Jedediah Smith

BIFX-545

Discussion Questions

2/1/2022

Questions to help me get to know you:

1. Write 5-6 sentences about someone you know personally or have learned about (could be someone from history) whom you feel is a good leader. What makes them a good leader? What could they do differently to improve?

Winston Churchill and George S. Patton are two important leaders from WWII. They were both strong, confident, and inspiring leaders that helped the Allies win the war. However, it could be said that they were not particularly diplomatic people. It was often “my way or the highway” with them. This caused trouble for both of them in the postwar years when they began to vehemently criticize disparage the Soviet regime that had helped the Allies win WWII.

1. Write 5-6 sentences about the type of biology/computation that you currently do or most want to do. What topics are most familiar to you?

I am most familiar with genetics and developmental biology. My final project for undergrad focused on fruit fly genetics and development But in addition to these topics, I am also personally interested in neuroscience and fungi. I think these topics are most interesting and familiar to me. It would be awesome to work more in these fields going forward, but if not that’s okay too.

1. For Bioinformatics MS students only: When do you plan to complete your capstone project? Do you have a subject or laboratory in mind?

Just started my second semester, not entirely sure when I am supposed to start working on this. I’ve heard lots of good things about Leidos, and am personally interested in working with the federal government in the future. So I’ll probably look at these places when the time comes.

BDS Chapter 1 Questions:

* 1. Large bioinformatics databases such as the Sequence Read Archive have grown exponentially in the last decade and new bioinformatics tools are continually being invented and improved. What are the downsides to the abundance and rapid development of bioinformatics tools?

Bioinformatics tools may not be adequately benchmarked, and when they are, it may only be for one organism. As a result, it becomes hard for scientists to find the best tools to analyze their data. Additionally, it puts strain on bioinformatics programs, forcing them to constantly adapt and update their programs or face irrelevancy.

* 1. What barriers exist to sharing data and methods?
     1. What barriers to sharing data and methods are true regardless of the type/size of the data?

Links to supplemental materials, methods, or data may break. Materials on faculty websites often disappear when faculty members move, leave, or update their site. Software projects become stale when developers leave or do not update their code.

* + 1. What barriers to sharing data and methods are specific to large datasets and complex computational methodology?

Longer supplementary methods, code, and data. Massive amount of storage that data may occupy. Reference genomes and annotation datasets are constantly updated from the form used in an experiment.

* 1. Do you agree or disagree with Karl Popper’s claim that “non-reproducible single occurrences are of no significance to science”? Why?

Disagree. I think they are in important part of the scientific process which help remined us of the fact that don’t have all the answers. While such flukes may not advance our understand of the world around us, this does not necessarily mean they are insignificant to the entire process as a whole.

* 1. Describe silent error.

Silent error may occur when code or programs continue to produce output despite the presence of an error or bug, rather than stopping or shutting down.

* 1. Why is silent error much more prevalent in scientific computing than in other kinds of programs such as video games?

Scientific code may only be run once or twice, then set aside once the desired output is acquired. This is in contract to video games, where it will be rigorously tested thousands of times by testers, and then played millions of times over by those who buy the game.

* 1. Why can it be especially difficult to identify errors when working on large and high dimensional datasets such as genomics datasets?

Output may be too large and high dimensional to easily inspect or visualize. If there are multiple steps, and it is typically not feasible to inspect after each step. Harder to form prior expectations about what an outcome will look like than wet lab biology. Programs that are vetted and used widely might not work for all types of organisms.

* 1. Why is it helpful to write code in a way that is easy for humans to read it, when it is best to record data in a way that is easy for computers to read it?

It is helpful to write code for humans because humans are the ones doing debugging. This also makes it more reproducible and easy to revisit old projects. To this end, code should be readable, broken down into small contained components, and reusable. Unless you are formatting data to present it, code should be optimized for computers to read, that way it is easier to run analysis.

* 1. Describe several reasons why it is important to treat data as read-only.

Modifying data in place can easily lead to corrupted results. It is easy to lose track of how a file is modified when done by hand instead of a program, making it irreproducible.

* 1. Why is it helpful to make figures and statistics the results of scripts in order to make a project reproducible?

It may take extra time to write out the code to generate figures and statistics, but it saves time if things need to be rerun later, and helps keep the project reproducible by ensuring that all figures or statistics will be generated the same way using the same code.