Foundations of Data Science
Lecture 1, Module 0
Fall 2019

Rumi Chunara, PhD

CS6053

About the Instructor

Rumi Chunara, PhD

Education:
Undergrad: Caltech, Electrical Engineering
Masters: MIT, Electrical Engineering and Computer Science
PhD: MIT, Electrical and Medical Engineering
Postdoc: Harvard Medical School, Computational Epidemiology

Research Interests:
Spatio-temporal statistics
Personally-generated and observational data
Disease surveillance

About the TA

Cheng (Alex) Shi

Alex is a second year CS Master student at NYU Tandon. His areas of interest are data science, visualization, and big data analysis. He is excited to help everyone!

About the Course

Staff Contact:

Instructor: Rumi Chunara, rumi.chunara@nyu.edu

Office hours: TH 10:00-11:00am, or after class

TA:

Alex Shi: TBD in 370 Jay St., TA area cs5615@nyu.edu

Use Class forum on NYU Classes for questions…

Goals of The Course

- Understand what Data Science is
- Approach problems from a data-driven, critical problem-based perspective
- Hands-on experience mining and analyzing data

Coding!

In order to succeed and participate in this class,
You will…

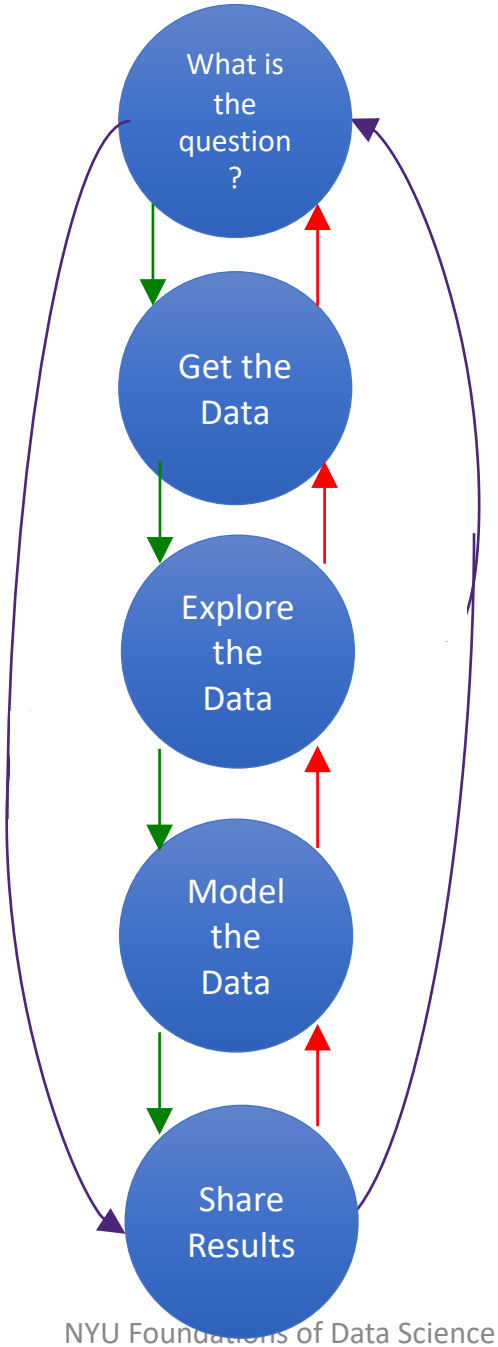
- Need access to a computer with admin privileges. It will be most helpful to bring this to class.
- Have to learn and use the Python programming language.

Please see me after class if this is an issue.

Course Outline

This order of topics in this course will follow the same order as a typical data science project.

We'll also peel the layers of data science like an onion, so the flow might not always be linear.



Suggested Readings

- Textbook: James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning with applications in R.* New York: Springer (available online for free)
- Python Data Science Handbook (good Python reference) by Jake VanderPlas, O'reilly (2017).
- Pratical Statistics for Data Scientists (good statistics reference) by Bruce and Bruce, O'reilly (2017).

Grading Breakdown

Grading

- In-class Quizzes (top 5) 25%
- Assignments (5) 30%
- Final Project 35%
- Participation 10%

Assignments

- Assignments will give you hands-on experience with data-driven problems and solutions. We'll introduce them in class and discuss what's expected
- Coding and open questions
- Examples in class will help
- Are all expected to be performed solo
  - Discussions are welcome and encouraged on the course forum
  - DON'T COPY CODE (from the web or other people). There are tools to detect this.
- All coding assignments must be completed in Python. Implement and comment your code so that anyone reading the file can reproduce the code easily
- Save the code as a notebook, and upload it to NYU classes.

Quizzes

- The idea of quizzes is to review material, and assess what topics the class may need more review on
- Dates not provided in advance

Project

- The final project will pull together all of the elements you learn from this class and will simulate the experience of being a professional data scientist.
- The idea is we want you to be able to
  - Identify a problem
  - Motivate and implement a solution including data and methods
  - Communicate and evaluate the solution
- The project will be evaluated via a proposal, short presentation and final report

Projects

Project teams should form and be reported by 9/19

Project proposals due by ~10/22

Project presentations prior to due date to explain proposed approach and get feedback 11/19 or 11/26

Final Report 12/10/2019

You can choose a project topic, but we will also provide a list of suggested projects from around campus

You need:

- A clear problem statement (motivated by literature, personal contact, other resource)
- An accessible dataset
- Modeling plan + appropriate tools

Project Milestones

1. Choose a group (2)
2. Pick a dataset and a problem, write a proposal
3. Explore and validate the utility of the data
4. Present to the class
5. Write a professional technical report

Project Proposal

- What is the problem?
- How will you learn the background?
- What kinds of data will you use?
- Almost anything is OK, except other predictions.
- Numerical or text?
- What kind of model will you build?
- What assumptions are safe to make?
- Proposal should be ~1000 words + figures
- Use the Data Science terminology we learn in class

Grading Scale

*This scale may change but I won't curve down

A: 90-100

B: 75-90

C: 60-75

D: <60

(lowest number in each range is inclusive)