Providing Students with Computational Tools for Working with Data

Center for Excellence and Innovation
in Teaching, Boston University
January 10, 2013

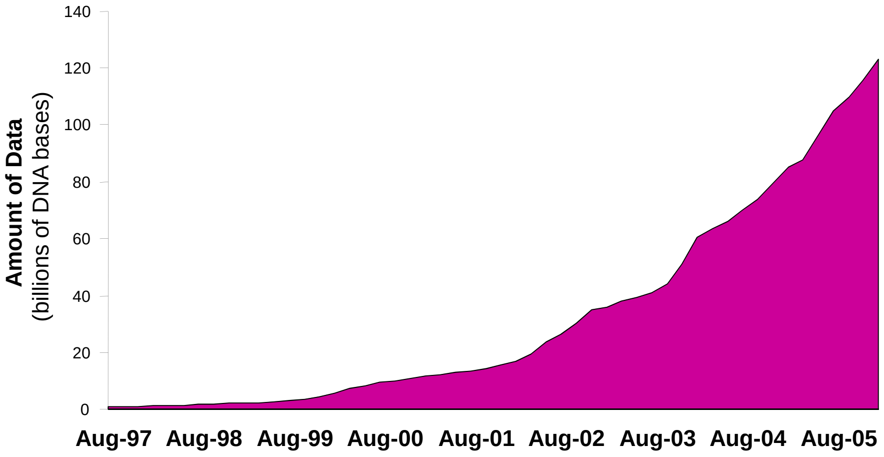
David G. Sullivan, Ph.D.
CAS Computer Science

Databases Are Everywhere

- Example collections of data:
  - account data: banks, credit-card companies, etc.
  - airline data: flights, reservations, etc.
  - biological data: DNA sequences, protein sequences, etc.
  - socioeconomic data
  - other examples?
- Some are managed by a *database management system* (DBMS) like Oracle, SQLServer, etc.
- Some are not.
  - text files (CSV files, tab-delimited, etc.)
  - etc.

The Amount of Data Is Exploding!

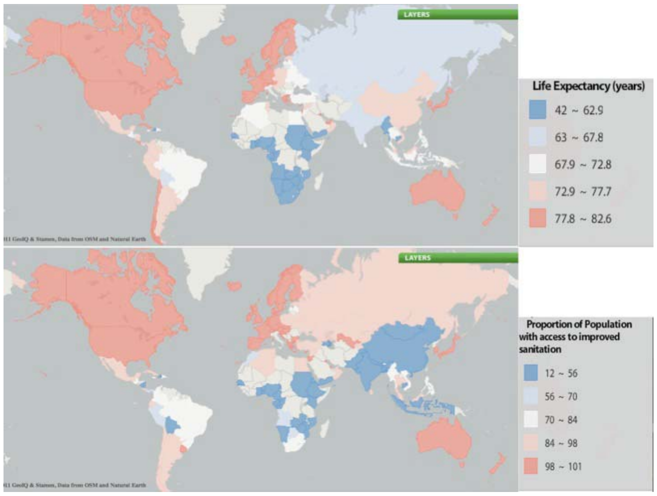
- Example: the GenBank database of genetic sequences



from: NCBI Field Guide presentation
(ftp://ftp.ncbi.nih.gov/pub/FieldGuide/Slides/Current/MtHolyoke.05.10.06/)

The Amount of Data Is Exploding!

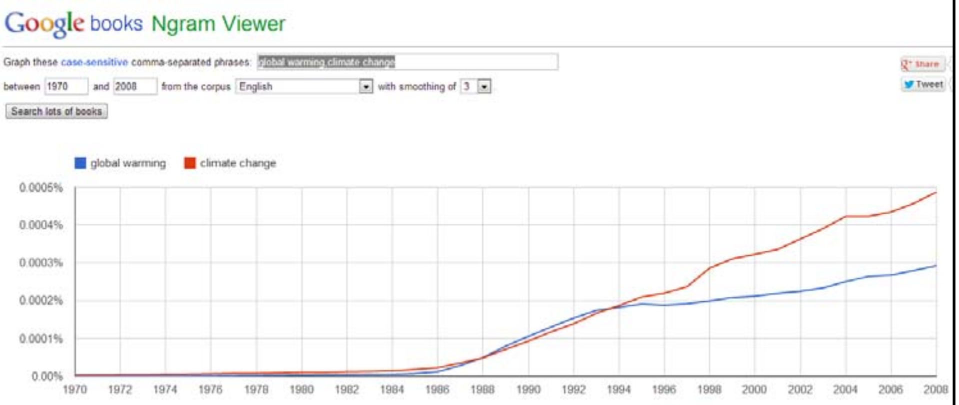
- Example: the UN Database (data.un.org)



