**Information Sheet**

**Summary and Goals:** Data that has relevance for managerial decisions is accumulating at an incredible rate due to a host of technological advances (big data!). Electronic data capture has become inexpensive and ubiquitous as a by-product of innovations such as the Internet, e-commerce, electronic banking, point-of-sale devices, bar-code readers, microarrays, genomic sequencing, and intelligent machines. Such data is often stored in data warehouses and data marts specifically intended for management decision support. Data mining is a rapidly growing field that is concerned with developing techniques to assist managers to make intelligent use of these repositories. A number of successful applications have been reported in areas such as credit rating, fraud detection, database marketing, customer relationship management, stock market investments, and bioinformatics. The field of data mining has evolved from the disciplines of statistics (multivariate analysis) and artificial intelligence (machine learning).

This course will examine methods that have emerged from both fields and proven to be of value in recognizing patterns and making predictions from an applications perspective. We will survey applications and provide an opportunity for hands-on experimentation with algorithms for data mining with easy-to-use software and cases.

Our objective is to develop an understanding of the strengths and limitations of popular data mining techniques and to be able to identify promising business applications of data mining. Students will be able to actively manage and participate in data mining projects that have been converted into cases. A useful takeaway from the course will be the ability to perform powerful data analyses in JMP PRO as well as other data-mining systems.

**Background:** Material on statistics at the level of 15.060 (Data, Models, and Decisions) or 15.074 (Statistical Reasoning and Data Modeling) or 15.075 (Statistical Thinking and Data Analysis) or by permission. Perhaps the most important topic is regression and you might want to review your notes.

**Instructor:** Professor Roy Welsch, E62-564 (x3-6601), rwelsch@mit.edu. Often, I am available after class, but the best way to see me is to schedule some time by email or phone. Please also feel free to email me with your questions or comments. I will do my best to respond in a timely manner. We will use Stellar for communication with the whole class.

**Establishing a Stellar Account:** Check first to see if you are already registered on Stellar for 15.062. If not, a message will be generated for us to grant you permission.

**Teaching Assistant:** Helen Wanqin Xie (hwx@mit.edu) who is Ph.D. student in Physical Chemistry at MIT. Office hours will be announced.

**Course Assistant:** Jennifer Challis, E62-571 (x4-4378), jchallis@mit.edu will have extra copies of handouts not posted on the website and can often get a message to me.

Lectures: These will be held in E51-345 from 4:05 to 5:25 on Mon. and Wed. I will use PowerPoint slides during lecture to provide an outline of what I want to cover. These will be available on Stellar (two slides per page) and will also be handed out (four slides per page) so that you can take notes during lecture. It helps if you skim the assigned textbook material before lecture in order to have some idea of what is coming even if you don't understand everything. Please ask questions as I go along.

Recitations: These will be held in room E51-315 on Tuesdays from 4:05-4:55. Generally the recitations will be conducted by the Teaching Assistant and cover some new material related to data analysis and statistical computing, e.g., JMP PRO. There will also be time to discuss homework problems, examples, and clear up any confusion from my lectures.

Exams: None. Grades based on homework (cases), participation, and project.

Term Project: A term project will be required. This usually involves exploring data with the methods in the course (best if you are interested in the data, but see the resources listed below) or picking some material outside of the book that we do not plan to cover and demonstrating that you have gained a working knowledge of it. Computing algorithms can be appropriate. More theoretical issues may also be addressed. The report should be about ten pages with additional material (e.g., computer output and extra plots) included as appendices. Note that a one-page project proposal is due on Nov. 28.

Homework: There will be homework about every ten days that will be graded and returned. (Sampling may be used, i.e. only a portion of the problems may be graded and the rest will be just counted. However, solutions will be provided to all of them.)

Grading: Our goal is to have everyone learn the material. If you are having problems, don't let them slide until the end. Talk to us. The homework will count 45%, class participation 10%, and the project 45%. Once we have handed out the solution sheet for a homework set, late homework will not be accepted.

Work Load: This is a 4-0-8 course for half a semester for a total of 6 units. We will have three main hours of lecture and one additional hour of recitation or demonstration each week. Homeworks should take the median student about 8 hours each week. If we have misjudged this load (most often because computing can sometimes take more time than we think), please let us know.

Feedback: Let me (or the TAs) know (anonymously, if you wish) what is going right and what is going wrong with lectures, homework, content, etc. I will occasionally invite a random sample of you to talk to me about the course and/or to fill out evaluation forms during the course.

