Improving Efficiency and Reliability of Building Systems Using Machine Learning and Automated Online Evaluation

Leon Wu, Gail Kaiser, David Solomon, Rebecca Winter, Albert Boulanger, Roger Anderson
School of Engineering and Applied Science
Columbia University
New York, NY 10027, USA

***Abstract*—A high percentage of newly-constructed commercial office buildings experience energy consumption that exceeds specifications and system failures after being put into use. This problem is even worse for older buildings. We present a new approach, 'predictive building energy optimization', which uses machine learning (ML) and automated online evaluation of historical and real-time building data to improve efficiency and reliability of building operations without requiring large amounts of additional capital investment. Our ML approach uses a predictive model to generate accurate energy demand forecasts and automated analyses that can guide optimization of building operations. In parallel, an automated online evaluation system monitors efficiency at multiple stages in the system workflow and provides building operators with continuous feedback.**

We implemented a prototype of this application in a large commercial building in Manhattan. Our predictive machine learning model applies Support Vector Regression (SVR) to the building's historical energy use and temperature and wet-bulb humidity data from the building's interior and exterior in order to model performance for each day. This predictive model closely approximates actual energy usage values, with some seasonal and occupant-specific variability, and the dependence of the data on day-of-the-week makes the model easily applicable to different types of buildings with minimal adjustment. In parallel, an automated online evaluator monitors the building's internal and external conditions, control actions and the results of those actions. Intelligent real-time data quality analysis components quickly detect anomalies and automatically transmit feedback to building management, who can then take necessary preventive or corrective actions. Our experiments show that this evaluator is responsive and effective in further ensuring reliable and energy-efficient operation of building systems.

***Index Terms*—green buildings, energy efficiency, prediction methods, reliability, machine learning, support vector machines, statistical analysis**

I. Introduction

According to the U.S. Department of Energy (DOE), commercial office buildings lead the industrial and transportation sectors in total energy consumption [1]. Although new buildings are often designed with energy efficiency and system reliability in mind, the use of energy-efficient materials and advanced Building Management Systems (BMS) does not always guarantee efficient and reliable building operation. A high percentage of new buildings consume energy at levels that exceed specifications and experience system failures after being put into use [2]. This problem is even worse for older buildings. We present here a new approach, applying machine learning (ML) and automated online evaluation to historical and real-time building Supervisory Control and Data Acquisition (SCADA) data and other building information to improve the efficiency and reliability of building systems without requiring large amounts of additional capital investment. We have developed a prototype of this application, which we implemented in a large multi-tenant office building in New York City.

Our ML approach, termed 'predictive building energy optimization', uses a model to produce accurate building energy demand forecasts as well as automated analyses that can aid in the tuning of building systems and operations schedules. It applies Support Vector Regression (SVR) on historical energy use of the building, along with temperatures and wet-bulb humidity data from the building's interior and exterior, to predict performance for each day. This does not require knowledge of the building's physical properties, such as size, heating, ventilation, air conditioning (HVAC) or electrical systems. It employs time-delay coordinates as a representation of past data in order to create the feature vectors for Support Vector Machine (SVM) training. Our experiments show that the predictive model closely approximates the actual values of energy usage with some seasonal and occupant-specific variability. The dependence of the data on day-of-the-week makes the model easily applicable to different types of buildings with minimal adjustments.

To ensure that the ML system works reliably 24x7, an automated online evaluator monitors the building's internal and external conditions (*e.g.*, temperature, humidity, electrical load, peak load, fluctuating electricity pricing and building work and maintenance schedules) control actions (*e.g.*, adjusting lighting, turning on/off the AC/heat and shutting off elevators) and the results of those actions. This evaluator employs intelligent real-time data quality analysis components to quickly detect anomalies, such as malfunctions of digital thermostats that interfere with temperature reading or introduce variances from normal HVAC set-points, and sends feedback to building management, who can then take appropriate preventive or corrective actions. Our experiments show that this automated online evaluator is responsive and effective in further ensuring that building systems continue to

run reliably and energy-efficiently.

