Welcome to CS 105!

- This course examines how collections of data are organized, stored, and processed.

- Topics include:
  - databases
  - programming
  - data mining
  - data visualization

- We'll consider applications from a variety of domains.
  - business, the arts, the life sciences, the social sciences
Broad Goals of the Course

• To give you computational tools for working with data

• To give you insights into databases and data mining
  • help you to understand their increasingly important role

• To expose you to the discipline of computer science
  • to how computer scientists think and solve problems

“Computer science is not so much the science of computers as it is the science of solving problems using computers.”

- Eric Roberts, Stanford

Data, Data Everywhere!

• financial data
• commercial data
• scientific data
• socioeconomic data
• etc.

AGCTTTTCATTCTGACTGCAACGGGCAATATGTCTCTGTGTGGATTAAAAAAG
AGTGTCTGATAGCAGCTTCTGAACTGGTTACCTGCCGTGAGTAAATTAAAATTTA
TTGACTTAGGTCACTAAATACTTTAACCAATATAGGCATAGCGCACAGACAGA
TAAAAATTACAGAGTACACAACATCCATGAACGCATTAGCACCACCATTACCA
CCACCATCACCATTACCACAGGTAAGGTGCGGGCTGACGCGTACAGGAAACAC
AGAAAAAAGCCCGCACCTGACAGTGCGGGCTTTTTTTTTCGACCAAAGGTAAC
GAGGTAACAACCATGGAAGTTCGGCGGTACATCAGTGGCAAATGCAGAACGTT
TTCTGCGTGTTGCCGATATTCTGGAAAGCAATGCCAGGCAGGGGCAGGTGGCC
ACCGTCCTCTC

Data, Data Everywhere!
Databases

• A database is a collection of related data.
  • example: the database behind StudentLink
  • other examples?

• Managed by some type of database management system (DBMS)
  • a piece of software (a program) that allows users to store, retrieve, and update collections of data

The Amount of Data Is Exploding!

• Example: the GenBank database of genetic sequences

  • on this graph, the data doubles every 12-14 months
  • as of May 2006, the doubling time was less than a year and getting shorter!

from: NCBI Field Guide presentation
(http://ftp.ncbi.nih.gov/pub/FieldGuide/Slides/Current/MyHolyoke.05.10.06)
The Amount of Data Is Exploding!

- Example: the UN Database (data.un.org)

from "An Analysis of Factors Relating to Energy and Environment in Predicting Life Expectancy", CS 105 Final Project by Valerie Belding '12

The Amount of Data Is Exploding!

- Example: the Google Ngrams Corpus

books.google.com/ngrams
The Amount of Data Is Exploding!

Relational Databases

- Most data collections are managed by a DBMS that employs a way of organizing data known as the *relational model*.
  - examples: IBM DB2, Oracle, Microsoft SQL Server, Microsoft Access

- In the relational model, data is organized into *tables* of *records*.
  - each record consists of one or more *fields*
  - example: a table of information about students

<table>
<thead>
<tr>
<th>id</th>
<th>name</th>
<th>address</th>
<th>class</th>
<th>dob</th>
</tr>
</thead>
<tbody>
<tr>
<td>12345678</td>
<td>Jill Jones</td>
<td>Warren Towers 100</td>
<td>2007</td>
<td>3/10/85</td>
</tr>
<tr>
<td>25252525</td>
<td>Alan Turing</td>
<td>Student Village A210</td>
<td>2010</td>
<td>2/7/88</td>
</tr>
<tr>
<td>33566891</td>
<td>Audrey Chu</td>
<td>Main Hall</td>
<td>2008</td>
<td>10/2/86</td>
</tr>
<tr>
<td>45678900</td>
<td>Jose Delgado</td>
<td>Student Village B300</td>
<td>2009</td>
<td>7/13/88</td>
</tr>
<tr>
<td>66666666</td>
<td>Count Dracula</td>
<td>The Dungeon</td>
<td>2007</td>
<td>11/1431</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
SQL

- A relational DBMS has an associated *query language* called SQL that is used to:
  - define the tables
  - add records to a table
  - modify or delete existing records
  - retrieve data according to some criteria
    - example: get the names of all students who live in Warren Towers
    - example: get the names of all students in the class of 2017, and the number of courses they are taking
  - perform computations on the data
    - example: compute the average age of all students who live in Warren Towers

Example Database

- A relational database containing data obtained from imdb.com

  ![Example Database](image)

- We'll use SQL to answer (or at least explore) questions like:
  - How many of the top-grossing films of all time have won one or more Oscars?
  - Does the Academy discriminate against older women?
Beyond Relational Databases

- While relational databases are extremely powerful, they are sometimes inadequate/insufficient for a given problem.

