

Testing to Improve User Response of Crowdsourced S&T Forecasting System

SYST-699 Capstone Project Final Report

May 5th, 2014

Sponsors:

Charles Twardy, GMU C4I Center
Adam Siegel, Inkling Markets

Team Members:

Kevin Connor
Andrew Kreeger
Neil Wood

Contents

1. Executive Summary.....	3
2. Introduction and Problem Statement.....	4
3. Scope.....	5
4. Project Requirements.....	8
5. Technical Approach.....	9
6. Experimental Results and Recommendations.....	11
7. Future Work.....	12
8. References.....	13
A. Appendices.....	13

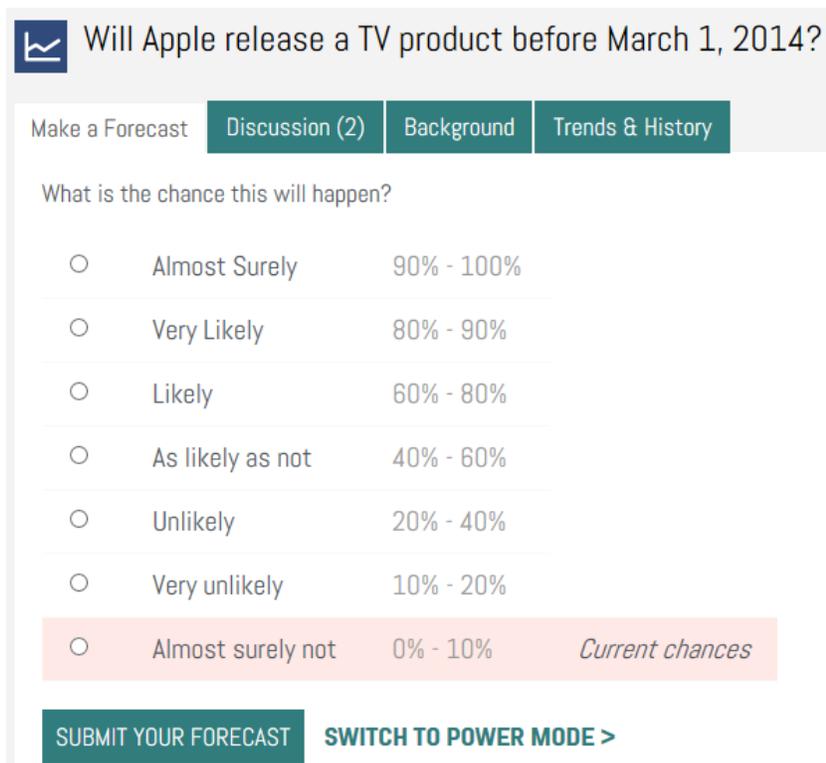
1. Executive Summary

Complete with Appropriate text.

2. Introduction and Problem Statement

The importance of the ability to accurately forecast outcomes of key issues in science and technology should be obvious to the majority of individuals. What is not as obvious is how one would accomplish this feat. SciCast was designed for this purpose. SciCast is a research project, which according to its website, is “run by George Mason University and sponsored by the U.S. Government to forecast the outcomes of key issues in science and technology.” The concept of SciCast is built around the notion that information collected from many informed individuals is often better at forecasting an outcome than that same information collected from a handful of experts in the subject. This concept is often referred to as crowdsourcing.

Users who wish to participate must first register with scicast.org, after which they can begin making forecasts. Once an individual has registered for SciCast, he or she is then free to search through questions on the site and make forecasts on those questions as desired. Figure 1 illustrates a forecasting example on the SciCast website.



The screenshot shows a forecasting interface on the SciCast website. At the top, there is a question: "Will Apple release a TV product before March 1, 2014?". Below the question, there are four tabs: "Make a Forecast", "Discussion (2)", "Background", and "Trends & History". The "Make a Forecast" tab is selected. Below the tabs, there is a question: "What is the chance this will happen?". There are seven radio button options with corresponding probability ranges:

Option	Probability Range
<input type="radio"/> Almost Surely	90% - 100%
<input type="radio"/> Very Likely	80% - 90%
<input type="radio"/> Likely	60% - 80%
<input type="radio"/> As likely as not	40% - 60%
<input type="radio"/> Unlikely	20% - 40%
<input type="radio"/> Very unlikely	10% - 20%
<input type="radio"/> Almost surely not	0% - 10%

The "Almost surely not" option is highlighted in a light red color and has the text "Current chances" next to it. At the bottom of the interface, there are two buttons: "SUBMIT YOUR FORECAST" and "SWITCH TO POWER MODE >".

Figure 1 – SciCast Forecasting Example

When utilizing the power of crowdsourcing, a key variable to developing accurate forecasts is the number of forecasts made by the “crowd”. In general, crowdsourced estimates are more accurate when more people participate. According to our sponsor, Dr. Charles Twardy, a key goal of SciCast is to provide enough forecast data for the SciCast team to study and analyze in order to increase SciCast’s forecasting accuracy. This can only be achieved by eliciting a significant number of user forecasts from an informed and diverse group of individuals working in the fields of science and technology.

The number of forecasts is determined by the number of registered users and the average number of forecasts made by each registered user. Our project sponsors would like us to propose, evaluate, and finally recommend ideas that will increase the number of registered users and the average number of forecasts made by each registered user. The remainder of this final report will outline the project’s scope and technical approach for achieving our sponsor’s goals. Experimental results will be provided to support our project team’s recommendations.

3. Project Scope

3.1 Overview and Rationale

3.1.1 Increasing User Registration

Throughout the process of developing this proposal, the group discussed several different methods that could be used to increase the number of registered users on SciCast. For instance, the SciCast user base could be increased by attracting participants to the website from Professional Societies, Universities, and from other interest groups. One approach for accomplishing this would be to advertise the SciCast website in the scientific journals of professional societies. Another approach would be to embed a description of, and link to, the SciCast website in the websites of various Universities and professional social media sites.

Potential users of SciCast may be directed to the SciCast website (via one of the methods mentioned above) and decide to leave the page without creating an account on SciCast. One approach for solving this problem, which would increase the SciCast user base, would be to reduce the website bounce rate and to provide motivations for potential users to stay and create an account on SciCast. Improving the web design aesthetics for the SciCast splash page could potentially solve these problems. Toward this end, the SciCast team has created a new version of the SciCast splash page. Figure 2 illustrates the original and updated versions of the SciCast splash page.

