COURSE NUMBER: CS 506  (note that some paperwork says CS 505, but the number has been adjusted to allow cross-listing in ECE).

INSTRUCTOR(S): Evimaria Terzi, George Kollios, Mark Crovella

TO BE FIRST OFFERED:  Sem./Year:  _Spring_ /__2017__

SHORT TITLE: The “short title” appears in the course inventory, on the Link University Class Schedule, and on student transcripts and must be 15 characters maximum including spaces. It should be as clear as possible.

COURSE DESCRIPTION: This is the description that appears in the CAS and/or GRS Bulletin and The Link. It is the first guide that students have as to what the course is about. The description can contain no more than 40 words.

Covers practical skills in working with data and introduces a wide range of techniques that are commonly used in the analysis of data, such as clustering, classification, regression, and network analysis. Emphasizes hands-on application of methods via programming.

PREREQUISITES: Indicate “None” or list all elements of the prerequisites, clearly indicating “AND” or “OR” where appropriate. Here are three examples: “Junior standing or CAS ZN300 or consent of instructor”; “CAS ZN108 and CAS ZN203 and CAS PQ206; or consent of instructor”; “For SED students only.”

1. State the prerequisites:

   CS 108 or CS 111;  CS 132 or MA 242 or MA 442;  CS 112 Recommended

2. Explain the need for these prerequisites:

   The requirement for CS 108 or CS 111 ensures that students have a sufficient level of programming ability. The recommendation for CS 112 is to help ensure that students are aware that significant programming ability is needed to complete the course. The requirement for CS 132, MA 242, or MA 442 is to ensure that students have the necessary grounding in linear algebra for this subject.
CREDITS: (check one)
□ Half course: 2 credits  □ Variable: Please describe.
X Full course: 4 credits  □ Other: Please describe.

Provide a rationale for this number of credits, bearing in mind that for a CAS or GRS course to carry 4 credits, 1) it must normally be scheduled to meet at least 150 minutes/week, AND 2) combined instruction and assignments, as detailed in the attached course syllabus, must anticipate at least 12 total hours/week of student effort.

This will be a standard course, with two 90 minute lectures per week. There are weekly homework assignments which, along with studying the lecture material, will require over 12 hours/week of student effort.

DIVISIONAL STUDIES CREDIT: Is this course intended to fulfill Divisional Studies requirements?
X No.
□ Yes. If yes, please indicate which division ______________________ and explain why the course should qualify for Divisional Studies credit. Refer to criteria listed here and specify whether this course is intended for “short” or “expanded” divisional list.

HOW FREQUENTLY WILL THE COURSE BE OFFERED?
X Every semester  □ Once a year, fall  □ Once a year, spring  □ Every other year
□ Other: Explain:

NEED FOR THE COURSE: Explain the need for the course and its intended impact. How will it strengthen your overall curriculum? Will it be required or fulfill a requirement for degrees/majors/minors offered by your department/program or for degrees in other departments/school/colleges? Which students are most likely to be served by this course? How will it contribute to program learning outcomes for those students? If you see the course as being of “possible” or “likely” interest to students in another departments/program, please consult directly with colleagues in that unit. (You must attach appropriate cognate comments using cognate comment form if this course is intended to serve students in specific other programs. See FURTHER INFORMATION below about cognate comment.)
This course is based on three successful offerings in trial versions in Fall 2015, Spring 2016, and Fall 2016. Each offering had to be capped at 75 students due to the very high level of interest in the subject matter.

This course serves as an introduction to Data Science from a Computer Science perspective. It emphasizes practical tools of data analysis and machine learning with an applied emphasis. Standard topics in machine learning, including clustering, classification, regression, and network analysis are presented. Emphasis is on the computational methods needed to obtain results efficiently on modern computer hardware, including distributed or cluster-based computing systems.

This course adds a key element to the department’s curriculum in the area of data science. The department currently offers CS 565 / Data Mining, which emphasizes the algorithmic and theoretical underpinnings of many topics covered in this course. The department also offers CS 560 / Databases and CS 562 / Advanced Database Applications, which cover the algorithms and systems required to store and manipulate data efficiently and securely. All three of those courses are primarily targeted at Computer Science concentrators and graduate students.

