# Crowdsourcing Predictors of Behavioral Outcomes

Josh C. Bongard, *Member, IEEE*, Paul D. H. Hines, *Member, IEEE*, Dylan Conger, Peter Hurd, and Zhenyu Lu

***Abstract*—Generating models from large data sets—and determining which subsets of data to mine—is becoming increasingly automated. However choosing what data to collect in the first place requires human intuition or experience, usually supplied by a domain expert. This paper describes a new approach to machine science which demonstrates for the first time that non-domain experts can collectively formulate features, and provide values for those features such that they are predictive of some behavioral outcome of interest. This was accomplished by building a web platform in which human groups interact to both respond to questions likely to help predict a behavioral outcome and pose new questions to their peers. This results in a dynamically-growing online survey, but the result of this cooperative behavior also leads to models that can predict user's outcomes based on their responses to the user-generated survey questions. Here we describe two web-based experiments that instantiate this approach: the first site led to models that can predict users' monthly electric energy consumption; the other led to models that can predict users' body mass index. As exponential increases in content are often observed in successful online collaborative communities, the proposed methodology may, in the future, lead to similar exponential rises in discovery and insight into the causal factors of behavioral outcomes.**

***Index Terms*—Crowdsourcing, machine science, surveys, social media, human behavior modeling**

## I. Introduction

There are many problems in which one seeks to develop predictive models to map between a set of predictor variables and an outcome. Statistical tools such as multiple regression or neural networks provide mature methods for computing model parameters when the set of predictive covariates and the model structure are pre-specified. Furthermore, recent research is providing new tools for inferring the structural form of non-linear predictive models, given good input and output data [1]. However, the task of choosing which potentially predictive variables to study is largely a qualitative task that requires substantial domain expertise. For example, a survey designer must have domain expertise to choose questions that will identify predictive covariates. An engineer must develop substantial familiarity with a design in order to determine which variables can be systematically adjusted in order to optimize performance.

The need for the involvement of domain experts can become a bottleneck to new insights. However, if the wisdom of crowds could be harnessed to produce insight into difficult problems, one might see exponential rises in the discovery of the causal factors of behavioral outcomes, mirroring the exponential growth on other online collaborative communities. Thus, the goal of this research was to test an alternative approach to modeling in which the wisdom of crowds is harnessed to both propose potentially predictive variables to study by asking questions, and respond to those questions, in order to develop a predictive model.

The authors are with the College of Engineering and Mathematical Sciences, University of Vermont, Burlington, VT USA (e-mail: jbongard@uvm.edu, paul.hines@uvm.edu).

This work was supported in part by the UVM Complex Systems Center, through NASA Grant #NNX09AJ18G. D. Conger was supported by the McNair Scholars Program.

This paper has been accepted for publication in a future edition of the *IEEE Transactions on Systems, Man, and Cybernetics*. Updated March 8, 2012.

### *Machine science*

Machine science [2] is a growing trend that attempts to automate as many aspects of the scientific method as possible. Automated generation of models from data has a long history, but recently robot scientists have been demonstrated that can physically carry out experiments [3], [4] as well as algorithms that cycle through hypothesis generation, experimental design, experiment execution, and hypothesis refutation [5], [1]. However one aspect of the scientific method that has not yet yielded to automation is the selection of variables for which data should be collected to evaluate hypotheses. In the case of a prediction problem, machine science is not yet able to select the independent variables that might predict an outcome of interest, and for which data collection is required.

This paper introduces, for the first time, a method by which non domain experts can be motivated to formulate independent variables as well as populate enough of these variables for successful modeling. In short, this is accomplished as follows. Users arrive at a website in which a behavioral outcome (such as household electricity usage or body mass index, BMI) is to be modeled. Users provide their own outcome (such as their own BMI) and then answer questions that may be predictive of that outcome (such as 'how often per week do you exercise'). Periodically, models are constructed against the growing data set that predict each user's behavioral outcome. Users may also pose their own questions that, when answered by other users, become new independent variables in the modeling process. In essence, the task of discovering and populating predictive independent variables is outsourced to the user community.

### *Crowdsourcing*

The rapid growth in user-generated content on the Internet is an example of how bottom-up interactions can, under some circumstances, effectively solve problems that previously required explicit management by teams of experts [6]. Harnessing the experience and effort of large numbers of individuals is frequently known as "crowdsourcing" and has been used effectively in a number of research and commercial applications [7]. For an example of how crowdsourcing can be useful, consider Amazon's Mechanical Turk. In this crowdsourcing tool a human describes a "Human Intelligence Task" such as characterizing data [8], transcribing spoken language

[9], or creating data visualizations [10]. By involving large groups of humans in many locations it is possible to complete tasks that are difficult to accomplish with computers alone, and would be prohibitively expensive to accomplish through traditional expert-driven processes [11].

Although arguably not strictly a crowdsourced system, the rapid rise of Wikipedia illustrates how online collaboration can be used to solve difficult problems (the creation of an encyclopedia) without financial incentives. Ref. [12] reviews several crowdsourcing tools and argues that direct motivation tasks (tasks in which users are motivated to perform the task because they find it useful, rather than for financial motivation) can produce results that are superior to financially motivated tasks. Similarly, ref. [12] reports that competition is useful in improving performance on a task with either direct or indirect motivation. This paper reports on two tasks with direct motivation: for the household energy usage task, users are motivated to understand their home energy usage as a means to improve their energy efficiency; for the body mass index task, users are motivated to understand their lifestyle choices in order to approach a healthy body weight. Both instantiations include an element of competition by allowing participants to see how they compare with other participants and by ranking the predictive quality of questions that participants provide.

There is substantial evidence in the literature and commercial applications that laypersons are more willing to respond to surveys and queries from peers than from authority figures or organizations. For example within the largest online collaborative project, Wikipedia, article writers often broadcast a call for specialists to fill in details on a particular article. The response rates to such peer-generated requests are enormous, and have led to the overwhelming success of this particular project. In the open source community, open source software that crashes automatically generates a debug request from the user. Microsoft adopted this practice but has found that users tend not to respond to these requests, while responses to open source crashes are substantially higher [13], [14]. Medpedia, a Wikipedia-styled crowdsourced system, increasingly hosts queries from users as to what combinations of medications work well for users on similar medication cocktails. The combinatorial explosion of such cocktails is making it increasingly difficult for health providers to locate such similar patients for comparison without recourse to these online tools.

Collaborative systems are generally more scalable than top-down systems. Wikipedia is now orders-of-magnitude larger than Encyclopedia Britannica. The climateprediction.net project has produced over 124 million hours of climate simulation, which compares favorably with the amount of simulation time produced by supercomputer simulations. User-generated news content sites often host as many or more readers than conventional news outlets [15]. Finally, many of the most recent and most successful crowdsourced systems derive their success from their viral [16], [17] nature: they are designed such that selective forces exerted by users lead to an exponential increase in content, automated elimination of inferior content, and automated propagation of quality content [18].

Citizen science [19], [20], [21] platforms are a class of crowdsourcing systems that include non-scientists in the scientific process. The hope is that participants in such systems are motivated ideologically, as their contributions forward what they perceive as a worthy cause. In most citizen science platforms user contributions are 'passive': they contribute computational but not cognitive resources [19], [22]. Some platforms allow users to actively participate by searching for items of interest [23] or solve problems through a game interface [24]. The system proposed here falls into this latter category: users are challenged to pose new questions that, when answered by enough of their peers, can be used by a model to predict the outcome of interest.

