School of Computing
UNIVERSITY OF GEORGIA

CSCI 4380/6380 DATA MINING

Fei Dou

Assistant Professor
School of Computing
University of Georgia

August 16, 2023

Course Overview

About CSCI 4380/6380

- Two sessions: A and B
  - Session A: Dr. Kimberly Van Orman
  - Session B: Dr. Fei Dou

Course Details

- **Instructor:** Fei Dou (fei.dou@uga.edu)
- **Lecture Times and Locations:**
  - TR 3:55–5:10 PM, Physics - Room 0221
  - W 4:10–5:00 PM, Boyd - Room 0208
- **Office Hours (tentative):**
  - Tuesday 8:30 am - 9:30 am
  - Also, by appointment
  - Location: 542 Boyd Research and Education Center
- **TA:** TBD
- **Prerequisites:** CSCI 2720 (Data Structures) Not Relevant for Grad Students
- **Course Website:** https://www.elc.uga.edu/.

Course Details

- **Texts:** *The course will follow:*
  - 1. Introduction to Data Mining by Pang-Ning Tan, Michael Steinbach, and Vipin Kumar, 2nd Edition, ISBN-10 : 0133128903
  - 2. Pattern Recognition and Machine Learning (Information Science and Statistics) by Christopher M. Bishop, ISBN-10: 0387310738
  - 3. Deep Learning (Adaptive Computation and Machine Learning Series) by Ian Goodfellow, Yoshua Bengio, and Aaron Courville, ISBN-10: 0262035618
- The texts are not technically required (but will be useful).
- **Lectures**:
  - Some straight lectures. Some working problems on the board.
  - At times, students will be asked to present material. E.g., discuss a conference paper.

Software

- A good portion of the course will be Python-based.
- We'll use Anaconda, which provides:
  - **Pycharm** (an IDE) and **Jupyter Notebook** (a Web-based environment)
  - **NumPy** (mostly for working with arrays of numerical data)
  - **SciPy** (for math and statistics)
  - **Matplotlib** (for graphing/visualization)
  - **Pandas** (for working with heterogeneous data)
  - **scikit-learn** (for developing ML models)

Homework

- Maybe 5-6 assignments.
- All homework must be submitted on time
- Lowest HW score will be dropped.

Unless explicitly stated otherwise:

1. Students must work alone on their HW. All work submitted by a student must be his or her own and not the result of collaboration or the improper utilization of outside sources.
2. HW must be typed and in PDF format – No handwritten submissions or phone pics.

Tests and Final Exam

- During the course of the semester, there will be:
  - Preliminary exam, **Next Tuesday**, (**not counted**, notations, calculus, linear algebra, statistics)
  - Two open-book open-notes tests
  - A final exam.

- The final exam is tentatively scheduled to be in-person

Presentations

- Students in the graduate section (6380) must also give an in-class presentation (or some suitable substitute).
- Presentations will be machine learning or data mining topics from conference or journal papers.
- A collection of potential papers will be posted to ELC, but students are also able to choose their own topic.

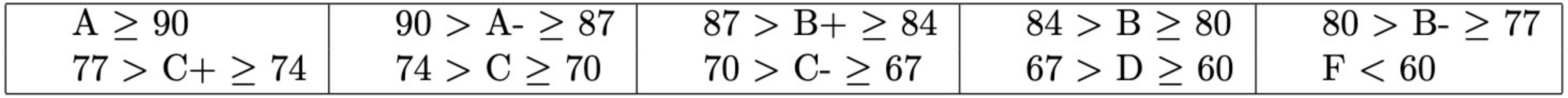
- Students in the undergraduate section (4380) are encouraged to ask questions during or after presentation.

Grading

4380 Section (Undergrads)
Homework (35%)
Test 1 (10%)
Test 2 (10%)
Final Exam (20%)
Term Project (25%)

6380 Section (Grads)
Homework (30%)
Test 1 (10%)
Test 2 (10%)
Presentation (10%)
Final Exam (20%)
Term Project (20%)

Final Letter Grades



$\text{A} \geq 90$	$90 > \text{A-} \geq 87$	$87 > \text{B+} \geq 84$	$84 > \text{B} \geq 80$	$80 > \text{B-} \geq 77$
$77 > \text{C+} \geq 74$	$74 > \text{C} \geq 70$	$70 > \text{C-} \geq 67$	$67 > \text{D} \geq 60$	$\text{F} < 60$

- Grades are based on work submitted.

