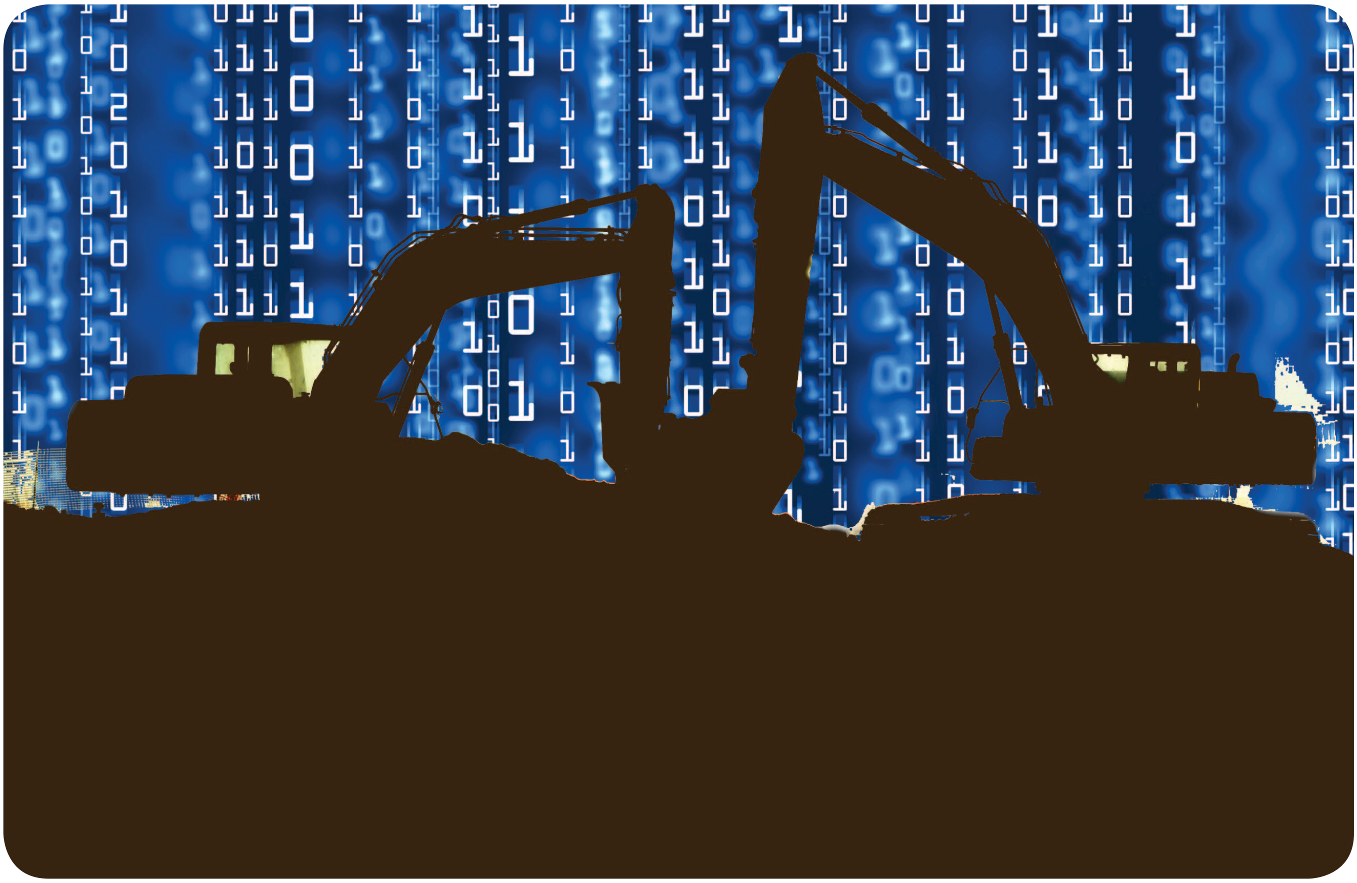
Data mining solutions for sustainability problems

Naren Ramakrishnan, Manish Marwah, Amip Shah, Debprakash Patnaik, M. Shahriar Hossain, Naren Sundaravaradan, and Chandrakant Patel



COURTESY OF STOCK.XCHNG/DAVE DYET. MINING COURTESY OF STOCK.XCHNG/SACHIN GHODKE

Nearly every aspect of modern life is laced with questions and choices regarding sustainability. Some questions are pervasive, e.g., should I print this *IEEE Potentials* article or should I read it online? Others are subtle and we might not think consciously about them, e.g., how much CO_2 does a Google search release into the atmosphere? Still others are knotty conundrums: how do we encourage and incentivize an entire city to "go green?"

Computational sustainability [Gomes (2009)] deals with answering questions such as the above using mathematical and algorithmic techniques. Its scope is broad: from designing environmentally friendly substitutes for everyday products, to reducing carbon emissions of data centers, to encouraging energy efficiency in homes, and finally to understanding the interplay between multiple systems at a societal level.