In the following section, we provide background building data. In section III, we describe the use of predictive building energy optimization to improve energy efficiency. In section IV, we describe automated online evaluation for improving system reliability. In section V, we present the results of our empirical study. We then compare some related work in section VI before providing our concluding remarks in section VII.

II. Background on Building Data

A. Building Energy Data

Building energy use is measured by total electricity consumption over a period of time, typically kilowatt-hours (kWh) per month. The kilowatt-hour is most commonly known as a billing unit for energy delivered to consumers by electric utilities. The energy demand of a building is the rate of energy consumption by the building; because energy use fluctuates during the week due to tenant activities and building operation schedule, energy demand is a more fine-grained measure of building energy use than the aggregate kilowatt-hours consumed during the whole period.

Large buildings commonly use Building Management Systems (BMS) to manage the interior environment and control mechanical and electrical equipment such as ventilation, lighting, power systems, fire systems and security systems. BMS provides a way to retrieve building energy-related data, such as data readings from sub-meters and sensors.

B. Weather Data

Climate factors include temperature, humidity, pressure, wind and cloud cover. In a highly condensed urban environment like New York City, different areas can have different weather measurements. This is often called *micro-weather*. The most commonly used weather data for New York City are collected from a weather station located in Central Park, because of its accuracy and stability. Historic hourly weather data can be obtained from some public websites, such as National Climatic Data Center [3] and Weather Underground [4].

Relative humidity and dew point temperature are also important weather data for buildings. Relative humidity is the ratio of the partial pressure of water vapor in the air to the saturated vapor pressure of water under given temperature and pressure conditions. Dew point is the temperature at which the air can no longer hold all of its water vapor, such that some of the water vapor must condense into water. The dew point is always lower than (or equal to) the air temperature. If the air temperature cools to the dew point, or if the dew point rises to equal the air temperature, then dew begin to form. When the dew point temperature equals the air temperature, the relative humidity is 100%.

C. Power Grid Data

Power grid data from utilities include electrical load, peak load, fluctuating electricity pricing during the day and power failure warning. The power grid data from the utilities are often communicated electronically via client web portal or email.

III. Predictive Building Energy Optimization

To improve the efficiency of building systems, we employed a new ML-based approach called predictive building energy optimization.

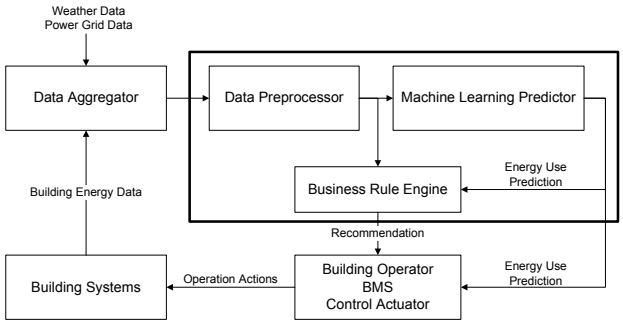

Fig. 1. Predictive building energy optimization workflow.

As illustrated in Fig. 1, predictive building energy optimization starts with data aggregation and preprocessing. External data, such as weather and power grid data, are combined with building energy data in the data aggregator, which passes the aggregated data to the data preprocessor for cleaning, formatting and normalization. The ML predictor uses historic energy use data as training data to build a model, which is then used to predict energy use in the present. This prediction is then passed to the building management and business rule engine. The business rule engine processes the aggregated data and the ML prediction in order to generate a set of recommended operation actions. The building management (*i.e.*, building operators or BMS or automatic control actuators) can then take action on the building systems, such as adjusting its HVAC schedule and set-points to achieve more efficient building operation. The modified building data will then be fed back to the data aggregator, thereby closing the loop.

A. Data Preprocessing

The data preprocessor receives various data streams from the data aggregator and restructures them so that all the data fit the format required by the ML predictor and business rule engine. For the specific predictive modeling technique we use, all the data need to be normalized to a value between 0 and 1 for equal weighting.