Academic Honesty: It is best to attempt the homework on your own and then ask us questions. In a pinch, talk to your classmates for clarification. What goes on your homework paper should be your own work. The project should, of course, be entirely your own work. Please see the
statement about MIT Sloan Academic Standards posted on the 15.062 Stellar site for further
details.

Computing: We will be using a statistics and data mining package called JMP PRO 13 that is
available free at http://ist.mit.edu/sas-jmp/license. Matlab (see below) is a possible substitute. R
(see below) is another alternative, but takes a while longer to learn. A nice interface for R is
called RStudio. We will only provide limited support for Matlab and R during the recitations.

Matlab (with the statistics toolbox plus other data mining toolboxes such as neural networks and
bioinformatics) can be obtained free at http://ist.mit.edu/software-hardware. R can be obtained
from http://www.r-project.org. This package is also free and can be used after you leave MIT.
RStudio can be obtained from https://www.rstudio.com.

Data for Projects:

Here are some places to get datasets for projects and other uses.

1. KDNuggets http://www.kdnuggets.com/datasets/index.html. For a great deal of additional
   information try the core site at http://www.kdnuggets.com.

2. UCI http://www.ics.uci.edu/~mlearn/MLRepository.html with summary descriptions at

3. DASL http://lib.stat.cmu.edu/DASL/Datafiles/. Other data sets and software may be found
   at the core link http://lib.stat.cmu.edu/.


5. Kaggle contest datasets http://www.kaggle.com

10/10/2016v1
Fall 2016   Data Mining: Finding the Data and Models that Create Value   15.062(IDS.145J)  
(Welsch)

**Tentative Schedule**

All readings listed here are in the book by Shmueli, Bruce, Stephens, and Patel, Data Mining for Business Analytics: Concepts, Techniques, and Applications with JMP PRO, 1st Edition (2017) (denoted as DMBA) and class notes.

<table>
<thead>
<tr>
<th>Date (L#)</th>
<th>Topics</th>
<th>Reading</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oct. 31 M (1)</td>
<td>What is Data Mining?</td>
<td>1, 2.1-2.3</td>
</tr>
<tr>
<td>Nov. 1 T</td>
<td>Rec.: Getting Started with JMP PRO (and other software)</td>
<td>2.4-2.8</td>
</tr>
<tr>
<td>2 W (2)</td>
<td>Data Visualization</td>
<td>3</td>
</tr>
<tr>
<td>7 M (3)</td>
<td>Evaluating Classification and Predictive Performance</td>
<td>5</td>
</tr>
<tr>
<td>8 T</td>
<td>Rec.: Visualization and Performance Computing</td>
<td>4.1-4.7</td>
</tr>
<tr>
<td>9 W (4)*</td>
<td>Near Neighbor and Naive Bayes Methods</td>
<td>7, 8</td>
</tr>
<tr>
<td>14 M (5)</td>
<td>Classification and Regression Trees</td>
<td>9</td>
</tr>
<tr>
<td>15 T</td>
<td>Rec: Near Neighbor, Naive Bayes, Trees Computing</td>
<td></td>
</tr>
<tr>
<td>16 W (6)</td>
<td>Regression Review and Selection of Variables</td>
<td>6</td>
</tr>
<tr>
<td>21 M (7)</td>
<td>Logistic Regression</td>
<td>10</td>
</tr>
<tr>
<td>22 T</td>
<td>Rec.: Regression Computing</td>
<td></td>
</tr>
<tr>
<td>23 W (8)*</td>
<td>Discriminant Analysis</td>
<td>12</td>
</tr>
<tr>
<td>28 M (9)</td>
<td>Neural Networks</td>
<td>11</td>
</tr>
<tr>
<td><strong>Project Proposal Due (one page)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>29 T</td>
<td>Rec.: Discriminant Analysis and Neural Net Computing</td>
<td></td>
</tr>
<tr>
<td>30 W (10)</td>
<td>Cluster Analysis</td>
<td>14</td>
</tr>
<tr>
<td>Dec. 5 M (11)</td>
<td>Dimension Reduction</td>
<td>4.8-4.10</td>
</tr>
</tbody>
</table>
There is no final examination in this course. Grades are based on homework, projects, and case studies.

* Denotes tentative homework or case due dates.