Principles of Green Data Mining

Johannes Schneider
University of Liechtenstein
johannes.schneider@uni.li

Marcus Basalla
University of Liechtenstein
marcus.basalla@uni.li

Stefan Seidel
University of Liechtenstein
stefan.seidel@uni.li

Abstract

This paper develops a set of principles for green data mining, related to the key stages of business understanding, data understanding, data preparation, modeling, evaluation, and deployment. The principles are grounded in a review of the Cross Industry Standard Process for Data mining (CRISP-DM) model and relevant literature on data mining methods and Green IT. We describe how data scientists can contribute to designing environmentally friendly data mining processes, for instance, by using green energy, choosing between make-or-buy, exploiting approaches to data reduction based on business understanding or pure statistics, or choosing energy friendly models.

1. Introduction

The use of computing power coupled with the unprecedented availability of data provide ample opportunity to improve energy efficiency [18]. However, they are also an increasingly relevant source of energy consumption and associated carbon emissions. Data centers consumed about 70 billion kWh in 2016 in the United States alone [50], and the total consumption of all IT is estimated to be close to 5% of total energy consumption [18]. In response to this increasing amount of energy used by IT, Greenpeace published the "Guide to Building the Green Internet" [10], promoting "a more widespread adaption in best practices" for energy efficient data center design. They demand that "data center operators and customers should regularly report their energy performance and establish transparent energy savings targets." Electricity consumption is costly—it involves various detrimental effects on nature and society, ranging from bird deaths by wind turbines, on to severe air pollution and CO2 emissions by coal power plants, and the risk of catastrophes stemming from nuclear power plants.

These concerns are partially addressed by current initiatives under notions such as green information systems (Green IS) or green information technology (Green IT) [34, 57], but environmentally friendly data mining is a novel topic.

Data scientists often leverage a large pool of computational resources using sophisticated and computationally costly machine learning techniques to extract knowledge and insights from data. Though existing processes such as the Cross Industry Standard Process for Data mining (CRISP-DM) [61] provide some guidance on how to execute a data mining project, the skills of a data scientist heavily rely on creativity [53], involving many degrees of freedom, often including the choice of tools, models, and data sources.

It is against this background that, in this paper, we develop guidelines for data scientists to implement more environmentally friendly practices that can complement technology-focused perspectives aiming to design more energy efficient IT-based systems. Specifically, we are focusing attention on one important area of data science—data mining. Data mining can be described as knowledge discovery from data [23] or in terms of different activities as collecting, cleaning, processing, analyzing and gaining useful insights from data [2]. We ask: *How can data scientists implement more environmentally friendly data mining processes?*

The remainder of this paper is structured as follows. We first describe our methodology. We then review the data mining process and develop a set of principles for green data mining. We conclude by discussing limitations and future work.

2. Methodology

We derived our principles by analyzing the CRISP-DM data mining process and literature on green IT and data mining. In a first step, we identified factors determining energy consumption. In a second step, we identified individual steps of the CRISP-DM process by investigating possibilities for reduction of each factor. We limited our analysis to those aspects that can be directly influenced by data scientists, including the choice of data, its representation, as well as processes and techniques used throughout the data

URI: https://hdl.handle.net/10125/59646
ISBN: 978-0-9981331-2-6

analysis process. We do not target the development of novel data mining algorithms for specific problems or improving hardware or software, though some of our insights might be helpful in guiding such developments.

We conducted a narrative literature review [25] on green IT, green IS, and data mining because our goal was to investigate elementary factors and research outcomes related to these areas of research. Green data science [59] is a novel field and, therefore, is more amenable to a qualitative approach such as narrative literature review than a more quantitative approach detailing the current-state-of-research, as done for a descriptive review. Our focus was on using established online databases from computer science as well as information systems such as IEEE Xplore, ProQuest (ABI/INForm), ScienceDirect (Elsevier), AIS electronic library and the ACM digital library. We did not limit ourselves to journals since new ideas are often presented first at academic conferences and a significant body of works, in particular in the field of computer science, only appear as conference articles.

3. The data mining process

There are multiple data mining processes [27], most of which share common phases. CRISP-DM [61] is arguably the most widely known and practiced model [41], attending to business and data understanding, data preparation, modelling, evaluation and deployment (Figure 1). The business understanding phase clarifies project objectives and business requirements, which are then translated into a data mining problem. There are unsupervised data mining problems including association pattern mining and clustering as well as supervised approaches like classification [2, 23]. Data understanding typically requires initial data selection or collection. Data is first analyzed in an exploratory fashion to get a basic understanding of the data in the business context. Exploratory analysis supports the development of hypothesis by identifying patterns in the data [3]. It allows to get first insights as well as to identify data quality problems. Data preparation includes using raw data to derive data that can be fed into the models. Activities include data selection, transformation, and cleaning. The data might have to be prepared separately for each model. The modelling phase consists of defining suitable models, selecting a model, and adapting the model, for instance, optimizing its parameters to solve the data mining problem. Computational evaluation of the model is part of the model selection process. Every data mining problem can be tackled using different strategies and models. Generally, there is no clear consensus about which model is best for a task. Consequently, some form of trial and error can often not be avoided. This is supported by the "no free lunch" theorem stating that any algorithm outperforms any other algorithm on some datasets [63] as well as by empirical studies [9, 26]. The choice of models depends on many factors such as data (dimensionality, number of observations, structuredness), data mining objectives (need for best possible expected outcome, need to explain results), and cost (focus on minimum human effort to build or operate). From the perspective of green data mining, performance is assessed in terms of energy consumption for model training and model use, for instance, for making predictions. For the evaluation phase the main goal is to review all steps involved in the construction of the model, and to verify whether the final model meets the defined business objectives. If the best model meets the evaluation criteria, then it is deployed. Deployment ranges from fabricating a report presenting the findings in an easy-to-comprehend manner to implementing a long running system. Such a system might learn continuously while often performing a prediction task.

