Research Statement

Manish Marwah

Physical systems, such as IT infrastructure (data centers, networking, printers), power infrastructure (power plants, distribution network, renewables), buildings, transportation infrastructure, and other urban infrastructure, are being augmented with sensors, actuators, computing and networking resources to transform them into cyber-physical systems (CPS) – integrated systems where physical and computational elements are tightly coupled for monitoring and control. These so called "smart" infrastructures are richly instrumented, and generate unprecedented amounts of data. My research is motivated by the fact that such systems are complex and difficult to model using "first principles," and thus data-driven methods can provide vital insights into their design and operation. Another motivation is sustainability, which has become a societal grand challenge due to the threat of climate change and depleting natural resources. Because CPS have a large environmental impact, my research seeks to answer:

How can data science improve sustainability of cyber-physical systems?

In particular, a large part of my research has centered around designing data mining techniques to improve sustainability of smart buildings and data centers. My work has tried to answer the following questions: How can users be provided with a breakdown of their electricity consumption without intrusive instrumentation? What are the optimal locations to place meters (sensors) in a building? How can anomalies be detected robustly in large scale sensor data? How can the cooling, power and computation infrastructure of a data center be made more sustainable? Answering these questions requires a multi-disciplinary effort. More recently, I have been exploring large scale analytics, in particular graph-based analytics and its applications to the Internet of Things (IoT) and security domains.

A common theme of my research is designing and adapting data mining techniques to solve real, impactful societal problems. Another common theme is multi-disciplinary research. While my background is in computer science, both in data mining and in computer systems, I have successfully collaborated with researchers in other disciplines such as mechanical engineering and civil engineering.

In the next three sections, I provide a summary of my research in smart buildings, sustainable data centers and automating lifecycle assessment, respectively. In the final section, I discuss future research directions.

1 Smart Buildings

Buildings are huge consumers of energy. In fact, in the U.S., buildings account for 40% of all energy use, and contribute about 8% of the global carbon dioxide emissions. At the same time, they present numerous opportunities for energy savings, e.g., it has been reported that poorly maintained, degraded, and improperly controlled equipment wastes 15-30% of energy in commercial buildings. How can data science help in realizing some of these opportunities? I present three pieces of research where I have attempted to address this question: 1) energy disaggregation, 2) meter placement, and 3) anomaly detection.

Energy disaggregation. A problem faced by electricity consumers is that utility bills, usually received at the end of a month, provide only aggregate usage with no insight on which appliances/devices were the top consumers. It has been shown that providing a breakdown of consumption helps users curtail their usage typically by 9-20%. Without installing a meter on every appliance, how can an appliance-wise breakdown be inferred from a single whole house electricity measurement that is provided, say, by a smart meter? This problem is called energy disaggregation or non-intrusive load monitoring (NILM), and in light of sustainability has recently attracted a lot of attention from data mining researchers and practitioners. We proposed an unsupervised approach to energy disaggregation [15] based on a factorial hidden Markov model (FHMM) that uses low frequency aggregate measurements from a smart meter. We proposed FHMM variants that better capture the probability distributions of appliance ON durations, and that allow additional contextual features such as hour of day, day of week, and input from other sensors to be incorporated into the model. We were the first to use a FHMM based approach, which has subsequently been used by several researchers.

In [28], we propose a motif-based approach to energy disaggregation for residential and commercial buildings. This paper received the best student paper award at AAAI 2013 computational sustainability track.

Meter placement. Energy disaggregation is hard to scale, especially in commercial buildings, which have a large number of loads. To monitor consumption and detect anomalous power usage, a building administrator may want to install a limited number of meters. What are the best locations in a building to place these electricity meters? Based on an information theoretic cost criterion, the general problem of finding the best locations is NP-hard. We propose a greedy meter placement algorithm for the electrical infrastructure in a building (a tree) that exploits submodularity of the cost function and provides a near-optimal solution guarantee [4, 5]. This solution can be extended to other similar infrastructures such as water and gas.

