Bus Inf Syst Eng 65(1):65–83 (2023)
https://doi.org/10.1007/s12599-022-00780-w

RESEARCH PAPER

Reuse, Reduce, Support: Design Principles for Green Data Mining

Johannes Schneider · Stefan Seidel · Marcus Basalla · Jan vom Brocke

Received: 16 April 2021 / Accepted: 30 August 2022 / Published online: 5 December 2022

Abstract This paper reports on a design science research (DSR) study that develops design principles for "green" – more environmentally sustainable – data mining processes. Grounded in the Cross Industry Standard Process for Data Mining (CRISP-DM) and on a review of relevant literature on data mining methods, Green IT, and Green IS, the study identifies eight design principles that fall into the three categories of reuse, reduce, and support. The paper develops an evaluation strategy and provides empirical evidence for the principles' utility. It suggests that the results can inform the development of a more general approach towards Green Data Science and provide a suitable lens to study sustainable computing.

Keywords Green Data Science · Green IT · Green IS · Data mining · Energy efficiency · Energy-saving · Design science research · Design principles

1 Introduction

The use of computing power, coupled with the unprecedented availability of data, provides ample opportunity to save energy and improve energy efficiency (Gelenbe and Caseau 2015). At the same time, the collection, storage, processing, and application of data constitute an increasingly relevant source of energy consumption and associated carbon emissions. In 2018, data centers were estimated to consume about 205 TWh or 1% of energy worldwide (Masanet et al. 2020), corresponding to the energy produced by about 100 nuclear power plants.

Data mining is at the core of data science. It contributes to this consumption and associated emissions. Data scientists use a large pool of computationally costly methods to extract knowledge and insights from data. Data mining refers to the processes of knowledge discovery from data (Han et al. 2011) through activities such as collecting, cleaning, processing, and analyzing. Using data mining is now prevalent across various application areas as data scientists have increasing amounts of data originating from multiple data sources at their disposal. Data mining is also one of several processes used in decision analytics to formally analyze and support important managerial decisions (Suhl and Voss 2014). Systems relying on learning from data are rapidly gaining importance. The International Data Corporation (IDC) predicts an annual growth of data beyond 20% from 2020 to 2025.[1]

The existing literature on data mining in general (e.g., Aggarwal 2015) and data mining *processes* in particular (e.g., Kurgan and Musilek 2006) partially addresses computational and storage efficiencies, such as data reduction and approximate algorithms. Van der Aalst (2016) uses the term "Green Data Science" to describe technological solutions that allow to mitigate the effects of data science on the social environment caused by "unfairness, undesired disclosures, inaccuracies, and non-transparency" (p. 9).

Accepted after two revisions by Stefan Lessmann.

J. Schneider (✉) · S. Seidel · M. Basalla · J. vom Brocke
University of Liechtenstein, Vaduz, Liechtenstein
e-mail: johannes.schneider@uni.li

S. Seidel
e-mail: stefan.seidel@uni.li

M. Basalla
e-mail: marcus.basalla@uni.li

J. vom Brocke
e-mail: jan.vom.brocke@uni.li

[1] https://www.idc.com/getdoc.jsp?containerId=prUS47560321.

Existing literature has also attended to the role of data, data storage, and data analytics from an environmental perspective, e.g., by identifying performance metrics for green data centers (Wang and Khan 2013) or addressing the energy efficiency in data center networks (Bilal et al. 2014; Wang and Khan 2013). Still, to the best of our knowledge, we are missing insights into ways by which data scientists can reduce the environmental impact of their practices *across their lifecycles*. Such knowledge, however, is becoming increasingly important as the use of data intensive methods becomes prevalent across industries. Furthermore, Hedman and Henningsson (2016) have shown that green initiatives in organizations result from a bottom-up process by individuals such as data scientists. Like Information Systems (IS), data science is also highly interdisciplinary, i.e., a data scientist needs both technical knowledge and business understanding. Thus, a data scientist might be stereotypical for IS and interesting to focus on. She might be primed as a "human information system," i.e., a person with socio-technical skills acting within an organization collecting, processing, and distributing information. In this study, we thus seek answers to the following research question:

> How can data scientists implement greener data mining processes?

To this end, we conducted a design research (DSR) study that develops and evaluates design principles for data scientists to implement green (i.e., environmentally sustainable) data mining processes. In the context of environmental sustainability, IS scholars have called for actionable solutions that reduce energy consumption and combat climate change (Gholami et al. 2016). Furthermore, sharing data on "what works and what does not " (Melville 2010, p.2) has been called for in the context of sustainability more than ten years ago, and it is still recited in current calls for action (Watson et al. 2021). Our design principles and their evaluation contribute to addressing these calls.

Design principles (Gregor et al. 2020) are a form of nascent design knowledge. They can contribute to the development of a more general design theory of Green Data Science (Gregor and Hevner 2013). We use the *Cross Industry Standard Process for Data Mining (CRISP-DM)* (Wirth and Hipp 2000) as the overarching framework for the development of these design principles. It is the most widely used process for data mining, which we can attribute to its flexibility – the model covers a wide range of data mining projects from different domains. At the same time, it is not too generic and thus provides sufficient guidance. Making decisions about a specific data science process requires the data scientist to explore a vast problem space as she chooses from a wide variety of methods. She needs to make decisions about data preparation, modeling, evaluation, and deployment. A set of design principles can assist her in traversing this problem space and in identifying feasible solutions efficiently and effectively. We hope to provide a set of foundational principles along with a first evaluation that can guide further research. Developing a first set of design principles provides an essential step for developing the emergent field of Green Data Science.

We will proceed as follows. We first give an overview of related literature on data mining as well as Green Information Systems (Green IS) and Green Information Technology (Green IT). We then describe our research methodology. Next, we construct a set of principles for Green Data Mining grounded in a review of the literature. We develop an evaluation strategy, provide empirical evidence of the principles' utility, and highlight general issues regarding the evaluation of such design principles. We then discuss our principles and highlight their implications for research and practice. We conclude by elaborating on limitations and future work.

2 Research Background

We begin by describing typical activities in a data mining process that involve energy usage. We then describe how our work relates to the existing discourse on Green IS, Green IT, and Green Software.

2.1 The Data Mining Process

Various conceptualizations of data mining processes (Kurgan and Musilek 2006) and machine learning processes exist (Amershi et al. 2019), most of which share common phases. CRISP-DM (Fig. 1) is the most widely known and practiced model, addressing business and data understanding, data preparation, modeling, evaluation, and deployment (Wirth and Hipp 2000). The business understanding phase clarifies project objectives and business requirements, which the data scientist then translates into a data mining problem such as clustering or classification (Aggarwal 2015; Han et al. 2011). Data understanding requires initial data selection or collection. The data scientist first analyzes data in an exploratory fashion to understand it in the business context. The exploratory analysis supports the development of hypotheses by identifying patterns in the data (Behrens 1997). This step allows the data scientist to get first insights as well as to identify data quality problems. Data preparation includes using raw data to derive data to be fed into models. Activities include data selection, transformation, and cleaning. The data scientist may have to prepare the data separately for each

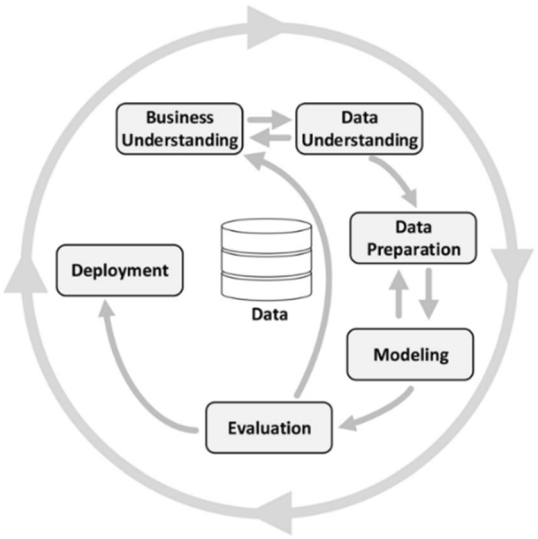

Fig. 1 Crisp DM

model. Modeling involves identifying suitable models as well as selecting and adapting a model, e.g., by optimizing its parameters. Computational model evaluation is part of the model selection process. No clear consensus exists about which model is best for a task, as indicated by the "no free lunch" theorem, stating that no algorithm outperforms all other algorithms on all datasets (Wolpert 1996). Consequently, the data scientist often cannot avoid some form of trial and error. The model choice depends on many factors such as data characteristics (dimensionality, number of observations, structuredness), data mining objectives (need for the 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 Science, the data scientist also assesses performance in terms of energy consumption for model training and model use, e.g., for making predictions. For the evaluation phase, the main goal is to review all steps involved in constructing the model and to verify whether the final model meets the defined business objectives. The model that best meets the evaluation criteria 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 performing a prediction task.

2.2 Green IS, Green IT, and Green Software

Analyzing data mining processes from a sustainability perspective is subject of the fields of Green IS, in particular the subfield of Green Software (Green SW) and Green IT. Green IT concerns itself with energy efficiency and equipment utilization of information technology throughout its lifecycle (Watson et al. 2008). Relevant topics include computational methods (Albers 2010), their implementation in software (Capra and Merlo 2009; Hindle 2016), hardware components of computers (Hsu et al. 2005; Roy et al. 2013), data centers (Molla and Cooper 2014), cloud computing (Gelenbe and Caseau 2015; Malhotra et al. 2013), and parallel data processing for big data (Goiri et al. 2012; Jin et al. 2017). Data centers (Molla and Cooper 2014) and their hardware as well as cloud platforms provide the necessary infrastructure for data scientists but are themselves not the subject of study in this work. Practices of (green) data scientists, as well as their preferred methodology, might inform the design of green data centers and hardware in general, e.g., green data centers might make common algorithms used in green data mining readily available or improve their energy efficiency through hardware and algorithm co-design. Our work relates to the type of Green SW that aims at software consuming less energy to run and being created by using a process that saves resources and reduces waste (Taina 2011). Data science differs from software engineering in that it is more computationally intensive since it relies on learning from or mining large amounts of data. Its outcomes such as models are often narrower in scope than software, i.e., a model is typically part of a software product, and a model has a smaller source code base because model behavior is learned rather than programmed.

