Final Project Proposal

Members:

Michael Rogers (With assistance from colleagues where needed in preprocessing)

Motivating Problem & Objectives: Currently, I am working for a Cloud technology consulting company and we have a rather unique problem that needs solving. First, a little bit of background on what we do: at my job, we work with our clients and we are partnered with Amazon Web Services (AWS) to migrate customer's websites or apps onto the AWS platform. We assist with building infrastructure on the cloud for them to host said apps or websites as well as security, among so many other things that can be done in the Cloud. We recently took on a fertility clinic as a client and they had a very interesting project that they needed assistance with. This clinic helps patients to determine weather embryos that have been fertilized contain fatal chromosomal disorders in their DNA. These disorders are classified by "gains" or "losses" in the copy number variation in the DNA of the embryo. Copy number variation is when sections of the genome are repeated. However, when there are copies of a chromosome, almost always, it causes a miscarriage or the death of the fetus. One of the only chromosomal disorders that an embryo can survive is Trisomy 21, commonly referred to as Down Syndrome. This is when a person has three copies of chromosome 21 in their genome. But how does this apply to Information Visualization? Visualization is incredibily important to solving this problem. This company has files that contain sections of the genome, and the proteins that are present in the genome, but these files only contain the raw data. It is impossible to see if there is a variation in a specific chromosome from looking at the data alone. The data must be visualized in a way that makes it easy to read, and easy to asses weather an embryo contains these fatal copy number variations. Luckily, the client has provided us a reference solution so we can know if we are doing the process correctly.

Rough Plan of Work: The bulk of work to be done in this process is preprocessing of the data, but I have a plan in mind to make the raw data readable so that I can make a graph that is easy to read. I have found a python library called cnvkit which assists with this exact problem in processing the human genome. These files are presented in a .BAM format. The objective is to get this .BAM file into a format that is visualizable such as a CSV, TXT or something along those lines. I will be using Amazon Sagemaker, which is essentially a Jupyter Notebook run in the cloud to do this preprocessing. Not only does the file need to be converted to a readable format, but it also needs to be aggregated into the chromosomes so we can visualize the gains or losses in the genome. I know this may seem like a lot of work, but with my colleagues help, and this python library, I think this is a very manageable project in the time provided. Not only this, but I have begun work on the project and cnvkit seems to be working as expected.

Deliverables: Once these files have been converted, I will be submitting the visualization of the gains and losses as well as the expected results. These results will be delivered in a simple

scatterplot made with cnvkit, or another python visualization library such a matplotlib. I will also provide the CSV or TXT that is made from our process, so you have an idea of the data that is being plotted.

I’m super excited to get started on this, and incredibly eager to hear your feedback.