4. Principles of green data mining

Grounded in concepts and ideas from the literature on Green IT as well as data mining and its processes,

Table 1: Factors and methods related to green data mining

Factor	Subfactors	Methods for Green Data Mining
Project Objectives and Execution	Performance specification; Make, buy, share	Transfer Learning
Data	Quantity; Quality; Representation; Data acquisition method; Data storage	Sampling, Active Learning, Dimensionality Reduction, Compression, Change of Data Representation, Data Aggregation
Computation (Analysis)	Structuring of computation; Choice/Training of models; Training of models	Reuse of intermediate results; Approximate Models/Algorithms
IT Infrastructure	Hardware, e.g., CPU, Storage	

we identified factors determining the ecological footprint of data mining and we developed principles for reducing this footprint (Table 1, Figure 1).

Green IT discusses institutional perspectives [39], the role of users, including their behavior and beliefs when using IT-based systems [38] as well as technical concerns [1,19,24]. Topics include computational methods [1], their implementation in software [8,21], hardware components of computers [24,44], datacenters [39], cloud computing [18,33], parallel data processing (for big data) [19,22,40], as well as organizational and business aspects such as sustainable value chains, green oriented procurement [7], and adoption of Green IT [28]. Loeser et al. [30] discussed constructs and practices from Green IT (and IS) with respect to sourcing, operations, disposal, governance and end products.

Current literature on data mining [2,38,59], in particular data mining processes [27], does not explicitly discuss environmental concerns of data mining but touches upon aspects related to computational efficiency and storage such as data reduction and approximate algorithms.

Next, we describe principles of green data mining related to the different steps of the CRISP-DM process. We first elaborate on those principles that pertain to all stages of the process (principles 1-3 in Figure 1), before we then turn to those which only address specific stages (principles 4-8).

Principle #1: Identify and focus on the most energy consuming phases

To maximize the outcome of time invested into making data mining more environmentally friendly, the focus should be on the most energy consuming factors. This analysis can be performed by investigating the factors listed in Table 1 and analyzing each process step shown in Figure 1. Which process steps and factors dominate energy consumption depends on the goals and particularities of the data mining endeavor. Project objectives such as predictive accuracy or required confidence in the analysis are very likely to have a profound impact on energy consumption, since they often indirectly influence the choice of computational methods and data. For example, recent "deep learning" [20] methods have outperformed other machine learning approaches for multiple classification tasks. A data scientist might turn to deep learning to meet certain project objectives, because it achieves state-of-the-art performance with respect to accuracy but, at the same time, requires lots of data and computation. Data preparation does often only require simple techniques, but it might be dominating in terms of energy consumption if complex computationally expensive methods are needed to extract features from the data that are used in later phases of the process. Deployment might be the dominating step if a system is built for continuous usage with large amounts of data. Still, deployment might contribute very little to the overall energy consumption compared to model selection, if the goal of the data mining project is to derive a report supporting a one-time decision.

Principle #2: Share and re-use data, models, frameworks and skills

A data scientist might control make-or-buy decisions. For example, for marketing purposes, she might choose to acquire data from social media

