Example: Undergraduate Research Project Design

CRA UR2PhD Program
Author: Kelly Shaw
Contact: ur2phd@cra.org

This worksheet provides suggestions on issues to consider when designing an undergraduate research project. It presents a series of questions to answer and steps to complete as you think about designing an undergraduate research experience. As with all plans, the plan you initially create will likely need to be adjusted as the research experience proceeds, but the process of thinking through these questions before the research experience begins will help lead to a more positive and productive experience for your student. When you need to modify your project plan in response to research realities, working collaboratively with the undergraduate researcher on those adjustments would be a value learning opportunity for the student.

This document provides a concrete example for using this worksheet to design an undergraduate research project. It is only a single example and should not be used as a rigid guide for using this worksheet; there can be very different levels of specificity in equally effective plans developed through the use of this worksheet. Working through the thought process presented by the worksheet is the most important aspect of using this resource.[2]

Brainstorming a meaningful, non critical[3] research question/project

One way to think about this is to think about the project creating a supporting or explanatory figure that would be nice to include in a paper or thesis, but is not essential to the success of the paper or thesis. Another possibility is to think of a project that would do some exploratory work that extends an existing idea into a new realm. Some possible questions to consider:

- Do you have some existing research results that are not fully explained and you would like to better understand by collecting and analyzing more data on an already existing system?

  Our research group's recently published paper implemented a new caching policy on a multiprocessor system. For 2 of the applications studied in the paper, the approach did not perform as well due to network congestion. Our research group would like to better understand the application characteristics that caused network congestion for those 2 applications when the new caching policy was used.

[1] UR2PhD: Graduate Student Mentor Training Course © 2023 by Computing Research Association's UR2PhD Program is licensed under Creative Commons Attribution-ShareAlike 4.0 International. To view a copy of this license, visit https://creativecommons.org/licenses/by-sa/4.0/

[2] If you're willing to share your completed version of this document to provide others with additional example approaches, please reach out to ur2phd@cra.org.

[3] The project should be non-critical for a first time student researcher because the student will initially be slow to produce results as they will be learning about research (including making lots of mistakes).

- Do you have some existing research results that have made you curious about how your tool or approach would apply or work in a different setting or for different inputs?
- Do you have existing collected data that you have not found time to analyze or visualize for specific characteristics, where the results could lead to deeper understanding of your approach or point to new problems to explore?
- Is there a small, relatively straightforward artifact (e.g, survey, software feature) that you need implemented and evaluated that builds on an already existing system or process?

Delineating the goals and steps of a research project

1. What is the precise research question you want this research project to answer?

   For the 2 poorly performing applications, what application characteristics resulted in network congestion when the new caching policy was used?

   - How is the answer to this question meaningful/helpful to your research?

     The answer to this question will provide insight into the situations that result in the new caching policy performing poorly and may lead to the design of an improved caching approach based on the understanding gained by examining these two applications' interaction with the caching policy.

   - List the sub questions associated with the larger research question. Identify which sub questions are complicated enough to be their own small research projects.

     - What is causing the network congestion? For example, is it the result of a large increase in specific types of messages or the result of an increase in a small number of network paths?
     - What application characteristics in conjunction with the caching policy trigger the sending of the messages causing the network congestion?

2. For a given research question or sub question, what is the set of small, measurable deliverables that must be achieved to answer it? (You may choose to structure this set of deliverables using an if/then organization if the work that needs to be done at a stage is dependent on the outcome of an earlier deliverable.) Completion of these individual deliverables should advance the project regardless of whether all of them are completed in the given time frame.

   - Implementation of data collection in the existing system simulator, which collects information about messages sent in the network and outputs raw data to files for post-processing.
   - Creation of several microbenchmark applications to test correctness of data collection implementation.

- Creation of Python scripts that read generated output files containing message data and analyze which messages (i.e., types of messages, number of messages, network paths) are resulting in network congestion.
- Data collection of message information for each of the two applications.
- Use of Python script to analyze causes of network congestion for each of the two applications.
- Mapping of messages causing network traffic to application characteristics for each of the two applications.

3. For each of these small, measurable deliverables, answer the following questions:
    1. What technical skills / knowledge are needed to complete each deliverable?