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

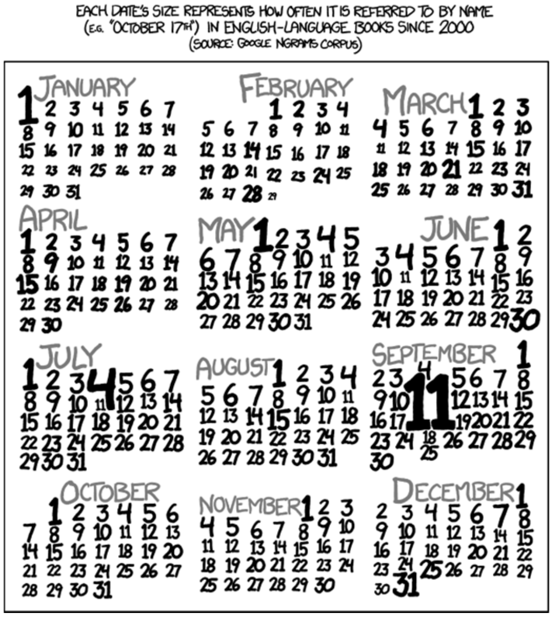


Graph these case-sensitive comma-separated phrases: global warming,climate change

between 1970 and 2008 from the corpus English with smoothing of 3

books.google.com/ngrams

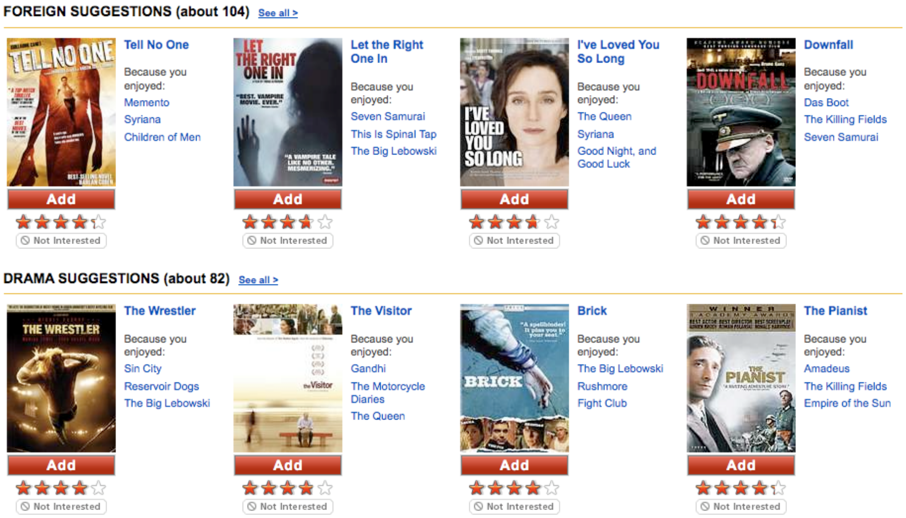
The Amount of Data Is Exploding!