The proposed course, while appropriate for Computer Science majors and graduate students, is also well suited to non-majors who need an improved ability to work with and draw conclusions from data. As such, it has proved to be popular with students in Engineering (ECE), in Math and Statistics, and in other programs. In fact, the course has recently been accepted for credit toward the ECE specialization in data analytics.

ENROLLMENT: How many undergraduate and/or graduate students do you expect to enroll in the initial offering of this course?

75

CROSS-LISTING: Is this course to be cross-listed or taught with another course? If so, specify. Chairs/directors of all cross-listing units must co-sign this proposal on the signature line below.

OVERLAP:

1. Are there courses in the UIS Course Inventory (CC00) with the same number and/or title as this course?
   X No.
   ☐ Yes. If yes, any active course(s) with the same number or title as the proposed course will be phased out upon approval of this proposal.
   NOTE: A course number cannot be reused if a different course by that number has been offered in the past five years.

2. Relationship to other courses in your program or others: Is there any significant overlap between this course and others offered by your department/program or by others? (You must attach appropriate cognate comments using cognate comment form if this course might be perceived as overlapping with courses in another department/program. See FURTHER INFORMATION below.)
FACILITIES AND EQUIPMENT: What, if any, are the new or special facilities or equipment needs of the course (e.g., laboratory, library, instructional technology, consumables)? Are currently available facilities, equipment, and other resources adequate for the proposed course? (NOTE: Approval of proposed course does not imply commitment to new resources to support the course on the part of CAS.)

Current facilities are adequate for the proposed course.

STAFFING: How will the staffing of this course, in terms of faculty and, where relevant, teaching fellows, affect staffing support for other courses? For example, are there other courses that will not be taught as often as now? Is the staffing of this course the result of recent or expected expansion of faculty? (NOTE: Approval of proposed course does not imply commitment to new resources to support the course on the part of CAS.)

We anticipate that this course will continue to be taught every semester in the near future, due to the currently-strong demand. Faculty who will teach this course include both existing faculty (Terzi, Kollios, Crovella), newly hired faculty (Tsourakakis), as well as potential future hires in the area of data science.

BUDGET AND COST: What, if any, are the other new budgetary needs or implications related to the start-up or continued offering of this course? If start-up or continuation of the course will entail costs not already discussed, identify them and how you expect to cover them. (NOTE: Approval of proposed course does not imply commitment to new resources to support the course on the part of CAS.)

No start up costs.

EXTERNAL PROGRAMS: If this course is being offered at an external program/campus, please provide a brief description of that program and attach a CV for the proposed instructor.

FURTHER INFORMATION THAT MUST BE ATTACHED IN ORDER FOR THIS PROPOSAL TO BE CONSIDERED:

- A complete week-by-week SYLLABUS with student learning objectives, readings, and assignments that reflects the specifications of the course described in this proposal; that is, appropriate level, credits, etc. (See guidelines on “Writing a Syllabus” on the Center for Teaching & Learning website.) Be sure that syllabus includes your expectations for academic honesty, with URL for pertinent undergraduate or GRS academic conduct code(s).

- Cognate comment from chairs or directors of relevant departments and/or programs. Use the form here under “Curriculum Review & Modification.” You can consult with Joseph Bizup (CAS) or Jeffrey Hughes (GRS) to determine which departments or programs inside and outside of CAS would be appropriate.

DEPARTMENT CONTACT NAME AND POSITION:

Mark Crovella, Professor and Chair

DEPARTMENT CONTACT EMAIL AND PHONE:
crovella@bu.edu, 3-8919

DEPARTMENT APPROVAL:

______________________________  ________________________
Department Chair                        Date

______________________________  ________________________
Other Department Chair(s) (for cross-listed courses)  Date
DEAN’S OFFICE CURRICULUM ADMINISTRATOR USE ONLY

CAS/GRS CURRICULUM COMMITTEE APPROVAL:

☐ Approved  Date: ________________
☐ Tabled  Date: ________________
☐ Not Approved  Date: ________________