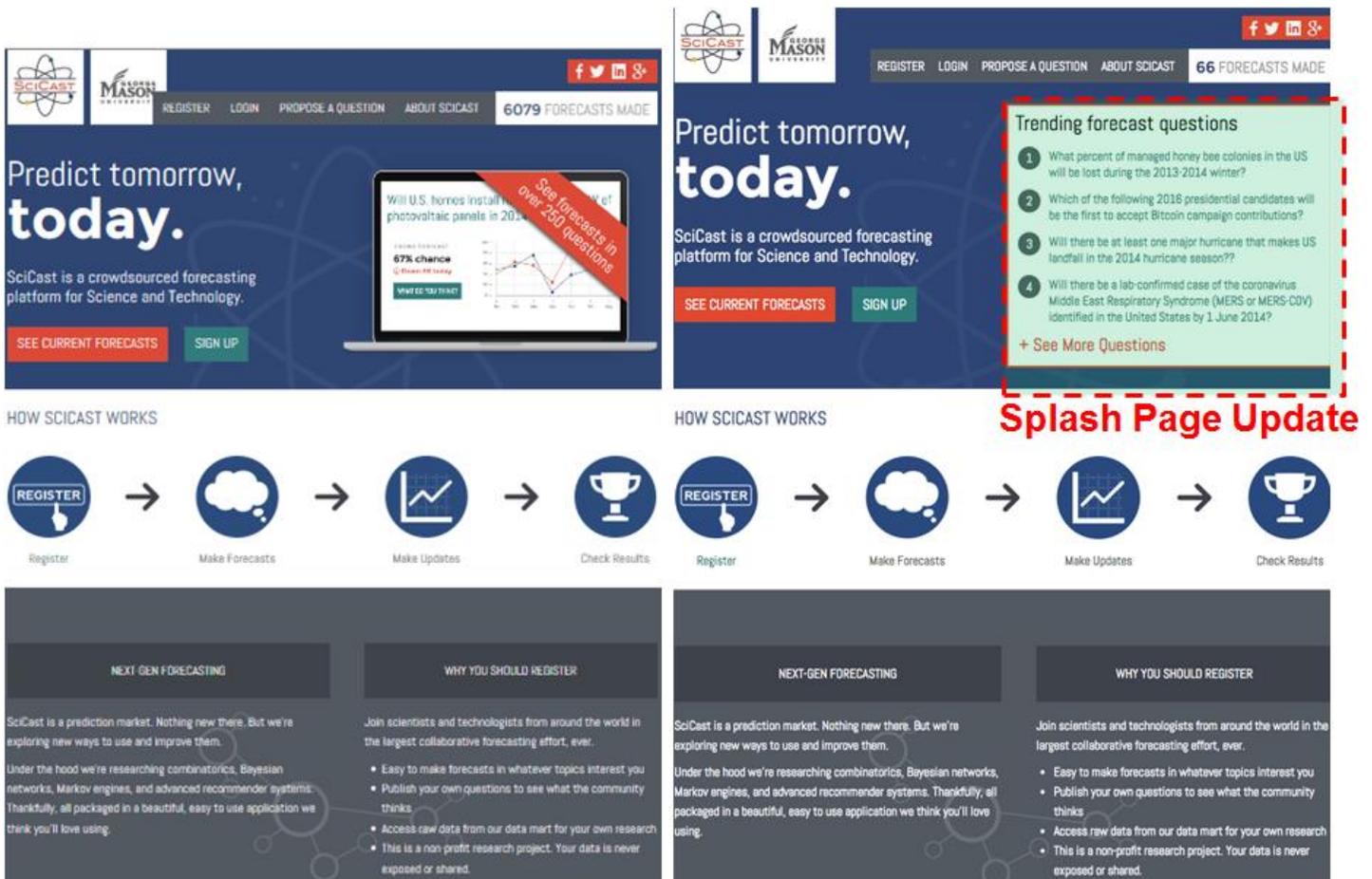


Figure 2 – Original SciCast Splash Page (Left) and Updated SciCast Splash Page (Right)

3.1.2 Increasing User Participation

Another approach for increasing the number of user forecasts would be to increase the average number of forecasts made by SciCast users. Simply increasing the number of registered users does not guarantee an increase in the number of forecasts made. Attracting new registrants without the proper background and/or who do not make forecasts will not improve the number of forecasts made. The SciCast team has proposed a variety of user interface (UI) design modifications for the SciCast web site which will hopefully increase the average number of forecasts made by each registered user. One UI design modification which was evaluated by our project team is the addition of a recommender box to the SciCast initial screen. The recommender box contains a sorted list of forecast questions considered relevant to the SciCast User. This list is determined based on an algorithm being developed by the SciCast team. The recommender box is placed in the SciCast initial screen as illustrated by Figures 3.

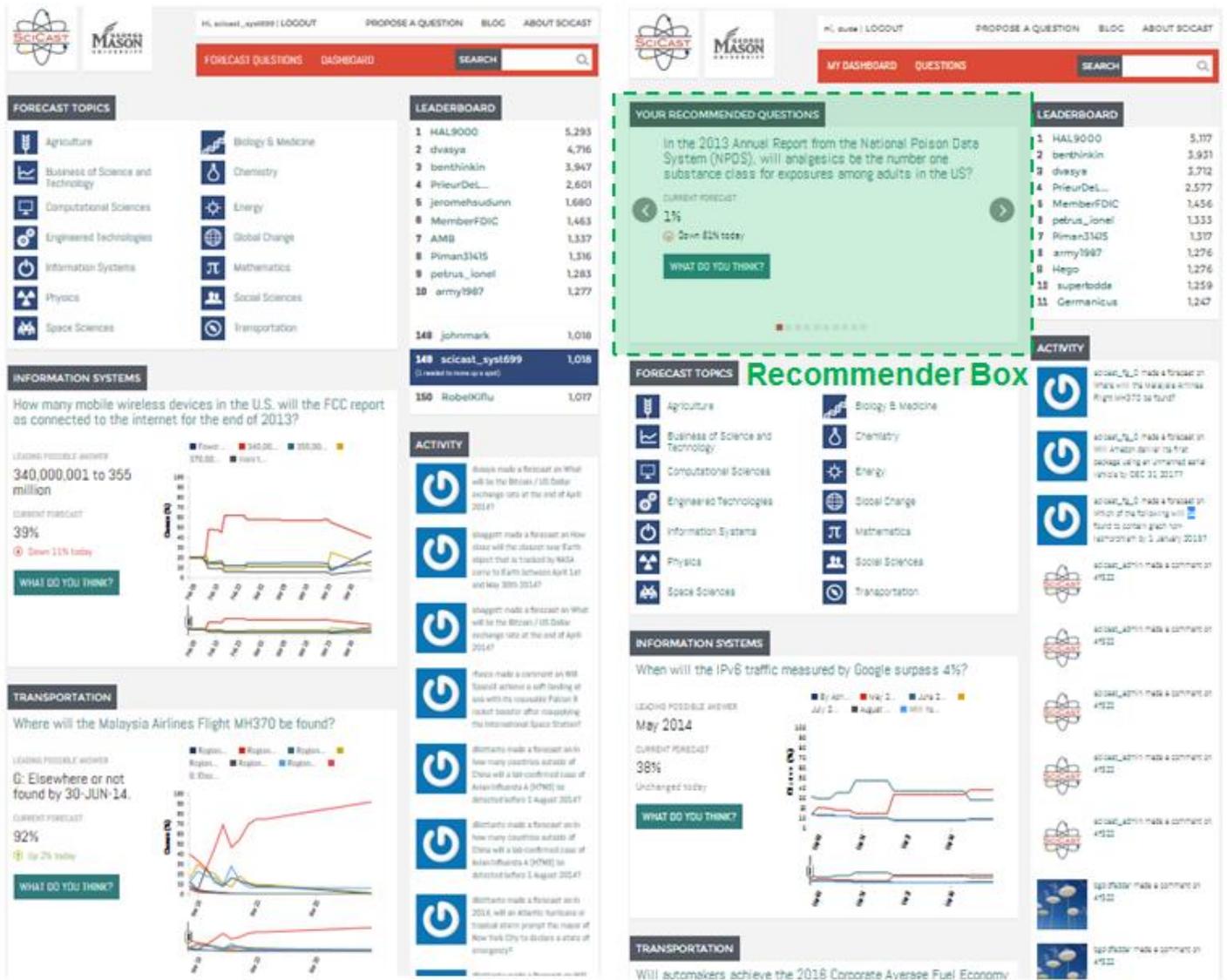


Figure 3 – SciCast Initial Screen (Left) and Proposed Insertion of Recommender Box (Right)

3.2 Summary

In the previous sections, two web design modifications have been proposed for increasing the number of registered users and the average number of forecasts made by each registered user. Our project sponsor has implemented both of these website modifications on test websites which our team has access to. Our project team has evaluated both of these modifications in order to determine if these modifications result in increased registration and user participation rates. Experiment details and results will be presented in the following sections.

