CS/STAT 242 — Introduction to Data Science

Spring 2020

Lectures: Monday & Wednesday / 11:30AM - 12:20PM

Labs Wednesday (HAAS 257) or Thursday (LWSN B146) / 3:30PM - 5:20PM / REC 313

Credit Hours: 3

Instructor: Vinayak Rao, PhD
Office: MATH 212
Office Hours: M 2:00 - 3:00 PM
E-mail: varao@purdue.edu

Teaching Assistants Yuanyang Shao (shao130@purdue.edu)
Haoyun Yin Shao (yin164@purdue.edu)
Office Hours: TBD

Course Website: Blackboard (mycourses.purdue.edu)

Piazza: (piazza.com/purdue/spring2020/cs24200)

Textbooks (available online via Purdue Libraries):

Bruce, P. and Bruce, A. (2017). *Practical statistics for data scientists: 50 essential concepts.*. O'Reilly Media.

Cairo A. (2016). *The truthful art : Data, charts, and maps for communication.*. New Riders.

Grus, J. (2016). *Data science from scratch.*. O'Reilly Media.

Kabaco , R. (2011). *R in action: Data analysis and graphics with R.*. Pearson Education.

Provost, F. and Fawcett, T. (2013). *Data science for Business.*. O'Reilly Media.

Schutt, R. and O'Neil, C. (2014). *Doing Data science.*. O'Reilly Media.

Course Description: This course provides a broad introduction to the field of data science. The course focuses on using computational methods and statistical techniques to analyze massive amounts of data and to extract knowledge. It provides an overview of foundational computational and statistical tools for data acquisition and cleaning, data manipulation, data analysis and evaluation, visualization and communication of results, data management and big data systems. The course surveys the complete data science process from data to knowledge and gives students hands-on experience with tools and methods.

Learning Outcomes: Upon completing the course, students should be able to:

- Use Python, R, and selected tools to scrape, clean, process, and visualize data
- Apply data management techniques to prepare, parse, manipulate, and store data

- Use statistical methods to summarize data and identify relationships
- Explain how to formulate new hypotheses and draw accurate conclusions from data
- Apply statistics and computational analysis to make predictions based on data
- Apply basic computer science concepts such as modularity, abstraction, and encapsulation to data analysis problems
- Effectively communicate the outcome of data analysis using descriptive statistics and visualizations

Prerequisites: CS18000, CS18200, CS38003 (grade of C or better)

Course Outline:

Introduction (1 week) What is Data Science? Examples, applications, and results obtained using data science techniques. Overview of the data science process.

Background and basics (2.5 weeks) Review of Python. Using Python notebooks. Types of data and data representations. How to acquire data (e.g., crawling), how to process and parse data. Data manipulation, data wrangling, and data cleaning.

Visualization and basic statistics (2.5 weeks) Introduction to R. Visualization principles and goals. Basic plots in R. The importance of communicating results. Visualizing distributions and relationships.

Hypothesis testing and causality (2.5 weeks) Introduction to statistical inference, populations/samples. Overview of hypothesis testing, A/B testing, and how to draw conclusions from data. Correlation vs causation.

Similarity and clustering (2 weeks) Definitions and examples of common similarity/distance measures. Overview of basic clustering methods and how to interpret/evaluate results. Dimensionality reduction.

Large scale analysis (1 week) Data engineering overview. Discussion of databases and SQL, mapReduce processing, Spark, and Hadoop.

Collaborative filtering (1.5 weeks) Recommender systems and collaborative filtering, including basic methods and applications.

Ethics (1 week) Overview of ethical issues of privacy, fairness, and bias in data science.

Course Work and Requirements:

	Percentage of Grade
Weekly labs:	20%
Homework:	15%
Programming projects:	20%
Midterm:	20%
Final exam:	25%

Assignments and exams There will be short exercises to be completed weekly in each lab. There will be five homework assignments and three projects that will be posted

on the schedule. Assignments should be submitted online in Blackboard or via turnin on data.cs.purdue.edu. Details will be provided in the assignments. Programming projects should written in Python 3, unless otherwise noted. In general, questions about the details of homeworks/projects should be directed to the TA on Piazza.

There will be a midterm and comprehensive final exam. Exams will be closed book and closed notes.

Academic honesty Please read the departmental academic integrity policy. This will be followed unless we provide written documentation of exceptions. We encourage you to interact amongst yourselves: you may discuss and obtain help with basic concepts covered in lectures or the textbook, homework specification (but not solution), and program implementation (but not design). However, unless otherwise noted, work turned in should reflect your own efforts and knowledge. Sharing or copying solutions is unacceptable and could result in failure. We use copy detection software, so do not copy code and make changes (either from the Web or from other students). You are expected to take reasonable precautions to prevent others from using your work.

Please note the Purdue Honors Pledge: "As a Boilermaker pursuing academic excellence, I pledge to be honest and true in all that I do. Accountable together - we are Purdue."

