# *Class Schedule Fall 2012* ***(Revised after Hurricane Sandy)***

| **Class Number** | **Date** | **Topics** (tentative) | **Readings** | **Deliverables** |
|---|---|---|---|---|
| 1 | Sept 19 | **Introduction**<br>Data science practice, Class objectives, Detailed example, Boring syllabus details, getting started lab | Material sent out via email | install python; know python to the level of "The Hard Way" |
| 2 | Oct 3 | **Data Representation I**<br>Objects, relationships, and information about them. File representations and file manipulation. XML, JSON, YAML. Reading/writing/transforming data. Command-line processing. Finding specific information: querying, filtering, regular expressions. Cleaning data.<br>*Example: extract specific information from a massive text file* | *handed out*<br>+<br>Book Ch. 1 | **HW#1 Due** |
| 3 | Oct 10 | **Data Representation II**<br>Relational databases, Key-value (noSQL) databases. Structured data: information architecture, ER representations, table structure. SQL. Indexes. NoSQL databases.<br>*Example: get data from a non-trivial external database, process locally to derive insight.* | *handed out*<br>+<br>Book Ch. 2 | **HW#2 Due** |
| 4 | Oct 17 | **Big Data & APIs**<br>Distributed file systems, Map/reduce, Hadoop (what's that? when is it useful?). Related big data technologies/platforms: Pig, HBase<br>Programmatic access to get (and post) data.<br>*Example: get data from web source, process locally, visualize using visualization API.* | *handed out* | **HW#3 Due** |

| Class Number | Date | Topics | Readings | Deliverables |
|---|---|---|---|---|
| 5 | Oct 24 | **Predictive models I.** Data representation for predictive modeling, models, data-driven-model applications. Evaluating models, metrics for model quality.<br>*Example: Evaluate several predictive models on "test" data, compute metrics, visualize results, compare methods. How to decide which model is best?* | *handed out*<br>+<br>Book<br>Ch.3 & 4 | **Take home assessment #1 due** |
| 6 | Oct 31 | ***CANCELLED -- HURRICANE SANDY*** | | |
| 7 | Nov 7 | **Systems and Experimentation**<br><br>Guest speaker<br>**Jason Davis**, Data Scientist at Etsy, Entrepreneur, Cowboy | *handed out*<br>+<br>Book Ch. 5 | **HW#4 Due** |
| 8 | Nov 14 | **Predictive models II**. Learning models from data. Training. How does that work for selected models? Overfitting, holdout evaluation, cross-validation, overfitting avoidance.<br>*Example: <u>build</u> predictive models from data; then apply evaluation methods from PM I, plus overfitting analysis.* | *handed out*<br>+<br>Book Ch. 6 | **Get projects in order** |
| 9 | Nov 28 | **Predictive models III** (topics continued)<br><br>+<br><br>**Making Data-driven Recommendations**<br>**Guest speaker:**<br>**Chris Volinsky, Director, Statistics Research**<br>**AT&T Research**<br>**Winner: $1Million NetFlix Prize** | *handed out* | **HW#5 Due** |
| 10 | Dec 5 | **Data Visualization**<br><br>**Guest speaker: Kristen Sosulski, Stern** | *handed out* | **Take home assessment #2 due** |
| 11 | Dec 12 | **Data Science Issues in Online Advertising**<br>**Guest speaker:**<br>**Claudia Perlich, Chief Scientist, M6D**<br><br>*display ad targeting, bid optimization, and more..* | *handed out* | |
| 12 | Dec 19 | **Wrap up** | | **Project due** |