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

MNIST database of handwritten digits
MNIST database publicly available at http://yann.lecun.com/exdb/mnist/ contains ~60,000 examples of handwritten digits in a 28x28 image.

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.

Single cell digital expression matrix from 10x genomes
Single cell expression data from 10x genomes is available at https://support.10xgenomics.com/single-cell-gene-expression/datasets (requires to enter contact info). It contains a sparse matrix of the read count for each barcoded cell and gene pair for thousands to millions of cells and tens of thousands of genes. A filtered digital gene expression matrix (e.g. http://cf.10xgenomics.com/samples/cell-exp/2.0.1/pbmc4k/pbmc4k_filtered_gene_bc_matrices.tar.gz , accessible after entering the contact info) can be used as a source for analysis, and they can be compared with the default analysis.

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
AWS : 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
Genome-wide association summary statistics:
https://www.ebi.ac.uk/gwas/downloads/summary-statistics

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

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!

3. Come up with a hypothesis for each of your questions

What answers do you expect?

4. Make an analysis plan

How will you read in the data? What commands will you need? What types of plots can you make to display the data? Write pseudocode and draw out examples of plots before you get started coding.

- How will you get the data and read it into Python?
- Specifically, how will you answer each question?
    - What part of the dataset will you need to answer this question?
    - How will you get the information you need out of the dataset?
    - What calculations will you need?
    - What libraries will you need?
    - What commands will you need?
    - This is where pseudocode is important!
- How will you visualize your results? This is one of the most important parts - you want to show other people what you found in a way that makes it easy for them to understand!
    - What is the best way to summarize your results?
    - What kinds of plots will be most informative and easy to understand?
    - How do you actually plot your results in Python?
- Note: Often, answering the question and visualizing your results go hand in hand - to answer the question, you usually need to visualize the results!

5. Start coding!

Use your plan from step 4 to start answering your questions in Jupyter Notebook! Sometimes you'll think of new questions during this process. In that case, you can always go back to previous steps and repeat the process again.

6. Make a presentation of your findings to share with the rest of the group

Document adapted from Dr. Hyun Min Kang, 2017

You should include background information, information about your dataset, your hypotheses, what questions you asked, your results, and an analysis of your results. Did your results match with your hypotheses? Do your results make sense? You can also be creative and include other things, too! What was fun about the project? What was hard? Did you find out anything surprising? The presentation should be about 10 minutes long and will be presented at the graduation ceremony.

GWC Capstone Project Outline

1. **Dataset:** What dataset are you going to use?

2. **Questions:** What questions are you going to ask?

3. **Hypothesis:** What answers do you expect?

4. **Analysis plan:** How are you going to analyze the data to answer your questions? How are you going to visualize your data? Write down ideas and pseudocode here or in your notebook.

Document adapted from Dr. Hyun Min Kang, 2017

5. **Coding:** Use Jupyter Notebook for this part! Feel free to write more pseudocode as you go.

6. **Presentation:** What are you going to include in your presentation? What are the main points you want to get across?

Document adapted from Dr. Hyun Min Kang, 2017