Many issues interplay in achieving sustainability goals. First, it is desirable to have an accurate model of the underlying process or product so that we can understand exactly where to focus our sustainability objectives. Second, we must systematically evaluate and assess alternatives alongside multiple (environmental and other) criteria. Finally, satisfactory implementation of sustainable alternatives requires a "buy-in" from all involved stakeholders.

Digital Object Identifier 10.1109/MPOT.2011.2181883
Date of publication: 31 October 2012

0278-6648/12/$31.00©2012IEEE

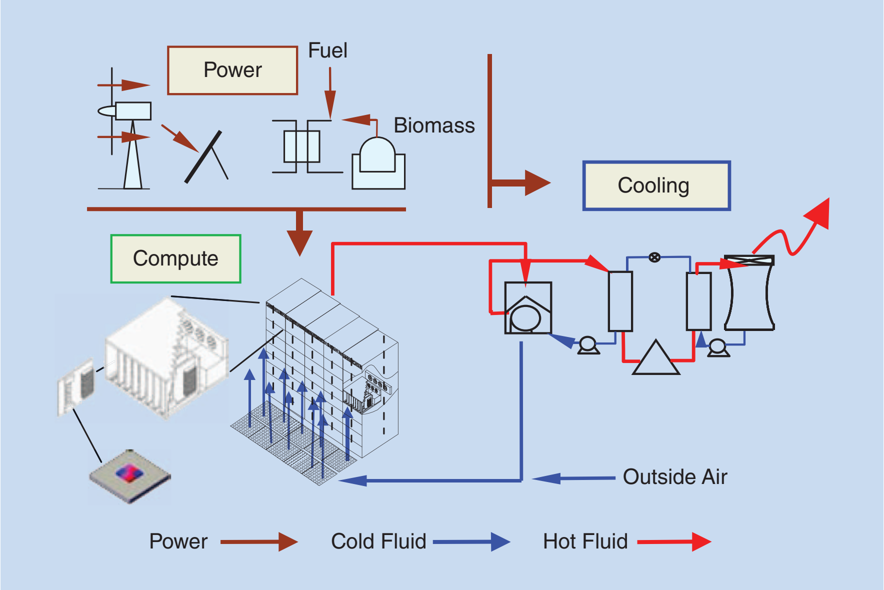


Fig. 1 Elements of a data center [Watson et al. (2009)].

Because sustainability involves complex systems interacting across various scales, "first-principles" models can be both costly to construct and infeasible to use in practice. This is where data mining becomes attractive. Data mining provides a powerful methodology to use inexpensively gathered data and build phenomenological models of the underlying system for possible optimization and reengineering.

Data mining, also referred to as knowledge discovery or machine learning, refers to the extraction of nontrivial and potentially actionable information from massive volumes of data. Established examples of data mining abound, (e.g., mining supermarket baskets for items frequently purchased together, studying customer reviews from product sites to understand opinions and sentiments, and finding patterns of gene expression in cells and tissues). Here, we use data mining techniques with a view toward extracting insights into how to design more sustainable systems. We illustrate our ideas by showcasing the use of data mining in three broad problem contexts: cloud computing, sustainable redesign of products, and urban infrastructure management.

Cloud computing

Cloud computing means different things to different people. For some, it is the "as-a-service" viewpoint that makes computing to be a true utility such as electricity or natural gas. For others, it is the networked view of computer and information resources that can be harnessed from anywhere, (e.g., smartphones). Still others emphasize the "pay for use" model with resources made available on demand rather than explicit provisioning or renting of computers.

Today, every major e-commerce portal or online social networking site runs on the cloud, meaning it is powered by data centers that have grown from housing a few hundred multiprocessor systems to tens of thousands of individual servers. Concomitantly, data centers have become an object of scorn for environmentalists. A news report [Leake and Woods (2009)] ignited a controversy by claiming that a single Web search query can use up to half of the equivalent energy of boiling a kettle of water! A more recent report [Markoff (2011)] has suggested that in recent times data centers have used less power than expected (partly due to reduced demand stemming from the sluggish world economy but also partly due to improved efficiencies in data center equipment and construction). Nevertheless, according to global estimates by the U.S. Environmental Protection Agency, data centers consume 1–2% of the world's electricity and are already responsible for more CO_2 emissions than entire countries such as Argentina or The Netherlands [Kaplan, Forest, and Kindler (2008)]. Hence, reducing the carbon footprint of cloud computing is an important goal to environmental sustainability.

Data centers constitute a mix of computing elements, networking infrastructure, and storage systems along with power and cooling infrastructure, all of which can contribute to energy inefficiency.