- **Regrades:** You can request reevaluation of any graded material. However, requests must be made ≤7 days after the material has been returned.

Late Submission and Regrading Policy

- For homework assignments, 20% is deducted for each late day for up to 48 hours (including weekends) after which submissions are not accepted.
- Late presentation materials and project reports not accepted.
- You may request a re-grade of any graded item any time within 7 days of receiving the grade on eLC.
- For grades posted on or after reading day, students must send a regrade request (using uga email account) within 2 days after the grade was posted on eLC.
- If a rubric is posted for assignments, then the regrade request must include which parts were incorrectly graded.
- Regrade requests may result in a lower grade.

Communication

- Important announcements will be made through eLC.

- Please turn on email notifications related to Announcements and Grades.
  - Log into eLC, click on your name/profile picture on the upper right hand corner, click “Notifications”, scroll down and check the appropriate boxes.

- Course materials from eLC must not be shared with anyone, posted on any websites, or used for commercial purposes.

Term Project

A typical project involves:

- Selecting a data set/datasets you have interests to analyze, and a problem.
- Preprocessing the data.
- Comparing several existing data mining techniques
- Analyzing and discussing the results.
- Discussing the current work in the context of related work.
- Discussing ways in which the current work could be improved.

Term Project Description

- Each team has 2 ~ 3 students.

- Consists of a presentation, a final report, and the executable source codes.
  - Presentations are scheduled before the final exam, reports and codes are due after the exam;

- Any topics related to the course. Topic selection suggestions:
  - Standard tasks: https://www.kaggle.com/.
    Real-world applications: Problems or applications that you are interested in.

- Detailed requirements to be introduced later.

Final Presentation

- November 26 - 28.
- All team members should present their slides.
- Each presentation is a talk (20 minutes) + QA (5 minutes) = 25 minutes.
- Slides should be emailed to the instructor before presentation.
- Content to be covered:
  - Background.
  - The formal problem definition.
  - The details of your solution.
  - Experimental results on the datasets, when compared with baseline methods, under the evaluation metrics.
  - Conclusion.

Final Report

- Latex template: To be provided later
- Length: ≥ 4 pages + unlimited references Format:
  - Introduction
  - Literature Review (short)
  - Problem Definition
  - The Proposed Method
  - Experiments
  - Conclusion
  - References
- Failing to following the template → 20% penalty.