Finally, problem solving through crowdsourcing can produce novel, creative solutions that are substantially different from those produced by experts. An iterative, crowdsourced poem translation task produced translations that were both surprising and preferable to expert translations [25]. We conjecture that crowdsourcing the selection of predictive variables can reveal creative, unexpected predictors of behavioral outcomes. For problems in which behavioral change is desirable (such as is the case with obesity or energy efficiency), identifying new, unexpected predictors of the outcome may be useful in identifying relatively easy ways for individuals to change their outcomes.

## II. Methodology

The system described here wraps a human behavior modeling paradigm in cyberinfrastructure such that: (1) the investigator defines some human behavior-based outcome that is to be modeled; (2) data is collected from human volunteers; (3) models are continually generated automatically; and (4) the volunteers are motivated to propose new independent variables. Fig. 1 illustrates how the investigator, participant group and modeling engine work together to produce predictive models of the outcome of interest. The investigator begins by constructing a web site and defining the human behavior outcome to be modeled (Fig. 1a). In this paper a financial and health outcome were investigated: the monthly electric energy consumption of an individual homeowner (Sect. III), and their body mass index (Sect. IV). The investigator then initializes the site by seeding it with a small set (one or two) of questions known to correlate with the outcome of interest (Fig. 1b). For example, based on the suspected link between fast food consumption and obesity [26], [27], we seeded the BMI website with the question "*How many times a week do you eat fast food?*"

Users who visit the site first provide their individual value for the outcome of interest, such as their own BMI (Fig. 1g). Users may then respond to questions found on the site (Fig. 1h,i,j). Their answers are stored in a common data set and made available to the modeling engine. Periodically the modeling engine wakes up (Fig. 1m) and constructs a matrix $\mathbf{A} \in \Re^{n \times k}$ and outcome vector $\mathbf{b}$ of length $n$ from the collective responses of $n$ users to $k$ questions (Fig. 1n). Each element $a_{ij}$ in $\mathbf{A}$ indicates the response of user $i$ to question $j$, and each element $b_i$ in $\mathbf{b}$ indicates the outcome of interest as entered by user $i$. In the work reported here linear regression

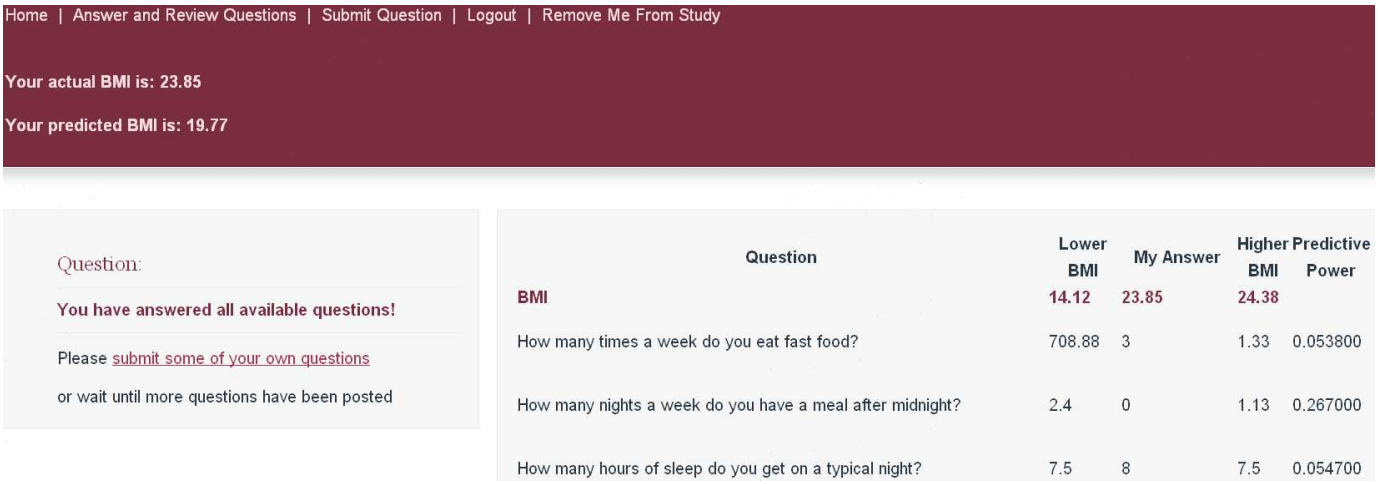



Figure 2. Screenshot from the Body Mass Index (BMI) website as seen by a user who has responded to all of the available questions. The user has the option to change their response to a previous question, pose a new question, or remove themselves automatically from the study.

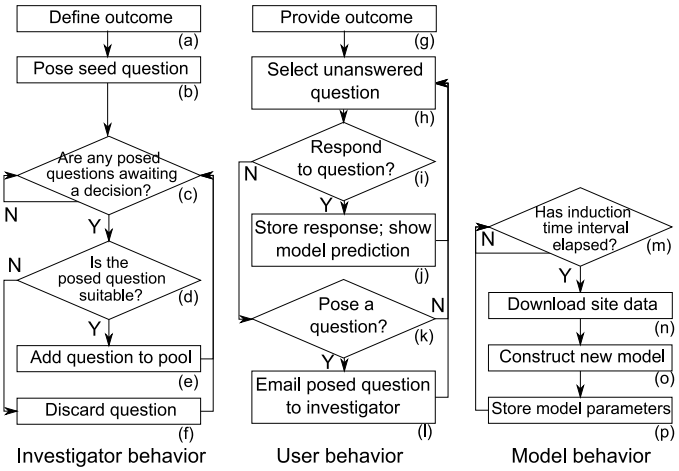



Figure 1. **Overview of the system.** The investigator (**a-f**) is responsible for initially creating the web platform, and seeding it with a starting question. Then, as the experiment runs they filter new survey questions generated by the users. Users (**g-l**) may elect to answer as-yet unanswered survey questions or pose some of their own. The modeling engine (**m-p**) continually generates predictive models using the survey questions as candidate predictors of the outcome and users' responses as the training data.

was used to construct models of the outcome (Fig. 1o), but any model form could be employed. The modeling process outputs a vector $\mathbf{c}$ of length $k+1$ that contains the model parameters. It also outputs a vector $\mathbf{d}$ of length $k$ that stores the predictive power of each question: $d_j$ stores the $r^2$ value obtained by regressing only on column $j$ of $\mathbf{A}$ against the response vector $\mathbf{b}$. These two outputs are then placed in the data store (Fig. 1p).

At any time a user may elect to pose a question of their own devising (Fig. 1k,l). Users could pose questions that required a yes/no response, a five-level Likert rating, or a number. Users were not constrained in what kinds of questions to pose. However, once posed, the question was filtered by the investigator as to its suitability (Fig. 1d). A question was deemed unsuitable if any of the following conditions were met: (1) the question revealed the identity of its author (e.g. "*Hi, I am John Doe. I would like to know if...*") thereby contravening the Institutional Review Board approval for these experiments; (2) the question contained profanity or hateful text; (3) the question was inappropriately correlated with the outcome (e.g. "*What is your BMI?*"). If the question was deemed suitable it was added to the pool of questions available on the site (Fig. 1e); otherwise the question was discarded (Fig. 1f).