EACH DATE'S SIZE REPRESENTS HOW OFTEN IT IS REFERRED TO BY NAME
(E.G. "OCTOBER 17TH") IN ENGLISH-LANGUAGE BOOKS SINCE 2000
(SOURCE: GOOGLE NGRAMS CORPUS)
JANUARY
FEBRUARY
MARCH
APRIL
MAY
JUNE
JULY
AUGUST
SEPTEMBER
OCTOBER
NOVEMBER
DECEMBER

xkcd.com/1140

Data Mining Is Increasingly Pervasive



FOREIGN SUGGESTIONS (about 104) See all >

Tell No One
Because you enjoyed:
Memento
Syriana
Children of Men

Let the Right One In
Because you enjoyed:
Seven Samurai
This Is Spinal Tap
The Big Lebowski

I've Loved You So Long
Because you enjoyed:
The Queen
Syriana
Good Night, and Good Luck

Downfall
Because you enjoyed:
Das Boot
The Killing Fields
Seven Samurai

DRAMA SUGGESTIONS (about 82) See all >

The Wrestler
Because you enjoyed:
Sin City
Reservoir Dogs
The Big Lebowski

The Visitor
Because you enjoyed:
Gandhi
The Motorcycle Diaries
The Queen

Brick
Because you enjoyed:
The Big Lebowski
Rushmore
Fight Club

The Pianist
Because you enjoyed:
Amadeus
The Killing Fields
Empire of the Sun

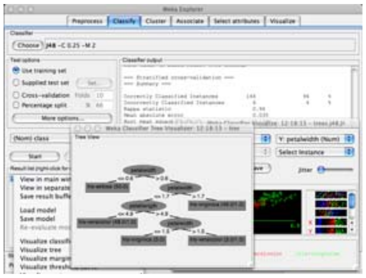
netflix.com

Data Mining Is Increasingly Pervasive

- Other examples:
  - detecting fraudulent credit-card purchases
  - targeted online advertising
  - retailers mining customer-purchase data

The Problem

- Courses on databases and data mining are typically limited to CS majors and grad students.



computer science

pyschology, political science, medicine, ...

- Students from other fields are left out.

Our Solution

- CAS CS 105: Intro. to Databases and Data Mining
- Designed for non-majors
- No prereqs
- Topics include:
  - relational databases (4 weeks)
  - programming in Python (4 weeks)
    - to process data stored in text files
  - data graphics/visualization (1 week)
  - data mining basics (4 weeks)
- Provides a *data-centric* introduction to computer science

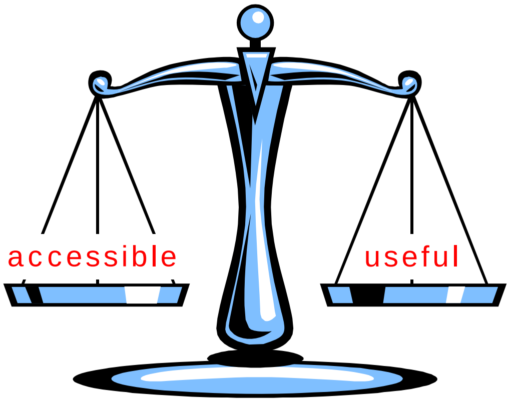
Broad Goals of the Course

- Give students computational tools for working with data
  - applicable skills = motivation
- Provide insight into the underlying concepts
  - abstraction
  - mathematical models
  - algorithmic thinking
- Expose them to the discipline of computer science

Computer science is not so much the science of computers as it is the science of solving problems using computers.
- Eric Roberts, Stanford

A Delicate Balance

- Allow students from non-technical backgrounds to succeed
- Provide sufficient challenge and coverage

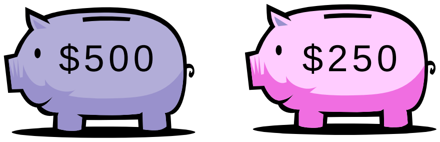


accessible

useful

Unit 1, part I: Database Fundamentals

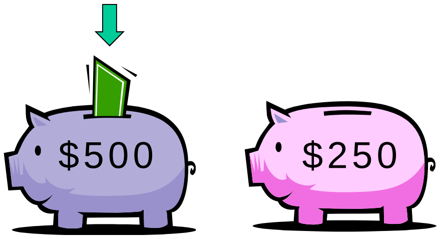
- How data is stored
- Key functions of a DBMS: just the big picture
- ***Example: transactions***
  - balance transfer: $50 from blue to pink
  - transaction = series of steps treated as a single operation
    - ***ensures all steps happen, or none do***



```
begin transaction
remove $50 from blue
add $50 to pink
end transaction
```