- Implementation of data collection in the existing system simulator, which collects information about messages sent in the network and outputs raw data to files for post-processing.
    i. Basic git skills
    ii. Extensive C programming and debugging skills
    iii. Knowledge of large, complicated simulator, including where and how to add data collection code
- Creation of several microbenchmark applications to test correctness of data collection implementation.
    i. Basic git skills
    ii. Knowledge of how and where to save artifacts
    iii. Basic C programming and debugging skills
    iv. Understanding of how to design microbenchmarks
    v. Understanding of how to read a technical paper
    vi. Understanding of new caching policy
    vii. Understanding of network modeled in system
- Creation of Python scripts that read generated output files containing message data and analyze which messages (i.e., types of messages, number of messages, network paths) are resulting in network congestion.
    i. Basic git skills
    ii. Knowledge of how and where to save artifacts
    iii. Python programming and debugging skills
    iv. Knowledge of Python libraries for data analysis
    v. Understanding of how to read a technical paper
    vi. Understanding of new caching policy
    vii. Understanding of network modeled in system
- Data collection of message information for each of the two applications.
    i. Basic git skills
    ii. Knowledge of how and where to save artifacts
    iii. Knowledge of how to execute applications on simulator to generate data collection output files

iv. Understanding of how to read a technical paper
v. High level understanding of 2 applications being examined

- Use of Python script to analyze causes of network congestion for each of the two applications.
  i. Basic git skills
  ii. Knowledge of how and where to save artifacts
  iii. Python programming and debugging skills
  iv. Understanding of how to read a technical paper
  v. Understanding of new caching policy
  vi. Understanding of network modeled in system
- Mapping of messages causing network traffic to application characteristics for each of the two applications.
  i. Basic git skills
  ii. Knowledge of how and where to save artifacts
  iii. Basic C programming and debugging skills
  iv. Understanding of how to read a technical paper
  v. Understanding of new caching policy
  vi. Understanding of network modeled in system
  vii. High level understanding of 2 applications being examined

2. What technical skills / knowledge will an undergraduate student have been required to learn in their coursework, given their current progress in the major?

- Python programming and debugging skills
- Basic C programming and debugging skills
- Understanding of memory systems, including caches

3. What is the difference between 1 and 2? These are the new skills the student will need to acquire to complete this deliverable.

- Implementation of data collection in the existing system simulator, which collects information about messages sent in the network and outputs raw data to files for post-processing.
  i. Basic git skills
  ii. Extensive C programming and debugging skills
  iii. Knowledge of large, complicated simulator, including where and how to add data collection code
- Creation of several microbenchmark applications to test correctness of data collection implementation.
  i. Basic git skills
  ii. Knowledge of how and where to save artifacts
  iii. Understanding of how to design microbenchmarks

    iv. Understanding how to read a technical paper
    v. Understanding of new caching policy
    vi. Understanding of network modeled in system
- Creation of Python scripts that read generated output files containing message data and analyze which messages (i.e., types of messages, number of messages, network paths) are resulting in network congestion.
    i. Basic git skills
    ii. Knowledge of how and where to save artifacts
    iii. Knowledge of Python libraries for data analysis
    iv. Understanding of how to read a technical paper
    v. Understanding of new caching policy
    vi. Understanding of network modeled in system
- Data collection of message information for each of the two applications.
    i. Basic git skills
    ii. Knowledge of how and where to save artifacts
    iii. Knowledge of how to execute applications on simulator to generate data collection output files
    iv. Understanding of how to read a technical paper
    v. High level understanding of 2 applications being examined
- Use of Python script to analyze causes of network congestion for each of the two applications.
    i. Basic git skills
    ii. Knowledge of how and where to save artifacts
    iii. Understanding how to read a technical paper
    iv. Understanding of new caching policy
    v. Understanding of network modeled in system
- Mapping of messages causing network traffic to application characteristics for each of the two applications.
    i. Basic git skills
    ii. Knowledge of how and where to save artifacts
    iii. Understanding of how to read a technical paper
    iv. Understanding of new caching policy
    v. Understanding of network modeled in system
    vi. High level understanding of 2 applications being examined

4. Of these new skills/knowledge that must be learned to complete the deliverable, which ones can be learned by an undergraduate sufficiently within a few days or up to a week? Which ones would take longer to acquire? For skills/knowledge that take longer to acquire, can you or a more experienced student complete the tasks requiring those more advanced skills/knowledge instead?

