

Spring 2009, Human-Computer Interaction Institute, Carnegie Mellon University
Joanna Bresee, Hajin Choi, Daniel Lee, Ellen Wu

Adaptive Personalized Information Management for Biologists

Final Report

1. Introduction	
1.1 Background	3
1.2 Process Overview	3
2. User Research	
2.1 User Interviews	4
2.2 Findings: Information Flow	6
2.2.1 Documenting Information	
2.2.2 Finding Information	
2.2.3 Conclusions	
2.3 Findings: Research Process	10
2.3.1 Research Process and Needs	
2.3.2 Conclusions	
3. Concept Creation	
3.1 Scenarios	12
3.2 Concept Validation: Findings and Implications	15
4. Technical Review	
4.1 Theme: Gaining a broad understanding of a research area	16
4.1.1 Scenario: Paper Tree	
4.1.2 Discussion	
4.1.3 Scenario: Instant Review Paper	
4.1.4 Discussion	
4.2 Theme: Finding new papers in your research area	17
4.2.1 Scenario: Weekly search	
4.2.2 Discussion	
4.3 Theme: Detecting new research trends and determining research areas to pursue	18
4.3.1 Scenario: Finding new areas from paper discussion sections	
4.3.2 Discussion	
4.4 Theme: Obtaining knowledge on experiment techniques	19
4.4.1 Scenario: The Recipe Book	
4.4.2 Discussion	
5. Conclusion	20

1. Introduction

1.1 Background

This report summarizes an investigation into how biologists search for related work information, and the development of design ideas that leverage machine learning. Biologists spend many years studying a single research topic, and must keep abreast of new research that advances work on the specific genes and proteins they study. Traditionally, biologists studied a single protein or process, but they have increasingly become interested in understanding how proteins fit into a larger community. As a result, biologist must be aware of a growing number of papers in their field, and they spend much of their time trying to discern relevant papers. Our three-month survey revealed opportunities for machine learning technology to help biologists easily discover work in related areas that may impact their research.

1.2 Process Overview

We conducted a survey of biologists searching behavior in three phases. First, we conducted a series of contextual interviews with researchers to understand when they need to find information in their research, and how they search for information. Using our findings from user research, we developed fifteen design ideas that leveraged machine learning to help biologists search for information. We reviewed our design ideas with biologists to determine the most successful concepts. Finally, we discussed the designs with machine-learning experts to determine which concepts were the most feasible applications for machine learning technology. The process, findings and implications of each phase are described in detail in the following sections.

2. User Research

To gain a broad range of perspectives, we conducted nine interviews with six Carnegie Mellon biology researchers, including professors, graduate students and postdoctoral researchers. We conducted interviews in the form of retrospective interviews and contextual inquiries. In the retrospective interviews, we sought to understand the different phases of a research project, and the information needs that biologists have at each phase. Through contextual inquiries, we sought to understand the methods biologists employ to find information by observing the biologists as they worked on research.

From the analysis of our interviews, we developed models to show the flow of information to and from different researchers and information resources in the lab. We also created sequence models to identify the informational needs at different phases during research.

