

CS 0155: Data Witchcraft – Spring 2017 (tentative)

Department of Computer Science, University of Pittsburgh

Instructor: Prof. Alexandros Labrinidis

Email: labrinid@cs.pitt.edu

Web: <http://labrinidis.cs.pitt.edu>

Office: 6105 Sennott Square

Phone: 412-624-8843

Office Hours: TBD (usually: one hour after each lecture)

Course Objective The course aims to provide a gentle introduction to Data Science principles, methodologies, technologies, and tools for non-CS majors. The name of the course comes from the famous Arthur C. Clarke quote “*Any sufficiently advanced technology is indistinguishable from magic.*”

Course Description: This course is an *Introduction to Data Science for Beginners*, designed as a first course in computer science, for non-CS majors and CS minors, aiming to introduce students to basic data management technologies and data analytics skills. The course will consist of about 1/3 of introduction to computer programming (using Python), 1/3 of introduction to data management technologies, and 1/3 of introduction to data analytics. The course will adopt the point of view of a user of data (e.g., who is just combining data and analyzing it using tools) and not a provider of data (e.g., who would be implementing a database-driven web site), as is typically be case for related courses for CS majors.

Course Format

- **Lectures (6%):** Two 75-minute meetings per week. Lectures will be fairly interactive and will include quizzes and student presentations. Class participation (discussions and in-class quizzes) will be a component of the students’ final grade (6%).
- **Recitations:** One 50-minute meeting per week (Friday). Needs to be in a computer lab, to provide hands-on experience, for the projects and assignments. Some of the recitations will follow the flipped classroom model, where additional material was given as reading/practice assignment ahead of time.
- **Exams (40%):** We will have two midterm exams (roughly timed at the one third and the two thirds time-points of the term), but no final exam, since there will be a term project.
- **Assignments (42%)** There will be six assignments, all of them being hands-on, using real data sets, whenever applicable. At least one of the assignment will optionally be a group assignment, for groups of two to three students. At least one assignment will include an in-class presentation component.
- **Term Project (12%):** The term project will be a group project (teams of two to three students), and it will involve hypothesis formation, data acquisition, data analysis, and data presentation.

Textbooks: There are two reference textbooks, both available for free for students, through the University’s subscription to O’Reilly’s *Safari Bookshelf*:

- **Data Science from Scratch – First Principles with Python**
By Joel Grus (O’Reilly), Published: April 2015
Purchase: <http://shop.oreilly.com/product/0636920033400.do>, or View online:
<http://proquest.safaribooksonline.com.pitt.idm.oclc.org/9781491901410>.
- **Think Python, 2nd Edition – How to Think Like a Computer Scientist**
By Allen B. Downey (O’Reilly), Published: December 2015
Purchase: <http://shop.oreilly.com/product/0636920045267.do>, or View online:
<http://proquest.safaribooksonline.com.pitt.idm.oclc.org/9781491939406>.

Prerequisites: Although there are no formal prerequisites, some familiarity with computers in general and with computer programming in particular are highly recommended.

Anti-requisites: This course should not be taken by CS majors or anyone who has passed four of the five CS04xx courses.

Class Web Page: <http://data-witchcraft.org>

All handouts and class notes will be published on the class web page. You are expected to check this page frequently (at least twice a week).

Class communications policies (NEW - please read carefully!):

- **Mailing List** – All students will be automatically subscribed to the class mailing list, so that they receive time-sensitive announcements from the instructor and TA(s).
- **In-class student responses** – we will use the **Socrative** system (<http://www.socrative.com>) to capture student responses to questions and record attendance. It is crucial that you provide your Pitt user account name (e.g., xyz123) at the name prompt, to properly record your answers.
- **Email to instructor and TA – instead of email**, we will use the **Piazza** system (which is essentially a web-based bulletin board) for questions and clarifications to assignments. More instructions will be posted on the class web site.
- **Confidential Email** – in case you need to communicate with the instructor and TA outside of the Piazza system (i.e., for confidential matters), you should send email to cs0155-staff@cs.pitt.edu. We will make every effort to respond to all email requests within one business day at the latest. **Due to spam filtering, you should always use your pitt email address when sending email and include your full name.**

E-mail Communication Policy: Each student is issued a University e-mail address (username@pitt.edu) upon admittance. This e-mail address may be used by the University for official communication with students. Students are expected to read e-mail sent to this account on a regular basis. Failure to read and react to University communications in a timely manner does not absolve the student from knowing and complying with the content of the communications. The University provides an e-mail forwarding service that allows students to read their e-mail via other service providers (e.g., GMail, Yahoo, Hotmail). Students that choose to forward their e-mail from their pitt.edu address to another address do so at their own risk. If e-mail is lost as a result of forwarding, it does not absolve the student from responding to official communications sent to their University e-mail address. To forward e-mail sent to your University account, go to <http://accounts.pitt.edu>, log into your account, click on Edit Forwarding Addresses, and follow the instructions on the page. Be sure to log out of your account when you have finished. (For the full E-mail Communication Policy, go to www.bc.pitt.edu/policies/policy/09/09-10-01.html)

Technology Policy (NEW - please read carefully!): Since this is the 21st century, the use of laptops, tablets, and other digital devices **is allowed** in class. **However**, when using digital devices in the classroom you must:

- **be mindful** – when you are emailing, tweeting, texting, surfing, etc, you are not paying attention. Research shows that no one can multitask that well, you included. Paying attention and taking good notes is essential to success in this course. Isn't that why you are here?
- **be respectful** – your use of digital devices should not distract other students in the class. It is unlikely that taking notes or searching class-related topics will be distracting to the other students. However, viewing videos of kittens or ice bucket challenges (gone well or gone wrong) will likely distract others. Complaints about inappropriate technology use in class will result in your privileges being curtailed or revoked.
- **be honest** – emailing, surfing, and the use of any other applications or technologies is not allowed during examinations. Doing so (unless explicitly allowed) is considered cheating in the exam and will be dealt accordingly.