4. Project Requirements

Below are the finalized requirements for this project. Due to sponsor input, these requirements have changed from the preliminary requirements given in our project proposal.

1. *Experiment Requirements*

- 1.1 A/B testing will be coordinated with the project sponsors.
- 1.2 Focus group testing will be conducted in accordance with the George Mason University and American Psychological Association standards.
- 1.3 CITI Training will be successfully completed.
- 1.4 Focus Group Volunteers will be recruited.

2. *Deliverable Requirements*

- 2.1 Two progress reports will be delivered on March 4th, 2014 and March 25th, 2014.
- 2.2 Final results will be delivered on May 5th, 2014 via a website created by the group.
- 2.2 Final results will be delivered in a final written report delivered on May 5th, 2014.
- 2.3 Final results will be presented to faculty and sponsors on May 9th, 2014.
- 2.4 Final Report / Final Presentation requirements:
 - 2.4.1 Final Report will be at least 20 pages long.
 - 2.4.2 Presentation will be approximately 25 minutes long.
 - 2.4.3 Final Report/Final Presentation will contain suggestions from users in the focus group that would make the recommender box more useful/effective.
 - 2.4.4 Final Report/Final Presentation will answer the following questions:
 - 2.4.3.1 Did the recommender box improve user participation?
 - 2.4.3.2 Is the recommender algorithm effective?
 - 2.4.3.3 Why or why not was the recommender box successful?
 - 2.4.3.4 Did the Splash Page update increase user interaction?

5. Technical Approach

5.1 Overview

As discussed in section 3, our team evaluated the web design modifications proposed by our project sponsor. The original plan was to design and conduct an A/B/C hypothesis test and a focus group test with a particular focus on the recommender box mentioned in section 3.1.2. Toward this end, an A/B/C hypothesis test and a focus group test were designed to evaluate the recommender box.

In late March, our sponsor informed us that the recommender box would not be available on the SciCast production site in time for us to perform our A/B/C test. Consequently, we finished the A/B/C experiment design (for a future team to use), developed a focus group test to evaluate the SciCast website (including the recommender box), and designed an A/B hypothesis test to evaluate the effect of the updated splash page on the user registration rate. The design of all of the aforementioned experiments will be given in the following sections.

5.2 Experiment Design

5.2.1 Recommender Box A/B/C Hypothesis Test

As discussed in section 3.1.2, the SciCast team has implemented a recommender box which is viewable by users logged into the test site (the recommender box will migrate to the production site after it has passed pre-production tests). The notion behind the recommender box is that it would suggest, or recommend, questions that the user may be interested in providing forecasts to. It is currently believed that providing interesting questions to the user via the recommender box will increase the number of forecasts that each user provides. The following paragraphs will outline the technical approach to addressing the following questions: 1) Does the recommender box increase the number of user forecasts? 2) Does the algorithm that creates the recommender work? and, 3) Why or why not is this the case?

The notion of determining the effects of a proposed change to a website falls directly in line with the concept of A/B testing and its generalized equivalents. A/B Testing is a controlled experiment for establishing a causal relationship between changes and their influence on user behavior¹. A well-conducted A/B, or A/B/C, test performed on the users of SciCast should provide the data necessary to conclude whether or not the implementation of the recommender box increases user forecasts.

When the A/B, or A/B/C, tests have been completed, the test results will be analyzed and a small, post A/B/C test focus group will be conducted if the analysis suggests that it will be beneficial.

The results of the A/B/C testing can potentially provide answers to whether or not the recommender box affects the number of forecasts performed by users and whether or not the recommender box algorithm provides useful recommendations. However, those results will not provide any insight into why the users acted as they did. If the A/B/C results are conclusive, a small focus group study could provide insight into the “why”.

Part of a well-designed A/B test is ensuring that there is only one change between the A and B group websites. This set up allows the experimenter to conclude that statistically significant differences between the two sites must be due to the singular change and not to any other factors. To this end, all test groups will be visually laid out the same, i.e. the same objects in the same position for all groups.

Per sponsor request, we have designed an A/B/C test in order to evaluate the recommender box. For this test, users will be assigned to either the control group (A), the treatment group (B), or a second treatment group (C). The control, or A group, will have only minimal changes with respect to the current site. The control group may use a “placebo” recommender box which does not provide any recommendations. The purpose of this group is to compare the new version of the site to a minimally modified version to see if the new version caused changes in user behavior. The B version of the site will implement the recommender. Finally, the C version will implement a recommender with random questions, not recommended questions. It can be concluded that having a recommender results in increased user activity if the B or C groups show increased user activity. It can also be concluded that the recommender outperforms a random question generator if the B version outperforms the C version.

In performing an A/B/C test, users can be randomly assigned or assigned by strata to anyone of the three test groups. For any A/B/C test, it is desired to prevent experimental outliers from randomly clustering in a single test group (i.e. all outliers land in the A or B or C group). Outlier clustering can become a significant possibility when user activity and user score are power-law or Pareto distributed. It is believed that stratified sampling can reduce the risk of outlier clustering and thus will be used to assign users to one of the three test groups.

Since registered users are being tested, users will be assigned to the A, B, or C group based on the users’ ID numbers. This will allow the users to remain in the same web site variant over multiple visits versus being randomly assigned to a group upon login, which may result in a user being measured in all groups across multiple site visits. Additionally, existing users will be stratified by user score, user activity, and potentially other strata in order to prevent outlier clustering.

In order to determine if the B or C group has increased (or decreased) user activity, metrics will be measured and recorded for each group. The final list of metrics, which are used for the experiment, will depend on which metrics can be made available for us to perform our A/B/C testing and evaluate the test results. This list will be vetted through our sponsor before the A/B/C Testing will begin. The following is a preliminary list of metrics:

- Number of times a user clicked on a question in the recommender box.
- Number of times a user provided a forecast on a question reached through the recommender box.
- Number of times a user provided a forecast for a question reached external to the recommender box.
- Recommender's "ranking" of questions selected via the recommender box.
- Recommender's "ranking" of questions selected external to the recommender box.

When enough experimental data points have been captured to draw statistically significant conclusions, the data will be analyzed and the results will be summarized. The length of the test and the number of users needed will be determined after the final metrics are decided upon and after some initial measurements have been completed. Hypothesis testing will be used to determine if there is any significant difference in user activity between the three groups. The current plan is to use a Student's t hypothesis test. The hypothesis test may switch to other techniques such as rank-sum or Kolmogorov-Smirnoff tests if the distributions do not meet the parametric assumptions for a normal distribution.

In an ideal case, the A/B/C test will be run for as long as needed in order to make conclusions that are statistically significant. Multiple power analysis simulations will be run in order to determine the number of experimental data points needed in order to reach conclusions that are statistically significant. If the available test time is too short, a "low-power" study will be conducted and the results will be analyzed in terms of confidence intervals on effect size.