Unit 1, part I: Database Fundamentals

- How data is stored
- Key functions of a DBMS: just the big picture
- ***Example: transactions***
  - balance transfer: $50 from blue to pink
  - transaction = series of steps treated as a single operation
    - ***ensures all steps happen, or none do***



```
begin transaction
remove $50 from blue
*** CRASH ***
restore state!
```

Unit 1, part II: Data Modeling

- The relational model
  - data is organized into *tables*
  - example: a table of student info

Id	***name***	***address***	***class***	***dob***
12345678	Jill Jones	Warren Towers 100	2013	3/10/95
25252525	Alan Turing	Student Village A210	2015	2/7/97
33566891	Audrey Chu	300 Main Hall	2014	10/2/96
45678900	Jose Delgado	Student Village B300	2016	7/13/98
66666666	Count Dracula	The Dungeon	2007	11/1431
...	...	...	...	...

- Other data-modeling topics:
  - keys, types, schema, etc.

Example Database

- Data obtained from imdb.com

- Tables with info about:
  - people
  - movies
  - Academy Awards

Unit 1, part III: SQL

- SQL is the *query language* used in relational databases.
- Include fairly advanced topics:
  - joins of two or more tables
  - simple subqueries
  - aggregates, GROUP BY, HAVING
  - outer joins

Unit 1, part III: SQL (cont.)

See our Golden Globes photo gallery, with over 750 images from the big night!

Today - 11:10am | 12:45pm | 1:45 | 3:20 | 4:20 | 6:20 | 7:20 | 9:20 | 10:20

Today - 1:00pm | 4:00 | 7:00 | 10:00

See the entire list of winners of the 68th Annual Golden Globe Awards and all the photos from the big night. We've got images galore from the red carpet, the press room, the parties, and the show itself. Find even more -- including a recap of the Globes show, as well as the Critics' Choice Award winners and photos -- in our Road to the Globes section.

- We answer (or at least explore) questions like:
  - How many of the top-grossing films have won Oscars?
  - Does the Academy discriminate against older women?

Making It Accessible

- Take a gradual approach:
  - start with queries on a single table
    - example: *Get the names of all movies rated PG-13.*
  - then introduce queries that ***join*** two or more tables
    - example: *Get the names of all people who won either Best Actor or Best Actress in the 1990s.*
- Provide hints as needed in the assignments.
- Be judicious in coverage of mathematical underpinnings.

Beyond Relational Databases

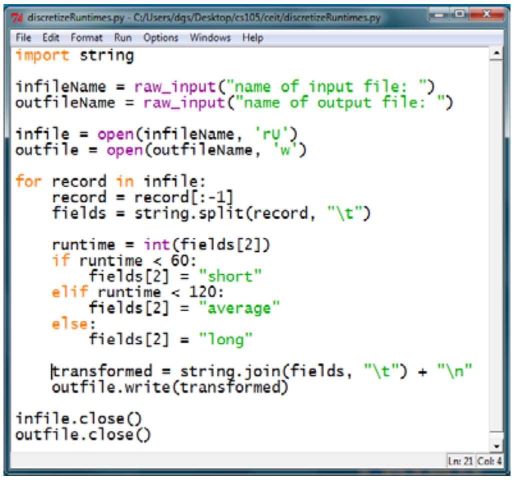
- Example: DNA sequence data

```
>gi|49175990|ref|NC_000913.2| Escherichia coli K12, complete genome
AGCTTTTCATTCTGACTGCAACGGGCAATATGTCTCTGTGTGGATTAAAAAAAGAGTGTCTGATAGCAGCTTCTGAACTGGTTACCTGCCGTGAGTA
AATTAAAATTTTATTGACTTAGGTCACTAAATACTTTAACCAATATAGGCATAGCGCACAGACAGATAAAAATTACAGAGTACACAACATCCATGAA
ACGCATTAGCACCACCATTACCACCACCATCACCATTACCACAGGTAACGGTGCGGGCTGACGCGTACAGGAAACACAGAAAAAAGCCCGCACCTGA
CAGTGCGGGCTTTTTTTTTCGACCAAAGGTAACGAGGTAACAACCATGCGAGTGTTGAAGTTCGGCGGTACATCAGTGGCAAATGCAGAACGTTTTC
TGCGTGTTGCCGATATTCTGGAAAGCAATGCCAGGCAGGGGCAGGTGGCCACCGTCCTCTCTGCCCCCGCCAAAATCACCAACCACCTGGTGGCGAT
GATTGAAAAAACCATTAGCGGCCAGGATGCTTTACCCAATATCAGCGATGCCGAACGTATTTTTGCCGAACTTTTGACGGGACTCGCCGCCGCCCAG
CCGGGGTTCCCGCTGGCGCAA
```