Anomaly detection. Buildings contain entities (e.g., devices, appliances) that consume a multitude of resources (e.g., power, water). Efficient operation of these entities is important for reducing operational costs and environmental footprint of buildings. Since it is impossible to manually sift through all the data generated, we developed unsupervised methods to detect anomalies and degradation in performance. In one approach, we proposed a method based on a low-dimensional embedding of the temporal consumption data [5, 4, 6]. In later work, we propose an entity characterization framework [7] based on a finite state machine abstraction. Each state in the state machine is characterized in terms of distributions of sustainability or performance metrics of interest. We demonstrate the usefulness of the framework using data from actual building entities, such as chillers and cooling towers, which are part of building HVAC systems.

2 Sustainable Data Center

Typically, over time the cost of power and cooling in a data center exceeds hardware and other costs, and thus energy management in a data center is vitally important. However, traditionally the three main infrastructural elements of a data center, namely, computation, cooling and power have been treated as silos and optimized separately. As part of the sustainable data center project at HP Labs, we proposed holistic, cross-layer management of data center resources to reduce energy consumption and increase sustainability [2, 17]. Below I describe the data science research threads I initiated and led related to this broader project.

Temporal Data Mining. The cooling infrastructure of a data center can consume anywhere from a third to a half of the total data center energy consumption. In [26, 27], we proposed a temporal data mining solution to model and characterize performance of an ensemble of data center chillers, top energy consumers of the cooling infrastructure. The method helps bridge raw time-series information from sensor streams toward higher level characterizations of chiller behavior, suitable for a data center engineer. Temporal data streams are encoded into a symbolic representation and mined for frequent motifs in the multivariate time series data. The discovered motifs are characterized using sustainability metrics, such as energy consumption and carbon footprint. The ultimate objective is to constrain the operational state of the system to the most desirable motifs. This system was prototyped using data from five chiller units at a Hewlett Packard data center. This work received the best application runner-up award at KDD 2009 [26]. In another temporal data mining project, we predicted solar (PV) panel power generation at a fine-grained level. This is useful in optimizing the use of solar power for a data center, e.g., delay tolerant workload can be scheduled when solar power is likely to be available. We proposed a novel Bayesian ensemble method involving three diverse predictors [9] – naive Bayes, k-NN, and motif-based. We demonstrated its success on real PV data from two locations with diverse weather conditions.

Rack sensor data mining. Data center equipment racks are beginning to be instrumented with temperature sensors to control and optimize the cooling infrastructure. The large amount of data from these sensors enables automated anomaly detection. We investigated methods to discover anomalies that threshold-based methods are unable to detect. We proposed a hierarchical principal component analysis based methodology for detection of anomalous thermal behavior, and demonstrate it on a large temperature sensor network in a production data center [24]. The hierarchical analysis performed on the temperature sensor data streams also identifies the location and scope of such anomalous behavior. These rack sensors are also used to manage computer room air conditioning (CRAC) units to provide cooling only to the extent needed and at the location and time needed. This requires knowledge of the relationship between a CRAC unit settings and its cooling impact at a sensor location in a data center. We proposed AutoTCI [18], where a machine learning model is built for estimating correlation values using historical CRAC settings and temperature sensor data. This non-intrusive mechanism simplifies computation of the correlation values, which was earlier a laborious process, without making simplifying assumptions.

IT Workload modeling. Virtualization technologies enable organizations to dynamically flex their IT resources based on workload fluctuations and changing business needs. However, absence of accurate models relating CPU resources and performance result in under-provisioned or over-provisioned IT resources. We built a probabilistic model relating application response time to CPU allocation [32]. The model is general enough to be applied to other virtualized resources as well. In related work, we proposed a hybrid approach to correctly allocate virtualized resources in a data center to prevent over or under provisioning [10].

