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 and systems rely on intelligent technology to analyze data systematically to improve decision-making. In many cases automating analytical and decision-making processes is necessary because of the volume of data and the speed with which new data are generated.

We will examine how data analysis technologies can be used to improve decision-making. We will study the fundamental principles and techniques of data mining, and we will examine real-world examples and cases 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.

After taking this course you should:

1. **Approach business problems data-analytically.** Think carefully & systematically about whether & how data can improve business performance, to make better-informed decisions for management, marketing, investment, etc.

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, consultants, etc. Envision 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. Focus and interaction

The course will explain through lectures and real-world examples the fundamental principles, uses, and some technical details of data mining techniques. The emphasis primarily is on understanding the business application of data mining techniques, and secondarily on the variety of techniques. We will discuss the mechanics of how the methods work as is necessary to understand the fundamental concepts and business application.

This is primarily a lecture-based course, but student participation is an essential part of the learning process in the form of active discussion. I will expect you to be prepared for class discussions by having satisfied yourself that you understand what we have done in the prior classes. You are expected to attend every class session, to arrive prior to the starting time, to remain for the entire class, and to follow basic classroom etiquette, including having all electronic devices turned off and put away for the duration of the class (this is Stern policy, see below) and refraining from chatting or doing other work or reading during class.

The Blackboard site for this course will contain lecture notes, reading materials, assignments, and late-breaking news. You should check the Blackboard site daily, and I will assume that you have read all announcements and class discussion.

You should use the Blackboard discussion board as the primary place to ask questions you have about class material. Please feel free to make full use of the office hours of the course assistants and me. However, if you have the question, someone else may too and everyone may benefit from the answers being available on Blackboard. Also, please try to answer your classmates' questions. In grading your class participation I will include your contributions to the discussion board. You will not be penalized for being wrong in trying to participate on the discussion board (or in class).

Worth repetition: The course assistants and I will use Blackboard as a primary means of communication. It is your responsibility to check Blackboard (and your email) at least once a day during the week (M-F), and you will be expected to be aware of any announcements within 24 hours of the time the message was sent.

I will check my email at least once a day during the week (M-F). I get a tremendous amount of email, and can not process it all daily. Your email will get priority if you include the special tag [DM Grad] in the email subject header as indicated in the information block on the first page of this syllabus. I sort/filter based on this tag in order to make sure to process class mail first. I cannot guarantee to be able to process your email promptly if you do not include the special tag, as I may not read it for a while.

In general, we will follow Stern default policies unless I state otherwise. I will assume that you have read them and agree to abide by them:

http://w4.stern.nyu.edu/academic/affairs/policies.cfm?doc_id=7511
3. Lecture Notes and Readings

Lecture notes: For most classes I will hand out lecture notes, which will outline the primary material for the class. You will be expected to flesh these out with your own note taking, and to ask questions about any material in the notes that is unclear after our class discussion. Depending on the direction our class discussion takes, we may not cover all material in the notes. If the notes themselves are not adequate to explain a topic we skip, you should ask about it on the discussion board.

I will hand out some required readings as we go along.

Other readings are intended to supplement the material we learn in class. They give alternative perspectives on and additional details about the topics we cover:

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

  - available as ebook for free: [http://site.ebrary.com/lib/nyulibrary](http://site.ebrary.com/lib/nyulibrary)
  - available from Amazon
  - Many students find this book to be an excellent supplemental resource. In the class schedule below I suggest the most important sections to read to supplement each class module.

- **“Weka Book” (optional)**: Data Mining: Practical Machine Learning Tools and Techniques, Third Edition by Ian Witten, Eibe Frank, Mark Hall
  - ISBN-10: 0123748569
  - available from Amazon
  - This book provides much more technical details of the data mining techniques and is a very nice supplement for the student who wants to dig more deeply into the technical details. It also provides a comprehensive introduction to the Weka toolkit.
  - Technical note: this book now is in its third edition; the corresponding versions of the Weka software may be different for the different books. I don’t think it will be a big deal—probably best to get the most recent edition of the book. The TAs have it if you have questions. You can always use a more recent version of the software if necessary.

4. Requirements and Grading

The grade breakdown is as follows:

1. Homeworks: 20%
2. Term Project: 30%
3. Participation & Class Contribution: 20%
4. Final Exam: 30%
At NYU Stern we seek to teach challenging courses that allow students to demonstrate differential mastery of the subject matter. Assigning grades that reward excellence and reflect differences in performance is important to ensuring the integrity of our curriculum. In my experience, students generally become engaged with this course and do excellent or very good work, receiving As and Bs, and only one or two perform only adequately or below and receive C’s or lower. Note that the actual distribution for this course and your own grade will depend upon how well each of you actually perform this particular semester.

**Homework Assignments**

The homework assignments are listed (by due date) in the class schedule below. Each homework comprises questions to be answered and hands-on tasks. **Except as explicitly noted otherwise, you are expected to complete your assignments on your own**—without interacting with others.

Completed assignments must be handed on blackboard at least one hour prior to the start of class on the due date (that is, by 5pm), unless otherwise indicated. Assignments will be graded and returned promptly. Answers to homework questions should be well thought out and communicated precisely, avoiding sloppy language, poor diagrams, and irrelevant discussion.