5.2.2 Focus Group Test

Hypothesis tests, such as the A/B or A/B/C tests mentioned in section 5.2.1, can provide quantitative answers with respect to an experiment. In our particular case, hypothesis testing can answer the following binary (i.e. True/False) questions: 1) Does the recommender box increase the number of user forecasts?, 2) Does the algorithm that creates the recommender work?, and potentially other binary questions. This hypothesis test can additionally provide a numeric confidence level associated with the answer to these questions. However, hypothesis tests cannot provide answers to qualitative or subjective questions such as: 1) Why or why not is this the case?, 2) Is the recommender box design well placed in the website and easily seen by SciCast users, 3) Does the recommender box supply interesting questions to users, etc. In order to overcome the limitations of hypothesis testing, our team designed and conducted a focus group test with the goal of answering questions which cannot be answered through hypothesis testing alone.

Originally, the focus group test was designed to be conducted after the Recommender Box Hypothesis Test had concluded. This test was designed to answer subjective questions about the recommender box such as those mentioned in the previous paragraph. However, since the recommender box would not be available on the SciCast production site in time for us to perform the hypothesis test, the focus group test was redesigned to evaluate the SciCast website as a whole and to include activities and questions associated with the recommender box.

While the recommender box was not available on the SciCast production site, it was available on the SciCast test site. The test site is used to test ideas for the SciCast website before transitioning these ideas to the production test site. In order to allow our focus group participants to be exposed to the recommender box, the focus group test was conducted on the SciCast test site instead of the SciCast production site. This would allow our team to collect qualitative test data on the recommender box.

Since the focus group test involves tests on human subjects, our team registered for and successfully completed the CITI (Collaborative Institutional Training Initiative) curriculum. Additionally, our focus group test was approved by the HSRB (Human Subjects Review Board) at GMU. The details of the HSRB approved experiment are contained in Appendix A.2. In order to comply with HSRB, the focus group test was conducted on human subjects who have consented to the conditions for the test and was designed to complete in less than an hour. The focus group participants were recruited by our project team from friends, family, co-workers, and professional connections interested in participating in the test. In order to accommodate the schedule constraints

of our volunteers, each team member conducted a separate focus group with the volunteers that they recruited.

The overall purpose of the focus group test is to use a “small group crowdsourcing” effort to discover problems with the current website and to elicit feedback which could result in an improved website. The activities and questions in the focus group test were designed to get feedback on the recommender box and to discover areas where users had difficulty in using the site, areas where a user could get unexpectedly stuck, and any sections of the site that negatively impact users.

The first activity that each focus group lead performed was to explain to the volunteers the purpose of the SciCast site and the focus group session. This was done by walking each volunteer through a Focus Group Overview Presentation which is contained in Appendix A.4. After the presentation concluded, each volunteer was asked to create a dummy account on the SciCast test site. Dummy accounts can be created on the SciCast test site by using the test email address, test@daggre.org, instead of using a personal email address. During account creation, we noted any problems that users had with creating an account.

After creating an account on the test site, users were asked to login with a dummy account and to explore the test site for 10 minutes. This would allow our project team to determine if any of the volunteers were particularly drawn to any area of the website and would allow the volunteers to get a general feel for the site and how it worked.

When the volunteer had finished exploring the website, we asked the volunteer to find a question of interest on the site and then to make a prediction. Since the recommender box was available for this activity, the user could select a question offered by the recommender box or choose a question through other means. The recommender box would not have access to the prior prediction history of the volunteer since the account being used would be newly created. Consequently, the recommender box would probably not provide useful recommendations to the volunteer. However, by asking the volunteer to “explore all options” for choosing a question, our team could get feedback from the volunteer on the value of the recommender box (i.e. usefulness, location in the website, graphic appeal, etc.) and the value of the other means used to select questions. If the volunteer had trouble selecting a question to predict, we could note that in our experimental write up. Once a question had been selected for prediction, we could observe the reactions of the volunteer when actually making the prediction. This would allow us to note any problems observed when the volunteer made a prediction.

Once the volunteer made a prediction, the volunteer was asked to fill out a questionnaire regarding their experience with SciCast (“SciCast Focus Group Questionnaire” contained in Appendix A.3). The answers contained in the questionnaire will be used as feedback to the SciCast team.

All of the activities that the volunteers participated in were timed by the focus group lead. This would allow our team to determine how long each activity took to complete and to determine if any of the activities were unexpectedly difficult for an inexperienced user. Activities that were found to take longer than desired could be reported to the SciCast team which could allow the SciCast team to improve the areas on the website which are related to these activities.

5.2.3 Splash Page A/B Hypothesis Test

As discussed in section 2, our project sponsor wanted our team to design an experiment to evaluate approaches for increasing SciCast user registration. One approach, which was implemented by the SciCast team, was an updated splash page (see Figure 2 in section 3.1.1). In order to determine if the updated splash page increased user registration, our team designed an A/B Hypothesis test. As shown in Figure 2, the proposed variation to the original splash page naturally lent itself to A/B testing due to the fact that only one change was being proposed to the page (the basis of A/B testing was covered in section 5.2.1 and will not be repeated here).

For this experiment, SciCast users will be placed in either the control group (A) or the treatment group (B). Upon reaching the SciCast website, the control, or A group, will see the original and unmodified SciCast splash page. The B group will see the updated splash page. Since users visiting the SciCast splash page could be registered or unregistered users, there is no straightforward way to collect information on these users which could be used to apply stratified sampling to the users. Therefore, in order to ensure that an appropriate number of users were assigned to (and remained in) the A and B groups, our team decided to use Google Analytics for this experiment. When an experiment is setup through Google Analytics, each user receives a unique identifier number which gets associated with cookies stored on the user’s machine. This mechanism ensures that each user remains in the same test group when accessing SciCast from the same machine. If the same user accesses SciCast from a different machine, that user may get assigned to another test group and this is acknowledged to be a limitation to this experiment.

The next step in designing this test was to determine a list of metrics which could be measured and would provide an answer to our fundamental question: Does the updated splash page actually increase the number of user registrations? One approach to answering this question is to measure the number of registered accounts before the introduction of the new splash page and the number of

registered accounts after the introduction of the new splash page. While this approach measures the metric of ultimate interest to our team, it is not capable, when used by itself, of determining what actually caused the increase (or decrease) in user registration.

One metric, which can be used to estimate a user's interest in the SciCast website is the splash page bounce rate. A bounce occurs when a user visits a page and does not interact with that page and either navigates away from the page or reaches a time threshold of inactivity. Measuring the bounce rate can help us answer our fundamental question because if a user bounces from the web site, it means that the user did not interact and register with the website. In other words, a higher bounce rate will result in a lower interaction and registration rate.

However, measuring the bounce rate does not provide us with a user registration rate. A user can visit the SciCast website, not bounce off the site, and still decide to not register with the site. This limitation can be overcome by measuring two additional metrics: 1) The number of registration attempts and 2) The number of successful registrations. Additionally, click-through rates and page visits to the SciCast question pages can provide additional insight on user interest (i.e. which areas did the user spend his/her time on the web site). When measured along with web site bounce rate, all four of these additional metrics can provide quantitative data used to answer our fundamental question.

During the initial planning stages of our hypothesis test for the recommender box, the team identified two potential sources of data for a hypothesis test. One source of data was accessible via Datamart which collects a variety of useful metrics on registered users. The other source of data was Google Analytics which is commonly used by website developers to measure and enhance the performance of their websites. Since user interaction with the splash page occurs before a user logs into their account, the data available from Datamart was not applicable for the splash page hypothesis test (however, it would have been applicable to the recommender box hypothesis test). Therefore, our team decided to leverage the capabilities provided by Google Analytics for our splash page hypothesis test. In order to capitalize on the features available to our team via Google Analytics, one of the team members completed training on Google Analytics via the Google Academy. This team member in turn educated the other team members on the capabilities of Google Analytics and on how to collect and analyze data through Google Analytics.