Visualization. Since our algorithm for finding motifs in chiller data usually finds tens to hundreds of them, we need to inspect and analyze the discovered motifs. For this purpose, we introduce three novel visual analytics methods [11, 13]: (1) motif layout, using colored rectangles for visualizing the occurrences and hierarchical relationships of motifs in a multivariate time series, (2) motif distortion, for enlarging or shrinking motifs as appropriate for easy analysis and (3) motif merging, to combine a number of identical adjacent motif instances without cluttering the display. To explore CPS time-series data and obtain actionable insights, we construct a Radial Pixel Visualization (RPV) system [12], which uses multiple concentric rings to show the data in a compact circular layout of small polygons (pixel cells), each of which represents an individual data value. RPV provides an effective visual representation of locality and periodicity of the high volume, multivariate data streams, and seamlessly combines them with the results of an automated analysis. In the outermost ring the results of correlation analysis and peak point detection are highlighted. Our explorations demonstrate how RPV can help administrators to identify periodic thermal hot spots, understand data center energy consumption, and optimize IT workload.

3 Automated Lifecycle Assessment

Lifecycle assessment is a widely used method to estimate the environmental impact, such as greenhouse gas emissions and toxicity, of a product (e.g., server, handheld). However, it is a manual and laborious process. We propose several methods to automate this process and reduce the cost and time associated with environmental assessments [23, 14]. The data set consists of the product's bill of materials, and a commercially available database of environmental impacts of commonly used components. The environmental impacts database is essentially a matrix, and we use collaborative filtering techniques to impute missing data [29]; we use a series of clustering, classification methods on the bill of materials to transform it to the impact factor space; determine the component tree in the impact factor space and the top contributors of a particular environmental impact [29, 16]; and finally, use disparate clustering to suggest alternative components for more sustainable design [14].

4 Future Research Directions: Data Science for Cyber-Physical Systems

My research so far only touches a small fraction of how data science can impact cyber-physical systems. Here I refer to CPS broadly – including large collections of interconnected CPS, and Internet of Things based systems. In the future, I plan to extend my current work in a number of directions. I want to explore: 1) how data science can improve *dependability* of CPS, where dependability comprises security, maintainability and availability; 2) the algorithmic and systems level challenges of scaling up CPS analytics to "big data"; and, 3) how data science can make CPS human-aware? These research topics will require collaboration with researchers with varied expertise in the CPS domains.

There are several data mining approaches that are likely to be applicable across CPS domains and problems. In prior work, I have looked at unsupervised and semi-supervised methods [4, 15] to address scarcity of labelled data; transfer learning [19, 14] for not having enough data for the exact task at hand (but enough for similar tasks); graphical models [15, 8, 3] to capture uncertainty and dependence in data. I plan to continue to explore and exploit these approaches.

Further, I would like to continue to collaborate with researchers in industry to obtain real data sets and validate my research from an industrial perspective. The following provides an overview of my research agenda for the next three to five years.

4.1 Dependability of CPS

As CPS become more pervasive and integrated into our society, they need to be dependable. This is especially true for CPS that provide critical services, e.g., the electrical grid, or healthcare infrastructure; if these CPS fail due to a malicious attack or other causes, they can put human life in danger. While dependability covers a whole gamut of attributes, I focus here on security, maintainability and availability. The basic question I address in this research thread is: *How can data science make CPS more dependable?*

- **Security:** Since several CPS provide critical infrastructure and services, protecting them from security threats is vital. However, security of CPS introduces several new challenges. Typically, CPS are designed with performance and cost in mind, with security being an afterthought. Compared to IT systems, many CPS have much longer lifespans, e.g., a building or data center chiller has a life expectancy of at least 15 to 20 years. They also use old insecure protocols (e.g. SCADA). Network and host data at the device and aggregate level is typically used to detect security threats such as malicious attacks and intrusions. A key observation is that the network and host data is a manifestation of a mixture of behaviors of devices and humans. I plan to investigate if network and host data can be used to capture these latent behaviors of CPS devices and humans. In fact, at some level this is similar to the energy disaggregation problem where the whole house measurements are derived from a combination of individual ones. I have used probabilistic graphical models in a semi-supervised setting for inferring malicious web sites [30]. This work was also showcased at the HP Discover conference. Currently, I am investigating graph-based methods for detecting insider threats. Since network and host data have a strong temporal dependence, I would like to explore use of recurrent neural networks for detecting unusual activity.