2.1 User Interviews

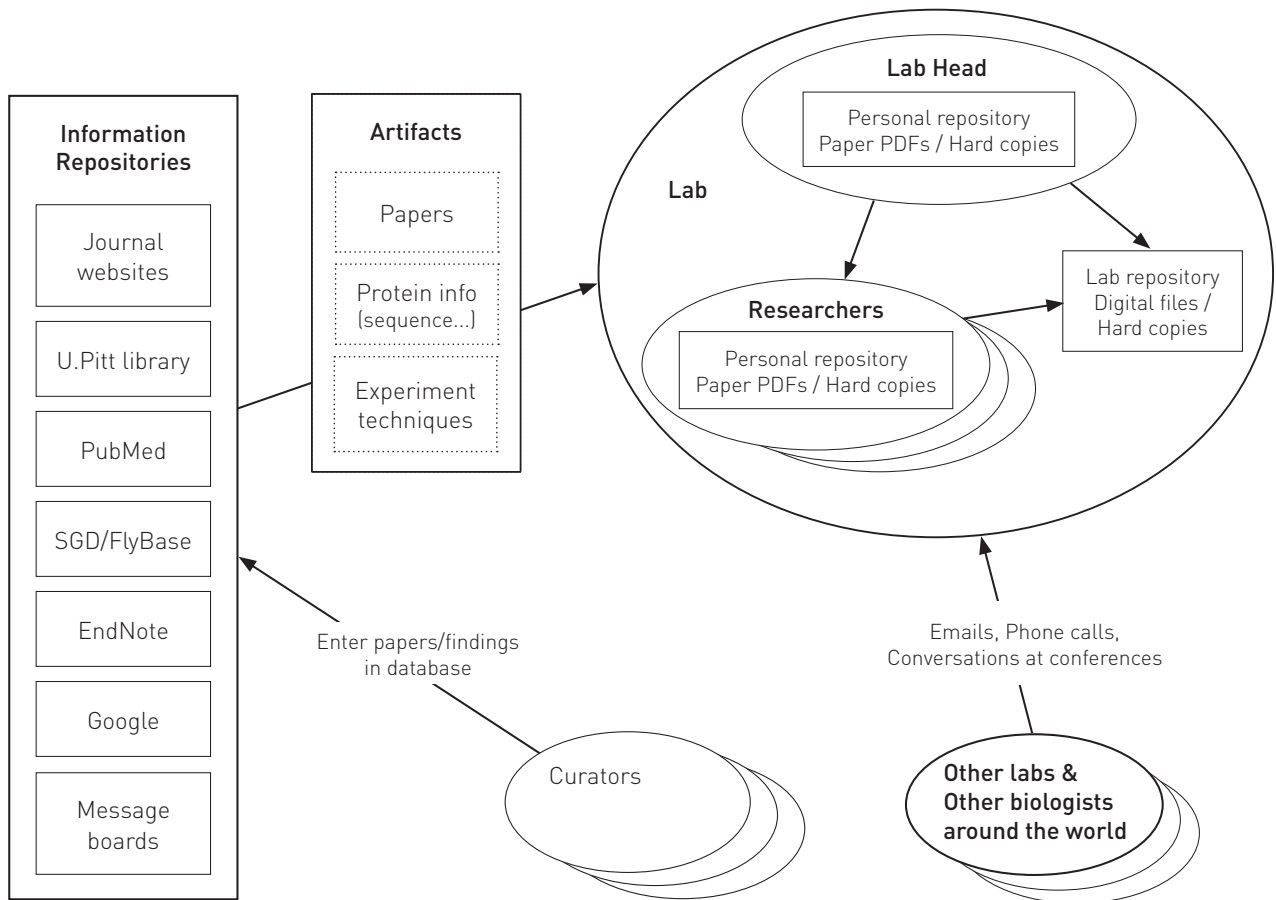
To gain a broad understanding of a lab's structure and practices, we interviewed the Principal Investigator (PI) of a biology research lab that studies yeast, and two graduate students in a lab that studies flies. We learned that labs are composed of graduate students, undergraduate students, post-doctoral researchers and technical or senior research scientists. From the PI, we learned that biologists in his lab primarily work individually, but when there is a larger project, the lab will form temporary teams to collaborate. Both labs have regular meetings to report research progress and discuss related papers published from other labs. On average, PhD students in both labs graduate in 5.5 years by following a specific research process.

The yeast lab members are interested in temporal and spatial aspects of the existence of certain proteins, and their interaction with other proteins. The PI pointed out that the research trend is moving from studying a single protein to understanding the whole picture. For example, one of the post-docs studying ribosome assembly focuses on the order of the steps, and what happens in specific steps. The post-doc informed us that they only studied one to three proteins as a graduate student, but she now studies five to fifteen proteins at one time. The fly lab members are interested in how cells change shape simultaneously to form a tube, and identifying and visualizing cell death signal in Golgi apparatus.

In order to keep up to date of relevant research, both labs search research journal websites, used the PubMed website for finding new articles, the ISI website for finding mostly cited papers, and online gene databases called SGD (*Saccharomyces Genome Database*) and FlyBase. They also use Google to search for general information and discussion boards.

Both labs are composed of approximately 10 people. Both the PI, the post-doctoral researcher in the yeast lab, and the graduate student in the fly lab emphasized that they seek out information and supplies from other labs saying, "Getting things from other labs is very, very, very common." There are three or four labs that conduct similar research to them, and they contact one another to ask about experiment details or exchange samples. Despite their close contact with other labs, we learned from the post-doctoral researcher that it is possible to do repeated work that another lab has done. For instance, she once discovered a new protein at around the same time as another lab and they both named it differently.

We modeled the structure of the labs, and confirmed our understanding of the structure with graduate students in both labs.