For this experiment, our team investigated the use of Goals and Experiments through Google Analytics. Goals are used to capture how users interact with websites. In the case of SciCast, goals were set up to capture when users clicked on the "register" button and when users successfully completed registration. Experiments in Google Analytics can be used to assist in running an A/B

hypothesis test. Through Google Analytics, our team could define which metrics to compare between the A and B splash page variants, how long to run the test, and how to split traffic between the two versions.

While a variety of useful metrics had been identified for the hypothesis test, the shortened time window for this test required our team to prioritize our experimental metrics and choose ones which could provide us with the most useful answers in the time frame we were given. Setting up a Google Analytics experiment to measure the bounce rate is straightforward and easy to implement. Measuring the number of registration attempts, the number of successful registrations, click-through rates, or pages visits requires more significant experimental setup and lead-time. Additionally, the experiment needs to run long enough to receive enough visitors in order to detect a statistically significant difference between the A and B test groups on the metrics of interest. Given the “calendar crunch” for our experiment, our team decided that measuring the splash page bounce rate for the A and B versions of the splash page should be the highest priority because it would provide us the most insightful answers.

Google Analytics provides a feature known as the Multi-Armed Bandit approach which can be used to decrease experimental time by splitting users between different website versions. The technical details for this approach can be found via Google Support⁵. Google has demonstrated that this approach will generally decrease the time needed to determine if a difference exists between two versions of a website when compared to an even, random split of visitors. The automation of the Multi-Armed Bandit approach splits traffic by determining which version of the site is performing “best” according to a single metric. This generally results in an uneven split of visitor traffic between A and B versions of a website, especially if one version is significantly outperforming the other version.

A power analysis was conducted for our experiment to determine if utilizing the Multi-Armed bandit approach was necessary. This analysis was conducted on our metric of interest, the website bounce rate, by using historical data collected via Google Analytics on the SciCast site. It was found that over a span of 98 days, the bounce rate had a mean of 5.5% and a standard deviation of 3.2%. Using this data and a desired power of 0.8 and an alpha of 5%, an online statistical calculator was used to investigate the approximate sample sizes needed to determine raw effect sizes of 0.15%, 0.25%, and 0.5% on the bounce rate. It was known that the SciCast site historically had between 500 and 700 visits per week. Based on this data, a standard hypothesis test with an even split of traffic would potentially take upwards of 4 weeks to collect enough data for the experiment, which would be too long considering the end date of the project.

Therefore, our team decided to run a single A/B hypothesis test on the website bounce rate using the Multi-Armed Bandit approach. Given more time, our team could have run multiple Multi-Armed Bandit hypothesis tests on additional metrics or a single random split hypothesis test while measuring multiple metrics simultaneously. While this approach reduced the number of metrics that could be captured and then analyzed, it increased the odds of running a statistically significant experiment on a single metric.

6. Experimental Results and Recommendations

6.1 Focus Group

Our project team conducted a focus group study on 7 volunteers. In order to work around the schedule constraints for our volunteers, each team member conducted a small focus group study at a time and location which was convenient for our volunteers. The focus group participants had a broad mix of professional backgrounds and all, except for two, had a background in science and technology. The gender mix of the focus group was balanced with approximately 57% of the volunteers being female. Only one participant was aware of the SciCast website before the study and only two participants were familiar with predictive sites such as SciCast. The focus group volunteers were chosen from a wide range of ages where the age of the volunteers ranged from 28 to 66.

The focus group study ran smoothly and none of the project team members reported difficulties associated with conducting the study. Even though the volunteer base as a whole was unfamiliar with SciCast and other predictive sites, none of them struggled with any of the activities assigned by the focus group leads. All of the participants were timed using stopwatches while performing assigned activities and were able to complete the assigned tasks within a reasonable period of time. None of the focus group volunteers struggled with finding the “sign up” button on the splash page and were thus able to easily create accounts on SciCast. All of the users easily and successfully logged into the SciCast website and were able to explore and become familiar with the website. We believe this result to be attributed to the effort made by the SciCast team to create a professional and simple to use website. At the end of the focus group study, all volunteers were asked to respond to the questionnaire in Appendix A.3. The answers to the questionnaire provided by the focus group participants are given in Appendix A.5.

We asked our focus group volunteers to identify strengths and weaknesses in the SciCast website in addition to suggested improvements. On the whole, participants were generally confused about the purpose and benefit of the SciCast website. One volunteer mentioned that he would not

have known what was expected of him unless he had seen the SciCast overview presentation first. Other volunteers suggested that the SciCast website could be improved by having a better explanation of its purpose and benefits. Once participants started looking for SciCast questions to predict, most struggled with finding questions that were relevant to them. One volunteer suggested that most of the questions are too narrow and specific and that a list of questions, which are more general in nature, should be added to the list. All participants appreciated finding and making predictions on questions which were in their respective fields of interest.

While exploring the SciCast website, we asked volunteers to pay attention to web site aesthetics and usability and to provide feedback. One volunteer complained that the site was slow while refreshing topic changes and suggested that a progress bar be displayed while the server is refreshing the page in order to ensure the user that the website is still functioning properly. Other users complained that the site contained too much text and that the chart/graph labels are too difficult to read. Volunteers appreciated being able to see:

- 1) Trending questions in each topic area
- 2) A list of possible outcomes for each question
- 3) How your answer affects your overall score
- 4) Exposure to information on new technology
- 5) The information and related links for each topic

Since most participants appreciated seeing questions that applied to them, we can conclude that a recommender box, if it provides applicable questions, would increase user participation. However, in this study, many users did not initially notice the recommender box. Some users suggested that the recommender box should be highlighted in order to grab the attention of SciCast users. For some users, the recommender box did not provide any recommended questions or any applicable questions. This result is not surprising since the recommender box was implemented on the test site instead of the production site.

Overall, the focus group study was very informative. Based on the results of this study, we would recommend that the SciCast team should modify the SciCast website to provide a better explanation of the websites benefit and value to society. We would also recommend that the SciCast team should incorporate the recommender box into the production site and investigate ways of making the recommender box more noticeable to SciCast users.

6.2 A/B Test

Using the Multi-Armed Bandit approach, the A/B hypothesis test was setup through Google Analytics to end if a winner was declared or if it were determined that running the experiment for a longer period of time would not significantly change the experimental outcome. Google Analytics suggests this experimental approach because its primary use is for business websites. Running an experiment longer than necessary could reduce profits by delaying the implementation of the most “profitable” version of the website. By default, an experiment is setup to run for a minimum of two weeks and a winner is declared when Google Analytics concludes that there is a 95% confidence level that the “winning” variant is the best choice for optimizing the experimental metric. Google Analytics can also end an experiment without declaring a winner when it determines that the experimental metric will change by less than 1% of the current value if the experiment continues to run.

In our case, the A/B hypothesis test on the SciCast splash page collected data for 15 days without declaring a winner. The experiment terminated because Google Analytics determined that running the experiment for a longer period of time would not have changed the bounce rates for the A and B versions by more than 1% of the rate (i.e. .01% increase/decrease to the bounce rate). In other words, running the experiment longer would not change the final results enough to warrant the extra time investment. The experimental results for the hypothesis test are contained in Figure 4 below.