- Common queries involve looking for similarities or patterns.
  - what genes in mice are similar to genes in humans?
  - SQL can't do this!

Unit 2: Programming in Python

- Main goal: to be able to process data stored in text files
- Python makes it easier.


discretizeRuntimes.py - C:/Users/dqs/Desktop/cs105/ceit/discretizeRuntimes.py
File Edit Format Run Options Windows Help
Ln: 21 Col: 4

```
import string

infileName = raw_input("name of input file: ")
outfileName = raw_input("name of output file: ")

infile = open(infileName, 'rU')
outfile = open(outfileName, 'w')

for record in infile:
    record = record[:-1]
    fields = string.split(record, "\t")

    runtime = int(fields[2])
    if runtime < 60:
        fields[2] = "short"
    elif runtime < 120:
        fields[2] = "average"
    else:
        fields[2] = "long"

    transformed = string.join(fields, "\t") + "\n"
    outfile.write(transformed)

infile.close()
outfile.close()
```

Making It Accessible

- Take advantage of Python:
  - simplified syntax
  - list and file processing
- Build on concepts learned in SQL
- Structure the assignments carefully
  - start by modifying an existing program
  - later, write programs similar to ones from lecture
  - provide "scaffolding"

Example Problem Without Scaffolding

Body Mass Index
A person's body mass index (BMI) is equal to the person's weight in pounds, multipled by 720, and then divided by the square of the person's height in inches. 19-25 is the range of healthy BMI values. Write a program that reads a person's weight and height, computes and prints the person's BMI to the nearest integer, and prints a message indicating whether they are below, above, or within the healthy range. You may assume that both inputs are positive.

Example Problem With Scaffolding

Body Mass Index
Body mass index (BMI) is a measure of body fat that is based on a person's weight and height. 19-25 is the range of healthy BMI values. Write a program named bmi.py that can be used to compute a person's BMI, and to determine whether it is below, above, or within the healthy range.

Step 1: The program should begin by getting the following inputs from the user:

- the person's weight, storing it in a variable named weight
- the person's height, storing it in a variable named height

Step 2: The program should then use the values of the variables weight and height to compute and print the person's BMI as a real number using the following formula:

```
             720 * weight
      BMI = ---------------
            height * height
```

…

Unit 3: Data Visualization

- A shorter unit taught by Wayne Snyder
- Based on the work of Edward Tufte
- Principles for creating data graphics that combine:
  - simplicity of design
  - complexity of data
- Show the value that computational tools can add

Unit 4: Data Mining

- The process of finding patterns in data.
  - "hidden knowledge"
  - vs. the "shallow", factual knowledge given by SQL queries
- Data mining applies *machine-learning* algorithms that:
  - operate on a set of *training data*
  - learn some type of *model*

Classification Learning

- One type of machine learning
- Learns a model that can classify/categorize
- Something that human beings have always done!
  - example: how do we learn to identify a dog?

Example: Medical Diagnosis

- Goal: diagnose a patient with cold-like symptoms
  - classify as: Strep throat, Allergy, or Cold
- Sample training data (Roiger & Geatz):

Patient ID#	Sore Throat	Fever	Swollen Glands	Congestion	Headache	Diagnosis
1	Yes	Yes	Yes	Yes	Yes	Strep throat
2	No	No	No	Yes	Yes	Allergy
3	Yes	Yes	No	Yes	No	Cold
4	Yes	No	Yes	No	No	Strep throat
5	No	Yes	No	Yes	No	Cold
6	No	No	No	Yes	No	Allergy
7	No	No	Yes	No	No	Strep throat
8	Yes	No	No	Yes	Yes	Allergy
9	No	Yes	No	Yes	Yes	Cold
10	Yes	Yes	No	Yes	Yes	Cold

- Can you see any patterns that would help the diagnosis?