Green IS concerns itself with the design and implementation of information systems that contribute to environmentally sustainable organizational processes (Watson et al. 2008). Relevant topics include sustainable procurement (Brooks et al. 2010), the role of information systems in organizational sustainability transformations (Seidel et al. 2013), the role of information systems in supporting more energy-efficient behavior (Loock et al. 2013), or the generation of organizational benefits by translating environmental strategies into Green IS initiatives (Löser, Recker, vom Brocke, Molla, Zarnekow 2017).

Considering these key goals of both Green IT and Green IS, our study is at the intersection of Green IT and Green IS. First, we seek to learn about how data scientists can improve the energy footprint of IT (in the context of data science), and our study thus falls into the category of Green IT. Second, data mining is an organizational practice, and we are interested in how information systems can help reduce the environmental impact of this organizational practice. Our work thus also falls into the category of Green IS. The implementation of our principles could also be supported by an environmental management information system (EMIS) (El-Gayar and Fritz 2006) that makes available relevant environmental information for and of data scientists.

Over the past years, researchers have repeatedly called for investigations into how information systems should be

designed that support environmental sustainability (Melville 2010; Seidel et al. 2017; Watson et al. 2010). Our work contributes to the growing body of design knowledge on how information systems should be designed for environmental sustainability (e.g., Hilpert et al. 2013; Seidel et al. 2018).

3 Methodology

We derive and evaluate design principles through a DSR process consisting of five key steps: (1) identify problem; (2) define solution objectives; (3) design and development; (4) demonstration and evaluation; and (5) communication (Peffers et al. 2007). We use the CRISP-DM model as well as relevant literature on data mining methods to guide our model development. CRISP-DM provides a general framework that highlights key data mining stages, each of which we aim to address by developing respective design principles. By addressing the key elements and stages of the data mining process, CRISP-DM serves as an analytical kernel theory for developing design knowledge (Gregor and Jones 2007). Kernel theory, broadly, serves as justificatory knowledge in the development of design knowledge[2] (Gregor and Hevner 2013), such as in the form of design principles (Gregor et al. 2020). The literature of Green IS and Green IT provides a dual perspective that sensitizes us towards considering both the effects of using data mining algorithms and the outcomes of using data mining algorithms to design and implement more sustainable work practices. Together, these inputs allow us to construct a set of design principles for Green Data Mining. Next, we briefly describe how we implemented the five key steps of our research design.

We first identified the problem (Step 1 – identify problem). Approaches to extracting knowledge from data through data mining are now key analytical processes in a broad area of applications, ranging from organizational decision making to the use of personal devices. Using these approaches consumes energy and thus contributes to an organization's environmental footprint. However, we did not find guidelines on how to consider environmental aspects in these processes. We thus aimed to address this problem by developing and evaluating a set of design principles that data scientists can use to decrease the environmental footprint of their work.

The solution objectives (Step 2 – define solution objectives) were that the design principles should (a) target a key element of data science processes – data mining – and (b) should be sufficiently abstract to be applied across contexts.

Our design and development stage ensued, consisting of the following key activities (Step 3 – design and development):

1. Identify general design principles to reduce the environmental footprint of data mining practices
2. Identify the key factors that determine energy consumption in data mining
3. Refine the principles using the CRISP-DM process by investigating possibilities for reducing energy consumption in relation to each factor

We limit our analysis to those aspects that data scientists can directly influence, including the choice of data, its representation, and processes and techniques the data scientist uses throughout the data mining 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 support doing so.

Our design and development stage is grounded in a narrative literature review (King and He 2005) on Green IT, Green IS, and data mining because our goal was to investigate elementary factors and research outcomes related to these research areas. Green Data Science is a novel field. Therefore, it is more amenable to a qualitative approach such as a narrative literature review than a more quantitative approach detailing the current state of research. Our focus was on using established online databases to set the context of our work, i.e., IEEE Xplore, ProQuest (ABI/INForm), AIS electronic library, and the ACM digital library. We did not limit ourselves to journals because scholars often present new ideas first at academic conferences. A significant body of works, particularly in computer science, only appear as conference articles. Our initial search targeted to identify recent works (from 2015 to 2021) that are at the intersection of two topics: (i) energy efficiency and reduction (in the context of Green IS/IT) and (ii) data mining (including machine and deep learning). Furthermore, we aimed to exclude applications of data mining to foster sustainability as our focus is on making data mining (and its outcomes such as computational models) green. Since terminology differs across disciplines, we did not use the same search terms for all databases. For the AIS electronic library and ProQuest (ABI/INForm) we filtered using ("data exploration" OR "data processing" OR "data mining" OR "machine learning" OR "deep learning"). In addition, we filtered the abstracts based on (sustainability OR "green IS" OR "green IT" OR energy), yielding 238 hits for the AIS library and 406 for ProQuest. For the IEEE and ACM libraries, where Green IS and Green IT are not so prevalent, we filtered the

[2] Kernel theory was proposed by Walls et al (1992) as "theories from natural science, social sciences and mathematics " encompassed in design theories (p. 41). We adopt Gregor and Hevner's (2013) perspective of kernel theories as „any descriptive theory that informs artifact construction " (p. 340).

abstract differently, i.e., we required an abstract to include either "sustainability OR green" and, in addition, "(power OR energy) AND (saving OR reduction OR efficiency)." We also had a set of exclusion terms that were not supposed to occur in the abstract such as grid and building or in the full text such as wireless, cellular, cache, city, quantum, FPGA, microarchitecture, dram, and physics. This yielded 1055 hits for IEEE and 318 for ACM. Still, most articles were either about applying data mining to a problem or addressed a specific energy reduction problem, e.g., in the context of hardware development. We excluded most articles based on title and abstract.

We investigated the full texts of 68 articles as a starting point to develop design principles. We identified important concepts through open coding (Wiesche et al. 2017) while we used the CRISP- DM model as a sensitizing device to guide our analysis. That is, while we used the model's overall phases as pointers to what matters in data mining processes, we were open to discovering codes that would help us generate prescriptive knowledge about "green" practices within those phases. This process of coding the literature under consideration of general data mining phases allowed us to identify ideas that then informed our analysis and provided the foundation to extract prescriptive knowledge, considering that our overall goal was to identify and describe design principles so that they can be applied by a green data scientist. We could divide the codes we identified into two broad groups, according to level of detail. First, high level codes provided the overall theme that was relevant for classifying codes into more abstract categories. Low level codes captured more detailed insights related to specific methods that mattered for a certain design principle. These codes thus captured the value of a method, e.g., in terms of energy savings and applicability, and often helped us refine the overall theme, for instance, by providing important ideas of the method. For factors related to energy consumption, we looked for various direct and indirect costs in the context of data mining and then reflected on the extent to which they relate to energy costs. For example, literature identified in our initial search mentioned labor costs as a financial cost, which caused us to investigate the relationship between labor and energy, e.g., how much energy a human consumes over time.

The final principles and factors related to energy consumption resulted from multiple rounds of searching (based on forward/backward search and previously identified concepts), coding, and further structuring of coded concepts into principles and factors related to energy consumption, drawing on CRISP-DM as our overall sensitizing model. Eventually, 36 articles informed the construction of our design principles.

In the last step (Step 4 – demonstration and evaluation) we empirically evaluated the design principles. We mainly discuss two evaluation cases, covering both text and image data. The evaluation reported in this paper provides a starting point for a research agenda for evaluating, refining, and continuously developing prescriptive knowledge on how to conduct green data mining.

Finally, we communicate the findings from our study (Step 5 – communication). We contend that design principles are an easy-to-access and appropriate way to communicate prescriptive knowledge about green data mining (Gregor et al. 2020). Through sharing the derived expertise in the form of design principles, we aim to contribute to the cumulative creation of knowledge, specifically in the fields of Green IT and Green IS.

4 Principles of Green Data Mining

This section identifies factors determining the ecological footprint of data mining and develops a set of design principles that data scientists can apply to reduce this footprint. We thus apply general ideas related to sustainability and Green IT (Murugesan 2008; Watson et al. 2010) and Green SW (Calero and Piattini 2015) to the context of Green Data Mining.

In a first step, we identified three general practices of greening data science approaches – reusing, reducing, and supporting (Fig. 2). We derive them from an analysis of the stages and associated practices of CRISP-DM as well as from the general idea that environmental sustainability requires us to consider the source function of any activity in terms of renewable and non-renewable resources that provide input to that activity (Goodland 1995). The categories of reusing and reducing impact the source function – and thus also the required computation and unwanted emissions – by decreasing the required energy for generating data and models and for computing. Finally, support relates to the application of resources to allow other data scientists to implement more environmentally friendly practices. Examples are data and model sharing. The idea of reuse, reduce, and support (i.e., sharing of knowledge) has also been discussed for specific cases in works on Green IT (Löser 2013).

In the second step, based on these three general-level practices, we identified key factors for energy consumption in data mining (Table 1) related to the CRISP-DM process and definitions of energy consumption (Chen et al. 2011). We derived these factors by investigating the constraints under which data scientists work and their choices when they follow the CRISP-DM process by analyzing existing literature, i.e., these factors are synthesized from the literature.