Data centers constitute a mix of computing elements, networking infrastructure, and storage systems along with power and cooling infrastructure (see Fig. 1), all of which can contribute to energy inefficiency. Many approaches are possible to stem energy usage across these categories. Servers are typically provisioned based on peak demand and thus are lightly used on average (believed to be in the single digits to at most 10–15%). The low server utilization problem is compounded by the fact that servers are not power proportional; that is, their power consumption is not proportional to their utilization. In fact, even energy efficient servers often consume more than 50% of maximum power at zero to low utilization levels. One approach to improving sustainability of information technology (IT) is to consolidate workload through intelligent scheduling and operating system (OS) virtualization [Tolia et al. (2008)], and turning off the idle servers. In other words, the number of servers deployed dynamically varies with workload. Similarly, dynamic management of an ensemble of chiller units in response to varying load characteristics is another strategy to make a data center more energy efficient. There are even end-to-end methodologies proposed that track inefficiencies at all levels of the IT infrastructure "stack" and derive measures of energy flow efficiencies during data center operation.

To understand what parts of a data center contribute to inefficiencies, it is helpful to conduct an energy breakdown of a data center's consumption. For every 100 W of total power utilized by a data center, often fewer than 50 W goes toward powering IT equipment. The rest of the power goes toward operating the cooling systems, lighting, uninterruptible power supplies (UPSs), server fans, and other subsystems. Of these, the cooling infrastructure, particularly the chiller units, consumes the bulk of the power and is the focus of our attention here.

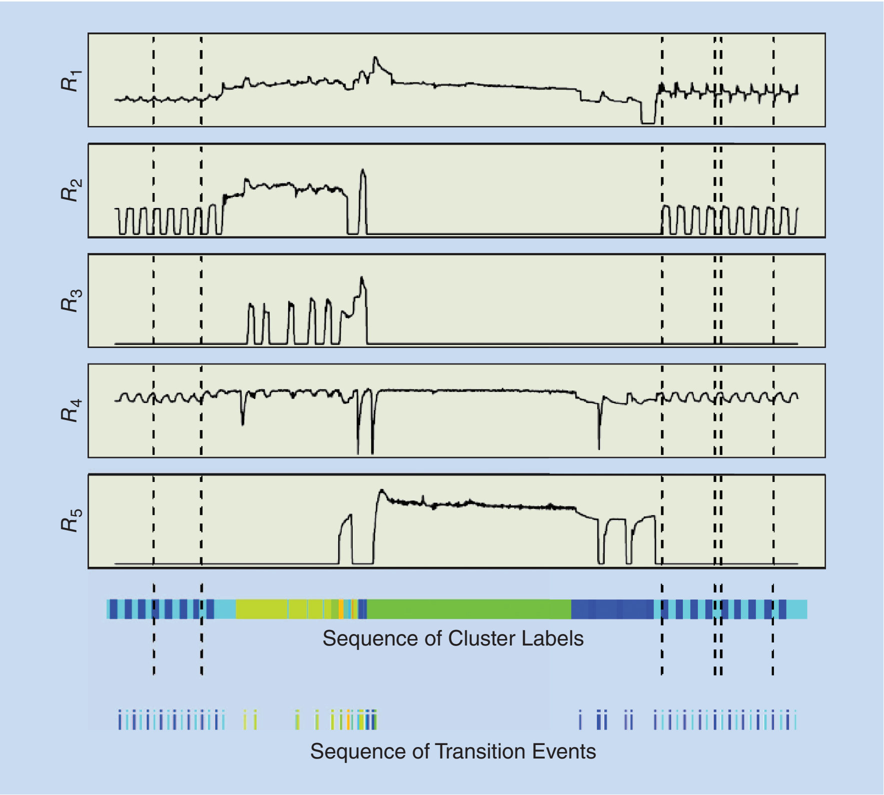


Fig. 2 Redescribing multivariate numeric chiller utilization data into an event sequence symbolic representation.

The cooling infrastructure can be viewed as a pipeline involving computer room air conditioner (CRAC) units, chillers, and cooling towers. CRAC units first cool the exhaust air from the server racks. The chilled water needed by the CRAC units is provided by chillers, where refrigerant loops transfer the heat extracted to the environment directly or through cooling towers. Modern data centers are cooled by an ensemble of chillers configured to dynamically respond to specific load conditions. However, such ensembles are difficult to configure optimally due to the unavailability, inadequacy, or infeasibility of theoretical models ("first principles" methodologies as described earlier). For instance, we can gain access to an operating curve for an individual chiller unit but translating such curves to multiple chillers is nontrivial.

By mining sensor streams from chiller installations, we can obtain a real-time perspective into system behavior and identify strategies to improve efficiency metrics. Due to the "firehose"-like nature of such data streams, data mining algorithms must be able to ingest and process data at rates necessary to yield real-time, actionable, insights.

One of the key aspects of interest to the data center engineer is to efficiently manage an ensemble of potentially heterogeneous chiller units. Hence our objective is to link multivariate, numeric, time series data—utilizations of units in a chiller ensemble–to sustainability metrics. We address this goal by composing a sequence of data mining algorithms [Patnaik et al. (2009)].

As shown in Fig. 2, we first perform clustering of the multivariate series (here utilization values from five chillers, R1–R5) and use the sequence of cluster identifiers as an abstract symbolic representation of the operating point of the overall system. Clustering is a data mining approach that groups nearby points into the same cluster and far-away points into different clusters. We further raise the level of abstraction of the symbol sequence by encoding the transitions from one symbol to another. The resulting event sequence is now mined for repetitive patterns, which we call motifs.