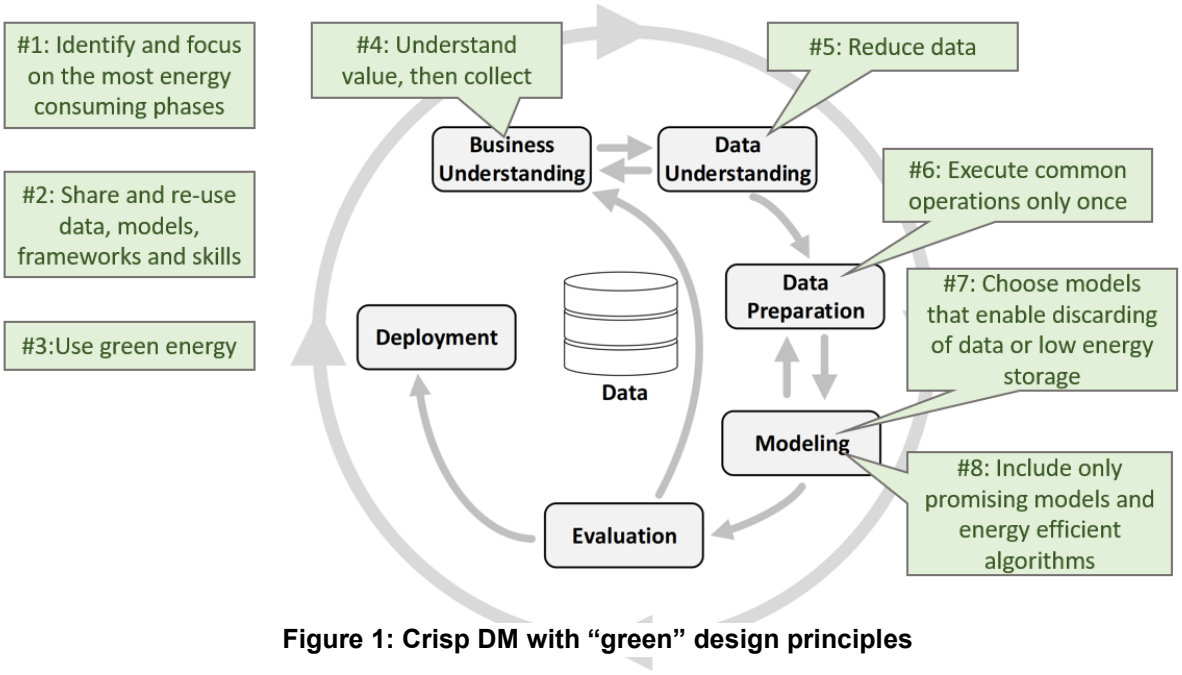

Figure 1: Crisp DM with "green" design principles

channels such as Twitter or Facebook and conduct the analysis by herself. She might also acquire models (implemented in software) to conduct the analysis. She might also decide to consult an external company to conduct the analysis or to obtain models. From an environmental perspective, outsourcing can be preferable if the contractor is more energy-efficient in extracting the demanded information, for instance, because of their prior experience and specialization, more energy efficient infrastructure, or even possession of relevant data. On a global scale, outsourcing of data analysis has the potential to involve less computation and to save energy.

Progress in the field of data science also relies on publicly available data, models, and development frameworks. Initiatives to make data available by research institutions [12] and by governments help create entire ecosystems [11]. State-of-the-art tools to develop (deep learning) models such as Google's Tensorflow are made freely available by large corporations. For such frameworks there are also numerous pre-trained models freely available, e.g., for image recognition based on the Imagenet dataset [12]. Transfer learning is a technique that enables using knowledge from existing models trained for a specific task and dataset on different tasks [42, 31]. The idea is that some "knowledge" of a model can be transferred to another domain. Deep learning networks might benefit from reusing parameters or layers of an already trained network [4, 64] to reduce time (and energy consumption) on developing a new model. Thus, a green data scientist should also contribute data, models, and potentially extensions to frameworks to encourage re-use.

Principle #3: Use green energy

The use of renewable ("green") energy such as solar or wind should be maximized. Conceptually, the idea is to align computation with the availability of green energy. Technical realizations for data processing tasks for distributed data processing platforms (e.g., Hadoop) have been investigated [19]. A system must predict the availability of green energy as well as brown energy and derive a schedule to maximize green energy use and to avoid using brown power at peak demand times. This strategy might also have a positive impact on energy costs as these increase with demand. The data scientist should identify the maximum possible slack in executing data processing tasks based on business objectives. More flexible scheduling allows for using more green energy.

4.1 Business understanding

The business understanding phase does typically not involve computation and as such generally does not contribute directly to the energy consumption. Still, understanding the business requirements and trends in the industry sector helps anticipate factors that influence energy consumption of later process steps, such as "What data are relevant and should be collected?" or "What precision of numbers is needed (over time)?" or "How frequently is a deployed system used?" or "How does the value of data change over time?"

Principle #4: Understand value, then collect and forget

Following the idea that "Data is the new oil"—a statement coined by Clive Humbly in 2006—it seems natural to collect as much data as possible, in particular given that storage is cheap and data might generate value "eventually." It is not uncommon that data can be obtained almost for free, for instance, in the form of trace data generated by users visiting a webpage. But, more data increases costs (due to storage and processing), requires more energy, impacts system performance and complexity and, additionally, enhances the risk of information overload. Query times to a database, for instance, increase with the amount of data stored in the database. The idea of collecting data only for the sake of collection has been criticized–"less data can be more value" [6]. The data scientist should thus try to determine what data is relevant for the business or task at hand [65]. Moreover, the quality of the data should be taken into consideration because data of inferior quality might require non-negligible effort for data cleaning [23].

Not all data has the same value. Even when data consists of a set of observations of the same kind, certain observations might be more valuable than others. For example, for observations, which should be split into classes, "difficult" to classify observations are often more helpful in training data mining models than "easy" to classify observations [56]. Though computational methods can often determine the relevance of data with respect to well-defined metrics, a holistic understanding of the business, its objectives, data, and analytical methodology is essential to limit the collection of data. Leading data analytics companies such as Google embrace the idea of computing on more "little" data, that is, samples [6]. This reasoning is well-founded not only based on statistical models, but also because models benefit from training data in a highly non-linear fashion with decreasing marginal gains given more data [16]. Therefore, in some scenarios, reducing the volume of data might be feasible with considerable impact on energy consumption but only minor changes for other

relevant metrics. Since each model comes with its own strengths and weaknesses related to interpretability, robustness, speed of learning, etc., the overall assessment of advantages and disadvantages must be carefully conducted and aligned with underlying business objectives.

4.2 Data understanding

Principle #5: Reduce data

The data scientist might face the choice of what data to collect (or store). This choice must be made with great foresight in order not to miss any opportunity for data-driven value creation. Business understanding as well as an in depth understanding of the data are necessary. However, there are also multiple helpful techniques based on computational and statistical methods that might be supportive. We describe strategies to minimize the amount of data to be collected or used for training such as sampling and dimensionality reduction. These strategies can be employed to limit the number of attributes or observations, reducing precision and changing the representation of data.

