1. Course Overview

This course will change the way you think about data and its role in business.

Businesses, governments, and individuals create massive collections of data as a by-product of their activity. Increasingly, decision-makers rely on intelligent technology to analyze data systematically to improve decision-making. In many cases automating decision-making processes is necessary because of the volume of data and the speed with which new data are generated. The course is suitable for those interested in working with and getting the most out of data, as well as those interested in understanding data mining from a strategic business perspective.

We will examine how data analysis technologies can be used to improve decision-making. We will study real-world examples and cases, from finance, marketing, operations, and electronic commerce, to place data-mining techniques in context, to develop data-analytic thinking, and to illustrate that proper application is as much an art as it is a science. In addition, we will work “hands-on” with data mining software.

Specifically, the goal of this course is three-fold; after taking this course you should:

1. **Approach business problems data-analytically.** Think carefully & systematically about whether & how data can improve business performance.

2. **Be able to interact competently on the topic of data mining for business intelligence.** Know the basics of data mining processes, algorithms, & systems well enough to interact with CTOs, expert data miners, and business analysts. Be able to envision data-mining opportunities.

3. **Have had hands-on experience mining data.** Be prepared to follow up on ideas or opportunities that present themselves, e.g., by performing pilot studies.
2. Instruction Method
This is primarily a lecture-based course, but student participation is an essential part of the learning process in the form of active technical and case discussion. The course will explain with real-world examples the uses and some technical details of various data mining techniques. The emphasis primarily is on understanding the application of data mining techniques, and secondarily on the variety of techniques and the mechanics of how they work. Each class session has materials you must read prior to class. You should be prepared to be called on to discuss the readings. You are expected to attend every class session, and to arrive prior to the starting time.

Homework Assignments
There will be a total of seven assignments, each comprising questions to be answered and some including hands-on tasks. The last assignment will be done in your teams (discussed below). Completed assignments must be handed in prior to the start of the class on the due date. If submitted by email, they must arrive at least one hour prior to the start of class. They will be graded and returned promptly.

The hands-on tasks will be based on data that we will provide. You will mine the data to get hands-on experience in formulating problems and in using the various techniques discussed in class. You will use these data to build and evaluate predictive models. The final assignment will include a “competition”: one part of the data will be held back to evaluate the models you mine.

For the hands-on assignments you will use the (award-winning) toolkit Weka, part of the Pentaho open source business intelligence suite:

http://www.cs.waikato.ac.nz/ml/weka/
http://www.pentaho.com

A demonstration of Weka will be given in class. In order to use Weka you must have access to a computer on which you can install software. If you do not have such a computer, please see me immediately so we can make alternative arrangements. The first hands-on assignment will be very easy, ensuring that you can install the software and get it running, before moving on to more challenging assignments.

Term Project
A term project report will be prepared by student teams. Student teams should be of 4 people. You should decide on your teams and submit them to me by the end of the third class.

Teams are encouraged to interact with me and with the TA electronically or face-to-face in developing their project reports. You will submit a pre-proposal for your project just after mid-term. Selected teams will present their projects in the last class meeting. We will discuss the project requirements and presentations in class.

Final Quiz
The scope and format of the final quiz will be discussed in class.
4. Requirements and Grading
You should attend all class sessions—the sessions build on previous discussions.

Answers to homework questions should be well thought out and communicated precisely. Homework reports should be professional, as if you were submitting them to a boss or potential client. Points will be deducted for sloppy language/figures/diagrams and irrelevant discussion.

The points to be addressed in the term-project analysis will be discussed in class. The material needed for the term project will be handed out during the term. The analysis should be about 20 double-spaced pages.

The grade breakdown is as follows:
1. Homeworks (7): 30%
2. Term Project (1): 30%
3. Participation and Class Contribution: 15%
4. Final Quiz: 25%

Late Assignments
Assignments are due prior to the start of the lecture on the due date. Turn in your assignment early if there is any uncertainty about your ability to turn it in on the due date. Assignments up to 24 hours late will have their grade reduced by 25%; assignments up to one week late will have their grade reduced by 50%. After one week, late assignments will receive no credit.

5. Communication, Text, etc.
The Blackboard site for this course will contain lecture notes, reading materials, assignments, and late breaking news. It is accessible via: http://sternclasses.nyu.edu/

Post questions regarding course content to the Blackboard site (unless you are uncomfortable doing so) so that others can benefit from the answers. You are encouraged to contribute by answering/following up on posted questions.

Readings:
1. Textbook:
   *Data Mining Techniques, Second Edition*
   by Michael Berry and Gordon Linoff, Wiley, 2004
   - available from bookstore
   - available from Amazon (usually; usually less expensive)
   - available as ebook for free: http://site.ebrary.com/lib/nyulibrary

2. Supplemental readings posted to blackboard or distributed in class. Note that some of these readings are accessible for free only from an NYU computer. If you can’t access a link from home, please try it from school.

