine enough, different user preferences might lead to a more wide-spread time span in which devices are started.

Many scientific studies have investigated the user acceptance and efficiency of dynamic pricing. A meta study [50] has investigated the results of 24 different pilots. The result in almost all of these studies is that consumers do accept dynamic prices and adapt their behavior to a certain extend. Concretely, these studies show that dynamic price schemes are an efficient demand-response measure – a median peak reduction of 12% could be achieved. Naturally, the user acceptance is particularly high if users are supported by intelligent technologies such as in a smart home *(see Scenario 4) [109].*

Scenario 3 (Market-Based Demand Response with Control Signals) *When energy retailers experience energy shortages during a day, they buy energy at intra-day exchanges (see Section 7.2.2.2). When renewable production is low, this can be very expensive and it might be a better option to ask their customers to consume less energy within a certain time frame. Similarly, grid operators monitor the electricity grid and may want to ask consumers to temporarily reduce their consumption in order to achieve grid stability. This scenario describes an alternative to dynamic prices (see Scenario 2) for demand response, with a focus on solving grid issues. It requires respective contracts between the consumers and the involved parties that describe the incentives for the consumers to participate in the respective measures (e.g., reductions of the energy bill). Concretely, consumers contract with specialized* demand-side management companies *which might be the local distribution system operators (DSOs). Further, an infrastructure is required that can execute demand-response measures. This scenario describes market mechanisms different from dynamic prices that can be used for trading flexible loads for demand-response purposes, relying on such an infrastructure.*

Energy retailers considering a demand-response measure can send their request to an electronic marketplace. Equally, if overloads or voltage problems are detected, the grid operator can issue a similar request. Then, all affected demand-side management companies receive these requests from the marketplace, and each company submits an offer for solving the issue. The marketplace selects a combination of offers which fulfills the request. If the retailer or grid operator accepts the assembly of offers (the retailer might prefer to alternatively buy energy at the exchange if this is cheaper), the respective demand-side management companies are then responsible to conduct the *demand-response request. The companies then send priority signals to smart-home control boxes of their contracted consumers. The control boxes send the signal to intelligent devices in the consumer's premises as well as charging infrastructure of electric vehicles.*

Scenario 4 (Smart Homes) Home automation *is known under the name* smart home *for quite some time. However, until lately, this meant mainly providing comfort features such as automatically shutting window blinds, switching the light on automatically or controlling the house's air conditioning system in accordance with current weather conditions. Smart homes are often equipped with some kind of energy production unit (e.g., photo voltaic or micro CHP), and they can employ a central optimization component which controls most generation and utilization of energy in the house. Recently, the potential of home automation for smart-grid applications has been recognized [17]. While control components in smart homes primarily act based on the user's presets, they could also take the current situation in the energy grid into account [23].*

Some applications in a household have the potential to shift their time of operation into the future, others might run earlier than they normally would. Shifting their power consumption in time realizes demand response. *Of course, not all applications in a house are suitable for such a purpose. Applications like lighting or cooking bear no potential for a time-variable application [60, 111]. The case is different for the dishwasher or the washing machine. These appliances normally have no need for an immediate start and in that way present a potential for shifting the power demand to the future. As one example, users want to have clean dishes for dinner. Whether they get cleaned right after lunch or sometime in the afternoon does not matter.*

Another possibility of shifting power demand is the variation of temperature in heating or cooling applications. Such applications include fridges, air conditioning and heating systems. They all have to maintain a temperature within a certain range. The tolerable range in a freezer would be around -18° C with a tolerance of ±2° C. In normal operation mode, the device would cool down to -20° C, then pause until the temperature reaches -16° C and then start over again. An intelligent system would be able to interrupt the cooling at -18° C or start cooling already at this temperature without waiting for a further rise, thus shifting the power demand of the cooling device. The same scheme could be applied to the room temperature, as the comfort range normally lies around 21° C. Extended knowledge of user preferences could expand this potential even further. If the resident wants the temperature to be 21° C on the return in the evening, the system could heat the house up to 25° C in the afternoon and let it cool down slowly to the desired 21° C. This would require more energy in total, but could still be feasible in a future-energy system where a lot of solar energy is available (and is thus cheap) during the day [73].

Profiles of typical user behavior could improve the demand shifting capabilities of a smart home even further if they were combined with an electric vehicle. Not only could the charging profile of the vehicle be matched to the user's and energy system's demands, but the battery of the EV could also be *used as temporary energy storage when the vehicle is not needed. This concept is known as* vehicle to grid *(*V2G*, see Scenario 5) [118].*

Scenario 5 (Energy Storage) *Storing electric energy becomes more important as the share of volatile energy production (like solar or wind power) increases. Although electric energy is hard to store, some technical solutions exist. In the context of storages for electrical power, one has to keep in mind some key parameters of such systems. The first is the* efficiency *and refers to the percentage of the stored energy which can be regained. It is always below 100%. Other parameters are the* capacity *and the* peak power *of a storage system, which refer to the performance of such systems.*

Today, the only storage systems with the ability of storing a relevant amount of energy for a reasonable price are pumped-storage water-power plants*. They store energy by pumping water from a low reservoir up into a high storage and regain this potential energy by letting the water run down through generators, just like in water-power plants. While these storages could store energy for an almost unlimited time, their land use is quiet high and they require a natural height difference, which makes it difficult to realize such storages in densely inhabited regions.*

Another option is the installation of large battery-based chemical storage facilities*. This is already done in some regions of Japan. Battery systems can be manufactured with almost any desired capacity and peak power. The drawback of battery systems is their relatively high price. With the anticipated rise in the market share of electric vehicles, this could change soon. Batteries loose capacity constantly during their lifespan. At a certain point in time, their power density is not high enough to use them as batteries for electric vehicles. However, power density is negligible in the context of immobile storage. Thus, battery storage facilities could benefit from a relevant market share of electric vehicles by reusing their old batteries.*

A further storage option would be the V2G concept [118]. As vehicles spend only about 5% of their lifetime on the road, an electric vehicle could potentially be used 95% of the time as local energy storage. Assuming one million electric vehicles in one country (the German government for instance aims at reaching this number by 2020 [3]), this could sum up to about 15 gigawatt hours (GWh) of storage capacity with a peak power of 3–20 gigawatt (GW), resembling 2–5 typical pumped-storage water-power plants.

The profitability of a storage system is, depending on the business model, correlated with its usage which nowadays is proportional to the number of store-drain cycles. As battery quality factors decrease not only with their lifetime but also with each use cycle, the utilization of such systems has to be considered with reservations.

Scenario 6 (Management Decisions Derived from Energy Data) *Energy has become a major cost factor for industry, public facilities, warehouses, small and medium enterprises and even for universities. Thus, the management in any of these organizations demands for decision support based on the* *energy consumption of the organization. Let's take the example of a university where the data center is planning to install a new generation of servers. Not only the acquisition costs, but also cooling, space and especially energy consumption have to be considered. While for most of such management decisions IT-solutions are available, energy consumption remains an open issue ignored by most enterprise resource-planning (ERP) tools. Energy management opens several new aspects to be considered. In particular with the availability of novel data sources, such as smart meters, we observe a high potential for such data-analysis toolkits.*

Automated data analysis or semi-automated data exploration can give an overview of the entire energy consumption of a university, a department, a single institute or in other dimensions detailed energy consumption for all servers, all personal computers and many more of such orthogonal views on the energy consumption. It is essential to provide such a multitude of views on different aggregation levels as management decisions might demand arbitrary selections of this huge information space. They require reports for such selected views, automated detection of suspicious consumption, comparison between different views, estimation of future consumption etc.

The essential challenge is the large variety of management actions that base their decisions on such energy data. Each of these decisions poses different challenges on data storage, data analysis and data interaction. Furthermore, they address different management levels, and thus, subparts of an organization. For example, individual reports might be required for each professor about the energy consumption of the institute in the past few months. Such reports have to show a detailed view on the energy consumption, distinguishing between different rooms or consumer classes. Optimally, interesting views for each institute would be selected (semi-)automatically. On the other side, some automated fault-detection algorithms might be required for the maintenance department of a university. Techniques require an intuitive description of failing situations in contrast to the regular behavior of the facility under observation. Going up to the dean of a department or any higher instance in the university one requires more general and aggregated views. Typically, such information is required for strategic planning of new facilities, new buildings, or the estimation of future energy consumption.

Overall, we have highlighted several – quite different – management decisions that pose novel challenges to data analysis. They could be realized by novel data acquisition with smart meters. However, neither data storage of such large data volumes, nor its analysis has been tackled by recent toolkits. It is an emerging application domain in database and data-mining research.

Further scenarios are described in the following chapters in this book. In particular, Chapter 8 describes a scenario which deals with finding the best mix of renewable demand management and storage, and Chapter 9 focuses on a scenario that deals with real-time identification of grid disruptions.

7.4 Data Analysis Challenges

With the rise of the *smart grid*, more data will be collected than before, at finer granularity. This facilitates innovative technologies and a better control of the whole energy system. As one example, the availability of both consumption/generation data and predictions facilitates the realization of demand-side-management techniques such as demand response. Ultimately, this allows a better integration of renewables and a more sustainable energy system. The new data sources and new technologies in the *future energy scenarios* (see Section 7.3) call for more advanced data management and data analysis as it has already been used in the traditional energy system (see Section 7.2). This section describes the data-analysis challenges in the energy area and presents first solutions. In particular, we look at data management (Section 7.4.1), data preparation (Section 7.4.2), the wide field of predictions, forecasts and classifications (Section 7.4.3), pattern detection (Section 7.4.4), disaggregation (Section 7.4.5) and interactive exploration (Section 7.4.6). Finally, we comment on optimization problems (Section 7.4.7) and the emerging and challenging field of privacy-preserving data mining (Section 7.4.8).

7.4.1 Data Management

Before addressing the actual data-analysis challenges, we now present some considerations regarding data management. As motivated before, the rise of the *smart grid* leads to many large and new data sources. The most prominent sources of such data are *smart meters* (see Scenario 1). However, there are many more data sources, ranging from dynamic prices to data describing demand-response measures, usage of energy storages and events in smart homes. In the following, we focus on smart-meter data. In Section 7.4.6.1, we deal with further data-management aspects in the context of exploration and comparison of energy datasets.

As described before, smart meters are able to measure energy consumption and/or generation at high resolution, e.g., using intervals of one second. Figure 7.1 provides an example of such measurements and shows a typical electricity-consumption curve in a two-person office with a resolution of one second (see Section 7.4.2 and [135] for more details on the data).

From a data-analysis point of view, storing data at finest granularity for long periods and for many smart meters would certainly be interesting. However, besides privacy concerns (see Section 7.4.8), this might not be possible from a technical point of view. Therefore, one has to decide which amounts of data need to be kept for which purpose. In many cases, not the finest granularity is needed and samples or aggregations of meter data suffice.

Table 7.1 illustrates the amounts of measurements and storage needs of smart-meter data, assuming that a single meter reading needs four bytes (B)

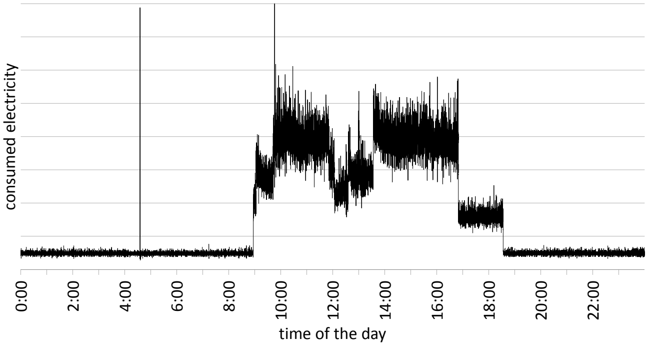


Figure 7.1: Typical electricity consumption in an office.

in a database.[2] In the rows, the table contains different granularities ranging from one second (finest granularity provided by many meters) to one year (period of manual meter readings frequently used today). In the columns, the table contains the number of measurements and the respective storage needs both for one day and one year. For instance, data at the one-second granularity sums up to 32 mio. meter readings per year, corresponding to 120 megabytes (MB). If one would like to collect such data for 40 mio. smart meters in one country (roughly in the size of Germany [143]), this would sum up to four petabytes (PB). As another example, the 15-minute granularity typically used for billing purposes still leads to five terrabytes (TB) in a whole country. Note that real memory consumption can easily be twelvefold as mentioned above [125]. Managing these amounts of data is still challenging.

metering		# measurements		storage need							
granularity		1 day	1 year	1 day		1 year		1 day		1 year	
1	second	86.400	31.536.000	338	kB	120	MB	13	TB	4	PB
1	minute	1.440	525.600	6	kB	2	MB	215	GB	76	TB
15	minutes	96	35.040	384	B	137	kB	14	GB	5	TB
1	hour	24	8.760	96	B	34	kB	4	GB	1	TB
1	day	1	365	4	B	1	kB	153	MB	54	GB
1	month		12			48	B			2	GB
1	year		1			4	B			153	MB
		1 smart meter						40 mio. smart meters			

Table 7.1: Storage-needs for smart-meter data (pure meter readings only).