Principle #5.1: Reduce number of data items

Often the data scientist can retrieve accurate results by looking at data samples or by using aggregated data. Data can also be categorized (or clustered) into groups, such that different attributes are relevant for some groups but not for others. A group might also be described using an average or median value. The grouping itself might be obtained by clustering algorithms, for instance, documents can be summarized using centroids obtained through clustering [43]. Intuitively, one should maintain data that is most relevant to achieve a certain task. Active learning [2] seeks to incrementally acquire relevant samples for learning. Thus, rather than having a passive model (or learner) that just uses the training data as given, an active learner might ask explicitly for data that is expected to yield maximal improvement in learning. Active learning is typically used in determining what data to collect. But the idea of active learning might also be used to assess the relevance of data and filter data accordingly. A model can be trained using active learning by incrementally adding the most important data items of the full dataset. The learning process might be stopped if there is no more data that improves the model beyond a small threshold. Unused data, which does not improve the model significantly, could then be discarded. Uncertainty sampling is the most prominent technique in active learning in the context of classification [49]. It seeks to obtain labelled data, where there is most uncertainty about the correct class labels. Uncertainty sampling has been employed successfully for margin-based classifiers such as Support Vector Machines (SVMs) [56]. Standard sampling techniques [52] can also be helpful to reduce the amount of data. One of the simplest, but often sufficient approaches is to conduct simple random sampling—choosing each data point with the same probability without replacement of selected data points. In a case study on predicting conversion probabilities for two online retailers, Stange and Funk [52] could show that only 1% of the data available to them was enough to achieve the optimal tradeoff between accuracy and the cost of collecting and processing the data. Stratified sampling is an appropriate sampling technique if groups are homogeneous, that is, data within groups has lower variance than data from distinct groups. One could also employ density-based sampling, for instance, assign samples with lower density a higher probability. This is useful if data from rare regions is highly important.

Principle #5.2: Reduce number or precision of attributes

The dataset might contain attributes that are irrelevant for the analysis. These attributes can be safely neglected. The relevance might depend on the type of data. For many text mining problems very frequent words—so-called stop words, such as “and”, ”the”, “is”, ”are”—can be ignored. In fact, removing unnecessary or noisy attributes such as stop words is often recommended [2]. More generally, dimensionality reduction can be achieved by feature selection and extraction as well as type transformation [51, 2]. Feature selection techniques encompass filter and wrapper methods as well as their combination. Filter models assess the impact of features by some criterion independent of the model. Wrapper models train the model using a subset of features. An example of a filter model is the use of predictive attribute dependence, where the idea is that correlated features yield better outcomes than uncorrelated ones. Therefore, the relevance of an attribute might be determined by assessing the classification accuracy when using all other attributes to predict the attribute. These techniques can be employed to remove attributes that do not reach a minimum relevance threshold. Since many of the techniques are of heuristic nature, the impact of the removal of data that is deemed irrelevant should be tested, for instance, by comparing models being trained on the full and the reduced attribute set. Attribute reduction can also lead to an increase in accuracy, e.g., for decision trees [59].

Feature extraction is often performed through axis rotations in a way that axes are sorted according to their ability to reconstruct data with minimal error [2].

Axes with negligible impact on data reconstruction can be removed. The derived dataset can often be used to train a model or it might be used to reconstruct the original data, which in turn is used for training. The prior approach is preferable, since a lesser volume of data must be processed. Prominent techniques include singular value decomposition (SVD), and a special case called principal component analysis (PCA).

SVD and similar techniques for feature extraction solve an optimization problem. This can be time consuming, making potential energy savings questionable. Random projections [51], where data is projected onto random manifolds, are a more simple and efficient dimensionality reduction technique. However, to achieve the same approximation guarantees more dimensions are needed than for SVD. Random projections preserve Euclidean distances according to the Johnson-Lindenstrauss Lemma as well as similarity computed using dot products [51], but random projections (as well as other dimensionality reduction techniques) do not preserve metrics such as the Manhattan distance. Therefore, some care is needed to ensure correct outcomes, when applying dimensionality reduction techniques. There is also empirical evidence comparing learning outcomes on the original data to outcomes on the data with reduced dimensionality [17]. Unfortunately, the comparison neglects metrics relevant to energy, e.g., computation time.

Aggarwal [2] describes dimensionality reduction with type transformation as the change of data from a more complex to a less complex type. For instance, graphs can be expressed as multidimensional data that might potentially be easier (and faster) to process. Time series can also be transformed to multi-dimensional data using the Haar Wavelet Transform or Fourier Transformation that both express the data using a (small) set of orthogonal functions. This form of data compression typically implies a loss of precision [46]. Often, a dataset might only contain a few informative attributes and, therefore, the loss of precision might be very small, while achieving a substantial amount of data reduction. A high level understanding of the data mining task helps the data scientist choose a suitable dimensionality reduction technique. A technique might distort some instances more than others, and a small number of instances that are very different in the original context can be very similar in the space with reduced dimensions. For tasks like outlier detection this can be inacceptable, since outliers might be transformed so that they are not identifiable in the transformed data. Other tasks such as segmenting data into unspecified groups (clustering) might be less impacted by altering a few instances in a non-desirable way.

Principle #5.3: Change data representation

Data can be described in many ways without any loss of information, using lossless compression algorithms [46]. This means that data is transformed among different representations without any effect on the minable knowledge. The green data scientist should prefer the representation that requires the least amount of storage, the least amount of computational effort to process throughout the data mining task, and the least amount of computation to create from the original data description.