A data scientist does not control all factors relevant to energy consumption (Table 1). Data scientists typically

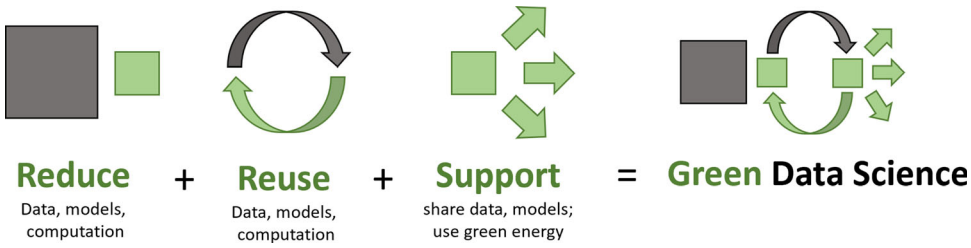

Fig. 2 Illustration of key practices of reuse, reduce, and support

Table 1 Factors related to energy consumption in data mining

Factor	Subfactors
Project objectives and execution	Performance specification; Make, Reuse, Share
Labor (energy used by data scientists)	Energy to conduct the entire project; Energy to implement design principles for green data mining
Model performance (energy costs such as rework due to model errors or waiting for model outputs)	Accuracy; Inference time
Data	Quantity; Quality; Representation; Accuracy; Data acquisition method; Data storage
Computation (analysis)	Structuring of computation; Choice and training of models
IT infrastructure	Hardware, e.g., CPU, GPU, Storage

have to use given hardware and implement a project according to the specification provided by business. Data scientists can influence the energy consumption of data mining processes by choosing computational methods, data sets, and models and structuring their work – all of which determine energy consumption.

Model performance, i.e., model accuracy and inference time to make predictions, might also influence energy costs in a more contrived way. A "green model" might trade prediction accuracy for energy consumption (Han et al. 2016; Li et al. 2017), leading to lower electricity usage due to less computation. At the same time, prediction errors can cause additional energy consumption, which only becomes visible upon model deployment. For example, consider an image recognition system that does not detect a defective part. A customer might return a defective part, leading to additional work associated with carbon emissions. Rework could have been avoided by immediately detecting the defect. Energy costs (due to rework) might also depend on the type and magnitude of the error – analogous to the lost value due to an incorrect decision as articulated in the expected value framework (Provost and Fawcett 2013).

A human in an industrial nation such as the US or Germany consumes more than 100 kWh per day (Snurr and Freude 2021). Working time by data scientists (and, thus, energy costs) might rise because considering environmental aspects in a data mining project poses an additional workload and leads to larger complexity, thereby increasing the data scientist's cognitive load. While we shall discuss such tradeoffs, we acknowledge that the nature of the impact of tradeoffs is highly problem specific.

Furthermore, energy costs vary across time, as shown in Fig. 3. Costs in the development phase (including the data understanding, preparation, modeling, and evaluation phases of CRISP-DM) are often characterized by high computational load due to training (many large) models (Strubell et al. 2020). However, energy costs are highly project specific. Computational energy costs can vary strongly during deployment, for example, energy costs for deployment could be zero if the project's outcome is insights from data (rather than a deployed model). Still, they could be much higher than energy costs during development if aggregated, but also per unit of time. Typically, models require updates from time to time (Tsymbal 2004). Updates are commonly triggered by a deterioration of model performance if the data distribution shifts over time (concept drift) or due to the availability of better models and more data that can improve performance.

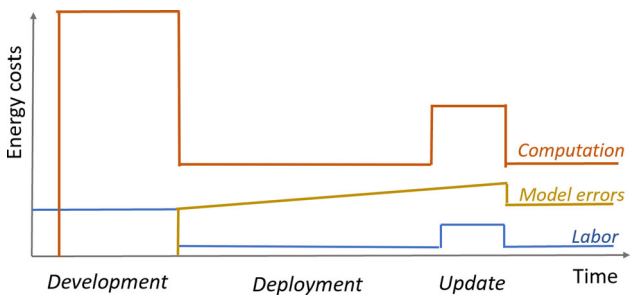

Fig. 3 Exemplary energy costs over time, i.e., across the project lifecycle, depending on direct costs such as computation and labor of data scientists as well as indirect costs such as model errors. Energy costs are highly project specific

Table 2 Overview of design principles for green data science

Reduce
Design Principle #1 – Reduce data
Design Principle #1.1 – Reduce number of data items
Design Principle #1.2 – Reduce number of attributes
Design Principle #1.3 – Reduce size of attributes
Design Principle #2 – Reduce operations
Design Principle #2.1 – Reduce trial and error
Design Principle #2.2 – Reduce model
Reuse
Design Principle #3 – Reuse own operations, data, and models
Design Principle #4 – Reuse from others
Support
Design Principle #5 – Share data, models, and skills

The energy consumption-related factors (Table 1) and the three general practices (Fig. 2) provide the conceptual foundation for our design principles. We structure the presentation of the design principles along these three key practices. Table 2 provides an overview. While the data scientist might have limited control over the project specification (i.e., objectives, etc.), she might have more freedom in controlling tradeoffs under consideration of multiple goals. Some principles involve tradeoffs between energy consumption and other metrics related to project outcomes. Key decisions the data scientist can make include decisions about data characteristics and data processing, which provide the primary scope of the principles (Table 2).

4.1 Design Principles of Reduce

Design Principle #1 – *Reduce data*

Store and process data with the minimum precision and volume needed while using a representation that minimizes storage and processing.

The choice of the quantity and precision of data to be processed is an important decision for energy usage. Opinions about data need often change as the business understanding evolves or as a result of the evaluation phase. For example, a data scientist might notice that the desired business questions cannot be addressed due to missing attributes or due to insufficient data quantity. Next, we describe strategies to limit the number of attributes or observations, reduce precision, and change data representation. We detail these strategies in a set of sub-principles.

4.1.1 Design Principle #1.1 – Reduce Number of Data Items

Models benefit from training data in a non-linear fashion with decreasing marginal gains through using increasing amounts of data (Figueroa et al. 2012). Therefore, reducing the data volume can considerably impact energy consumption in some scenarios but only lead to a minor impact on other relevant metrics. Standard sampling techniques (Stange and Funk 2015) can help reduce the amount of data. One of the simplest but often sufficient approaches is random sampling – choosing each data point with the same probability without replacement. For instance, only one-fourth of the available data was enough for a large dataset to achieve the optimal tradeoff between accuracy and data collecting and processing cost (Stange and Funk 2015). Active learning (Aggarwal 2015), a more elaborate technique, seeks to acquire relevant samples for learning incrementally. An active learner can ask explicitly for data labels that are expected to yield maximal improvement in learning. This is most beneficial in case such data must be collected based on energy intense (physical) experiments. Data valuation strategies, e.g., using data Shapley, allow removing data that actually diminishes the accuracy of classifiers (Ghorbani and Zou 2019). That is, removing data can lead to both energy savings and improvements in performance. In the context of big data, data deduplication has been proposed to improve energy efficiency (ur Rehmann et al. 2016). Data stream mining has also elaborated on instance reduction by selecting representative samples (Ramírez-Gallego et al. 2017).

4.1.2 Design Principle #1.2 – Reduce Number of Attributes

The data scientist can reduce the dimensionality using feature selection, feature extraction, and type transformation (Aggarwal 2015; Sorzano et al. 2014). Prominent dimensionality reduction techniques include singular value decomposition (SVD), principal component analysis (PCA), and simple random projections (Sorzano et al. 2014). There are also specific techniques for big data (ur Rehmann et al. 2016) and data streams (Ramírez-Gallego et al. 2017). This form of data compression typically implies a loss of precision (Sayood 2017). A technique might distort some instances more than others. A small number of different samples in the original context can be very similar in the lower-dimensional space. This can be unacceptable for tasks such as outlier detection, where one needs to identify a few specific data points. Other tasks such as clustering are typically less impacted by corrupting a few instances. Many of these techniques require that the data scientist solves an optimization problem, a process

that can be computationally intensive, thus making potential energy savings case-dependent.

4.1.3 Design Principle #1.3 – Reduce Size of Attributes

Whereas discrete attribute values stem from a fixed set of values, attributes with continuous values are stored with a specific precision level. The data scientist can define the precision of individual attributes and the set of possible values through the attribute type, e.g., using a so-called "domain constraint" in database systems (Elmasri and Navathe 2010). She can choose a data type that meets these requirements and at the same time uses a minimal amount of storage. For example, selecting 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 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 following years. For financial data, the needed accuracy might be provided by the smallest unit, i.e., one cent or one dollar. The idea of attribute size reduction has also been expressed as feature space simplification in the context of data stream mining (Ramírez-Gallego et al. 2017), where continuous attributes are discretized and placed into a few distinct categories.

Design Principle #2 – Reduce operations

Minimize model evaluations and parameter tuning. Conduct as few operations on each data item as possible. Each operation should be as simple as possible.

Only the most promising parameters should be used, models, and energy-efficient algorithms. That is, the data scientist should only consider model candidates, parameter settings, and optimization algorithms that are likely to yield good results in terms of the desired metrics, including energy efficiency. The principle addresses mainly the modeling and deployment phase in CRISP-DM. It also includes the data preparation and evaluation phases, where a significant amount of overall computation might take place. The data scientist can preselect models and algorithms based on results from small data samples and theoretical and empirical findings. The data scientist should also consider energy consumption due to training, usage of models, and data storage as a model assessment metric. A specific strategy to save energy is "early stopping." The idea is to discontinue the training of a model once it becomes evident that it cannot be better than the best solution found so far.

4.1.4 Design Principle #2.1 – Reduce Trial and Error

In the context of the "no free lunch theorem" (Wolpert 1996), no one model exists that fits all tasks. Only through a certain amount of preferably principled experimentation can one find the best model for a given problem.

Existing literature only gives limited advice on selecting the best methods for a dataset without trying them on the dataset at hand. For instance, Manning and associates (2010) advocate using high bias classifiers in situations where little data is available. The number of hyperparameters and the effort to optimize these parameters also impact energy consumption and performance (Thornton et al. 2013). For optimal performance, hyperparameters must be tuned. This can be done either manually or automatically by using hyper-optimization algorithms (Luo 2016). Two common algorithms are random search (Bergstra and Bengio 2012) and Bayesian optimization (Snoek et al. 2012). Finding the best hyperparameters is still an active research area. One theoretical insight is that evident and intuitive techniques such as a systematic grid search might be inferior even to an unstructured random search (Bergstra and Bengio 2012) because parameters are often un- or weakly correlated. For standard classification problems, existing knowledge on model and algorithm selection for a particular dataset can be provided by cloud-based solutions, e.g., Google Cloud AutoML. AutoML techniques (He et al. 2019) leverage a rich set of knowledge from optimizing similar classification problems. Using these techniques can help reduce the need for carrying out energy intensive experimentation to identify suitable models. There are also AutoML techniques that focus on finding models with minimal resource consumption (Fedorov et al. 2019).