	Energy Demand	Temperature	Dew Point Temperature	Pressure	Wind Speed	Humidity
Test Set	?	0.65	0.52	0.40	0.72	0.68
Training Set	0.65	0.63	0.50	0.38	0.75	0.71
	.	.	.	.	.	.
	.	.	.	.	.	.
	.	.	.	.	.	.

Fig. 2. Sample training and test dataset.

As illustrated in Fig. 2, the training set is used to build the predictive model (*i.e.*, a function that can be used for predicting unknown values). The test set includes data for every column except energy demand, which needs to be predicted. The training set is laid out in descending order, such that one hour before prediction is top-most, two hours before prediction is next, and so forth. Some data rows at the bottom of the training set will lack 'prior value' data due to the ordering system, and those rows are ignored. For any given column, such as temperature, it is possible to expand it such that multiple prior temperatures over a sequential series of time-points become properties of the same data row. In this way, we can construct a normalized dataset with time-delayed coordinates for use by the ML predictor.

B. ML Predictive Modeling

Predictive analytics or predictive modeling deals with extracting information from data and using it to predict future trends and behavior patterns. An SVM is a supervised learning method for predictive modeling. It constructs a hyperplane or set of hyperplanes in a high or infinite-dimensional space, which can be used for classification, regression or other tasks [5]. An SVR is a version of SVM for regression [6]. The model produced by SVR depends only on a subset of the training data, because the cost function for building the model ignores any training data close to the model prediction.

A nonlinear kernel function allows the algorithm to fit the maximum-margin hyperplane in a transformed feature space. We selected a Gaussian radial basis function (RBF), $K(X_i, X_j)$, as the SVM kernel function:

$$K(X_i, X_j) = e^{-(\epsilon||X_i - X_j||)^2}, (\epsilon > 0).$$

The RBF kernel nonlinearly maps samples into a higher dimensional space and, unlike the linear kernel, can handle the case where the relationship between class labels and attributes is nonlinear [7]. The nonlinear, dynamic nature of the influence of weather and other data on energy demand in building systems excludes the possibility of using a linear kernel.

In order to measure how well future outcomes are likely to be predicted by the model, we used the coefficient of determination R^2, which is used in the context of statistical models whose main purpose is the prediction of future outcomes on the basis of other related information. This statistical model accounts for the proportion of variability in a dataset [8].

$$R^2 = 1 - \frac{SS_{err}}{SS_{tot}} = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \bar{y})^2},$$

where values y_i are the observed values and values f_i are the modeled values or predicted values in the dataset. SS_{err} is called the residual sum of squares and SS_{tot} is called the total sum of squares. The closer the R^2 is to 1, the more accurate the predicted values are and the better the predictive model is.

C. Business Rule Engine

The business rule engine receives aggregated data and ML prediction output. It then applies the business knowledge that supports rules, constraints, priority, mutual exclusion, preconditions and other functions onto the data to derive executable recommendations such as work schedule and preventive actions. The business rule engine consists of a BPM (Business Process Management) component and a BRM (Business Rules Management) component. Both components interact with each other responding to events or executing business judgments that are defined by business rules.

The set of business rules is initially defined and incrementally improved by experienced building operators and property managers. It includes both forwarding-chaining (*e.g.* IF something happens THEN do something) and backward-chaining rules (*e.g.*, IF I want to achieve this goal THEN something has to happen). These collected rules can also serve as the learning metrics for the more advanced adaptive stochastic controller (ASC) driven by approximate dynamic programming (ADP) to derive action or policy recommendations [9].

IV. Automated Online Evaluation

To ensure that the ML system works reliably 24x7, the internal and external conditions, control actions, and the results of those actions are evaluated using an automated online evaluator. As illustrated in Fig. 3, the automated online evaluation system receives data at multiple stages in the system workflow. The evaluator employs intelligent real-time data quality analysis components to quickly detect data anomalies (*e.g.*, malfunctions of digital thermostats that interfere with temperature reading or introduce variances from normal expected HVAC set-points) and gives feedback to building management, who can then respond appropriately.