As illustrated by Table 7.1, smart meters might lead to huge amounts of data. This is similar for other data sources in future energy systems. As mentioned before, every actor involved in the energy system will only be responsible for certain subsets of the existing data. This might still lead to amounts of data which challenge data-management infrastructure. Concrete

[2] This does not include meta data such as date, time and location; [125] reports that the size including such data could be much larger, i.e., by a factor of twelve.

challenges for the respective actors are the selection of relevant data as well as aggregation and sampling of such data without loss of important information.

Several researchers have investigated storage architectures for smart-meter data: [19] has investigated centralized and distributed relational databases, key-value stores and hybrid database/file-system architectures, [125] presents results with in-memory databases [112], [26] presents experiences with the Hadoop [139] MapReduce [41] framework, and [120] has investigated further cloud-storage techniques. Apart from that, smart-meter readings can be managed with techniques from *data streams* [10], as fine-grained readings can be seen as such a stream (see Section 7.4.4).

One approach to deal with huge amounts of data is compression. It might ease the storage using one of the architectures mentioned previously. [125] for example reports a compression factor of eight when using *lossless* compression techniques in database technology (on metering data including meta data). Using *lossy* compression techniques on fine-grained data (see Figure 7.1 for an example) that approximate the original data seems to make compression factors of several hundred possible, depending on the required accuracy and granularity of data. Such an approximation of *time-series data* can be done with various *regression models*, for instance using straight line functions [35, 47], linear combinations of basis functions or non-linear functions (using respective approximation techniques, e.g., described in [127, 130]). However, the authors of this chapter are not aware of any studies which investigate the trade-off between compression ratio, computational costs and usefulness of lossy compressed data for different applications based on smart-meter data of different temporal granularities. Further, lossy compression techniques would need to be integrated with data-management technology. Investigating respective techniques and validating their deployment in realistic scenarios – in particular with regard to the trade-off mentioned – is an open research problem.

Data from smart meters belongs to the group of *time-series data* [86]. Besides compression via regression techniques and the actual storage of such data, many other data-management aspects are of importance. This includes indices and similarity-based retrieval of time series (surveys of these techniques can be found in [52, 64, 134]). Such techniques are of importance for many analytical applications that are based on such data. For example, indexes and similarity search can be used to retrieve consumers with a similar electricity demand, which is important in classification and clustering (see Sections 7.4.3.2 and 7.4.4.1, respectively). Investigating the usage of the mentioned techniques from time-series analysis in the context of energy data would be promising as they are rarely mentioned in the literature.

7.4.2 Data Preprocessing

Data preprocessing is an essential step in data-analysis projects in any domain [30]. It deals with preparing data to be stored, processed or analyzed and with cleaning it from unnecessary and problematic artifacts. It has been stated that preprocessing takes 50% to 70% of the total time of analytical projects [30]. This certainly applies to the energy domain as well, but the exact value is obviously highly dependent on the project and the data. In the following, we highlight some preprocessing challenges that are specific for the energy domain. – Many further data-quality issues as they are present in many other domains might be important here as well (see, e.g., textbooks such as [24, 61, 140] for further issues and techniques).

Data from smart meters frequently contains *outliers*. Certain outliers rather refer to measurement errors than to real consumption, as can be seen in the raw data visualized in Figure 7.1: The peaks roughly at 04:30 and at 10:00 happening at single seconds are caused by a malfunction of measurement equipment. The smart meter has malfunctioned for some seconds resulting in an accumulated consumption reported at the next measurement point. Such outliers have to be eliminated if certain functions need to be applied afterwards. For example, calculating the maximum consumption of uncleaned data in Figure 7.1 would not be meaningful. Other outliers might refer to untypical events or days: Consumption patterns of energy might differ significantly when there is, e.g., a world-cup final on TV or if a week day becomes a public holiday. (Figure 7.2(b) illustrates that load profiles at weekdays and weekends are quite different.) Such exceptional consumption patterns should not be used as a basis for predictions of 'normal' days, but analyzing them might be particularly interesting if one approaches a similar special event. We elaborate a bit more on unsupervised learning techniques for preprocessing – in particular cluster analysis and outlier detection – in Section 7.4.4.

Another common problem with smart meter data are *timing issues*: It might happen that (i) a smart meter operating at the one-second granularity produces a few measurements too much or too little during one day (or week, month etc.) or (ii) that a meter operating at the 15-minutes granularity does its measurements not exactly on the quarter hour. Both cases might be negligible in certain situations, but need to be tackled in other situations. While one missing second might be quite meaningless, ignoring it might be problematic in the light of laws on weights and measurements. Billing in the presence of dynamic energy prices (see Scenario 2) might require measurements at exact points in time. If measurements are, say, five minutes delayed, this could make significant differences (e.g., when the start of energy-intensive processes are scheduled to start when a cheap time span starts). A possible solution for the first problem would be to add/subtract the missing/additional measurements to/from the neighboring ones. The second problem might be solved using regression techniques that enable estimations of measurements at arbitrary points in time.

7.4.3 Predictions, Forecasts and Classifications

Predictions and classifications belong to the most important data-analysis challenges in the energy domain. This is not only true for future scenarios as described in Section 7.3, but is already crucial today: Predictions of consumption and generation are essential for making profits at today's energy markets (see Section 7.2). In the following, we elaborate on time-series forecasting first. Then, we focus on predictions and classifications of consumers and their behavior before we discuss time-dependent events.

7.4.3.1 Time-Series Forecasting

As mentioned beforehand, numerical predictions of time-dependent data – also called *time-series forecasting* – are crucial in today's and future energy scenarios. In the following, we list a number of scenarios where this is the case:

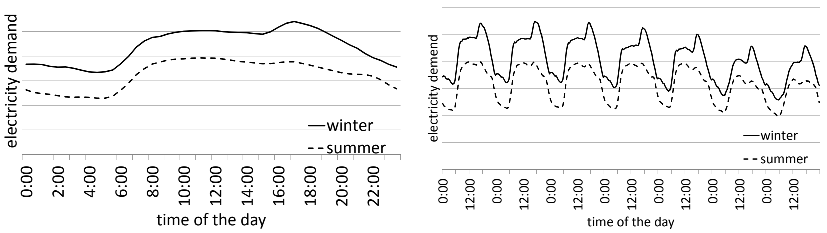
- **Predicting consumer demand** is needed in many different scenarios: (1) In *energy trading* (see Section 7.2.2), retailers are interested in predicting the demand of their customers. The more precise this prediction is, the more energy can be bought at potentially cheaper long-term markets instead of buying it at the intra-day market or to pay for energy balancing. Buying more 'cheap' energy than needed is also unprofitable, as retailers have to pay for it even if their customers do not use it. (2) To realize *dynamic pricing* (see Scenario 2), retailers need to know the predicted consumption of their customers. This is to derive the prices in a way that the customers might shift loads to other time spans in order to align consumption with previously procured energy or predicted renewable production. (3) To *avoid grid issues*, balance responsible parties need to plan their grid capacities based on the predicted load in the respective areas. If predicted high loads (e.g., when charging electric vehicles) are supposed to cause grid issues, *demand-response scenarios* could ease the situation (see Scenarios 2 and 3). (4) *Smart homes* (see Scenario 4) typically plan their energy consumption and generation in advance, requiring the predicted consumption. (5) An operator of *energy-storage facilities* (see Scenario 5) needs to know the predicted consumption in order to plan its operation accordingly. (6) Deriving *management decisions* (see Scenario 6) frequently requires not only information on actual consumption, but also on the forecasts. As an example, this allows assessing if a certain unit within an organization consumes less or more than predicted.
- **Predicting renewable generation** is needed in exactly the same scenarios as predicting consumer consumption of conventional energy. This is as generation and consumption have to be equal at all times. Thus all mechanisms requiring predicted consumption also need to know the predicted generation of energy.
- **Predicting grid loads** is important on the short run and on the long run: Knowing the predicted grid load for certain segments for several hours in advance is important for planning *demand-response measures* (see Scenarios 2 and 3) and operating *energy-storage facilities* (see Scenario 5). Having an estimate for the grid load in several years is impor-

tant for electricity grid planners (i.e., the distribution system and transmission system operator, see Section 7.2.1) who need "to guide their decisions about what to build, when to build and where to build" [85].

- **Predicting flexible capacities** is needed in several scenarios and is related to the prediction of consumer demand, as an energy demand can only be shifted if it actually exists. Concretely, the requirements of the *demand-response scenarios* considered in this chapter are as follows: (1) To bid for demand shifting, a demand-side manager in the control-signal scenario (see Scenario 3) needs to know the load-shifting potential of its customers as precisely as possible. (2) A retailer who uses *dynamic prices* (see Scenario 2) needs to estimate the number of customers who react to price incentives, along with the respective volumes. This requires knowing how much load can potentially be shifted.
- **Predicting energy storage capacities** is helpful in *storage scenarios* (see Scenario 5). As storage operators typically aim to maximize profit by means of *energy trading* (see Section 7.2.2), they need to know the future capacities. This can be an input for optimization algorithms which determine the scheduling of filling-up and emptying an energy storage.
- **Predicting energy prices** is certainly not easy, but there might be some regularity in energy prices which facilitate forecasting. Concretely, the following two directions are of interest: (1) If one knows the future energy prices with a certain probability in *energy trading* (see Section 7.2.2), one can obviously make large benefits. For example, in the presence of *demand response* (see Scenarios 2 and 3), one can shift loads of customers to cheaper points in time. (2) In the presence of *dynamic prices* (see Scenario 2) which are not known long in advance, one can make its own predictions of the energy price and speculatively adjust the consumption. This could be done in particular in highly automated *smart homes* (see Scenario 4).

All these scenarios are different, but they deal with the same problem from a technical point of view: *time-series forecasting*. However, the different scenarios require different data. Historical generation and consumption data from smart meters – possibly aggregated from many of them – is the basis for most scenarios. Other scenarios rely on historical storage capacities, data on demand-response measures conducted in the past, energy prices or they require external data such as weather forecasts for predicting renewable generation. In the following, we focus on predictions of consumption. The other predictions mentioned before can be treated in a similar way with their own specific data.

Predictions and forecasts can generally be done by learning from historical data. In the case of energy consumption, this is a promising approach, as there are certain regularities in the data: (1) The consumption within one day is typically similar. People get up in the morning and switch the light on, they cook at noon and watch TV in the evening. Figure 7.2(a) illustrates two typical demand curves during two different days, aggregated for all consumers



(a) day

(b) week

Figure 7.2: Typical aggregated demand curves. Data taken from [8].

in the UK. (2) The consumption at weekdays is typically similar, while it is different at weekends and national holidays where, e.g., factories are not running. Figure 7.2(b) describes the typical energy consumption in the course of one week. (3) The electricity consumption in winter is different from the summer. This is caused by differing usage of electrical light and possibly heating. Figure 7.2 illustrates this lower demand in summer as well as different consumption patterns in winter and summer days.

The probably easiest approach for predictions of consumption is to average the curves of a certain number of similar days in the past, which do not refer to special events. As one example, to predict the demand of a particular Sunday, one could average the demand from the past four Sundays where no special events occurred. This could be improved by increasing the influence of Sundays having a similar weather forecast.

