

Placeholder syllabus (class in development)

Data Wrangling: From Excel to SQL and Beyond.  
IST 385T Special Topics class with Unique # 27800, Spring 2015  
Instructor: James Howison.  
Wed and Fri 10:30-12. (Yes, 2 x 1.5 hour sessions!)

This coming semester I'm going to be teaching an alternative to Database Management that I've called "Data Wrangling." It's a different course, but this semester at least, you won't want to take it if you've already taken Database Management.

tl;dr comparison of Database management and Data Wrangling:

Database management (1 x 3hr/week):

- data modeling and sql queries
- php to display/edit database data on webpages

Data Wrangling (2 x 1.5hr/week):

- data modeling and sql queries
- python to import/export excel/csv to/from database
- managing data processes

**Both are great if you have no previous experience with databases or programming.**

How to choose:

The current Database Management class structure is probably better if you are heading for web development, the Data Wrangling perhaps better if you see yourself handling data and data processes in organizations, including making reports with graphs etc. Both courses will provide a gentle introduction to basic programming.

**Longer version:**

Database management to date has been: database design or data modeling, learning sql queries and then trying to learn sufficient php to make web pages to display/edit data in the database. The use cases are small, from scratch, designed databases, only accessed from the web pages. I've also taught it in our standard 3 hour class sessions. This semester I'm changing it up. First, I'm replacing about half of the material, and second, I'm teaching it in 1.5 hour classes twice a week (Wed and Fri mornings).

Second, I'm itching to teach more of what I think people encounter in the workplace: lots of data scattered around in various places (typically a mix of text and excel files and a few different database schemes, usually managed by a few different people). The challenges, then, are a mix of understanding existing schemas and processes

and figuring out how to keep everything running while simultaneously bringing some higher level order to things and managing the questions of power and compliance that come up.

So we'll cover database design and sql queries, as in the regular Database Management course. But we'll also cover how to handle and design basic data processes, e.g., how to have multiple people entering data into a database? How can I deal with a boss that only wants things in excel? Why do excel people love pivot tables so much? What are reporting tools like Crystal Reports or Microsoft SRSS anyway?

As far as the programming is concerned, we won't learn php nor will we build web-pages or forms to edit databases. What we'll cover instead is enough python to read in, and export, excel, csv, and text files (even when they don't align with our database design). That means we don't have to worry about html, CSS, and forms, and can concentrate on the basic elements of talking to databases with scripts (including useful stuff like splitting fields with "regular expressions"). I'm still finalizing the task for the final projects, but I think it will be to take a group of disparate files and existing databases, write code to bring them together in a database, and produce a report that mixes text, tables, and graphs, drawn from the data, all working as a workflow, so that changes in the source files show up in the report.

Why twice a week? I find that, for technology heavy classes, three hours is a bit mind-numbing; we learn much more in the first half than the second. So I'm going to experiment with having two "first halves" :) You'll notice that the class is scheduled "inside" our 3 hour blocks (9-12), so that does mean that if you take this class you can't take a class Wed 9-12 nor Fri 9-12.

Honestly, the "wrangling" thing isn't UT Austin specific, I swear! I will minimize Bevo and cowboy jokes in general, I promise.