User Testing of Crowd Sourced S&T Forecasting System

SYST-699 Capstone Project Proposal

February 11, 2014

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

Team Members:
Kevin Connor
Andrew Kreeger
Neil Wood

Contents

1. 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 crowd sourcing.

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.

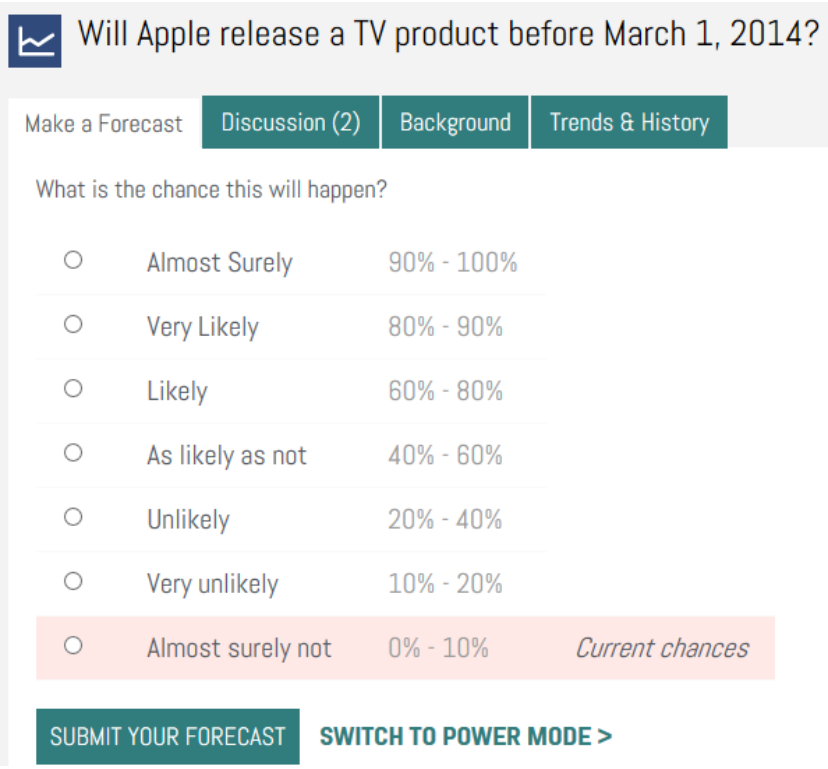

Figure 1 – SciCast Forecasting Example

When utilizing the power of crowdsourcing, a key variable to developing accurate forecasts is the number of forecasts made by the "crowd". In general, crowsourced 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.

It has been determined that two metrics have a direct influence on the number of user forecasts that are made in SciCast. The first metric is the number of registered users and the second metric is 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 proposal will outline the scope, preliminary requirements, technical approach, and expected results for addressing SciCast's need to increase the number of user forecasts that are made in SciCast.

2. Project Scope

2.1 Overview and Rationale

Throughout the process of developing this proposal the group discussed several different methods that could be used to increase user forecasts made in SciCast. For instance, the SciCast user base could be improved 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 on professional social media sites.

The SciCast user base could also be improved by reducing the website bounce rate. Potential users of SciCast may be directed to the SciCast website and decide to leave the page without creating an account on SciCast. One approach for solving this problem is to change the web design aesthetics for the SciCast web site by making appropriate modifications to the SciCast splash page, registration screen, or login screen. Figure 2 illustrates the current version of the SciCast splash page and SciCast registration screen.