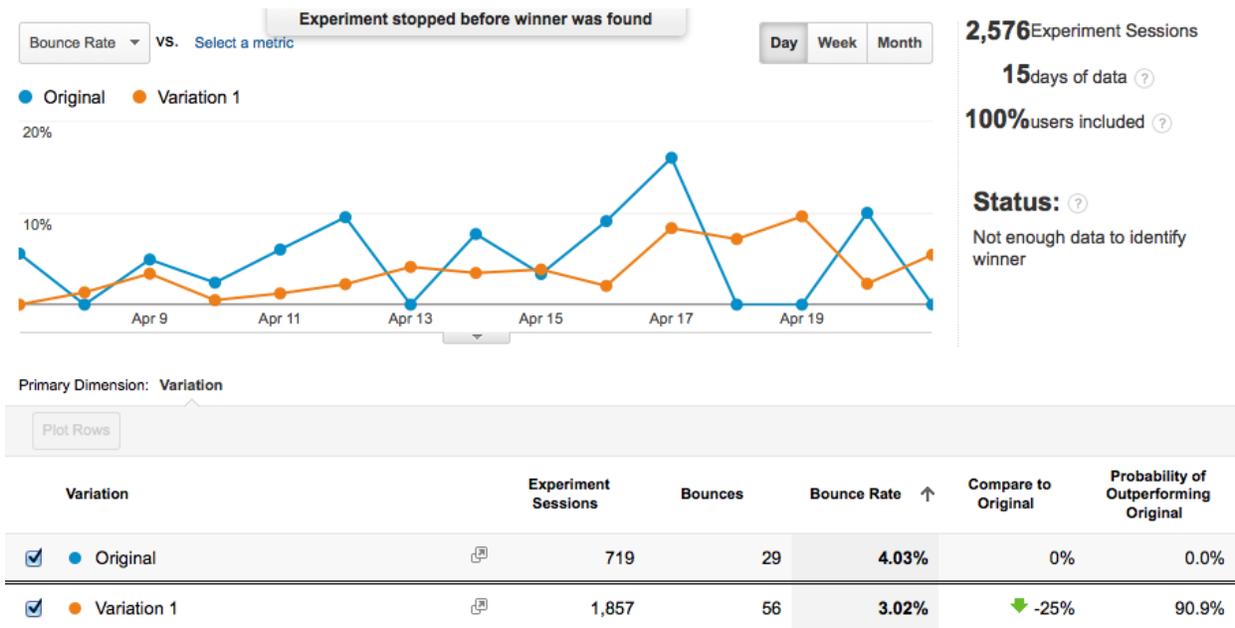


Figure 4 – Splash Page A/B Test Results

Over the 15-day period, a total number of 2,576 sessions were analyzed by Google Analytics. The Multi-Armed Bandit approach split the visitor traffic by directing 719 users to the original splash page and 1,857 users to the updated splash page (i.e. Variation 1). The final results of the experiment showed that the original splash page had a user bounce rate of 4.03% and the updated splash page had a user bounce rate of 3.02%. This represents a 25% reduction in the bounce rate which according to Google Analytics is significant at a 90.9% confidence level. In other words, we are 90.9% confident that adding questions to the splash page will reduce user bounce rates. Decreasing user bounce rate implies that users are interacting with the site more frequently which is one result that the SciCast team was hoping for.

Given the high confidence level associated with our experimental results, our team considers the A/B hypothesis test to have been conclusive and recommends our sponsor replace the original splash page with the splash page variation containing forecast questions (refer back to Figure 2). This should result in a reduced bounce rate for the SciCast site. Additionally, we recommend that our sponsor use Google Analytics for any future A/B testing for the SciCast website. Our team's experimental design can be easily leveraged to quickly implement other A/B tests for future website modifications. As demonstrated by our team, A/B testing is a useful tool for constantly improving the SciCast site and ensuring that design decisions for the site result in their intended impacts, thus reducing some of the risk associated with making changes to a website.

7. Future Work

While our team regards this project as being successful, we believe that running additional experiments and conducting additional focus groups would provide beneficial results to the SciCast team. Once the recommender box has been successfully integrated into the production SciCast site, we recommend that our sponsor run the A/B/C Hypothesis test which was designed in section 5.2.1. This test could be implemented in Google Analytics and the Multi-Armed Bandit approach could be used on a single metric of interest if the schedule is tight. The approach provided in section 5.2.1 can also be expanded to consider additional recommender box variations representing changes in recommender box location, colors schemes, font sizes, etc.

We also recommend that our sponsor conduct additional focus group studies once the recommender box is up and running on the production site. This will allow the sponsor to gain additional insight with respect to the recommender box. Finally, we suggest that the sponsor expand

the focus group study by recruiting a larger number of volunteers with diverse backgrounds and stratify test results on volunteer background and interest level.

8. References

- 1) Kahavi, Ron; Longbotham, Roger; Sommerfield, Dan; Henne, Randal M. (2009). "Controlled experiments on the web: survey and practical guide". *Data Mining and Knowledge Discovery* (Berlin: Springer) 18 (1): 140-181
- 2) Charles Twardy; Robin Hanson; Kathryn Laskey; Adam Siegel (2014). "SciCast: Collective Forecasting of Innovation". *Collective Intelligence 2014*
- 3) James Surowiecki. 2005. "The Wisdom of Crowds." Reprint. Anchor.
- 4) Wei Sun; Robin Hanson; Kathryn Laskey; Charles Twardy. (2012). "Probability and Asset Updating Using Bayesian Networks for Combinatorial Prediction Markets." *In Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence (UAI-2012)*. Catalina, CA: AUAI Press.
<http://mason.gmu.edu/~wsun/publications/uai2012.htm>.
- 5) Steven L. Scott, "Multi-armed bandit experiments",
<https://support.google.com/analytics/answer/2844870?hl=en>

A. Appendices

A.1 Project Management / EVMS

For this project, we broke out the tasks and assigned each task an expected completion date as well as an expected number of hours each task would take to complete. Our projected time to complete each axis as a function of duration of the project is shown below and is represented as our Budgeted Cost of Work Scheduled (BCWS). The actual completion rate of these tasks is the Budgeted Cost of Work Performed (BCWP) which represents how ahead or behind schedule we are. The Actual Cost of Work Performed, or ACWP, shows how many hours we actually worked and represents how over or under budget we are.

Figure 5 below shows how our ACWP, BCWS and BCWP evolved over the course of the project. Generally, tasks took more work to complete than planned meaning had allotted hours, we would have been over. The sharp spike in both our budgeted cost of work scheduled as well as budgeted cost of work complete is due to our multiple final tasks being completed and delivered.