- **Maintainability and Availability:** Maintainability refers to how quickly a system or an equipment can be restored to its working state following a failure, while availability which depend on maintainability, is the fraction of time a system or service is available. My PhD research focused on systems software mechanisms to improve availability [20, 21, 22]. I would like to investigate how data science can complement systems architectures to further improve maintainability and availability in CPS through failure (anomaly) detection and prediction. Sensors in CPS constantly provide information on the operational state of its various components. This information can be used to determine if any component is degraded or in danger of imminent failure. However, typically quality of the data collected is poor, with likely missing data. While there has been extensive work on anomaly detection methods in the past, most methods are local, that is, use local data, and are not robust to missing or erroneous data. In many instances, there is a large population of same/similar CPS (e.g., appliances, equipment, and vehicles). In these cases, in addition to local data, global relationships between these entities can be leveraged. How can the local model be augmented so the global dependency structure can be captured and exploited? What would be the best way to represent the relationship between these entities? I have done some initial work on this problem using a Markov random field to model a large collection of sensors, and performing inference on this model to detect anomalies [8]. Furthermore, in several instances, data available to build models is insufficient, or over a narrow range of operation of a CPS; in these situations, incorporation of domain knowledge during training results in higher accuracy models [25].

4.2 Scaling up to "Big Data" in CPS

As we scale data mining to increasingly large data sets, parallel and distributed architectures are necessary. To evaluate and compare performance for large scale IoT data, we proposed a benchmark toolkit called IoTAbench [3], and provide an initial benchmark for a smart meter use case. A key part of our toolkit is a realistic synthetic data generator, a generative model that estimates its parameters from a small seed data set, and produces large scale data with similar properties. Our experimental study was performed on over 700 TB of meter data. While the analytics considered in the initial smart meter use case are data parallel, a significant number of data mining algorithms are not, and are thus harder to distribute and scale.

In some cases subsampling and/or independence assumptions can be used to reduce problem size, however, these assumptions may decrease solution accuracy. As an example, consider using a graphical model for billions of sensors distributed globally. How would you run, say, belief propagation, a commonly used inference algorithm, over such a large model so it doesn't take days or weeks to run? Most existing graph analytics platforms do not scale well to this size. What algorithm level changes can be made to run it faster on a given distributed architecture? What systems level changes are required to run a particular algorithm faster? What are the performance/accuracy trade-offs? These are some interesting questions I would like to address as data science is applied to large scale data in CPS. In fact, we have implemented a version of loopy belief propagation for large scale data on Apache Spark/GraphX, and open-sourced it as project Sandpiper [1, 31].

Although deep learning has been very successful in feature learning of images, speech and text, it is supervised. Unsupervised feature learning from large-scale multivariate, spatio-temporal sensor data is still an open research problem that I have started to work on.

4.3 Human-Aware CPS

Most CPS provide some functionality that is used by humans. By human-aware, I imply two main properties: 1) the CPS models some aspect of human behavior for control and optimization; and, 2) the CPS provides interactive visualization for understanding and feedback.

- **Model human behavior:** By being able to detect/predict human behavior, CPS can not only provide better service to humans, but also perform better resource optimization. For example, modeling occupancy behavior of individuals in buildings[6] can be used to optimize control of lighting and HVAC. Similar behavioral modeling can be used in a transportation infrastructure to predict the demand, for say public transportation, during a particular time, or for determining traffic congestion. Based on the data that is available, how can such behavior, e.g., location, use of a resource over time, be best modeled? Dynamic graphical models have proved useful for such modeling, and allow a principled way of integrating heterogeneous, uncertain data. However, several factors, such as number of variables, their domain sizes, coupling between individual models due to behavioral correlations, make parameter estimation and inference hard in these models. Such behavior models can also be used to detect anomalous behavior, which, for instance, could point to a security threat.

- **Model Interpretability:** The outputs of simpler models, such as linear regression, logistic regression or decision trees, are easier to explain and interpret compared to more complex models such as random forests, or neural networks, which pretty much function as black boxes. Simpler models are deployed more often in industry, since often times in addition to providing an output, a model also needs to provide a human understandable explanation for a particular model output. However, this interpretability comes at the cost of accuracy. In essence, there is a trade-off between model interpretability and model accuracy. I am currently exploring novel ways of providing explanations, aimed at helping operators, for outputs of CPS anomaly detection models.

