Capstone I Ideas
July 9, 2018 Cohort

Below are three (3) ideas I am considering for my Capstone Project I. If you have time, please feel free to take a look and let me know if you have any questions or would like any additional information. Your feedback is greatly appreciated!

1. Pitch Prediction

Goal: Predict the next pitch thrown by MLB Pitcher - Benefit Hitter by knowing the next pitch to come
Alternate Goal: Identify and predict the next pitch likely to get a hitter out - Benefit Pitchers

Data Sets

PITCH*f/x*: http://baseball.physics.illinois.edu/pitchtracker.html
Pitch tracking system, created by Sportvision, and is installed in every Major League Baseball (MLB) stadium since around 2006. This system tracks the velocity, movement, release point, spin, and pitch location for every pitch thrown in baseball, allowing pitches and pitchers to be analyzed and compared at a detailed level.
Note -- The link doesn't really take you to the data, but does provide some good information. The data can be accessed using R and/or Python and stored in SQL database approximately 6 GB in size.

The Lahman Database: http://www.seanlahman.com/baseball-archive/statistics/
The updated version of the database contains complete batting/pitching statistics from 1871 to 2017

Major League Baseball Website: http://www.MLB.com

I already see where there may be some issues with the data. I read that the data may have some errors with mis-identifying the type of pitch thrown or not being able to identify the pitch at all. I have an idea of how to determine if some pitches are mis-labeled, but may require another resource provided by **Brooks Baseball** (http://www.brooksbaseball.net/) which I understand has already done a great job of correcting the data, but only seems to provide access to the data through interactive web queries.

Some reference articles and links I am currently using to understand the data provided through PITCH*f/x*:

- http://baseball.physics.illinois.edu/FastPFXGuide.pdf
- http://www.sloansportsconference.com/wp-content/uploads/2012/02/98-Predicting-the-Next-Pitch_updated.pdf
- https://fivethirtyeight.com/features/baseballs-new-pitch-tracking-system-is-just-a-bit-outside/
- https://www.beyondtheboxscore.com/2014/7/22/5919581/why-pitchfx-is-important

I have only become interested in Baseball in the last year and a half and literally had to look up the difference between a Fastball and Fastball Cutter but I am learning. I initially considered only how this

could benefit the Hitter to know the next Pitch to be thrown by the Pitcher, however, we can easily flip the information to benefit the Pitcher to know which pitch is more likely to get the Hitter out.

2. Predict disease using Human Gut Microbiome Genomic Data

Goal: Predict the health or disease of an individual based on their Human Gut Microbiome Genomic Data

Data Sets
AWS Human Microbiome Project: https://registry.opendata.aws/human-microbiome-project/
NIH Human Microbiome Project: https://commonfund.nih.gov/hmp/databases

I have not spent much time on this project but I do know that there is a ton of data available through AWS, NCBI, and other data repositories. The format of this data would most likely come as BAM files which are compressed FASTQ/FASTA genomic data files. These FASTQ files include sequence data and their corresponding quality scores. I would need to use a third-party tools/applications to analyze the data, identify operational taxonomic unit (OTUs), perform other steps to clean the data.

There are a ton of papers published on this subject, but a few that I am including discussing Human Gut Microbiota and its impact on health:

- https://www.sciencedirect.com/science/article/pii/S0092867412001043
- http://diabetes.diabetesjournals.org/content/62/10/3341.full-text.pdf
- https://www.nature.com/articles/nature08821.pdf
- https://reader.elsevier.com/reader/sd/94E15663841DEDA4AE12E57E473B89466ADDB9F22CE4034130632489C9403F3946E1D3D76E3783CE9E66C41F00687271

I think that this project may require more data cleaning and formatting than time allows for the first Capstone Project but if anyone has suggestions of better approach or better question to ask using this data, I would love to discuss it!

3. Identify key indicators of a student loan going into default status

Problem: The increased number of student loans and the ever-growing amount has placed a choke hold on many students looking to move forward in their careers and lives but are held back by debt.

While many students are working to pay off their loans, some if not many end up in default. How can we determine why some loans get paid and others go into default? Is it related to the schools (private, public, technical, 2-year, 4-year institutions), field of study, students economic background, whether they complete their education or not? Is there is minimum salary to debt ratio necessary to avoid defaulting?

These are all the questions I would like to answer with data provided by the following sources:
https://studentaid.ed.gov/sa/about/data-center/student/default
https://www.ed.gov/news/press-releases/national-student-loan-cohort-default-rate-declines-steadily