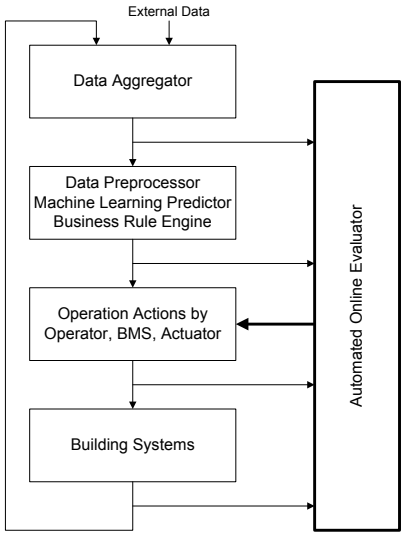

Fig. 3. Automated online evaluation workflow.

A. Real-time Data Quality Analysis

1) Thresholds: The thresholds define the normal working range of the specific data points. If the data reading exceeds the thresholds at either the lower or upper bound, the data record will be flagged as anomalous and a corresponding warning will be communicated back to the building operator electronically.

2) Online Anomaly Detection Using Incremental Local Outlier Factor (LOF) Algorithm: Anomaly detection finds data instances that are unusual and do not fit any established pattern. It concentrates on modeling normal behavior in order to identify unusual data points. This system processes the continuously updated data-streams to detect anomalies, using an incremental LOF algorithm. This uses k-nearest neighbor on each inserted data record to instantly compute LOF value, which is the degree to which a data record represents an outlier or an indicator of abnormality [10]. The LOF values for existing data records can be updated on the fly if necessary.

3) Visualization: Visualization provides an easy way to obtain additional verification of a data anomaly. This component is also a useful communication channel to help building management understand where issues are arising.

V. Empirical Study

A. Implementation

We implemented a prototype application of our predictive building energy optimization approach at 345 Park Avenue, a 634 ft (193 m) tall skyscraper in midtown Manhattan, New York City. Designed by Emery Roth & Sons and completed in 1969, the building has 44 floors and more than 2 million square feet of tenant space. Rudin Management, one of the largest private real estate companies in New York City, is the building owner and property manager. Approximately 5,000 people work in the building, and there are about 1,000 visitors to the building daily. The building's regular hours are 7:00 AM to 7:00 PM Monday through Friday, and 8:00 AM to 1:00 PM on Saturdays. The estimated energy cost of running the HVAC system of 345 Park for an hour amounts to approximately $2,000 to $2,500 in 2011. The building uses electricity, steam and natural gas supplied by Con Edison, the main utilities company in New York City, for heating and cooling in the building. Management has installed a state-of-the-art energy monitoring system, which provides an archived data log of energy demand that can be used for predictive building energy optimization.

B. Results of Predictive Building Energy Optimization

A more detailed empirical study of our data can be found in our technical report [11]. In order to identify the SVR parameters (*i.e.*, C and γ values), and number of time delays that yield the most accurate and efficient model, we used a step-wise search method. The step-wise method works by running regressions using values of different orders of magnitude for a specific parameter, calculating the R^2 value to assess accuracy, then evaluating on finer scales until the appropriate value is established. We used the same method variable selection of C, γ and time delay values, where the test file incorporating real values as classifiers in order to compare the model's accuracy at predicting for those values. First we evaluated the C value at 1, 10, 100 and 500, and then at 200, 300, 400. We evaluated γ at 1, 0.1, 0.01, 0.001 and 0.0001. We evaluated the number of time delays at 24, 48, 96, 144, 192, 240, 288 and 336; these are all multiples of 24 to ensure that we did not disrupt daily cyclicity.

Based on the results of the R^2 statistical tests, the best combination of variables for a February regression would be to use one year of energy data. For May, the best combination of variables would be two years of energy and temperature. While these statistical tests proved the accuracy of these models, we opted to use two years of energy, temperature and humidity for all regressions. The reason for this is that response of the HVAC to weather is very dynamic. In studying the physical HVAC plant at 345 Park Avenue, the operations management indicated that they employ the next day's heat index in order to determine their heating and cooling load for the day. Heat index is determined based on a combination of temperature and humidity in an attempt to estimate how the air temperature feels to humans. We therefore decided that it was important to include those variables in the creation of our model. A model using fewer variables produces smooth, highly cyclical curves, while the addition of more variables creates curves with more noise and statistically poorer fits. However, the inclusion of more variables allows the model to adapt more dynamically to changes in weather that occur within a single day or week, and it aids the model in predicting minimal and maximal energy demand values.