A sequence of 0,0,0,0,99,99 can be written more compactly as 0:4, 99:2. Another form of encoding is difference encoding, where differences between two elements are stored, e.g., 0,0,0,0,99,0. Difference encoding is often beneficial for time-series data, where commonly there is a strong dependency between consecutive data points. It is also possible to store only non-zero elements with indexes, e.g., the sequence 0,0,0,0,99,99 becomes 4:99, 5:99. In multiple dimensions such data structures are called sparse matrices. There are many applications where zero entries are common, e.g., document-term matrices representing textual documents and user-item matrices used to derive recommendations.

Numerous compression algorithms can be used to alter the data representation: General purpose algorithms such as Lempel-ziv as well as algorithms tailored to specific types of data. Sakr [45], for instance, surveys algorithms for XML data compressions. A dataset can be compressed in such a way that the entire dataset must be decompressed to access a single element. A compressed dataset might also allow for even faster access and manipulation of data than non-compressed data. For large matrices in a sparse matrix representation, for instance, some manipulations such as multiplication of two matrices are often faster. Compression and decompression also consume energy and, thus, data compression might or might not be beneficial depending on the number of required compress and decompress operations. General purpose algorithms allow to specify how much effort they should invest into finding the representation that minimizes space. Some algorithms take advantage of compressed representations and work on them directly, whereas others require an uncompressed representation. In case data is transferred across networks or is infrequently accessed, compression is even more appealing.

Principle #5.4: Accurate specification of attribute requirements

Whereas discrete attribute values stem from a fixed set of values, attributes with continuous values are

stored with a specific precision. The precision of individual attributes as well as the set of possible values can be defined by specifying an attribute type. For example, for an attribute containing temperature measurements, a data scientist might specify a precision of 0.001 degrees and a range of feasible values such as [0,100] as so called "domain constraint" in database systems [15]. As a next step a data type can be chosen that meets these requirements and uses the least amount of storage—for instance, databases provide a set of data types according to the SQL standard [15], whereas programming languages usually follow the IEEE standards for floating point, integer, and other data types. The data type also determines the amount of storage and impacts the time and energy to conduct operations on data. The green data scientist should specify reasonable requirements. Choosing inappropriate types might more than double the amount of needed storage. For example, choosing an integer type (64 bits) rather than a (single) byte type (8 bits) for an array of many values leads to an increase of a factor of almost eight in memory demand.

Domain constraints depend on the data source, the range of the data, and the intended application: For sensor data, the accuracy is given by the maximum precision that seems achievable in the next years. For financial data, the needed accuracy might be given by the smallest unit, that is, one cent or one dollar. For time information, a precision up to milliseconds might not yield better outcomes than maintaining timestamps with hourly precision. For images, accuracy can be translated to the maximal resolution in terms of number of pixels or color depth that is beneficial for the analysis.

4.3 Data preparation and modeling

Principle #6: Execute common operations only once

Data preparation should be structured in such way that common preparation operations for multiple models are executed only once. For example, it can be reasonable to store a version of pre-processed data after general transformation and cleaning steps have been performed. The principle of factoring out common operations is already known, for instance, in the context of the Extract-Transform-Load (ETL) process optimization for data warehouses [58]. The idea of storing temporary results has also been applied in the context of ETL processes [58] and it is an integral part of the distributed data processing for Map-Reduce jobs. In both cases, the goal is fault tolerance rather than energy optimization. Strategies for identifying data processing results likely to be reused and thus worth storing have been investigated, too—for instance, for Map-Reduce jobs [14].

Principle #7: Choose models that enable discarding of data or low energy storage

Data lifecycle management has embraced the idea of moving data from high-cost to low-cost storage [35], for instance, moving data between storage tiers based on the value of data [8]. Energy consumption and accessibility of stored data are typically negatively correlated: The easier it is to access data the more energy is required to maintain the data. Keeping data on a (magnetic) tape storage is much more energy efficient than keeping the same amount of data in the main memory of a computer. The former consumes energy only upon access, whereas main memory consumes energy even if no data is accessed. By her choices the data scientist determines the level of accessibility to data and thereby also the type of storage and amount of energy needed. The data scientist should thus be able to assess the relevance of data (over time) and assess the possibility to discard (older) data, compress (older) data, or work on summarized data. The availability of (old) data impacts the methodology that can be chosen, and the chosen methodology might also impact the data that must be stored. This is a key concern for long running systems, where data accumulates over time and models can be adjusted from time to time using newly available data. Some models can be trained incrementally using online learning algorithms, while others require the full dataset including all prior data, even in case only minor updates should be made due to new data using offline learning algorithms. For some models online as well as offline algorithms exist. Consider a system that classifies messages as spam or not spam. Such a system can be built by training a model based on previously classified messages. Since spammers adjust their strategies and style of messages, the system needs continuous updates—that is, learning. Whereas in an online learning scenario, data might be discarded after training the model, in the offline learning scenario it has to be kept.

Minimizing data access and thereby allowing to move data to energy friendly mediums is a viable option. But discarding data is a risky endeavor. What if the existing model should be replaced by a new model? Is it possible to change a model when all historic training data has been discarded? A careful assessment and management of risks is necessary. Various techniques from the domain of machine learning support reducing the need to keep data. One way is to use transfer learning [42, 31] by generating training data from the existing model for a new model, that is, to create labeled data in case unlabeled data is

available or can itself be generated. The disadvantage of this approach is that the generated labels are usually less accurate than the labels of the original dataset. Training data for the new model might still be highly beneficial despite transferred knowledge, but transfer learning can help reduce the amount of data needed to achieve good performance. Furthermore, training data can be enhanced by artificial training data that are a modification of existing data, thereby leading to improved results [4, 37]. Marginal returns decrease with additional data [16], and the impact in performance of having to retrain a new model might be small, even if just a small fraction of all data is retained.