A motif is a pattern of the form that occurs frequently in the event stream.

A motif is a pattern of the form, e.g., "symbol A followed by symbol B followed by symbol C" (not necessarily consecutively), that occurs frequently in the event stream. To mine such patterns, we use serial episode discovery algorithms. An overview is shown in Fig. 3, which uses a levelwise approach popular in many areas of data mining. First, we evaluate patterns of length one-symbol for their frequency, and retain only those that pass a user-specified frequency threshold. These frequent one-symbol patterns are then composed to form candidate two-symbol patterns, which are in turn evaluated and pruned for frequency. For instance, because

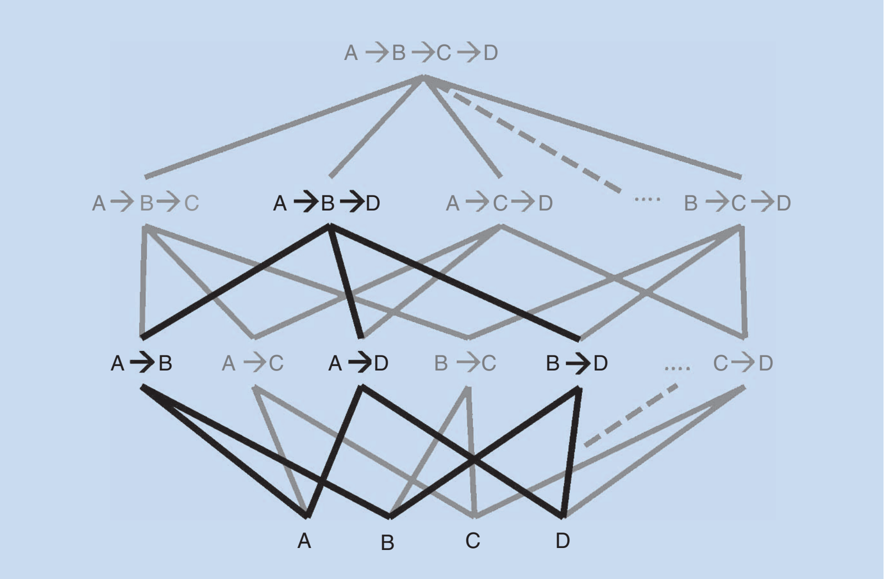


Fig. 3 Levelwise search for motifs.

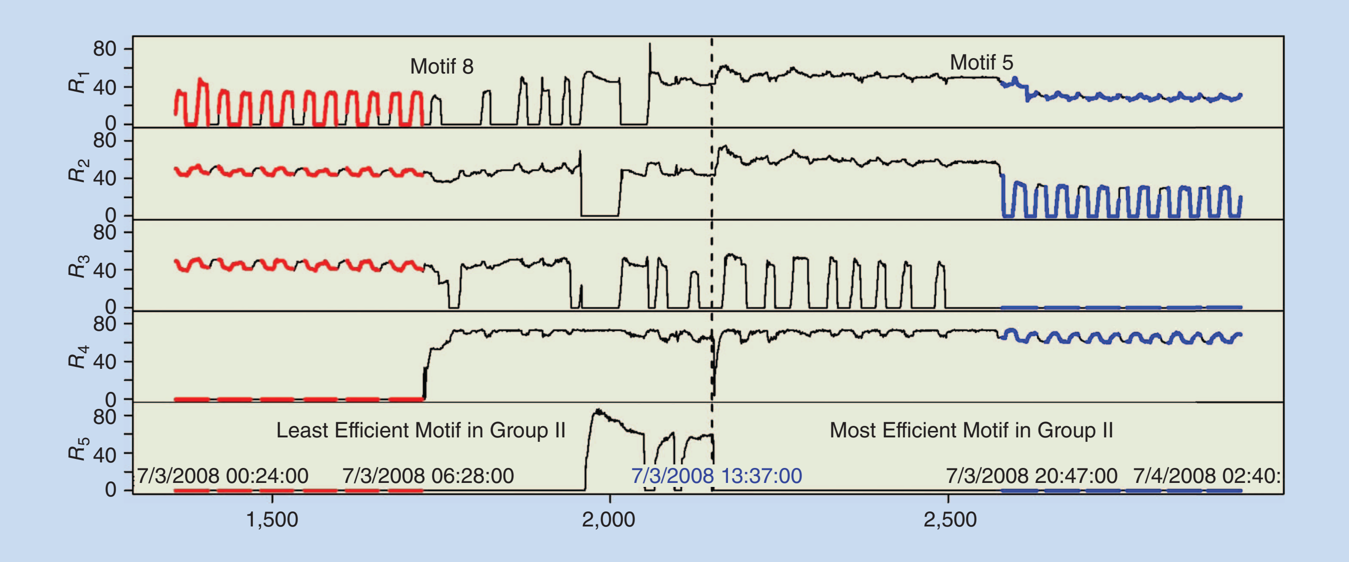


Fig. 4 Two repetitive motifs mined from chiller utilization data. Switching from the left motif to the one on the right yields an estimated power savings of 10%.

symbols A and D are both frequent, we create the candidate pattern "A→D" and evaluate it. Conversely, because symbol C is not frequent, we do *not* consider candidates such as "A→C," "C→D," etc. This process continues until we can no longer compose patterns. We see that "A→B→D" is the longest frequent motif. Because we allow "don't care" symbols inside a motif and because such don't care symbols can span different lengths, this framework allows for robustness to noise and scaling in terms of finding matching motifs.