Each time a user responded to a question, they were shown a new, unanswered question as well as additional data devised to maintain interest in the site and increase their participation in the experiment. Once a user had answered all available questions, they were shown a listing of the questions, their responses, and contextual information to indicate how their responses compared to those of their peers. Fig. 2 shows the listing that was shown to those users who participated in the BMI site; the individual elements are explained in more detail in Sect. IV.

The most important datum shown to each user after responding to each question was the value of their actual outcome as they entered it ($b_i$) as well as their outcome as predicted by the current model ($\hat{b}_i$). Fig. 2 illustrates that visitors to the BMI site were shown their actual BMI (as entered by them) and their predicted BMI. The models were able to predict each user's outcome before they had responded to every question by substituting in missing values. Thus after each response from a user

$$\hat{b}_i = c_0 + c_1 a_{i1} + c_2 a_{i2} + ... + c_k a_{ik} + \epsilon_i \qquad (1)$$

where $a_{ij} = 0$ if user $i$ has not yet responded to question $j$ and $a_{ij}$ is set to the user's response otherwise.

## III. Energy efficiency instantiation and results

In the first instantiation of this concept, we developed a web-based social network to model residential electric energy consumption. Because of policy efforts to increase energy efficiency, many are working to provide consumers with better information about their energy consumption. Research on consumer perception of energy efficiency indicates that electricity

customers often misjudge the relative importance of various activities and devices to reducing energy consumption [28]. To provide customers with better information, numerous expert-driven web-based tools have been deployed [29], [30], [31]. In some cases these tools use social pressure as a means of improving energy efficiency [32], [33], however the feedback provided to customers typically comes from a central authority (i.e., top-down feedback) and research on risk perception [34] indicates that the public is often skeptical of expert opinions. A recent industry study [35] indicates that customers are notably skeptical of large online service providers (e.g., Google, Microsoft) and (to a lesser extent) electric utilities as providers of unbiased information about energy conservation. Therefore, information generated largely by energy consumers themselves, in a bottom-up fashion, may have value in terms of motivating energy efficient behavior.

Thus motivated, we designed the "EnergyMinder" website to predict and provide feedback about monthly household (residential) electricity consumption. Participants were invited to join the site through notices in university e-mail networks, a university news letter, and reddit, a user-generated content news site. The site was launched in July of 2009, and gradually accumulated a total of 58 registered users by December of 2009. The site consisted of a simple login page and five simple, interactive pages. The *Home Page* (after login) contained a simple to-do list pointing users to tasks on the site, such as, enter bill data, answer questions, check their energy efficiency ranking, etc. The *Energy Input Page* showed a time series trend of the consumer's monthly electricity consumption and asked the user to enter the kilowatt hours (kWh) consumed for recent months. This value became the output variable ($b_i$) in the regression model (Eq. 1) for a particular month. The *Ask-A-Question Page* allowed users to ask questions of the group, such as "How many pets do you have?" (Question 10, Table I). When typing in a new question, users were instructed to specify the type of answer expected (numeric, yes/no, agree/disagree) and to provide their own response to the question. The *Answer Page* asked participants to respond to questions, and provided them with information about each answered question including the distribution of answers within the social network. Finally, a *Ranking Page* showed users their energy consumption, relative to that of others in the group. In addition the Ranking Page reported the predictive power (the percentage of explained variance) for each statistically significant question/factor. This final page was intended to provide information to participants that might help them in choosing behaviors that would reduce electricity consumption.

In total the site attracted 58 participants, of whom 46 answered one or more questions, and 33 (57%) provided energy consumption data. Eight new questions were generated by the group, after the seed questions ($Q_1$ and $Q_2$ in Table I) were placed there by the investigators. The fact that only about half of the participants provided energy data was most likely due to the effort associated with finding one or more electricity bills and entering data into the site. This low response rate emphasized that the utility of this approach depends highly on the ease with which the user can access the outcome data.

Despite the small sample size, this initial trial resulted in a statistically significant predictive model, and provided insight into the nature of the method. Of the 33 participants, 24 provided data for the months of June, July or August. Because this was the largest period for which common data were available, the mean outcome for these three months was used as the outcome variable $b_i$. One participant reported kWh values that were far outside of the mean (46,575 kWh per month) and one did not answer any questions. These two data sets were discarded as outliers. The $N = 22$ that remained comprised the sample-set used to produce the results that follow.

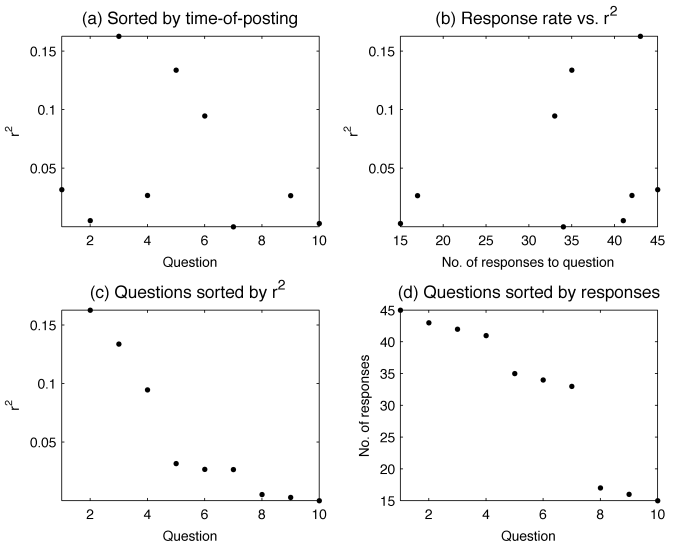



Figure 3. **EnergyMinder Question Statistics**. Panel (**a**) shows the $r^2$ value for each question as numbered in table I. (**b**) shows that there is a mild correlation between the response rate and the $r^2$ values. (**c**) shows the questions sorted by their $r^2$ value, and (**d**) shows the number of responses for each question, sorted by the number of responses.

Table I shows results from two predictive models. Model 1 included all questions that had 18 or more answers ($Q_1$-$Q_7$). The total explained variance for Model 1 was $r^2 = 0.63$. Model 1 indicated that the number of adults in the home ($Q_3$) significantly increased monthly electricity consumption ($P < 0.05$) and the ownership of a natural gas hot water heater ($Q_6$) significantly decreased electricity consumption ($P < 0.05$). Note that this second result is not consistent with the fact that owning an electric hot water heater increases electricity consumption. It appears either that this correlation was due to chance, or that ownership of a gas hot water heater correlates to some other factor, such as (for example) home ownership. Model 2 tested the removal of the least significant predictors, and included only $Q_3$, $Q_5$, and $Q_6$. Model 2 showed the same pair of statistically significant predictors ($Q_3$ and $Q_6$).