Furthermore, model investigation using explainability techniques (Meske et al. 2021) can help in identifying model weaknesses, which can be used to select or improve models rather than relying on assessing different hypotheses by gathering data through computation in a trial-and-error manner.

4.1.5 Design Principle #2.2 – Reduce Model

A data scientist can optimize models for energy efficiency. Optimizing a model for energy efficiency is often not beneficial for a model serving a one-time decision-making problem. But it is likely of value for models that are intended for repeated use. For example, Han et al. (2016) showed energy savings of more than a factor of 1000 for deep learning networks using model compression. While the optimizations in their work might be beyond the capabilities of an ordinary data scientist due to their reliance on low-level hardware knowledge, other methods are relatively easy to implement but can still yield considerable gains. For instance, eliminating neurons with minor contributions to neural networks can save more than 30% (Li et al., 2017). For deep learning models, one can store

internal state representations at a low precision with appropriate rounding strategies, leading to only a minor loss in performance (Gupta et al. 2015; Hubara et al. 2016). One can also apply knowledge distillation to train a small, energy-efficient model based on an ensemble of large models with little performance loss (Goel et al. 2020). Models apart from deep learning can also be optimized, e.g., Kumar et al. (2020) showed a reduction of about 14.4% in energy at the cost of up to 0.48% in accuracy using automated code optimization of data mining algorithms towards lower energy consumption.

4.2 Design Principles of Re-use

Design Principle #3 – Reuse own operations, data, and models

Reuse temporary results rather than recompute.

Common operations should only be executed once. A data scientist should structure computation on data so that results of identical operations on the same data during preparation and modeling are reused and are thus only executed once. The principle of factorizing out common operations is already known, e.g., in the extract-transform-load (ETL) process optimization for data warehouses (Vassiliadis 2009), but also in the context of low-level implementations of deep learning (Kwon et al. 2019). In ETL processes, it can be reasonable to store a version of pre-processed data after the data scientist performs general transformation and cleaning steps. Storing intermediate results is also an integral part of the distributed data processing for map-reduce jobs. In both cases, the goal is fault tolerance rather than energy optimization. Writing to and reading data from a persistent medium such as a hard drive might require more energy than (re-)computing results. Therefore, partial results should primarily be stored when the results stem from complex operations and are reused frequently.

Design Principle #4 – Reuse from others

Reuse existing knowledge, data, and models from third-party sources.

The data scientist might control "make-or-buy" and "make-or-reuse" decisions. The principle is of primary concern for the modeling phase. It also bears relevance for the data preparation and understanding phases, when the data scientist makes decisions about what data is needed. For example, a data scientist could choose to acquire data from social media channels such as Twitter or Facebook and analyze it herself for marketing purposes. She could also acquire models (i.e., software) to conduct the analysis or consult an external company. From an environmental perspective, one might prefer outsourcing if the contractor is more energy-efficient in extracting the demanded information, e.g., because of their prior experience and specialization or more energy-efficient infrastructure. At a global scale, outsourcing data analysis can involve less computation and save energy. But even if outsourcing is not an option, the data scientist might still leverage empirical and theoretical knowledge made available by others.

Existing theory on data mining algorithms provides some, though limited, guidance. Researchers have discussed models that describe the energy efficiency of systems and algorithms from different perspectives, such as power management (Albers 2010), energy per low-level operation (e.g., low-level operations per Watt (Hsu et al. 2005), or involving hardware components such as CPUs and memory (Roy et al. 2013). However, data scientists usually work on a higher level of abstraction than considering individual hardware components and low-level CPU instructions. Theoretical computer science analyzes algorithms in terms of running time: the count of abstract, higher-level operations needed to solve a task. The notion of time complexity can be applied to a single computer and a cluster of computers, taking into account costs for communication and idling (waiting for work). For many data mining models, the number of parameters (i.e., model capacity) correlates with time complexity and, thus energy consumption. Theoretical bounds might be coarse and derived for worst-case behavior limiting their practical relevance.

It is generally preferable to adjust existing models rather than to develop models from scratch. The data scientist can reuse knowledge or parts from existing models trained for a specific task using a technique called "transfer learning" (Lu et al. 2015; Pan and Yang 2010). For example, deep learning networks can benefit from reusing parameters or layers of an already-trained network (Bengio et al. 2014) to reduce time (and energy consumption) spent on developing a new model. Selecting an initial model that likely performs well based on performance metrics such as accuracy can be done, for instance, based on academic publications (Caruana and Niculescu-Mizil 2006).

To the best of our knowledge, a thorough comparison of learning algorithms for model parameters concerning energy-related concerns does not exist. Some works provide empirical results for running time of a few models, such as in the field of density-based clustering (Schneider and Vlachos 2017). Running time seems to be a viable surrogate metric for measuring energy consumption.

4.3 Design Principle of Support

Design Principle #5 – Share data, models, and skills

A data scientist should support fellow data scientists in saving energy, e.g., by sharing (cleaned) data, fitted models, code of analysis and algorithms, experience, and learnings.

Progress in the field of data science relies on publicly available data, models, and development frameworks. This principle applies to all phases of the CRISP-DM process. Initiatives to make data available by research institutions (Deng et al. 2009) and governments help create entire ecosystems (Najafabadi and Luna-Reyes 2017). The green data scientist should also contribute data, models, and extensions to frameworks to encourage reuse.

A data scientist can also share her experience to prevent others from learning through energy-consuming trial-and-error approaches. Commonly, data scientists share knowledge on dedicated Internet platforms for Q&A, and reputation schemes can incentivize this behavior (Khansa et al. 2015).

5 Evaluation

This section describes our evaluation strategy and provides empirical evidence for the design principles.

5.1 Evaluation Case Description

We use two scenarios covering two different types of data and models to evaluate the design principles – image and text classification. Computer vision and text mining tasks are popular in academia because of their accessibility as well as high practical relevance. For instance, medical applications include image classification to identify health issues on MRI scans, applications in manufacturing include assessing the quality of produced parts, and autonomous driving benefits from identifying obstacles. Furthermore, applying text and image classification to real-world problems has gained considerable popularity in recent years because of much better performance through deep learning approaches. The machine learning community typically treats problems from the perspective that the dataset is given, and the goal is to improve performance by finding (or tuning) the best-performing model. This view is highly relevant in practice (and in our evaluation). On its own, it is too narrow in two ways: (i) it focuses only on optimization towards performance, neglecting energy consumption; (ii) it discusses only the modeling part in CRISP-DM, and only for a single iteration – e.g., measures related to long-term impact on data collection or model deployment are neglected. Therefore, we consider model-related design options and more general design parameters also covering data quantity and data quality, thereby highlighting trade-offs between performance and energy consumption as well as strategies that might not impact performance at all but help in preserving energy. We propose and use different evaluation strategies for our design principles (Table 3).

Green data mining often involves a tradeoff between performance and energy costs. We state the objective to optimize as:

> Minimize energy consumption while achieving at least a minimum performance level.

This objective demands a minimum performance threshold but does not necessitate improvements beyond that threshold. Labor costs involved in creating a more environmentally friendly setup are a secondary concern. The above objective is adequate for initial feasibility studies or competitive assessments to answer questions such as "Can we build a smart service that is better than existing services?" It is also suitable for models that implement services based on Service Level Agreements (SLA), where the goal is to avoid penalties due to the violation of a minimum guaranteed performance level.

Next, we describe the evaluation setup, including data and models, and then detail the design principles' evaluation according to Table 3.

5.1.1 Data

We used images from a public dataset, the (tiny)-ImageNet.[3] This data set consists of images of 64 × 64 pixels. We used 27,500 images from 50 classes, i.e., 22,500 for training and 5000 for testing. We used the IMDB (Maas et al. 2011) dataset for text data, which consists of 50,000 movie reviews and sentiment ratings (positive or negative). We used 40,000 samples for training and the rest for testing. Figure 4 illustrates the two types of data involved in the two cases – image data (Case A) and text (Case B).

5.1.2 Models and Training

For both tasks, we employed deep learning models for two reasons. First, they are the best-performing models for these tasks. Second, they are also known to involve high energy consumption. We decided to use the smallest and simplest networks that should allow illustrating general behavior. For image recognition, we used a visual geometry group VGG-style network (Simonyan and Zisserman 2014) with eleven layers (VGG-11) and a single dense classification layer. We used the stochastic gradient descent optimizer for 100 epochs with a common learning rate schedule, i.e., decaying the initial rate of 0.1 by a factor of

[3] https://www.kaggle.com/competitions/tiny-imagenet.

Table 3 Evaluation strategies

Design principles	Evaluation strategy	Description/justification
1.1, 1.2, 2.1, 2.2 (Reduce)	Assess energy-performance tradeoffs through empirical investigation	Understanding the impact of tradeoffs among energy use and data as well as model-related performance metrics helps with assessing strategic questions that require a holistic understanding, such as: "Is it worth spending energy on model tuning, or should more data be collected?" They allow to predict the outcome of measures for performance improvement and, therefore, enable selection of measures without energy-consuming trial and error
1.3 (Reduce)	Leverage domain understanding and basic technical understanding	We demonstrate how to reduce energy costs by a more concise specification of attribute types and their ranges
3 (Reuse)	Improve reuse of computational results through overall process inspection	Model optimization is a repetitive process with many operations being performed anew in each iteration. We assess the benefits and effort of reusing partial results. This resorts to understanding the processing of data and model optimization on a higher level to identify repeated applications of the same operations and assess whether it is beneficial to store them for reuse
4 (Reuse)	Use academic and non-academic literature covering theoretical and empirical findings	We use existing empirical works to narrow the choice of hyperparameters related to network architecture and optimization procedures that might otherwise require energy due to experimentation
5 (Support)	Assess sharing of methods and insights with others	Supporting others incurs effort, which poses a significant obstacle to its application. We assess the effort to support others applying two commonly used platforms for Q&A and sharing code and models

10 after 50% and again after 75% of all training epochs. For classifying the textual data, we used a network provided by Kim (2014). We used the Adam optimizer with default settings. For hyperparameter optimization we used the Tree of Parzen Estimators as available through Python's Hyperopt library.