Divisional Studies Credit:
☐ Endorsed
☐ Not endorsed

______________________________________________________________
Curriculum Committee Chair Signature and Date

Comments:

PROVISIONAL APPROVAL REQUESTED for Semester/Year  ________________

______________________________________________________________
Dean of Arts & Sciences Signature and Date

Comments:

CAS FACULTY:  Faculty Meeting Date: ________________  ☐ Approved  ☐ Not Approved

______________________________________________________________
Curriculum Administrator Signature and Date

Comments:
Boston University College and Graduate School of Arts & Sciences
Undergraduate Academic Program Office
725 Commonwealth Avenue, Room 102

Date:  Sept 8, 2016

Cognate Comment Request

TO:  Name:  Clem Karl
Department:  ENG / ECE

FROM:  Name:  Mark Crovella
Department:  CAS / Computer Science
Telephone:  3-8919  E-mail: crovella@bu.edu

Course Number:  CS 505 506
Course or Program Title:  Computational Tools for Data Science

Our Department/Program would like to request cognate comments on this course (or program). A complete proposal is attached for your review. If you need further information, please do not hesitate to contact me.

Kindly return the signed original to me by  Sept 15  so that I may include your comments when submitting our proposal for review and approval. Please do not send any cognate letters directly to the address above. Thank you.

COMMENTS:  CS 505/CS 591 will be a welcome complementary effort to ENG EC 503 which can benefit ECE students given its focus and current demand. The ECE department would like to co-list the course in ECE.

________________________________________________________________________
________________________________________________________________________

________________________________________________________________________
________________________________________________________________________

__________________________________________
Signature:  Professor and Chair

Date:  9/16/16

Please explain fully any objections.
TO: Tasso Kaper  
Department: Mathematics and Statistics  
FROM: Mark Crovella  
Department: CAS / Computer Science  
Telephone: 3-8919  
E-mail: crovella@bu.edu  

Course Number: CS 505  
Course or Program Title: Computational Tools for Data Science  

Date: Sept 18, 2016

Our Department/Program would like to request cognate comments on this course (or program). A complete proposal is attached for your review. If you need further information, please do not hesitate to contact me.

Kindly return the signed original to me by [date] so that I may include your comments when submitting our proposal for review and approval. Please do not send any cognate letters directly to the address above. Thank you.

COMMENTS: We are happy to provide this cognate comment on the proposed CS 505 course. We thank you very much for having provided us detailed information earlier about the recent and current offerings of CS591, which has the same content as the proposed course, and we are pleased to see for your sake that this will now become a regular offering.

The methods for clustering, classification, regression, and network analysis covered in this course have their theoretical grounding in the fields of computer science and statistics. Also, they are some of the many important tools for the broad field of data analysis.

From the proposal and syllabus, we understand that emphasis of this course will be on the practical implementation of the methods, via a 'hands on' understanding of some classical data analysis techniques and proficiency in Python. As such, we think that the course complements existing offerings and will lead to increased demand from students to learn many of the other statistical techniques and computer science techniques that are central to, and useful for, data science and data analytics. We plan to list this course, once it's approved, as an elective for our Masters students.

Finally, we greatly appreciate your commitment that this course will become part of the type of proposed data science programs at both the Masters and Undergraduate levels that our departments are currently jointly developing.

Signature:

PROFESSOR & CHAIR

Date: 9/18/16
Meeting Place: SCI 117

Meeting Time: TR 11-12:30

Instructor: Prof. Mark Crovella

- Office: MCS-140E
- Office Hours: M 2-3:30, R 3-4:30
- Email: crovella@bu.edu

Teaching Fellow: Ms. Katherine Missimer

- Office Hours: W 4-5:30, F 5-6:30
- Office Hours Location: Undergrad Lab, EMA 302
- Lab Tutoring Hours: F 3-5.
- Email: kzhao@bu.edu

Overview of the Course

This course is targeted at students who require a basic level of proficiency in working with and analyzing data. The course emphasizes practical skills in working with data, while introducing students to a wide range of techniques that are commonly used in the analysis of data, such as clustering, classification, regression, and network analysis. The goal of the class is to provide to students a hands-on understanding of classical data analysis techniques and to develop proficiency in applying these techniques in a modern programming language (Python).