Information Flow Model

2.2 Findings: Information Flow

2.2.1 Documenting Information

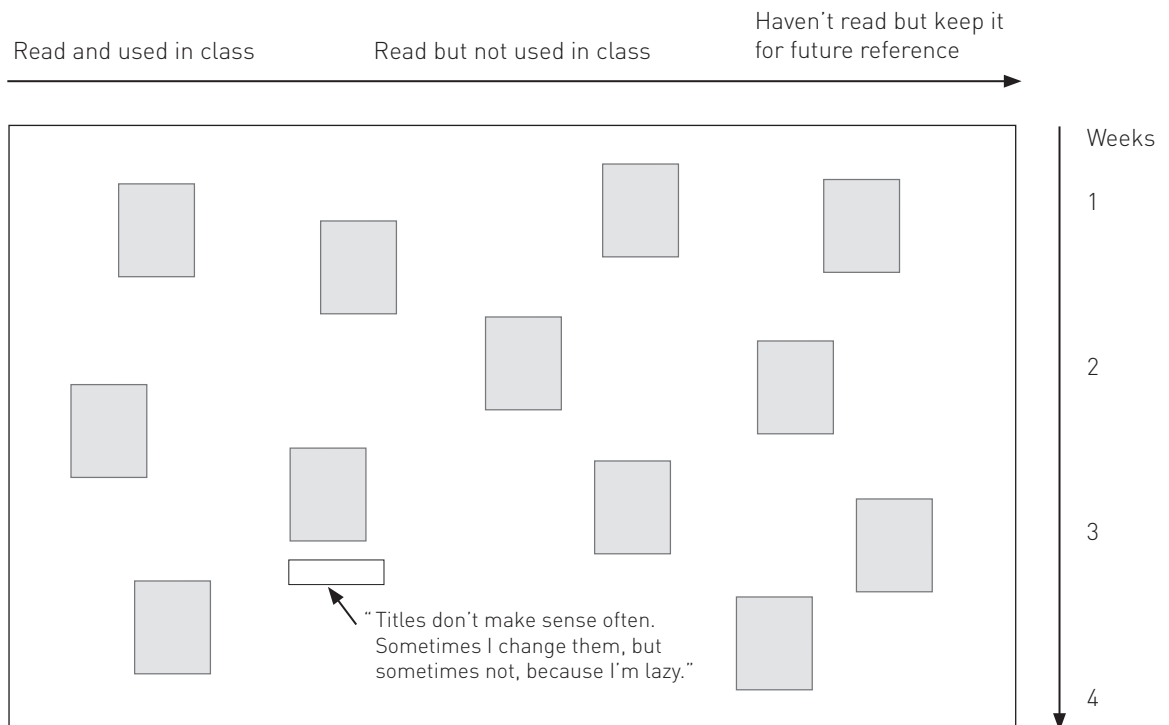
Each lab has its own way of keeping literature. The PI of the yeast lab organizes his files and folders on his computer in a way that is very similar to a physical desktop. In one of his folders for storing papers for teaching purposes, the research paper files are approximately aligned to invisible grids where the vertical axis represents weeks, and the horizontal axis categorizes the files into: papers read and used in class, papers read but not used yet, papers that haven't been read, but kept for future reference. On his computer's desktop, folders are on the right, and files are on the left. He tries to organize the files to the folders they belong to weekly.

There are two shared computers in the yeast lab, one PC and one Mac, each for running certain data analysis programs. The lab stored a list of related papers on one of the shared computers using a 50-page Word document, sorted by protein name alphabetically and review papers at the end. There was no link to the actual research paper files. No one is responsible for maintaining the file and it's rarely updated and used. There is a physical documentation of the lab publication, which is sorted by year and author.

The fly lab used EndNote, a reference management software package, to document related papers digitally. One can search PubMed through EndNote, download articles and create links to the PDF files. The PI or senior researchers often share their personal EndNote libraries with new graduate students.

2.2.2 Finding Information

In order to learn how researchers find information, we asked the Principal Investigator and three graduate students to walk us through how they search for a paper of protein information. The PI browses the University of Pittsburgh website every Monday and searches through the table of contents of twenty related journals. He used the University of Pittsburgh website as it had a larger collection of journals. If he finds a title that appears to be relevant he opens and quickly scans the paper's abstract. If the paper still appears to be relevant he looks at the introduction, results and discussion sections. Papers that are relevant to him are written by an author that he is familiar with, address a protein or method that he's interested in, or are related to subject matter in a course that he teaches. If he finds a relevant paper he saves the PDF file to his desktop and then sorts the files on his desktop into folders. He organized his folder according to the time in the semester, and the extent to which he has read the paper. The following illustration represents a typical folder of his desktop:



File Arrangement in a Folder on PI's Desktop

If the paper is related to one of his student's research, he will often print out the paper and put it on the student's desk. The PI expressed that he was concerned that we "might miss some important paper in an obscure journal."

A graduate student in the yeast lab walked us through what he does when the PI gives him a paper. First, he looks up the protein in the paper on the SGD database. Second, he looked for a review paper that relates to that gene. However, there is no indication which paper is a review paper. The student told us that based on the title "this looks like a review paper", but when he opened the he saw that it was not. The graduate student did not read the paper that the PI handed him, and told us that he doesn't "have enough time" to read all of the papers the PI thinks are relevant to his research.