Figure 3 shows the relative predictive power of the 10 questions. The results show that the most highly correlated factors ($Q_3$, $Q_5$, and $Q_6$) were posed after the initial two seed questions (Fig. 3a) and a weak correlation between the response rate and the $r^2$ values, indicating that more answers to questions would have likely produced improved results. Panels (c) and (d) show the distributions of $r^2$ values and the number of responses, to facilitate comparison with the BMI

Table I
QUESTIONS ENTERED INTO THE ENERGYMINDER WEB SITE.

| | Question | Type | # of answers | answers in $G$ | Model 1** $c_i$ | Model 1** $P$ | Model 2** $c_i$ | Model 2** $P$ |
|---|---|---|---|---|---|---|---|---|
| 1. | What is the square footage of your house?* | numeric | 45 | 22 | 0 | 0.52 | - | - |
| 2. | How many children do you live with?* | numeric | 41 | 22 | 109 | 0.47 | - | - |
| 3. | How many adults do you live with? | numeric | 43 | 22 | 303 | 0.03 | 297 | 0.01 |
| 4. | How many south facing windows do you have? | numeric | 42 | 22 | -11 | 0.77 | - | - |
| 5. | Do you have an electric clothes dryer? | yes/no | 35 | 19 | 430 | 0.23 | 240 | 0.28 |
| 6. | Do you have an electric water heater? | yes/no | 33 | 18 | -577 | 0.04 | -535 | 0.01 |
| 7. | Do you have gas heating? | yes/no | 34 | 18 | 188 | 0.44 | - | - |
| 8. | Do you have geothermal heating? | yes/no | 16 | 10 | - | - | - | - |
| 9. | How many adults are typically home throughout the day? | numeric | 17 | 10 | - | - | - | - |
| 10. | How many pets do you have? | numeric | 15 | 9 | - | - | - | - |
| | $r^2$ value for predictive models | | | | 0.63 | | 0.57 | |

* Questions 1 and 2 were seed questions placed on the site by the investigators.
** In Model 1 and Model 2, $c_i$ is the parameter estimate (kWh $\cdot$ month$^{-1}$ $\cdot$ unit$^{-1}$) and $P$ is the significance level of the parameter estimate.

results (Fig. 6).

While the small sample size in this study limits the generality of these results, this initial trial provided useful information about the crowdsourced modeling approach. Firstly, we found that participants were reluctant or unable to provide accurate outcome data due to the challenge of finding one's electric bills. Our second experiment corrects this problem by focusing on an outcome that is readily accessible to the general public. Secondly, we found that participants were quite willing to answer questions posed by others in the group. Questions 1-4 were answered by over 70% of participants. This indicated that it is possible to produce user-generated questions and answers, and that a trial with a larger sample size might provide more valuable insight. Finally, questions that were posed early in the trial gained a higher response rate, largely because many users did not return to the site after one or two visits. This emphasizes the importance of attracting users back to the site to answer questions in order to produce a statistically useful model.

## IV. BODY MASS INDEX INSTANTIATION AND RESULTS

In order to test this approach with an outcome that was more readily available to participants a second website was deployed in which models attempted to predict the body mass index of each participant. Body mass index (BMI) is calculated as mass(kg) / (height(m))$^2$ and, although it is known to have several limitations [36], is still the most common measure for determining a patient's level of obesity. Each user's BMI could readily be calculated as all users know and are thus able to immediately enter their height and weight. A second motivator for investigating this behavioral outcome is that obesity has been cited [37] as one of the major global public health challenges to date, it is known to have myriad causes [38], [39], and people with extreme BMI values are likely to have intuitions as to why they deviate so far from the norm.

Participants arriving for the first time at the BMI site were asked to enter their height and weight in feet, inches and pounds respectively, as most of the visitors to the site resided in the U.S. Participants were then free to respond to and pose new questions.

In order to further motivate the participants, in addition to displaying their predicted outcome, users were also shown how their responses compared to two peer groups. For each user the peer groups were constructed as follows. The first peer group was composed of 10 other users who had BMI values as close to but below that of the user; the second group was composed of 10 other users who had BMI values as close to but above that of the user. If $N < 10$ users could be found the peer group was composed of those $N$ users. The average BMI for each of the two peer groups, as well as the user's own BMI, were displayed (see Fig. 2). Also, the responses to each question, within each peer group, were averaged and shown alongside the user's response to that question. Finally, the 'predictive power' of each question was shown. Predictive power was set equal to the $r^2$ obtained when the responses to that question alone were regressed against the outcome.

The peer group data were meant to help users compare how their lifestyle choices measured up to their most similar peers who were slightly more healthy than themselves, and slightly less healthy than themselves. This approach in effect provides individualized suggestions to each user as to how slight changes in lifestyle choices may lead to improvements in the health indicator being measured. Presenting the user with the predictive power of each question was designed to help them learn what questions tend to be predictive, and thus motivate them to formulate new or better questions that might be even more predictive. For example one user posed the question "*How many, if any, of your parents are obese?*". Another user may realize that the 'predictive power' of this question (which achieved an $r^2$ in the actual experiment of 0.23 and became the sixth-most predictive question out of a total of 57) may be due to it serving as an indirect measure of the hereditary component of obesity. This may cause the user to pose a new question better tailored to eliciting this information, such as "*How many, if any, of your **biological** parents are obese?*" (a question of this form was not posed during the actual experiment).

The BMI site went live at 3:00pm EST on Friday, November 12, 2010, stayed live for slightly less than a week, and was discontinued at 10:20am EST on Thursday, November 18, 2010. During that time it attracted 64 users who supplied at least one response. Those users proposed 56 questions (in addition to the original seed question), and together provided 2021 responses to those questions.

Table II
LISTING OF THE 20 MOST PREDICTIVE QUESTIONS FROM THE BMI SITE.

| Index | Question | $r^2$ | Responses |
|---|---|---|---|
| 1 | Do you think of yourself as overweight? | 0.5524 | 43 |
| 2 | How often do you masturbate a month? | 0.3887 | 32 |
| 3 | What percentage of your job involves sitting? | 0.3369 | 57 |
| 4 | How many nights a week do you have a meal after midnight? | 0.2670 | 67 |
| 5 | You would consider your partner/boyfriend/girlfriend/spouse etc to be overweight? | 0.2655 | 24 |
| 6 | How many, if any, of your parents are obese? | 0.2311 | 57 |
| 7 | Are you male? | 0.2212 | 32 |
| 8 | I am happy with my life | 0.2062 | 31 |
| 9 | How many times do you cook dinner in an average week? | 0.2005 | 44 |
| 10 | How many miles do you run a week? | 0.1865 | 28 |
| 11 | Do you have a college degree? | 0.1699 | 12 |
| 12 | Do you have a Ph.D. | 0.1699 | 12 |
| 13 | Would you describe yourself as an emotional person? | 0.1648 | 30 |
| 14 | How often do you eat (meals + snacks) during a day | 0.1491 | 33 |
| 15 | How many hours do you work per week? | 0.1478 | 46 |
| 16 | Do you practice a martial art? | 0.1450 | 31 |
| 17 | What is your income? | 0.1419 | 55 |
| 18 | I was popular in high school | 0.1386 | 31 |
| 19 | Do you ride a bike to work? | 0.1383 | 64 |
| 20 | What hour expressed in 1-24 on average do you eat your last meal before going to bed? | 0.1364 | 30 |

