IST687 Applied Data Science  
School of Information Studies  
Syracuse University  
Spring / 2014 – Document version 1.0  
(IS 687 – M001 / Class number 82992)

Instructor: Carlos E. Caicedo, Ph.D.  
Office: 215 Hinds Hall  
Office Hours: TBA

Phone: 315-443-1363 (Office)  
Email: ccaicedo@syr.edu

Class Sessions: Wednesdays 2:15 p.m. – 5:00 p.m.

Catalog Description  
Introduces fundamentals about data and the standards, technologies, and methods for organizing, managing, curating, preserving, and using data. Discusses broader issues relating to data management and use as well as quality control and publication of data.

1. Course Description  
Data scientists play important roles in the four A's of data: data architecture, data acquisition, data analysis and data archiving. A data scientist provides input to system architects on how data needs to be routed and organized to support analysis, visualization and presentation of the data to the appropriate people. Data acquisition, from a data scientist's standpoint, is a critical phase in which data are obtained in the right format, linked to other relevant data and screened to ensure appropriateness for analysis. Analysis is where the data scientists are most heavily involved: summarization of the data, using portions of data (samples) to make inferences about the larger context and visualization of the data by presenting it in tables, graphs and even animations. Finally, the data scientist must become involved in the archiving of the data. Preservation of collected data is a form that makes it highly reusable - what you might think of as "data curation" - is a difficult challenge because it is so hard to anticipate all of the future uses of the data. This course provides an overview of all four areas, as well as the opportunity to ramp up on the popular open source data science tool, the “R” open source statistical analysis and visualization system. R is reckoned by many to be the most popular choice among data analysts worldwide; having knowledge and skill with using it is considered a valuable and marketable job skill for most data scientists.

The course will include applied examples of data collection, processing, transformation, management, and analysis to provide students with hands-on experience. The course will generally use openly available data sets – both small and large – as the basis for demonstrations, homework, and projects. Students will have opportunities to practice the methods and tools for planning a data analysis project from data collection to the end data products and reports with data quality management and data product development in between. The evaluation of each student will arise from Activities, reports, class discussion, and a course project.
1.1. Learning Objectives

At the end of the course, students are expected to understand:

- Essential concepts and characteristics of data
- Scripting/code development for data management using R, R-Studio, and a variety of add-on packages
- Principles and practices in data screening, cleaning, and linking
- Sampler of technologies used to manage, manipulate, and analyze data
- Communication of tabular results and basic visualizations to decision makers

At the end of the course, students are expected to be able to:

- Identify a problem and the data needed for addressing the problem
- Perform basic computational scripting using R and other optional tools
- Transform data through processing, linking, aggregation, summarization, and searching
- Organize and manage data at various stages of a project lifecycle

1.2. What does it take to succeed in the course?

- An interest and passion in data science career in the corporate, academic, or government sector
- Curiosity about business, science, education, health or another substantive area
- Essential computer skills particularly around spreadsheets
- Close familiarity with algebra, geometry, and trigonometry
- Basic understanding of simple descriptive statistics
- Motivation to learn and achieve a high degree of professional preparation

2. Textbooks:

*Introduction to Data Science (2013)*, by Jeffrey M. Stanton. Available for free in the iTunes bookstore or as a PDF download at [http://jsresearch.net](http://jsresearch.net)

The instructor will provide additional and supplemental readings in Blackboard as electronic documents for downloading and printing. Students are expected to read the assigned materials for discussions and coursework.

3. Contributions to Grade

The work for this class will involve a mixture of individual assignments, reports, and a final project.

- **Activities** (11 x 5% = 55%) are designed for you to practice the necessary skills in carrying out data processing, analysis, and management tasks.
- **Mid-term quiz** (15%) is designed to evaluate your mastery of concepts, methods, and tools in data analysis and management.
• **Final individual project** (30%): For the final project you will identify a data set or a data source, transform the data as needed, and provide an analysis with visualizations.

4. **Grading Policy**

- Each assigned work will be graded on the scale as specified for the component, which will be summed at the end of the semester.
- Grade levels follow the scales below:

<table>
<thead>
<tr>
<th>Highest</th>
<th>Lowest</th>
<th>Letter</th>
<th>Expectation</th>
</tr>
</thead>
<tbody>
<tr>
<td>100.00 %</td>
<td>93.00 %</td>
<td>A</td>
<td>Your work is outstanding</td>
</tr>
<tr>
<td>92.99 %</td>
<td>90.00 %</td>
<td>A-</td>
<td></td>
</tr>
<tr>
<td>89.99 %</td>
<td>87.00 %</td>
<td>B+</td>
<td>Your work is about what would be expected of a serious student</td>
</tr>
<tr>
<td>86.99 %</td>
<td>83.00 %</td>
<td>B</td>
<td></td>
</tr>
<tr>
<td>82.99 %</td>
<td>80.00 %</td>
<td>B-</td>
<td></td>
</tr>
<tr>
<td>79.99 %</td>
<td>77.00 %</td>
<td>C+</td>
<td>Your work falls below what is expected but is adequate</td>
</tr>
<tr>
<td>76.99 %</td>
<td>73.00 %</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>72.99 %</td>
<td>70.00 %</td>
<td>C-</td>
<td></td>
</tr>
<tr>
<td>69.99%</td>
<td>0.00 %</td>
<td>F</td>
<td>Your work is out of the picture</td>
</tr>
</tbody>
</table>