To summarize, the raw, multivariate, time series data from chillers is first transformed to one discretized sequence using clustering. The time points where cluster labels change are noted as transition events. This process can be noisy depending on the noise level in the raw sequences and the clustering algorithm used. However, the flexibility allowed by the episode mining framework allows us to control this noise by overlooking cluster transitions that are noisy and occur fewer times than true motif patterns. Once motifs are mined, we can translate their occurrences back to the original time series to observe them in their original setting.

Having discovered many motifs, the next step is to categorize them as "good" or "bad" to provide guidance to an administrator or a management system regarding the most efficient configurations of the chiller ensemble under a particular load. There are several sustainability metrics such as power consumed, carbon footprint, and exergy loss to study motifs. Note that optimizing a sustainability metric, such as power consumed, may also minimize the total cost of operation. In our study, we estimate two sustainability metrics for each motif: the average coefficient of performance (COP) of the motif, and a measure reflecting the frequency and amplitude of oscillations in utilization values. The COP of a motif quantifies the cooling effectiveness of the ensemble during that motif occurrence. In order to estimate the frequency of oscillations in a motif, we compute the number of mean crossings, that is, the number of times the utilization crosses the mean value. This is very similar to the number of zero-crossings that is commonly used in speech processing for estimation of frequency.

Fig. 4 describes results from one installation involving an ensemble of five chiller units. The ensemble consists of two types of chillers: three air-cooled chillers and two water-cooled chillers. From the analysis, we found several frequent motifs that repeatedly occurred throughout the data. For example, we found two motifs with very similar load levels (motifs 5 and 8) but which

Table 1. Sample nodes from a PCB BOM. In practice, the number of nodes can easily run into tens of thousands.

Type	Part #	Description
Capacitor	71211838211Y	cap-chip-270pf-50v-k-x7r-0603-tap
	71211858231V	cap-chip-470pf-50v-j-x7r-0603-tap
	7121B1159312	capacitor-al,220uf,16v,m,-55˜+105c
Resistor	7124A1235812	res-chip-976-1%-1/10w-0603-tap
	7124A1235812	res-chip-976-1%-1/10w-0603-tap
	7124B1216112	resistor-ar,4p2r,0,5%,1/16w,1616,tr
Inductor	7125A1123812	idut-4.7uh-20%-43mhz-650ma-smd
	7125B1147812	inductor,0.22uh,+/-10%,25mhz,250ma
	7125B1147812	inductor,0.22uh,+/-10%,25mhz,250ma

Table 2. Impact factors of some nodes in the EI database. In practice, the number of impact factors runs into hundreds.

Description	SO_2 (kg)	CO_2 (kg)
Capacitor, electrolyte type, > 2-cm height	0.21549	47.78
Resistor, SMD type, surface mounting	13.123	11.204
Inductor, miniature RF chip type, MRFI	0.38215	54.542
Integrated circuit, IC, memory type	2.6046	505.92

differed considerably in the COP level. These are shown in Fig. 4. Note that while both motifs 5 and 8 have three chillers turned on, they are of different types. In motif 8, all three operating chillers (C1, C2, and C3) are air-cooled. In motif 5, two air-cooled (C1 and C2) and one water-cooled chiller (C4) are running. In motif 5, one chiller runs at high utilization (C4 at 66.5%), while the other two run at low utilizations (11.3% and 33.8%). In motif 8, one chiller runs at low utilization (17.6%) whereas the other two operate at the medium range (49.1% and 44.3%). If the operational state of the chiller units could be transformed from motif 8 to motif 5, an overall power savings of nearly 10% can be achieved. This directly translates to a cost savings of nearly US$40,000 annually (41 kW savings × 11 cents per kWh × 24 hr × 365 days). Extrapolating this cost saving to other similar motifs gives us an idea of the utility of data mining algorithms in helping achieve cost effectiveness. Moreover, saving 1 kWh of energy is equivalent to preventing 0.8 kg of carbon dioxide release for this data center. The above energy savings would result in a carbon footprint reduction of 287,328 kg of CO_2 released into the atmosphere.