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 using the various techniques discussed in class. You will use these data to build and evaluate predictive models, and to find patterns in the data.

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/ download the “book version” – see HW#1  
http://www.pentaho.com

**IMPORTANT:** 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 is very easy, ensuring that you can install the software and get it running, before moving on to more challenging assignments. If you need additional help with using the data mining software, please see the Course Assistants.

Generally the Course Assistants should be the first point of contact for questions about and issues with the homeworks.

**Late Assignments**

As stated above, assignments are to be submitted on Blackboard at least one hour prior to the start of the class 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. Please turn in your assignment early if there is any uncertainty about your ability to turn it in on time.
Term Project
A term project report will be prepared by student teams. Student teams should comprise 3-4 students. You should decide on your teams by the end of the third class, and submit them to me. Teams are encouraged to interact with the instructor and TA electronically or face-to-face in developing their project reports. You will submit a proposal for your project about half way through the course. Each team will present its project at the end of the semester. We will discuss the project requirements and presentations in class.

Final Quiz
The final quiz will be a take-home to be completed during the week following the last class. The subject matter covered and the exact dates will be discussed in class.

Participation/Contribution/Attendance/Punctuality
Please see Section 2.

Regrading
If you feel that a calculation, factual, or judgment error has been made in the grading of an assignment or exam, please write a formal memo to me describing the error, within one week after the class date on which that assignment was returned. Include documentation (e.g., a photocopy of class notes). I will make a decision and get back to you as soon as I can. Please remember that grading any assignment requires the grader to make many judgments as to how well you have answered the question. Inevitably, some of these go “in your favor” and possibly some go against. In fairness to all students, the entire assignment or exam will be regraded.

FOR STUDENTS WITH DISABILITIES: If you have a qualified disability and will require academic accommodation during this course, please contact the Moses Center for Students with Disabilities (CSD, 998-4980) and provide me with a letter from them verifying your registration and outlining the accommodations they recommend. If you will need to take an exam at the CSD, you must submit a completed Exam Accommodations Form to them at least one week prior to the scheduled exam time to be guaranteed accommodation.

Please read the policies for Stern courses
http://w4.stern.nyu.edu/academic/affairs/policies.cfm?doc_id=7511

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

<table>
<thead>
<tr>
<th>Class Number</th>
<th>Date</th>
<th>Module</th>
<th>Topics</th>
<th>Book Sections (Optional)</th>
<th>Deliverables</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>Introduction</td>
<td>What is DM? Why DM now? DM process, relation to other BI techniques, different data mining tasks</td>
<td>Ch. 1 &amp; 2</td>
<td>Info Sheet</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td>How do I produce a focused segmentation? What is a model? basic terminology, predictive modeling, classification, regression, tree induction, class-probability estimation, brief toolkit demo</td>
<td>Ch. 4 pp. 116-120 Ch. 6 pp. 165-194, 209</td>
<td>HW#1 due Try HW#2 hands-on before this class</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>Data Mining Fundamentals: Predictive Modeling</td>
<td>How do I know my model is any good? evaluation, in-sample versus out-of-sample, overfitting, cross-validation, ROC analysis, expected value framework, domain knowledge validation, geometric interpretation, linear model versus tree induction, logistic regression</td>
<td>Ch. 3</td>
<td>HW#2 due Group Lists due</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td>Bayesian &amp; memory-based reasoning, nearest neighbors, variable normalization, text classification, “Naïve” Bayes, spam filtering</td>
<td>Ch. 8 pp. 257-271</td>
<td>HW#3 due Project Proposal due by…</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>SPRING BREAK</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>Data Mining Fundamentals: Descriptive/Unsupervised Data Mining</td>
<td>descriptive data mining, unsupervised methods, associations, clustering</td>
<td>Ch. 9 Ch. 11</td>
<td>HW#4 due by…</td>
</tr>
<tr>
<td>Class Number</td>
<td>Date</td>
<td>Module</td>
<td>Topics</td>
<td>Book Sections (Optional)</td>
<td>Deliverables</td>
</tr>
<tr>
<td>--------------</td>
<td>------</td>
<td>--------</td>
<td>--------</td>
<td>---------------------------</td>
<td>--------------</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td>Toward Analytical Engineering</td>
<td>Ch. 7 pp. 211-243</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Possible Applications: Fraud Detection</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Targeted Marketing</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Customer Retention</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>variable selection, feature engineering, social networks</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td>Application: On-line Advertising</td>
<td>Ch. 4 pp. 87-110 (skip pp.90-93)</td>
<td>HW#5 due</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ethics of data mining, privacy, what can/do firms know?, what should they do?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>Data Mining in Action: cases, applications, and practical insight</td>
<td>knowledge-engineering bottleneck, rule-based systems, knowledge in action</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>data mining for credit management</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>data mining process in action, expected value in action, clustering revisited</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td>TBD</td>
<td></td>
<td>HW#6 due</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td>DM and ecommerce (Amazon), recommender systems, collaborative filtering, DM and competitive advantage</td>
<td>Ch 8 pp 282-285</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td>wrapup &amp; review</td>
<td>revisit pp.60-64 &amp; 233</td>
<td>Project reports due</td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td>project presentations</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Final Quiz:** Take Home (on Blackboard)