- It is unethical to allow some students additional opportunities, such as extra credit assignments, without allowing the same options to all students.
- To discuss a grade, arrange for a private meeting in which you identify the sources of your concern. It is important to bring with you to that meeting the relevant materials (e.g., marked papers). Except for extraordinary circumstances, no appeal for an individual assignment or project will be considered later than two weeks after the graded assignment was returned. For final grades, no appeal will be considered after one week of final project submission date.

5. **Schedule**

<table>
<thead>
<tr>
<th>Unit</th>
<th>Topics</th>
<th>Readings</th>
<th>Activities/Assignments</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>General introduction to the course.</td>
<td>Intro to DS, pp. ii-7 (Section 1: Data Science, Many Skills)</td>
<td>Install the R open source software package on your computer. Submit a screenshot of R to the instructor.</td>
</tr>
<tr>
<td>2</td>
<td>Essential concepts of data.</td>
<td>Intro to DS, pp. 8-13 (Chapter 1: About Data)</td>
<td>Activity 1: Boolean logic.</td>
</tr>
<tr>
<td>3</td>
<td>Using R to manipulate data.</td>
<td>Intro to DS, pp. 14-23 (Chapter 2: Identifying Data Problems; Chapter 3: Getting Started with R)</td>
<td>Activity 2: myFamily data structure (Note to instructors: Same as DSM1 Lab 1)</td>
</tr>
<tr>
<td>Activity</td>
<td>Topic</td>
<td>Page and Section</td>
<td>Notes</td>
</tr>
<tr>
<td>----------</td>
<td>-------</td>
<td>-----------------</td>
<td>-------</td>
</tr>
<tr>
<td><strong>4</strong></td>
<td>Data frames and other data structures</td>
<td>Intro to DS, pp. 24-36 (Chapter 4: Follow the Data; Chapter 5: Rows and Columns)</td>
<td>Activity 3: Manipulating data frames</td>
</tr>
<tr>
<td><strong>5</strong></td>
<td>Descriptive Statistics, Distributions, and Histograms</td>
<td>Intro to DS, pp. 37-48 (Chapter 6: Beer, Farms, and Peas)</td>
<td>Activity 4: Generating distributions (Note to instructors: Same as DSM1 Lab 2)</td>
</tr>
<tr>
<td><strong>6</strong></td>
<td>Introduction to sampling and inferential statistics</td>
<td>Intro to DS, pp. 49-49 (Chapter 7: Sample in a Jar)</td>
<td>Activity 5: Standard error and alpha thresholds</td>
</tr>
<tr>
<td><strong>7</strong></td>
<td>Identifying a data source for your final project</td>
<td>Intro to DS, pp. 50-64 (Chapter 8: Big Data: Big Deal)</td>
<td>Submit &lt;1 page proposal for the dataset or data source you plan to use for your final project Midterm Quiz</td>
</tr>
<tr>
<td><strong>8</strong></td>
<td>Code creation with R studio</td>
<td>Intro to DS, pp. 65-75 (Chapter 9: Onward with R Studio)</td>
<td>Activity 6: Code your own MySamplingDistribution()</td>
</tr>
<tr>
<td><strong>9</strong></td>
<td>Connecting with external data sources</td>
<td>Intro to DS, pp. 132-147 (Chapter 14: Storage Wars)</td>
<td>Activity 7: Running SQL queries from R Submit &lt;2 page proposal outlining data analysis plan</td>
</tr>
<tr>
<td><strong>10</strong></td>
<td>Working with map data</td>
<td>Intro to DS, pp. 148-159 (Chapter 15: Map Mash-Up)</td>
<td>Activity 8: Color code map points</td>
</tr>
<tr>
<td><strong>11</strong></td>
<td>Linear modeling</td>
<td>Intro to DS, pp. 160-169 (Chapter 16: Line Up, Please)</td>
<td>Activity 9: Evaluating linear models</td>
</tr>
<tr>
<td><strong>12</strong></td>
<td>Association Rules Mining</td>
<td>Intro to DS, pp. 170-182 (Chapter 17: Hi Ho, Hi Ho-Data Mining We Go)</td>
<td>Activity 10: Apply association rules mining to a novel data set Submit &lt; 3 page project report describing results of data screening, cleaning, and linking</td>
</tr>
<tr>
<td><strong>13</strong></td>
<td>Support Vector Machines</td>
<td>Intro to DS, pp. 183-193 (Chapter 18: What’s your vector, Victor?)</td>
<td>Activity 11: Apply support vector analysis to a novel data set</td>
</tr>
<tr>
<td><strong>14</strong></td>
<td>Final project problem diagnosis and support</td>
<td>No reading</td>
<td>No Activity</td>
</tr>
<tr>
<td><strong>15</strong></td>
<td>Final project problem diagnosis and support</td>
<td>No reading</td>
<td>Final project submissions due</td>
</tr>
</tbody>
</table>
6. Class Expectations