Behavior and Academic Policies

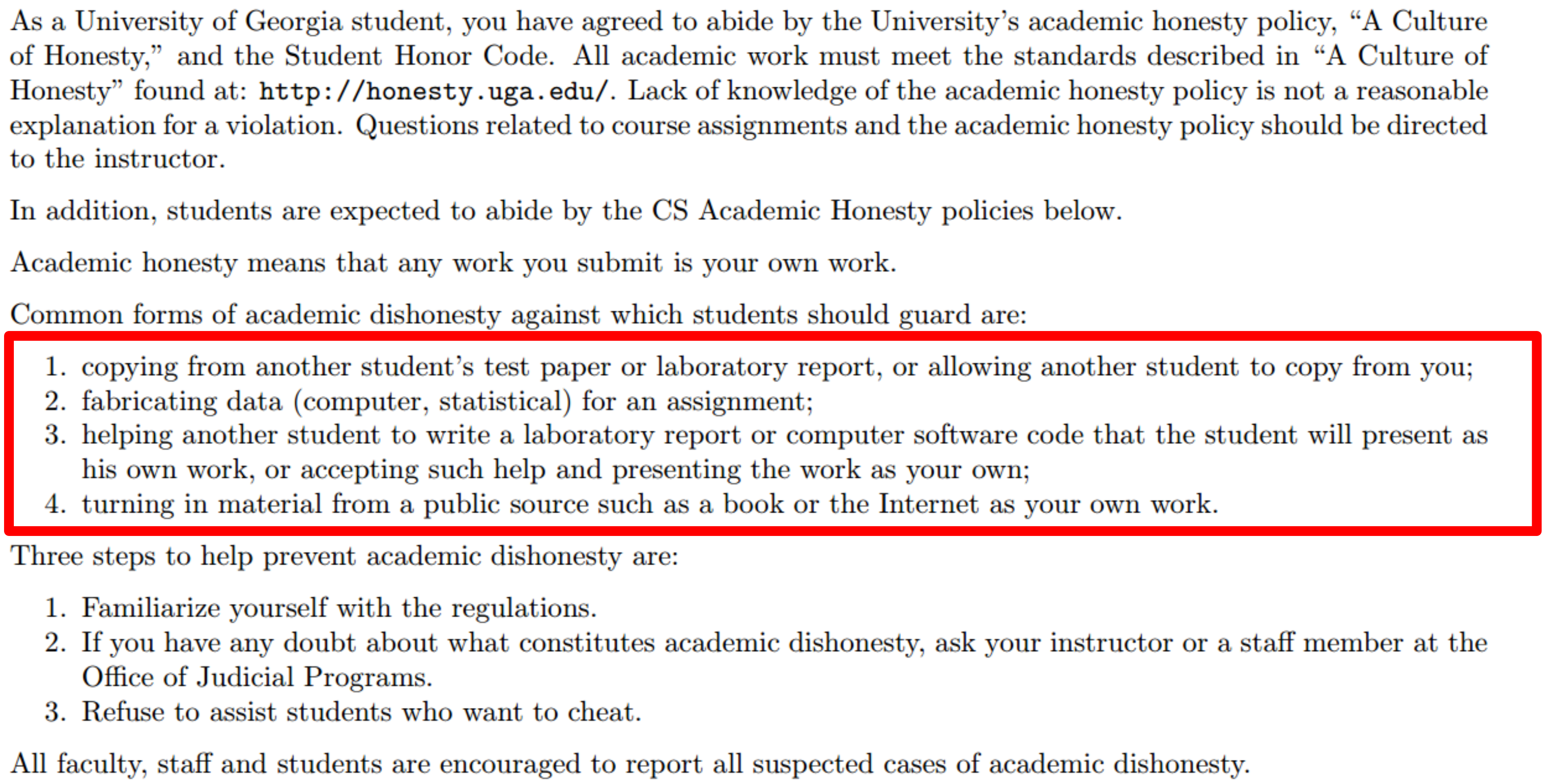
- Review the academic honesty policies of UGA and CS department.
  - See http://honesty.uga.edu/ and the syllabus.
- Review the policies of the course syllabus.

In general, everything you submit should be your own and not the result of collaborating with others or utilizing external sources without approval.

Also

- Students who require special accommodations should talk to me.
- The syllabus is a tentative outline of the course. Alterations might be necessary.

Academic Honesty Policy



As a University of Georgia student, you have agreed to abide by the University's academic honesty policy, "A Culture of Honesty," and the Student Honor Code. All academic work must meet the standards described in "A Culture of Honesty" found at: `http://honesty.uga.edu/`. Lack of knowledge of the academic honesty policy is not a reasonable explanation for a violation. Questions related to course assignments and the academic honesty policy should be directed to the instructor.

In addition, students are expected to abide by the CS Academic Honesty policies below.

Academic honesty means that any work you submit is your own work.

Common forms of academic dishonesty against which students should guard are:

1. copying from another student's test paper or laboratory report, or allowing another student to copy from you;
2. fabricating data (computer, statistical) for an assignment;
3. helping another student to write a laboratory report or computer software code that the student will present as his own work, or accepting such help and presenting the work as your own;
4. turning in material from a public source such as a book or the Internet as your own work.

Three steps to help prevent academic dishonesty are:

1. Familiarize yourself with the regulations.
2. If you have any doubt about what constitutes academic dishonesty, ask your instructor or a staff member at the Office of Judicial Programs.
3. Refuse to assist students who want to cheat.

All faculty, staff and students are encouraged to report all suspected cases of academic dishonesty.

Accommodations for Disabilities

- Students with a disability or health-related issue who need a class accommodation should make an appointment to speak with the instructor as soon as possible.

- Paperwork from the Disability Resource Center (https://drc.uga.edu/) will be required.

