

Fall 2020

CS 301-101: Introduction to Data Science

Connor Watson

Follow this and additional works at: <https://digitalcommons.njit.edu/cs-syllabi>

Recommended Citation

Watson, Connor, "CS 301-101: Introduction to Data Science" (2020). *Computer Science Syllabi*. 156.
<https://digitalcommons.njit.edu/cs-syllabi/156>

This Syllabus is brought to you for free and open access by the NJIT Syllabi at Digital Commons @ NJIT. It has been accepted for inclusion in Computer Science Syllabi by an authorized administrator of Digital Commons @ NJIT. For more information, please contact digitalcommons@njit.edu.

CS 301- Introduction to Data Science

Fall 2020 Course Syllabus

Instructor: Connor Watson

Email: cww5@njit.edu

Office Hours: Tuesdays 6-7:30 PM, or by appointment (request via email / Canvas)

Grader: N/A

Email: N/A

Office Hours: N/A

Course Description: This course will introduce data science by touching many areas within the field so that students can understand what kind of problems exist and why a data science solution may be relevant to those problems. In the first part of the course, students will understand “{who uses / what is / when to use / why / how to apply} data science”, and learn the tools used during typical data science workflows. This includes a review of some underlying mathematics concepts, an exploration of data, preprocessing data for most workflows, and an introduction to popular tools used in both research and industry. The second part of the course focuses on the algorithms and models involved with machine learning and analytics, including the theory and applications of such topics. Students will gain exposure to different data sets and will learn to create data description documents.

Primary Course Objectives:

- Conceptualization, visualization, and implementation of data science algorithms.
- Learn different sub-topics of data science (e.g., big data, machine learning).
- Work with existing data science tools (e.g., numpy, scikit-learn, matplotlib).
- Develop problem solving skills for data science objectives.
- Feel comfortable with structured data, and learn how to approach unstructured data.
- Create end to end solutions from preprocessing to analysis and prediction.

Prerequisite Courses: CS 114, Math 333, (Math 337 is helpful but not required)

Course Resources:

- Datasets and competitions: <https://www.kaggle.com/>
- Open source Python packages: <https://www.anaconda.com/distribution/>

Primary Textbooks (available with free public license online):

- Think Stats (2nd ed) by Allen B. Downey
 - <https://greenteapress.com/wp/think-stats-2e/>
- Python Data Science Handbook by Jake VanderPlas
 - <https://jakevdp.github.io/PythonDataScienceHandbook/>

Recommended Supplemental Books (non-mandatory):

The following list is recommended for students if they would like to gain additional knowledge:

- Applied Statistics for Engineers and Scientists (3rd ed 14) Devore & Farnum
- Introduction to Data Science: A Python Approach to Concepts, Techniques and Applications Laura Igual, et. al
- Think Bayes by Allen B. Downey
 - <https://greenteapress.com/wp/think-bayes/>
- Introduction to Probability by Charles M. Grinstead and J. Laurie Snell
 - http://www.dartmouth.edu/~chance/teaching_aids/books_articles/probability_book/amsbook.mac.pdf
- Designing Data-Intensive Applications by Martin Kleppmann
- Data Mining: Practical Machine Learning Tools and Techniques Frank & Witten 4th ed
- Deep Learning with Python (1st ed) by François Chollet
- Deep Learning by Ian Goodfellow et al
 - <http://www.deeplearningbook.org/>

Course Content:

There will be several aspects to this course:

1. Instructor/guest lectures
2. Individual homework assignments
4. A semester-long project will be completed in teams and presented at the end of term
5. Midterm/final exams
6. Class participation and attendance
7. Weekly lab assignments

Class Etiquette:

Please silence all technologies during class time and during exams. Students may receive positive points for their participation, or negative points for their lack of participation. Students are also expected to be prepared to take notes, either with pen/pencil and paper or electronically.

For online class, please stay muted unless otherwise called on. Within Webex, there is a hand raise feature; please utilize it to ask a question, or chat a question. In addition, please be sure to remove inappropriate / offensive imagery from the view in which your camera captures.

All assignments are to be submitted to Canvas by the date due and time (EST), unless otherwise specified by the instructor. If a physical (paper, USB, etc.) submission is needed, make sure that your name is on each sheet/USB that is handed in. If a USB copy is needed, please submit a README file with all members' names, as well as a summary of the files contained in the USB.

The use of offensive language during the course of the semester is strictly prohibited, through all channels of communication in person and online. Offensive language includes (but is not

limited to) the use of words/phrases which may target (in or out of context) people of different race, ethnic background, mental or physical condition, social or economic status, etc. Anyone observed using the use of any such insults during the course of the semester is subject to penalties. Depending on the occurrence of, or the intensity of the language used, the instructor holds the right to report the person to the Dean of Students Office (hereby referred to as DSO) for further investigation.

Attendance:

It is mandatory to come to class. Attendance will be counted via Webex attendance. If a student should miss a class, they must receive proper documentation through the DSO. Only then will an absence be excused.