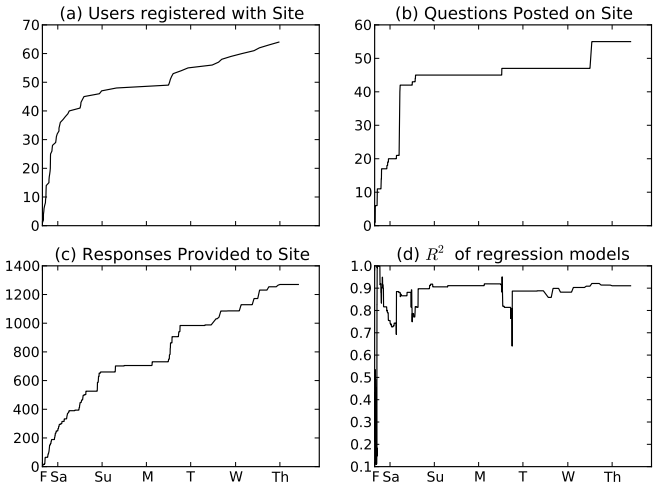



Figure 4. **User behavior on the BMI site.** The BMI site was maintained for slightly less than seven days. During that time it attracted 64 users (**(a)**) who together posted a total of 57 questions **(b)** and 2021 responses to those questions **(c)**. Every five minutes a regression model was constructed against the site's data: The quality of these models are shown as a function of their $R^2$ value **(f)**.

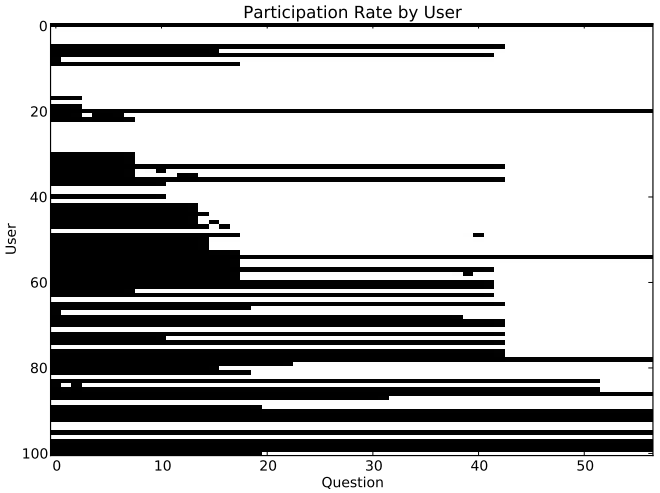



Figure 5. **Participation Rate by User of the BMI site.** Each row corresponds to a user of the BMI site, sorted by time of registration. Each column corresponds to one of the questions, sorted by time of posting. A black pixel at row $i$ column $j$ indicates that user $i$ responded to question $j$; a white pixel indicates they did not.

Users were recruited from reddit.com and the social networks of the principal investigators. Fig. 4a shows an initial burst of new users followed by a plateau during the weekend, and then a steady rise thereafter until the termination of the experiment. Fig. 4b shows a similar, initially rapid increase in the number of questions, and no significant increase until one user submits 8 new questions on day 6. Fig. 4c shows a relatively steady rise in the number of responses collected per day. This can be explained by the fact that although fewer users visit the site from the third day onward, there are more questions available when they do and thus, on average, more responses are supplied by later users than earlier users.

This increase is supplemented by a few early users who return to the site and respond to new questions, as shown in Fig. 5. It shows that of the 100 users who registered, only 57 supplied at least one response. The triangular form of the matrix is due to the fact that for the majority of users, they only visited the site once and answered the questions that were available at that time. This led to a situation in which questions posted early received disproportionally more responses than those questions posted later.

For the first several hours of the experiment the modeling engine (Fig. 1m-p) was run once every minute. At 5:30pm on November 12 the modeling engine was set to run once every five minutes. With the decrease in site activity the modeling engine was set to run once an hour starting at 2:20pm on November 16 until the termination of the experiment. Fig. 4d reports the $r^2$ value of the regression models as the experiment proceeded. During the first few hours of the experiment when there were more users than questions (see Fig. 4a,b), the early models had an $r^2$ near 1.0, suggesting that overfitting

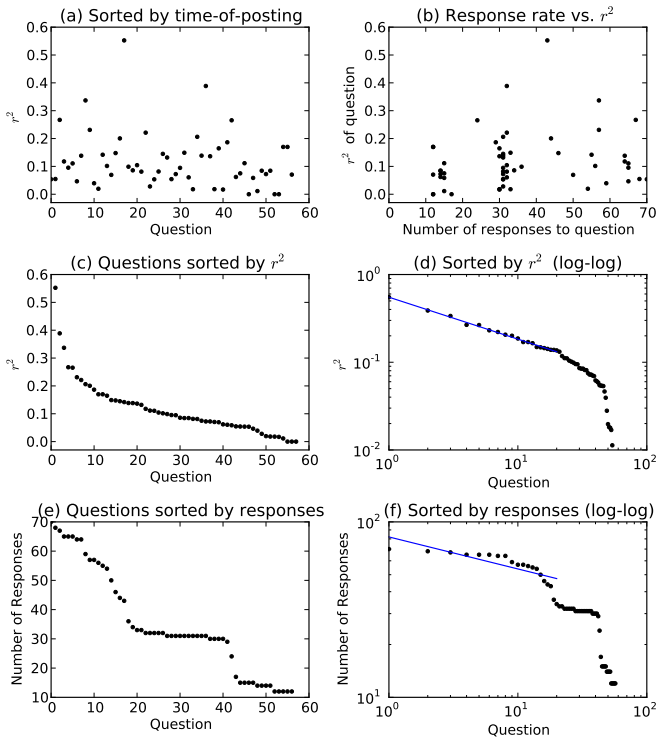



Figure 6. **BMI Question Statistics.** (**a,b**): No relationship was found between questions' time of posting, response rate or predictive power. However a power law relationship was discovered among questions' predictive power (**c,d**) but not for their response rate (**e,f**).

of the data was occurring. However at the termination of the experiment when there were more users (64) than questions (57)—and many users had not responded to those questions—the models were still performing well with an $r^2$ near 0.9. There is still a possibility though that the models overfit the data as the site was not instrumented with the ability to create a testing set composed of users whose responses were not regressed against.

Fig. 6 reports statistics about the user-posed questions. Fig. 6a shows that there is no correlation between when a question was posed and how predictive it became: the second- and fifth-most predictive question were posed as the 35th and 42nd question, respectively. Similarly, Fig. 6b reports the lack of correlation between the number of responses a question receives and its final predictive power. Although a slight positive correlation may exist, several of the most predictive questions (including the second- and fifth-most) received less than half of all possible responses.

Fig. 6c reports the questions sorted in order of decreasing $r^2$, and reveals that this distribution has a long tail: a large number of questions have low, but non-zero $r^2$ when regressed alone against the outcome. This distribution is replotted in Fig. 6c on a log-log scale. Linear regression was performed on the 20 most predictive questions (indicated by the line), and the resulting fit was found to be highly correlated, with $r^2 = 0.994$. This finding suggests that a power law relationship exists among these predictive questions[1]. It is possible that the power law exists because of an underlying power law relationship in the number of responses these questions attracted. However, Fig. 6b indicates there is little or no correlation between the number of responses a question attracts and its predictive power. Further, Fig. 6e reports the questions sorted by number of responses, and, when plotted on a log-log scale (Fig. 6f) shows that there is no power law ($r^2 = 0.65$) among the 20 most responded-to questions. This suggests the power law relationship among the most predictive questions has some other cause.