The different approaches for time-series forecasts differ not only in the techniques involved, but also in the time span for the predictions: Are predictions needed for the next couple of hours, for the next day, next month or next year? In general, time-series forecast techniques can be categorized as follows [36]:

- **Auto regression** is a group of techniques using mathematical models that utilize previous values of the time series. Some of these techniques, called *moving average*, rely on sliding-window approaches using historical time series. Various enhancements are used to deal, e.g., with seasonal effects in energy data.
- **Exponential-smoothing techniques** are moving-average approaches, but use a weighting with factors decaying exponentially over time.
  Many of the concrete approaches for auto regression and exponential smoothing rely on parameter estimation, for which various techniques can be used.
- Several techniques from **machine learning** have as well been adapted to time-series forecasting. This includes *artificial neural networks*, *Bayesian networks* and *support-vector machines*. See, e.g., textbooks such as [24, 61, 95, 140] for descriptions of these algorithms.

[36] is an extensive review of all the previously mentioned techniques in the context of energy, [16] is another one. [66] particularly reviews *neural-network approaches*, which can be combined with similar-day approaches mentioned beforehand [93]. While a large number of papers focuses on predictions of consumption of energy, many of them can be used for other predictions as well. Besides more general reviews [16, 36, 66], [14] reviews *price-forecasting techniques* in particular. Another direction of work is the forecast of wind-power production [20, 84, 89]. The application of some of the above-mentioned time-series forecast techniques has been investigated in this context for both short and long-term predictions, based on data from wind-energy production and meteorological observations.

[37] is a study on hierarchical *distributed forecasting* of time-series of energy demand and supply. Such approaches tackle explicitly the huge amounts of data that might need to be considered when making forecasts at higher levels such as a whole country (see Section 7.4.1). Besides distributed forecasting, the authors also deal with the important problem of *forecast model maintenance* and reuse previous models and their parameter combinations [38].

Time-series forecasting seems to be quite a mature field, but it is still a challenge for the future energy domain. It has been applied to forecasting demand, generation and prices, but there is little literature available regarding the other future-energy scenarios listed above. Particularly in the light of dynamic pricing (see Scenario 2), other demand-response measures (e.g., Scenario 3), energy storage (see Scenario 5) and distributed and volatile small-scale generation (see Section 7.1), predictions of consumer demand, grid usage etc. become much more challenging. This is as many more factors than pure historical time series are needed to make accurate predictions. The following Section 7.4.3.2 sheds some light on the human factor, but many further factors need to be integrated in an appropriate way to achieve high-quality forecasts which are needed in the smart grid. (Many future energy scenarios require extremely high accuracies of predictions, i.e., even small deviations from the optimal result may cause huge costs.) This calls for more research on the question which factors are useful in which situation and which forecast model (or ensemble thereof) to use for which task when certain data is available. These questions can certainly not be answered in general and need to be addressed individually. However, some guidance and experience would be of high practical relevance for new smart-grid scenarios.

7.4.3.2 Predicting and Classifying User Behavior

Predicting and classifying users and their behavior is one of the most popular applications of data mining. This is as well the case in the energy domain. We assemble an exemplary list of respective challenges in the following:

- Electricity retailers (see Section 7.2.1) acting in a very competitive market want to **classify customers for marketing reasons**. For example, if they would like to introduce a new tariff scheme targeting a certain group of consumers, say families living in apartments, they would like to select this target group for marketing campaigns, based on the energy consumption patterns.

- In demand-response scenarios relying on dynamic prices or control signals (see Scenarios 2 and 3), the respective parties would like to **predict which consumers will participate** in a certain demand-response measure (e.g., a price incentive) and **how much demand could be shifted** with this particular measure. Similar predictions are of relevance in smart homes.
- In smart homes (see Scenario 4), **user behavior classification** can decide whether a user will go to work, will stay at home, will use an electric vehicle etc. This is important for scheduling the energy generation and consumption. Similar classifications are important in the field of electric vehicles (see the paragraph on *new loads* in Section 7.1 and Scenario 5). Intelligent charging and vehicle-to-grid (V2G) mechanisms [118] need to know, e.g., whether the user will behave as usual and will solely drive to work and back or if the user might plan any longer or further trips.

Again, individual challenges require different data ranging from general customer data and smart-meter readings to data describing demand-response measures, events in a smart home and electric-vehicle usage. As prediction and classification are very mature fields in data mining and machine learning, a large number of potentially relevant techniques is available. This includes *decision tree classifiers*, *neural networks*, *support vector machines*, *naïve Bayes classifiers* and *k-nearest neighbor algorithms*. More information can be found in the relevant literature, e.g., [24, 61, 95, 140]. However, such classifiers cannot be applied directly to all kinds of relevant data in order to predict behavior. If, e.g., sequential data of behavioral events is available, combined approaches [29] might be needed to deal with the data. To cope with time-series data from smart meters, aggregates have to be calculated to feed the data into standard classifiers. Alternatively, more specific time-series techniques [86] can be applied (see Section 7.4.1), e.g., specialized *time-series classification* [56, 76].

A few works on classifying electricity consumers are available in the literature. [116] first uses clustering techniques to identify different groups of customers (see Section 7.4.4.1). Then, the authors assemble feature vectors and use a standard decision-tree classifier to learn these groups and to automatically assign new consumers to them. They assemble the features from averaged and normalized daily load profiles of the consumers by defining a number of aggregates. These aggregates include ratios between peak demand and average demand, ratios of energy consumed at lunch time, at night etc. [90] extracts its features differently. The authors use the average and the peak demand of a consumer, as well as coefficients from time-series-forecasting techniques [36]. For classification, the authors employ linear discriminate analysis.

Predicting and classifying the behavior of customers has been an important application of data analysis in the past. Surprisingly, not so much research has been conducted in the context of energy consumer behavior. However, as more market roles are arising (see Section 7.2.1) and potentially more data will be collected, the need of such analytics will increase. Some analytic challenges can certainly be solved by means of established techniques from data mining and machine learning. As data might be complex and come from different sources, there is also a need for further developing specific algorithms and to combine different analysis techniques (see, e.g., [29]). Prediction and classification of energy customer behavior is therefore an important field in *domain-driven data mining* [28].

7.4.3.3 Predicting and Classifying Consumption Events

In future energy systems, there are a number of challenges involving prediction and classification of events and consumption patterns:

- Optimized control and planning in a **smart home** (see Scenario 4) requires the detection of load profiles and the prediction of events, together with the respective forecasts [117, 141].
- In the **smart-meter scenario** (see Scenario 1), the visualization could be enhanced by displaying not only a household's total consumption, but to disaggregate the load curve into the different appliances and highlight them in the visualization. This would increase user awareness and boost energy efficiency. Pattern-recognition algorithms can be used to identify appliances within the households load curve [63, 87].
- Another use case for load-pattern recognition are **cross-selling activities** conducted by, e.g., value-added service providers. By analyzing single appliances, special sales offers could be triggered in cases where new energy-efficient appliances are available.
- **Energy-efficiency** effects become more important with the size of the loads that are considered. Therefore, identifying consumption patterns is of importance **in complex environments** as described in Scenario 6.
- **Charging electric vehicles** will become a major load in the electricity grids. To illustrate, driving 10 km to work and back resembles four loads of a washing machine. These loads need to be predicted in a reliable way in order to allow the future energy system to make appropriate schedules both within a smart house and in the whole electricity grid [118].
- Detection and prediction of **user behavior events in electric mobility** (an event could be to start a certain trip) is quite complex as a vehicle is often used by multiple persons. Such knowledge and predictions are essential to facilitate smart charging of electric vehicles and V2G (see Scenario 5).
- Massively distributed generation and new loads (see Section 7.1) can lead to **problematic grid situations**. Detecting and predicting such events is a major topic in smart grids. Chapter 8 in this book elaborates this in a comprehensive way.

From a technical point of view, the mentioned challenges can be divided into two parts: Prediction of events and classification of consumption patterns. Abundant research has been conducted in the field of pattern detection from smart-meter data. This has partly been published in the privacy domain [96, 113] (see Section 7.4.8). Pattern detection is also a basic block for *disaggregation techniques* which we describe in more detail along with the techniques in Section 7.4.5. Early works have already shown that the electricity consumption of a whole house can be disaggregated into the major appliances with high accuracy [49]. [108] is a survey of load profiling methods. Event prediction has received less attention in the context of energy. While traditional techniques like *sequence mining* [62] can be used in principle to predict discrete events [46], further techniques from *machine learning* have been adapted recently. For instance, [124] performs event prediction in the field of electricity consumption with neural networks and support vector machines.

To sum up, there is a huge need for the prediction of events and for classification and prediction of consumption patterns. On the one side, quite some research has been conducted in pattern detection (classification of patterns), partly in the context of *disaggregation* (see Section 7.4.5). On the other side, techniques for predicting consumption patterns and events of user behavior can still be improved for application in the field of future energy. As the demand for accurate techniques is clearly given, respective research would be a chance to support the developments of the smart grid significantly.

7.4.4 Detection on Unknown Patterns

In many of the described energy scenarios, data analysis is needed to detect novel, unknown and unexpected knowledge. Such knowledge is represented by hidden patterns describing correlation of energy measurements, groups of similar consumers or deviating objects such as a single household with unexpected energy consumption. In all of these cases no information is given about the type of pattern, its characteristics and there are no example instances known for this pattern. Thus, this unsupervised learning is clearly different from the prediction techniques described in Section 7.4.3. In the following, we describe pattern detection techniques focusing on *clustering*, *outlier mining* and *subspace analysis*. We will highlight the applicability of these techniques in the energy domain. However, we will also point out open challenges not yet addressed by these data analysis paradigms.

Let us start with a brief overview of clustering applications on energy data:

- **Unsupervised learning as preprocessing step.** In most cases the proposed techniques, such as clustering and outlier mining, are used as preprocessing steps to other data-analysis tasks. For example, for prediction tasks (see Section 7.4.3) it is essential to know about substructures in the data. One can train specific classifiers for each individual cluster of customers. In other cases one can extract novel features by cluster analysis and use these features for prediction of unknown objects. Outlier analysis can be used to clean the data, it removes rare and unexpected objects that hinder the learning process. Pattern detection can assist in all of the previously mentioned scenarios (see Section 7.3) as a data preprocessing step. However, it is also of high value for knowledge discovery as described in the following two cases.

- **Pattern detection for enhanced demand response.** In demand response (e.g., for dynamic prices in Scenario 2), one tries to match energy production with energy consumption, which requires a deep understanding of both sides. For the generation side, one has proposed prediction techniques that are used to forecast wind or solar energy production (see Section 7.4.3.1). For the consumption side, besides forecasts, one is interested in customer profiles that provide insights about their daily behavior (see Section 7.4.3.2). As behaviors change dramatically over time, one cannot always rely on historic data and learning algorithms. Thus, unsupervised methods (e.g., clustering or outlier mining) are means for this kind of data analysis. Clustering algorithms detect groups of customers showing highly similar behavior, without any prior knowledge about these groups. In particular for demand-side management, these clusters can be used for specific strategies in each customer group. While some customers will not be able or willing to participate in some management actions, others will show high potential for shifting parts of their energy consumption. It is essential to be aware of such groupings to utilize the overall potential of demand-side management.
- **Automatic smart home surveillance.** As one part of smart homes (see Scenario 4), we discussed demand response in the previous example. However, smart homes have further potential for data analysis tasks. Having all the energy consumption data of smart homes available, one can design automated surveillance mechanisms assisting, e.g., elderly people in their daily living. Energetic profiles are very detailed and reveal a lot of information about our daily behavior and can be used for tracking, warning and assistance systems. For example in assisted home living, it is crucial to know if elderly people change their daily habits. A youngster who typically moves around and uses many electric devices throughout the day will become highly suspicious if she or he stops this behavior for one day. Such dramatic changes can be detected as unexpected patterns and used as warnings for medical or other assistance parties. This example highlights the requirements for unsupervised learning techniques. Although some patterns might be learned with supervised techniques, most unexpected behavior will be new for the system and hard to be learned. In particular, we observe outlier mining to be one of the key techniques in the area of energy data analysis.
- **Cluster customers for marketing reasons.** Corresponding to the case of a classification of the customers (see Section 7.4.3.2), cluster analysis can detect the different groups of customers of an electricity retailer. This promises interesting insights of the customer base and is a basis for the design of tariffs.

