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 and data science. The emphasis primarily is on understanding the fundamental concepts and business applications of data mining. We will discuss the mechanics of how the methods work as is necessary to understand and illustrate the fundamental concepts and business applications. This is not a data mining algorithms course. However, many techniques are the embodiment of one or more of the fundamental principles.

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. The assigned readings will cover the fundamental material. The class meetings will be a combination of lectures on the fundamental material, discussions of business applications of the ideas and techniques, guest lectures from practitioners, case discussions, and student exercises.

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. 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](http://w4.stern.nyu.edu/academic/affairs/policies.cfm?doc_id=7511)

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.

If you have questions about class material that you do not want to ask in class, or that would take us well off topic, please detain me after class, come to office hours to see me or the TAs, or ask on the discussion board. The discussion board is much better than sending me email, which I have a hard time keeping up with. Also, 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: 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). Your email will get priority if you include the special tag [DM Grad] in the email subject header.** I use this tag to make sure to process class email first. If you do not include the special tag, I may not read the email for a while (maybe long). If you forget and send without the tag and then remember, just send it again with the tag. I won’t mind the repetition.
3. Lecture Notes and Readings

**Book:** The textbook for the class will be:

*Data Science for Business: Fundamental principles of data mining and data analytic thinking* Provost & Fawcett (2013).

I will give you a copy of this book in class. I wrote it over the past couple years, in response to feedback from this course—in particular, that the available books were not adequate. This book covers the fundamental material that will provide the basis for you to think and communicate about data mining for business analytics. We will complement the book with discussions of applications, cases, and demonstrations.

**Lecture notes:** For most classes I will hand out lecture notes. I intend that the notes on the fundamental material will follow the book very closely. I expect you to ask questions about any material in the notes that is unclear after our class discussion and reading the book. I wrote the book to follow the class closely, and to free us up for more discussion of applications, etc.—so many of your questions may be answered in the book. If not, please let me know! Depending on the direction our class discussion takes, we may not cover all material in the class notes for any particular session. If the notes and the book are not adequate to explain a topic we skip, you should ask about it on the discussion board. I will be happy to follow up.

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

For those interested in going further, these following supplemental books give alternative perspectives on and additional details about the topics we cover. These are completely optional; you will not be required to know anything in these readings that are not in the primary materials or lectures. I have many other books that I can recommend, for example if you want a reference to a more mathematical treatment of the topics. Please don’t hesitate to come and talk to me about what supplemental material might be best for you, if you want to go further.

- **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.*
- **Supplemental book (optional):**
  - available as ebook for free: [http://site.ebrary.com/lib/nyulibrary](http://site.ebrary.com/lib/nyulibrary)
  - Many students find this book to be an excellent supplemental resource
  - The Third Edition just came out in the past year. I have not read it yet. Berry says it has been improved substantially. I have a copy in my office if you want to talk a look at it before buying it.
  - available from Amazon
4. Requirements and Grading

The grade breakdown is as follows:
1. Homeworks: 20%
2. Term Project: 30%
3. Participation & Class Contribution: 20%
4. Final Quiz: 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/or 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.

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 “latest stable” version (3.6.6)
(which is the version associated with the 3rd edition of the Weka Book)
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. You should bring your computer to the second class. During the class we will have a “lab session” during which we will install and configure the software, get it running, and dealing with the inevitable glitches that a few of you might experience. If you need additional help with using the data mining software, please see the Course Assistant.

Generally the Course Assistant should be the first point of contact for questions about and issues with the homeworks. If they cannot help you to your satisfaction, please do not hesitate to come see me.

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 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., pages in the book, a copy of class notes, etc.). 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
<table>
<thead>
<tr>
<th>Class Number</th>
<th>Date</th>
<th>Topics</th>
</tr>
</thead>
</table>
| 1            |      | Introduction  
What is DM? Why DM now? DM process, relation to other data analytic techniques, different data mining tasks  
Application: Churn Management |
| 2            |      | Introduction to Predictive Modeling  
How do I produce a focused segmentation? What is a model? basic terminology, predictive modeling, classification, regression  
Hands-on “Lab” session – *Bring your laptop*  
Application: Targeted Marketing for Charitable Donations |
| 3            |      | Predictive Modeling II: Supervised Segmentation  
attributes and information, tree induction, class-probability estimation, logistic regression  
geometric interpretation  
Application: Operations Support (diagnosis & technician dispatch) |
| 4            |      | Model performance analytics I: Fitting the data and overfitting the data, holdout testing, cross-validation, learning curves  
Application: Targeted Marketing Revisited |
| 5            |      | Model performance analytics II: Profit, Lift  
ROC analysis, expected value framework, domain knowledge validation  
Guest speaker: Roger Stein, President, Moody’s Research Labs |
|              | SPRING BREAK | (No class Mar 13) |
| 6            |      | Similarity, Distance, Nearest Neighbors  
Application: IBM business customer analytics  
Application: Recommendation systems I |

**Readings**  
Ch. 1 & 2  
Ch. 3  
Ch. 4  
Ch. 5  
Ch. 6  
Ch. 7

**Deliverables**  
Info Sheet (in class)  
HW#1 due  
Team Choices due  
HW#2 due  
Project Proposal due  
HW#3 due
<table>
<thead>
<tr>
<th>Class Number</th>
<th>Date</th>
<th>Topics</th>
<th>Readings</th>
<th>Deliverables</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td></td>
<td>Prediction via evidence combination, Bayesian reasoning, text classification, “Naïve” Bayes Application: Internet Advertizing</td>
<td>Handed out</td>
<td>Project Update Due</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Guest: Brian Dalessandro, Vice President - Research/Data Science M6D</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>Descriptive data mining, unsupervised methods, associations, clustering Demo: Facebook Associations MiniCase: Whiskey Classification Case: GE Capital</td>
<td>Handed out</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td>Toward Analytical Engineering Applications: Revisit Targeted Marketing, Churn, Online Advertising</td>
<td>Handed out</td>
<td>HW#4 due</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Guest: Chris Volinsky Winner of the $1 Million Netflix Challenge Director, Statistics Research AT&amp;T Labs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>Issues in Deployment Data Science Team Development</td>
<td></td>
<td>HW#5 due</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Guest: Claudia Perlich Chief Scientist, M6D Formerly of IBM Research</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td>Wrapup Data Mining Proposal Evaluation A few words on competitive advantage from data science and/or Data Science Job Picture</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td>Project Presentations</td>
<td></td>
<td>Project report due</td>
</tr>
</tbody>
</table>

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