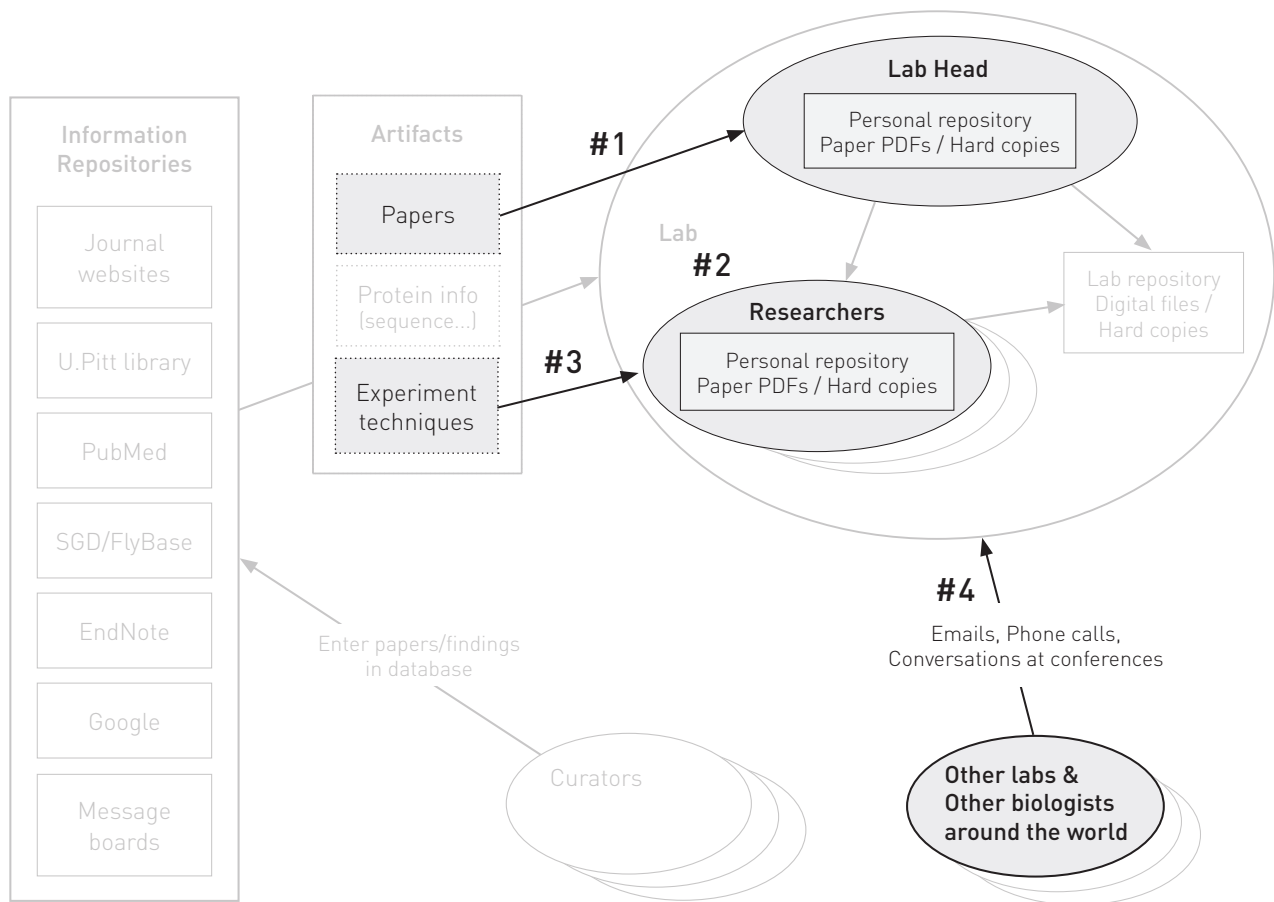
Graduate students in the fly lab told us they their PI did not print relevant papers for them. Instead they search for the papers themselves. Several students in the fly lab subscribed to email alerts from journals. The journal would send them any articles that referenced authors they were interested in.

Besides literature, online databases are another main source of information, which provide facts about different proteins. The SGD and FlyBase are the two databases frequently used by the labs we interviewed.

A fifth year PhD student in the yeast lab walked us through how she searches in the SGD database. She uses the database “almost every day” to access crucial information for an experiment process she was using. To find information for the process, she typed in the name of a protein she was studying and browsed to the page with information about that protein. Most of the information on the page was irrelevant to her, so she skipped to information such as protein size, sequence, map and interaction.