7.4.4.1 Clustering

Let us now abstract from these individual scenarios and discuss some well-known techniques in pattern detection. Clustering is an unsupervised data-mining task for grouping of objects based on their mutual similarity [61]. Algorithms can detect groups of similar objects. Thus, they separate pairs of dissimilar objects into different groups. A large variety of approaches has been proposed for convex clusters [43, 92], density-based clusters [48, 65] and spectral clustering [106, 107]. Further extensions have been proposed for specific data types such as time series [76, 114]. All of these approaches differ in their underlying cluster definitions. However, they have one major property in common: They all output a single set of clusters, i.e., one partitioning of the data that assigns each object to a single cluster [102].

Let us discuss this single clustering solution for customer segmentation based on smart meter data. One has given a database of customers (objects) that are described by a number of properties (attributes). These attributes can be various types of information derived from smart-meter measurements (see Scenario 1). For example, each customer has a certain set of devices. For each device one may detect its individual energy consumption and additional information about the time points these devices are used in the household [78] (see Sections 7.4.3.3 and 7.4.5 for further details about the identification of devices). Obviously, one can detect groups of customers owning different types of devices. This grouping can be used to separate customers in different advertisement campaigns (expensive devices, low-budget devices, energy-efficient devices and many more). However, in contrast to this simple partitioning one might be interested in several other groupings: Each customer is part of groups with respect to the daily profile (early leaving, home office, part-time working), or with respect to current living situation (single households, family without children, with children, elderly people). This example highlights the need for multiple clustering solutions on a single database [102]. In particular, with the large number of attributes given, it is unclear which of them are relevant. It is an additional challenge for data analysis to select these attributes.

Dimensionality reduction techniques have been proposed to select a set of attributes. They tackle the *'curse of dimensionality'* which hinders meaningful clustering [25]. Irrelevant attributes obscure the cluster patterns in the data. Global dimensionality techniques such as *principle components analysis (PCA)*, reduce the number of attributes [71]. However, the reduction may obtain only a single clustering in the reduced space. For locally varying attribute relevance, this means that some clusters will be missed that do not show up in the reduced space. Moreover, dimensionality reduction techniques are unable to identify clusterings in different reduced spaces. Objects may be part of distinct clusters in different data projections. Our customer segmentation example highlights this property. Each cluster requires an individual set of relevant dimensions.

Recent years have seen increasing research in clustering in high dimensional spaces. Projected clustering or subspace clustering aims at identifying the locally relevant reduction of attributes for each cluster. In particular, subspace clustering allows identifying several possible subspaces for any object. Thus, an object may be part of more than one cluster in different subspaces [101, 128].

For customer segmentation based on energy profiles, several open challenges arise in the detection of object groups and the detection of relevant attributes for each of these groups. Many private and public organizations collect large amounts of energy measurements, however their relevance for individual patterns is still unclear. Neither clustering on the full set of attributes is a solution, nor a pre-selection of relevant attributes by dimensionality reduction techniques. Costly search in all possible projections of the data has to be performed to identify multiple clustering solutions with respect to different attribute combinations. Thus, scalability of such data-mining techniques will be a major challenge that has been highlighted by a recent study [101]. As described in Section 7.4.1, smart-meter readings will provide huge databases. Only a few publications have focused recently on such scalability issues [34, 98, 99]. However, most subspace clustering models are still based on inefficient processing schemes [15, 72, 100]. Further challenges arise in the stream processing of such data [11]. It raises questions on the detection of clusters, but also on their tracking over time [59, 129]. Tracking the change of customer profiles will be an essential means for tracking the energy demands in demand-site management systems or online adjustment of energy prices in future energy markets.

The literature in the field of energy data analysis has focused only on clustering similar consumers or consumption profiles, making use of similar preprocessing techniques as in classification (see Section 7.4.3.2). Examples of such works include [90, 116, 132]. However, they do not address the mentioned challenges in tracing, multiple clustering solutions and local projection of data, which leave a high potential for enhanced clustering results.

7.4.4.2 Outlier Mining

In contrast to clusters (groups of similar objects), outliers are highly deviating objects. Outliers can be rare, unexpected and suspicious data objects in a database. They can be detected for data cleaning, but in many cases they provide additional and useful knowledge about the database. Thus, pattern detection considers outliers as very valuable patterns hidden in today's data. In our previous example, suspicious customers might be detected that deviate from the residual customers. Considering the neighboring households, one might observe very high energy consumption for heating devices. While all other households in this neighborhood use oil or gas for heating, the outlier is using electric heating. There have been different outlier detection paradigms proposed in the literature to detect such outliers. Techniques range from deviation-based methods [119], distance-based methods [80] to density-based methods [27]. For example, density-based methods compute a score for each object by measuring its degree of deviation with respect to a local neighborhood. Thus, one is able to detect local density variations between low-density outliers and their high density (clustered) neighborhood. Note that in our example the neighborhood has been literally the geographic neighborhood of the household. However, it can be an arbitrary neighborhood considering other attributes (e.g., similarity to other customers with respect to the set of devices used).

Similarly to the clustering task, we observe open challenges in online stream analysis for outlier detection [9], the detection of local outliers in subspace projections [12, 103] and the scalability to large and high dimensional databases [40]. An additional challenge is the description of such outlier patterns. Most approaches focus only on the detection of highly deviating objects. Only few consider their description [18, 79]. Similar to subspace clustering, it seems very promising to select relevant attribute combinations as descriptions. Based on this general idea of subspace mining in arbitrary projections of the data, several preprocessing techniques for the selection of subspaces have been proposed [33, 74]. They try to measure the contrast between outliers and the clustered regions of a database. A first approach proposes a selection based on the entropy measure [33]. A subspace is selected if it has low entropy, i.e., if it shows a large variation in the densities. More recent approaches have focused on statistical selection of high contrast regions [74]. They compare the deviation of densities and utilize only the most significant subspaces. Such subspaces can be seen as the reasons for high deviation. In our example, high contrast subspaces might be ‘energy consumption of heating devices’ and ‘age of the refrigerator’. This combination might be characteristic for the distinction of modern vs. old households and might reveal some unexpected cases with old refrigerators (that should be exchanged) in an energy-efficient house. This example shows that it is important to detect these cases. However, it is even more important to provide good explanations why these cases show high deviation.

Looking at the future users of such outlier mining techniques in, e.g., smart homes, we observe that most of them will be people with no background in data analysis. This will raise new visualization and explanation requirements for result presentation. It requires novel data-mining techniques that are able to highlight the differences between patterns. For instance, such techniques could reveal the reason for a high deviation of a single object or the difference between two groups of customers. First techniques in this direction have been proposed [22]. Given two different data sets, they try to measure the difference and output characteristic properties to distinguish between these sets. Very promising instantiations have been applied to emerging pattern detection or novelty detection in the context of stream data [44].

As outlier mining is an established technique in data mining, it has been used in quite some works in the field of energy data. To name some examples, [90] uses outlier mining to detect abnormal energy use. [32] defines so-called load cleansing problems for energy data and develop techniques similar to outlier mining. [67] presents specialized outlier detection algorithms dedicated to power datasets collected in smart homes. All of these approaches can be seen as first instantiations of simple outlier models. Further potential of energy data has to be exploited by more enhanced outlier detection techniques.

7.4.5 Disaggregation

For achieving energy efficiency, deep knowledge about the distribution of the consumed power among the devices within a facility is important (see Scenarios 1 and 6). In practice, this is often achieved by installing metering devices directly at single devices, which is expensive, time-consuming and usually not exhaustive. It would be easier to derive the power distribution from the metered data at the interface to the grid (see as well Section 7.4.3.3).

Smart metering, i.e., high-resolution metering and remote transmission of metered data, promises to provide that deep look into the infrastructure at all metering points. Techniques for achieving this are commonly called *non-intrusive (appliance) load monitoring (NILM*, sometimes also *NALM)* or *disaggregation* of power data. This has potential applications in achieving better energy efficiency (see Scenarios 1 and 6) and in facilitating demand response (see Scenarios 2 and 3) and load management (e.g., in a *smart home*, see Scenario 4). Thus, the topic has sparked increased interest recently [31, 53, 55, 78, 82, 91, 142] after quite some research (including, e.g., [49, 87]) that has been done since the first paper was published in 1992 [63].

Common smart meters in residential and industrial environments are placed at the interface to the distribution grid. They measure the active and reactive energy used by all the devices that are connected to the electric circuit that originates at the meter. Additional values can be measured, such as peak loads. Multiple meters can be installed at a single facility, which is usually the case if separated billing of the consumed energy is required. For billing purposes, such meters pick up the consumed energy typically in intervals of 15 minutes. However, an interface with a higher temporal resolution is usually provided at the meter itself that can be accessed locally.

As these meters are increasingly available, it is tempting to use the metered data for analytical purposes as well. In a residential setting, transparency of energy consumption may lead to energy conservation (see Scenario 1). NILM has also been proposed as a tool for verifying the effectiveness of demand-response measures. In industrial or commercial settings, an energy audit is a valuable tool for identifying potentials for energy efficiency (see Scenario 6). Such audits can be executed more thoroughly the more detailed information is available. The (temporary) installation of sub-meters is therefore commonly practiced and could be, at least partially, substituted by NILM.

An example of real energy data available to load disaggregation is visualized in Figure 7.1. If this load curve would represent what one has been doing throughout that day, one would be able to assign labels to certain patterns in the load curve. These labels would describe events or activities of that day.

However, if somebody else looks at the load curve, they cannot directly infer information about one's daily activities. They might be able to identify certain devices, such as an oven. The lack of contextual information limits the usage of this data. This calls for automated disaggregation and visualization as described in Section 7.4.3.3.

The load curve of a factory or a commercial building is much more complex than that of a household or an individual person (see Scenario 6). Many more individuals and devices are contributing to the load curve, many of them with individual behavior. Complex industrial processes are executed at the same time. However, there is a lot of contextual information available that can be used to identify individual devices and processes.

Besides the curve representing the real power over time, also values such as reactive power, peak current and possibly other electrical features can be used. From these time series, a lot of higher-level information can be deduced, such as features in the frequency domain, the instantaneous admittance waveform and more [91].

A sophisticated energy auditing system should be enabled to map process information to load patterns. It should be able to identify recurring patterns. It might be necessary to initialize this system with additional knowledge. Load disaggregation should be as accurate as possible and with a tunable relation between false positives and false negatives.

7.4.5.1 Approaches to Load Disaggregation

In the following, we will describe the fundamental problem of load disaggregation. Further, we will describe the fundamental approach to the problem, and discuss some recent work. For a thorough description of historic and recent work in this field, we refer to [142] and to [53] for an overview of device characteristics that can be useful for disaggregation.

A load curve is a function L which describes the complex load (real and reactive energy) over time. In each discrete time step t, the real energy consumed by all devices r_i and their respective reactive energy q_i is summed up. Noise r_b and q_b is added as well:

$$L(t) = \sum_{i=0}^{N} (r_i(t), q_i(t)) + (r_b, q_b)$$

Given only the resulting sum value over time, we are looking for a state matrix which contains the state of each device at any discrete point in time. The state spaces of the devices are independent of each other. For most practical devices, there exist several constraints on the possible states and the state transitions that are caused by their internal electrical structure and their usage modes. For example, all practical devices are operating between a minimum load and a maximum load, and they have finitely many operating states.

There are two fundamental steps to be made for load disaggregation. The first step is *feature recognition*, which extracts features from the observed meter data. The second step is the application of an *optimization* algorithm that assigns values to the state matrix.

Pattern recognition is being applied to the observed values (see Section 7.4.3.3), in its simplest form to a change in the real power load. The objective of this step is to identify a set of devices that may exhibit the observed pattern. A naïve algorithm could map a load change to the device that exhibits the closest step size of the observed change. An ideal algorithm would perfectly identify the cause of an observed event as either a fluctuation not caused by a state change, or the very device and its state change that caused the change. However, no such perfect algorithm exists today and false positive matches and false negatives are unavoidable.

There is a variety of features that can be used to find a valid disaggregation. The most basic feature of a device is its load variance, which was used in [63]. Based on this feature, four classes of devices can be identified: *permanent*, *on-off*, *multi-state* and *variable*. Permanent devices are single-state and are consuming the same load at all times (e.g. alarming systems that are never switched off). On-off devices have an additional off-state, where consumption is (near-)zero. Multi-state devices have multiple operating modes, which are usually executed in certain patterns, e.g., a washing machine has certain modes such as heating water, pumping or spinning. Variable load devices expose arbitrary, irregular load patterns, which may depend on their actual usage mode. It is important to note that most practical devices cannot be fully characterized by one of these classes alone. Usually, a device exhibits behavior that is a complex mixture of these classes. The challenge of disaggregating such loads is complicated by the fact that, of course, the complex load profiles of devices are superimposed on each other, which makes an accurate, non-ambiguous disaggregation hard to achieve.