Please keep in mind the Stern Honor Code
http://www.stern.nyu.edu/mba/studact/mjc/hc.html
# Class Schedule from Spring 2008

*Topics, structure & deliverables may change for 2009*

<table>
<thead>
<tr>
<th>Class Number</th>
<th>Date</th>
<th>Module</th>
<th>Topics</th>
<th>Examples</th>
<th>Book Sections</th>
<th>Homework due dates</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Tuesday</td>
<td>Introduction</td>
<td>What is DM?, why DM now?, DM process, relation to other BI techniques, different data mining tasks</td>
<td>Retailing (Walmart); Customer retention (churn/attrition control)</td>
<td>Ch. 1 &amp; 2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>February 12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Tuesday</td>
<td></td>
<td>How do I produce a useful segmentation? what is a model?, basic terminology, predictive modeling, classification, regression, tree induction, class-probability estimation</td>
<td>Consumer credit default prediction (simplified)</td>
<td>Ch. 4 pp. 116-120 Ch. 6 pp. 165-194, 209</td>
<td>HW#1 due</td>
</tr>
<tr>
<td></td>
<td>February 19</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Tuesday</td>
<td>Predictive Modeling</td>
<td>How do I know my model is any good? evaluation, in-sample versus out-of-sample, overfitting, cross-validation, domain knowledge validation, geometric interpretation, linear model versus tree induction, logistic regression, toolkit demo</td>
<td>Fraud detection</td>
<td>Ch. 3 pp. 43-54</td>
<td>HW#2 draft due Team Lists due</td>
</tr>
<tr>
<td></td>
<td>February 26</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Tuesday</td>
<td></td>
<td>knowledge-engineering bottleneck, rule-based systems, knowledge in action, evaluation, error costs</td>
<td>Operations Support; Credit Risk</td>
<td>Ch. 3 pp. 54-86</td>
<td>HW#2 due</td>
</tr>
<tr>
<td></td>
<td>March 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Tuesday</td>
<td></td>
<td>Bayesian &amp; memory-based reasoning, nearest neighbors, variable normalization, text classification, “naïve” Bayes</td>
<td>Spam filtering</td>
<td>Ch. 8 pp. 257-271</td>
<td>HW#3 due</td>
</tr>
<tr>
<td></td>
<td>March 11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

March 18  

SPRING BREAK
<table>
<thead>
<tr>
<th>Class Number</th>
<th>Date</th>
<th>Module</th>
<th>Topics</th>
<th>Examples</th>
<th>Book Sections</th>
<th>Homework due dates</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>Tuesday March 25</td>
<td>Predictive Modeling (cont.)</td>
<td>variable selection, feature engineering, neural networks</td>
<td>Spam revisited; Image Classification; Corporate default prediction</td>
<td>Ch. 7 pp. 211-243</td>
<td>Project pre-proposal due</td>
</tr>
<tr>
<td>7</td>
<td>Tuesday April 1</td>
<td>Descriptive/Unsupervised Data Mining</td>
<td>unsupervised/descriptive data mining, unsupervised algorithms, associations, clustering</td>
<td>Whisky clustering; e-commerce; Consumer credit management revisited (GE Capital)</td>
<td>Ch. 9 Ch. 11</td>
<td>HW#4 due</td>
</tr>
<tr>
<td>8</td>
<td>Tuesday April 8</td>
<td>Data Mining and Electronic Commerce</td>
<td>DM and competitive advantage, DM and ecommerce, recommender systems, collaborative filtering, and other e-com applications</td>
<td>Amazon.com; Farecast; NetFlix</td>
<td>Ch 8 pp 282-285</td>
<td>HW#5 due</td>
</tr>
<tr>
<td>9</td>
<td>Tuesday April 15</td>
<td>Ethics of Data Mining and Privacy</td>
<td>ethics of data mining, privacy, what can/do firms know?, what should they do?</td>
<td>Targeted Marketing</td>
<td>Ch. 4 pp. 87-110 (skip pp.90-93)</td>
<td>HW#6 due</td>
</tr>
<tr>
<td>10</td>
<td>Tuesday April 22</td>
<td>Professor Vasant Dhar, formerly of Morgan Stanley</td>
<td>data mining in finance, genetic algorithms</td>
<td>More examples from Data Mining in Finance</td>
<td>Ch. 13</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Tuesday April 29</td>
<td></td>
<td>wrapup and review</td>
<td>revisit pp.60-64 &amp; 233</td>
<td>2 things due: project reports, HW#7</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Tuesday May 6</td>
<td></td>
<td>project presentations and competition results</td>
<td></td>
<td></td>
<td>Final Quiz: Take Home</td>
</tr>
</tbody>
</table>