Kenneth Thompson
CS 461
CAPSTONE
Fall Term

Abstract: In 2016-2017, three computer science seniors from OSU developed software to detect the damage caused by coal mining. Their work was impressive, they developed a suite of algorithms designed to detect and track damage using satellite imagery. They even managed to win an award and get mentioned in the OSU alumni newspaper. However there were several key shortcomings in their work that was unresolved at the time of graduation. Our project, and our goal, is to take their work, and improve upon it. We will be doing it in 4 ways. 1: We will take the existing algorithms and make them more efficient and run faster. 2: We will add in the Global Reservoir and Dam Database (GRaND) and other water system datasets such as the National Hydrography Dataset (NHD) to make it more accurate. 3: We will develop a set of key metrics to determine the damage caused by coal mining, giving us something to compare other areas too and see how much worse they are then a generic similar area. 4: We will extend coal as a set of reusable components which can be used in cloud based platforms such as XSEDE HPC resources.

In 2016-2017, three computer science seniors from OSU developed software to detect the damage caused by coal mining. Their work was impressive, they developed a suite of algorithms designed to detect and track damage using satellite imagery. They even managed to win an award and get mentioned in the OSU alumni newspaper. However there were several key shortcomings in their work that was unresolved at the time of graduation. Our project, and our goal, is to take their work, and improve upon it. We will be doing it in 4 ways. 1: We will take the existing algorithms and make them more efficient and run faster. 2: We will add in the Global Reservoir and Dam Database (GRaND) and other water system datasets such as the National Hydrography Dataset (NHD) to make it more accurate. 3: We will develop a set of key metrics to determine the damage caused by coal mining, giving us something to compare other areas too and see how much worse they are then a generic similar area. 4: We will extend coal as a set of reusable components which can be used in cloud based platforms such as XSEDE HPC resources.

At the end of the project, we will improve the algorithm, provide the metrics, and make it run on High powered clusters.

Our proposed system will be to break each issue down into multiple steps. For the first one, we will take the current algorithms and run benchmarks to get the speeds, this will give us some data to work with. We will go through the existing code, algorithm by algorithm, and create goals for each one that we can try to strive for. It will be important to be both realistic and optimistic, as well as ensure that the integrity of the code, that is to say the actual information generated, is not degraded in any way. We will then start testing, improving, and trying to reach the self set goals, each step making our code more efficient. And finally, at the end, we will be analyzing the generated results to ensure we are still collecting accurate information. This process should take several months, as the code base for the existing algorithm was rather large.

For the second issue, we will be analyzing the existing code, and figuring out ways to implement in the databases. This will be done concurrently with the first goal, since as we are testing and improving the code, our understanding of it will be increased. Its also important we dont complete this step after step 1, since after this is implemented in our code, we will need to ensure that it is as efficient as possible in order to make sure we arent wasting resources on the high performance computer (HPC.) The end state will be having data be generated that is more accurate than before thanks to these new resources of data.

For the third issue, this will be done after step 2, but it does not necessarily need to be done after step 1. While we dont want to waste HPC resources, the metrics should be fairly easy to produce, and unlike other scans which might be ongoing, this should in theory be a one time sunk cost. We will analyze a statistically significant amount of land, use it to determine some key metrics for average coal damage, average environmental variables, basically anything that could be impacted by coal mining in some way, we will be wanting information on what the readings should look like, if the coal mining hadnt taken place. The end state of this section will be a sheet of known values that all other areas can be compared to as being either 'worse' or 'better.' If the area of land that is deemed necessary for 'statistical significance' is too large, then we will conduct this step last, so that valuable resources on the HPC will not be wasted.

For the fourth and final issue will be bringing it to the HPC, converting the existing framework of algorithms into something that can be used to crunch through huge data sets. This will be either the last step, or the second to last step, depending on the area required for statistical significance. After the algorithms have been made more efficient, more accurate, then we will work on making them compatible with the cloud, and able to be reused for numerous projects. This section will reach its end state once their is a fully functioning set of algorithms able to run on the HPC without issue.

Performance metrics will be depending on the section. For the first section, we will be generating our own perfomrance metrics. But to summarize, they will be the percentage gain in efficiency and computer run time over the existing code. The metrics for the second will be the increase in accuracy. This is a rather hard thing to measure, but we should be able to use known values and on the ground collected data to see if we created something more accurate. For the third section, the performance metric will be the final sheet of baseline metrics that we produce, once we have that we will know we are done with this step, although its always possible to refine and improve the metrics to ensure they are representative of the area as a whole.

We have not yet talked to our client, but once we do we will have several key questions in regards to how the integration of the databases should take place, how much HPC resources we have access to, and what kind of metrics he is interested in creating.