Since basic features, which are also referred to as *macroscopic*, such as consumption or real and reactive power, have their limitations, features on the *microscopic* level have been studied in order to obtain more accurate results (see [142]). Microscopic features refer to characteristics of the underlying electrical signal, which can be measured at frequencies of at least in the kHz-range. This allows the identification of waveform patterns and the harmonics of the signal. Using these features yields better results than disaggregation based on basic features alone. However, such measurements require dedicated hardware and additional processing capacities, which limits their practical use.

The *optimization* step (which is a common task in data analysis, see Section 7.4.7) tries to find an assignment to the state matrix that best matches the observed values. This answers the question which device was active during which period and at which power level.

A common approach to finding the state matrix is to create a *hidden Markov model (HMM)* of the system [31, 78, 82]. Each device is represented by a HMM, which is a flexible structure that can capture complex behavior. Roughly, a sharp change in power consumption corresponds to a state change within a device HMM. The challenge is to extract the HMM parameters from the observed meter data. This is often supported by a supervised training phase where known features are being used.

7.4.5.2 Practical Applications

Accurate load disaggregation could replace sub-metering, at least for some applications. But even with the currently available level of accuracy, useful applications seem feasible. For example, [31] is using meter data from water consumption to identify activities such as showering or washing. This work improves results by evaluating the specific context in which load disaggregation is being used. Usage patterns depending on time of day, household size, and demographics help to derive statistical information about appliance use, such as the distribution of washing machine usage. Reportedly, it also helped people to make decisions about more efficient resource usage, e.g. by replacing appliances with more efficient ones.

It remains a challenge to improve the accuracy of NILM for practical applications. Many studies assume that the features of the involved devices are known in advance. In such *supervised* settings, it is required to determine the features of individual devices in a controlled environment. In contrast, *unsupervised* techniques have recently been proposed [55, 78]. This class of techniques does not rely on a given decomposition of power signals from individual devices but instead automatically separates the different consumption signals without training. Although unsupervised techniques seem to work in practice, research shows that the quality drops when increasing the number of devices [78].

The accuracy of the existing approaches has only been tested under individual lab conditions so far. A common methodology for evaluation is missing, for example no systematic testing on a common dataset has been performed yet. Only recently, a set of test data has been proposed [83], which comprises residential appliances. Similarly, test data for industrial applications are required, but are not freely available. Some notion of accuracy is usually used to assess the quality of an approach. As discussed in [142], this might not be the desirable measure. Thus, the authors propose to use *receiver operating characteristic (ROC) curves* as a quality measure (see, e.g., [24, 61, 140]).

Notwithstanding the deficiencies of the measure itself, it is clear that none of the existing approaches is suitable to completely substitute sub-metering due to inaccuracy. Thus, disaggregation is not suitable for billing and other applications that require accurate and precise measures. It is likely that sub-metering or separate metering will be required to satisfy these demands at least in the near future. For 'soft' applications like energy-efficiency auditing, the accuracy of load disaggregation might be sufficient in many cases. However, no evaluation in a working environment has been reported so far. Sub-metering still is the state-of-the-art when it comes to accurate load disaggregation. Research is still required to demonstrate NILM's practicability as there seem to be no reports on large-scale field tests of NILM. Furthermore, the literature mentioned in this section employs data at the one-second granularity or at even finer temporal resolutions. Further research is needed to investigate if and under which conditions disaggregation techniques can be applied to data at coarser granularities which is frequently available in practice.

In the future, a semi-automatic approach to load disaggregation might be practically valuable for energy audits. Graphical, interactive exploration tools could be used to validate the automatic recognition of devices and correct for errors. After the consumption patterns of individual devices (or classes of devices) are identified, the next step would be to correlate these patterns with additional data, such as operational data from production runs, working hours, or out-of-order events. By doing so, higher accuracies could be obtained.

7.4.6 Exploration and Comparison of Energy Datasets

In the previous sections, the focus has been on automatic learning: (1) for models in prediction tasks (Section 7.4.3), (2) extraction of unexpected and novel patterns (Section 7.4.4) and (3) disaggregation of devices by meta data extraction (Section 7.4.5). In contrast to these automatic techniques, many of the energy scenarios such as Scenarios 1 and 6 require a manual exploration of data. Users want to understand the underlying data and use data management and analysis techniques to get an overview of their data. In many cases they try to derive knowledge out of the data by comparing two or more different data sets. Assisting these manual or semi-automated exploration tasks will be the main focus in the following.

Let us give some brief examples of semi-automated and manual tasks on energy data related to exploration and comparison:

- **Exploration of energy trades.** For the energy market, it is crucial to know about the amount of trades with specific conditions. All market participants are interested in such manual selections to understand the market. They explore the trades with respect to some manually defined condition. For example, how many trades have been made with solar energy, in the last month, overall in Germany. Others might be interested in the overall volume of energy produced in wind parks in off-shore regions located around the coast of Denmark. A third example might be the average capacity of energy storage facilities in Europe and how it evolved from 2010 to 2011. All these examples are user-driven queries on aggregated subparts of the data. Human experts design these queries and data management techniques have to be designed for an efficient processing.
- **Comparison of different customer behaviors.** Many case studies [57] have looked at the difference in energy consumption for two or more given databases describing the energy behavior of different customers. One is interested in the characteristic behavior of one group of people (or devices, facilities etc.) compared to a second group. These characteristic differences are used to understand the customer population. In other

cases the two contrasting data states are 'before' and 'after' an energy saving campaign. Thus, comparison is required to measure the success of such a campaign. For example, semi-automated techniques can derive the reduced energy consumption for lighting and cooling devices. This energy saving can be an important result for future campaigns and might reveal some more potentials in energy saving.

- **Manual verification of unexpected events.** As discussed in Section 7.4.3, there are some rare events such as the world-cup final on TV which effect the typical energy consumption dramatically. In simple cases, such as the world-cup final, the understanding of this event is quite easy. Experts will not need any assistance in the verification of this event. However, for both energy production and energy consumption there is a large variety of factors influencing the system. Many of the unexpected events, e.g. the detected outliers as described in Section 7.4.4.2, or the emerging events hindering good prediction in Section 7.4.3, will require assistance in their verification. Providing a time stamp and the unexpected high energy consumption might be very limited information for the human expert. Understanding and verifying such events means that we have to enrich the set of information provided to the user.

7.4.6.1 Extending Data Management Techniques

We observe the efficient aggregation of energy data as one of the main tasks for manual exploration. Abstracting from our toy example based on energy trades, the users are interested in various aggregations of the raw production, consumption, distribution and sales data. In general, such processing is well known in the database community as *online analytical processing (OLAP).* Based on a user-specified hypothesis, the system has to provide aggregated information with respect to a specific set of conditions. The conditions are described by the attributes (e.g., location, time, production type etc.) and are structured based on a given hierarchy of granularities (e.g., weeks, months, years etc.). Such OLAP systems have been proposed for sales analysis in retail companies. They provide the essential means for decision making but do not address the specific scenario of decisions based on energy data (see Scenario 6).

Essential properties of energy production, distribution and consumption are missed by these techniques. Modern data-management techniques (see Section 7.4.1) try to overcome these challenges. In particular, large data volumes have to be aggregated in main memory. This processing can be assisted by modern column-based data storage [112, 125]. In contrast to row-based data representation, the column-based storage allows an efficient aggregation over a single (or multiple) attribute without accessing the entire database. This selectivity allows very efficient processing of OLAP queries. In addition, we can utilize automated techniques such as disaggregation (see Section 7.4.5) to enrich the set of available attributes. This results in large and high dimensional databases, which pose novel scalability issues to both manual (and semi-automated) exploration as well as to automatic data analysis. For future energy data, we have to extend traditional OLAP and data-mining techniques to achieve such a scalable data analysis.

Currently, most energy case studies rely on the traditional techniques in OLAP and information management [57]. They are not able to cope with the entire set of information available. They design their interactive exploration on a small subset of attributes with quite rough aggregation levels (e.g., allowing exploration of energy data only on a daily basis). Further restrictions are made for query types and visualization methods. Overall, we consider such systems only as first steps to future energy information systems. The state-of-the-art has not reached the complexity of data, user exploration and interaction required by most of the energy scenarios envisioned.

7.4.6.2 Guided Exploration to Unexpected Patterns

One major challenge of OLAP is its manual search for interesting patterns in the data. It is highly depending on the expert using the OLAP system. If she or he knows a lot about the energy data it will be easy to find the right aggregation level, the appropriate set of attributes and the conditions on these attributes. Thus, one might be able to reveal the required information out of the huge database. However, in most cases this information is unexpected such that even experts do not know where to search. Furthermore, if lay users are involved in the OLAP system, they do not have any idea where to start with the aggregation. Thus, it is very important to provide semi-automated techniques that guide the user through the database to the unexpected aggregates and the right attribute combinations.

In recent years, there have been some interesting approaches for the so-called discovery-driven OLAP systems. They add automatic techniques to the OLAP system, which guide the users according to unexpected data distributions [122]. Comparing the mean and variance of each column of the database, one can simply detect unexpected cells in an OLAP cube. For example, if we look at the energy production in each month, one could detect a high peak in August, which deviates from the residual months due to some unexpected high-energy production. The same statistics can be applied for all August months over several years and highlight a specific year. This leads to a very promising selection of attribute combinations, each with a high deviation in their energy production. Overall, these unexpected measures can be seen as candidates for a manual exploration. One can provide some of these attribute combinations to the user and he or she will be able to refine these selections.

Further techniques have been proposed for guided OLAP [121, 123], they focus more on the interaction with the users and provide additional means for the step-by-step processing through the OLAP cube and additional descriptions on the deviation of data. However, all these techniques are expensive in terms of computation. Similar to other automatic data-analysis techniques they do not scale to energy databases with many attributes and millions of measurements on the very fine-grained level. Applications of such techniques are always limited by the efficiency, and energy data poses one of the most problematic application areas with respect to scalability issues.

7.4.6.3 Contrast Analysis and Emerging Patterns

Another automated approach for pattern exploration is contrast analysis. This technique has its focus on the extraction of descriptive, distinguishing, emerging and contrasting patterns for two or more given classes in a database [22, 44, 79]. Contrast analysis techniques provide subsets of attributes (and attribute values) as contrasting patterns. For example, given a database with more than two persons living in the same household and another database of energy profiles for single households, one might be interested in the comparison of these two groups of customers. Such automatic comparison can provide the characteristic differences in the behavior of people. On the one side, these differences can be used as input to any learning task. On the other side, it provides the essential mean for human exploration. We will focus on the later one and highlight the technical challenges in contrast analysis.

For human exploration it is always essential to have outputs that are easy to understand. In contrast analysis, one research direction is based on so-called contrast sets [22]. They form characteristic attribute combinations that show high deviation in the two databases. For example, the energy consumption with respect to washing machines could be one of these characteristic differences between single and family households. Contrast analysis detects such deviations and outputs a set of these contrasting properties for further investigation by the user. It is quite similar to the previous discovery-driven OLAP techniques. However, it is based on prior knowledge about the two classes that is not given in OLAP. Hence, it is based on some prior knowledge and provides a specific insight to these two classes, while discovery-driven OLAP highlights any unexpected data distribution. Further relations can be observed to subspace analysis (see Section 7.4.4), which is quite similar to the extraction of influential attributes [79].

Overall, we observe a high demand for such exploration and comparison techniques. For energy databases with many unexpected events, it is essential to have some descriptive information about the differences to other databases or the deviation of an object inside a database. In all of these cases, automatic techniques are guidance for humans in their manual exploration. Only the combination of manual and automatic exploration seems to be able to reveal the hidden knowledge out of complex databases. With many of the proposed techniques for prediction, pattern detection and disaggregation one can perform some fully-automated data analyses. However, in most cases users are not willing to accept these black-box techniques, in which they do not understand the derived models, patterns and separation. Furthermore, similar to other domains such as health surveillance, we observe many regulations by law in the energy domain. This enforces the manual verification of automatically detected patterns. Modern data-analysis techniques should be aware of this additional requirement and provide more descriptions as outputs of their algorithms. For example, it is more or less useless to detect unexpected energy consumption in a single household if one has no information about *why* this consumption profile is unexpected compared to other households.