Principle #8: Include only promising models and energy efficient algorithms

The traditional model selection process focuses almost exclusively on picking the model that yields the best results in terms of data mining-task-specific metrics such as accuracy or F-score for classification. A data scientist can base her model selection by comparing such metrics using empirical and theoretical comparisons (on similar datasets). The green data scientist, however, should also take into account energy consumption due to training, operating, and potentially data storage. Minor differences in task specific metrics might still be tolerable according to overall business objectives. It is not recommended to use all model and optimization algorithms as part of the computational selection process, because this leads to high energy costs. Ideally, the model candidates (and optimization algorithms) are limited to models that are likely to yield good results in terms of the desired metrics including energy efficiency. To this end, theoretical and empirical evidence should be leveraged.

A data scientist faces the choice of selecting model candidates and (hyper)parameter optimization algorithms. Energy costs are often determined by the effort to train and apply the model, that is, for predictions.

Principle #8.1: Leverage theoretical insights

Existing literature only gives limited advice on how to select the best methods for a dataset without trying them on the dataset at hand. Manning et al. [36] advocate the use of high bias classifiers if little data is available. Properties of the learning algorithm are not the only factor impacting energy consumption. The number of hyperparameters and the effort to optimize these parameters also play a vital role. There are little theoretical foundations with respect to the best choice of hyperparameter optimization methods. The field is subject to current research [32]. One theoretical insight is that obvious and intuitive techniques such as a systematic grid search might be inferior even to unstructured random search [5].

Models to describe the energy efficiency of systems and algorithms have been discussed from different perspectives such as power management [1], energy per low level operation (e.g., low level operations per Watt [24]), or models involving hardware components such as CPUs and memory [44]. However, none of these metrics seems suitable for quantifying the energy efficiency of models in the context of data mining. A data scientist usually works on a higher level of abstraction than individual hardware components and low-level CPU instructions that are the focus of many of these metrics. Theoretical computer science analyses algorithms in terms of running time. Running time, or time complexity, is the count of abstract, higher level operations needed to solve a task. The notion of time complexity can be applied to a single computer but also to a cluster of computers. In the field of parallel computing, one might simply aggregate the operations of all computers. This neglects costs due to information exchange between computers. Distributed systems such as clusters running data analytics frameworks such as Hadoop or Spark can also involve significant costs due to communication or idling (waiting for work). Generally, costs for communication, computation, and idling are tradeable [47, 29]. Many existing data mining algorithms are analyzed using the classical time complexity metric for a single computer, where the running time is often expressed as a function of the number of observations in a dataset and the number of dimensions. From the perspective of a green data scientist, algorithms with small time complexity seem preferable. But theoretical bounds might be coarse and, furthermore, often they neglect constants as part of the analysis process that might be of practical relevance. Therefore, empirical investigation might be more meaningful.

Principle #8.2: Leverage empirical knowledge

To the best of our knowledge a thorough comparison of learning algorithms for model parameters with respect to energy related concerns does not exist. Some works do provide empirical results for running-time of a few models, e.g., [48] in the field of density based clustering. Running time seems to be a viable surrogate metric for measuring energy consumption of models for training and operation. For other metrics such as accuracy, multiple publications provide comparisons [9, 26].

Hyperparameters often have a profound impact on model performance [55]. To optimize hyperparameters multiple strategies exist [32]. Some techniques try to reduce the time (and energy) for model selection by

training models on samples of data and predicting performance on the full dataset. Some optimization techniques allow to specify time constraints that guide the model selection process [57]. Unfortunately, empirical comparisons [13] do not report on the overall energy consumption for training, but rather focus on other metrics such as accuracy.

5. Conclusion and future work

We introduced principles for green data mining based on the CRISP-DM methodology. Our principles apply to various phases of the process, impacting managerial decisions (e.g., make-or-buy) as well as technical questions (e.g., which model to use to conserve energy?). Creating a platform allowing to share information on model performance based on hyperparameter settings and datasets will not only be valuable for fellow data scientists, but also for improving hyperparameter learning algorithms [32]. Aside from empirical contributions, theoretical insights related to model selection could advance the field of green data mining. Furthermore, a detailed evaluation of the proposed principles can help in their application.

6. References

[1] Albers, S., "Energy efficient algorithms", Communications of the ACM, 53(5), 2010, pp. 86-96.

[2] Aggarwal, C. C., Data mining: The Textbook, 2015.

[3] Behrens, J. T., "Principles and procedures of exploratory data analysis", Psychological Methods, 1997.

[4] Bengio, Y., "Deep learning of representations for unsupervised and transfer learning", Proc. of ICML Workshop on Unsupervised and Transfer Learning, 2012.

[5] Bergstra, J. and Bengio, Y., "Random search for hyper-parameter optimization", Journal of Machine Learning Research, 13(Feb), 2012, pp. 281-305.

[6] Borra, S., and Di Ciaccio, A., "Measuring the prediction error. A comparison of cross-validation, bootstrap and covariance penalty methods", Computational statistics & data analysis, 54(12), 2010, pp. 2976-2989.