5.1.3 Hard- and Software Setup

For the software implementation, we used Google's open-source framework TensorFlow (Version 1.15). Our hardware consisted of a server with 64 GB of RAM, 16 cores of an AMD Threadripper 2950X CPU, and an NVIDIA RTX TI 2080 GPU.

5.1.4 Metrics for Empirical Evaluation

Our model performance metric is *accuracy*. To measure *energy consumption,* we considered static and variable parts. Static energy consumption of a computer is caused by, e.g., memory (RAM) as well as the minimum level of operation of components (mainboard, CPU, GPU). The variable part depends on I/O operations (hard drive usage, network communication) and computation (CPU, GPU, and cooling). The model we used to measure power consumption hence consists of a variable part and a fixed part. It is given by:

$$\text{used power [Wh]} = \text{static power [W]} \cdot \text{run time [h]} + \Delta t\,[\text{h}] \cdot \sum_i \text{variable power at time } (i \cdot \Delta t)\,[\text{W}]$$

For deep learning in general and in our setup, the GPU is the most energy-consuming component – the GPU we used can draw up to 260 W.[4] It also exhibits most fluctuation across models with different configurations and changes in data. It should thus be considered variable. We measure GPU power consumption in Watts using NVIDIA's "smi" tool. The tool allows for a sampling interval of about 100 ms. Our AMD CPU with 32 threads remained largely idle: only one thread was fully (100%) used for text data and only two threads for the image data analysis. The CPU requires about 30 W in idle mode and about 55 W in comparable settings to ours.[5] Our mainboard and RAM add another 45 W of static power consumption. The SSD hard drive uses about 1 W for our read-dominated usage. Any operating system controlling the computer's hardware also consumes power. We verified that no process caused a CPU or GPU usage of more than 1% by investigating running processes before and after running data mining algorithms. We used the server exclusively for performing measurements related to this study.

5.2 Evaluation of Design Principles #1.1, #1.2, #2.1, #2.2 Using Optimization

First, we describe means to obtain information on tradeoffs between energy usage and model performance for each

[4] https://www.nvidia.com/en-me/geforce/graphics-cards/rtx-2080-ti/.

[5] https://www.tomshardware.com/reviews/amd-ryzen-threadripper-2-2990wx-2950x,5725-13.html.

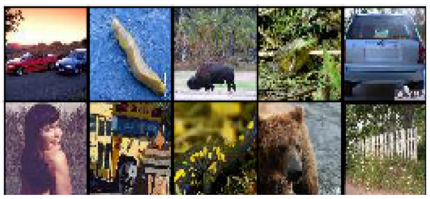

Fig. 4 Image data (Case A) and text (Case B)

principle. Second, we discuss how data scientists can assess these tradeoffs when developing (i) strategies for data collection and processing and (ii) model optimization. The idea is to collect measurements to understand tradeoffs computed using energy-efficient configurations. In turn, these measurements are used to predict model performance and energy consumption for less energy-efficient (but more performant) configurations.

To assess the impact of an energy-relevant design parameter, we vary one parameter and evaluate its effects on outcome metrics. We use these measurements to predict behavior for non-observed parameter values and to decide on which parameters to alter. For instance, decisions might be between collecting more data or improving data quality.

The conducted measurements themselves are specified in a way to keep energy consumption small. For example, to predict the value of collecting more data, we measure performance when using ½, ¼, 1/8 of all available data (rather than all data). We then predict an estimate of performance when using all data and even twice as much data as available. While we vary one parameter, the other parameters remain fixed. We chose parameter values that result in reduced energy consumption but are not expected to give the best performance. That is, only 50% of the data samples are used. This contributes to energy reduction. But it increases inaccuracy in the estimates of model behavior when a parameter is varied. However, since we are primarily interested in a coarse understanding, some level of inaccuracy is tolerable. Aside from altering parameters, one might also opt for using a faster optimizer that might not result in the best generalization behavior or discarding data augmentation, which also requires computation and prolongs convergence.

Our proposed procedure relies on the assumption that model parameters exhibit predictable behavior and that many (hyper)parameters are independent (Bergstra, and Bengio 2012). In machine learning, the amount of data and model size (as described by hyperparameters) exhibit diminishing returns (Figueroa et al. 2012).

Figures 5, 6 and 7 visualize tradeoffs. Each data point is the mean of five runs, with standard deviations indicated by shaded areas. Note that standard deviations are often small, i.e., barely visible. For each of these plots, we varied a parameter related to energy consumption, i.e., related to data (number of attributes and number of samples) and training (number of training epochs). Figures 5, 6 and 7 show that all design parameters exhibit diminishing marginal utility on performance while (often) showing approximately linear growth in energy consumption. It is even possible that the utility of a further increase in a parameter yields adverse outcomes. For instance, training for too long can lead to overfitting of the data.

For the reduction of the number of attributes (Design Principle #1.1 – Fig. 5), we can observe that energy consumption increases in parts non-linearly when we change the number of words. We kept the most common words to address the task of text classification. The distribution of words follows a power-law distribution, meaning that a few words are very common, and the vast number of words is rare. Thus, increasing the number of words from 2000 to 4000 has a larger impact than increasing them from 10,000 to 12,000 because the former reduces the overall dataset size more than the latter.

Next, we leveraged tradeoffs for strategic considerations, i.e., to answer questions like "Is it worth investing a lot of energy into model tuning, or should one collect more data or improve the data quality?" Let us assume that the data scientist decides to further improve the initial models (We shall focus on image recognition in the discussion). Figures 5, 6 and 7 indicate that a 20% or more increase in performance is unlikely by varying any of the depicted parameters. We can see this through visual extrapolation in Figs. 5, 6 and 7 or, more formally, by fitting a model using the collected performance measurements. In this case, the data scientist might not even try to improve the model by hyperparameter tuning but instead fundamentally alter the model or input data. For medium improvements of about 5–10%, neither higher image resolution nor longer training time are likely to be sufficient. Additional data samples yield clear improvements. Therefore, data related techniques for model improvement might be highly beneficial, including collecting more data, data augmentation, or using a pre-trained model on a similar dataset (transfer learning). For minor improvements of 1–3%, both increasing training data and image resolution are feasible strategies. Higher image resolution also implies higher energy consumption during system operation, while increasing the number of training samples does not impact energy costs during operation. Doubling data roughly doubles the cost of energy spent on training (Fig. 6). Data collection might

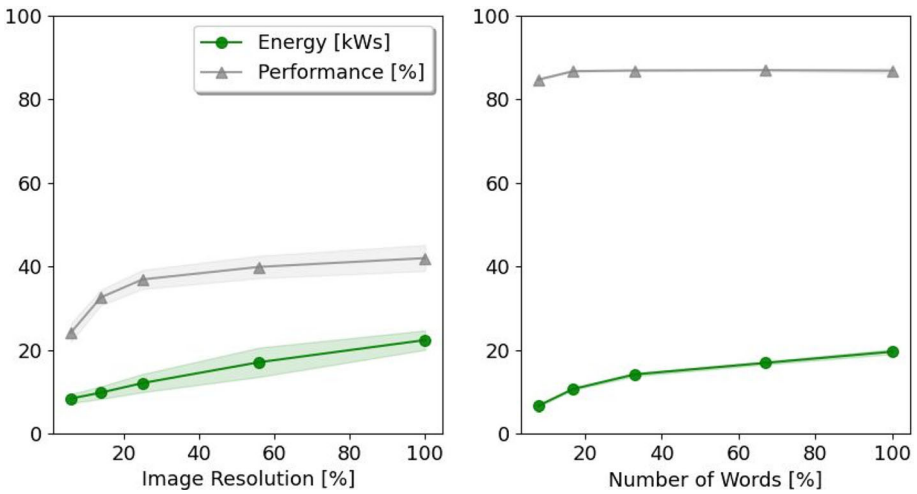

Fig. 5 Impact of attribute reduction on energy and model performance

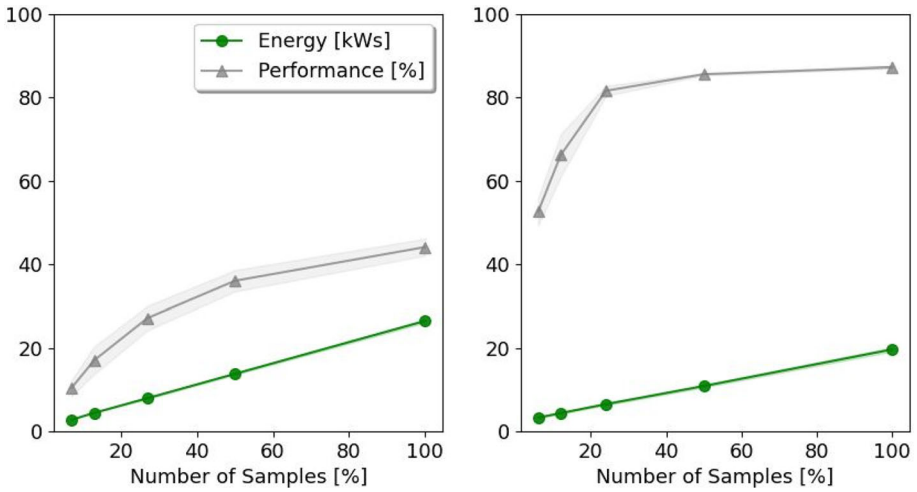

Fig. 6 Impact of samples reduction on energy and model performance

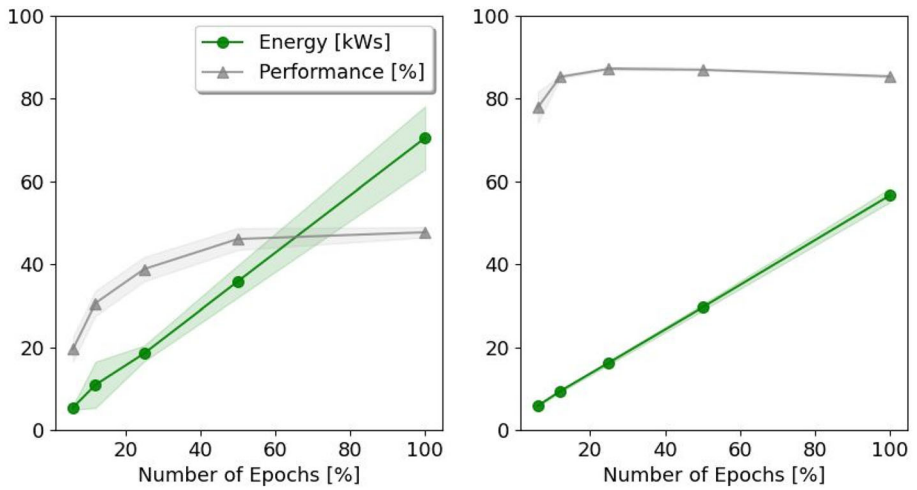

Fig. 7 Impact of training time reduction on energy and model performance

also incur energy costs. Doubling image resolution leads to an energy increase that is slightly less than double because some parts of the model (upper layers, to be precise) require the same amount of computation irrespective of the input size.