Cell Phone Use (NEW - please read carefully!): Answering a cell phone or texting is very disruptive and hence any use of a cell phone to make or receive calls or text messages **is not permitted** in the class or recitation. Cell phones must be switched to silent mode and if you have a phone call which cannot wait until the end of the class, you need to step out of the class and then answer it.

Audio/Video Recording: To ensure the free and open discussion of ideas, students may not record classroom lectures, discussion and/or activities without the advance written permission of the instructor, and any such recording properly approved in advance can be used solely for the student's own private use.

Copyrighted Material All material provided through this web site is subject to copyright. This applies to class and recitation notes, slides, handouts, assignments, solutions, project descriptions, etc. You are allowed (and expected!) to use all the provided material for personal use. However, you are strictly prohibited from sharing the material with others in general and from posting the material on the Web or other file sharing venues in particular.

Grading Policy: Unless explicitly noted otherwise, the work in this course is to be done independently. Discussions with other students on the assignments should be limited to understanding the statement of the problems (except when assignments are to be done in groups in which case it is expected of members of the same group to work together). **Cheating in any way, including giving your work to someone else, will result in an F for the course and a report to the appropriate University authority.** Submissions that are alike in a substantive way will be considered to be cheating by ALL involved parties. Please protect yourselves by only storing your files in private directories, and by retrieving all printouts promptly.

Students are expected to abide by the Dietrich School of Arts and Sciences' Academic Integrity code of conduct, posted at <http://www.as.pitt.edu/fac/policies/academic-integrity>

All assignments must be submitted electronically. Grades can be appealed up to two weeks after they have been posted; no appeals will be considered after that time.

Academic Integrity Policy: Cheating/plagiarism will not be tolerated. Students suspected of violating the University of Pittsburgh Policy on Academic Integrity, noted below, will be required to participate in the outlined procedural process as initiated by the instructor. A minimum sanction of a zero score for the quiz, exam or paper will be imposed. (For the full Academic Integrity policy, go to www.provost.pitt.edu/info/ail.html)

Late Policy: A late assignment will receive a deduction of 5 points if it is up to one day past the deadline and 15 points if it is up to two days past the deadline. Assignments that are past two days late will not be accepted.

Make-up Policy: Students are expected to be present for all exams and quizzes. Make-up exams will only be given in the event of an emergency, and only if the instructor is informed **in advance**. Failure to notify the instructor prior to missing an exam will result in a zero for the exam.

Final Exam Conflict Policy: In case you have a final exam conflict (i.e., have more than two exams scheduled on the same date during finals week), you need to notify the instructors of all classes involved in order to resolve the conflict by the sixth week of classes, according to the University policy (posted at <http://www.registrar.pitt.edu/classroomscheduling.html>).

Students with Disabilities: If you have a disability for which you are or may be requesting an accommodation, you are encouraged to contact both your instructor and the Office of Disability Resources and Services, 140 William Pitt Union, 412-648-7890, as early as possible in the term. Disability Resources and Services will verify your disability and determine reasonable accommodations for this course. Their web site is <http://www.drs.pitt.edu>.

Religious Observances: In order to accommodate the observance of religious holidays, students should inform the instructor (by email) of any such days that conflict with scheduled class activities **within the first two weeks of the term**.

Tentative Schedule: A detailed reading guide will be published on the web page, along with the class notes and additional online articles and resources. Time permitting, the tentative class schedule (for lectures, recitations, assignments, and exams) will be as follows:

- Week 1** Introduction to Data Science and Big Data
- Week 2** Introduction to Python Programming (Control flow, Variables, Basic data structures)
- * **Week 3** Python Programming (User-Defined Functions)
- Week 4** Python Programming (File Input/Output)
- * **Week 5** Python Programming (Parsing Popular data exchange formats: CSV, JSON, RSS)
- ✓ **Week 6** Introduction to Data Mining (Clustering, Association Rule Mining)
- * **Week 7** Intro to Data Visualization (Powerpoint, Tableau, Google Fusion Tables)
- Week 8** Data Visualization using Python
- * **Week 9** Introduction to Recommendation Systems
- ✓ **Week 10** Introduction to Data Warehousing and (Statistical) Data Summarization
- Week 11** Introduction to Relational Databases
- * **Week 12** Introduction to SQL (Select, Project, Join queries)
- Week 13** SQL Programming using Python
- * **Week 14** Advanced Topics (Graph Databases, REST APIs)
- Week 15** Term Project Presentations

An * denotes an assignment that week, whereas a ✓ denotes a midterm examination. There is no final examination.

Ideas for term projects:

- Utilizing open data from different cities (e.g., Pittsburgh, Chicago, and Washington DC)
- Identifying, combining, and analyzing interesting datasets from <http://data.gov>
- Using publicly available *challenge data* (e.g., from Kaggle, DEBS, etc)
- Utilizing movie rating data from <http://grouplens.org/datasets/movielens/>
- Analyzing historical weather data

[Last updated on June 21, 2016 at 4:01pm EST]