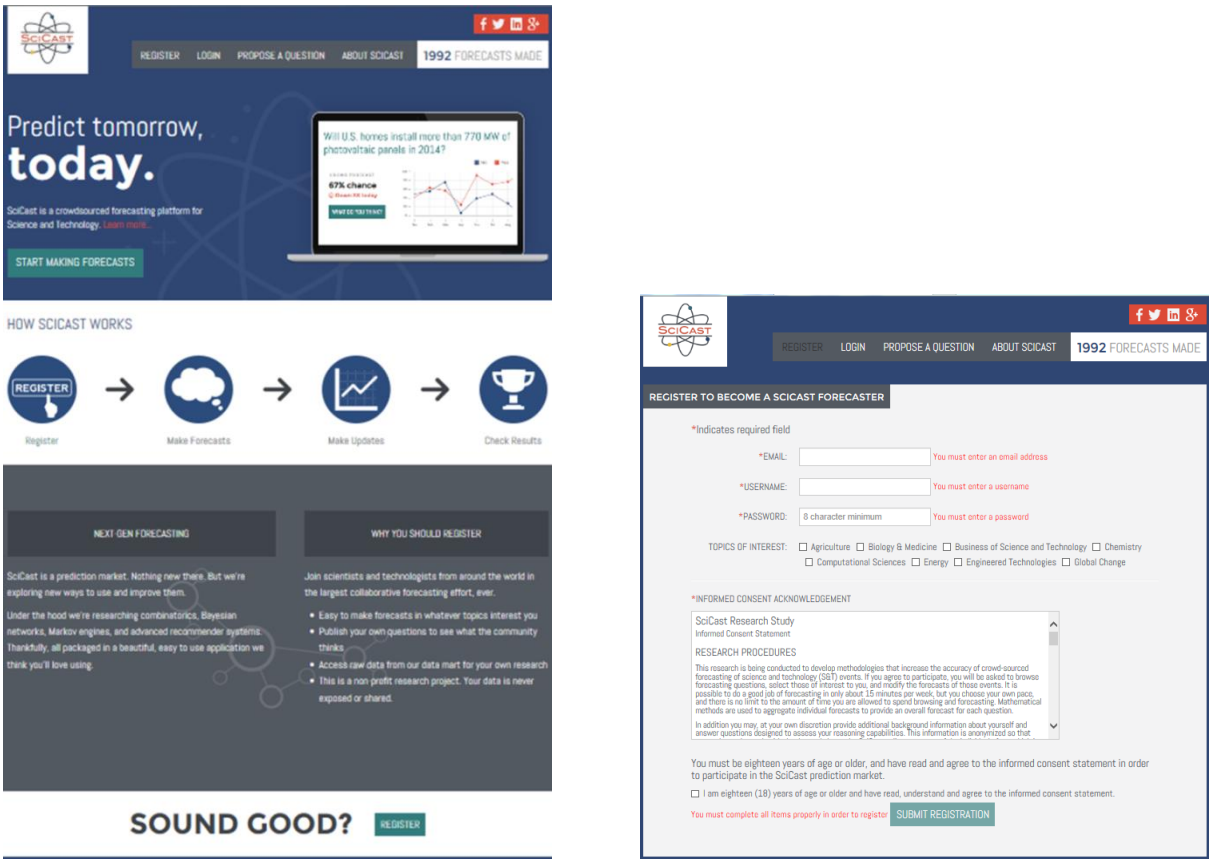

Figure 2 – SciCast Splash Page (Left) and Registration Screen (Right)

Increasing the number of new registrants and the number of forecasts made by each registered user are both worthwhile goals. However, our team has determined that increasing the number of forecasts made by currently registered users should be the priority for this project. 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 that do not make forecasts will not improve the number of forecasts made.

In order to improve the average number of forecasts made by each registered user, various options for modifying the user interface (UI) design of the SciCast web site will be proposed, evaluated, and finally recommended for implementation. These options will be prioritized based on perceived effectiveness and the ability to properly evaluate them within the schedule constraints of our class project.

One UI design modification that will be proposed and evaluated is the addition of a recommender box to the SciCast initial screen. The recommender box would contain a sorted list of forecast questions considered relevant to the SciCast User. This list would be determined based on an algorithm being developed by the SciCast team. The recommender box could be placed in the middle of the SciCast initial screen as illustrated by Figures 3.