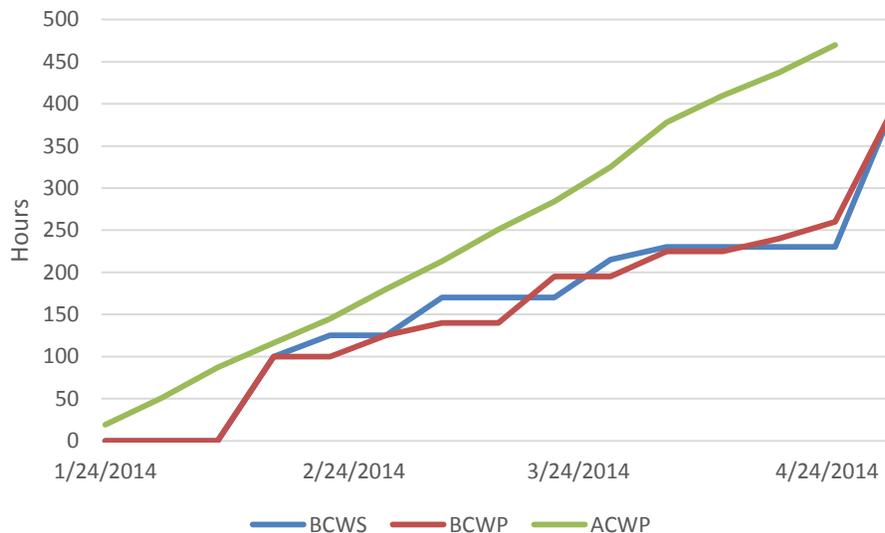


Figure 5 – EVMS Parameters Across the Project Lifetime

A.2 HSRB Approved SciCast Focus Group Experiment

Section 1: Introduction and Purpose

Crowdsourcing is based on the idea that predictions made by aggregating the opinions or guesses from a large group of people are more accurate than the predictions made by more traditional means, such as polling experts or using models. The SciCast team is hoping to use crowdsourcing as a method to gain insight into emerging technology and scientific developments in the near future. The hope is that this website will be used as a method to inform government and industry personnel who will use the information to make investments in their fields.

Kevin Connor, Andrew Kreeger, and Neil Wood will be conducting this focus group and will be referred to in this document as the experimenters. The objective of this focus group study is to receive feedback from users on current and upcoming site features. Users will be asked by the experimenters their opinion of the site and how, they the users would recommend improving it. While the users are exploring the site, the experimenters will take notes of any times it appears the user gets lost or confused navigating the site. The data being collected for this study will be explicitly used for the purpose of improving the site's usability and appeal to outside users. Ideally, changes made to the site as a result of focus group recommendations will cause more individuals to sign up and use the site more frequently in the future. The data collected will not be used to for the primary SciCast forecasting research.

Section 2: Overview and General Plan

. The focus group will be conducted with the consent of volunteers recruited by our project team who will be referred to as volunteers throughout this document. For more info on the recruitment process, see Section 5. The experimenters and volunteers will meet in the Johnson Center. Volunteers will bring laptops that they will use to access the SciCast site in order to participate in the focus group test.

The experimenters' first activity will be to explain to the volunteers the purpose of the SciCast site and the focus group session. After this, the focus group experiment will commence. The volunteers will first create an account on the live site, and then perform the following tasks:

- 1) Log in to the test site (using dummy accounts which we create before the focus group test)
- 2) Explore the test site.
- 3) Select a prediction that is of interest to the volunteer.
- 4) Make a prediction.

The experimenters will assist the volunteers (as needed) in successfully completing the aforementioned tasks. Additionally, the experimenters will use stop watches to time volunteers as they complete these tasks. The only task that will not be timed is task #3. A defined period of time will be given to each volunteer to explore the test site with the experimental purpose of collecting the volunteer's feedback. One of the purposes for timing the volunteer is to discover if any of these tasks are unduly difficult for an inexperienced user. Following these activities, the experimenters will ask the volunteers the questions listed in Section 4 about their experience with the SciCast site. The focus group session is expected to take approximately an hour to complete.

Section 3: Activities and Estimated Timing

Intro:

Explain purpose of the SciCast site. (5 minutes -- Estimated)

Explain purpose of the focus group testing. (5 minutes -- Estimated)

Activities:

Create an account on the live site (5 minutes -- Estimated)

Log in with dummy account on the test site (5 minutes -- Estimated)

Explore the test site (10 minutes – As per section 2, this is a defined time period/limit for this task)

Find a prediction that's interesting to them (10 minutes -- Estimated)

Make a prediction (10 minutes -- Estimated)

Answer questionnaire (10 minutes -- Estimated)

Section 4: Questionnaire – Contained in Appendix section A.3

Section 5: Recruitment Plans

In order to obtain volunteers while spending minimal time and resources recruiting, the experimenters are hoping to recruit friends to volunteer in the focus group. In order to ensure that the experimenters impose no undue influence on participation, the experimenters will not have an authoritative relationship with respect to the prospective volunteer (i.e. a manager to employee relationship), and the prospective volunteers will be assured that a decision to decline to volunteer will have no ill affect on the existing relationship between the experimenter and the prospective volunteer. The experimenters will strive to have a volunteer base with a diverse gender, ethnic, and age background. Additionally, the experimenters will only recruit from their respective social networks based on the willingness of prospective volunteers to participate. During recruitment, the prospective

General Questions:

1. Do you have any suggestions for improving to the site? What would you change to make it better?
2. What is your least favorite thing about the website?
3. What is your favorite thing about the website?
4. What would make you more likely to make predictions using this site?

Usability Questions:

1. Were any of the activities confusing? Did you have trouble completing them?
2. What would make the process of completing an account easier?
3. What would make xxx easier (if they struggled on XXX)?

Recommender Based Questions:

1. Would seeing questions in fields that interest make you more like to make a prediction?
2. Why did you select the prediction that you selected?
3. Did you feel that the recommended predictions applied to you?
4. How do you feel about the look and position of the recommender box? Are your eyes drawn to it?
Would a different position make you more likely to click on a prediction in the box?

Closing Question:

After you leave here today, will you use the account you made? Why or why not?

A.4 SciCast Focus Group Overview Presentation



Testing to Improve User Response of Crowdsourced S&T Forecasting System

Introduction to SciCast/ Focus Group Test

SYST 699 / Spring 2014

Sponsors:
Charles Twardy, GMU C4I Center
Adam Siegel, Inking Markets

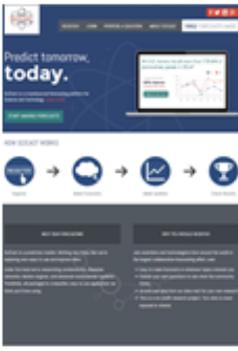
Team Members:
Kevin Connor
Andrew Kreeger
Neil Wood



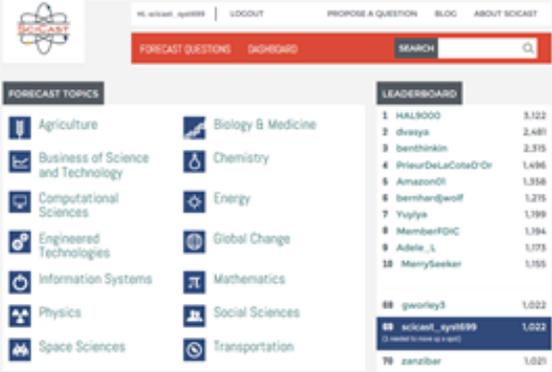
Title Slide

Introduction to SciCast

SciCast central premise: "the collective wisdom of an informed and diverse group is often more accurate at forecasting the outcome of events than that of one individual expert."



SciCast Introduction Screen



SciCast Initial Screen

SciCast is a research project that forecasts the outcomes of key issues in science and technology*.

**SciCast is run by George Mason University and sponsored by the U.S. Government.*

Slide 1

SciCast Forecast Screen

What will be the total value of U.S. cherry production for the 2013 marketing season, in thousands of USD?

Make a Forecast **Discussion (0)** Background Trends & History

What is the chance any of these possible answers will happen?