Course Policies

Addressing Questions via E-mail: Your first resource for asking questions should be the class Piazza page, where you are likely to get the fastest response. Please feel free to e-mail questions to the instructor or TAs. Reserve those that involve extensive computation or mathematical expressions for office hours. If your question involves programming, please be sure to e-mail a minimal working example of your code to the instructors.

Incompletes: Incompletes will only be given under emergency circumstances, e.g., a serious auto accident, death of family member, etc. (see the grief absence policy for further information). Incompletes will not be given to students failing the course.

Grading: The grading scale is predetermined so as to eliminate competition with other students, and to ensure that you always know your grade in the class. Your grade is based upon *your* performance only. Grades will not be curved.

Grade	Numerical range
A	90.00-100.00
B	80.00-89.99
C	70.00-79.99
D	60.00-69.99
F	0.00-59.99

Re-grading: All grade disputes are to be made on paper, and submitted *directly* to Dr. Rao. Discussions or arguments for re-grades will *not* be done in person. A student has until one week after receiving his/her grade to dispute the grade (in writing). Handling re-grades in this manner eliminates the "end of the semester" digging for points.

When disputing a grade, you should state the question, the dispute, and the number of points you feel you should have received for the question. If you do not state the number of points you think are reasonable for the re-grade, zero points will be given as the re-grade. Please note that when you ask for a question to be re-graded, the entire assignment may be re-graded, and there is a possibility of losing points.

Dropping the Course: The instructors reserve the right to *not* sign anyone out of the course once the deadline for dropping without the instructors signature has passed. Please take care to pay attention to these dates.

Attendance and Participation:

Students: You are expected to attend lectures. You are expected to arrive on time, or before. You are expected to stay until the end of lecture unless you have asked in advance to leave early. You are expected to be prepared and participate. On the rare occasion that a student is extremely close to the cut-off value between letter grades, attendance and class participation may help.

When conflicts or absences can be anticipated, such as for many University sponsored activities and religious observations, you should inform the instructors of the situation as far in advance as possible. For unanticipated or emergency absences when advance notification is not possible, you should contact the instructors as soon as possible by e-mail, or the Department of Statistics main office. When you are unable to make direct contact with the instructors and unable to leave word with the Department of Statistics because of circumstances beyond your control, and in cases of bereavement, you or your representative should contact the Office of the Dean of Students. The instructors will try to accommodate you either by excusing you or by allowing you an extension when possible. Ultimately, you are responsible for all required coursework and bear full responsibility for any academic consequences that may result due to your absence.

Links to the complete attendance policy and implications can be found at

www.purdue.edu/advocacy/students/absences.html and
www.purdue.edu/studentregulations/regulations_procedures/classes.html.

Instructors: You can expect that we will attend lectures. We will arrive in the lecture room prior to the start of lecture, and will end lecture on time. You can expect that we will be prepared for lecture, try our best to convey the information for the course, and show respect for all students.

If we are unable to attend lecture you will know in advance, and we will either cancel class or provide a guest instructor. We will be present for our office hours, and available for scheduled appointments.

The amount of material covered in each lecture is governed by the speed with which we complete the material. Every group of students is different, and we would rather teach the material well (and have you learn it) than speed through the topics for the purpose of covering a preset number of topics. Accordingly, the course outline is subject to change as the course progresses.

Grief Absence Policy for Students: Purdue University recognizes that a time of bereavement is very difficult for a student. The University therefore provides the follow-

ing rights to students facing the loss of a family member through the Grief Absence Policy for Students (GAPS). Students will be excused for funeral leave and given the opportunity to earn equivalent credit and to demonstrate evidence of meeting the learning outcomes for missed assignments or assessments in the event of the death of a member of the student's family.

Counseling and Psychological Services Information: Purdue University is committed to advancing the mental health and well-being of its students. If you find yourself beginning to feel some stress, anxiety, and/or feeling slightly overwhelmed, try Well-Track. Sign in and find information and tools at your fingertips, available to you at any time. If you need support and information about options and resources, please see the Office of the Dean of Students for drop-in hours (Monday - Friday, 8:00 AM - 5:00 PM). If you or someone you know is feeling depressed and/or in need of mental health support, services are available. For help, such individuals should contact Counseling and Psychological Services (CAPS) at (765) 494-6995, and `www.purdue.edu/caps` during and after hours, on weekends and holidays, or by going to the CAPS office on the second floor of the Purdue University Student Health Center (PUSH) during business hours.

University Emergency Information: A safety briefing will be conducted on the first day of class. In the event of a major campus emergency or temporary suspension of classes, course requirements, deadlines and grading percentages are subject to changes that may be necessitated by a revised semester calendar or other circumstances beyond the instructors' control. You can get information about changes in this course by means of the course web page, or contacting the instructors via e-mail or phone. You are expected to read your Purdue e-mail on a frequent basis.

