COMS E6111
Advanced Database Systems
Fall 2016

Computer Science Department
Columbia University

Your Instructor: Luis Gravano

- Ph.D. in Computer Science, Stanford U.
- Professor, Computer Science Department
  (at Columbia U. since Fall 1997)
- Senior Research Scientist at Google in 2001
  (on leave from Columbia U.)

- Research interests: Databases, Web Search,
  Information Extraction, Social Media
Class Resources

- Class web page: http://www.cs.columbia.edu/~gravano/cs6111/
- Discussion board: Piazza, which you can access from CourseWorks, at https://courseworks.columbia.edu/
- Announcements from class staff: on CourseWorks

Your Instructor: Luis Gravano

- Various addresses and numbers:
  - gravano@cs.columbia.edu
  - 706 Schapiro CEPSR
  - 1-212-939-7064
- Office hours:
  - Thursdays 2-4 p.m.
  - By appointment by email
Your “Instructional Assistants,” or IAs

- Shubham Bansal
- Terra Blevins
- Rohit Gernapudi
- Aaron Zakem

- All IA office hours are held in the CS IA Room (directions from class web page)
- IA office hours and coordinates will be posted on class web page

Class Information: Prerequisites

- COMS W4111—Introduction to Databases (equivalent courses taken elsewhere are acceptable as well)
- Fluency in Java or Python

You need permission from the instructor if you don’t have the prerequisites.
Note that COMS W4112 is not a prerequisite.
Class Information: Lectures

- Fridays, 2:10-4:00 p.m.
- 1127 Mudd

Grading Information

- **Midterm exam** (Fri Oct 21, in class): 25%
  Covers all lectures and required readings; closed book
- **Final exam** (Fri Dec 16, 1:10-4:00 p.m.): 25%
  Covers all lectures and required readings; closed book; **not** cumulative
- **Projects (3)**: 50%
  - In Java or Python (student’s choice)
  - All projects are equally weighted
Grading Information (cont.)

- Median grade will be a B+ or slightly higher.
- Alternative or make-up exams will not be given.

Grading Information (cont.)

- To be fair to all students in the class, I will grant no extensions or exceptions for project submission.
- Instead, you have three grace late days total for projects that you can use as you wish throughout the semester. Weekends and university holidays are not counted.
- After using all grace days, you will get a 25% grade deduction for each additional late day.

Check full details on web site.
Projects: In Groups of 2

- You get to choose your group partner.
- Further group-related information will be given with first project.
- You will need one **CS** account per group. You can open one ($50 fee) from:


Collaboration Policy

- Please check “Lateness and Collaboration Policy” page from the main web page for the class.
- Exams are to be done individually. Projects are to be done in teams, and different teams cannot collaborate with each other.
- We will not tolerate cheating. Check the **CS** Department policies and procedures regarding academic honesty at:
  http://www.cs.columbia.edu/education/honesty
  They fully apply to this course.
- Contact the instructor right away if you have any questions.
Ongoing Feedback

- Don’t wait until the end-of-semester course evaluations to complain or give feedback on how to improve the course. (It’s too late then!)
- Come see me early on during my office hours or send me email with your concerns and suggestions, or use the IAs to forward them to me.

Topics Covered

- Information Retrieval
- Web Search
- Information Extraction
- Data Mining
- Data Warehousing, OLAP, Decision Support
- Time Series Mining
- Spatial Data Management
- …
Information Retrieval:
Text Databases

• Objects are text documents
• User queries are usually less “precise” than in the relational world

Examples:
• The archive of a newspaper
• A web search engine

Answering Queries?

Key issue:
Ranking documents in order of expected relevance for a given query
Indexes?

Inverted files:

• “ramakrishnan”: doc_1, doc_3, ...
• “databases”: doc_1, doc_4, ...
• ...

Web Search Engines

• How do they work?
• What are their latest tricks and trends?
• How can we exploit the link structure of the web?

Subject of first project
Data Mining

Goal:
To find interesting trends or patterns in large datasets

Examples:
• Identify target customers for junk mail
• Decide what to place next to beer on supermarket shelves
• Recommend products to online shoppers

Multidimensional DBs, Decision Support, OLAP

Data Warehouse:
• Large repository of data about an organization (~ a view)
• Complex queries common
• Goal: support of OLAP, data mining, ...
OLAP: On-Line Analytic Processing

• Multidimensional data:
  Product, Location, Time: Sales value
• Complex queries, with group-by and aggregate operators:
  Total Sales, Total Sales per City, Total Sales per State
• Dimensionality reduction, roll-up & drill-down, slice, dice, ...

Required Readings
(No Textbook Required)

• Mostly papers/material available on line
• Occasionally, chapters from the cs4111 textbook, on reserve in Science and Engineering Library:

Also, some chapters from:
• Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze: Introduction to Information Retrieval, Cambridge University Press, 2008
  Available electronically at http://www.mmds.org/