Table II reports the 20 most predictive questions, sorted by decreasing $r^2$. The questions span many of the classes of factors known (or hypothesized) to influence obesity, including demographic (q. 7 [41]), social or economic (qs. 5, 11, 12, 15, 17, 18 [39]), genetic (q. 6 [42]), psychological (qs. 1, 8, 13 [43], [44]) dietary (qs. 4, 9, 14, 20 [45]), and physical activity-related (qs. 2, 3, 10, 16, 19 [46]). This indicates that although the majority of participants are unlikely to be experts in the domain of interest, collectively they uncovered many of the classes of known correlates of obesity, and responded sufficiently honestly so that these correlates became predictive of BMI on the site. It could be argued that the most predictive question should not have been accepted as it is highly likely that it correlates with the outcome: people who perceive themselves as overweight are likely to be overweight. However, it is known that for those suffering from body image disorders the opposite is often the case: those that perceive themselves incorrectly as overweight can become extremely underweight [47]. Separating the auto- and anti-correlated components of this broad question could be accomplished by supplementing it with more targeted questions (eg., "*Do you think you are overweight but everyone else tells you the opposite?*").

Despite the lack of filtering on the site there were only a few cases of clearly dishonest responses. Fig. 2 indicates that at least one member of this user's peer group answered the fast food question dishonestly. It is interesting to note that this dishonest answer (or answers) was supplied for the seed question, and this question—despite collecting the most responses (70)—had nearly no individual correlation ($r^2 = 0.054$) and thus contributed negligibly to the predictions of the models. Questions 3, 4, 6, 9, 15, and 20 as shown in Table II have maximum possible values (qs. 3 max=100; qs. 4 and 9 max=7; qs. 6 max=2; qs. 15 max=168; qs. 20 max=24), and together collected 301 responses. Of those responses, none were above the maximum or below the minimum (min=0 for all qs.) indicating that all responses were not theoretically impossible. This suggests that clear dishonesty (defined as supplying a response below or above the theoretical minimum or maximum, respectively) was quite rare for this experiment. Conversely, unlike the popular yet corrupted seed question, these questions became significantly predictive as the experiment progressed. Further investigation into whether or how the

[1]The close linear fit for these questions does not guarantee that a power law exists among these questions, however [40]. Subsequent work and a larger data set will be required to confirm if power law relationships do indeed exist among user-generated questions predictive of a behavioral outcome.

rare cases of clear dishonesty (and the possibly larger amount of hidden dishonesty) affect modeling in such systems remains to be investigated.

## V. Discussion/conclusions

This paper introduced a new approach to social science modeling in which the participants themselves are motivated to uncover the correlates of some human behavior outcome, such as homeowner electricity usage or body mass index. In both cases participants successfully uncovered at least one statistically significant predictor of the outcome variable. For the body mass index outcome, the participants successfully formulated many of the correlates known to predict BMI, and provided sufficiently honest values for those correlates to become predictive during the experiment. While, our instantiations focus on energy and BMI, the proposed method is general, and might, as the method improves, be useful to answer many difficult questions regarding why some outcomes are different than others. For example, future instantiations might provide new insight into difficult questions like: "Why do grade point averages or test scores differ so greatly among students?", "Why do certain drugs work with some populations, but not others?", "Why do some people with similar skills and experience, and doing similar work, earn more than others?"

Despite this initial success, much work remains to be done to improve the functioning of the system, and to validate its performance. The first major challenge is that the number of questions approached the number of participants on the BMI website. This raises the possibility that the models may have overfit the data as can occur when the number of observable features approaches the number of observations of those features. Nevertheless the main goal of this paper was to demonstrate a system that enables non domain experts to collectively formulate many of the known (and possibly unknown) predictors of a behavioral outcome, and that this system is independent of the outcome of interest. One method to combat overfitting in future instantiations of the method would be to dynamically filter the number of questions a user may respond to: as the number of questions approaches the number of users this filter would be strengthened such that a new user is only exposed on a small subset of the possible questions.

### A. User Fatigue

Another challenge for this approach is user fatigue: Fig. 5 indicates that many of the later users only answered a small fraction of the available questions. Thus it is imperative that users be presented with questions that most require additional responses first. This raises the issue of how to order the presentation of questions. In the two instantiations presented here, questions were simply presented to all users in the same order: the order in which they were posted to the site. It was possible that this ordering could have caused a 'winner take all' problem in that questions that accrue more responses compared to other questions would achieve a higher predictive power, and users would thus be attracted to respond to these more predictive questions more than the less predictive questions. However, the observed lack of correlation between response rate and predictive power (Fig. 6b) dispelled this concern.

In future instantiations of the method, question ordering will be approached in a principled way. Instead of training a single model $m$, an ensemble of methods $m_1, ..., m_k$ will be trained on different subsets of the data [48], [49]. Then, query by committee [50] will be employed to determine question order: The question that induces maximal disagreement among the $k$ models as to its predictive power will be presented first, followed by the question that induces the second largest amount of disagreement, and so on. In this way questions that may be predictive would be validated more rapidly than if question ordering is fixed, or random.

### B. User Motivation

Typically, human subjects play a passive role in social science studies, regardless of whether that study is conducted offline (pen-and-paper questionnaire) or online (web-based survey): They contribute responses to survey questions, but play no role in crafting the questions. This work demonstrates that users can also contribute to the hypothesis-generation component of the discovery process: Users can collectively contribute—and populate—predictors of a behavioral outcome.

It has been shown here that users can be motivated to do this without requiring an explicit reward: The subjects were unpaid for both studies. Much work remains to be done to clarify under what conditions subjects will be *willing* and *able* to contribute predictors.

We hypothesize that *willingness* to generate candidate predictors of a behavioral outcome may be stimulated under several conditions. First, if subjects are incurring a health or financial cost as a result of the outcome under study, they may be motivated to contribute. For example a user that has an above average electricity bill or body mass index, yet has similar lifestyle attributes as his fellow users, may wish to generate additional attributes to explain the discrepancy.

Conversely, a user that posts a superior outcome (i.e. a low electricity bill or very healthy body mass index) may wish to uncover the predictor that contributes to their superior outcome (i.e. a well-insulated house or good exercise regimen) and thus advertise it to their peers. This may act as a form of online 'boasting', a well known motivator among online communities.

In the current studies, some participants may have been motivated to contribute because they were part of the authors' social networks. However, a substantial number of users were recruited from online communities outside of the authors' social networks, indicating that some online users are motivated to contribute to such studies even if they do not know those responsible for the study. The exact number of users in these two groups is not clear on account of the anonymity requirements stipulated for these human subject studies.

Similarly, a non domain expert's *ability* to contribute a previously unknown yet explanatory predictor of a behavioral

outcome may rely on them suffering or benefiting from a far-from-average outcome. For example consider someone who is extremely underweight yet their outcome is not predicted by the common predictors of diet and exercise: this user has a high caloric intake and does not exercise. This user may be able to generate a predictor that a domain expert may not have though of, yet is predictive for a certain underweight demographic: this user may ask her peers: "Are you in an abusive relationship?"