7.4.7 Optimization

In the context of future energy and smart grids, there is a large number of different optimization problems that need to be solved. As elaborating on all these problems would be beyond the scope of this chapter, we limit ourselves to highlight the most important problems.

Optimization problems in the field of electricity can be roughly partitioned in the demand side and the supply side:

- On the **demand side**, intelligent devices (e.g., in a *smart home*, see Scenario 4) need to *react to dynamic prices* (see Scenario 2) and optimize their demand planning accordingly. If consumers own micro CHP units, they have to find optimized schedules for their operation. A further challenge is *charging of electric vehicles* and possibly *V2G scenarios* (see Scenario 5). This is not only of importance in consumer premises, but also in so-called *smart car parks* [115]. They have a particularly high impact on the energy systems as they display high power consumption. *Disaggregation* is a further technique which makes use of optimization (see Section 7.4.5).
- On the **supply side**, the probably most prominent optimization problem is *finding dynamic prices* [70]. In scenarios with *control signals*, optimization is needed to select offers from demand-side managers (see Scenario 3). Another field where optimization is of relevance is the *management of energy storages* (see Scenario 5) where it needs to be decided when to charge and when to discharge a storage.

Many of the mentioned optimization problems become quite complex. This is in particular due to frequently many parameters to be considered:

- Planning of **micro CHP units** has constraints concerning their profitability. This includes minimum runtime, uptime and cycle costs. Similar constraints apply to **central storages** such as pumped-storage water-power plants.
- **Smart charging of electric vehicles** and **V2G** requires to consider user preferences (when does the car need to be charged to which level?) and economic interests of the owner or car park operator. Furthermore, the current situation of the grid and possibly dynamic prices are of importance [115].
- **Finding dynamic prices** aims at achieving the desired profits and realizing demand response to prevent grid issues with a low economical

risk. Furthermore, the predicted generation and demand needs to be taken into account (see Section 7.4.3.1), together with the predicted willingness and ability of customers to react accordingly (see Section 7.4.3.2). Obviously, the current market prices and possibly existing long-term contracts are further parameters.
- In **control-signal scenarios**, available demand-shifting offers need to be selected, taking into account that they are cost-efficient, reliable and located in the correct grid segments.
- In **energy storages**, the operator has to consider the (predicted) future generation, demand and prices, as well as the storage-system parameters capacity and peak power.

The result of the mentioned conditions and constraints are often high-dimensional, multivariate optimization problems. Besides classical solving methods [51], heuristic methods [42, 75, 88, 97] have been an important field of research in recent years. For smart-charging scenarios for instance, multi-objective evolutionary optimization algorithms have been investigated [115].

7.4.8 Privacy-Preserving Data Mining

As discussed in this chapter, an increasing number of actors in the liberalized electricity markets (see Section 7.2.1) collect more and more data (see Section 7.4.1) when realizing the current and future energy scenarios (see Section 7.3). Many types of data can be mapped to real persons and bear potential privacy risks. Smart-meter data (see Scenario 1) is probably the most common example, but other types of data such as participation in demand-response measures (see Scenario 3), user behavior in a smart home (see Scenario 4) or from an electric vehicle might disclose private data as well.

As privacy is a wide field, we concentrate on illustrating the possibilities of analyzing smart-meter data in the following. Depending on the temporal resolution of such data, different user behaviors can be derived. Having smart-meter data at one-minute granularity for example enables identifying most electric devices in a typical household [113]. Having data at half-second granularity might reveal whether a cutting machine was used to cut bread or salami [21]. Needless to say, disclosing such data would be a severe privacy risk as one could derive precisely what a person does in which moment. Furthermore, recent research suggests that it is even possible to identify which TV program (out of a number of known programs) someone is watching using a standard smart meter at the temporal granularity of half seconds [58]. Interestingly – and maybe frighteningly – even data at the 15-minute granularity frequently used for billing scenarios can be a privacy threat. Such data is sufficient to identify which persons are at home and at what times, if they prepare cold or warm breakfast, when they are cooking and when they watch TV or go to bed [96]. The authors of [69] furthermore show that consumption curves of a household are typically unique and can be used to identify a household.

There is a bunch of work which identifies the different scenarios of privacy risks and attacks in the field of energy – an overview can be found, e.g., in [77]. A smaller number of studies propose particular solutions, mostly for specific problems such as billing in the presence of smart meters [68]. However, this field is still quite young, and effective methods to provide privacy protection are still needed, which can easily be applied in the field. Besides privacy of consumers, such methods need to ensure that all actors in the energy market can obtain the data they need in order to efficiently fulfill their respective role in the current and future energy scenarios. This calls for further developments and new techniques in the fields of *security research* and *privacy-preserving data mining* [13, 131, 133] for which future energy systems and markets are an important field of application.

7.5 Conclusions

The traditional energy system relying on fossil and nuclear sources is not sustainable. The ongoing transformation to a more sustainable energy system relying on renewable sources leads to major challenges and to a paradigm shift from *demand-driven generation* to *generation-driven demand*. Further influential factors in the ongoing development are liberalization and the effects of new loads such as electric vehicles. These developments in the future energy domain will be facilitated by a number of techniques which are frequently referred to as the *smart grid*. Most of these techniques and scenarios lead to new sources of data and to the challenge to manage and analyze them in appropriate ways.

In this chapter we have highlighted the current developments towards a sustainable energy system, we have given an overview on the current energy markets, and we have described a number of future energy scenarios. Based on these elaborations, we have derived the data-analysis challenges in detail. In a nutshell, the conclusion is that there has been a lot of research but that there are still many unsolved problems and the need for more data-analysis research. Existing techniques can be applied or need to be further developed to be used in the smart grid. Thus, the future energy domain is an important field for applied data-analysis research and has the potential to contribute to a sustainable development.

Acknowledgements. We thank Pavel Efros for his help and Anke Weidlich and many colleagues at SAP Research and acteno energy for fruitful discussions and proof-reading (parts of) the chapter.

Bibliography

[1] Directive 2009/72/EC of the European Parliament and of the Council of 13 July 2009 Concerning Common Rule for the Internal Market in Electricity. *Official Journal of the European Union*, L 211:56–93, 2009.

[2] E-Energy Glossary. Website of the DKE – Deutsche Kommission Elektrotechnik Elektronik Informationstechnik im DIN und VDE, Germany: `https://teamwork.dke.de/specials/7/Wiki_EN/WikiPages/Home.aspx`, 2010.

[3] Energy Concept for an Environmentally Sound, Reliable and Affordable Energy Supply. Publication of the German Federal Ministry of Economics and Technology and the Federal Ministry for the Environment, Nature Conservation and Nuclear Safety, September 2010.

[4] MeRegio – Project Phase 2. Homepage of the MeRegio project: `http://www.meregio.de/en/?page=solution-phasetwo`, 2010.

[5] Annual Report 2010. Publication of the German Federal Motor Transport Authority, 2011.

[6] connecting markets. EEX Company and Products brochure, European Energy Exchange AG, October 2011.

[7] Federal Environment Minister Röttgen: 20 Percent Renewable Energies are a Great Success. Press Release 108/11 of the German Federal Ministry for the Environment, Nature Conservation and Nuclear Safety, August 2011.

[8] Metered Half-Hourly Electricity Demands. Website of National Grid, UK: `http://www.nationalgrid.com/uk/Electricity/Data/Demand+Data/`, 2011.

[9] Charu C. Aggarwal. On Abnormality Detection in Spuriously Populated Data Streams. In *Int. Conf. on Data Mining (SDM)*, 2005.

[10] Charu C. Aggarwal, editor. *Data Streams: Models and Algorithms*, volume 31 of *Advances in Database Systems*. Springer, 2007.

[11] Charu C. Aggarwal. On High Dimensional Projected Clustering of Uncertain Data Streams. In *Int. Conf. on Data Engineering (ICDE)*, 2009.

[12] Charu C. Aggarwal and Philip S. Yu. Outlier Detection for High Dimensional Data. In *Int. Conf. on Management of Data (SIGMOD)*, 2001.

[13] Charu C. Aggarwal and Philip S. Yu, editors. *Privacy-Preserving Data Mining: Models and Algorithms*, volume 34 of *Advances in Database Systems*. Springer, 2008.

[14] Sanjeev Kumar Aggarwal, Lalit Mohan Saini, and Ashwani Kumar. Electricity Price Forecasting in Deregulated Markets: A Review and Evaluation. *International Journal of Electrical Power and Energy Systems*, 31(1):13–22, 2009.

[15] Rakesh Agrawal, Johannes Gehrke, Dimitrios Gunopulos, and Prabhakar Raghavan. Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications. In *Int. Conf. on Management of Data (SIGMOD)*, 1998.

[16] Hesham K. Alfares and Mohammad Nazeeruddin. Electric Load Forecasting: Literature Survey and Classification of Methods. *International Journal of Systems Science*, 33(1):23–34, 2002.

[17] Florian Allerding and Hartmut Schmeck. Organic Smart Home: Architecture for Energy Management in Intelligent Buildings. In *Workshop on Organic Computing (OC)*, 2011.

[18] Fabrizio Angiulli, Fabio Fassetti, and Luigi Palopoli. Detecting Outlying Properties of Exceptional Objects. *ACM Transactions on Database Systems*, 34(1):1–62, 2009.

[19] Mariá Arenas-Martíandnez, Sergio Herrero-Lopez, Abel Sanchez, John R. Williams, Paul Roth, Paul Hofmann, and Alexander. Zeier. A Comparative Study of Data Storage and Processing Architectures for the Smart Grid. In *Int. Conf. on Smart Grid Communications (SmartGridComm)*, 2010.

[20] Thanasis G. Barbounis, John B. Theocharis, Minas C. Alexiadis, and Petros S. Dokopoulos. Long-TermWind Speed and Power Forecasting Using Local Recurrent Neural Network Models. *Energy Conversion, IEEE Transactions on*, 21(1):273–284, 2006.

[21] Gerald Bauer, Karl Stockinger, and Paul Lukowicz. Recognizing the Use-Mode of Kitchen Appliances from Their Current Consumption. In *Smart Sensing and Context*, 2009.

[22] Stephen D. Bay and Michael J. Pazzani. Detecting Group Differences: Mining Contrast Sets. *Data Mining and Knowledge Discovery*, 5(3):213–246, 2001.

[23] Birger Becker, Florian Allerding, Ulrich Reiner, Mattias Kahl, Urban Richter, Daniel Pathmaperuma, Hartmut Schmeck, and Thomas Leibfried. Decentralized Energy-Management to Control Smart-Home Architectures. In *Architecture of Computing Systems (ARCS)*, 2010.

[24] Michael R. Berthold, Christian Borgelt, Frank Höppner, and Frank Klawonn. *Guide to Intelligent Data Analysis: How to Intelligently Make Sense of Real Data*, volume 42 of *Texts in Computer Science*. Springer, 2010.

[25] Kevin Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. When Is Nearest Neighbors Meaningful. In *Int. Conf. on Database Theory (ICDT)*, 1999.

[26] Christophe Bisciglia. The Smart Grid: Hadoop at the Tennessee Valley Authority (TVA). Blog of Cloudera, Inc., USA: `http://www.cloudera.com/blog/2009/06/smart-grid-hadoop-tennessee-valley-authority-tva/`, 2009.

[27] Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. LOF: Identifying Density-Based Local Outliers. In *Int. Conf. on Management of Data (SIGMOD)*, 2000.

[28] Longbing Cao, Philip S. Yu, Chengqi Zhang, and Yanchang Zhao. *Domain Driven Data Mining*. Springer, 2010.

[29] Longbing Cao, Huaifeng Zhang, Yanchang Zhao, Dan Luo, and Chengqi Zhang. Combined Mining: Discovering Informative Knowledge in Complex Data. *IEEE Transactions on Systems, Man, and Cybernetics*, 41(3):699–712, 2011.

[30] Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rüdiger Wirth. CRISP-DM 1.0. Step-by-step data mining guide, SPSS, August 2000.