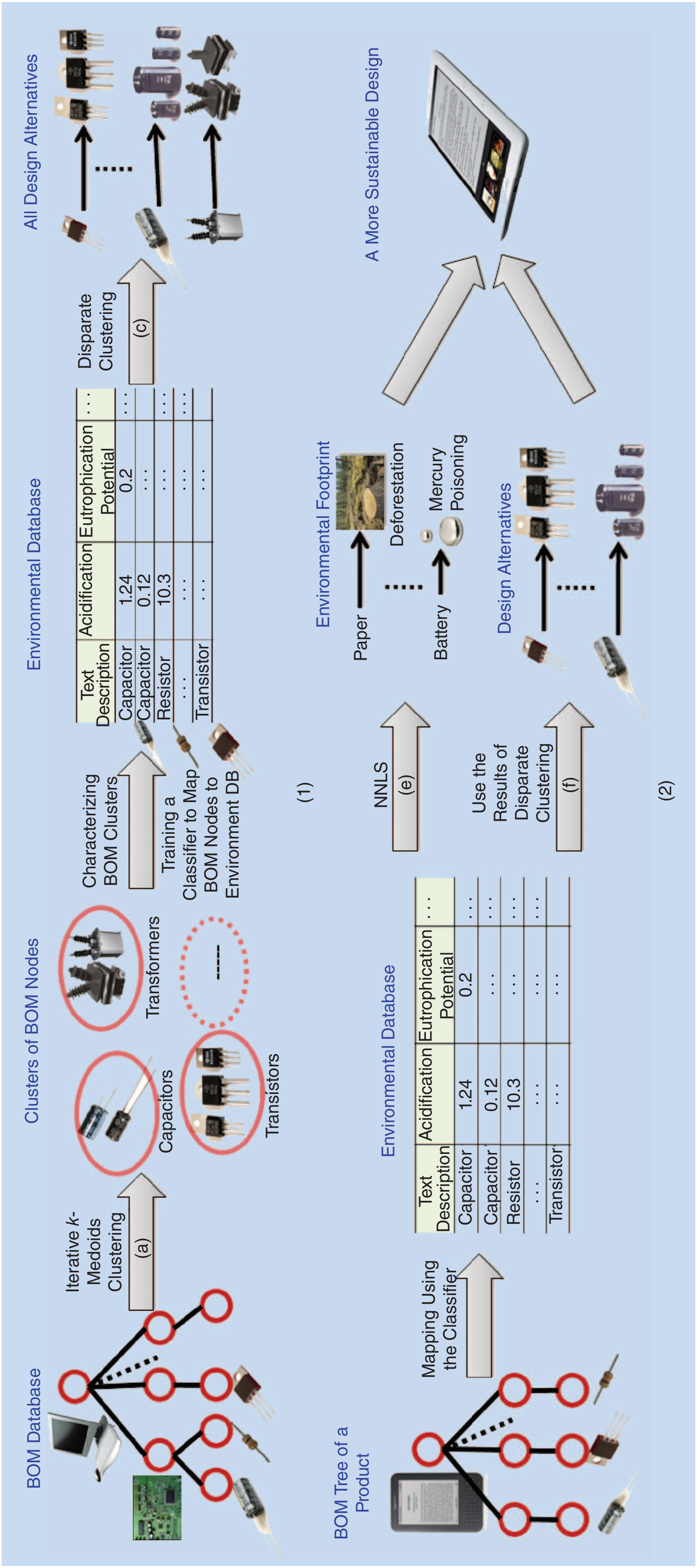


Fig. 5 Architectural overview of our sustainable redesign framework. (1) The steps required to build a classifier for mapping BOM nodes to environmental DB nodes and finally preparing a database of possible alternatives of components. (2) How we can find a more sustainable design of a product using our framework.

Sustainable redesign of products

We now turn to a second illustration of a data mining application to a sustainability problem, namely to design sustainable products. Due to increasing public consciousness about sustainability, companies are ever more eager to introduce ecofriendly products and services.

In a 2010 article in *The New York Times*, Goleman and Norris (2010) investigate whether an e-reader or a printed book is more environmentally friendly. After considering the lifecycle of both products (including materials, manufacturing, transportation, even the light bulb energy used for reading, and finally discard) they conclude that the impact of one e-reader is somewhere between 50–100 paper books. This type of analysis is known as life cycle assessment (LCA) because it requires analysis of each component of a product from “cradle to grave.” Similarly, Toffel and Horvath (2004) compare reading a traditional newspaper versus wirelessly receiving it on a personal ditial assistant and conclude that from a lifecycle perspective the latter results in 32–140 times lower carbon impact and 26–67 times lower water use.

Assessing environmental footprints in this manner and designing sustainable products are challenging tasks since they require analysis of each component of a product through its life cycle. To achieve a sustainable design of products, companies need to evaluate the environmental impact of their system, identify the major contributors to the footprint, and select the design alternative with the lowest environmental footprint.

To understand what is involved, consider a computer manufacturer conducting an LCA of a printed circuit board (PCB). It begins with a bill of materials (BOM) that outlines the composition of the product (see Table 1). The manufacturer must first map the nodes from the BOM into an environmental impacts (EI) database that quantifies multiple environmental impacts of components (see Table 2).