Broadly speaking, the course breaks down into three main components, which we will take in order of increasing complication: (a) unsupervised methods; (b) supervised methods; and (c) methods for structured data.

Lectures will present the fundamentals of each technique; focus is not on the theoretical underpinnings of the methods, but rather on helping students understand the practical settings in which these methods are useful. Class discussion will study use cases and will go over relevant Python packages that will enable the students to perform hands-on experiments with their data.

Prerequisites: Students taking this class must have some prior familiarity with programming, at the level of CS 105, 108, or 111, or equivalent. CS 132 or equivalent (MA 242, MA 442) is required. CS 112 is also helpful.
Learning Outcomes

Students who successfully complete this course will be proficient in data acquisition, manipulation, and analysis. They will have good working knowledge of the most commonly used methods of clustering, classification, and regression. They will also understand the efficiency issues and systems issues related to working on very large datasets.

Readings

There is no text. Lecture notes will be posted online.

Some recommended texts are:

1. Python for Data Analysis (http://shop.oreilly.com/product/0636920023784.do)

Web Resources

The slides I use are actually executable python scripts, using the jupyter notebook. You can download and execute the lectures on your own computer, and you can modify them any way you’d like, play around with them, experiment, etc.

The slides I use in lecture are published on github. The repository is https://github.com/mcrovella/CS505-Data-Science-in-Python. If you want to access the repository using git, please feel free. If you find a bug, feel free to submit a pull request.

Homeworks and Project

1. There will nine homework assignments. In a typical assignment you will analyze one or more datasets using the tools and techniques presented in class.

   Homeworks will be submitted via github. For this, we need your github account (create one if you don’t already have it). After you have created it, fill out the form at https://goo.gl/forms/8W0S0dvMn07UKdip2 to let us know what it is.

   You are expected to work individually on homeworks.

2. In addition, there will be a final project. For the project you will extract some knowledge or conclusions from the analysis of dataset of your choice. The analysis will be done using a subset of the methods we described in class. The final project will require a proposal, two progress reports, and a final presentation in poster form.

   The project will have three essential components: 1) a data collection piece (which may involve crawling or calls to an API, combining data from different sources etc), 2) a data analysis piece (which will involve applying different techniques we described in class for the analysis) and 3) a conclusion component (where the results of the data analysis will be drawn). The students will submit a 5-page report explaining clearly all the three components of their project. Finally a poster presentation will be required where the students will be prepare to present their effort and results in front of their poster.
As an example, you may choose to collect data from Twitter related to a specific topic (e.g., Ebola virus) and then measure the intensity of posts about a topic in different areas of the world etc. Other examples of projects may include (but are not limited to): analysis of MBTA data, analysis of NYC data, crawling of YouTube (or other social media data) and analysis of social behavior like trolling, bullying etc.

The project is due by the last day of class (December 8). The project presentations will be given in the form of a final poster explaining components 1, 2 and 3 of the project.

You are expected to work in teams of two on the final project. I will leave it up to you to form teams on your own, but everyone must work in a team.

**Piazza**

We will be using Piazza for class discussion. The system is really well tuned to getting you help fast and efficiently from classmates, Ms. Missimer, and myself. Rather than emailing questions to the teaching staff, I encourage you to post your questions on Piazza. Our class Piazza page is at: https://piazza.com/bu/fall2016/cs505/home. We will also use Piazza for distributing materials such as homeworks and solutions.

When someone posts a question on Piazza, if you know the answer, please go ahead and post it. However please don’t provide answers to homework questions on Piazza. It’s OK to tell people where to look to get answers, or to correct mistakes; just don’t provide actual solutions to homeworks.

**Programming Environment**

We will use python as the language for teaching and for assignments that require coding. Instructions for installing and using Python are on Piazza.

**Course and Grading Administration**

Homeworks are due at 7pm on Fridays. Assignments will be submitted using github. Ms. Missimer will explain how to submit assignments.

*NOTE: IMPORTANT:* Late assignments **WILL NOT** be accepted. However, you may submit one homework up to 3 days late. You must email Ms. Missimer before the deadline if you intend to submit a homework late.