- **Visualization:** CPS generate a lot of multi-modal data, and it is humanly impossible to manually scan all of it. However, CPS managers, operators and users are interested in looking at "important" parts of the data, where an importance score can be defined by context of the data and the particular user. This leads to following high-level research questions: How can large-scale multi-modal data (e.g., video, text, audio, other sensor data) be effectively filtered and summarized based on context and the user? Further, how can it be made interactive to allow the user to tweak parameters and pose "what if" queries?

References

[1] https://github.com/hewlettpackard/sandpiper.

[2] Arlitt, M., Bash, C., Blagodurov, S., Chen, Y., Christian, T., Gmach, D., Hyser, C., Kumari, N., Liu, Z., Marwah, M., et al. Towards the design and operation of net-zero energy data centers. In *Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm), 2012 13th IEEE Intersociety Conference on* (2012), IEEE, pp. 552–561.

[3] Arlitt, M., Marwah, M., Bellala, G., Shah, A., Healey, J., and Vandiver, B. IoTAbench: An internet of things analytics benchmark. In *Proceedings of the 6th ACM/SPEC International Conference on Performance Engineering* (New York, NY, USA, 2015), ICPE '15, ACM, pp. 133–144.

[4] Bellala, G., Marwah, M., Arlitt, M., Lyon, G., and Bash, C. Following the electrons: methods for power management in commercial buildings. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* (2012), ACM, pp. 994–1002.

[5] Bellala, G., Marwah, M., Arlitt, M., Lyon, G., Bash, C., and Shah, A. Data analytics for managing power in commercial buildings. *TCPS 1*, 4 (2017), 22:1–22:25.

[6] Bellala, G., Marwah, M., Arlitt, M., Lyon, G., and Bash, C. E. Towards an understanding of campus-scale power consumption. In *Proceedings of the Third ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings* (2011), ACM, pp. 73–78.

[7] Bellala, G., Marwah, M., Shah, A., Arlitt, M., and Bash, C. A finite state machine-based characterization of building entities for monitoring and control. In *Proceedings of the Fourth ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings* (2012), ACM, pp. 153–160.

[8] Chakrabarti, A., Marwah, M., and Arlitt, M. Robust anomaly detection for large-scale sensor data. In *Proceedings of the 3rd ACM International Conference on Systems for Energy-Efficient Built Environments* (New York, NY, USA, 2016), BuildSys '16, ACM, pp. 31–40.

[9] Chakraborty, P., Marwah, M., Arlitt, M. F., and Ramakrishnan, N. Fine-grained photovoltaic output prediction using a bayesian ensemble. In *AAAI* (2012).

[10] Gandhi, A., Chen, Y., Gmach, D., Arlitt, M., and Marwah, M. Minimizing data center sla violations and power consumption via hybrid resource provisioning. In *Green Computing Conference and Workshops (IGCC), 2011 International* (2011), IEEE, pp. 1–8.

[11] Hao, M., Marwah, M., Janetzko, H., Sharma, R., Keim, D., Dayal, U., Patnaik, D., and Ramakrishnan, N. Visualizing frequent patterns in large multivariate time series. In *IS&T/SPIE Electronic Imaging* (2011), International Society for Optics and Photonics, pp. 78680J–78680J.

[12] Hao, M., Marwah, M., Mittelstadt, S., Janetzko, H., Keim, D., Dayal, U., Bash, C., Felix, C., Patel, C., Hsu, M., et al. Exploring cyber physical data streams using radial pixel visualizations. In *Visual analytics science and technology (VAST), 2012 IEEE Conference on* (2012), IEEE, pp. 225–226.

[13] Hao, M. C., Marwah, M., Janetzko, H., Dayal, U., Keim, D. A., Patnaik, D., Ramakrishnan, N., and Sharma, R. K. Visual exploration of frequent patterns in multivariate time series. *Information Visualization 11*, 1 (2012), 71–83.

[14] Hossain, M. S., Marwah, M., Shah, A., Watson, L. T., and Ramakrishnan, N. Autolca: A framework for sustainable redesign and assessment of products. *ACM Transactions on Intelligent Systems and Technology (TIST) 5*, 2 (2014), 34.