2.2.3 Conclusions

Graduate students need access to a lot of information in order to conduct successful research. They need to learn everything they can about their chosen subject, keep up with all new research papers that are relevant to their work, and gather information on how to carry out research techniques. Most students don't have a broad background understanding of their specific research area, so they have difficulty determining what information is relevant to them. In addition, they are often too busy conducting their own research to gather all the information they need, and they rely on the PI to help them. The PI in the yeast lab summarized the problem by saying “The people who understand the information the least are having to cope with information the most.” We present the problems we discovered in both labs in the following model:



- #1** Lab head spends hours and hours searching for papers. Sometimes misses important papers.
- #2** Overwhelmed by the number of papers they have to read. Unable to determine what's important and relevant to them.
- #3** Hard to find detailed info on techniques: Have to rely on trials and errors to find out.
- #4** Other biologists aren't always available to help you: Information doesn't get transmitted.

Problems Identified in the Information Flow

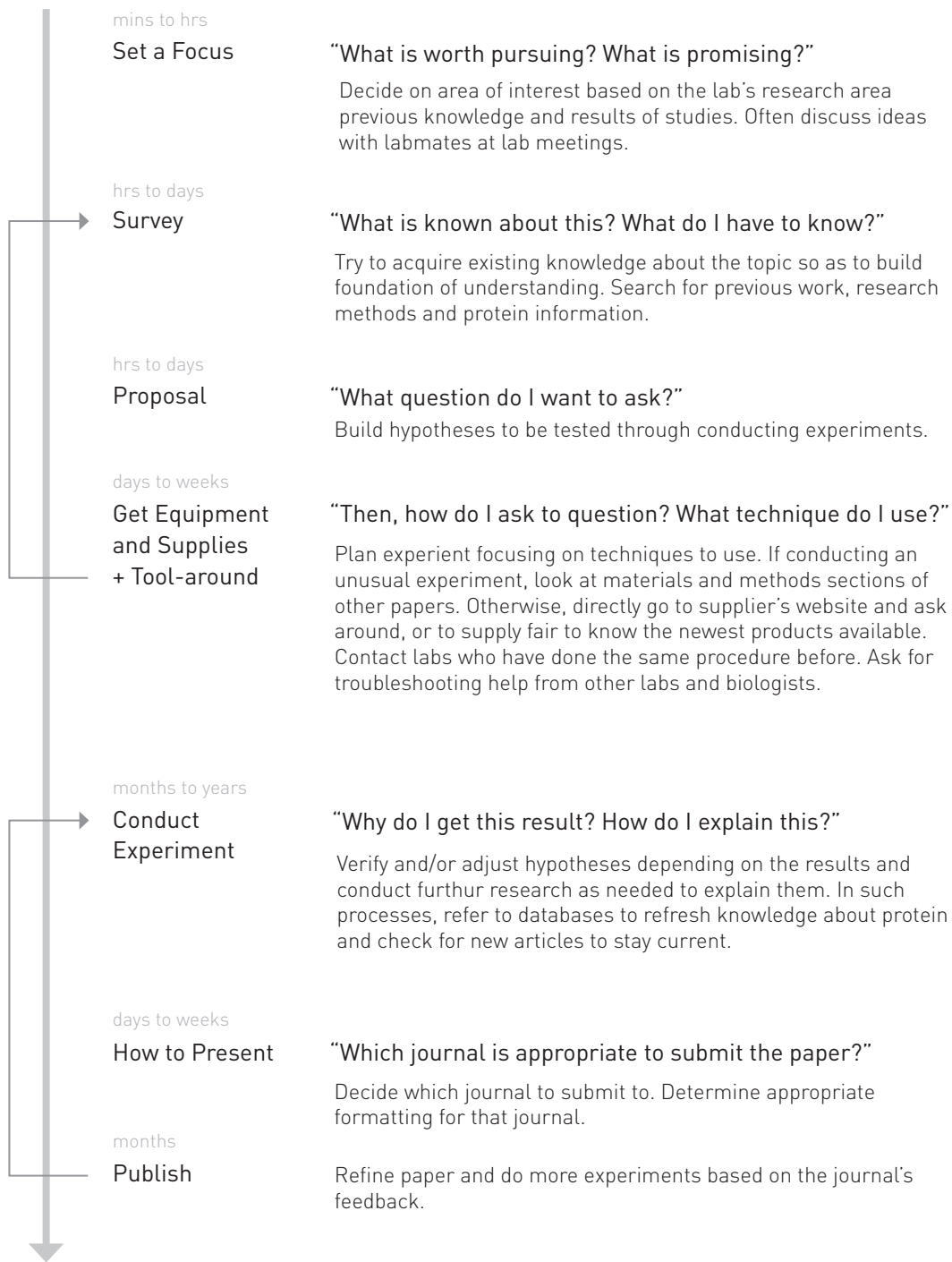
2.3 Findings: Research Process

2.3.1 Research Process and Needs

Through our interviews with a post-doc in the yeast lab and a graduate student in the fly lab, we modeled the research process for biologists. Each step shows a different phase in the research process and the information needs within each phase. As shown in the model on the next page, biologists go through an iterative process of adjusting their hypothesis or conducting further experiments in order to build their findings into a presentable unit.

