

# DATA SCIENCE

**CSC 405/605**



THE UNIVERSITY *of* NORTH CAROLINA  
**GREENSBORO**

# COURSE INFO

- **Course: Data Science**
  - CSC 405/605
  - Tuesday and Thursday 12:30 pm - 1:45 pm
  - Prerequisite:
    - CSC 339 (Programming Languages) OR Programming experience (Instructor Permission Required)
    - *Mostly programming experience*
- **Instructor: Dr. Somya Mohanty**
  - Office: Petty 152
  - Office Hours: Tuesday and Thursday 10:00 am - 11:00 am
    - Email for appointment
  - Email: [mohanty.somya@uncg.edu](mailto:mohanty.somya@uncg.edu)



# COURSE INFO

**What is the course about?**

- **Programming your way into Data Science**
- **Theory - 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

- **Experience in:**
  - **Programming skills Python**
    - *We will go through Introduction to Python*
      - *You would have to work hard in the early weeks to get comfortable with Python*
  - 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

- **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 Final Presentation: 15%
      - Graduate Students
        - (Paper and Presentation) –
          - 5% (Presentation)
          - 10% (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 for assignments (own account)
    - Link to the assignment submission via Canvas for submission



# COURSE INFO

- Grading

- **Final Project:**

- Most of the grade is based on the Final Project
    - ***Take a look at the syllabus***
    - Project progress
      - 3 min presentation by each member 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
        - Data, Methods, Novelty, Visualization
        - Everyone is invited – Department, Mentors, Other students
      - Graduate Students
        - Paper (6 pages minimum) – IEEE/ACM Standard



# COURSE INFO

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





# COURSE INFO – TIMELINE (TENTATIVE)

- **Introduction to Data Science: (Week 1)**
  - Class Syllabus, Grading, Expectations, and Getting to know each other.
  - Introduction to Data Science.
- **Startup Tools and Programming (Weeks 2-3)**
  - Class Project discussion and assignment
  - Programming
    - Git (Github)
    - Re/Introduction to Python
    - IPython, IPython-Notebook
- **Data Munging, Wrangling, Cleaning (Week 3-4)**
  - Pandas
  - NumPy
  - ***Project Review – 1 (Understand the data and project)***



# COURSE INFO – TIMELINE (TENTATIVE)

- **Data and Statistics (Week 5-6)**
  - Basics
  - Distributions and Point Estimates
  - Statistical Hypothesis Testing
  - Correlation
  - Regression
  - ***Project Review – 2 (Data Statistics)***



# COURSE INFO – TIMELINE (TENTATIVE)

- **Introduction to Applied Machine Learning: (Weeks 8-11)**
  - Overview
  - Python Libraries for Machine Learning: Sk-Learn, Scikit, Sci-Py, Gensim
  - Basic Principles
  - Classification
  - Validation
  - Dimensionality
  - Clustering
  - Time-Series
  - End-to-end Machine Learning
  - Clustering, Topic Modeling, Classification, Regression, Feature Selection and Dimensionality Reduction
  - NLP, Text Processing and Feature Extraction: Review of NLTK, Gensim
  - ***Project Review – 3 (Learning from Data)***



# COURSE INFO – TIMELINE (TENTATIVE)

- **Data Analytics and Visualization: (Weeks 12-14) \***
  - 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)

- **Project Presentations: (Week 15)**
  - *Class Presentations*
  - Project Paper (Graduate students only)



# COURSE INFO – DISCLAIMER

- Read the syllabus.
- Take regular notes.
- **Class is encouraged to participate and discuss/ask questions – Class Participation Points!**
- **On team projects**
  - Start early
  - 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.
- **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



# COURSE INFO – DISCLAIMER

- **Do not cheat in the course – Result will be an ‘F’ grade.**
  - Assignment solutions are unique, differs from student to student. No collaboration on Assignments whatsoever.
  - 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 – *worked locally on my computer will not count.*
- **Utilization of resources found on the Internet is allowed for project accomplishment, with caveats**
  - Any code/library used should be referenced/cited and thoroughly understood
  - If you use code without understanding, that counts as plagiarism



# COURSE INFO – DISCLAIMER

- **More on team projects**
  - You will get critical comments from me, both on presentation and project progress
    - Its geared towards making your projects awesome!
  - You will be presenting at the end to the department and external attendees.
    - We are trying to achieve a great presentation *made by you for your project.*
- ***Use office hours – I am here to help you through your project.***





# QUESTIONS

