Lecture 11:
Reproducibility of Science and Project Progress

COSC 526: Introduction to Data Mining
Spring 2020

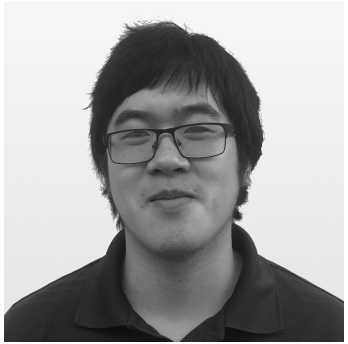
THE UNIVERSITY OF TENNESSEE KNOXVILLE

BIG ORANGE. BIG IDEAS.®

Instructor:

Michela Taufer

GRA:



Nigel Tan

Experts:

Leobardo Valera

Mike Wyatt

Project

Define your project (March 13)

- Which dataset will you be using? How are you obtaining the data? (We will provide you the appropriate data from NHANES or Medicaid, if you choose to use either of these datasets.)
- What is (are) the scientific question(s) that you want to answer? Be as specific as possible.
- What is your strategy to answer the question(s)? Define a set of steps that, if implemented with your code, will allow you to answer the question(s). Be as specific as possible.
- What is the tentative title of your project?

Create a new notebook with your solution (March 27)

- Write down the steps of your solution in distinct text cells; add one or multiple cells (as needed) to hold your code for each step. You can leave these software cells empty for the moment. Expand the text cells describing your solution.
- Add visualization cells that allow you to visualize results. You can leave these software cells empty for the moment.
- Add software to the code cells that upload data from source and pre-process data.
- Push your notebook into your GitHub repository as frequently as needed.

Finalize software and run tests within your notebook (April 3)

- Add the software that implements the method (or methods) to analyze your data.
- Add visualization cells that allow you to visualize results
- Push your notebook into your GitHub repository as frequently as needed.

Build a set of 15 slides that describe your work and get feedback (April 10)

- Build a set of ppt slides (use template provided) that summarize your work; use text slides to tell the story of your project and figures with the key results of your work.
- Make sure your slides include: motivation and problem definition, related work and background, your methodology (e.g., with flowcharts and code sections), your results, summary, and conclusions.

Create your poster and get feedback (April 17)

- Copy and paste your slides into the poster template.
- Shuffle as needed, extend and fill gaps, embellish fonts and text, enlarge text and figures to make them readable.
- **Submit your poster in GitHub for printing (April 20)**
- **Present your poster at the virtual annual EECS COSC 562 poster session (April 23)**

Project Updates

Project updates

- Detecting Trends in Twitter Health News
  - Nasib and Burcum
- Measurement of probable vulnerability of self-harm in different demographic using CDC survey data
  - Mohammad
- A Deeper Look into Scalable Methods for Creating Food Groups Using NHANES Dataset
  - Samuel

Project updates (II)

- Authorship Identification Using N-Gram Feature Classification
  - Austin
- Medicaid Data Set from Delaware
  - Elizabeth and Aileen
- Stock Prediction
  - Abhijeet

Project updates (III)

- Impact of soil moisture in wildfire simulations
  - Kae
- Freesound General-Purpose Audio Tagging
  - Pengxiag and Bohan
- Performance Comparison of Different MPIs using Hatchet
  - Ian

Project updates (IV)

- Neutron Events Detection Using Clustering
  - Rebecca and SuAnn
- Using Different Clustering Methods on Single Cell RNA Sequencing Gene Expression Data to Determine Genes of High Importance For Human and Mouse Cells
  - Angelica

Project updates (V)

- Classification of distribution power system faults
  - Haoyuan
- Accelerating the Dimer Method with Machine Learning Techniques
  - Liubin

Question Queue

- Enter your questions here:

  https://tinyurl.com/s9xeqwh