- Skills/knowledge that can be acquired quickly
    i. Basic git skills

ii. Knowledge of how and where to save artifacts
iii. Knowledge of Python libraries for data analysis
iv. Understanding of how to design microbenchmarks
v. Understanding of how to read a technical paper
vi. Understanding of new caching policy
vii. Understanding of network modeled in system
viii. High level understanding of 2 applications being examined
ix. Knowledge of how to execute applications on simulator to generate data collection output files

- Skills/knowledge that require significant time to acquire. (A graduate student could quickly complete tasks requiring these skills.)
  i. Extensive C programming and debugging skills
  ii. Knowledge of large, complicated simulator, including where and how to add data collection code

5. Are there resources that provide an example solution to a similar problem that the student can learn from?

- Knowledge of Python libraries for data analysis - example Python scripts for analyzing other data generated by system
- Understanding of how to design microbenchmarks - example will be created for student
- Understanding of new caching policy - published paper and presentation
- Understanding of network modeled in system - network textbook and network course materials can be used
- Knowledge of how to execute applications on simulator to generate data collection output files - examples scripts exist for executing applications on simulator

6. Have you identified tools/techniques/papers/documentation that enable the student to acquire those technical skills / knowledge, and is there someone who can answer questions on those materials?

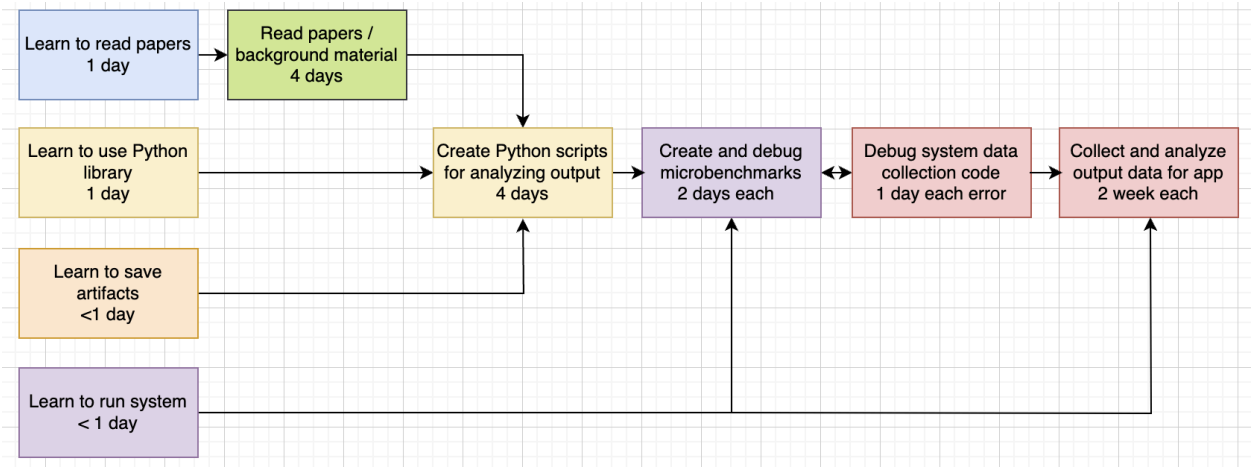
- git online tutorial
- Python libraries for data analysis
- Papers and talks on reading a technical paper
- New caching policy paper and talk
- White paper describing system
- Chapters of networking textbook and network course lecture slides for modeled system

