International Journal of Research (IJR) Vol-1, Issue-6, July 2014 **ISSN 2348-6848**

# CrowdTransfaring Mediators of Behavioral Results

**B.hyndavathi[1] & V.lakshmichaitanya[2]**

1. Dept. of MCA, JNTUA University, Balapanuru road, Kurnool (dist.), AP, Santhiram Engg college, Hyndavathi7@gmail.com.

2. (Asst. prof) Santhiram Engg College, Dept. of CSE, JNTUA University, Balapanuru road, Kurnool (dist.), AP, Chaitu223@gmail.com

Abstract—

*Generating models from large data sets—and deter-mining 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*—Crowdtranfaring, 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. For example, a survey designer must have domain expertise to choose questions that will identify predictive covariates.

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.

Machine science 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 as well as algorithms that cycle through hypothesis generation, experimental design, experiment execution, and hypothesis refutation.. 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 project 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.

*Crowdtranfaring*

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. Harnessing the experience and effort of large numbers of individuals is frequently known as “crowdtranfaring” and has been used effectively in a number of research and commercial applications . For an example of how crowdtranfaring can be useful, consider Amazon's Mechanical Turk. In this crowd-sourcing tool a human describes a “Human Intelligence Task” such as characterizing data transcribing spoken linguae or creating data visualizations. 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.

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.

. 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. Some platforms allow users to actively participate by searching for items of interest or solve problems through a game interface. 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 crowdtranfaring can pro-duce novel, creative solutions that are substantially different from those produced by experts. An iterative, crowdtransfered poem translation task produced translations that were both surprising and preferable to expert translations. We conjecture that crowdtransfaring the selection of predictive variables can reveal creative, unexpected predictors of behavioral out-comes. 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.

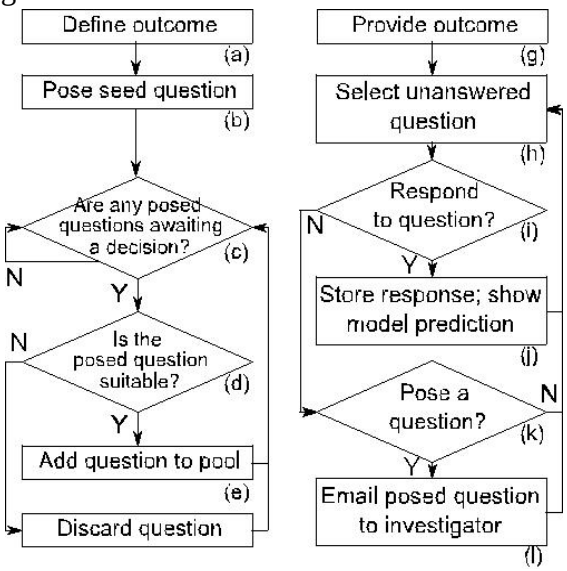
## II. METHODOLOGY

The system described here wraps a human behavior modeling paradigm in cyber infrastructure 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 we seeded the BMI website with the question "How *many times a week do you eat fast food?*"

Algorithm:-

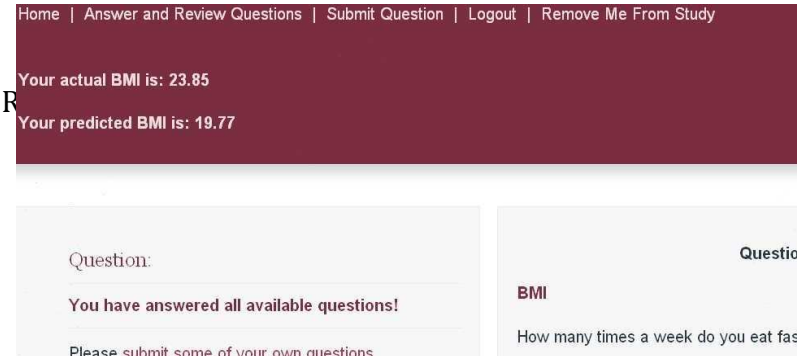


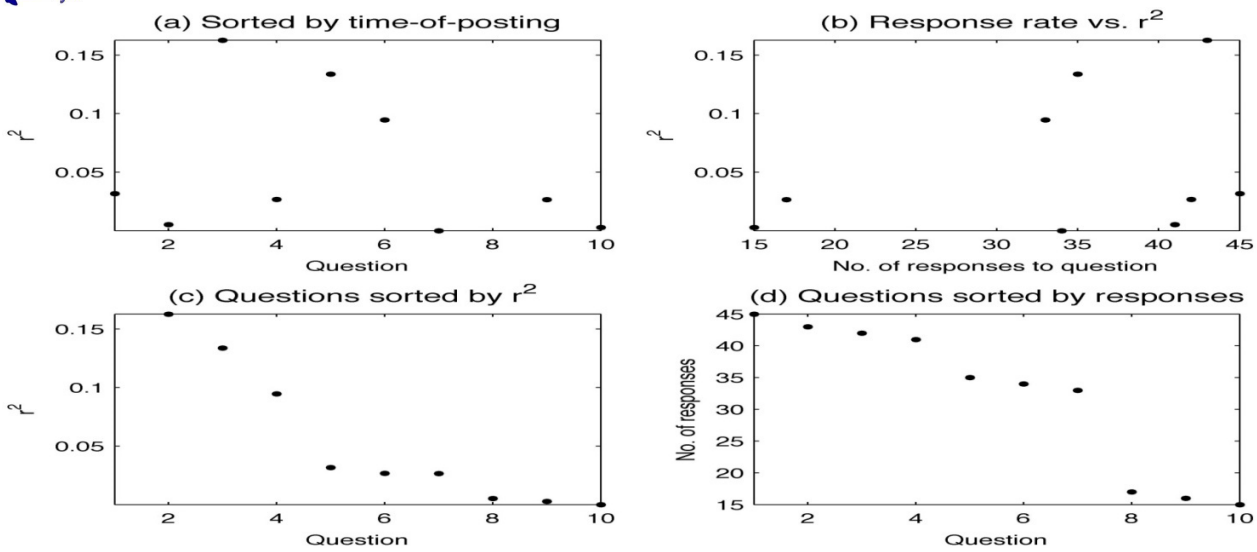

Users who visit the site first provide their individual value for the outcome of interest, such as their own BMI. Users may then respond to questions found on the site. Their answers are stored in a common data set and made available to the modeling engine. Periodically the modeling engine wakes up (Fig.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

Figure 1. **Overview of the figure.** 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 C of length k + 1 that contains the model parameters. It also outputs a vector 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 A against the response vector 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 otherwise the question was discarded. 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 ($\hat{}^i$). Fig. 2 illustrates that visitors to the b
BMI site were shown their actual BMI (as entered by them) and their predicted BMI. The models were able to predict each consumers themselves, in a bottom-up fashion, may have value in terms of motivating energy efficient behavior.

Thus motivated, we designed the "Energy Minder" 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 newsletter, 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

user's outcome before they had responded to every question by substituting in missing values. Thus after each response from a user

$$\hat{b}_i = c_0 + c_1 a_{i1} + c_2 a_{i2} + \ldots + c_k a_{ik} + \varrho_i \quad (1)$$

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.
Therefore, information generated largely by energy

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.

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 predictor s, and included only $Q_3$, $Q_5$, and $Q_6$. Model 2 showed the same pair of statistically significant predictors (Q3 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.

While the small sample size in this study limits the generality of these results, this initial trial provided useful information about the crowdtranfered 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.

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 | |

## V1. 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 correlate s 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.

## V. 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.