girls who code

at the University of Michigan Department of
Computational Medicine and Bioinformatics

2017-2018 Capstone Project

REMEMBER, the Facilitators are here to give you guidance and help whenever you need it. Please let them know when you have questions or want to discuss something!

1. Choose a data set

List of public datasets by subject:
https://github.com/caesar0301/awesome-public-datasets

MEDLINE/PubMed citation records
This database containing publication record and abstract is available in XML format at https://www.nlm.nih.gov/databases/download/pubmed_medline.html . This data can serve as useful resource for text mining. The abstract text is typically modeled as a bag of words, and can be clustered by topics, disease of interests, or authors for example.

Lahmann's baseball database
A comprehensive database of baseball records from 1996 to 2016 is available at http://www.seanlahman.com/baseball-archive/statistics/ .

Twitter live feed
Twitter provides users with python APIs to stream tweets based on specific keywords. An official description of twitter API is available at https://developer.twitter.com/en/docs, and DataCamp provides an introduction on how to import data using Twitter APIs at https://www.datacamp.com/courses/importing-data-in-python-part-2.

Biological data sets are available from DCMB faculty members if you have a particular topic of interest please let a Facilitator know.

Other public datasets
Fivethirtyeight: https://github.com/fivethirtyeight/data
BuzzFeedNews: https://github.com/BuzzFeedNews

Document adapted from Dr. Hyun Min Kang, 2017

Amazon Web Server : https://aws.amazon.com/datasets/?_encoding=UTF8&jiveRedirect=1
Kaggle: https://www.kaggle.com/datasets
US Government Open Data: https://www.data.gov/
Reddit datasets: https://www.reddit.com/r/datasets/top/?sort=top&t=all

2. Identify questions to ask of the data

Asking scientifically and/or methodologically interesting questions on your dataset of interest is the most important decision you need to make. There are many generic questions that can be asked across many different datasets, as well as questions specific to certain datasets.

Examples:

- What baseball player had the most homeruns in his career?
- What gene has the most SNPs associated with disease?
- What words were most commonly used in abstracts in PubMed in 2016?
- What is the most expensive health care cost?

There are a couple ways you can come up with a question:

1. You can pick a dataset that interests you and try to think of interesting questions to ask with it.
2. You can think of an question you're interested in and search for a dataset that allows you to answer it. You might not be able to find the perfect one, but there's a lot of data out there, so it's worth looking!

If you're stuck, look at a few different datasets that sound interesting and try to think of 3 different questions relevant to that topic. Then go back and see if any of them sound interesting enough that you'd like to pursue them for a project!

Document adapted from Dr. Hyun Min Kang, 2017