Final grades will be computed based on the following:

- 50%  Homework assignments.
- 50%  Final Project

The exact cutoffs for final grades will be determined after the class is complete.
**Academic Honesty**

You may discuss homework assignments with classmates, but you are solely responsible for what you turn in. Collaboration in the form of discussion is allowed, but all forms of cheating (copying parts of a classmate's assignment, plagiarism from books or old posted solutions) are NOT allowed. We – both teaching staff and students – are expected to abide by the guidelines and rules of the Academic Code of Conduct (which is at [http://www.bu.edu/dos/policies/student-responsibilities/](http://www.bu.edu/dos/policies/student-responsibilities/)).

You can probably, if you try hard enough, find solutions for homework problems online. Given the nature of the Internet, this is inevitable. Let me make a couple of comments about that:

1. If you are looking online for an answer because you don’t know how to start thinking about a problem, talk to Ms. Missimer or myself, who may be able to give you pointers to get you started. Piazza is great for this – you can usually get an answer in an hour if not a few minutes.

2. If you are looking online for an answer because you want to see if your solution is correct, ask yourself if there is some way to verify the solution yourself. Usually, there is. You will understand what you have done much better if you do that. So ... it would be better to simply submit what you have at the deadline (without going online to cheat) and plan to allocate more time for homeworks in the future.
<table>
<thead>
<tr>
<th>Date</th>
<th>Topics</th>
<th>Reading</th>
<th>Assigned</th>
<th>Due</th>
</tr>
</thead>
<tbody>
<tr>
<td>9/6</td>
<td>Introduction to Python</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9/8</td>
<td>Essential Tools (Git, Jupyter Notebook, Pandas)</td>
<td></td>
<td></td>
<td>HW 0</td>
</tr>
<tr>
<td>9/13</td>
<td>Probability and Statistics Refresher</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9/15</td>
<td>Linear Algebra Refresher</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9/20</td>
<td>Numpy, Scikit-learn, Distance and Similarity Functions</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9/22</td>
<td>Intro to Timeseries</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9/27</td>
<td>Clustering, k-means</td>
<td>HW 1.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9/29</td>
<td>Clustering II</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9/30</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9/27</td>
<td>Hierarchical Clustering</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10/4</td>
<td>Expectation Maximization and GMM</td>
<td>HW 2.1, 2.2</td>
<td></td>
<td>HW 1.2</td>
</tr>
<tr>
<td>10/6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10/7</td>
<td>NO CLASS; Monday Schedule</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10/11</td>
<td>DB Clustering and Comparing Clustering Algorithms</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10/18</td>
<td>Dimensionality Reduction - SVD I</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10/20</td>
<td>SVD II and Web Scraping</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10/21</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10/25</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10/27</td>
<td>Open</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10/28</td>
<td>Classification: Decision Trees</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11/1</td>
<td>Classification: SVM, Naive Bayes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11/3</td>
<td>Regression: Linear Regression</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11/4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11/8</td>
<td>Logistic Regression</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11/10</td>
<td>Linear Regression II</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11/11</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11/15</td>
<td>Recommendation Systems</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11/17</td>
<td>Network Analysis I</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11/18</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11/22</td>
<td>Network Analysis II</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11/24</td>
<td>NO CLASS; Thanksgiving Break</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11/29</td>
<td>Graph Clustering</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12/1</td>
<td>Text Analysis and Topic Modeling</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12/2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12/6</td>
<td>Wrapup</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12/8</td>
<td>Poster Session</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12/12</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12/8</td>
<td></td>
<td></td>
<td></td>
<td>HW 5</td>
</tr>
</tbody>
</table>

*HW 0* indicates homework due on the assigned date.

*HW 1.1* indicates homework due on the following date.

*HW 1.2* indicates homework due on the subsequent date.

*HW 2.1* indicates homework due on the following date.

*HW 2.2* indicates homework due on the subsequent date.

*HW 3.1* indicates homework due on the following date.

*HW 3.2* indicates homework due on the subsequent date.

*HW 4* indicates homework due on the following date.

*HW 5* indicates homework due on the following date.

*Hw 5* indicates homework due on the following date.