(Tentative) Course Schedule:

The course will approximately follow the guide below:

Number	Topic
1	Introduction to Data Science, Math Review, NumPy
2	Data Collection, Querying Data, Pandas
3	Data Wrangling, Anomaly Detection, Outlier Analysis
4	Understanding Data Models, Visualization
5	Dimensionality Reduction, Correlation
6	Big Data, Distributed (Cloud) Computing
7	Midterm Exam: Topics 1-6
8	Supervised Learning
9	Unsupervised Learning
10	Assessing Predictive Models
11	Methods in Artificial Intelligence: Neural Networks
12	Applied / Advanced Topics in Data Science Workflows
13	(Thanksgiving Recess - No Class - Continue Work on Projects)
14	Student Project Presentations
15	Final Exam: All Topics

Grading Scale:

A: 100%-90%	B+: 89%-85%	B: 84%-80%	C+: 79%-75%
C: 74%-70%	D: 69%-60%	F: 59%-0%	

Students' final grades are based on the above tentative grading scale. We reserve the right to change the scale only on an as needed basis. Below is the breakdown of how the final average is calculated:

Grading Breakdown:

10% Attendance	20% Midterm Exam	25% Final Exam
20% Homework Assignments	25% Project	

A submission is considered late if it is submitted after the due date/time. Submissions can be submitted up to 48 hours after the due date for a maximum of 50% credit, and submissions made after that will not be counted.

We reserve the right to make class-wide adjustments to grade weights, or to add/subtract assignments as the need arises.

Labs:

In addition to lecture, students will have weekly lab assignments (not counted in the grading breakdown). The purpose of the labs in substitute for an additional lecture, is to connect the concepts learned in class to a practical assignment. Students will gain exposure to popular tools in both industry and academic data workflows. Each lab will be provided electronically in addition to a video which will explain the solutions found in the lab. It is encouraged that students attempt the lab first without looking at the solutions found in the video.

Course Project:

Students will have the opportunity to work on a project throughout the course of the semester. The project have five total deliverables:

1. Outline for Project / Project Scope / Identify Research Questions
2. Data Manipulation / Data Statistics / Descriptions
3. Data Visualization
4. Final Documentation
5. Presentation

Students will work in teams with an opportunity to pick preferences on a project topic and group members. Ultimately the assignment of groups is at the discretion of the instructor. The goal of

the project is to provide an active learning experience while learning the weekly topics. This project will be a good opportunity to apply the topics learned on a larger scale with team mates, while gaining exposure to a typical project workflow. By the end of the semester, students will have a large document cumulative of each deliverable, and a presentation.

Version Control:

Students may gain exposure to Version Control through tools like Git and GitHub Desktop. By maintaining code through this collaborative software, students can keep their code available and transparent with their teammates and the instructor.

Exams:

Online exams will be conducted via Canvas and Respondus Lockdown Browser. This requires the use of a webcam. For more information, please visit the following web page:

<https://ist.njit.edu/respondus>

Students with special needs are advised to make arrangements with the Office of Accessibility Resources and Services (hereby referred to as OARS).

<https://www.njit.edu/studentsuccess/accessibility/>

If allowed by the instructor, students may be allowed a non-graphing calculator, and/or a one-page cheat sheet (8.5"x11", one-sided), and/or a formula sheet for an exam. They must be brought in on the day of the exam so that the instructor may verify the calculator and cheat sheet matches the above criteria.

If it is found that the student is suspected of cheating during the exam, the instructor reserves the right to rescind the students' exam and nullify the grade, and/or submit to the DSO.

Cheating is defined as using unauthorized means to gain knowledge during the exam including (but is not limited to) having more than one page cheat sheet, using a prohibited technology, looking at another student's paper, talking during the exam, searching for additional information outside of the classroom, etc.

No extra Internet access / other technologies are allowed during the exam, unless the student is granted permission from the OARS, or unless specified by the instructor.

In the case where a student may miss an exam due to a well documented reason, with approval by the DSO, they may receive a make-up exam which may or may not have the same content as the exam the other students receive.

Honor and Policy:

In the exam, each student is required to sign the Honor Code Agreement "On my honor, I pledge that I have not violated the provision of the NJIT Student Honor Code."

If you believe that you deserve more credit than you have been awarded on a particular exam, you may request (within 48 hours of the exam being returned) that it be re-graded. Your entire exam will be regraded, which may result in points being added or subtracted.

Students found cheating or plagiarizing will be immediately referred to the DSO and the NJIT Committee on Professional Conduct and subject to Disciplinary Probation, and will receive a permanent mark on their record, possible dismissal, and an "F" grade in the course. All submitted assignments will be checked for similarities, and plagiarism / guilty students will be identified and submitted to the above offices.

You should read the University Code on Academic Integrity:

<https://www.njit.edu/policies/sites/policies/files/academic-integrity-code.pdf>.

It describes infractions of academic integrity and penalties for violations, including, for the most serious violations, an XF grade in the course or expulsion. All work that you represent as your own must, in fact, be your own. Work done by others must be given proper credit.