Proposal 1: New Introductory Data Science Course

Proposal: Introduce a new course at the 300 level, called “Introduction to Data Science”, that covers basic data management (SQL, MapReduce) and basic stats/ML, with a focus on usage and functionality.

Why: With increased emphasis on data and deriving knowledge from data, it would be useful for our students to know this basic material. Many students would be better served by a course that covers data science end-to-end, but without going into as much depth as 422 or 424. Such a course would also allow us to cover more material in 424 (Databases) and 422 (Machine Learning), both of which would have this as a pre-requisite. At the same time, it is expected that this course will cover sufficient topics in data science/analytics for many students, without taking the other two courses.


Grading: 6 programming assignments (30%), 4 written homeworks (20%), midterm + final (50%). To simplify the labs, a virtual image with all the software and skeleton code preloaded would be provided.

Pre-requisites: Same as CMSC330/CMSC351 (i.e., CMSC250 and CMSC216).

Draft Syllabus: Introduction to Data Science Course

Introduction to the class (1 lecture)
https://github.com/umddb/datascience-fall14/blob/master/lecture-notes/intro.md

Python primer (1 lecture)
The main programming language for the class would be Python, since many of the relevant libraries are in Python (including Pandas, matplotlib, scikit etc). Only very basic syntax/features of the language would be needed for the class.

Basic Statistics, Probability Distributions, Hypothesis Testing, Fallacies (2 lectures)
https://github.com/umddb/datascience-fall14/blob/master/lecture-notes/stats1.md
Material from early chapters in MMD = Mining Massive Datasets (Ullman et al.)
Lab in Python + Matplotlib
ML/Stats/Data Mining Algorithms: Regression, Clustering (K-Means), Decision Trees, Association Rules, Finding Similar Items, ... (6-8 lectures)

What is the goal for each of these, and how to interpret the models/results. Some real-world applications. Basic naive algorithms for some of those.

Lab on SciKit or some other package like Shogun

Visualization using D3.js, maybe some other Javascript stuff, Web applications, Building Webapps using Flask or some other MVC (1-2 lectures)

Not really central to DS, but useful for students to know

Instead of doing web-app stuff, could cover infoviz in some more detail

Lab using D3.js and/or some Python MVC

Data Models, Relational Databases, SQL, Python Pandas (3-4 lectures)

Why data models are needed. Brief overview of E/R, Relational, XML, and JSON data models. Notion of data independence and its importance. Basic SQL. Equivalence between Pandas and SQL.

Lab on Databases/SQL and Python Pandas

Data scraping and wrangling, Unix tools, Data cleaning Issues (2-3 lectures)

Key steps in going from unstructured, noisy, incomplete data to well-structured data that can be analyzed.

https://github.com/umddb/datascience-fall14/blob/master/lecture-notes/wrangling.md

Lab on Unix tools and Python Data Wrangling Functionality

Map-Reduce Framework (4 lectures)

What is MR. How to write MR programs. Work through several examples, including several of the ML algorithms covered earlier. Here the focus is on understanding how to write data-parallel programs for distributed execution (and not on any specific MR implementation like Hadoop or Spark).

Lab using Spark

Some Implementation details (4 lectures)

Goal is to convey some overall issues in implementation so that the students know how to reason about the performance.

Basic Relational Database Operators: How to implement Joins, and Aggregates

Scaling up and scaling out: differences and challenges

Map and reduce operators -- parallel execution of those and overall framework

NoSQL Key-Value Stores, Consistency Issues (1 lecture)

This could be covered along with Web Apps
FAQ

When would the new course be taught for the first time?
Our goal is Spring 2016.

Would students have the background to take the course w/o having taken 330 or 351?
We believe so. Similar courses are often taught as service courses to non-CS majors, who also wouldn’t have taken those courses. Another example: Washington’s course on data management only has our 250 as a pre-req, and it covers all the data management topics listed here (and more). [http://courses.cs.washington.edu/courses/cse344/](http://courses.cs.washington.edu/courses/cse344/)

Would we need to change our major requirements?
No. Our official major requirements are that, students must have 27 credit hours at the 300-400 levels, which must include 330, 351, and at least 15 at the 400 level. The students can use the new 3xx course toward the remaining 6 credit hours. [http://www.umd.edu/catalog/index.cfm/show/content.section/c/1/s/103](http://www.umd.edu/catalog/index.cfm/show/content.section/c/1/s/103)
The description on our webpage is slightly simplified to reflect the current offerings, and would need to be adapted.

Why not introduce the course at the 400-level?
Several reasons: with data analysis increasingly becoming a key piece in CS, it would be useful to expose the students to the basic principles of data management and analysis, such as data modeling, databases/SQL, basic stats algorithms, large-scale analysis, etc., as early as possible. This course will also be a pre-requisite for 424 and 422 (and potentially other courses) -- the pre-req chain for those courses would become too long if this course were to be at 400-level (and have 330 and 351 as pre-requisites).