Example: Medical Diagnosis (cont.)

Patient ID#	Sore Throat	Fever	Swollen Glands	Congestion	Headache	Diagnosis
1	Yes	Yes	Yes	Yes	Yes	Strep throat
2	No	No	No	Yes	Yes	Allergy
3	Yes	Yes	No	Yes	No	Cold
4	Yes	No	Yes	No	No	Strep throat
5	No	Yes	No	Yes	No	Cold
6	No	No	No	Yes	No	Allergy
7	No	No	Yes	No	No	Strep throat
8	Yes	No	No	Yes	Yes	Allergy
9	No	Yes	No	Yes	Yes	Cold
10	Yes	Yes	No	Yes	Yes	Cold

- Different algorithms learn different types of models.

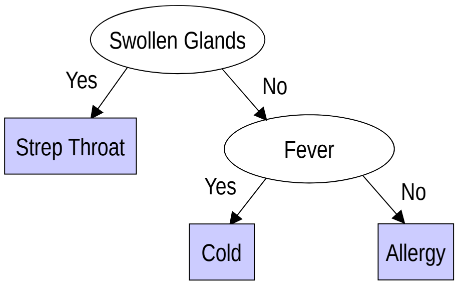
- One possible model is a set of rules:

```
if Swollen Glands == Yes
then Diagnosis = Strep Throat

if Swollen Glands == No and Fever == Yes
then Diagnosis = Cold

if Swollen Glands == No and Fever == No
then Diagnosis = Allergy
```

Example: Medical Diagnosis (cont.)

Patient ID#	Sore Throat	Fever	Swollen Glands	Congestion	Headache	Diagnosis
1	Yes	Yes	Yes	Yes	Yes	Strep throat
2	No	No	No	Yes	Yes	Allergy
3	Yes	Yes	No	Yes	No	Cold
4	Yes	No	Yes	No	No	Strep throat
5	No	Yes	No	Yes	No	Cold
6	No	No	No	Yes	No	Allergy
7	No	No	Yes	No	No	Strep throat
8	Yes	No	No	Yes	Yes	Allergy
9	No	Yes	No	Yes	Yes	Cold
10	Yes	Yes	No	Yes	Yes	Cold
11	**No**	**No**	**No**	**No**	**Yes**	**?**

- Another possible model is known as a *decision tree*:

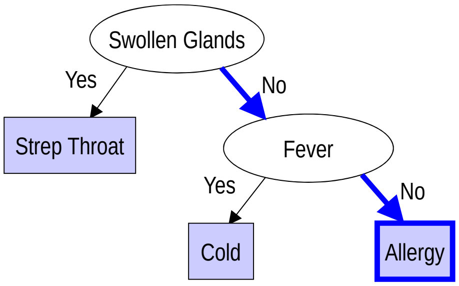


- ***what diagnosis would it give for patient 11 above?***

Example: Medical Diagnosis (cont.)

Patient ID#	Sore Throat	Fever	Swollen Glands	Congestion	Headache	Diagnosis
1	Yes	Yes	Yes	Yes	Yes	Strep throat
2	No	No	No	Yes	Yes	Allergy
3	Yes	Yes	No	Yes	No	Cold
4	Yes	No	Yes	No	No	Strep throat
5	No	Yes	No	Yes	No	Cold
6	No	No	No	Yes	No	Allergy
7	No	No	Yes	No	No	Strep throat
8	Yes	No	No	Yes	Yes	Allergy
9	No	Yes	No	Yes	Yes	Cold
10	Yes	Yes	No	Yes	Yes	Cold
11	**No**	**No**	**No**	**No**	**Yes**	***Allergy***

- Another possible model is known as a *decision tree*:



- ***what diagnosis would it give for patient 11 above?***

Unit 4: Data Mining (cont.)