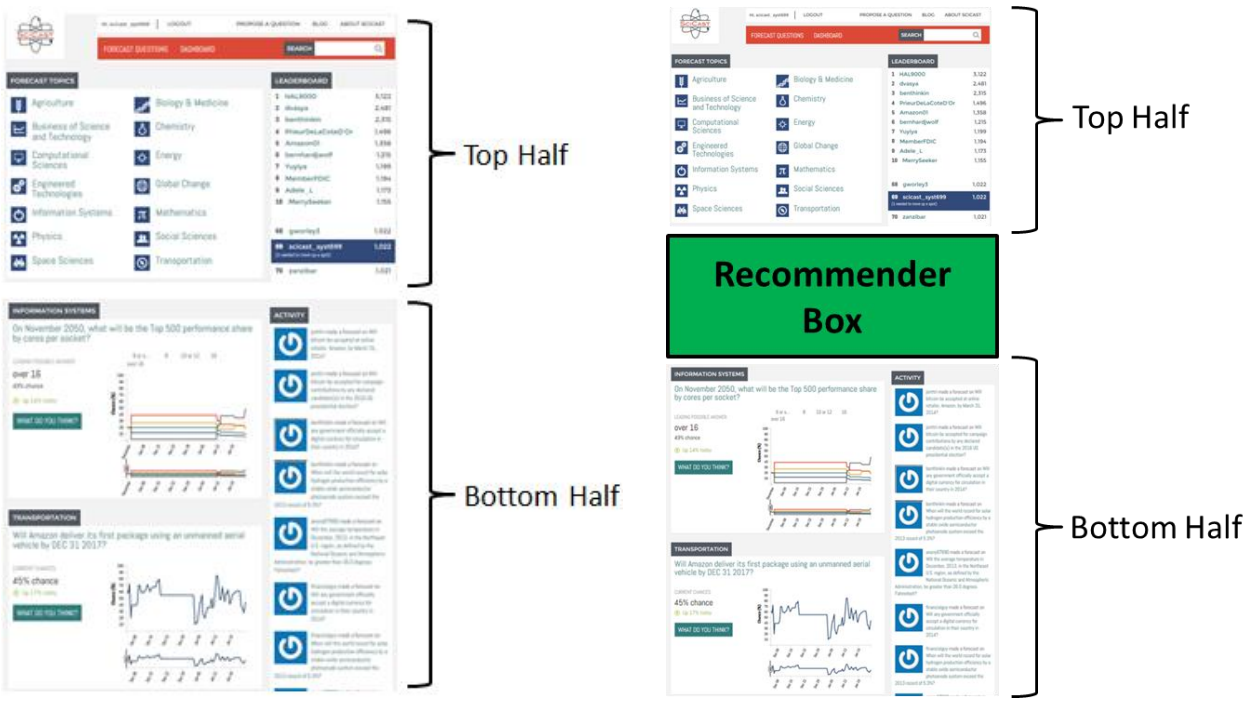

Figure 3 – SciCast Initial Screen (Left) and Proposed Insertion of Recommender Box (Right)

2.2 Summary

In the previous section, a variety of ideas have been suggested for increasing the number of registered users and the average number of forecasts made by each registered user. Our team has determined that increasing the number of forecasts made by currently registered users should be our priority for this project. To this end, our team will evaluate the recommender box and determine if adding the recommender box results in increased user participation. Our team's effort will be focused on testing the recommender box to determine if it causes significant changes in user participation rate. All website design and programming will be performed by the sponsor.

3. Preliminary Requirements

Below are the preliminary requirements for this project. These requirements are subject to change based on feedback from our sponsor and class professor in addition to lessons learned from our experiments.

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.

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?

4. Technical Approach

As discussed in section 2.1, this project will determine if adding the recommender box will increase user participation. As illustrated by Figure 3, the recommender box would be inserted into the site and viewed by all registered users once they login to the SciCast site. 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 our 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 their generalized equivalents. A/B Testing is a controlled experiment for establishing a causal relationship between changes and their influence on user behavior[1]. We believe that a well-conducted A/B or A/B/C test performed on the users of SciCast will provide us with the data necessary to conclude whether or not the implementation of the recommender box increases user forecasts. Once the A/B or A/B/C test has been concluded and the results are analyzed, we will follow-up with a small focus group study if the analysis suggests that it would be beneficial and if our project schedule permits.

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, we propose that all test groups be visually laid out the same, i.e. the same objects in the same position for all groups.

Per sponsor request, we intend on designing and conducting an A/B/C test on the users of SciCast. 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. If the B or C groups show increased user activity, then we can conclude that having a recommender results in increased user

activity. If the B version outperforms the C version, then we will be able to conclude that the recommender outperforms a random question generator.

In performing an A/B/C test, users can be randomly assigned or assigned by strata to anyone of the three test groups. Ideally, our team would like to prevent experimental outliers from accidently clustering in a single test group (ie, 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 per sponsor request, we intend on using that approach.

Since we are testing registered users, we plan on using the users' ID number to assign them to the A, B, or C group. 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, we need to measure and record metrics for each group. The metrics that we will use will depend on which metrics can be made available in time for us to perform our A/B/C testing and evaluate the test results. The following is a brief, and initial, list of identified metrics the group would like to measure:

- 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.