Possible Answer

POSSIBLE ANSWERS AND CURRENT CHANCE

- Less than 694245 **23% chance**
- Between 694245 and 772774 **15% chance**
- Between 772775 and 851304 **15% chance**
- Between 851305 and 929834 **15% chance**
- Between 929835 and 1008364 **15% chance**
- More than 1008364 **17% chance**

What is the chance any of these possible answers will happen?

Less than 694245

YOUR FORECAST	CHANCES OF HAPPENING
<input type="radio"/> Almost surely	90% - 100%
<input type="radio"/> Very Likely	80% - 90%
<input type="radio"/> Likely	60% - 80%
<input type="radio"/> As likely as not	40% - 60%
<input checked="" type="radio"/> Unlikely	20% - 40% Current chance
<input type="radio"/> Very unlikely	10% - 20%
<input type="radio"/> Almost surely not	0% - 10%

What is the chance any of these possible answers will happen?

POSSIBLE ANSWERS AND CURRENT CHANCE

Less than 694245 **23% chance**

CURRENT CHANCE **23%** YOUR FORECAST **23%** CHANCE CHANGED TO **23%**

DEBIT 0 POINTS

Safe Mode

Power Mode

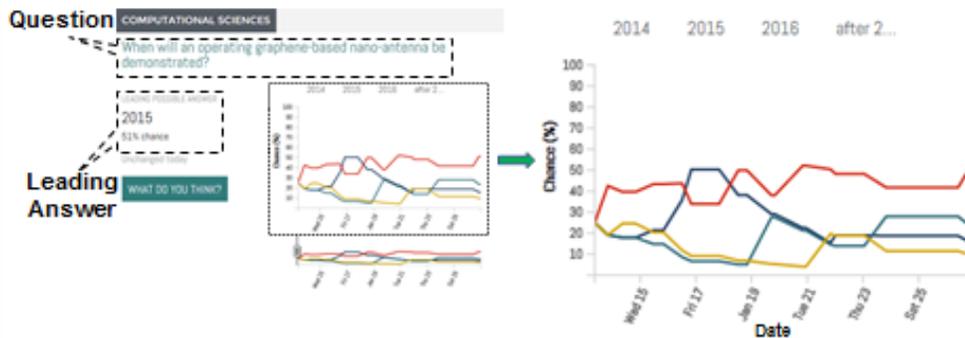
SciCast users can currently provide a probability associated with each forecast by using the Safe Mode or the Power Mode.

Via the SciCast Forecast Screen, participants can make and change their forecasts at any time on a published question.

Slide 2

Real-Time Predictions

SciCast functions like a real-time indicator of what our participants think is going to happen.



Forecasts made by SciCast Users are aggregated to provide predictions on given questions.

Slide 3

Focus Group Test

- Purpose of Focus Group test is to elicit feedback from volunteers with respect to the SciCast web site.
- Volunteers will perform a few activities on the SciCast website under the supervision of our project team.
- After the activities have been completed, the volunteers will answer questions with respect to their experience with SciCast.

Slide 4

A.5 Focus Group Answers

Questions	User 1	User 2	User 3	User 4	User 5	User 6	User 7
Age	28	66	28	28	56	58	39
Gender	F	M	M	F	M	F	F
Occupation	Policy Analyst	Veterinarian	Data Analyst	Banker	Software Engineer	Registrar	Computer Scientist
Heard of Website?	No	No	No	No	No	No	Yes
Experience w/ Prediction sites?	No	No	Yes, a little bit.	No	No	No	Yes, a little bit.
Interested in Site?	No	No, not interested in predictions that aren't based on objective data. Questions were highly specialized and would require a lot of background in order to answer objectively.	A little.	No	Yes	Not really	Somewhat interested, because of the pool of information and the discussions available in the site.
Background in S&T?	No	Yes	Yes	No	Yes	Yes	Yes

<p>Suggestions for improving site?</p>	<p>Remove the highlights from current forecasts. Have topic selections made during registration automatically connect to individual user dashboard.</p>	<p>A better explanation of the site, what it is for and why use it? Also on the first page it says “data <u>is</u> never shared”, it should say “data <u>are</u>”.</p>	<p>It would help to make finding question I’m interested in easier to find. Also, questions that will have answers more near-term would help.</p>	<p>Upfront description on the goal/mission statement of the website</p>	<p>During sign up, a select all for the check boxes. More information on the purpose of the site, wants to know how predictions are validated</p>	<p>More background information on the question, explaining some of the issues involved</p>	<p>Improve on the navigation. For example, provide a “back” function to return to the list of questions in the same category. Alternatively, provide a mini column to show other relevant questions at the side while I am diving into one specific question of that category.</p>
<p>Least favorite thing?</p>	<p>Chart/Graph label difficult to read. I would prefer seeing the full description vs. having to roll-over w/ cursor to see chart labels.</p>	<p>It was not apparent what was being asked of the user. Without the pre-explanation from you, hard to discern the purpose.</p>	<p>Not really sure what the value is if I make a prediction.</p>	<p>Too much text</p>	<p>Not knowing benefit of using the site</p>	<p>Questions were obtuse, difficult to find question she felt she could understand</p>	<p>It takes too long (about 1-3 seconds) to refresh when I change a topic setting on the page. While the server is refreshing the date, something like a progressive bar may be shown, so that the user may wait. I almost walked away thinking there is no question on the topic.</p>

Favorite thing?	Seeing trending questions in each issue/topic area.	Liked listing the possible outcomes, helped think about what makes sense but it might bias the answers.		How you are able to see how answer affects overall score	Exposure to information on new technology	Graphics on the site, leaderboard	The information and related links for each topic.
Make you more likely to make predictions?	Mix of scientific and pop-culture questions.	Incentives	Not sure.	If the website demonstrated correlation between prediction and past results, also, if the site was more related to her occupation/interests	Getting feedback about predictions, knowing when you would get feedback, or what the predictions are used for	Having a better understanding of the questions	Provide more questions on general subject matters, on which expert knowledge may cast insight. As it is, the current topics are so specific that an expert may only be knowledgeable enough to predict one or two questions.
Any activities confusing?	Answering questions, explanation on web site was a little unclear and follow-up question wording a little confusing.	No	None	No	No	No	No
Make creating account easier?	Nothing. Very simple.	Nothing	None.	If your username was your email account	Nothing	Nothing	Nothing

Seeing questions in fields of interest help?	Yes	Absolutely	Yes	Yes	Yes	Yes	Yes
Why did you pick the question you did?	Best guess.	It was in a field I was familiar with and was a subject I had general knowledge of the current events.	Seemed interesting, and didn't want to keep looking.	It was a topic she had background knowledge in	Because it was a topic that interested him	Thought question was intellectually interesting	I am informed in the field related to that question.
Did you feel the recommend predictions applied?	Did not see the recommended question, maybe highlight w/ different color.	Only saw the one question, did not know I could interact with it.	No	Did not see recommender box	Yes	Did not see the recommender box	No. Did not see the recommended questions.
Look and position of recommender?	See above. Did not see – good location, needs to be highlighted.	Not sure.	It needs more interesting questions. Something that would draw my attention.	Did not see recommender box	Eyes were drawn to it, liked the multiple ways to navigate through questions (arrows and boxes)	Did not see recommender box	Currently quite good.
Will you create your own account?	No, no topics or questions that drew me in.	No, not of interest to me.	Maybe, but probably not.	No	Possibly	No	Yes, sometimes. The site is very informative. But I won't do it too often, because the topics are too specific and can be boring.