**General Requirement for Assignment Submissions**
Assignments must be professionally prepared with computer applications. Unless otherwise stated, assignments must be submitted electronically to Blackboard. No hand-written assignments will be accepted. An assignment must be in one document when it is submitted to Blackboard. If you have additional supporting materials that are in physical forms or hard copies (e.g., business forms or some images), you must scan them into JPG or TIFF format and embed them into your MS Word or PDF document. If you use MS Word to prepare your assignments, use Times New Roman style with 12-pt font, 1.5 line spacing, and 1 inch margin around.

For each assignment, be sure to include the a title page with the following information:
- Course number (i.e. IST 687)
- Name(s) of the author(s) or the team members
- Name of the assignment (e.g. “Homework 2”)
- Date the assignment is due and date in which it is submitted
- Number pages in the body of the assignment (not necessary on the first page)

You must submit all assignments to the professor in class on the deadline specified for each assignment. If you are late, I will deduct 10% of the original grade for the first day of lateness plus 15% for every subsequent day. You will not receive full credit for topics/assignments presented in an unprofessional manner. Professionalism includes the proper use of grammar, punctuation, and limiting spelling mistakes. Professionalism also means adhering to given instructions. Failure to adhere to the assignment instructions will result in a reduction of the grade. If English is not your first language set up an appointment with the writing program so they can help you improve your writing.

**Respect and Disruption**
It is expected that we all treat people’s contributions and differences of opinion with respect. There are certain actions that can be disruptive not only to your own learning experience but to everybody else’s as well. Examples include talking to neighbors during class, arriving late, cell phone ringing during class, text messaging, falling asleep, reading newspapers or magazines, lack of civility and respect in comments made, etc. Your repeated disruption will reduce your final grade. In extreme cases, you can be asked to leave the class and even excluded from the course.

7. Course LMS
This course uses the University’s Blackboard LMS. This system is available at http://blackboard.syr.edu. The LMS is a required and integral part of the course. Teaching
materials (syllabus, class notes, assignments, resources, etc.), grading information and progress will be made available on Blackboard.

8. University and School Policies

Academic Integrity
The academic community of Syracuse University and of the School of Information Studies requires the highest standards of professional ethics and personal integrity from all members of the community. Violations of these standards are violations of a mutual obligation characterized by trust, honesty, and personal honor. As a community, we commit ourselves to standards of academic conduct, impose sanctions against those who violate these standards, and keep appropriate records of violations. The academic integrity statement can be found at: http://supolicies.syr.edu/ethics/acad_integrity.htm.

Faith-based Observances
Syracuse University recognizes the diverse faith traditions represented among its campus community and supports the rights of faculty, staff, and students to observe according to these. This link http://supolicies.syr.edu/studs/religious_observance.htm provides a description of SU’s religious observance policy. Under this policy, students are provided an opportunity to make up examination, study, or work requirements that may be missed due to religious observance provided they notify the university and their instructors before the end of the second week of classes. Students will have access to an online notification system for this purpose on MySlice during the first two weeks of the semester. The make up of an activity affected by a religious observance will be scheduled to be completed within a week of the missed deadline in agreement with the instructor.

Disabilities
If you believe that you need accommodations for a disability, please contact the Office of Disability Services (ODS), http://disabilityservices.syr.edu, located in Room 309 of 804 University Avenue, or call (315) 443-4498 for an appointment to discuss your needs and the process for requesting accommodations. ODS is responsible for coordinating disability-related accommodations and will issue students with documented disabilities Accommodation Authorization Letters, as appropriate. Since accommodations may require early planning and generally are not provided retroactively, please contact ODS as soon as possible.

Ownership of Student Work
This course may use course participation and documents created by students for educational purposes. In compliance with the Federal Family Educational Rights and Privacy Act, works in all media produced by students as part of their course participation at Syracuse University may be used for educational purposes, provided that the course syllabus makes clear that such use may occur. It is understood that registration for and continued enrollment in a course where such use of student works is announced constitutes permission by the student. After such a course has
been completed, any further use of student works will meet one of the following conditions: (1) the work will be rendered anonymous through the removal of all personal identification of the work’s creator/originator(s); or (2) the creator/originator(s)’ written permission will be secured. As generally accepted practice, honors theses, graduate theses, graduate research projects, dissertations, or other exit projects submitted in partial fulfillment of degree requirements are placed in the library, University Archives, or academic departments for public reference.