[7] Brooks, S., Wang, X., and Sarker, S., "Unpacking green IT: A review of the existing literature", In Proc. of the Americas Conf. on Information Systems (AMCIS), 2010.

[8] Capra, E., and Merlo, F., "Green IT: Everything starts from the software", In Proc. of European Conf. on Information Systems (ECIS), 2009.

[9] Caruana, R., and Niculescu-Mizil, A., "An empirical comparison of supervised learning algorithms", In Proc. of Int. Conf. on Machine learning, ACM, 2006.

[10] Cook, G., Pomeranz, D., Rohrbach, K., and Johnson, B., Clicking Clean: A Guide to Building the Green Internet. Greenpeace Inc., Washington, D.C, 2015.

[11] M Najafabadi, M., and Luna-Reyes, L., "Open Government Data Ecosystems: A Closed-Loop Perspective", In Proc. of the Hawai Int. Conf. on System Sciences (HICSS), 2017.

[12] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., and Fei-Fei, L., "Imagenet: A large-scale hierarchical image database". In Conf. on Computer Vision and Pattern Recognition. IEEE, 2009.

[13] Eggensperger, K., Feurer, M., Hutter, F., Bergstra, J., Snoek, J., Hoos, H., and Leyton-Brown, K., "Towards an empirical foundation for assessing bayesian optimization of hyperparameters", In NIPS workshop on Bayesian Optimization in Theory and Practice (Vol. 10), 2013.

[14] Elghandour, I., and Aboulnaga, A., ReStore: reusing results of MapReduce jobs. In Proc. of the VLDB Endowment, 5(6), 2012, pp. 586-597.

[15] Elmasri, R., and Navathe, S. B., "Fundamentals of database systems", 6th edition, Pearson, 2011.

[16] Figueroa, R. L., Zeng-Treitler, Q., Kandula, S., and Ngo, L. H., "Predicting sample size required for classification performance", BMC medical informatics and decision making, 12(1), 2012.

[17] Fradkin D., and Madigan, D., "Experiments with random projections for machine learning", In Proc. of the Int. Conf. on Knowledge discovery and data mining, ACM, 2003.

[18] Gelenbe, E., and Caseau, Y., "The impact of information technology on energy consumption and carbon emissions", Ubiquity, 1, 2015.

[19] Goiri, Í., Le, K., Nguyen, T. D., Guitart, J., Torres, J., and Bianchini, R., "GreenHadoop: leveraging green energy in data-processing frameworks", In Proc. of the European Conf. on Computer Systems, ACM, 2012.

[20] Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y., Deep learning (Vol. 1), Cambridge: MIT press, 2016.

[21] Hindle, A., "Green software engineering: the curse of methodology", In Conf. on Software Analysis, Evolution, and Reengineering (SANER), IEEE, 2016.

[22] Jin, C., de Supinski, B. R., Abramson, D., Poxon, H., DeRose, L., Dinh, M. N. and Jessup, E. R., "A survey on software methods to improve the energy efficiency of parallel computing", The Int. Journal of High Performance Computing Applications, 2016.

[23] Han, J., Pei, J., and Kamber, M., Data mining: concepts and techniques, Elsevier, 2011.

[24] Hsu, C. H., Feng, W. C., & Archuleta, J. S., "Towards efficient supercomputing: A quest for the right metric", In Proc. of the Parallel and Distributed Processing Symposium, IEEE, 2005.

[25] King, W.R., and He, J., "Understanding the Role and Methods of Meta-Analysis in IS Research", Communications of the Association for Information Systems (16) Article 32, 2005, pp. 665-686.

[26] Kotsiantis, S. B., Zaharakis, I. D., and Pintelas, P. E., "Machine learning: a review of classification and combining techniques", Artificial Intelligence Review, 26(3), 2006.

[27] Kurgan, L. A., and Musilek, P., "A survey of Knowledge Discovery and Data Mining process models", The Knowledge Engineering Review, 21(1), 2006.

[28] Lei, C. F., and Ngai, E. W. T., "Green IT Adoption: An Academic Review of Literature", In Proc. of Pacific Asia Conf. on Information Systems (PACIS), 2013.

[29] Li, S., Maddah-Ali, M. A., Yu, Q., and Avestimehr, A. S., "A fundamental tradeoff between computation and