Violent Behavior Policy: Purdue University is committed to providing a safe and secure campus environment for members of the university community. Purdue strives to create an educational environment for students and a work environment for employees that promote educational and career goals. Violent behavior impedes such goals. Therefore, violent behavior is prohibited in or on any University Facility or while participating in any university activity.

Academic Dishonesty: Purdue prohibits "dishonesty in connection with any University activity. Cheating, plagiarism, or knowingly furnishing false information to the University are examples of dishonesty." [Part 5, Section III-B-2-a, University Regulations] Furthermore, the University Senate has stipulated that "the commitment of acts of cheating, lying, and deceit in any of their diverse forms (such as the use of substitutes for taking examinations, the use of illegal cribs, plagiarism, and copying during examinations) is dishonest and must not be tolerated. Moreover, knowingly to aid and abet, directly or indirectly, other parties in committing dishonest acts is in itself dishonest." [University Senate Document 72-18, December 15, 1972].

Academic integrity is one of the highest values that Purdue University holds. You are encouraged to alert university officials to potential breeches of this value by either e-mailing `integrity@purdue.edu`, calling 765-494-8778, or contacting the Office of the Dean of Students (`www.purdue.edu/odos`). While information may be submitted anonymously, the more information that is submitted provides the greatest opportunity for the university to

investigate the concern. Bonus points will be given to students who report instances of cheating.

Incidents of academic misconduct in this course will be addressed by the course instructor and referred to the Office of Student Rights and Responsibilities (OSRR, www.purdue.edu/odos/osrr) for review at the university level. Any violation of course policies as it relates to academic integrity will result minimally in a failing or zero grade for that particular assignment, and at the instructor's discretion may result in a failing grade for the course. In addition, all incidents of academic misconduct will be forwarded to OSRR, where university penalties, including removal from the university, may be considered. Use of instructor solution manuals or related resources will not be tolerated.

Use of Copyrighted Materials: Among the materials that may be protected by copyright law are the lectures, notes, and other material presented in class or as part of the course. Always assume the materials presented by the instructors are protected by copyright unless the instructors have stated otherwise. Students enrolled in, and authorized visitors to, Purdue University courses are permitted to take notes, which they may use for individual/group study or for other non-commercial purposes reasonably arising from enrollment in the course or the University generally.

Notes taken in class are, however, generally considered to be "derivative works" of the instructors' presentations and materials, and they are thus subject to the instructors' copyright in such presentations and materials. No individual is permitted to sell or otherwise barter notes, either to other students or to any commercial concern, for a course without the express written permission of the course instructor. To obtain permission to sell or barter notes, the individual wishing to sell or barter the notes must be registered in the course or must be an approved visitor to the class. Course instructors may choose to grant or not grant such permission at their own discretion, and may require a review of the notes prior to their being sold or bartered. If they do grant such permission, they may revoke it at any time, if they so choose.

Students with Disabilities: Purdue University is required to respond to the needs of the students with disabilities as outlined in both the Rehabilitation Act of 1973 and the Americans with Disabilities Act of 1990 through the provision of auxiliary aids and services that allow a student with a disability to fully access and participate in the programs, services, and activities at Purdue University.

Purdue University strives to make learning experiences as accessible as possible. If you anticipate or experience physical or academic behaviors based on disability, you are welcome to let the instructors know so that they can discuss options. You are also encouraged to contact the disability resource center at drc@purdue.edu or by phone 765-494-1247. If you have a disability that requires special academic accommodation, please make an appointment to speak with the instructors within the first three (3) weeks of the semester in order to discuss any adjustments. It is important to talk about this at the beginning of the semester. It is the student's responsibility to notify the Disability Resource Center (www.purdue.edu/drc) of an impairment/condition that may require accommodations and/or classroom modifications.

Nondiscrimination: Purdue University is committed to maintaining a community that

recognizes and values the inherent worth and dignity of every person; fosters tolerance, sensitivity, understanding, and mutual respect among its members; and encourages each individual to strive to reach his or her own potential. In pursuit of its goal of academic excellence, the University seeks to develop and nurture diversity. The University believes that diversity among its many members strengthens the institution, stimulates creativity, promotes the exchange of ideas, and enriches campus life. Purdue University's nondiscrimination policy can be found at www.purdue.edu/purdue/ea_eou_statement.php.

Purdue University prohibits discrimination against any member of the University community on the basis of race, religion, color, sex, age, national origin or ancestry, genetic information, marital status, parental status, sexual orientation, gender identity and expression, disability, or status as a veteran. The University will conduct its programs, services and activities consistent with applicable federal, state and local laws, regulations and orders and in conformance with the procedures and limitations as set forth in Executive Memorandum No. D-1, which provides specific contractual rights and remedies. Any student who believes they have been discriminated against may visit www.purdue.edu/report-hate to submit a complaint to the Office of Institutional Equity. Information may be reported anonymously.