2.3.2 Conclusions

A typical biology research cycle is five years on average, which is much longer than the research cycle in computer science. Biologists turn to multiple resources for information, and they look for different types of information at various stages of their research. For example, during the survey period biologists do extensive background research by searching for papers and information about related proteins in databases. In the tool-around period, however, biologists look for information that can help them conduct their experiment successfully by reading about research techniques in papers, or emailing other biologists for questions about a technique.



Research Sequence Model

3. Concept Creation

We conducted a brainstorming session to generate ideas on how to solve biologists' existing problems and needs. Afterwards, we created an affinity diagram with the brainstorming ideas to group similar ideas and consolidate them into fully developed solutions. We wrote 15 scenarios based on our design concepts and presented the scenarios to two of our interview subjects to discover if our ideas resonated with biologists. We recorded their reactions, and asked them to rate the helpfulness of our design on a scale of 1 to 5.

3.1 Scenarios

1

Debra picked a research focus, and was trying to learn as much as possible about her research area before she did her study. She typed in an author that has done work in her area and saw a bunch of papers in a tree diagram, seeing: all papers written by that author, papers that reference that author, and papers that reference the papers that reference that author. She changed the option to show by protein name, biological process or labs that did related research. The papers that were most likely to be directly relevant to her were at the top, and the tree spanned out to include papers that reference things that were related to her research. Review papers were also shown for related topics. She downloaded all of the papers that were near the top of her reference trees, and downloaded a few of the papers towards the bottom that explore interesting things related to her research.

2

Steve, a new graduate student arriving in the lab, wanted to quickly catch up with all of the important papers in his lab's research area. He went to his research-paper program, and typed in the area that the lab studied. From there he could visualize all of the papers that related to the lab's topic. The papers that were the most recent and referenced the most times were visually large on his screen. He downloaded and read all of the papers that were most frequently referenced.

3

Sara, a graduate student at the lab, was having trouble with an experiment technique. She was not sure how to do one of the steps, so she went to the "recipe book" on her laptop, typed in the kind of experiment she was running, and type of equipment she was using. The recipe book showed her step-by-step instructions, compiled from all of the research papers that discussed that technique. In addition it contained troubleshooting advice from other colleagues who had done that technique before. She quickly found the help she

needed and successfully carried out the experiment. Later, when she finished the experiment and was writing up her paper, she went back to the website to note the papers she referred to.

4

Steve, a new graduate student in the lab, needed to find an interesting and unique area to pursue with his research. He looked at the “discussion section” tool on his computer, and typed in a protein and biological process that he was interested in. From there, he was able to see all of the research questions that research papers had mentioned in their discussion sections. He looked for a cluster of questions that weren’t answered and that he was interested in. He mentioned his topic idea to the lab head, and he thought it was a great idea.

5

Sara, a graduate student at the lab, was struggling to find a new and interesting topic to research. She opened her computer program, typed in a protein and biological process that she was interested in, and saw a visualization that showed all of the papers that biologist were referencing in their papers. Most everyone was referencing that same papers, but a few authors were referencing unique papers. She checked out the papers with unique references and saw that they were exploring a different and interesting area. She decides to expand on the work that they were doing in her own research.

6

Debra, a graduate student at the lab, found some very exciting data, wrote a paper, and submitted it to a top research journal. Upon being accepted to the journal, a computer quickly analyzed the paper for important content and added the relevant finding to an on-line database. Debra received an email linking her to the work that had been incorporated into the database. She visited the database and saw how her research had added to what was known about that protein.

7

The PI for the lab needed to keep up with the most recent papers in his area. He kept a folder of important papers that related to his lab’s work. His computer analyzed those papers, and sent him paper-suggestions whenever new papers were published that contained similar biological processes, protein types, techniques, or referenced those authors (and the authors that those authors referenced).

8

The PI for the lab was tired of spending hours searching for new papers every week. He entered his preferences into a search program. He mentioned the biological processes,

protein types, techniques, and authors he was interested in. The program searched through all new biology journals and sent him abstracts for the papers that met his criteria. Based on the papers that James decided to download, the search program refined its criteria in order to send him the most relevant papers.

9

Recently, a colleague mentioned a protein to Debra that had a similar behavior to the one that she was studying, so she looked up information about the protein on SGD and PubMed. A computer program noted the information that she was looking up. A few days later she received an email informing her of a paper that just came out that discussed a new finding about the protein that her colleague recommended to her.

10

It's Monday morning and Debra, a post-doc in the lab, decided to look for new papers that related to things she was studying. After finishing her search she received an email that described papers that she might also be interested in, based on the papers she decided to download that day.