Users may also be motivated to contribute to such studies because it provides entertainment value: users may view the website as a competitive game in which the 'goal' is to propose the best questions. In a future version we plan to create a dynamically sorted list of user-generated questions: questions bubble up to the top of the list if (1) it is a question that many other users wish to respond to, (2) it is orthogonal to the other questions, and (3) it is found to be predictive of the outcome under study. Users may then compete by generating questions that climb the leaderboard and thus advertise the user's understanding of the outcome under study.

### C. Rare Outcomes

Obesity and electricity usage are well-studied behavioral outcomes. It remains to be seen though how the proposed methodology would work for outcomes that affect a small minority of online users, or for which predictors are not well known.

We hypothesize that for rare outcomes, online users who have experience with this outcome, could be encouraged to participate, and would be intrinsically motivated to contribute. For example if the outcome to be studied were a rare disease, users who suffer from the disease would be attracted to the site. Once there, they may be in a unique position to suggest and collectively discover previously unknown predictors of that disease. Moreover, a user who suffers from the disease is likely to know more people who suffer from that disease and would be motivated to advertise the site to them. Finally, even if a user discovers the site and does not suffer from the disease, he may know someone who does and thus introduce the site to that person. Such a person may serve as a caregiver for someone suffering from the disease, such as a family member. A caregiver may be able to contribute novel predictors that are different from those proposed by the sufferer himself.

Thus, a website that hosts such a rare outcome may serve as a 'magnet' for people who exhibit the outcome or know people that do. In future we will study the 'attractive force' of such websites: if such a website experiences increased user traffic as the study goes forward, and the average outcome of users on the site drifts away from the global population's mean value for this outcome, that would indicate that a growing number of people with such an outcome are being attracted to the site.

In closing, this paper has presented a novel contribution to the growing field of machine science in which the formulation of observables for a modeling task—and the *populating* of those observables with values—can be offloaded to the human group being modeled.

## VI. Acknowledgement

The authors acknowledge valuable contributions from three anonymous reviewers, and useful discussions with collaborators in the UVM Complex Systems center.

## References

[1] J. Bongard and H. Lipson, "Automated reverse engineering of nonlinear dynamical systems," *Proceedings of the National Academy of Sciences*, vol. 104, no. 24, pp. 9943–9948, 2007.

[2] J. Evans and A. Rzhetsky, "Machine science," *Science*, vol. 329, no. 5990, p. 399, 2010.

[3] R. D. King, K. E. Whelan, F. M. Jones, P. G. K. Reiser, C. H. Bryant, S. H. Muggleton, D. B. Kell, and S. G. Oliver, "Functional genomic hypothesis generation and experimentation by a robot scientist," *Nature*, vol. 427, pp. 247–252, 2004.

[4] R. King, J. Rowland, S. Oliver, M. Young, W. Aubrey, E. Byrne, M. Liakata, M. Markham, P. Pir, L. Soldatova *et al.*, "The automation of science," *Science*, vol. 324, no. 5923, p. 85, 2009.

[5] J. Bongard, V. Zykov, and H. Lipson, "Resilient machines through continuous self-modeling," *Science*, vol. 314, pp. 1118–1121, 2006.

[6] J. Giles, "Internet encyclopedias go head to head," *Nature*, vol. 438, no. 15, pp. 900–901, 2005.

[7] D. C. Brabham, "Crowdsourcing as a model for problem solving," *Convergence*, vol. 14, pp. 75–90, 2008.

[8] A. Sorokin and D. Forsyth, "Utility data annotation with amazon mechanical turk," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2008.

[9] M. Marge, S. Banerjee, and A. Rudnicky, "Using the amazon mechanical turk for transcription of spoken language," in *Proc. IEEE International Conference on Acoustics Speech and Signal Processing*, 2010.

[10] N. Kong, J. Heer, and M. Agrawala, "Perceptual guidelines for creating rectangular treemaps," *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, no. 6, 2010.

[11] A. Kittur, E. Chi, and B. Suh, "Crowdsourcing user studies with mechanical turk," in *Proc.Twenty-sixth annual SIGCHI conference on human factors in computing systems*, 2008.

[12] D. Wightman, "Crowdsourcing human-based computation," in *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries*, 2010.

[13] B. Fitzgerald, "The transformation of open source software," *Management Information Systems Quarterly*, vol. 30, no. 3, pp. 587–598, 2006.

[14] J. Howe, *Crowdsourcing: Why the Power of the Crowd is Driving the Future of Business*. Crown Business, 2009.

[15] N. Thurman, "Forums for citizen journalists? adoption of user generated content initiatives by online news media," *New Media and Society*, vol. 10, no. 1, 2008.

[16] C. DiBona, M. Stone, and D. Cooper, *Open Source 2.0: The Continuing Evolution*. O'Reilly Media, 2005.

[17] J. Leskovec, L. Adamic, and B. Huberman, *The Dynamics of Viral Marketing*. New York: ACM Press, 2007.

[18] K. Lerman, "Social networks and social information filtering on digg," arXiv: cs/0612046v1, 2006.

[19] D. Anderson, J. Cobb, E. Korpela, M. Lebofsky, and D. Werthimer, "Seti@home: an experiment in public-resource computing," *Communications of the ACM*, vol. 45, no. 11, pp. 56–61, 2002.

[20] J. Cohn, "Citizen science: Can volunteers do real research?" *BioScience*, vol. 58, no. 3, pp. 192–197, 2008.

[21] J. Silvertown, "A new dawn for citizen science," *Trends in Ecology & Evolution*, vol. 24, no. 9, pp. 467–471, 2009.

[22] A. Beberg, D. Ensign, G. Jayachandran, S. Khaliq, and V. Pande, "Folding@home: Lessons from eight years of volunteer distributed computing," in *IEEE International Symposium on Parallel Distributed Processing*, May 2009, pp. 1–8.

[23] C. Lintott, K. Schawinski, A. Slosar, K. Land, S. Bamford, D. Thomas, M. Raddick, R. Nichol, A. Szalay, D. Andreescu *et al.*, "Galaxy zoo: morphologies derived from visual inspection of galaxies from the sloan digital sky survey?" *Monthly Notices of the Royal Astronomical Society*, vol. 389, no. 3, pp. 1179–1189, 2008.

[24] S. Cooper, F. Khatib, A. Treuille, J. Barbero, J. Lee, M. Beenen, A. Leaver-Fay, D. Baker, Z. Popović *et al.*, "Predicting protein structures with a multiplayer online game," *Nature*, vol. 466, no. 7307, pp. 756–760, 2010.

[25] A. Kittur, "Crowdsourcing, collaboration and creativity," *XRDS*, vol. 17, no. 2, pp. 22–26, 2010.

[26] S. Bowman, S. Gortmaker, C. Ebbeling, M. Pereira, and D. Ludwig, "Effects of fast-food consumption on energy intake and diet quality among children in a national household survey," *Pediatrics*, vol. 113, no. 1, p. 112, 2004.

[27] J. Currie, S. DellaVigna, E. Moretti, and V. Pathania, "The effect of fast food restaurants on obesity and weight gain," *American Economic Journal: Economic Policy*, vol. 2, no. 3, pp. 32–63, 2010.