[15] Kim, H., Marwah, M., Arlitt, M. F., Lyon, G., and Han, J. Unsupervised disaggregation of low frequency power measurements. In *SDM* (2011), vol. 11, pp. 747–758.

[16] Kim, J., Ramakrishnan, N., Marwah, M., Shah, A., and Park, H. Regularization paths for sparse nonnegative least squares problems with applications to life cycle assessment tree discovery. In *2013 IEEE 13th International Conference on Data Mining* (2013), IEEE, pp. 360–369.

[17] Liu, Z., Chen, Y., Bash, C., Wierman, A., Gmach, D., Wang, Z., Marwah, M., and Hyser, C. Renewable and cooling aware workload management for sustainable data centers. In *ACM SIGMETRICS Performance Evaluation Review* (2012), vol. 40, ACM, pp. 175–186.

[18] Marwah, M., Bash, C., Zhou, R., Felix, C., Shih, R., and Christian, T. Estimating data center thermal correlation indices from historical data. In *Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm), 2012 13th IEEE Intersociety Conference on* (2012), IEEE, pp. 344–352.

[19] Marwah, M., and Mahajan, R. L. Building neural network equipment models using model modifier techniques. *IEEE transactions on semiconductor manufacturing 12*, 3 (1999), 377–381.

[20] Marwah, M., Mishra, S., and Fetzer, C. Tcp server fault tolerance using connection migration to a backup server. In *DSN* (2003), pp. 373–382.

[21] Marwah, M., Mishra, S., and Fetzer, C. A system demonstration of st-tcp. In *2005 International Conference on Dependable Systems and Networks (DSN'05)* (2005), IEEE, pp. 308–313.

[22] Marwah, M., Mishra, S., and Fetzer, C. Enhanced server fault-tolerance for improved user experience. In *2008 IEEE International Conference on Dependable Systems and Networks With FTCS and DCC (DSN)* (2008), IEEE, pp. 167–176.

[23] Marwah, M., Shah, A., Bash, C., Patel, C., and Ramakrishnan, N. Using data mining to help design sustainable products. *IEEE Computer 44*, 8 (2011), 103–106.

[24] Marwah, M., Sharma, R., Lugo, W., and Bautista, L. Anomalous thermal behavior detection in data centers using hierarchical pca. In *Proceedings of the 4th International Workshop on Knowledge Discovery from Sensor Data (SensorKDD 2010)* (2010).

[25] Muralidhar, N., Islam, R., Marwah, M., Karpatne, A., and Ramakrishnan, N. Incorporating prior domain knowledge into deep neural networks. In *IEEE Big Data* (2018).

[26] Patnaik, D., Marwah, M., Sharma, R., and Ramakrishnan, N. Sustainable operation and management of data center chillers using temporal data mining. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (2009), ACM, pp. 1305–1314.

[27] Patnaik, D., Marwah, M., Sharma, R. K., and Ramakrishnan, N. Temporal data mining approaches for sustainable chiller management in data centers. *ACM Transactions on Intelligent Systems and Technology (TIST) 2*, 4 (2011), 34.

[28] Shao, H., Marwah, M., and Ramakrishnan, N. A temporal motif mining approach to unsupervised energy disaggregation: Applications to residential and commercial buildings. In *AAAI* (2013).

[29] Sundaravaradan, N., Patnaik, D., Ramakrishnan, N., Marwah, M., and Shah, A. Discovering life cycle assessment trees from impact factor databases. In *AAAI* (2011).

[30] Ulanov, A., and Marwah, M. Malicious site detection with large-scale belief propagation, Mar. 2017.

[31] Ulanov, A., Marwah, M., Kim, M., Dathathri, R., Zubieta, C., and Li, J. Sandpiper: Scaling probabilistic inferencing to large scale graphical models. In *IEEE Big Data* (2017).

[32] Watson, B. J., Marwah, M., Gmach, D., Chen, Y., Arlitt, M., and Wang, Z. Probabilistic performance modeling of virtualized resource allocation. In *Proceedings of the 7th international conference on Autonomic computing* (2010), ACM, pp. 99–108.