Undergraduate Research Project in Data Science (Computer Science)

Faculty Advisors: Ravi Gandham (Computer Science), Emilia Gan (Computer Science)

Proposed length of UGR Project: 2 quarters/student

Introduction

As it becomes ever easier and cheaper to collect and store vast amounts of data, all fields of human endeavor are transforming into data-intensive fields. No matter what career path a student selects, he or she will need to be adept at manipulating, analyzing, and interpreting data at scales never required previously. As William Buxton describes, technology has advanced to the point where the limiting factor currently is human capability (Buxton, 2001). Today’s "big data" requires more processing in order to extract knowledge from data sets that are often noisy, in inconvenient formats, and, simply, too large to useful unless somehow presented on a more “human” scale. This proposal presents a project that will provide students with experience in framing a research question and seeking out appropriate publically-available data to answer that question. The student(s) will learn how to approach a data set, how to perform initial exploratory data analysis on the data, how to conduct simple statistical analyses, including model building, to test one or more hypotheses relevant to the data, and how to interpret and present the results of the investigation.

Data science is a field that is getting a lot of attention at present. Hal Varian (chief economist, Google) stated, “The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—that’s going to be a hugely important skill in the next decades... (Varian, 2009). According to a market analysis performed by the McKinsey Global Institute, in the United States alone, there is a shortage of close to 200,000 people with deep analytical skills and a shortage of 1.5 million managers and analysts who can effectively use the results of big data analytics (Manyika et al., 2011).

Data science is an interdisciplinary field requiring at least some level of expertise in three main areas. As demonstrated in the figure below (Conway, 2010), data science involves math and statistics knowledge, combined with expertise in a specific domain (the area from which the research question arises), and supplemented by hacking (command line/ text manipulation/ scripting/ programming) skills.

Figure 1: Data science Venn diagram

This Venn diagram was constructed during a discussion of how to create a data science curriculum and represents the skills considered to be essential for any data scientist.

While not all data scientists need to be equally expert in all three areas, a minimum level of competency is required in each. Participating in this project will show students how data science requires the integration of knowledge, as students will need to draw on their backgrounds in each area or else learn how to educate themselves as needed as their projects progress.

Hacking skills are generally taken to include the ability to search out answers to questions and troubleshoot problems that arise over the course of a data science project. New software tools and platforms are being developed constantly, so a data scientist who is not able to learn new skills on his or her own will soon become obsolete. Although the faculty advisors will always be available to consult with students, it is expected that over the course of the research project, students will become more self-sufficient in determining what directions to take their analyses and in learning how to use the relevant tools.

Research Questions

Given the abundance and variety of publically available data sets, the student(s) involved in this project will have considerable freedom in the selection of their research question. The methods needed to conduct the project will be largely similar, regardless of the domain area of the question. In any data science project, the student(s) will need to formulate a research question, find data likely to provide insights into the question of interest, clean the data, perform exploratory analyses of the data, develop hypotheses to test and perform some basic statistical analyses to address these hypothesis, interpret the results obtained and present research findings. While certain tools or algorithms may be more suitable to one field or another, or to one dataset or another, an important part of learning data science is learning when to apply a specific method or approach.

Research questions could originate from any of a multitude of study areas. In medicine/healthcare datasets could be obtained from sites such as HealthData.gov (http://www.healthdata.gov/dataset/search) or the Centers for Disease Control website (http://www.cdc.gov/nchs/data_access/sets/available_data.htm). Students with an interest in biology or genetics/genomics could work with datasets from organizations such as the National Cancer Institute (http://epi.grants.cancer.gov/dac/) and the Broad Institute (MIT) (http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi). Students passionate about a particular sport could search the data sets that are available at http://databasesports.com/, a site which claims to have the largest sports statistics database online, while students who prefer focusing more on business issues could choose to work with data from sites such as USA.gov (http://www.usa.gov/Business/Business-Data.shtml), the National Bureau of Economic Research (http://www.nber.org/data/), or The Data Page of the Stern School (http://pages.stern.nyu.edu/~adamodar/New_Home_Page/data.html). For research involving social media, freely available data sets can be obtained via the Stanford Network Analysis Project (SNAP) (http://snap.stanford.edu/data/index.html) or from Yelp (https://www.yelp.com/academic_dataset). In short, once the student has come up with a question of interest, there should be sufficient options for finding appropriate data to examine.