5.3 Evaluation of Design Principle #1.3

We also assessed Principle #1.3 for the specification of attributes in the context of our evaluation cases. For images, one might reduce color depth. For example, reducing color depth from 24 bits ($\sim$ 4 Mio. Colors) to 16 bits ($\sim$ 65 k colors) saves about 1/3 of storage capacity with little impact otherwise. Furthermore, common GPUs like newer models of NVIDIA allow for training with reduced precision of numeric value without compromising final accuracy. For example, reducing from 32 to 16 bits precision is claimed to half computation time for deep learning models without negative side effects on model performance.[6] Public repositories often do not leverage this option of reducing precision. This is surprising since it can be implemented with less than 5 lines of code, i.e., within a matter of minutes. While for image data, we could achieve an energy reduction of about 45% for both training and evaluation, for textual data, it was only about 25% for training but 60% for evaluation. Differences were confirmed to be statistically significant (p-value < 0.01) for both datasets by using a t-test based on 10 samples, i.e., energy measurements of 10 models trained with and without mixed precision. Our evaluation confirmed that there is no degradation in performance, i.e., a one-tailed t-test yielded p-values larger 0.2.

5.4 Evaluation of Design Principle #3

We repeatedly down sampled the data in our setup, e.g., from 64 $\times$ 64 pixels image size to smaller sizes. We also frequently performed the same pre-processing on textual data. The simplest implementation applies these operations in every epoch anew for a total of about 50–200 times per training a neural network. Storing the data once rather than recomputing it can have a significant impact. We found that for images resized to 32 $\times$ 32 pixels, energy savings are between 0.1 and 0.5% of the overall energy spent on training the network. One can implement storing and loading data in a few lines of code.

Other ideas require significantly more effort to implement. For instance, during training with our image data, a sample is modified for every usage due to data augmentation. That is, an image might be slightly rotated, shifted, etc. It might be possible to use the same (modified) sample more than once before creating another modification. However, this is not readily supported by our framework (Tensorflow). Thus, it requires more implementation effort.

[6] https://docs.nvidia.com/deeplearning/performance/mixed-precision-training/index.html.

5.5 Evaluation of Design Principle #4

Identifying the best model requires training a model with many different hyperparameter configurations. We select these configurations using a hyper-optimization algorithm. This still leaves the burden of defining adequate ranges for hyperparameters with a data scientist. A data scientist without adequate knowledge must specify large ranges for each parameter, including values that could be excluded by a more knowledgeable data scientist. Large ranges lead to a large search space and training of models with hyperparameter configurations, which yield poor outcomes. We assessed reducing computation by narrowing hyperparameter ranges based on existing empirical works. Our evaluation focused on two hyperparameters related to model regularization. Regularization is commonly done since it often yields significant accuracy gains. We considered dropout and L2 regularization, which ensure that deep learning models do not make decision based on few decision criteria by setting the outputs of some neurons to 0 (dropout) or by penalizing large weights, indicating strong dependence on individual characteristics (L2 regularization). The two hyperparameters, i.e., dropout rate and the (L2) regularization constant, are lower bounded by 0. The upper bound for dropout is 1, and for the L2 constant a conservative bound is also 1. Thus, a data scientist might simply use a range of [0,1] for both hyperparameters. However, by investigating existing literature, she might arrive at more narrow bounds that are very likely to contain the best hyperparameters. We gained insights from the original paper on VGG (Simonyan and Zisserman 2014) for image classification and from the paper of Kim (2014) for text classification. Simonyan and Zisserman (2014) stated for the L2 constant a value of 0.0005 and a dropout rate of 0.5, whereas Kim (2014) only used dropout with the same rate of 0.5. An Internet search for common dropout rates revealed that they do not seem to be above 0.7. Thus, we narrowed the range for dropout to [0,0.8]. For the L2 constant we assumed a factor of at most 100 difference from the original, i.e., we chose a range of [0,0.05]. But we have no evidence that the narrow hyperparameter ranges still contain the optimal values. If they do not, the returned hyperparameters would most certainly be close to one of the limits of the intervals for a parameter. Thus, a data scientist could simply rerun the optimization with a larger range. This demonstrates that there is little risk involved in achieving an inferior model, but too aggressive narrowing of hyperparameter ranges might yield higher energy consumption overall.

The hyperparameter optimization algorithm was configured to train 150 models with hyperparameters chosen automatically for each trained model from the provided ranges. The accuracy of the 150 models is shown in Fig. 8

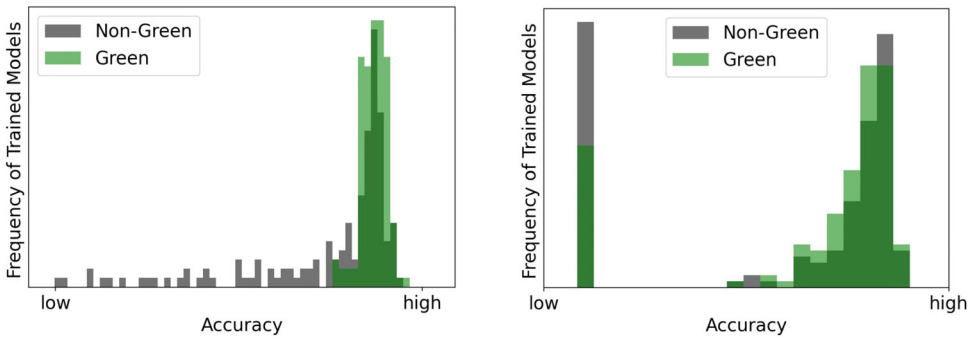

Fig. 8 Impact of Hyperparameter Ranges (left panel: image data; right: text): "Non-Green" ranges yield more trained models with poor performance

for both datasets in the form of a histogram. It becomes apparent that using large ranges containing unreasonable values leads much more often to hyperparameter settings that yield poor performance. In particular, for text classification strong regularization tends to render models useless, i.e., no better than guessing. Thus, the training of these poorly performing models constitutes unnecessary or wasted effort that can be avoided by narrowing hyperparameter ranges.

5.6 Evaluation of Design Principle #5

We also performed a qualitative analysis. Following Principle #5, advocating sharing of code and models, we made some of our code publicly available.[7] While the creation of the public repository was a matter of minutes, we spent a few hours documenting, generalizing, and structuring the code to make it easier to use. While sharing a code is often a delicate issue due to intellectual property concerns and a lack of maintenance can be a problem, contributing to Q&A platforms such as StackOverflow to answer generic questions is of lesser concern. We replied to 11 questions related to data mining in 2020, which required a few working hours. We required on average more time than the reported median of 11 min in Mamykina et al. (2011). About 22,000 people viewed our replies. However, it is difficult to assess how much energy such shared knowledge saves. As a positive return for the respondents to questions, the platform provides the privilege to set bounties to users, which motivates others to answer the respondent's questions. Thus, the invested time might be beneficial beyond saving energy.

6 Discussion

The principles we developed are a form of nascent design knowledge (Gregor and Hevner 2013). Such knowledge is formal enough to be applicable across contexts so that researchers can further develop it into more abstract, mature, and complete design knowledge, e.g., in the form of a design theory for data mining. To move toward a comprehensive and integrated prescriptive framework for Green Data Mining, we will next discuss the interaction among principles and project goals.

6.1 Interaction among Principles and Project Goals

In data mining, several conflicting goals exist, resulting in tradeoffs such as costs vs. quality and completion time vs. costs (Kerzner and Kerzner 2017). Following Green Data Mining principles can positively affect project goals because the fastest way is often the least energy-consuming. Fast algorithms typically require less energy, are often simpler, and are more interpretable and transparent but sometimes approximate, leading to lower model performance. Model performance is a key goal for any data mining project. Thus, it is likely that a data scientist cannot make compromises with respect to model performance for the sake of energy savings due to project objectives set by business stakeholders. Regulatory intervention or more pressure from customers is likely needed to change this. Similarly, sharing knowledge is generally not within the company's interest. Furthermore, while energy costs due to computation seem relatively easy to measure, overall energy costs, including costs due to lowered model performance as well as the energy gains due to sharing of knowledge, can be challenging to estimate. Furthermore, following green data mining increases project complexity for a data scientist. Thus, it might lead to higher labor costs, prolonged projects, and an increase in the risk of errors. For example, the principle of reducing often relies on previous knowledge gathering from large experiments by using statistical analysis and estimates. For instance, model performance can be estimated on all data when training only on a subset of data. Statistical estimates come with a small risk of exceptionally large deviations.

In this light, we have critically reflected on our evaluations (see summary in Table 4) as examples of possible concrete implementations of our design principles within a

[7] https://github.com/MBasalla/Principles-of-Green-Data-Science/.