[31] Feng Chen, Jing Dai, Bingsheng Wang, Sambit Sahu, Milind Naphade, and Chang-Tien Lu. Activity Analysis Based on Low Sample Rate Smart Meters. In *Int. Conf. on Knowledge Discovery and Data Mining (KDD)*, 2011.

[32] Jiyi Chen, Wenyuan Li, Adriel Lau, Jiguo Cao, and Ke Wang. Automated Load Curve Data Cleansing in Power Systems. *IEEE Transactions on Smart Grid*, 1(2):213–221, 2010.

[33] Chun-Hung Cheng, Ada Waichee Fu, and Yi Zhang. Entropy-based Subspace Clustering for Mining Numerical Data. In *Int. Conf. on Knowledge Discovery and Data Mining (KDD)*, 1999.

[34] Robson Leonardo Ferreira Cordeiro, Agma J. M. Traina, Christos Faloutsos, and Caetano Traina Jr. Finding Clusters in Subspaces of Very Large, Multi-Dimensional Datasets. In *Int. Conf. on Data Engineering (ICDE)*, 2010.

[35] Marco Dalai and Riccardo Leonardi. Approximations of One-Dimensional Digital Signals Under the l^∞ Norm. *IEEE Transactions on Signal Processing*, 54(8):3111–3124, 2006.

[36] Lars Dannecker, Matthias Böhm, Ulrike Fischer, Frank Rosenthal, Gregor Hackenbroich, and Wolfgang Lehner. State-of-the-Art Report on Forecasting – A Survey of Forecast Models for Energy Demand and Supply. Public Deliverable D4.1, The MIRACLE Consortium (European Commission Project Reference: 248195), Dresden, Germany, June 2010.

[37] Lars Dannecker, Matthias Böhm, Wolfgang Lehner, and Gregor Hackenbroich. Forecasting Evolving Time Series of Energy Demand and Supply. In *East-European Conf. on Advances in Databases and Information Systems (ADBIS)*, 2011.

[38] Lars Dannecker, Matthias Schulze, Robert andBöhm, Wolfgang Lehner, and Gregor Hackenbroich. Context-Aware Parameter Estimation for Forecast Models in the Energy Domain. In *Int. Conf. on Scientific and Statistical Database Management (SSDBM)*, 2011.

[39] Sarah Darby. The Effectiveness of Feedback on Energy Consumption: A Review for DEFRA of the Literature on Metering, Billing and Direct Displays. Technical report, Environmental Change Institute, University of Oxford, UK, April 2006.

[40] Timothy de Vries, Sanjay Chawla, and Michael E. Houle. Finding Local Anomalies in Very High Dimensional Space. In *Int. Conf. on Data Mining (ICDM)*, 2010.

[41] Jeffrey Dean and Sanjay Ghemawat. MapReduce: Simplified Data Processing on Large Clusters. In *Symposium on Operating Systems Design and Implementation (OSDI)*, 2004.

[42] Kalyanmoy Deb, Samir Agrawal, Amrit Pratap, and T Meyarivan. A Fast Elitist Non-dominated Sorting Genetic Algorithm for Multi-objective Optimization: NSGA-II. In *Int. Conf. on Parallel Problem Solving from Nature (PPSN)*, 2000.

[43] Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.

[44] Guozhu Dong and Jinyan Li. Efficient Mining of Emerging Patterns: Discovering Trends and Differences. In *Int. Conf. on Knowledge Discovery and Data Mining (KDD)*, 1999.

[45] Karen Ehrhardt-Martinez, Kat A. Donnelly, and John A. “Skip” Laitner. Advanced Metering Initiatives and Residential Feedback Programs: A Meta-Review for Household Electricity-Saving Opportunities. Technical Report E105, American Council for an Energy-Efficient Economy, Washington, USA, June 2010.

[46] Frank Eichinger, Detlef D. Nauck, and Frank Klawonn. Sequence Mining for Customer Behaviour Predictions in Telecommunications. In *Workshop on Practical Data Mining: Applications, Experiences and Challenges*, 2006.

[47] Hazem Elmeleegy, Ahmed K. Elmagarmid, Emmanuel Cecchet, Walid G. Aref, and Willy Zwaenepoel. Online Piece-wise Linear Approximation of Numerical Streams with Precision Guarantees. In *Int. Conf. on Very Large Data Bases (VLDB)*, 2009.

[48] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases. In *Int. Conf. on Knowledge Discovery and Data Mining (KDD)*, 1996.

[49] Linda Farinaccio and Radu Zmeureanu. Using a Pattern Recognition Approach to Disaggregate the Total Electricity Consumption in a House into the Major End-Uses. *Energy and Buildings*, 30(3):245–259, 1999.

[50] Ahmad Faruqui and Jennifer Palmer. Dynamic Pricing and Its Discontents. *Regulation Magazine*, 34(3):16–22, 2011.

[51] Michael C. Ferris and Todd S. Munson. Complementarity Problems in GAMS and the PATH Solver. *Journal of Economic Dynamics and Control*, 24(2):165–188, 2000.

[52] Eugene Fink and Kevin B. Pratt. Indexing of Compressed Time Series. In Last et al. [86], chapter 3, pages 51–78.

[53] Jon Froehlich, Eric Larson, Sidhant Gupta, Gabe Cohn, Matthew S. Reynolds, and Shwetak N. Patel. Disaggregated End-Use Energy Sensing for the Smart Grid. *Pervasive Computing*, 10(1):28–39, 2011.

[54] Drew Fudenberg and Jean Tirole. *Game Theory.* MIT Press, 1991.

[55] Hugo Gonçalves, Adrian Oceanu, and Mario Bergés. Unsupervised Disaggregation of Appliances Using Aggregated Consumption Data. In *Workshop on Data Mining Applications in Sustainability (SustKDD)*, 2011.

[56] Carlos J. Alonso González and Juan J. Rodríguez Diez. Boosting Interval-based Literals: Variable Length and Early Classification. In Last et al. [86], chapter 7, pages 149–171.

[57] Jessica Granderson, Mary Piette, and Girish Ghatikar. Building Energy Information Systems: User Case Studies. *Energy Efficiency*, 4:17–30, 2011.

[58] Ulrich Greveler, Benjamin Justus, and Dennis Löhr. Multimedia Content Identification Through Smart Meter Power Usage Profiles. In *Int. Conf. on Computers, Privacy and Data Protection (CPDP)*, 2012.

[59] Stephan Günnemann, Hardy Kremer, Charlotte Laufkötter, and Thomas Seidl. Tracing Evolving Subspace Clusters in Temporal Climate Data. *Data Mining and Knowledge Discovery*, 24(2):387–410, 2011.

[60] Duy Long Ha, Minh Hoang Le, and Stéphane Ploix. An Approach for Home Load Energy Management Problem in Uncertain Context. In *Int. Conf. on Industrial Engineering and Engineering Management (IEEM)*, 2008.

[61] Jiawei Han, Micheline Kamber, and Jian Pei. *Data Mining: Concepts and Techniques.* Morgan Kaufmann, 2011.

[62] Jiawei Han, Jian Pei, and Xifeng Yan. Sequential Pattern Mining by Pattern-Growth: Principles and Extensions. In W. Chu and T. Lin, editors, *Studies in Fuzziness and Soft Computing*, volume 180 of *Foundations and Advances in Data Mining*, pages 183–220. Springer, 2005.

[63] George W. Hart. Nonintrusive Appliance Load Monitoring. *Proceedings of the IEEE*, 80(12):1870–1891, 1992.

[64] Magnus L. Hetland. A Survey of Recent Methods for Efficient Retrieval of Similar Time Sequences. In Last et al. [86], chapter 2, pages 27–49.

[65] Alexander Hinneburg and Daniel Keim. An Efficient Approach to Clustering in Large Multimedia Databases with Noise. In *Int. Conf. on Knowledge Discovery and Data Mining (KDD)*, 1998.

[66] Henrique Steinherz Hippert, Carlos Eduardo Pedreira, and Reinaldo Castro Souza. Neural Networks for Short-Term Load Forecasting: A Review and Evaluation. *IEEE Transactions on Power Systems*, 16(1):44–55, 2001.

[67] Vikramaditya Jakkula and Diane Cook. Outlier Detection in Smart Environment Structured Power Datasets. In *Int. Conf. on Intelligent Environments (IE)*, 2010.

[68] Marek Jawurek, Martin Johns, and Florian Kerschbaum. Plug-In Privacy for Smart Metering Billing. In *Int. Symposium on Privacy Enhancing Technologies (PETS)*, 2011.

[69] Marek Jawurek, Martin Johns, and Konrad Rieck. Smart Metering De-Pseudonymization. In *Annual Computer Security Applications Conference (ACSAC)*, 2011.

[70] Andrej Jokić, Mircea Lazar, and Paul P. J. van den Bosch. Price-based Control of Electrical Power Systems. In Negenborn et al. [105], chapter 5, pages 109–131.

[71] Ian Joliffe. *Principal Component Analysis.* Springer, New York, 1986.

[72] Karin Kailing, Hans-Peter Kriegel, and Peer Kröger. Density-Connected Subspace Clustering for High-Dimensional Data. In *Int. Conf. on Data Mining (SDM)*, 2004.

[73] Andreas Kamper. *Dezentrales Lastmanagement zum Ausgleich kurzfristiger Abweichungen im Stromnetz.* KIT Scientific Publishing, 2009.

[74] Fabian Keller, Emmanuel Müller, and Klemens Böhm. HiCS: High Contrast Subspaces for Density-Based Outlier Ranking. In *Int. Conf. on Data Engineering (ICDE)*, 2012.

[75] James Kennedy and Russel Eberhart. Particle Swarm Optimization. In *Int. Conf. on Neural Networks*, 1995.

[76] Eamonn Keogh and Shruti Kasetty. On the Need for Time Series Data Mining Benchmarks: A Survey and Empirical Demonstration. *Data Mining and Knowledge Discovery*, 7(4):349–371, 2003.

[77] Himanshu Khurana, Mark Hadley, Ning Lu, and Deborah A. Frincke. Smart-Grid Security Issues. *IEEE Security and Privacy*, 8(1):81–85, 2010.

[78] Hyungsul Kim, Manish Marwah, Martin F. Arlitt, Geoff Lyon, and Jiawei Han. Unsupervised Disaggregation of Low Frequency Power Measurements. In *Int. Conf. on Data Mining (SDM)*, 2011.

[79] Edwin M. Knorr and Raymond T. Ng. Finding Intensional Knowledge of Distance-Based Outliers. In *Int. Conf. on Very Large Data Bases (VLDB)*, 1999.

[80] Edwin M. Knox and Raymond T. Ng. Algorithms for Mining Distance-Based Outliers in Large Datasets. In *Int. Conf. on Very Large Data Bases (VLDB)*, 1998.

[81] Koen Kok, Martin Scheepers, and René Kamphuis. Intelligence in Electricity Networks for Embedding Renewables and Distributed Generation. In Negenborn et al. [105], chapter 8, pages 179–209.

[82] J. Zico Kolter and Tommi Jaakkola. Approximate Inference in Additive Factorial HMMs with Application to Energy Disaggregation. In *Int. Conf. on Artificial Intelligence and Statistics (AISTATS)*, 2012.

[83] J. Zico Kolter and Matthew Johnson. REDD: A Public Data Set for Energy Disaggregation Research. In *Workshop on Data Mining Applications in Sustainability (SustKDD)*, 2011.

[84] Andrew Kusiak, Haiyang Zheng, and Zhe Song. Short-Term Prediction of Wind Farm Power: A Data Mining Approach. *IEEE Transactions on Energy Conversion*, 24(1):125–136, 2009.

[85] National Energy Technology Laboratory. A Vision for the Modern Grid. In *Smart Grid*, chapter 11, pages 283–293. The Capitol Net Inc., 2007.

[86] Mark Last, Abraham Kandel, and Horst Bunke, editors. *Data Mining in Time Series Databases*, volume 57 of *Series in Machine Perception and Artificial Intelligence.* World Scientific, 2004.

[87] Christopher Laughman, Kwangduk Lee, Robert Cox, Steven Shaw, Steven Leeb, Les Norford, and Peter Armstrong. Power Signature Analysis. *Power and Energy Magazine*, 1(2):56–63, 2003.

[88] Yiu-Wing Leung and Yuping Wang. An Orthogonal Genetic Algorithm with Quantization for Global Numerical Optimization. *IEEE Transactions on Evolutionary Computation*, 5(1):41–53, 2001.