- Example: DNA sequence data
  >gi|49175990|ref|NC_000913.2| Escherichia coli K12, complete genome
  AGCTTTTCATTCTGACTGCAACGGGCAATATGTCTCTGTGTGGATTAAAAAAAGAGTGTCTGATAGCAGCTTCTGAACTGGTTACCTGCCGTGAGTA
  ACGCATATGACCACTACATTACACACACACATCTTATACACACCAGATAGTACGCGCCGCGTACAGGAAACACAGAAAAAAGCCCGCACCTGAC
  AGTGCGGGCTTTTTTTTCGACCAAAGGTAACCGGTAACACCTGCGCCCGCCAGGAAACACAGAAAAAAGCCCGCACCTGACAGTGCGGGCTTTTTTT
  TCAGGTAACCGGTAACACCTGCGCCCGCCAGGAAACACAGAAAAAAGCCCGCACCTGACAGTGCGGGCTTTTTTTTCGACCAAAGGTAACCGGTAACACCT
  TCGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG

- common queries involve looking for similarities or patterns
  - what genes in mice are similar to genes in humans?
- need special algorithms (problem-solving procedures)
- biologists store this data in text files and use simple computer programs to process it
- we'll learn to write simple programs using Python

Data MINING Everywhere!

- Informally, data mining is the process of finding patterns in data.

- Example: customized recommendations

  **DRAMA SUGGESTIONS (about 82)**

  ![The Wrestler](image)
  *The Wrestler*
  Because you enjoyed:
  - Sir City
  - Reservoir Dogs
  - The Big Lebowski

  ![The Visitor](image)
  *The Visitor*
  Because you enjoyed:
  - Gandhi
  - The Motorcycle Diaries
  - The Queen

  ![Brick](image)
  *Brick*
  Because you enjoyed:
  - The Big Lebowski
  - Rushmore
  - Fight Club

- Example: detecting credit-card fraud
How Companies Learn Your Secrets
By CHARLES DUHIGG

Andrew Pole had just started working as a statistician for Target in 2002, when two college marketing department officials stopped by his desk to ask an odd question: "If we wanted to figure out if a customer is pregnant, even if she didn't want us to know, can you do that?"

Pole has a master's degree in statistics and another in economics, and has been obsessed with the intersection of data and human behavior throughout his life. His parents were teachers in North Carolina, while other kids were going to 4-H. Pole was doing algebra and writing computer programs even when he was a toddler, he told me when I spoke with him last year. "I kind of grew up as an evangelizing analytic."

• Target may know that your friend is pregnant before you do!

How the U.S. Uses Technology to Mine More Data More Quickly

WASHINGTON — When Americans analysts hunting terrorists sought new ways to comb through the troves of phone records, emails and other data piling up as digital communications exploded...
Structure of the Course

- databases (4 weeks)
- programming in Python (4 weeks)
- data graphics/visualization (1 week)
- data mining (4 weeks)

Requirements

- Attendance at and participation in lectures and labs (10%)
  - everyone has an allowance of 3 missed classes;
    do not email unless extreme circumstances
  - labs begin next week
  - held in the CS teaching lab, EMA 304
  - complete Lab 0 sometime this week
    (see course website)
- Nine homework assignments (30%)
- Final project (10%): done in teams of three
  - use the techniques covered in the course to explore a
    dataset that interests you
- Three quizzes (20%)
- Final exam (30%)
Textbooks

- **optional:** *Database Concepts, 7th edition* by Kroenke & Auer (Prentice Hall, 2015)


- **required:** *The CS 105 Coursepack*
  - contains all of the lecture notes
  - will be available at Fedex Office at the corner of Comm Ave and Cummmington Mall (the Warren Towers building)

Course Staff

- Instructor: Dave Sullivan (dgs@cs.bu.edu)
- Teaching fellow: Baichuan Zhou (baichuan@bu.edu)
- Office hours and contact info. will be available on the course Web site:
  - http://www.cs.bu.edu/courses/cs105
- *For general course-related questions, email:*
  - cs105-staff@cs.bu.edu
  - *which will forward your question to the full course staff.*
**Algorithm for Finding My Office**

1. Go to the entrance to the MCS (math/CS) building at 111 Cummington Mall – behind Warren Towers.
   
   Do not enter this building!

2. Turn around and cross the street to the doors across from MCS.

3. Enter those doors and take an immediate right.
   
   (continued on next slide)

---

**Algorithm for Finding My Office (cont.)**

4. As you turn right, you should see the door below. Open it and go up the stairs to the second floor.

5. As you leave the stairs, turn right and then go left into a small hallway. My office is the first door on the left (PSY 228D).
Other Details of the Syllabus

- Collaboration

- Policies:
  - lateness
  - please don't request an extension unless it's an emergency!
  - grading

- Please read the syllabus carefully and make sure that you understand the policies and follow them carefully.

- Let us know if you have any questions.