- Teach two simple classification-learning algorithms
  - students apply them by hand
- Other topics include:
  - two other types of machine learning
  - preparing data for mining
  - assessing the goodness of a learned model
  - the possibility of overfitting
- Introduce students to Weka
  - freely available toolkit for data mining

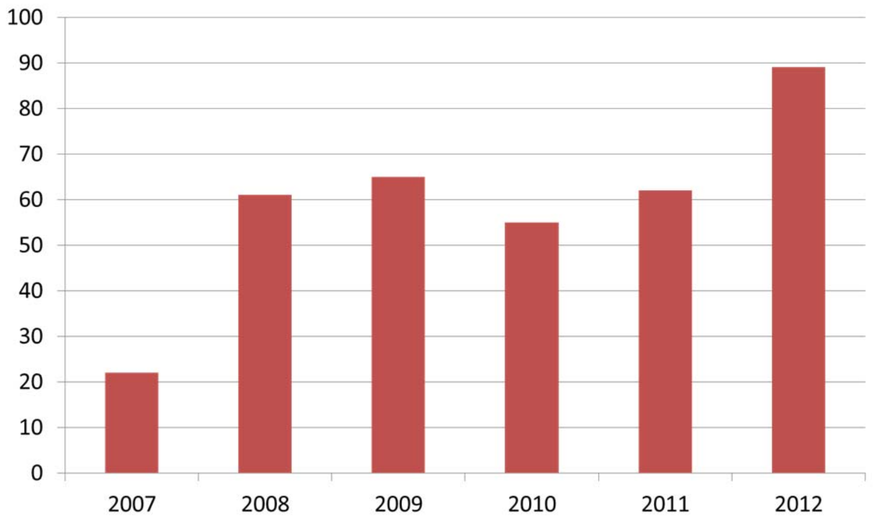
Making It Accessible

- Stick with simple algorithms and straightforward math
- Apply the algorithms to "toy" datasets
  - provides concrete illustrations of the key concepts
- Throughout the course, hold weekly lab sessions
  - hands-on practice, assisted by a TF
- Use the Piazza online learning environment

Student Assessment

- Nine problem sets
- Three 50-minute "quizzes"
- Final exam
- Final project
  - choose a dataset of interest
  - analyze it using techniques learned from the course
  - written report
  - brief in-class presentation
  - work alone or in pairs
- ***Hall of Fame on course website***
  http://cs-people.bu.edu/dgs/courses/cs105/hall_of_fame/

Enrollment



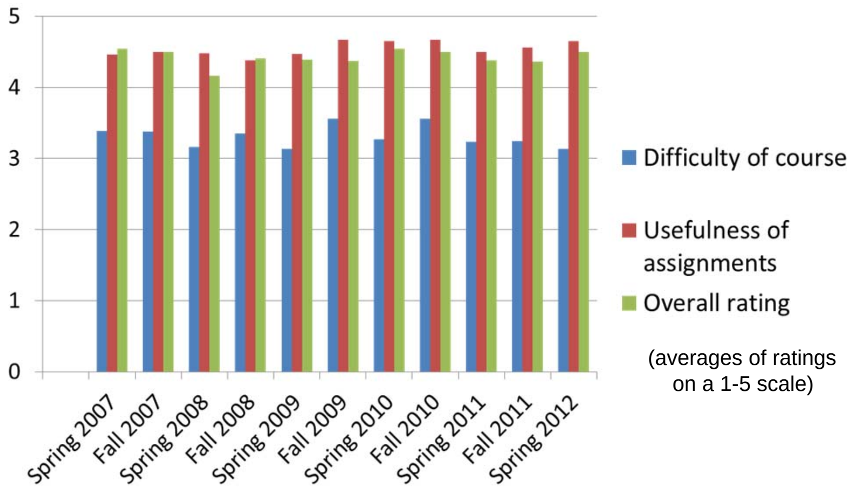
fall and spring semesters combined

Most Common Majors of Enrollees

Major	Number of Students
Economics	32
Business Admin / Management	30
Computer Science	21
International Relations	19
Archeology	16
Mathematics	16
Anthropology	14
Undeclared	13
Political Science	13
English	11

fall 2007 – spring 2011

End-of-Semester Evaluations



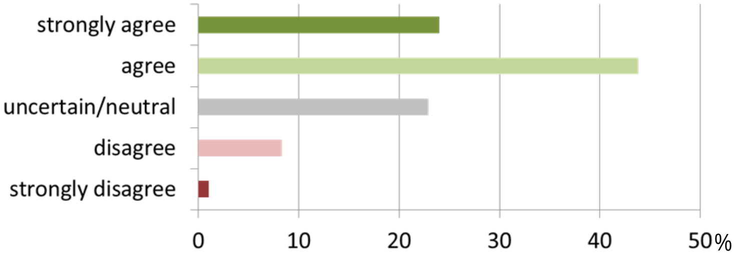
(averages of ratings on a 1-5 scale)