[89] Shuhui Li, Donald C. Wunsch, Edgar O'Hair, and Michael G. Giesselmann. Comparative Analysis of Regression and Artificial Neural Network Models for Wind Turbine Power Curve Estimation. *Journal of Solar Energy Engineering*, 123(4):327–332, 2001.

[90] Xiaoli Li, Chris P. Bowers, and Thorsten Schnier. Classification of Energy Consumption in Buildings With Outlier Detection. *IEEE Transactions on Industrial Electronics*, 57(11):3639–3644, 2010.

[91] Jian Liang, Simon K.K. Ng, Gail Kendall, and John W.M. Cheng. Load Signature Study – Part I: Basic Concept, Structure, and Methodology. *IEEE Transactions on Power Delivery*, 25(2):551–560, 2010.

[92] J. MacQueen. Some Methods for Classification and Analysis of Multivariate Observations. In *Berkeley Symposium on Mathematical Statistics and Probability*, 1967.

[93] Paras Mandal, Tomonobu Senjyu, Naomitsu Urasaki, and Toshihisa Funabashi. A Neural Network based Several-Hour-Ahead Electric Load Forecasting using Similar Days Approach. *International Journal of Electrical Power and Energy Systems*, 28(6):367–373, 2006.

[94] Friedemann Mattern, Thorsten Staake, and Markus Weiss. ICT for Green: How Computers Can Help Us to Conserve Energy. In *Int. Conf. on Energy-Efficient Computing and Networking (E-Energy)*, 2010.

[95] Tom Mitchell. *Machine Learning.* McGraw Hill, 1997.

[96] Andrés Molina-Markham, Prashant Shenoy, Kevin Fu, Emmanuel Cecchet, and David Irwin. Private Memoirs of a Smart Meter. In *Workshop on Embedded Sensing Systems for Energy-Efficiency in Building (BuildSys)*, 2010.

[97] Sanaz Mostaghim and Jürgen Teich. Strategies for finding good local guides in multi-objective particle swarm optimization (MOPSO). In *Swarm Intelligence Symposium (SIS)*, 2003.

[98] Emmanuel Müller, Ira Assent, Stephan Günnemann, and Thomas Seidl. Scalable Density-Based Subspace Clustering. In *Int. Conf. on Information and Knowledge Management (CIKM)*, 2011.

[99] Emmanuel Müller, Ira Assent, Ralph Krieger, Stephan Günnemann, and Thomas Seidl. DensEst: Density Estimation for Data Mining in High Dimensional Spaces. In *Int. Conf. on Data Mining (SDM)*, 2009.

[100] Emmanuel Müller, Ira Assent, and Thomas Seidl. HSM: Heterogeneous Subspace Mining in High Dimensional Data. In *Scientific and Statistical Database Management (SSDBM)*, 2009.

[101] Emmanuel Müller, Stephan Günnemann, Ira Assent, and Thomas Seidl. Evaluating Clustering in Subspace Projections of High Dimensional Data. In *Int. Conf. on Very Large Data Bases (VLDB)*, 2009.

[102] Emmanuel Müller, Stephan Günnemann, Ines Färber, and Thomas Seidl. Discovering Multiple Clustering Solutions: Grouping Objects in Different Views of the Data. In *Int. Conf. on Data Mining (ICDM)*, 2010.

[103] Emmanuel Müller, Matthias Schiffer, and Thomas Seidl. Statistical Selection of Relevant Subspace Projections for Outlier Ranking. In *Int. Conf. on Data Engineering (ICDE)*, 2011.

[104] Daniel Müller-Jentsch. The Development of Electricity Markets in the Euro-Mediterranean Area: Trends and Prospects for Liberalization and Regional Intergration. Technical Paper 491, The World Bank, Washington, USA, 2001.

[105] Rudy R. Negenborn, Zofia Lukszo, and Hans Hellendoorn, editors. *Intelligent Infrastructures*, volume 42 of *Intelligent Systems, Control and Automation: Science and Engineering.* Springer, 2010.

[106] Andrew Ng, Michael Jordan, and Yair Weiss. On Spectral Clustering: Analysis and an Algorithm. In *Conf. on Neural Information Processing Systems (NIPS)*, 2001.

[107] Donglin Niu, Jennifer G. Dy, and Michael I. Jordan. Multiple Non-Redundant Spectral Clustering Views. In *Int. Conf. on Machine Learning (ICML)*, 2010.

[108] Anisah H. Nizar, Zhao Y. Dong, and J.H. Zhao. Load Profiling and Data Mining Techniques in Electricity Deregulated Market. In *Power Engineering Society General Meeting*, 2006.

[109] Alexandra-Gwyn Paetz, Birger Becker, Wolf Fichtner, and Hartmut Schmeck. Shifting Electricity Demand with Smart Home Technologies - An Experimental Study on User Acceptance. In *USAEE/IAEE North American Conference*, 2011.

[110] Peter Palensky and Dietmar Dietrich. Demand Side Management: Demand Response, Intelligent Energy Systems, and Smart Loads. *IEEE Transactions on Industrial Informatics*, 7(3):381–388, 2011.

[111] Peter Palensky, Dietrich Dietrich, Ratko Posta, and Heinrich Reiter. Demand Side Management in Private Homes by Using LonWorks. In *Int. Workshop on Factory Communication Systems*, 1997.

[112] Hasso Plattner and Alexander Zeier. *In-Memory Data Management – An Inflection Point for Enterprise Applications.* Springer, 2011.

[113] Elias Leake Quinn. Smart Metering and Privacy: Existing Laws and Competing Policies. Report for the colorado public utilities commission, University Colorado Law School (CEES), Boulder, USA, May 2009.

[114] Thanawin Rakthanmanon, Eamonn Keogh, Stefano Lonardi, and Scott Evans. Time Series Epenthesis: Clustering Time Series Streams Requires Ignoring Some Data. In *Int. Conf. on Data Mining (ICDM)*, 2011.

[115] Maryam Ramezani, Mario Graf, and Harald Vogt. A Simulation Environment for Smart Charging of Electric Vehicles Using a Multi-objective Evolutionary Algorithm. In *Int. Conf. on Information and Communication on Technology for the Fight against Global Warming (ICT-GLOW)*, 2011.

[116] Sérgio Ramos and Zita Vale. Data Mining Techniques Application in Power Distribution Utilities. In *Transmission and Distribution Conference and Exposition*, 2008.

[117] Sira Panduranga Rao and Diane J. Cook. Identifying Tasks and Predicting Actions in Smart Homes using Unlabeled Data. In *The Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining*, 2003.

[118] Ulrich Reiner, Thomas Leibfried, Florian Allerding, and Hartmut Schmeck. Potential of Electrical Vehicles with Feed-back Capabilities and Controllable Loads in Electrical Grids under the Use of Decentralized Energy Management. In *Int. ETG Congress*, 2009.

[119] Peter J. Rousseeuw and Annick M. Leroy. *Robust Regression and Outlier Detection*. Wiley, 1987.

[120] Sebnem Rusitschka, Kolja Eger, and Christoph Gerdes. Smart Grid Data Cloud: A Model for Utilizing Cloud Computing in the Smart Grid Domain. In *Int. Conf. on Smart Grid Communications (SmartGridComm)*, 2010.

[121] Sunita Sarawagi. User-Adaptive Exploration of Multidimensional Data. In *Int. Conf. on Very Large Data Bases (VLDB)*, 2000.

[122] Sunita Sarawagi, Rakesh Agrawal, and Nimrod Megiddo. Discovery-Driven Exploration of OLAP Data Cubes. In *Int. Conf. on Extending Database Technology (EDBT)*, 1998.

[123] Gayatri Sathe and Sunita Sarawagi. Intelligent Rollups in Multidimensional OLAP Data. In *Int. Conf. on Very Large Data Bases (VLDB)*, 2001.

[124] Domnic Savio, Lubomir Karlik, and Stamatis Karnouskos. Predicting Energy Measurements of Service-Enabled Devices in the Future Smartgrid. In *Int. Conf. on Computer Modeling and Simulation (UKSim)*, 2010.

[125] Matthieu-P. Schapranow, Ralph Kühne, Alexander Zeier, and Hasso Plattner. Enabling Real-Time Charging for Smart Grid Infrastructures using In-Memory Databases. In *Workshop on Smart Grid Networking Infrastructure*, 2010.

[126] Joachim Schleich, Marian Klobasa, Marc Brunner, Sebastian Gölz, and Konrad Götz. Smart Metering in Germany and Austria: Results of Providing Feedback Information in a Field Trial. Working Paper Sustainability and Innovation S 6/2011, Fraunhofer Institute for Systems and Innovation Research (ISI), Karlsruhe, Germany, 2011.

[127] Raimund Seidel. Small-Dimensional Linear Programming and Convex Hulls Made Easy. *Discrete & Computational Geometry*, 6(1):423–434, 1991.

[128] Kelvin Sim, Vivekanand Gopalkrishnan, Arthur Zimek, and Gao Cong. A Survey on Enhanced Subspace Clustering. *Data Mining and Knowledge Discovery*, 2012.

[129] Myra Spiliopoulou, Irene Ntoutsi, Yannis Theodoridis, and Rene Schult. MONIC: Modeling and Monitoring Cluster Transitions. In *Int. Conf. on Knowledge Discovery and Data Mining (KDD)*, 2006.

[130] Asher Tishler and Israel Zang. A Min-Max Algorithm for Non-linear Regression Models. *Applied Mathematics and Computation*, 13(1/2):95–115, 1983.

[131] Jaideep Vaidya, Yu Zhu, and Christopher W. Clifton. *Privacy Preserving Data Mining*, volume 19 of *Advances in Information Security*. Springer, 2006.

[132] Sergio Valero Verdú, Mario Ortiz García, Carolina Senabre, Antonio Gabaldón Marín, and Francisco J. García 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.

[133] Vassilios S. Verykios, Elisa Bertino, Igor Nai Fovino, Loredana Parasiliti Provenza, Yucel Saygin, and Yannis Theodoridis. State-of-the-Art in Privacy Preserving Data Mining. *SIGMOD Record*, 33(1):50–57, 2004.

[134] Michail Vlachos, Dimitrios Gunopulos, and Gautam Das. Indexing Time-Series under Conditions of Noise. In Last et al. [86], chapter 4, pages 67–100.

[135] Harald Vogt, Holger Weiss, Patrik Spiess, and Achim P. Karduck. Market-Based Prosumer Participation in the Smart Grid. In *Int. Conf. on Digital Ecosystems and Technologies (DEST)*, 2010.

[136] Horst F. Wedde, Sebastian Lehnhoff, Christian Rehtanz, and Olav Krause. Bottom-Up Self-Organization of Unpredictable Demand and Supply under Decentralized Power Management. In *Int. Conf. on Self-Adaptive and Self-Organizing Systems*, 2008.

[137] Anke Weidlich. *Engineering Interrelated Electricity Markets*. Physica Verlag, 2008.

[138] Anke Weidlich, Harald Vogt, Wolfgang Krauss, Patrik Spiess, Marek Jawurek, Martin Johns, and Stamatis Karnouskos. Decentralized Intelligence in Energy Efficient Power Systems. In Alexey Sorokin, Steffen Rebennack, Panos M. Pardalos, Niko A. Iliadis, and Mario V.F. Pereira, editors, *Handbook of Networks in Power Systems: Optimization, Modeling, Simulation and Economic Aspects*. Springer, 2012.

[139] Tom White. *Hadoop: The Definitive Guide*. O'Reilly, 2009.

[140] Ian H. Witten, Eibe Frank, and Mark A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2011.

[141] G. Michael Youngblood and Diane J. Cook. Data Mining for Hierarchical Model Creation. *IEEE Transactions on Systems, Man, and Cybernetics*, 37(4):561–572, 2007.

[142] Michael Zeifman and Kurt Roth. Nonintrusive Appliance Load Monitoring: Review and Outlook. *Transactions on Consumer Electronics*, 57(1):76–84, 2011.

[143] Roberto V. Zicari. Big Data: Smart Meters – Interview with Markus Gerdes. ODBMS Industry Watch Blog: `http://www.odbms.org/blog/2012/06/big-data-smart-meters-interview-with-markus-gerdes/`, 2012.