The above list of metrics is only a subset needed to accurately draw conclusions from the experiment. We do not intend to draw conclusions from only one or two metrics, but it is acknowledged that calendar time is a significant constraint in this project and that not all of the desired metrics may be made available. The final list of metrics to capture will be vetted through our sponsor before the A/B/C Testing would begin.

Once we have captured enough experimental data points to draw statistically significant conclusions, our team will analyze the data and summarize the results. 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. We are currently considering a Student's t hypothesis test, but may switch to other hypothesis test techniques such as rank-sum or Kolmogorov-Smirnoff tests if the distributions do not meet the parametric assumptions for a normal distribution.

When the A/B/C tests have been completed, our team will analyze the test results and recommend a small focus group test if the analysis suggests that it will be beneficial and if calendar time will permit. The results of the A/B/C testing can potentially provide us the 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".

One challenge in performing a focus group study is that we need the results of the A/B/C testing prior to completing the design for the focus group study. These results, if conclusive, will allow us to leverage the knowledge gained from the A/B/C tests in order to focus the time spent in a focus group study to answering a limited set of questions. This is needed since a single focus group study can only last for about 60 minutes.

Since the production site will not be available for A/B/C testing until early April, it is possible that our project schedule will not allow us to hold a meaning post A/B/C focus group test. In order to mitigate this risk, a focus group study will be performed prior to (or in parallel with) the A/B/C testing. To this end, our team will prepare the questions and processes that we will use for the focus group testing and will vet these through our sponsor and the human testing review process. Currently, we are planning to conduct a pre-A/B/C focus group test on a test version of the SciCast website which will enable us to complete this testing using the recommender box before the recommender launches on the production site. When the pre-A/B/C focus group test has been completed, the results will be analyzed and the conclusions will be folded into our Final Report/Final Presentation. Additionally, these results will be folded into the design of any post A/B/C focus group test that we conduct.

Ideally, we would like to run the A/B/C testing for as long as needed in order to make conclusions that are statistically significant. We plan on running multiple power analysis simulations in order to determine the number of experimental data points needed in order to reach conclusions that are statistically significant. Given the late start date of the A/B/C tests, it is possible that we will need to run a "low-power" study and then analyze the results in terms of confidence intervals on effect size.

5. Expected Results

The purpose of the A/B/C testing is to determine if changing the website causes a change in the behavior of the people accessing the website. The goal of the experiment is to determine the answer to the three questions shown in the preliminary requirements section. The first two questions will be answered during A/B/C testing (requirement 2.4.3.1 and 2.4.3.2). The second question will be answered in a focus group by observing and discussing the user experience on the SciCast website (requirement 2.4.3.3).

The first question being examined is whether or not the recommender box causes a user to make more forecasts. Currently, many users of the site create an account and never make any predictions. It is hoped that the recommender box will cause these users to make at least one and hopefully more forecasts. It is expected that this experiment will demonstrate that the recommender box will increase user participation by a statistically significant amount. The A/B/C testing will also try to answer the question of whether or not the algorithm used by the recommender is effective. Again, this answer will be determined by comparing the recommender to a box that displays random questions. We expect that the recommender will outperform the random display of questions.

The final question will be answered through the use of a focus group. By observing users as they use the SciCast site, and asking them why or why not they used the recommender box, it will be possible to determine some of the reasons that motivate users to participate and to possibly find new ways to increase user participation. A/B/C testing makes it possible to observe how users act, and how changes to a website change these actions. However, it is impossible to figure out why they acted that way, and why changes made to the website in the A/B testing succeeded or failed to make a change. Observing and speaking to users makes it possible to understand why the changes affected their behavior. It is expected that the focus group will demonstrate that users used the recommender box because they instantly see questions that appeal to them and then work to answer those questions.

6. Project Plan

The tasks and project milestones for this project are shown in Figure 4 below. This figure additionally contains a high-level Work Breakdown Structure (WBS) beside each task name. For the sake of work load planning, it is assumed that each team member contributes 10-15 hours/week (which includes any time spent in the classroom). Minor schedule/WBS changes may occur during the course of this project as long as the due dates for major project deliverables are not adversely affected. An Earned Value Methodology System (EVMS) will be used for tracking project progress.

