CAS CS 105: Introduction to Databases and Data Mining
Boston University
Spring 2012

Syllabus

Description: Databases and other collections of data are everywhere. Retailers use
data about customers and their purchases to make decisions that increase profits.
Researchers analyze data about the human genome to find treatments for diseases.
Policymakers analyze socioeconomic data to gain insights that guide their decisions.
Online music and video services perform data mining to deliver customized
recommendations. How does all this work? CS 105 explores the ways in which
collections of data are organized, stored, and analyzed. Topics include relational
databases and the SQL query language, the writing of simple programs to process
data, data visualization and the graphical display of data, and data-mining techniques
for discovering patterns in data. Applications from a variety of domains (including
business, the arts, the life sciences, and the social sciences) are used to illustrate the
course's key concepts. The course counts towards the math/CS divisional studies
requirement.

Prerequisites: none

Instructor
David G. Sullivan, Ph.D. (dgs @ cs . bu . edu – omitting the spaces)
office hours: Mon., 3-5 p.m.; Wed., 2:30-4:30 p.m., and by appointment
office: Psychology Building (PSY), room 228D
64 Cummington Street (behind Warren Towers)

Teaching Fellow
Haohan Zhu (zhu @ cs . bu . edu – omitting the spaces)
See the staff page of the course website for his office hours.

Meeting Times and Places
lectures: MWF, 1-2 p.m., CAS 222
labs: Thursdays, 4-5 p.m. (A2)
       Thursdays, 5-6 p.m. (A3)
       All labs meet in the CS teaching lab, EMA 304.

Course Website: http://www.cs.bu.edu/courses/cs105
Please consult this site frequently (at least once per week) for important
announcements and updates to the course materials handed out in lecture.
Requirements
1. Nine problem sets
2. A final project, which will involve using the techniques covered in the course to organize and analyze a collection of data that interests you, to draw conclusions based on your analysis, and to present your results in a clear and compelling way
3. Three closed-book quizzes
4. A closed-book final exam
5. Attendance at and participation in both the lectures and labs

Policies
Lateness policy: Assignments must be submitted by the date and time listed on the assignment. There will be a 10% deduction for homework that is up to 24 hours late, and a 20% deduction for homework that is 24-48 hours late. We will not accept any homework that is more than 48 hours late.

Plan your time carefully, and don’t wait until the last minute to begin an assignment. Starting early will give you ample time to ask questions and obtain assistance from members of the course staff.

Determining the final grade:
- attendance/participation 10%
- problem sets 30%
- final project 10%
- quizzes 20%
- final exam 30%

Extensions and makeup quizzes/exams will only be given in documented cases of serious illness or other emergencies.

You cannot redo or complete extra work to improve your grade. Incompletes will not be given.

Collaboration Policy*
The collaboration policy for this class is as follows.

- You are strongly encouraged to collaborate with one another in studying the textbook and the lecture material.

- As long as it satisfies the following conditions, collaboration on the homework assignments is encouraged and will not reduce your grade:
  1. Before discussing each homework problem with anyone else, you must give it an honest half-hour of serious thought.
  2. You may discuss ideas and approaches with other students in the class, but not share actual code or solutions to other types of problems. In other words, the work you submit must be entirely your own, which you must complete without looking at other people's work, and you must not permit others to copy your work. You must also acknowledge clearly in the appropriate portion of your solutions (e.g., in the comments of your code) people with whom you discussed ideas for that portion.
3. You may get help from the teaching staff and from tutors in the lab for specific homework problems. Don't expect them to do it for you, however.

4. If you get really stuck with a bug in a program (defined roughly as over an hour of frustration), you are allowed to get help from a friend as long as you acknowledge that help clearly in your solutions (e.g., in the comments of your code).

5. You may not work with people outside this class (but come and talk to us if you have a tutor), seek online solutions, get someone else to do it for you, etc.

- You are not permitted to collaborate on exams.

The last point is particularly important: if you don't make an honest effort on the homework but always get ideas from others, your exam score will reflect it.

Violations of Collaboration Policy*

Violations of collaboration policy fall into two categories: ones that are acknowledged at the time they occur (for example, in clearly marked comments in your code) and ones that are unacknowledged.

Acknowledged violations (e.g., using someone else's code for a method you didn't know how to write yourself, and stating clearly in your code that this is not your own work) will result in an appropriate reduction in the grade, but will not be considered cheating. You should send an email to cs105-staff@cs.bu.edu about all such violations, or anything that you think may possibly be considered as such.

Unacknowledged violations of the collaboration policy—for example, not stating the names of your collaborators, or any other attempt to represent the work of another as your own—will result in an automatic failing grade and will be reported to the Academic Conduct Committee (ACC). The ACC often suspends or expels students deemed guilty of plagiarism or other forms of cheating. We will assume that you understand the CAS Academic Conduct Code, copies of which are available in room CAS 105.