communication in distributed computing", IEEE Transactions on Information Theory, 2017.
[30] Loeser, F., "Green IT and Green IS: Definition of constructs and overview of current practices", In Pro. of the Americas Conf. on Information Systems (AMCIS), 2013.
[31] Lu, J., Behbood, V., Hao, P., Zuo, H., Xue, S., and Zhang, G., "Transfer learning using computational intelligence: a survey. Knowledge-Based Systems", 80, 2015.
[32] Luo, G., "A review of automatic selection methods for machine learning algorithms and hyper-parameter values", Network Modeling Analysis in Health Informatics and Bioinformatics, 5(1), 2016, p.18.
[33] Mastelic, T., Oleksiak, A., Claussen, H., Brandic, I., Pierson, J. M., and Vasilakos, A. V., "Cloud computing: Survey on energy efficiency", ACM computing surveys, 47(2), 2015, p. 33.
[34] Malhotra, A., N. Melville, et al., "Spurring Impactful Research on Information Systems for Environmental Sustainability", MIS Quarterly 37(4), 2013.
[35] Malik, P., "Governing Big Data: Principles and practices", IBM Journal of Research and Development, 57(3/4), 2013, pp. 1-13.
[36] Manning, D. C., Raghavan, P and Schuetze, H., Introduction to information retrieval. Cambridge Press, 2008.
[37] Melville, P., & Mooney, R. J., "Constructing diverse classifier ensembles using artificial training examples", In IJCAI, Vol. 3, 2003, pp. 505-510.
[38] Melville, N. P., "Information systems innovation for environmental sustainability", MIS Quarterly 34(1), 2010, pp. 1-21.
[39] Molla, A., & Cooper, V., "Greening data centres: The motivation, expectancy and ability drivers", In Proc. of European Conf. on Information Systems (ECIS), 2014.
[40] Orgerie, A. C., Assuncao, M. D. D., and Lefevre, L., "A survey on techniques for improving the energy efficiency of large-scale distributed systems", ACM Computing Surveys (CSUR), 46(4), 2014, p. 47.
[41] Piatsky, G., "CRISP-DM: Still the top methodology for analytics, data mining, or data science projects", KDNuggets, http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html, 2014, (accessed 2018-05-20).
[42] Pan, S. J., and Yang, Q., "A survey on transfer learning", IEEE Transactions on knowledge and data engineering, 22(10), 2010, pp. 1345-1359.
[43] Radev, D. R., Jing, H., Styś, M., and Tam, D., "Centroid-based summarization of multiple documents", Information Processing & Management, 2004.
[44] Roy, S., Rudra, A., & Verma, A., "An energy complexity model for algorithms", In Proc. of Conf. on Innovations in Theoretical Computer Science, ACM, 2013.
[45] Sakr, S., "XML compression techniques: A survey and comparison", Journal of Computer and System Sciences, 75(5), 2009, pp. 303-322.
[46] Sayood, K., Introduction to Data Compression. 4th Edition, Elsevier, 2012.
[47] Schneider, J., and Wattenhofer, R., "Trading bit, message, and time complexity of distributed algorithms", In Int. Symposium on Distributed Computing. 2011.
[48] Schneider, J., & Vlachos, M., "Scalable density-based clustering with quality guarantees using random projections", Data Mining and Knowledge Discovery, 2017, pp. 1-34.
[49] Settles, B., "Active learning", Synthesis Lectures on Artificial Intelligence and Machine Learning, 6(1), 2012, pp. 1-114.
[50] Shehabi, A., Smith, S.J., Sartor, D.A., Brown, R.E., Herrlin, M., Koomey, J.G., Masanet, E.R., Horner, N., Azevedo, I.L., Lintner, W., "US Data Center Energy Usage Report", 2016, https://eta.lbl.gov/publications/ united-states-data-center-energya (accessed 2018-05-15)
[51] Sorzano, C.O.S., Vargas, J., & Montano, A. P., "A survey of dimensionality reduction techniques", arXiv preprint arXiv:1403.2877, 2014.
[52] Stange, M. and Funk, B., "How Much Tracking Is Necessary? - The Learning Curve in Bayesian User Journey Analysis", ECIS 2015 Completed Research Papers, 2015.
[53] Swan, A., & Brown, S., "The skills, role and career structure of data scientists and curators: An assessment of current practice and future needs", University of Southampton, 2008.
[54] Thompson, S. K., Sampling, John Wiley & Sons, Hoboken NJ, 2012.
[55] Thornton, C., Hutter, F., Hoos, H.H. and Leyton-Brown, K., "Auto-WEKA: Combined selection and hyperparameter optimization of classification algorithms", In Proc. of It. Conf. on Knowledge discovery and data mining, 2013.
[56] Tong, S., & Koller, D., "Support vector machine active learning with applications to text classification", Journal of machine learning research, 2001.
[57] Van Rijn, J. N., Abdulrahman, S. M., Brazdil, P., and Vanschoren, J., "Fast algorithm selection using learning curves", In Int. Symposium on Intelligent Data Analysis, Springer, 2015.
[58] Vassiliadis, P., "A survey of Extract–transform–Load technology", Int. Journal of Data Warehousing and Mining (IJDWM), 5(3), 2009, pp. 1-27.
[59] Van der Aalst, W. M., ".Green Data Science: Using Big Data in an "Environmentally Friendly" Manner", Proc. of the 18th Int. Conf. on Enterprise Information Systems (ICEIS), 2016 pp. 9-21.
[60] Watson, R. T., M.-C. Boudreau, Chen, A. J., « Information systems and environmentally sustainable development: Energy informatics and new directions for the IS community", MIS Quarterly 34(1), 2010, pp. 23-38.
[61] Wirth, R., & Hipp, J., "CRISP-DM: Towards a standard process model for data mining", In Proc. of the Int. Conf. on the practical applications of knowledge discovery and data mining, 2000, pp. 29-39.
[62] Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J., Data Mining: Practical machine learning tools and techniques, Morgan Kaufmann, 2016.
[63] Wolpert, D. H., "The lack of a priori distinctions between learning algorithms", Neural computation, 8(7), 1996, pp. 1341-1390.
[64] Yosinski, J., Clune, J., Bengio, Y., and Lipson, H., "How transferable are features in deep neural networks?", In Advances in neural information processing systems, 2014.
[65] Yu, J., "Big Data vs. Relevant Data: Intelligence That Matters", Huffington Post, 2014 (2014-26-03), https://www.huffingtonpost.com/jim-yu/big-data-vs-relevant-data_b_5022792.html (accessed 2018-05-11)