Follow-Up Survey

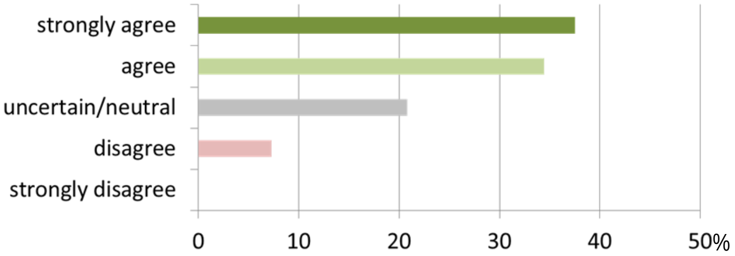
- Online survey of alums of the course from 2007-2011
- Received 96 replies (35% response rate)
- Very positive responses overall

CS 105 Has Been Useful…

- …in subsequent classes

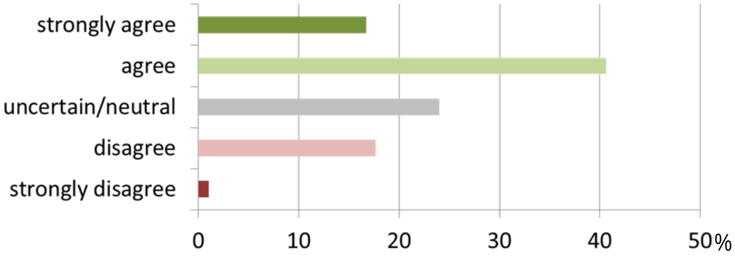


- …in non-academic pursuits like internships and jobs

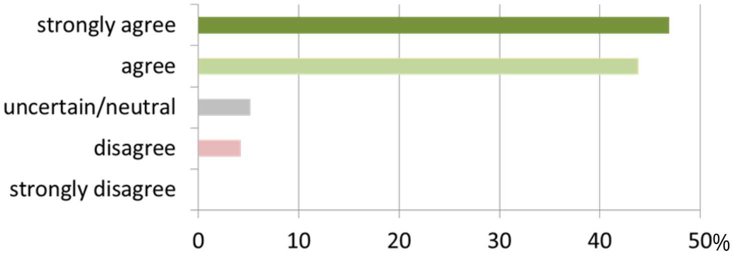


Computer Science Is Relevant to My Pursuits

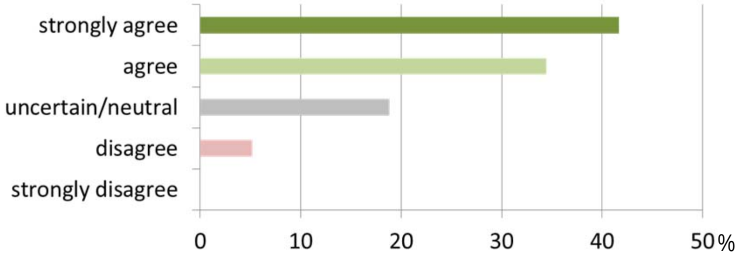
- I thought that ***before*** taking the course.



- I think that now, ***after*** taking the course.



CS 105 Increased My Interest In Learning More About Computer Science



Lessons Learned

- Take steps to keep the material accessible.
  - gradually increase the level of difficulty
  - provide hints/scaffolding in the assignments
- Be willing to experiment and adapt.
  - list of topics
  - sequencing of topics
  - number and difficulty of assignments
- You *can* teach non-majors practical tools for working with data.
  - introduce key concepts at the same time
  - show them a way of thinking and solving problems that underlies much of the modern world