Mental Health and Wellness Resources

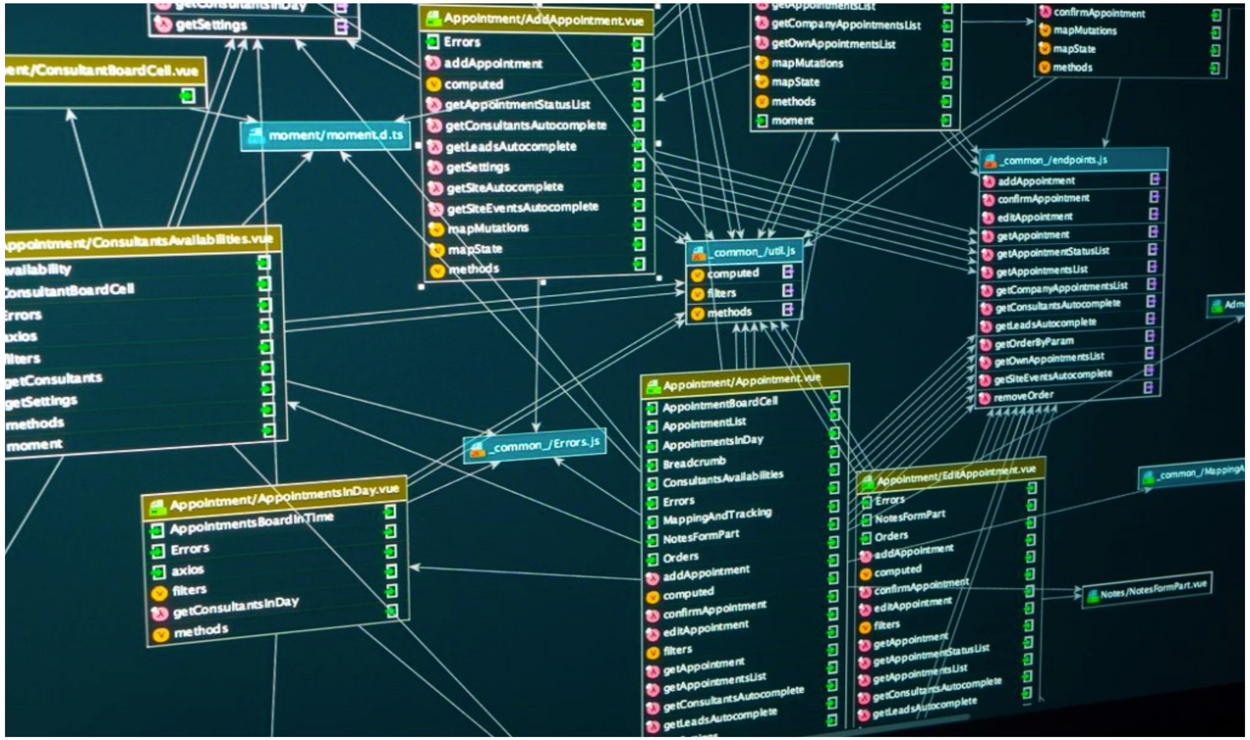
- If you or someone you know needs assistance, you are encouraged to contact Student Care and Outreach in the Division of Student Affairs at 706-542-7774 or visit https://sco.uga.edu/. They will help you navigate any difficult circumstances you may be facing by connecting you with the appropriate resources or services.
- UGA has several resources for a student seeking mental health services (https://www.uhs.uga.edu/bewelluga/bewelluga) or crisis support (https://www.uhs.uga.edu/info/emergencies).
- If you need help managing stress anxiety, relationships, etc., please visit BeWellUGA (https://www.uhs.uga.edu/bewelluga/bewelluga) for a list of FREE workshops, classes, mentoring, and health coaching led by licensed clinicians and health educators in the University Health Center.
- Additional resources can be accessed through the UGA App.

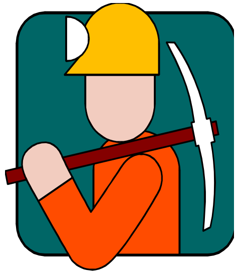
Survey

- 1. Are you undergraduate or graduate student?
- 2. Did you take any course about linear algebra, probability & statistics? If yes, please list those courses.
- 3. Did you take any course about machine learning, and big data analysis before? If yes, please list those courses.
- 4. What’s your preferred programming language?
- 5. Have you ever used Pytorch/Tensorflow or any other deep learning tool before? If yes, please list it (them) and rate your experience from 1 to 5 with 5 being the strongest.
- 6. What do you expect to learn from this class?

Data Mining - Overview

What is Data Mining?





After years of data mining, there is still no unique answer to this question.

A tentative definition:

Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data.