Figures 4 and 5 show regression results of SVR prediction versus actual energy demand for two different five-month datasets at different times of the year. The spring graph is closer to the actual energy consumption of the building, with an R^2 value of about 0.95, while the winter graph is less accurate, with an R^2 of about 0.71. We hypothesize the reason for the less accurate winter regression is that the SVR predictive model may need additional features in its dataset in order to better handle low winter temperature values, which cause increased energy demand for heating.

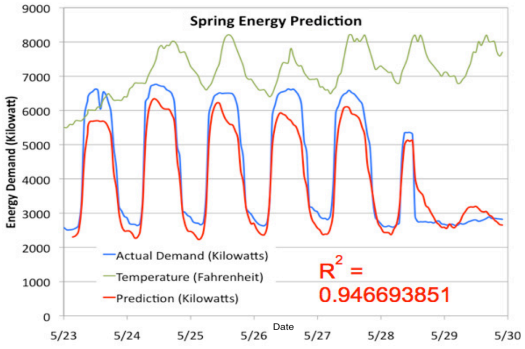

Fig. 4. Predicted versus actual energy demand in May 2011 [11]

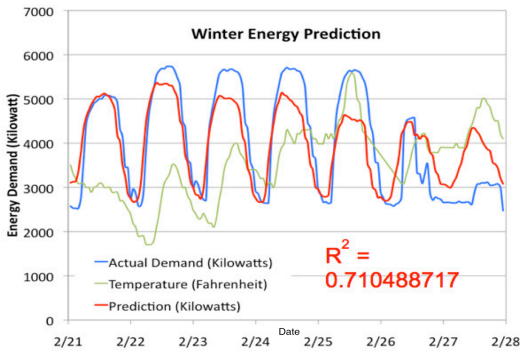

Fig. 5. Predicted versus actual energy demand in Feb 2011 [11]

C. Results of Automated Online Evaluation

Our experiments showed that the automated online evaluator is responsive and effective in ensuring that building systems continue to run reliably and energy-efficiently. We identified more than 10 suspicious data anomalies among 2480 building data-points obtained over a two-month period (December 2011 to January 2012) and investigated the related sensor or SCADA data sources. Fig. 6 and Fig. 7 show the dynamic real-time visualization charts with selectable data-points. Fig. 6 also shows the building system shutdown during New Year's Eve and the subsequent reactivation after the holiday. If not for the holiday, this kind of dip would have been detected as anomalous behavior and a warning would be triggered and sent to building management from the automated online evaluator.

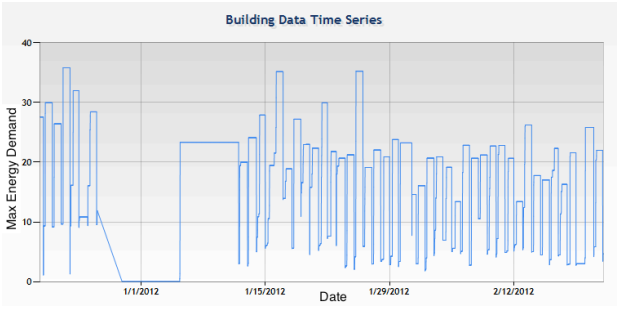

Fig. 6. Maximum energy demand time series.

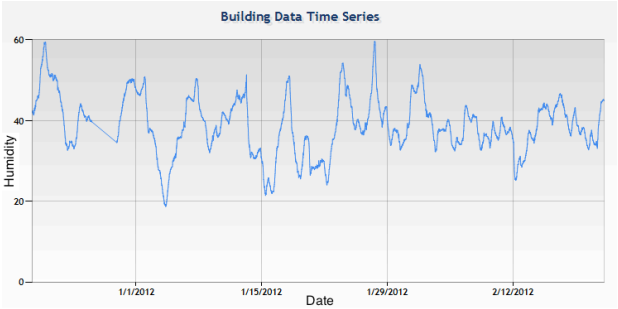

Fig. 7. Building internal wet-bulb humidity time series.

