UNIVERSITY OF MARY WASHINGTON -- NEW COURSE PROPOSAL

Electronically submit this completed form with attachments in one file to the Chair of the College Curriculum Committee.

<table>
<thead>
<tr>
<th>COLLEGE (check one):</th>
<th>Arts and Sciences</th>
<th>Business</th>
<th>X</th>
<th>Education</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposal Submitted By: Stephen Davies, Chris Garcia</td>
<td>Date Prepared: Fall 2014</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Course Title: Foundations for Data Science

Department/discipline and course number*: BUAD/CPSC 219

*This course number must be approved by the Office of the Registrar before the proposal is submitted.

<table>
<thead>
<tr>
<th>Number of credits proposed:</th>
<th>3</th>
<th>Prerequisites:</th>
<th>None</th>
</tr>
</thead>
<tbody>
<tr>
<td>Will this be a new, repeatable “special topics” course? (Do you want students to be able to take this new course more than once if the topic changes?)</td>
<td>NO</td>
<td>X</td>
<td>YES</td>
</tr>
</tbody>
</table>

| Date of first offering of this new course: FALL SEMESTER, 2015 |
| Proposed frequency of offering of the course: Once per year |
| List the faculty who will likely teach the course: Davies, Garcia, Hohman/Marchette/Solka (NSWC) |

Are ANY new resources required? NO | YES | X | Document in attached impact statement |

This new course will be (check all that apply):

| Required in the minor | X | General Elective | X |
| Elective in the major | General Education** |

**AFTER the new course is approved, a separate proposal must be sent to the General Education Committee.

Catalog Description:

BUAD/CPSC 219 -- Foundations for Data Science (3)

Prerequisite: none. Skills and tools in acquiring, parsing, manipulating, and preparing data for statistical analysis.

COURSE HISTORY

Was this course taught previously as a topics or experimental course? YES | X | NO |

<table>
<thead>
<tr>
<th>Course Number and Title of Previous Course</th>
<th>Semester Offered</th>
<th>Enrollment</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPSC 270C Foundations for Data Science</td>
<td>Spring 2015</td>
<td>TBA</td>
</tr>
</tbody>
</table>

X CHECK HERE if the proposed course is to be equated with the earlier topics or experimental offerings. This means that students who took the earlier “topics” course will only be able to take the new course if they made a C- grade or lower in the earlier course.

NOTE: If the proposed course has not been previously offered as a topics or experimental course, explain in the attached rationale statement why the course should be adopted even though it has not been tried out.

REQUIRED ATTACHMENTS:

1. Rationale Statement (Why is this course needed? What purposes will it serve?)
2. Impact Statement (Provide details about the Library, space, budget, and technology impacts created by adding this new course. Include supporting statements from the Library, IT Department, etc. as needed.)
3. Sample Syllabus

Department Chair Approval: Ken Machande Date: 11/17/14

CCC Chair Approval: T. Nichole Phillips Date: 12/4/14

UCC Chair Approval: Date: ____________________

New Course Proposal Cover Sheet (July 2013)
Data Science is about computational methods to enable the statistical analysis of data. It uses algorithms that can detect and quantify patterns in data sets, which are essential for both prediction and understanding.

Before any meaningful analysis can take place, however, the data itself has to exist in electronic form. Nearly all interesting information has its origin in the awkward, inconsistent, and ambiguous “real world,” and must be imported and represented electronically in order to be analyzed en masse. Variables must be appropriately encoded; data points combined or separated; information from different sources intelligently merged. Doing this in a way that is faithful to the data’s original meaning and context, and which simultaneously satisfies data storage constraints, is a non-trivial and challenging task. Further, even when available electronically, data must be structured and represented in a way that conforms to what the data analysis algorithms require. For example, to perform a linear regression, pairs of coordinates (or more generally, tuples of more than two dimensions) must exist in the form of two (or more) numeric vectors of identical length. For most classification algorithms, a single spreadsheet-like table is required in which every observation appears in exactly one row, and every quantitative or qualitative feature appears in exactly one column. Useful data comes from diverse sources such as databases, websites, spreadsheets, plain-text documents, public APIs, and serialized formats like XML or JSON. Each brings with its own methods and challenges. Hierarchical structures may need to be “flattened,” variables transformed, scales adjusted, quantitative values recoded into labeled intervals. In short, a prerequisite for meaningful data analysis is often a shrewd and artistic combination of practical techniques used to wrestle a precious data set into a form where it can be dealt with. This vital activity goes by many names in the industry: data preparation, cleansing, munging, fusing,” and wrangling.

The need for the course
In our recent collaboration with the Dahlgren Naval Surface Warfare Center, in which Naval researchers have helped us jointly mentor over a dozen individual study students on real-world data science projects, we quickly identified this area as a gap in our curriculum. Math and Computer Science students, excited to dive into an interesting data set and make predictions and discoveries, immediately hit a wall because the data does not already exist in the form needed for their analysis tool. And students, having no experience with methods for dealing with this situation, are stymied as to how to proceed. In our first semester of projects, we had one group of students essentially playing “catch-up” for the entire semester dealing with data importing and transformation issues.