11

Steve downloaded a paper about a protein he was researching. His computer then offered him a few other suggestions stating, "biologists in your area who downloaded this paper also downloaded the following papers:". He saw many papers that he was unfamiliar with, but he downloaded one in particular that appeared to be very popular and very closely related to something he was studying.

12

Steve downloaded a paper about a protein he was researching. His computer then offered him a few other suggestions stating, "biologists in your area who enjoyed this paper also rated the following papers highly:". He downloaded a paper that was highly rated and that he was unfamiliar with.

13

Debra, a graduate student in the lab, finished her experiment and was ready to write a paper. The system recommended journals that might be appropriate for the type of results she had. After selecting the journal she wanted to submit to, the system told her the amount of text she needed to provide, the resolution of images, and the appropriate way to format references.

14

Steve wants to read “the most important papers in his area” but he can’t tell what papers are more important than other when he searches on PubMed. However, many review paper authors rate a paper’s importance using a 1, 2 and 3 dot system. Steve’s search program accounts for author’s rating and highlights the papers that biologists in his area find more important.

15

Steve tries to keep up with all of the journals that come out in this area, but he doesn’t have the time to sit down and read every article. His computer program scans all of the relevant journals and pulls out all of the preview sections- a description all of the findings that the journal finds most noteworthy. Steve receives an email of the relevant preview stories in his inbox.

3.2 Concept Validation: Findings and Implications

Presenting the scenarios to our subjects helped us identify the solutions that resonated the most and develop new concepts based on their feedback. The ideas that subjects did not like were those that they perceived as providing them little timesavings, and those that appeared to mimic “subjective” human judgment. For instance, the subjects did not think that formatting their papers took very much time, so our solution to automatically format papers was not important to them. Participants were also concerned about design solutions. In addition, subjects were uncomfortable with the design that recommended appropriate papers for submission, because choosing a paper is a “judgment call” based on how optimistic the researcher was about the chances of his work being accepted to a particular journal.

Ideas that subjects rated highly were those that they perceived as saving time and reducing frequent unpleasant tasks. Biologists are overwhelmed with information that they need to sift through, and our participants reacted very favorably to scenarios that reduced the time they spent looking for papers. Participants were especially attracted to design solutions that helped new biologists establish a big picture understanding of their field. In addition, they enjoyed scenarios that helped them discover new and unique research topics to pursue. For instance, the Principal investigator enjoyed the discussion section scenario, and said that it mimics what he tells his students: “look for the wave, not the one that just broke, but the one that’s 50 yards out in the ocean. Can you listen to the place that the smart people say the waves are going to be?”

4. Technical Review

We selected a set of five scenarios that resonated the most with biologists and reviewed them with machine-learning researchers to evaluate the technical feasibility. We discussed opportunities of applying existing technology to our design solutions, and discussed other variations of our designs. The following section introduces each of the scenarios and the discussion we had during the technical review meeting.

4.1 Theme: Gaining a broad understanding of a research area

Through our user research, we realized that the process of searching for papers related to one's research was unstructured and time consuming. Researchers searched for the earliest papers on a topic as these usually described the fundamental information, but even finding these papers was not an easy task. Also, we learned that understanding which papers referenced each other and which succeeded or preceded each other in the order of research helped identify which papers would be valuable to one's research. Visually displaying a hierarchy of the relationship between research papers could assist biologists in finding the information they needed.

Reading through a number of papers to gain a broad understanding on a topic was also a tedious process. Biologists valued review papers which summarized a number of research papers as these papers helped them quickly understand the topic, but review papers on certain topics weren't always available. We learned through our concept validation session with biologists that providing summaries of papers similar to review papers would be valuable to biologists.

These concepts are described through the following scenarios.

4.1.1 Scenario: Paper Tree

Steve, a new graduate student in a research lab, was trying to learn as much as possible about the lab's research area. He typed in a protein name and biological process that the lab was studying and saw many papers in a tree diagram. The tree modeled the flow of ideas in the field, with each branch showing the growth of a new idea. Given that he didn't have enough time to read all of the papers, he decided to download all of the earliest papers that expressed a new idea.

4.1.2 Discussion

Implementing a solution like this is relatively feasible, and there is related research currently being done on how to detect concepts that flow from paper to paper. There could be

different ways to define the relationship between papers such as using references, identical or related words in the content, relevance of authors, proteins or experiment methods. Using citations and determining why each citation was used could help identify the relationship between two papers. Also, because review articles are valuable and biologists have a difficult time finding them, clearly showing review articles in search results could be a simple, but useful solution.

4.1.3 Scenario: Instant Review Paper

Steve had a ton of reading to do to understand all of the important work in the area that the lab is interested in. Luckily, when his professor, James, sent him 40 papers to read, he asked his computer to generate an “instant review paper”. The computer looked for all of the important findings in the papers and eliminates the any repeated information. It provided him a 20-page review document. He read over the document that afternoon and felt like he had a much better understanding of the lab and the work in the lab’s area. Later when he wanted to read more in-depth about one particular area, he clicked on the part of the review paper related to that protein and downloaded the three papers about that topic.