VI. Related Work

Some prior research has been done to predict and analyze the energy demand of buildings. Dong *et al.* used SVM to predict building energy consumption in a tropical region [7]. However, our application of SVR is different from their approach in architecture and applied domains. The DOE-2 model, created by the U.S. Department of Energy, uses physical aspects of the building such as construction materials to predict its energy needs [12]. Our approach, on the other hand, makes exclusive use of operational building data to model building energy demand. A sensor data-based building information measurement and actuation profile for building data management was discussed in [13]. The building information collected from sensors can serve as input to our ML predictor and automated online evaluator.

VII. Conclusion

This paper presents a new approach using ML and automated online evaluation of historical and real-time building data to improve efficiency and reliability of building systems without requiring large amounts of additional capital investment. The ML component generates a predictive model of building energy demand forecasts and applies automated analyses to aid in the tuning of building systems and operations schedules. The automated online evaluation works in parallel with the ML and existing BMS to conduct continuous evaluations at multiple stages in the system workflow, and provides building operators with continuous feedback that can be used to improve reliability and performance. Our experiments show that this SVR model is accurate in predicting energy demand and that the automated online evaluation is effective and responsive.

Acknowledgment

Wu and Kaiser are members of the Programming Systems Laboratory, funded in part by NSF CNS-0717544, CNS-0627473 and CNS-0426623, and NIH 2 U54 CA121852-06. Wu, Boulanger, and Anderson are members of the Energy Research Group in the Center for Computational Learning Systems at Columbia University, supported in part by General Electric, FedEx, Consolidated Edison, and Rudin Management Company.

References

[1] U.S. Department of Energy, "Energy efficiency & renewable energy: Building technologies program," 2012, available at www.eere.energy.gov/buildings/.

[2] U.S. Energy Information Administration, "Annual energy review 2010," October 2011.

[3] National Oceanic and Atmospheric Administration, "National climate data center," 2012. [Online]. Available: http://www.ncdc.noaa.gov/oa/ncdc.html

[4] Weather Underground, "Weather underground," 2012. [Online]. Available: http://www.wunderground.com/

[5] V. N. Vapnik, *The nature of statistical learning theory*. New York: Springer-Verlag, 1995.

[6] H. Drucker, C. J. C. Burges, L. Kaufman, A. J. Smola, and V. N. Vapnik, "Support vector regression machines," *Advances in Neural Information Processing Systems 9, NIPS 1996*, pp. 155–161, 1997.

[7] B. Dong, C. Cao, and S. E. Lee, "Applying support vector machines to predict building energy consumption in tropical region," *Energy and Buildings*, vol. 37, 2005.
[8] R. G. D. Steel and J. H. Torrie, *Principles and Procedures of Statistics*. McGraw-Hill, 1960.
[9] R. N. Anderson, A. Boulanger, W. B. Powell, and W. Scott, "Adaptive stochastic control for the smart grid," *Proceedings of the IEEE*, vol. 99, no. 6, pp. 1098–1115, June 2011.
[10] D. Pokrajac, A. Lazarevic, and L. J. Latecki, "Incremental local outlier detection for data streams," in *In Proceedings of IEEE Symposium on Computational Intelligence and Data Mining*, 2007, pp. 504–515.
[11] D. Solomon, R. Winter, A. Boulanger, R. Anderson, and L. Wu, "Forecasting energy demand in large commercial buildings using support vector machine regression," Department of Computer Science, Columbia University, Tech. Rep. CUCS-040-11, September 2011.
[12] Simulation Research Group, Lawrence Berkeley National Laboratory, University of California, *Overview of DOE-2.2*. University of California, June 1998.
[13] S. Dawson-Haggerty, X. Jiang, G. Tolle, J. Ortiz, and D. Culler, "sMAP a simple measurement and actuation profile for physical information," in *Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems (SenSys'10)*, November 2010.