If you are uncertain as to whether a particular kind of interaction with someone else constitutes illegal collaboration or academic dishonesty, please ask Dr. Sullivan before taking any action that might violate the rules; if you can't reach him in time, then at the very least include a clear explanation of what happened in your homework write-up to avoid being treated as a cheater. Citing your sources is usually the easiest way out of trouble.

* Thanks to Prof. Leo Reyzin, who wrote the original versions of the sections describing the collaboration policy and violations to that policy. I have made only minor modifications.
Textbooks
We will provide lecture notes that fully cover all of the material you are expected to learn as part of the course. However, you may wish to consider purchasing:


These books will be available for purchase at Barnes & Noble.

Schedule (tentative)

<table>
<thead>
<tr>
<th>week</th>
<th>lecture dates</th>
<th>topics, exams, and special dates</th>
<th>optional readings</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1/18, 1/20</td>
<td>Course overview and introduction Database fundamentals <em>No lab meetings this week</em></td>
<td>K (Kroenke): chap. 1</td>
</tr>
<tr>
<td>2</td>
<td>1/23, 1/25, 1/27</td>
<td>The relational model: How is data organized in a typical database? SQL: a language for creating, modifying, and extracting data from a database</td>
<td>K: chap. 2, 3</td>
</tr>
<tr>
<td>3</td>
<td>1/30, 2/1, 2/3</td>
<td>SQL (cont.) <em>Problem Set 1 due on 2/1 1/30: last day to add a class</em></td>
<td>K: chap. 2, 3</td>
</tr>
<tr>
<td>4</td>
<td>2/6, 2/8, 2/10</td>
<td>SQL (cont.) case study: an astronomical database <em>Problem Set 2 due on 2/8</em></td>
<td>see above</td>
</tr>
<tr>
<td>5</td>
<td>2/13, 2/15, 2/17</td>
<td>Programming in Python: intro., working with numbers <em>Problem Set 3 due on 2/15</em></td>
<td>Z (Zelle): chap. 1, 2, 3</td>
</tr>
<tr>
<td>6</td>
<td>2/21, 2/22, 2/24</td>
<td>More Python: making decisions <em>Quiz 1 on 2/22 Problem Set 4 due on 2/24 (Friday) No lecture on 2/20 (Presidents’ Day) Lecture on Tues. 2/21 (Mon. schedule) 2/21: last day to drop without a 'W', and to change from credit to audit status</em></td>
<td>Z: sect. 7.1-7.3</td>
</tr>
<tr>
<td>7</td>
<td>2/27, 2/29, 3/2</td>
<td>More Python: working with strings/text and lists; accessing a database <em>Problem Set 5 due on 2/29</em></td>
<td>Z: sect. 4.1-4.5</td>
</tr>
<tr>
<td>8</td>
<td>3/5, 3/7, 3/9</td>
<td>More Python: working with data stored in a text file <em>Problem Set 6 due on 3/7</em></td>
<td>Z: sect. 4.6</td>
</tr>
<tr>
<td></td>
<td>3/10-3/18</td>
<td><em>Spring break</em></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>3/19, 3/21, 3/23</td>
<td>Data visualization: how to create compelling and useful data graphics <em>Problem Set 7 due on 3/21</em></td>
<td></td>
</tr>
<tr>
<td>Week</td>
<td>Dates</td>
<td>Topics</td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>-------</td>
<td>--------</td>
<td></td>
</tr>
</tbody>
</table>
| 10   | 3/26, 3/28, 3/30 | Data mining: what is it? How can a computer learn to categorize things?  
**Quiz 2 on 3/26**  
3/30: last day to drop a class with a 'W' |
| 11   | 4/2, 4/4, 4/6 | More data mining: categorizing things (cont.), finding patterns in numeric data  
**Problem Set 8 due on 4/4**  
**Final-project proposal due on 4/6** |
| 12   | 4/9, 4/11, 4/13 | More data mining: preparing data for mining, discovering arbitrary relationships in your data  
**Work on final project** |
| 13   | 4/18, 4/20 | More data mining: discovering relationships (cont.); case study  
**Problem Set 9 due on 4/20 (Friday)**  
No lecture on 4/16 (Patriots' Day)  
4/18 is a Mon. schedule |
| 14   | 4/23, 4/25, 4/27 | Case studies and/or project presentations  
**Quiz 3 on 4/25** |
| 15   | 4/30, 5/2 | Case studies and/or project presentations  
**Final projects due on 5/2**  
**Final exam: Friday, 5/11, 3-5 p.m.** |