Dr. Debbie Hydorn (Mathematics) jokes that traditionally, courses in statistical methods often treat the data itself as something presented to them as a gift, wrapped up with a bow.” The mysterious process by which the data was somehow magically created and handed to the analyst in exactly the form needed to run statistical tests is rarely explored or understood. Obviously, in the real world no magic is involved. And in our Data Science program, we aim to produce graduates who do not depend on such “magic”, and who are capable of taking the initial step themselves. A data scientist with the skills necessary to and, retrieve, import, massage, and structure the information he/she needs to analyze has a tremendous advantage over one who can
perform a principled analysis, but only once some other contributor has done the dirty part of the work. This signature characteristic of the UMW data scientist will thus bring advantages in employability, exibility, and overall project perspective.

We ultimately view the field of “Data Science" as consisting of three parts: (1) input (the data ‘wrangling’ referred to, above), (2) processing (statistical analysis), and (3) output (information visualization to interpret results). All three of these components will be featured in the Data Science minor program, and this course (“Foundations for Data Science) will focus squarely on the first.

Place in the curriculum
Interested students will take CPSC/BUAD 219 as the gateway to the minor. Because this program seeks to draw students from business as well as arts and sciences, we are cross-listing it as both Computer Science (CPSC) and Business Administration (BUAD). Students will learn the R programming language and statistical environment, an ubiquitous and state-of-the-art tool versatile enough to accommodate any data analytics function. (This skill alone will add considerable oomph to students’ resumes.) They will be introduced to a wide variety of information representation methods, input formats, and data sources. They will explore a number of publicly available data sets, driven in part by their own interests, and truly “get to know” the data as they explore its structure, transform it as needed, and bring it under their control. They will also learn how to probe those data sets for preliminary insights. Students will exit the course fully prepared to engage in later courses focused on statistical analysis, having participated in the process that produces the data used in that analysis. Students will build confidence and perspective in seeing the entire end-to-end data science workflow emerge before them.

Impact statement | BUAD/CPSC 219
The resource requirements to support BUAD/CPSC 219 are fairly minimal, and have been identified as follows:

1. Staffing the course. We anticipate offering this course once per year to support the Data Science minor as well as interested students who wish to take it as an elective. We believe this is easily accomplished by rotating between the following sources:

-CPSC full-time faculty. Drs. Stephen Davies and Karen Anewalt are both willing and able to teach the course on a periodic basis. Too, we are actively seeking the Data Science specialization in our current tenure-track job ad, so we may even have a third contributor in this area.

-BUAD full-time faculty. Dr. Chris Garcia, co-author of this proposal, is able and willing to teach the course on a periodic basis.

-CPSC adjunct faculty. Drs. Dave Marchette, Elizabeth Hohman, and Jeff Solka are Ph.D. statisticians and scientists who work for the U.S. Navy at the Dahlgren Naval Surface Warfare Center (NSWC). Dr. Davies and Dr. Debbie Hydorn (Mathematics) have been collaborating with these three a great deal over the past couple of years, and they have been instrumental in contributing ideas for the Data Science program. Drs. Hohman and Solka have
both adjuncted for the CPSC department in the past, have indicated they are willing to do so again, and have specifically expressed interest in this course. Dr. Marchette is also interested in joining our adjunct faculty at some point. Given these three pools to pull from, it shouldn't be difficult to staff the course at the required levels. If and when the minor becomes popular enough to increase demand, we will address the resources required to increase the frequency of offering.

2. Active learning lab/classroom. The course lends itself to hands-on learning in which students can collaboratively work with data and programs during lecture time. We are excited to offer the pilot “special topics” version of the course this spring in the ITCC’s room 327, which is exactly suited to that purpose. Unless that room and similar learning spaces become too scarce and thus unable to handle the University’s needs, we anticipate no troubles with this resource requirement either.
“BUAD/CPSC 219 — Foundations for Data Science”
Professor: Stephen Davies
Fall semester 2015
Class: MWF 10am–11am, ITCC 327
Final exam: Dec. 11, 12pm-2:30pm

Course Description
Data science involves applying advanced techniques from statistics, machine learning, operations research, and information visualization to predict, classify, and find interesting insights in data. In order to apply these types of analytical methods, however, it is necessary to collect, clean, and merge data from multiple sources including social media, databases, geospatial sources, and many others. This requires both a firm understanding of data manipulation techniques as well as programming capabilities in order to execute these techniques. In this course we will be learning how to do this type of “data wrangling” using the R statistical computing environment.