Leon Wu (M'07) is a PhD candidate at the Department of Computer Science and a Senior Research Associate at the Center for Computational Learning Systems of Columbia University. He received his MS and MPhil in Computer Science from Columbia University and BSc in Physics from Sun Yat-sen University.

Gail Kaiser (M'85-SM'90) is a Professor of Computer Science and the Director of the Programming Systems Laboratory in the Computer Science Department at Columbia University. She was named an NSF Presidential Young Investigator in Software Engineering and Software Systems in 1988, and she has published over 150 refereed papers in a range of software areas. Her research interests include software testing, collaborative work, computer and network security, parallel and distributed systems, self-managing systems, Web technologies, information management, and software development environments and tools. She has consulted or worked summers for courseware authoring, software process and networking startups, several defense contractors, the Software Engineering Institute, Bell Labs, IBM, Siemens, Sun and Telcordia. Her lab has been funded by NSF, NIH, DARPA, ONR, NASA, NYS Science & Technology Foundation, and numerous companies. Prof. Kaiser served on the editorial board of IEEE Internet Computing for many years, was a founding associate editor of ACM Transactions on Software Engineering and Methodology, chaired an ACM SIGSOFT Symposium on Foundations of Software Engineering, vice chaired three of the IEEE International Conference on Distributed Computing Systems, and serves frequently on conference program committees. She also served on the Committee of Examiners for the Educational Testing Service's Computer Science Advanced Test (the GRE CS test) for three years, and has chaired her department's doctoral program since 1997. Prof. Kaiser received her PhD and MS from CMU and her ScB from MIT.

David Solomon is an undergraduate student at the Department of Earth and Environmental Sciences at Columbia College.

Rebecca Winter is an undergraduate student at the Department of Earth and Environmental Engineering at Columbia University Fu Foundation School of Engineering and Applied Science.

Albert Boulanger received a B.S. in physics at the University of Florida, Gainesville, Florida USA in 1979 and a M.S. in computer science at the University of Illinois, Urbana-Champaign, Illinois USA in 1984. He is a co-founder of CALM Energy, Inc. and a member of the board at the not-for-profit environmental and social organization World Team Now and founding member of World-Team Building, LLC. He is a Senior Staff Associate at Columbia University's Center for Computational Learning Systems, and before that, at the Lamont-Doherty Earth Observatory. For the past 12 years at Columbia, Albert has been involved in far reaching energy research and development in oil and gas and electricity He is currently a member of a team of 15 scientists and graduate students in Computer Sciences at Columbia who are jointly developing with Con Edison and others the next generation Smart Grid for intelligent control of the electric grid of New York City. He held the CTO position of vPatch Technologies, Inc., a startup company commercializing a computational approach to efficient production of oil from reservoirs based on time-lapse 4D seismic technologies. Prior to coming to Lamont, Albert spent twelve years doing contract R&D at Bolt, Beranek, and Newman (now Raytheon BBN Technologies). His specialties are complex systems integration and intelligent computational reasoning that interacts with humans within large-scale systems.

Roger Anderson (M'09) has been at Columbia University for 35 years, where he is a Senior Scholar at the Center for Computational Learning Systems in the Fu School of Engineering and Applied Sciences (SEAS). Roger is Principal Investigator of a team of 15 scientists and graduate students in Computer Sciences at Columbia who are jointly developing the next generation Smart Grid for intelligent control of the electric grid of New York City with Con Edison and others in New York City. Previously at the Lamont-Doherty Earth Observatory of Columbia, Roger founded the Borehole Research, Global Basins, 4D Seismic, Reservoir Simulation, Portfolio Management, and Energy Research Groups. Roger also teaches Planet Earth, a science requirement course in the core curriculum at Columbia College from his position in the Department of Earth and Environmental Sciences. He co-founded the Alternative Energy program at the School of International and Public Affairs at Columbia, and is a director of the Urban Utility Center at the Polytechnic Institute of New York University. Roger received his Ph.D. from the Scripps Institution of Oceanography, University of California at San Diego. He is the inventor of 16 Patents, and has written 3 books, & more than 200 peer-reviewed scientific papers.