WBS	Task Name	Duration	Start	Finish
1	**Capstone Course Project**	**78 days**	**Tue 1/21/14**	**Fri 5/9/14**
1.1	**Project Kickoff**	0 days	Tue 1/21/14	Tue 1/21/14
1.2	**Perform Initial Research and Develop Project Proposal**	**14 days**	**Wed 1/22/14**	**Mon 2/10/14**
1.3	**Project Management and Control**	**30 days**	**Tue 2/11/14**	**Tue 3/25/14**
1.3.1	**Deliver Project Proposal**	0 days	Tue 2/11/14	Tue 2/11/14
1.3.2	**Discuss Project Proposal with Dr. Laskey**	0 days	Tue 2/18/14	Tue 2/18/14
1.3.3	**1st Progress Report**	0 days	Tue 3/4/14	Tue 3/4/14
1.3.4	**2nd Progress Report**	0 days	Tue 3/25/14	Tue 3/25/14
1.4	**Finalize Experiments and Testing Techniques**	15 days	Tue 2/11/14	Mon 3/3/14
1.5	**Conduct Testing**	20 days	Tue 3/4/14	Mon 3/31/14
1.6	**Analyze Results and make Recommendations**	20 days	Tue 4/1/14	Mon 4/28/14
1.7	**Develop Final Report and Final Presentation**	**33 days**	**Tue 3/25/14**	**Fri 5/9/14**
1.7.1	Generate Rough Draft of Final Report and Final Presentation	25 days	Tue 3/25/14	Mon 4/28/14
1.7.2	**Dry Run of Final Presentation**	0 days	Tue 4/29/14	Tue 4/29/14
1.7.3	Finalize Final Report	4 days	Tue 4/29/14	Fri 5/2/14
1.7.4	**Deliver Final Report**	0 days	Mon 5/5/14	Mon 5/5/14
1.7.5	Finalize Presentation	8 days	Tue 4/29/14	Thu 5/8/14
1.7.6	**Give Final Presentation**	0 days	Fri 5/9/14	Fri 5/9/14
1.8	**Develop Website**	**44 days**	**Tue 3/4/14**	**Mon 5/5/14**
1.8.1	Generate Website	44 days	Tue 3/4/14	Fri 5/2/14
1.8.2	**Deliver Final Website**	0 days	Mon 5/5/14	Mon 5/5/14

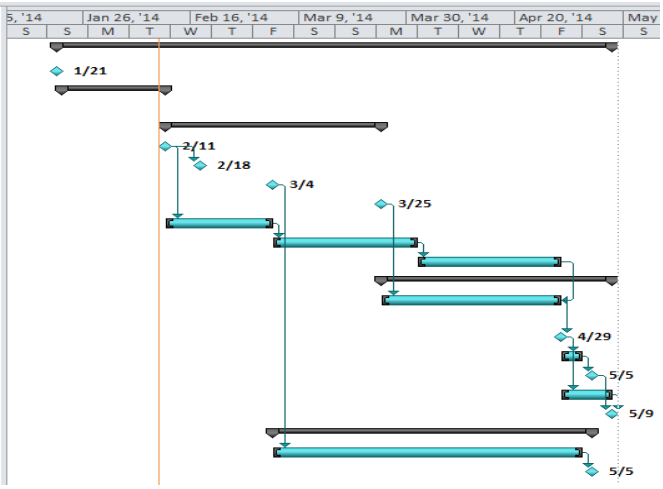

Figure 4 – Project Schedule

7. Risk Mitigation

There are two primary risks that could interfere with the completion of this project. The first risk is that the recommender box may not be implemented early enough for us to start A/B/C testing in time to gather a statistically significant data set. The second risk is that there may be insufficient participation in the experiment which would also prevent us from gathering a statistically significant data set.

To mitigate both of these risks, our group will work to ensure that there is as much time as possible for the SciCast team to implement the A/B/C test and allow the experiment to run so that we are able to collect data. The earlier we complete our critical path tasks and hand off work to the SciCast team, the more likely the SciCast team will be able to complete website implementation on a favorable time scale. In addition to accomplishing critical tasks early, our group will also work to complete additional tasks in parallel while waiting for testing to commence. These tasks include preparing for a focus group experiment as well as creating the group website and working on the final report and documentation.

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.