Table 4 Evaluation summary

Design principles	Summary of evaluation
Reduce	The evaluation demonstrated that understanding existing tradeoffs among performance metrics, input, and model parameters can improve model performance in a principled, energy-efficient manner by focusing on promising parameters and replacing trial and error through a more informed approach. Tradeoffs can be assessed in an energy-efficient manner by reducing all energy contributing parameters to the extent that still suffices to make reliable predictions. Attribute specification is arguably the easiest strategy to implement, with considerable potential for energy saving
Reuse and support	Reuse of results that are computed multiple times was shown to save energy with little effort. (Re)using theoretical insights is often not possible since they are scarce. Empirical knowledge can be helpful to limit the parameter search space and start from promising model architectures. While sharing insights is technically easy, it also requires time for a data scientist, and it might not be in the interest of a company

company in terms of energy savings, impact on model performance, and costs to implement. We conclude that principles #1.1, #1.2, #2.1, #2.2, all aiming to "reduce," tend to lead to lower model performance. They are also non-trivial to implement, and energy cost savings are not guaranteed. In contrast, principles #1.3(reduce), #3(reuse), #4(reuse) tend not to impact model performance, and they take at most a few hours to implement with very likely, often even large energy savings.

Thus, some of our evaluations are of immediate practical value. But going beyond the proposals and finding new implementations of the principles is often far from trivial. This requires a mindset that focuses on energy saving. Data scientists must also think differently and more holistically to adapt the given principles. They must factor in energy costs and break with the prevalent paradigm that more data and computation are always better. They must anticipate the implications of project requirements for models as well as what happens beyond model deployment, e.g., to understand the implications of model errors for energy costs as illustrated in Fig. 3. Data scientists need to understand the impact of their implementation choices on energy consumption, which is rather uncommon in data mining. They must also be able to apply a range of evolving techniques to implement the principles suggested in this paper, ranging from active learning to hyperparameter optimization.

To move Green Data Mining forward we envision an EMIS, i.e., a recommendation engine supporting a data scientist, e.g., by suggesting energy-efficient models and data processing analogous as done for data structures in software engineering (Oliveira et al. 2021). An EMIS might also support transparency of models' energy consumption to raise awareness as EnergyVis (Shaikh et al. 2021).

6.2 Contribution to Green IS / Green IT Literature

Developing prescriptive knowledge, next to explanatory knowledge, is a key obligation in the IS field in general (Gregor and Hevner 2013) and for Green IS scholarship in particular, as environmental degradation requires concrete solutions to remedy the problem (Seidel et al. 2018). With this paper, we contribute to the emergent stream of literature on prescriptive knowledge of Green IS. We contend that *Green Data Science* has two key aspects that fall into the categories of first and second order effects (Hilty et al. 2011). First, Green Data Science needs to consider the environmental impact of the involved resources (first order effects). Second, Green Data Science can help design more environmentally sustainable processes (second order effects). Grounded in our analysis of Green Data Mining, we define the broader area of Green Data Science as follows:

> Green Data Science is the application of methods to extract knowledge from data that (a) seeks to minimize the environmental impact of this activity and (b) aims to leverage this knowledge to implement more environmentally sustainable business processes and afford more environmentally sensible decision making.

Our study has some limitations. First, we are addressing one specific area of data science – data mining. Still, we contend that data mining involves a set of key data science practices. Future research can thus seek to extend our principles towards a more general understanding of data science. Our principles can also be enhanced beyond the scope of a data scientist. For example, a company might provision computing infrastructure to data scientists only when green energy (e.g., from solar panels or wind) is amply available (Niu et al. 2016), or a company might introduce incentive mechanisms for being green.

Deriving principles from literature also has its limitations since it does not allow to foresee or propose principles that cannot be derived from considered literature. Thus, our principles are likely non-exhaustive. For example, we did not include principles such as "design for maintainability and longevity of a model" since there was

no literature with content specific to data mining but only for software.

We conducted the evaluation of the principles based on only two cases. Still, it is important to highlight that we grounded the principles on findings from previous literature that have been empirically tested. We thus build upon previous works. Further, we have designed a technical setup, evaluated principles empirically, and provided some learning based on our analysis. We have thus prepared the ground for more research on evaluating, refining, and extending our set of design principles.

The field of data mining is quickly evolving. Thus, while we believe that our principles are generic enough to be of long-term value, details of our technical evaluation are likely to become outdated within a few years.

Our work is limited to those actions a green data scientist can take during her work. As such, it does not deal with all possible aspects that can be supportive of Green Data Mining but are not part of the daily routine of the green data scientist. For instance, it does not consider the development of tools facilitating the implementation of our principles. Besides, our focus is on the data scientist rather than organizational aspects, such as incentivization and adoption barriers of Green Data Mining practices. Our work provides a foundation for future studies using empirical cases involving data mining practitioners that can refine, further develop, and deepen our findings.

7 Conclusion

Environmental sustainability, and climate change in particular, are key societal challenges of our time. As we live in the digital age where more and more data are available, it is crucial to understand how one can design the processes involved in collecting, storing, processing, and extracting knowledge from this data in an environmentally responsible way and use this data to make better, more environmentally friendly decisions. While the second field yields greater potential, we deem it necessary to attend to immediate impacts of data science, which is where the present study contributes.

Funding Open access funding provided by University of Liechtenstein.

References

Aggarwal CC (2015) Data mining: the textbook. Springer, Berlin. https://doi.org/10.1007/978-3-319-14142-8

Albers S (2010) Energy-efficient algorithms. Commun ACM 53(5):86–96. https://doi.org/10.1145/1735223.1735245

Amershi S, Begel A, Bird C et al (2019) Software engineering for machine learning: a case study. In: International conference on software engineering: software engineering in practice, pp 291–300. https://doi.org/10.1109/ICSE-SEIP.2019.00042

Bengio Y (2012) Deep learning of representations for unsupervised and transfer learning. In: Proceedings of ICML workshop on unsupervised and transfer learning, pp 17–30

Bergstra J, Bengio Y (2012) Random search for hyper-parameter optimization. J Mach Learn Res 13:281–305

Bilal K, Malik SUR, Khalid O et al (2014) A taxonomy and survey on Green Data center networks. Fut Gen Comput Syst 36:189–208

Brooks S, Wang X, Sarker S (2010) Unpacking green IT: a review of the existing literature. In: 16th Americas Conference on Information Systems, Lima, pp 749–759. https://aisel.aisnet.org/amcis2010/398

Calero C, Piattini M (eds) (2015) Green in software engineering, vol 3. Springer, Heidelberg

Capra E, Merlo F (2009) Green IT: everything starts from the software. In: European conference of information systems, pp 62–73

Caruana R, Niculescu-Mizil A (2006) An empirical comparison of supervised learning algorithms. In: International conference on machine learning. https://doi.org/10.1145/1143844.1143865

Chen Q, Grosso P, van der Veldt K, de Laat C, Hofman R, Bal H (2011) Profiling energy consumption of VMs for green cloud computing. In: International conference on dependable, autonomic and secure computing. https://doi.org/10.1109/DASC.2011.131

Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L (2009) Imagenet: a large-scale hierarchical image database. In: IEEE conference on computer vision and pattern recognition, Miami, pp 248–255. https://doi.org/10.1109/CVPR.2009.5206848

El-Gayar O, Fritz BD (2006) Environmental management information systems (EMIS) for sustainable development: a conceptual overview. Commun Assoc Inf Syst. https://doi.org/10.17705/1CAIS.01734

Elmasri R, Navathe S (2010) Fundamentals of database systems. Addison-Wesley

Fedorov I, Adams RP, Mattina M, Whatmough PN (2019) SpArSe: sparse architecture search for CNNs on resource-constrained microcontrollers. Adv Neur Inf Proc Syst 32:4977–4989

Figueroa RL, Zeng-Treitler Q, Kandula S, Ngo LH (2012) Predicting sample size required for classification performance. BMC Med Inform Decis Making 12(1):8. https://doi.org/10.1186/1472-6947-12-8

Gelenbe E, Caseau Y (2015) The impact of information technology on energy consumption and carbon emissions. Ubiquity 2015:1

Gholami R, Watson RT, Hasan H, Molla A, Bjorn-Andersen N (2016) Information systems solutions for environmental sustainability: how can we do more? J Assoc Inf Syst 17(8):2

