Ideas:

I think once you have an idea of which project appeals to you, then we can easily scope out the credits and the work you will be doing.

Usually start with Google scholar to see what papers might touch on the topic. Once I find one that looks good, I generally prowl through the papers that they cite and see who cited that paper. I wouldn’t go back more than 3 or 4 years right now. If we decide to look at more “historical” stuff, we can do that later.

1. Text mining: can we identify the top terms in job posting for different data jobs and see if there are particular tokens that would help classify the job title? For example, is “reporting” or “Tableu” more often mentioned in job descriptions that are “BI” oriented? What data activities are found in non-data jobs? (Creating a measure of expectations of data literacy would be AMAZING!). This would involve a LOT of scraping of LinkedIn/Indeed, etc or the discovery of a data set with the content. Ideally it would be from pre-covid but that’s asking a lot. We could use various classification techniques.

2. How has the need for data skills changed over time? What did job descriptions look like 10 years ago compared to today? Is there any relationship between those changes and the median salary for jobs with “data” work in them? Again a fair amount of parsing but for this we might be able to use some government info like the BLS Occupational Handbook if they have versions from today versus 10 years ago. It would also involve combining text and numeric data to get the relationship to wages. We could also do some stuff with Google searches over time? This sounds like a regression analysis unless we bucket the wages into categories then it’s a classification problem.

3. How has the education for “data” stuff evolved? When did “analytics” and other data oriented topics first appear in course catalogs? How prevalent is it now? This may be the most ambitious given the size of the higher education field but limiting things to, say, business schools or even undergrad programs would likely cut down on the amount of text to parse. And then we could compare the educational offerings with the job postings. Even as just a correlation exercise, I think this could be cool.