ACM Data Science Task Force Course Example

DA 210/CS 181: Data Systems

Denison University, Granville, OH 43023
Thomas C. Bressoud
bressoud@denison.edu

Materials are provided as part of a set of resources allied with a textbook created for the course:
https://datasystems.denison.edu

Note that these materials are being assembled and organized as an ongoing effort that is likely to continue to the end of the calendar year (2020).

Knowledge Areas that contain competencies (knowledge, skills, and dispositions) covered in the course

Knowledge Area	Total Number of Contact Hours
Data Acquisition, Management, and Governance (DG)	24
Computing and Computer Fundamentals (CCF)	9
Programming, data structures and algorithms (PDA)	8
Analysis and Presentation (AP)	1

Where does the course fit in your undergraduate Data Science curriculum?
This course is required for the Data Analysis major. It is the second of the required courses coming from the computer science department, and has a single prerequisite of the introductory computer science course. It is commonly taken in the sophomore year. As a required course, all majors take it, which amounts to 35 or 40 students per year, in sections capped at 24 students.

Is this course from or used in other curricula/majors?
This course was designed to fit both the Data Analytics major as well as to give an early experience exposure to data systems, management, and "data-aptitude" for computer science majors and minors and is also required for CS majors and minors.

What is covered in the course?
This course provides a broad perspective on the access, structure, storage, and representation of data. It encompasses traditional database systems, but extends to other structured and unstructured repositories of data and their access/acquisition in a client-server model of Internet computing. Also developed are an understanding of data representations amenable to structured analysis, and the algorithms and techniques for transforming and restructuring data to allow such analysis. The course focuses on two

dimensions: the forms of data, conveyed through the framework of data models (structure, operations, constraints), and the sources of data, from local filesystems to remote relational databases, to web servers and web scraping, to provider APIs over HTTP.

What is the format of the course?
The class is face-to-face with four 50-minute class sessions per week, with each class session using an integration of lecture with the use of notebooks to encourage engagement and hands-on activity.

How are students assessed?
There are many homework assignments, with 2 to 3 per week, each consisting of programmatic application of the ideas through Jupyter Notebook questions, many of which allow feedback through assertion testing cells and autograding through the nbgrader framework. There are also typically two projects that each focus on a data model and a synthesis project at the end of the term that brings it all together. We use 3-4 test assessments over the semester. Students are expected to spend approximately 10 hours of work outside of class in the completion of their assignments.

Course tools and materials

- Textbook: **Introduction to Data Systems: Building from Python**, Thomas Bressoud and David White, ISBN 978-3-030-54371-6. Springer-Nature. https://www.springer.com/us/book/9783030543709. This is a textbook written specifically for this class, with final manuscript submitted and expected availability around January 1, 2021.
- The course uses Python and Jupyter Notebooks installed via Anaconda.
- For the unit on relational databases, we use both SQLite and MySQL.
- Data sets used include open data from various sources as well as a relational database based on anonymized school/class data from Denison and approved for distribution through the Denison Data Governance committee.

Why do you teach the course this way?
This course was designed this way because of the confluence of working with the Park City Math Institute 2017 workshop for data science curricula, a revision in our CS major, and an external review that advocated for data-centric and systems topics early in the CS curriculum, along with Denison's introduction of a Data Analytics major. We decided that to become proficient as *users* of data systems, students could be introduced to such topics early in their careers and could then use the knowledge and skills in downstream projects, internships, and research experiences.

Body of Knowledge coverage

Total Contact Hours: 200 min/week x 14 weeks = 46 total hours

KA	Sub-domain	Competencies Covered	Hours
DG	Data Acquisition	Sources of data; pull based approaches	5
DG	Information Extraction	Extracting structured from unstructured	2

DG	Working with various types of data		3
DG	Data Integration		2
DG	Data transformation	Pipeline; standardization; normalization; encoding	4
DG	Data cleaning		5
DG	Data privacy and security	Authentication and Authorization for data systems clients to access protected resources owned by others (OAuth)	3
CCF	File Systems	Organization; directories; access and processing	2
CCF	Networks	Layered network structure; Basic protocols, HTTP	4
CCF	The Web and Web programming	Web programming languages; web scraping	3
PDA	Algorithmic Thinking and Problem Solving		2
PDA	Programming	ADTs and Object-oriented packages appropriate for data science	3
PDA	Data Structures	Native 2D data structures for tabular data representation; ADTs	3
AP	Visualization	Software; dashboards; chart types	1

Additional topics
[List notable topics covered in the course that you do not find in the ACM Draft 2 Computing Competencies for Undergraduate Data Science]

Other comments
[optional]