DATA SCIENCE

CSC 495/663

THE UNIVERSITY *of* NORTH CAROLINA
GREENSBORO

COURSE INFO

- **Course: Data Science**
  - CSC 495/663
  - Monday and Wednesday 3:30 pm - 4:45 pm
  - Prerequisite:
    - CSC 339 (Programming Languages) OR Programming experience (Instructor Permission Required)
    - *Mostly programming experience*
- **Instructor: Dr. Somya Mohanty**
  - Office: Petty 152
  - Office Hours: Monday and Wednesday 2:00 pm - 3:00 pm
    - Or by appointment
  - Email: mohanty.somya@uncg.edu

COURSE INFO

- **Experience in:**
  - **Programming skills Python**
    - *If not there, you would have to work hard in the early weeks to get there*
  - Linux
  - Terminal, Command-Line

- **Books:**
  - Nothing is required
  - Recommended
    - Building Machine Learning Systems with Python (Richert and Coelho)
    - Python for Data Analysis: Data Wrangling with Pandas, NumPy, and Ipython (Wes McKinney)

COURSE INFO

What is the course about?

- Programming your way into Data Science
- Some theory, mostly programming
- It is not a statistics or an AI or a visualization course
- The course contains parts of everything
- Learn about lot of tools and how to use them in innovative ways
- We will work with real-world data
- Hopefully develop some cool projects

COURSE INFO

- **Grading**
  - **Class Participation: 5%**
  - **Class / Homework Assignments (4-6): 30%**
  - **Final Project: 65%**
    - Project Progress Presentation and Report: 15%
    - Project Final Submission Code: 20%
    - Project Class Presentation: 15%
      - Graduate Students
        - (Paper and Presentation) – 7.5% (Presentation) and 7.5% (Paper)
      - Undergraduates – 15%
- **No Exams**

COURSE INFO

- **Grading**
  - **Class Participation**
    - Most of the activities in class are interactive
    - Asking questions and participating in discussion gets you bonus points!
    - Show off your programming skills by finding better approaches.
  - **Homework Assignments (4-6):**
    - Utilization of tools learned in class
    - Mostly programming and data analysis
    - The submission will be on IPython notebooks
    - Utilize Github / Bitbucket for assignments (own account)
    - Link to the assignment submission via email for submission

COURSE INFO

- **Grading**
  - **Final Project:**
    - Most of the grade is based on the Final Project
    - Novel ideas have 5% bonus
    - Project progress
      - 3-5 min presentation and 1 page report
      - End of each course topic (discussion later)
      - Utilization of git is important, will look at the commit logs of every member (hosted by the team)
    - Final Presentation on Completion
      - 20 min presentation to class
      - Poster presentation to Department
        - Data, Methods, Novelty, Visualization
      - Graduate Students
        - Paper (5 pages minimum) – IEEE/ACM Standard

COURSE INFO

A+	100%	to	99%
A	< 99%	to	94%
A-	< 94%	to	90%
B+	< 90%	to	87%
B	< 87%	to	84%
B-	< 84%	to	80%
C+	< 80%	to	77%
C	< 77%	to	74%
C-	< 74%	to	70%
D+	< 70%	to	67%
D	< 67%	to	64%
D-	< 64%	to	60%
F	< 60%	to	59%

COURSE INFO – TIMELINE (TENTATIVE)

- **Introduction to Data Science: (Week 1)**
  - Class Syllabus and Introduction
  - Class Project discussion and assignment

- **Data Resources and Mining: (Weeks 2-3)**
  - Programming
    - Re/Introduction to Python
    - IPython, IPython-Notebook
    - Git
  - Resources
    - Web-Resources, Data-Repositories
    - API's – Streaming and REST

COURSE INFO – TIMELINE (TENTATIVE)

- **Data Munging, Wrangling, Cleaning (Week 4)**
  - Pandas
  - NumPy
  - *Project Review*
- **Efficient Programming and Storage (Week 5-6)**
  - Python HPC
    - Parallel Programming, Multi-Processing, IPython Cluster
  - Storage
    - Data Structures (Hashes, JSON, XML, Lists)
    - Relational vs Non-relational Databases, NoSQL
    - Query optimization and Indexing
    - Utilizing Cloud Resources (Amazon EC2 and S3)
  - *Project Review*

COURSE INFO – TIMELINE (TENTATIVE)

- **Data and Statistics (Week 7-8)**
  - Statistical Hypothesis Testing
  - Bootstrapping
  - Correlation
  - Regression
  - Bayesian
  - Distribution
  - *Project Review*

COURSE INFO – TIMELINE (TENTATIVE)

- **Data Analytics and Visualization: (Weeks 9-11) ***
  - Graph Generation and Tools
    - MatplotLib
    - Plotly
    - Pandas
    - Bokeh
  - Spatial and Temporal Analysis:
    - Google Maps
    - Basemap
    - CartoDB
  - Network Analysis:
    - NetworkX
    - Gephi, CytoScape
  - *Project Review*

COURSE INFO – TIMELINE (TENTATIVE)

- **Introduction to Applied Machine Learning: (Weeks 12-13) ***
  - Clustering, Topic Modeling, Classification, Regression, Feature Selection and Dimensionality Reduction
  - Python Libraries for Machine Learning: Sk-Learn, Scikit, Sci-Py, Gensim
  - NLP, Text Processing and Feature Extraction: Review of NLTK, Gensim
  - *Project Review*
- **Privacy and Ethics in Data Science: (Week 14)**
  - If possible
- **Project Presentations: (Week 15)**
  - Class Presentations
  - Project Paper (Graduate students only)

COURSE INFO – DISCLAIMER

- **The course is going to be tough, especially for people with limited programming experience**
  - Work hard, be rewarded with a good data science experience
  - Will talk about the benefits later in course intro
- **Do not cheat in the course**
  - I will run the code through plagiarism detection software- single incident reporting to honor committee
  - In team project, do not think that you can get away without contributing.
    - I will be monitoring repositories for work done
    - Any work done should be reported on the repository – I worked locally on my computer will not count.

COURSE INFO – DISCLAIMER

- **Utilization of resources found on the Internet is allowed for project accomplishment, with caveats**
  - Any code/library used should be referenced and thoroughly understood
  - If you use code without understanding, that counts as plagiarism
- **On team projects**
  - The team creation can be random or self-assigned, we will discuss it
  - In the project review presentations, two members must present each time. All members must know their counterparts work.
  - Class is encouraged to participate and discuss/ask questions

COURSE INFO – DISCLAIMER

- **On team projects**
  - You will get critical comments from me both on presentation and project progress
  - If you are not able to take critical comments on the progress, this course is not suitable for you.
  - You will be presenting at the end to the department and external attendees.
    - We are trying to achieve a good presentation of your project for that.

QUESTIONS