Learning Outcomes
Upon successful completion of this course:
• Students will be proficient in basic programming techniques using the R language
• Students will be able to programmatically manipulate raw data in both tabular and hierarchical formats from sources including files, API’s, and databases
• Students will be knowledgeable about data cleaning, imputation, and re-coding principles
• Students will be able to recognize and work with standard data set types including time series, graphs/networks, GIS coordinates, and feature sets
• Students will be proficient in exploratory data analysis and basic information visualization

Assignments and Grading
There will be two types of graded assignments: 1) mini-exercises, and 2) projects. Miniexercises will occur approximately weekly and will help you get your feet wet with the week's featured packages and functions. They'll be scripted, connect-the-dots affairs. Projects will be more substantial, and will allow you to choose data source(s) that you're interested in from a variety of possibilities.

Activity Contribution
Mini-exercises 20%
Project #1: file-based data sources 15%
Project #2: API-based data sources 15%
Project #3: data fusion from multiple sources 15%
Project #4: exploratory data analysis 15%
Final exam 20%

Rules of the game
• There are absolutely, positively, NO stupid questions!! Your job is not to already know everything before you start the course. Your job is to try hard to learn, and part of that involves asking questions. I'm a nice guy, and I will not ever belittle you, snub you, or make fun of you; and if anyone else does so I will personally
break both of their arms.

• This class will be interactive. When I point at you in class, say your first name, and be prepared to try and answer questions. (Don't worry if you don't know all the answers.)

• You will be held responsible only for the material I cover in class. Throughout the semester I will recommend several books, websites, and other materials, and encourage you to choose from them for extended study. But the final exam, and the various assignments, will only explicitly require you to know lecture material.

• Don't skip class. Just don't. It's bad form. I work hard to prepare for class, to make it compelling and relevant. It hurts my feelings when you don't come. Plus you miss out on important stuff, and you'll end up falling behind if you skip lecture. So come every time. Come happy, fresh, excited, ready to think and to participate.

• Don't cheat. Cheating is heinous, rude, and bad karma. It really makes me mad, and it will also eat away your character like hydrochloric acid if you're not careful. If you ever feel tempted to cheat, in this class or any other, come and talk to me about it. It's not wrong to feel tempted, and we can find other ways out of whatever dilemma you're facing without compromising your moral character.

The Honor Code and this course
I strongly believe in UMW's honor code and scrupulously adhere to it. The majority of the work you will do outside of class involves writing programs in the R data analysis environment. You are not only permitted, but encouraged to discuss all assignments with your fellow students, weigh alternative solutions, work side-by-side in the lab, etc. Collaboration is, in fact, an important part of data science. However, you must turn in your own code. This will be impossible to fake on the projects, since each student will be selecting their own data set and their own “angle” to the analysis. For the mini-exercises it will be easier to fake, but you are honor-bound not to fake it. You must not simply copy a fellow classmate's solution and turn it in as your own, even if you worked with them at least partially.

Ethics aside, you'll quickly realize that it's not in your best interests to do this anyway. The point of the mini-exercises is to help you gain proficiency in the techniques. If you don't do that, (1) the projects will be near impossible, and (2) you won't get anything meaningful out of the class.

Late policy
No late work will be accepted this semester. Get your stuff in on time!

Basis for determining mid-semester reports
For midterm progress reports, I'll be looking mostly at your mini-exercises: whether they are correct, complete, and on time. If you're keeping up with those, and understanding what they're teaching you, that's a really good sign. I'll also consider your project #1, which should be of adequate quality. If either or both of these categories are lacking, it's a sign of danger, and I will give you a "U" for your mid-semester grade. Please don't hesitate at all to come talk to me about this so we can figure out how you can do better in
the course.

**Guidelines for class participation**

I believe that students learn best when they participate wholeheartedly in all aspects of the learning process. Hence while your grade will not be partially determined by any “class participation score” *per se*, it is very much to your advantage, and very much recommended, that you join in during class discussions, ask questions, and make comments.

Note also that this will be an active, hands-on learning experience in which we take advantage of what the Convergence Center has to offer. You'll be getting your hands dirty with code and data nearly every class period. Come prepared to explore!

**Disabilities**

If you have a documented disability, please present me your letter from the Office of Disability Resources and I'll be happy to accommodate you.

**Tentative schedule**

I **Programming Concepts**

2 Intro to R and Data Science  
2 R programming (part 1)  
3 R programming (part 2)  
4 Data Processing Methods

II **Data “in the wild” – overview of formats, availability, licensing, and data integrity issues**

5 Working with tabular data and file sources  
6 Tidy data - Data formatting, cleaning, and imputation  
7 Working with databases and API’s (part 1)  
8 Working with databases and API’s (part 2)  
9 Working with GIS and graph data  
10 Data fusion and dimension reduction strategies (part 1)  
11 Data fusion and dimension reduction strategies (part 2)  
12 Exploratory Data Analysis

III **Principles of exploratory data analysis**

13 Data visualization methods (part 1)  
14 Data visualization methods (part 2)  
15 Using data to form hypotheses and frame further research  
16 Final Exam Week