Ghorbani A, Zou J (2019) Data shapley: equitable valuation of data for machine learning. In: International conference on machine learning, pp 2242–2251
Goel A, Tung C, Lu YH, Thiruvathukal GK (2020) A survey of methods for low-power deep learning and computer vision. In: World Forum on Internet of Things, pp 1–6
Goiri Í, Le K, Nguyen TD, Guitart J, Torres J, Bianchini R (2012) GreenHadoop: leveraging green energy in data-processing frameworks. In: ACM European conference on computer systems. https://doi.org/10.1145/2168836.2168843
Goodland R (1995) The concept of environmental sustainability. Ann Rev Ecol Syst 26:1–24. https://doi.org/10.1146/annurev.es.26.110195.000245
Gregor S (2006) The nature of theory in information systems. MIS Q 30(3):611–642. https://doi.org/10.2307/25148742
Gregor S, Hevner AR (2013) Positioning and presenting design science research for maximum impact. MIS Q 37(2):337–355
Gregor S, Jones D (2007) The anatomy of a design theory. J Assoc Inf Syst 8(5):313–335
Gregor S, Chandra Kruse L, Seidel S (2020) The anatomy of a design principle. J Assoc Inf Syst
Gupta S, Agrawal A, Gopalakrishnan K, Narayanan P (2015) Deep learning with limited numerical precision. In: International conference on machine learning, pp 1737–1746
Han J, Pei J, Kamber M (2011) Data mining: concepts and techniques. Elsevier
Han S, Liu X, Mao H, Pu J, Pedram A, Horowitz MA, Dally WJ (2016) EIE: efficient inference engine on compressed deep neural network. In: International symposium on computer architecture. https://doi.org/10.1145/3007787.3001163
He X, Zhao K, Chu X (2019) AutoML: a survey of the state-of-the-art. Knowl-Based Syst 212:106622. https://doi.org/10.1016/j.knosys.2020.106622
Hedman J, Henningsson S (2016) Developing ecological sustainability: a green IS response model. Inf Syst J 26(3):259–287. https://doi.org/10.1111/isj.12095
Hevner AR, March ST, Park J, Ram S (2004) Design science in information systems research. MIS Q 28(1):75–105
Hilpert H, Kranz J, Schumann M (2013) Leveraging green IS in logistics. Bus Inf Syst Eng 5(5):315–325. https://doi.org/10.1007/s12599-013-0285-1
Hilty L, Lohmann W, Huang E (2011) Sustainability and ICT – an overview of the field. Politeia 27(104):13–28
Hindle A (2016) Green software engineering: the curse of methodology. In: International conference on software analysis, evolution, and reengineering. https://doi.org/10.1109/SANER.2016.60
Hsu C-H, Feng W-C, Archuleta JS (2005) Towards efficient supercomputing: a quest for the right metric. In: International parallel and distributed processing symposium. https://doi.org/10.1109/IPDPS.2005.440
Hubara I, Courbariaux M, Soudry D, El-Yaniv R, Bengio Y (2016) Binarized neural networks. In: The advances in neural information processing systems, pp 4114–4122
Jin C, de Supinski BR, Abramson D et al (2017) A survey on software methods to improve the energy efficiency of parallel computing. Int J High Perf Comput Appl 31(6):517–549
Kerzner H, Kerzner HR (2017) Project management: a systems approach to planning, scheduling, and controlling. Wiley
Khansa L, Ma X, Liginlal D, Kim SS (2015) Understanding members' active participation in online question-and-answer communities: a theory and empirical analysis. J Manag Inf Syst 32(2):162–203
King WR, He J (2005) Understanding the role and methods of meta-analysis in IS research. Commun Assoc Inf Syst 16(1):32
Kumar M, Zhang X, Liu L, Wang Y, Shi W (2020) Energy-efficient machine learning on the edges. In: Parallel and distributed processing symposium workshops. https://doi.org/10.1109/IPDPSW50202.2020.00153
Kurgan LA, Musilek P (2006) A survey of knowledge discovery and data mining process models. Knowl Eng Rev 21(1):1–24
Kwon H, Chatarasi P, Pellauer M, Parashar A, Sarkar V, Krishna T (2019) Understanding reuse, performance, and hardware cost of dnn dataflow: a data-centric approach. In: International symposium on microarchitecture, pp 754–768
Li H, Kadav A, Durdanovic I, Samet H, Graf HP (2017) Pruning filters for efficient convnets. In: International conference on learning representations. arXiv:1608.08710
Loock C-M, Staake T, Thiesse F (2013) Motivating energy-efficient behavior with Green IS: an investigation of goal setting and the role of defaults. MIS Q 37(4):1313–1332
Löser F, Recker J, vom Brocke J, Molla A, Zarnekow R (2016) How IT executives create organizational benefits by translating environmental strategies into Green IS initiatives. Inf Syst J 27(4):503–553
Löser F (2013) Green IT and Green IS: definition of constructs and overview of current practices. In: Americas conference on information systems. https://doi.org/10.13140/2.1.3065.6962
Lu J, Behbood V, Hao P, Zuo H, Xue S, Zhang G (2015) Transfer learning using computational intelligence: a survey. Knowl-Based Syst 80:14–23
Luo G (2016) A review of automatic selection methods for machine learning algorithms and hyper-parameter values. Netw Model Anal Health Inform Bioinforma. https://doi.org/10.1007/s13721-016-0125-6
Malhotra A, Melville N, Watson RT (2013) Spurring impactful research on information systems for environmental sustainability. MIS Q 37(4):1265–1274
Mamykina L, Manoim B, Mittal M, Hripcsak G, Hartmann B (2011) Design lessons from the fastest Q&A site in the west. In: SIGCHI conference on human factors in computing systems. https://doi.org/10.1145/1978942.1979366
Manning C, Raghavan P, Schütze H (2010) Introduction to information retrieval. Nat Lang Eng 16(1):100–103
Masanet E, Shehabi A, Lei N, Smith S, Koomey J (2020) Recalibrating global data center energy-use estimates. Sci 367(6481):984–986
Melville NP (2010) Information systems innovation for environmental sustainability. MIS Q 34(1):1–21
Meske C, Bunde E, Schneider J, Gersch M (2021) Explainable artificial intelligence: objectives, stakeholders, and future research opportunities. Inf Syst Eng. https://doi.org/10.1080/10580530.2020.1849465
Molla A, Cooper V (2014) Greening data centres: the motivation, expectancy and ability drivers. In: European conference of information systems
Murugesan S (2008) Harnessing green IT: principles and practices. IT Prof 10(1)
Najafabadi M, Luna-Reyes L (2017) Open government data ecosystems: a closed-loop perspective. In: Hawaii international conference on system sciences. https://doi.org/10.24251/HICSS.2017.327
Niu Z, He B, Liu F (2016) Not all joules are equal: towards energy-efficient and green-aware data processing frameworks. In: International conference on cloud engineering. https://doi.org/10.1109/IC2E.2016.17
Oliveira W, Oliveira R, Castor F, Pinto G, Fernandes JP (2021) Improving energy-efficiency by recommending Java collections. Emp Softw Eng 26(3):1–45
Pan SJ, Yang Q (2010) A survey on transfer learning. IEEE Trans Knowl Data Eng 22(10):1345–1359

Peffers K, Tuunanen T, Rothenberger MA, Chatterjee S (2007) A design science research methodology for information systems research. J Manag Inf Syst 24(3):45–77

Provost F, Fawcett T (2013) Data science for business: what you need to know about data mining and data-analytic thinking. O'Reilly Media

Ramírez-Gallego S, Krawczyk B, García S, Woźniak M, Herrera F (2017) A survey on data preprocessing for data stream mining: current status and future directions. Neurocomput 239:39–57

Roy S, Rudra A, Verma A (2013) An energy complexity model for algorithms. In: Conference on innovations in theoretical computer science. https://doi.org/10.1145/2422436.2422470

Sayood K (2017) Introduction to data compression. Morgan Kaufmann

Schneider J, Vlachos M (2017) Scalable density-based clustering with quality guarantees using random projections. Data Min Knowl Discov 31(4):972–1005

Seidel S, Recker J, vom Brocke J (2013) Sensemaking and sustainable practicing: functional affordances of information systems in green transformations. MIS Q 37(4):1275–1299

Seidel S, Bharati P, Fridgen G et al (2017) The sustainability imperative in information systems research. Commun Assoc Inf Syst 40(1):3

Seidel S, Chandra Kruse L, Szekely N, Gau M, Stieger D (2018) Design principles for sensemaking support systems in environmental sustainability transformations. Eur J Inf Syst 27(2):221–247

Shaikh O, Saad-Falcon J, Wright AP, Das N, Freitas S, Asensio O, Chau DH (2021) EnergyVis: interactively tracking and exploring energy consumption for ML models. In: Conference on human factors in computing systems, pp 1–7

Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556

Snoek J, Larochelle H, Adams RP (2012) Practical Bayesian optimization of machine learning algorithms. In: Adv Neur Inf Proc Syst. arXiv:1206.2944

Snurr R, Freude (2021) Energy fundamentals – Daily energy needs. https://home.uni-leipzig.de/energy/energy-fundamentals/04.htm. Accessed 10 Oct 2021

Sorzano COS, Vargas J, Montano AP (2014) A survey of dimensionality reduction techniques. arXiv preprint arXiv:1403.2877

Stange M, Funk B (2015) How much tracking is necessary? The learning curve in Bayesian user journey analysis. In: European conference of information systems. https://doi.org/10.18151/7217484

Strubell E, Ganesh A, McCallum A (2020) Energy and policy considerations for modern deep learning research. In: Proceedings of the AAAI conference on artificial intelligence. https://doi.org/10.1609/aaai.v34i09.7123

Suhl L, Voß S (2014) An introduction to the special focus issue "Decision Analytics." Bus Inf Syst Eng 6(3):129. https://doi.org/10.1007/s12599-014-0324-6

Taina J (2011) Good, bad, and beautiful software. In search of green software quality factors. Cepis Upgrade 12(4):22–27

Thornton C, Hutter F, Hoos HH, Leyton-Brown K (2013) Auto-WEKA: combined selection and hyperparameter optimization of classification algorithms. In: International conference on knowledge discovery and data mining. https://doi.org/10.1145/2487575.2487629

Tsymbal A (2004) The problem of concept drift: definitions and related work. Computer Science Department, Trinity College Dublin 106(2):58

ur Rehman MH, Liew CS, Abbas A, Jayaraman PP, Wah TY, Khan SU (2016) Big data reduction methods: a survey. Data Sci Eng 1(4):265–284

van der Aalst WM (2016) Green data science: using big data in an environmentally friendly manner. In: International conference on enterprise information systems, pp 9–21

Vassiliadis P (2009) A survey of extract–transform–load technology. Int J Data Wareh Min 5(3):1–27

Walls JG, Widmeyer GR, El Sawy OA (1992) Building an information system design theory for vigilant EIS. Inf Syst Res 3(1):36–59. https://doi.org/10.1287/isre.3.1.36

Wang L, Khan SU (2013) Review of performance metrics for green data centers: a taxonomy study. J Supercomput 63(3):639–656

Watson RT, Boudreau M-C, Chen AJ (2010) Information systems and environmentally sustainable development: energy informatics and new directions for the IS community. MIS Q 34(1):23–38

Watson RT, Elliot S, Corbett J et al (2021) How the AIS can improve its contributions to the UN's sustainability development goals: towards a framework for scaling collaborations and evaluating impact. Commun Assoc Inf Syst 48(1):42

Watson RT, Boudreau M-C, Chen AJ, Huber M (2008) Green IS: building sustainable business practices. In: Watson RT (ed) Information systems: a global text, pp 247–261

Wiesche M, Jurisch MC, Yetton PW, Krcmar H (2017) Grounded theory methodology in information systems research. MIS Q 41(3):685–701

Wirth R, Hipp J (2000) CRISP-DM: towards a standard process model for data mining. In: Conference on the practical applications of knowledge discovery and data mining

Wolpert DH (1996) The lack of a priori distinctions between learning algorithms. Neur Comput 8(7):1341–1390

Yosinski J, Clune J, Bengio Y, Lipson H (2014) How transferable are features in deep neural networks? In: Neural information processing systems, pp 3320–3328