4.1.4 Discussion

Summarizing information is an advanced cognitive process which computers are not currently capable of. One solution could be to build an information website based on social networking which encourages creating and sharing opinions on research papers. Graduate students often write hundreds of paper abstracts during the course of research and this information could become the basis for a project like this. Another possibility would be to consider methods similar to the Cognitive Atlas project.

4.2 Theme: Finding new papers in your research area

In addition to searching for papers when information is required, biologists frequently check journal websites and use search engines to find recently published literature that might be related to their research. This is currently a time-consuming process and biologists are concerned that they might not notice an important paper hidden in one of the many journals in their field. Machine-learning and search technology could be of great help to biologists with this problem.

4.2.1 Scenario: Weekly search

James, a professor and head of a research lab, was tired of spending hours searching for new papers every week. He entered his preferences into a search program. He mentioned the biological processes, proteins, techniques, and authors he was interested in. The program searched through all new biology journals and sent him abstracts for the papers that met his criteria. Based on the papers that James decided to download, the search program refined its criteria in order to send him the most relevant papers.

4.2.2 Discussion

Preferences could be explicitly defined by the user or learned by the system based on the user's previous search behavior. It seems highly feasible to implement a system that learns personal preferences using the currently available machine-learning technology. Other possible directions could be suggesting research papers that other researchers in your area are interested in or using a recommendation system similar to the "Faculty of 1000" website.

4.3 Theme: Detecting new research trends and determining research areas to pursue

The length of a research project in biology is relatively long, often three to five years. Selecting a good topic for research is therefore very important and researchers sometimes spend several months reading existing literature to fully understand the ongoing research and selecting an appropriate topic. We learned that the discussion section of research papers mention future areas to explore and additional questions that arise, and these often become the starting points of exploration. Researchers felt that having a compiled list of questions and areas to explore from multiple papers could assist them in finding promising research topics.

4.3.1 Scenario: Finding new areas from paper discussion sections

Steve needed to find an interesting area to pursue with his research. He looked at the "discussion section" tool on his computer, and he typed in a protein and biological process that he was interested in. From there, he was able to see all of the research questions that research papers had mentioned in their discussion sections. He looked for a cluster of questions that he was interested in. He mentioned his topic idea to his professor, James, and he thought it was a great idea.

4.3.2 Discussion

Considering how long research projects take to complete, helping researchers find an appropriate and promising topic is an important problem. However, there are some issues that need further investigation in order to determine the feasibility of this solution. For example, we need to see how well research questions are identified in a paper so that they can be extracted and organized. One valuable aspect of this solution is “forward reference” - linking the current paper with other papers published afterwards that cite the current paper. Forward reference would identify which questions in the discussion section have been answered and in which areas research is actually being done.

4.4 Theme: Obtaining knowledge on experiment techniques

Biologists spend a great amount of time during the experiment phase figuring out why an experiment result is different from what they expected and this is often due to a mistake in the experiment method. Figuring this out is a very time consuming process and this could be a reason why researchers are very hesitant in attempting to use a new experiment method they are not familiar with. Also, researchers rely on experts when they encounter problems during the experiment, but contacting experts is time-consuming for both ends and the information is not well shared with others. A solution for sharing detailed experiment methods and for collaborating with other researchers on problems during experiments could make this process much easier.

4.4.1 Scenario: The Recipe Book

Steve was having trouble with an experiment technique. He was not sure how to do one of the steps, so he went to the “recipe book” on his laptop and typed in the kind of experiment he was running. The recipe book showed him step-by-step instructions, compiled from all of the research papers that discussed that technique. In addition, it contained troubleshooting advice from other colleagues who had done that technique before. He quickly found the help he needed and successfully carried out the experiment.

4.4.2 Discussion

Researchers who have used an experiment method for their research usually have a detailed “recipe” for conducting the experiment, but this information is not usually published in the paper or shared effectively. Encouraging paper authors to create a web page with this information and linking it to their research paper could be a solution. If this took the form of a wiki, the information could naturally accumulate and become a self-growing repository. We could find similar recipes by using citations in research papers or pattern matching.

5. Conclusion

We discovered opportunities for machine-learning technology to help biologists gain a broad understanding of their research area, detect new research trends, and obtain knowledge on experiment techniques. By discussing our ideas with machine-learning experts, we learned that several of our design suggestions are within the realm of what is currently possible with machine-learning technology. In the next stage of the project, we will refine our design and create a high-level prototype. Working closely with biology researchers and machine-learning experts, we will continuously evaluate the feasibility and desirability of our design ideas.