Note that nodes in the BOM database outline attributes such as a part number and a short, unstructured text description (e.g., Table 1) whereas nodes in the EI database provide a textual description of the node and a set of impact factor values (e.g., Table 2). BOM databases are supplied by the manufacturer whereas the EI databases are created by other organizations such as environmental regulation and certification bodies. The description columns across the BOM database (Table 1) and the EI database (Table 2) are hence different and may not have exact resemblance to form a mapping. This is the reason why we need a classifier to map BOM nodes to the nodes of the environmental database [depicted as Fig. 5(2) (top)].

Thus, one important task of data mining here is to learn a mapping from nodes in the BOM database to nodes in the EI database. This mapping serves two purposes: first, it provides an automated mechanism for environmental assessment of BOM components. Second, it helps identify the components where redesign efforts should be focused. We induce a naïve Bayes classifier to learn this mapping which is a technique that computes the posterior probability of classes by assuming conditional independencies of features. Once such a mapping is learned, we apply a disparate clustering technique [Hossain et al. (2010)] to find components that are functionally similar but have disparate environmental impact factors, thus providing candidates for more sustainable design recommendations. This process is depicted in Fig. 5(1).

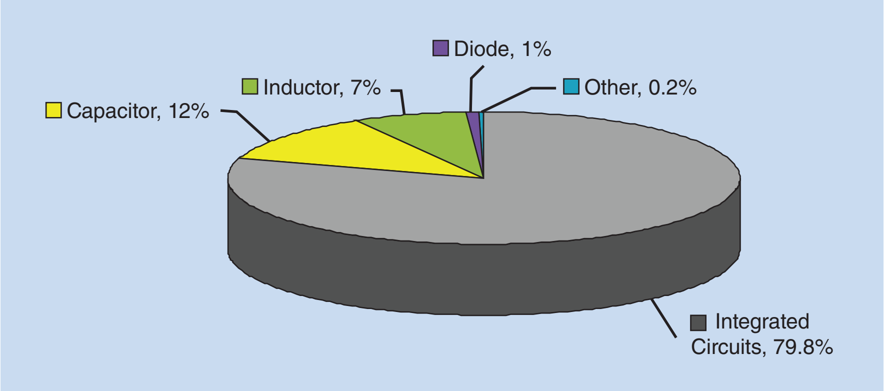


Fig. 6 Hotspot analysis of the carbon footprint of enterprise computer PCB components.

Next, Fig. 5(2) shows how we obtain a design alternative of a specific product. At first, we use the classifier we trained in Fig. 5(1) to map each node of the bill of materials of the product to the EI DB nodes. Then we use a nonnegative least-squares (NNLS) fit to assess the environmental footprints of each component, to identify the top contributors. We also find the design alternatives from the list generated earlier. We suggest replacements for the components that have high environmental footprint with similar but more environmentally friendly components to design a more sustainable product.

> **One important task of data mining here is to learn a mapping from nodes in the BOM database to nodes in the EI database.**

We have applied the above methods on real data from a large computer manufacturer. We performed a case study on an enterprise computer PCB BOM that contained about 560 components, including a mix of resistors, capacitors, application-specific integrated circuits (ASICs), and logic devices. We used our framework to 1) estimate environmental footprints of the BOM components, 2) identify the top contributors to a particular impact (carbon emissions), and 3) suggest design alternatives for the top impact contributors. Note that we prepared a list of possible design alternatives by the methods described in Fig. 5(1). We use NNLS fit [Fig. 5(e)] to assess environmental footprint and identify top contributors. Finally, we used the results generated at the end of Fig. 5(1) to propose design alternatives of the top contributors of a specific BOM in Fig. 5(f). As the hotspot analysis in Fig. 6 shows, the primary culprit to carbon emissions in a PCB are the ICs and our suggested design alternatives could reduce the carbon footprint of the PCB by 4 to 7%. Although this might seem a modest improvement, the millions of PCBs routinely purchased across the globe can add up to a sizable contribution to sustainability.

Sustainability in urban infrastructure

Finally, we look at sustainability issues involving an entire city or urban area. For instance, in the summer of 2011, a significant portion of the United States was reeling under a heat wave, placing significant demands on utility companies in cities. This situation is reminiscent of heat waves recorded (and studied) in the past. For instance, in the book by Klinenberg (2009), which discusses the social and infrastructural issues of the 1995 Chicago heat wave, the author draws attention to the inability of the infrastructure to meet peak demand and how two adjoining neighborhoods (Little Village and North Lawndale) were statistically identical but one had ten times the fatality rates of the other. Empirical models of urban infrastructure are hence critical to understanding such discrepancies.

While urban infrastructure research has typically employed techniques from supply chain management, asset management, logistics, and planning, we are beginning to use data mining techniques to understand the complex

There are many other issues to successfully realizing sustainability goals including the human element of how to encourage and incentivize consumers to conserve resources and the economic and public policy aspects of making sustainable products succeed in the marketplace.

relationships and interactions between entities whose dynamics are evolving over time.

