SML312 Homework #1 Data Science Project Questions Due Mon 9/12/2021

1. Data Science Past: The following questions ask you to describe the data science projects you have worked on, and to think about your process, choices, and challenges. They will help you better understand the process of doing a data science project, and prepare for the next one.

1a) What Data Science Project(s) have you worked on? What did you like about them?
One of my main projects is designing an algorithm that identifies when and where crop residue burning occurs in Punjab, India based on 4-band high resolution satellite imagery. This burning produces extensive pollution and emissions, yet is necessary for Indian farmers to keep up with their seasonal crop rotation. Addressing the consequences of crop residue burning in Punjab requires precise characterization of its spatio-temporal distribution. My work is on developing an ML model that takes a time series of satellite images and can classify the date upon which each pixel of farmland becomes burned. I've had several other smaller projects but I don't particularly consider them worth remembering or talking about. I like working with geographic data and having the power to create and uncover useful information and knowledge from such data.

1b) What high-level steps did you go through to complete the project(s) from start to finish?(i.e. gathered data, processed data, chose models, etc.)
For this project I had to go through pretty much the entire data science workflow aside from asking the question (as it was an already ongoing area of research in my lab). To gather data, my team and I had to purchase and download data from multiple satellite imagery products; then we did extensive processing on the data including removing missing or faulty images, extracting spectral band values, calculating spectral indices, and interpolating missing values. We are in the process of choosing an appropriate model; we have tried different kinds of classification models such as RF, different numbers of features, changing and improving our inputs into the model, the proportion of training and validation data, etc.

1c) What choices did you have to make in the project(s)? Was anything difficult?
We had to make a ton of choices, including: choosing satellite product(s), choosing how to interpolate missing data, choosing our model, choosing what indices are best for input into the model, and choosing whether to focus on spatial or temporal detection of burning. The most difficult choices were choosing how to interpolate missing data and choosing the model itself (both of which are still in progress) because there are no clear answers or precedents in literature, and burning is quite unpredictable.

2. Data Science Present/Future: These questions will help to guide you through the process of designing a data science project for yourself, which will be the main project to pursue this semester! They follow the general data science development process of trying to find and use data to *answer a compelling question*. Your project/data/approach will likely evolve as you explore -- that's encouraged and a perfectly normal part of the process. Your thoughtful choices will help to tell a compelling story about your question and the data related to it.

2a) What compelling question is your project trying to answer?
What regions in the Front Range foothills of Colorado are at greatest risk for wildfire damage?
2b) What data might be helpful in answering this question? Why?
Satellite imagery, vegetation and forest data, digital elevation models (DEM), shapefiles on power line distribution and conditions, regional climate data, regional residential property data, locations and status of fire stations and emergency services, water bodies and rivers. These factors are all very influential in determining potential wildfire severity, spread, damage, and the capacity of a response.
2c) How might you use this data if you had it?
I would use satellite imagery for classifying vegetation types and susceptibility to wildfire and DEMs for possibly calculating how a wildfire might spread. I'd use power line data to evaluate areas of risk of fire ignition from faulty lines and regional climate data to assess what times of year, weather, and seasons may pose the greatest fire risk. I'd use the locations of fire stations and water bodies to evaluate how quickly and effectively a response could be initiated and fire lines could be built. Property data would be used to estimate the cost of potential wildfires.
2d) Where could you look for this data? (Please take at least 30 minutes and search online, and describe any data sets you find that might be interesting to look at. Include links to whatever you find that seems interesting.)
USGS EarthExplorer https://earthexplorer.usgs.gov/ has a large suite of accessible satellite imagery. OpenStreetMap https://www.openstreetmap.org/ may be useful for street-level data such as roads and property that can influence a potential fire response. ESRI Open Data Hub http://opendata.arcgis.com/ provides a very large suite of miscellaneous geographic data sets from various organizations. There is also Open Topography, which can provide LiDAR data to make a DEM. NASA Earth Observations provides real-time climate observations: https://neo.gsfc.nasa.gov/. USGS and USFS also provide a large array of geographic data on geology, forests, land use, etc.
2e) What types of Data Science tools will help you to understand your question? How?
(e.g. Classification / Regression / Clustering / Prediction / NLP / etc.)
For identifying and classifying vegetation types with satellite imagery, I would primarily use classification, and possibly regression and clustering. As this question motivates a risk analysis kind of project, there could also be a lot of statistical estimation of risk based on all the factors and training of models on large geographic datasets. In general, classification, regression, and clustering could help me identify what regions are exposed to the greatest risk based on the surrounding risk factors.
Design/Data Note: If you already have a dataset in mind, that's great, but it's *generally better to start from a question* than from a dataset! **Your answer is only as good as your data** (at best!), and there may be additional data available that can offer deeper/different perspectives on the same question. It's always good to reassess your data choices as you continue to investigate your question! Thoughtful choices here can turn a good project into a great one!