1. Sam Blumer | [blumer60@gmail.com](mailto:blumer60@gmail.com) | 303-416-0664
2. The Impact of Population Growth and Urbanization on Food Availability
3. High level description of the project: what question or problem are you addressing?

Currently, there are over 7 billion people on Earther, and by the year 2050, it is estimated that there will be another 2 billion more. According to a survey by the United Nations in 2009 and the International Organization for Migration in 2015, it is “estimated that around 3 million people are moving to cities every week. Approximately 54% of people worldwide now live in cities, up from 30% in 1950. Sources estimate this will grow to 2/3 of world population in the next 15-30 years.”

Ultimately, ensuring that there is enough food, fiber, and infrastructure for for this mass urbanization will require in-depth analysis and understanding of trends and patterns in information. I would like to utilize a variety of data sources to analyze population growth, urbanization, and crop/animal food production/stores.

1. What type of data science task is it? (some example answers but not limited to)

I would like to utilize to unsupervised machine learning to classify or cluster counties in the United States that are most at risk of becoming food deserts due to poor, or fewer food choices. This would include visualizations.

1. Data:

I expect that this will be a big data project, as it will require data from multiple sources, including:

Census Data, Historical Migration data, historical crop/animal product yield, USDA data, and data from the [Food and Agriculture Organization of the United Nations.](http://www.fao.org/faostat/en/#data/EL)

I intend on using Sparklyr or Microsoft Azure to store and process the big data.

1. How will you analyze the data? What machine learning methods do you plan to use, and/or what business intelligence aspect do you plan on incorporating?

I expect that a large part of my project will revolve around exploratory analysis of the data to identify patterns. This will include using either Principal Component Analysis or Random Forest Variable importance to identify the most influential factors. Then I plan on using a probit regression to offer a probability on a county-by-county basis for the likelihood of becoming a food dessert in a designated number of years.

1. Describe any anticipated difficulties and problems. Discuss how you may overcome the problems.

I think one of the largest problems is going to be data collection and storage. However, as mentioned above, I intend on taking advantage of Microsoft Azure of the Spark Framework for quicker storage, processing, and recall.

Additionally, while I have always had a strong interest in agriculture, it is a newer space for me, and that means there will be a large amount of knowledge gathering that I will need to do to make sure that I am asking the right questions and analyzing the data correctly. As a solution to this, I have been reading about the current problems relative to urbanization and food shortages, and how data science teams are currently working to offer solutions to both.

1. Suggest a timeline for the project.  This should be a weekly breakdown of what you plan on doing each week.

Week 1 – Project Proposal

Week 2 – Data Collection

Week 3 – Data Collections

Week 4 – Exploratory Analysis

Week 5 – Exploratory Analysis

Week 6 – Model Development

Week 7 – Model Improvement

Week 8 - Visualizations

Resources:

<https://graylinegroup.com/urbanization-catalyst-overview/>

<http://americannutritionassociation.org/newsletter/usda-defines-food-deserts>

<https://www.usda.gov/topics/data>

<http://www.fao.org/faostat/en/#data>