For instance, given gross metered usage data from homes, we can use data mining techniques to disaggregate and reconstruct the energy demand profiles for various home appliances across time. Unsupervised methodologies exist [e.g., Kim et al (2011)] that can break down a power load into its constituents using the aggregate load and contextual information such as time of day, environmental conditions, and usage of other resources. Studies have shown that fine-grained feedback on usage obtained by such methodologies can help curtail peak use by up to 50%. The main advantage of disaggregation is that it allows aggregate load to be split up into its constituents without requiring each individual device or appliance to be instrumented and metered. This provides insights into component-wise resource consumption.

Going further, we can infer a model of realistic urban usage: what utilities are being used during which time periods in predominantly which regions? On top of such usage models, we can impose dynamics so that we can model movements of people between locations, variations in usage with respect to the day of the week, holidays, and other "distractions" such as accidents and closures. This will allow us, to create a synthetic test bed of urban utility consumption. Such synthetic test beds are more privacy-preserving than methods that require intrusive knowledge of people and their habits. Further, they enable us to pose critical "what-if" scenarios that would not be possible through other means. Finally, we can integrate models of multiple physical and social organizational sectors such as electricity, water supply, surface transport, gas supply, drainage, waste management, and telecommunications to arrive at sustainability models for entire regions and cities.

Conclusion

Sustainability issues permeate all aspects of modern life. We have shown how computational sustainability through data mining techniques can serve as a key enabling technology in creating an environmentally friendly future. There are many other issues to successfully realizing sustainability goals that we have not considered here, including the human element of how to encourage and incentivize consumers to conserve resources, and the economic and public policy aspects of making sustainable products succeed in the marketplace. Nevertheless, as more and more complex systems are studied through a sustainability lens, automated methodologies such as those presented here will become more important.

Read more about it

- D. Goleman and G. Norris, "How green is my iPad?" *NY Times*, 4 Apr. 2010.
- C. P. Gomes, "Computational sustainability: Computational methods for a sustainable environment, economy, and society," *Front. Eng.*, vol. 39, no. 4, pp. 5–13, 2009.
- M. S. Hossain, S. Tadepalli, L. T. Watson, I. Davidson, R. F. Helm, and N. Ramakrishnan, "Unifying dependent clustering and disparate clustering for non-homogeneous data," in *Proc. SIGKDD Conf. Knowledge Discovery and Data Mining (KDD'10)*, 2010, pp. 593–602.
- J. M. Kaplan, W. Forrest, and N. Kindler, "Revolutionizing data center energy efficiency," McKinsey and Company Rep., July 2008.
- H. Kim, M. Marwah, M. F. Arlitt, G. Lyon, and J. Han, "Unsupervised disaggregation of low frequency power measurements," in *Proc. SIAM Int. Conf. Data Mining (SDM'11)*, 2011, pp. 747–758.
- E. Klinenberg, *Heat Wave: A Social Autopsy of Disaster in Chicago*. Chicago, IL: Univ. of Chicago Press, 2003.
- J. Leake and R. Woods, "Revealed: The environmental impact of Google searches," *The Sunday Times*, 11 Jan. 2009.
- J. Markoff, "Data centers' power use less than was expected," *NY Times*, July 31, 2011.
- D. Patnaik, M. Marwah, R. K. Sharma, and N. Ramakrishnan, "Sustainable operation and management of data center chillers using temporal data mining," in *Proc. SIGKDD Conf. Knowledge Discovery and Data Mining (KDD '09)*, 2009, pp. 1305–1314.
- N. Tolia, Z. Wang, M. Marwah, C. Bash, P. Ranganathan, and X. Zhu, "Delivering energy proportionality with non energy-proportional systems—optimizing the ensemble," in *Proc. HotPower*, 2008, p. 2.
- M. Toffel and A. Horvath, "Environmental implications of wireless technologies: news delivery and business meetings," *Environ. Sci. Technol.*, vol. 38, no. 11, pp. 2961–2970, 2004.
- B. J. Watson, A. J. Shah, M. Marwah, C. E. Bash, R. K. Sharma, C. E. Hoover, T. W. Christian, and C. D. Patel, "Integrated design and management of a sustainable data center," in *Proc. ASME InterPACK*, July 2009, pp. 635–644.

About the authors

Naren Ramakrishnan (naren@cs.vt.edu) is the Thomas L. Phillips professor of engineering at Virginia Tech in Blacksburg.

Manish Marwah (manish.marwah@hp.com) is a senior research scientist at HP Labs in Palo Alto, California.

Amip Shah (amip.shah@hp.com) is a principal research scientist at HP Labs in Palo Alto, California.

Debprakash Patnaik (debprakash@gmail.com) is a software engineer at Amazon, Inc. in Seattle, Washington.

M. Shahriar Hossain (mshossain@vsu.edu) is an assistant professor in the Mathematics and Computer Science Department at Virginia State University in Petersburg.

Naren Sundaravaradan (narens@vt.edu) is a Ph.D. student at Virginia Tech in Blacksburg.

Chandrakant Patel (chandrakant.patel@hp.com) is a senior fellow and interim director at HP Labs in Palo Alto, California.