[28] S. Z. Attari, M. L. DeKay, C. I. Davidson, and W. B. de Bruinc, "Public perceptions of energy consumption and savings," *Proceedings of the National Academy of Sciences*, Aug. 16 2010.

[29] Microsoft. (2011) Microsoft hohm. [Online]. Available: http://www.microsoft-hohm.com/

[30] E. Mills, "The home energy saver: Documentation of calculation methodology, input data, and infrastructure," Lawrence Berkeley National Laboratory, Tech. Rep. LBNL-51938, 2008.

[31] H. Allcott, "Social norms and energy conservation," *Journal of Public Economics*, 2011.

[32] J. E. Petersen, V. Shunturov, K. Janda, G. Platt, and K. Weinberger, "Dormitory residents reduce electricity consumption when exposed to real-time visual feedback and incentives," *International Journal of Sustainability in Higher Education*, vol. 8, no. 1, pp. 16–33, 2007.

[33] L. Kaufman, "Utilities turn their customers green, with envy," *The New York Times*, Jan. 30 2009.

[34] P. Slovic, "Trust, emotion, sex, politics, and science: Surveying the risk-assessment battlefield," *Risk Analysis*, vol. 19, no. 4, 1999.

[35] G. S. Guthridge, "Understanding consumer preferences in energy efficiency: Accenture end-consumer observatory on electricity management," Accenture, Tech. Rep. ACC10-0229, 2010.

[36] A. Romero-Corral, V. Somers, J. Sierra-Johnson, R. Thomas, M. Collazo-Clavell, J. Korinek, T. Allison, J. Batsis, F. Sert-Kuniyoshi, and F. Lopez-Jimenez, "Accuracy of body mass index in diagnosing obesity in the adult general population," *International Journal of Obesity*, vol. 32, no. 6, pp. 959–966, 2008.

[37] L. Barness, J. Opitz, and E. Gilbert-Barness, "Obesity: genetic, molecular, and environmental aspects," *American Journal of Medical Genetics Part A*, vol. 143, no. 24, pp. 3016–3034, 2007.

[38] T. Parsons, C. Power, S. Logan, and C. Summerbell, "Childhood predictors of adult obesity: a systematic review." *International journal of obesity and related metabolic disorders: journal of the International Association for the Study of Obesity*, vol. 23, p. S1, 1999.

[39] Y. Wang and M. Beydoun, "The obesity epidemic in the United States—gender, age, socioeconomic, racial/ethnic, and geographic characteristics: a systematic review and meta-regression analysis," *Epidemiologic reviews*, vol. 29, no. 1, p. 6, 2007.

[40] A. Clauset, C. Rohilla Shalizi, and M. Newman, "Power-law distributions in empirical data," *SIAM review*, vol. 51, no. 4, pp. 661–703, 2009.

[41] P. Boumtje, C. Huang, J. Lee, and B. Lin, "Dietary habits, demographics, and the development of overweight and obesity among children in the United States," *Food Policy*, vol. 30, no. 2, pp. 115–128, 2005.

[42] A. Herbert, N. Gerry, M. McQueen, I. Heid, A. Pfeufer, T. Illig, H. Wichmann, T. Meitinger, D. Hunter, F. Hu *et al.*, "A common genetic variant is associated with adult and childhood obesity," *Science*, vol. 312, no. 5771, p. 279, 2006.

[43] M. Friedman and K. Brownell, "Psychological correlates of obesity: Moving to the next research generation." *Psychological Bulletin*, vol. 117, no. 1, p. 3, 1995.

[44] M. Van der Merwe, "Psychological correlates of obesity in women," *International Journal of Obesity*, vol. 31, pp. S14–S18, 2007.

[45] R. Bonow and R. Eckel, "Diet, obesity, and cardiovascular risk," *N Engl J Med*, vol. 348, no. 21, pp. 2057–2058, 2003.

[46] R. Ewing, T. Schmid, R. Killingsworth, A. Zlot, and S. Raudenbush, "Relationship between urban sprawl and physical activity, obesity, and morbidity," *Urban Ecology*, pp. 567–582, 2008.

[47] S. Grogan, *Body image: Understanding body dissatisfaction in men, women, and children*. Taylor & Francis, 2008.

[48] M. Skurichina and R. Duin, "Bagging, boosting and the random subspace method for linear classifiers," *Pattern Analysis & Applications*, vol. 5, no. 2, pp. 121–135, 2002.

[49] Z. Lu, X. Wu, and J. Bongard, "Active learning with adaptive heterogeneous ensembles," in *Data Mining, 2009. ICDM'09. Ninth IEEE International Conference on*. IEEE, 2009, pp. 327–336.

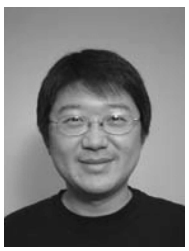
[50] H. Seung, M. Opper, and H. Sompolinsky, "Query by committee," in *Proceedings of the fifth annual workshop on Computational learning theory*. ACM, 1992, pp. 287–294.


**Josh C. Bongard (M'06)** is an Assistant Professor in the Department of Computer Science at the University of Vermont. Prior to this appointment, he was a postdoctoral researcher in the Sibley School of Mechanical and Aerospace Engineering at Cornell University. He received the B.Sc. honors degree in computer science from McMaster University, Canada, in 1997, and a M.Sc. in evolutionary and adaptive systems from the School of Cognitive and Computing Sciences at University of Sussex in 1999. He obtained his Ph.D. from the Artificial Intelligence Laboratory at the University of Zurich, for research in the field of evolutionary robotics. He was named a Microsoft New Faculty Fellow, one of the top 35 innovators under the age of 35 by MIT's Technology Review Magazine, and is a National Science Foundation CAREER award recipient.

**Paul Hines (M'07)** is an Assistant Professor in the School of Engineering at the University of Vermont. He is also a member of the Carnegie Mellon Electricity Industry Center Adjunct Research Faculty and a commissioner for the Burlington Electric Department. He received the Ph.D. in Engineering and Public Policy from Carnegie Mellon U. in 2007 and the M.S. degree in Electrical Engineering from the U. of Washington in 2001. Formerly he worked at the US National Energy Technology Laboratory, where he participated in Smart Grid research, the US Federal Energy Regulatory Commission, where he studied interactions between nuclear power plants and power grids, Alstom ESCA, where he developed load forecasting software, and for Black and Veatch, where he worked on substation design projects. His main research interests are in the areas of complex systems and networks, cascading failures in power systems, wind integration and energy security policy.

**Dylan Conger** received the B.S. degree in Computer Science at the University of Vermont in 2011. He worked on the BMI project as an undergraduate research assistant in 2010.

**Peter Hurd** received the M.S. degree in Computer Science from the University of Vermont in 2010. He worked on the EnergyMinder project as a graduate research assistant in 2009.



**Zhenyu Lu** is a Ph.D candidate in Computer Science at the University of Vermont. His main research interests are active learning and ensemble learning. He has published in various data mining forums, including ACM SIGKDD, IEEE ICDM and SIAM SDM. He is currently working as a machine learning specialist at Sears Holdings Corporation (SHC) with focuses on Web information exploration.