7. Are any of these deliverables decomposable into repetitive subparts that can be tackled independently?

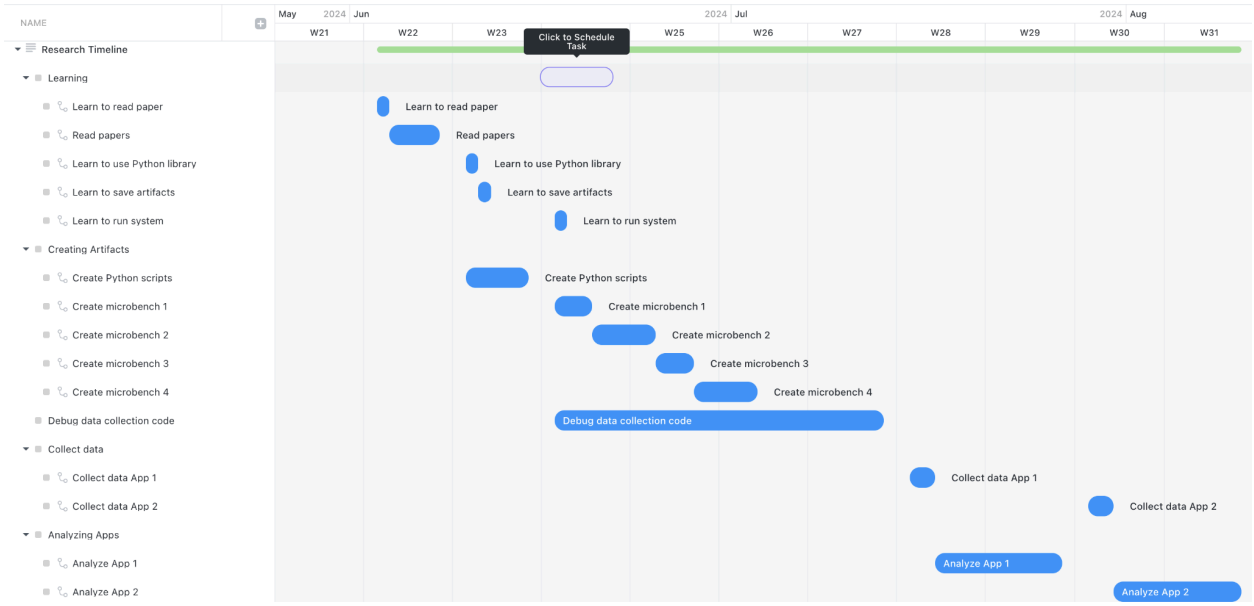
   - Creation of multiple microbenchmarks can each be tackled independently
   - Data generation and analysis for each of the 2 applications can be tackled independently

Creating a project timeline

1. Create a dependency graph where each item is either
   a. a small, measurable deliverable as defined above
   b. a technical skill or knowledge that must be acquired
2. For each item in the dependency graph, determine a **realistic** amount of time it will take for an undergraduate student to complete. For deliverables with decomposable parts, determine a time estimate for each independent part. (Be slightly pessimistic in your time estimates)



3. Map the tasks from 1 onto a Gantt chart using the information from 1 and 2.



4. **If the resulting timeline is too long** for the research experience's duration, consider places you can cut tasks or parts of tasks in order to revise the Gantt chart. For example,
    a. Can repetitive tasks be reduced to fewer repeats?

- Fewer microbenchmarks could be created
- Only 1 application could be studied

    b. Can a team of students work in parallel on repetitive tasks?

- The microbenchmarks could be created in parallel
- The 2 applications could be studied in parallel

    c. Can one of the defined deliverables become the new final goal, with the expectation that the remaining deliverables will be completed by the same or another student in the future?

- Creation of the Python analysis scripts could become final goal
- Any of the microbenchmarks could become a final goal
- Generation of an application's network traffic data could be a final goal
- Analysis of the first application could be a final goal

5. Keep in mind that this initial timeline should be viewed as a *draft* plan that will likely need to be adapted during the student's actual research experience based on the progress being made and the challenges encountered. As the research experience proceeds, keep in mind ways to expand or shrink the set of tasks as required

Stitching Multiple Research Projects Together to Answer the Original Research Question

- Have you established a shared space for each deliverable and its supporting data, artifacts, analyses, documentation, experimental notes?
- Have you established the organization and protocol the student must use to share each deliverable? This includes when items should be shared.
- Have you established a protocol for reviewing the student's deliverables and providing feedback for revision? This includes how quickly feedback should be given.
- Have you ensured that you will retain access to these documents when the student leaves the group and/or graduates?
- Have you shared these expectations with the student and provided any necessary documentation for performing these steps?