While our preference would be for students to investigate questions of their own, and will meet with students to provide assistance in formulating research questions, we have prepared a number of “ready-to-go” options to serve as examples. These are described briefly below:
1. **Medicine:**
   a. Question: Can gene expression data be used to gain insights into inducing differentiation in leukemia cells?
   b. Data sets: Data on gene expression in chemically treated untreated leukemia cells from the Broad Institute website

2. **Social Media:**
   a. Question: Can sentiment analysis of Yelp reviews be used to predict star ratings?
   b. Data set: The Yelp academic dataset

3. **Economics:**
   a. Question: How do employment status and income level correlate with expectations of crime victimization?
   b. Data set: The Survey of Economic Expectations data available on the National Bureau of Economic Research website

Some sample project proposals are included in Appendix A of this document. Students participating in this UGR project may choose to work with one of these proposals or may develop questions of their own to investigate. We will encourage students to gain experience searching for publically available data to analyze, as this experience will be useful to them if they wish to pursue further investigations in data science on their own in the future.

**Experimental Methods**

By participating in this UGR Project, students will become familiar with the basics of how a data science project is conducted and with many of the basic tools of the data science researcher. The methods that all students will learn are: formulating a question – the most important step of any data science project, collecting data and preparing it for use, statistical analyses of the data, interpretation of results, data visualization and communicating research findings. These steps, and the tools likely to be used at each stage, are described in detail in following sections.

**Experimental Design / Proposed Schedule**

The proposed timeframe for this UGR Project is two quarters. If an interested student is only available for one quarter (e.g. due to graduation schedule), an abbreviated version of the project may be arranged on a case-by-case basis. A student working on this UGR Project can expect to follow a schedule similar to the one listed below (modeled after the life cycle of a data science project as described by Dr. Vincent Granville (Granville, 2014):

1. **Identify a research problem (2 weeks):** The student will meet with faculty advisors to explore areas of interest and to determine what background reading should be done in preparation for starting the research project
2. **Identify available data sources (2 weeks):** The student will submit possible data sources to faculty and describe how the proposed data may assist in answering the research question
3. **Examine the data sources and clean data as needed (2 weeks):** The student will use a variety of tools (text editors, Excel, R, Python) to perform exploratory data analysis and data “munging” (parsing and formatting data) to determine the adequacy of the data for the research question being investigated and to get ensure that the data are in usable formats.
4. **Formulate hypotheses to test (2 weeks):** Based on the initial overview of the data, the student will identify one or more hypotheses to test.

5. **Perform statistical analyses of the data (4 weeks):** The student will use the R programming language to perform statistical analyses of the data. Depending on the student’s background, more or less time will need to be spent learning to use R and learning the theory behind the statistical tests. At a minimum, basic hypothesis testing will be done. More advanced students may have time to work on clustering their data or doing more extensive building and testing of models.

6. **Interpret results of data analysis (4 weeks):** The student will learn to examine research results critically to decide whether or not the results obtained are as relevant to the question under investigation as was anticipated. The student will consider the potential differences between “statistical significance” and “practical significance.”

7. **Write up results for presentation (4 weeks):** Since data visualization is a critical subfield of data science students will use the graphics capabilities of R and/or Python packages to create publication-quality graphics to highlight their research results. These data visualizations will be included in any written or oral presentations prepared by the student(s). In addition, both R and Python have report generating capabilities that students will learn to use in order to share their code and their project documentation and to create their research portfolios. Students will be encouraged to make use of on campus resources such as the Writing Center to get feedback on drafts of their written research reports, and should gain an appreciation for the importance of being able to effectively communicate their research findings.

**Total time: 20 weeks (2 quarters)**

**Ongoing throughout entirety of project period:** developing foundational research skills such as careful record keeping (facilitated by the use of git (software configuration management system) for version control) and working with collaborators (facilitated via the use of GitHub to host student code and share code with faculty advisors). If a student is not familiar with git and GitHub, the faculty advisors will assist students in setting up the necessary accounts and getting started. In addition, there are numerous online resources, such as the short tutorial at https://www.kernel.org/pub/software/scm/git/docs/v1.5.0.7/core-tutorial.html. By keeping careful research records and using version control, this project aims to promote good practices for reproducible research.

**Analysis of Results**

Students will perform basic descriptive statistical analyzes to gain insights into their data. These may include, but will not be limited to, looking at subsets of their data, calculating summaries of their data (central tendency, variability, quantiles, etc.), looking at graphical representations of their data (e.g. scatter plots, box plots, histograms). Students will learn how to subset their data to explore specific properties of their dataset. As their projects progress, students will likely formulate simple models to address features of their research questions (e.g. ordinary least squares regression). They may use subsets of their data to evaluate how well their models seem to match the “reality” represented by their data and to evaluate their models’ potential usefulness in making predictions. Given time and student interest, multivariate methods or elementary machine learning algorithms may be explored as potentially helpful analytic tools.
All results will be considered critically, and faculty advisors will meet with students to discuss possible implications of research findings and to brainstorm on where the students might expand their research, if they were to continue exploring their research questions in the future.

**Ethical Concerns**

Since this UGR Project, as proposed, will use only publically available datasets, there should be no ethical concerns regarding the use of the data. However, as was recently addressed at a Meetup attended by 150 people in Brooklyn (DataKind, 2014), the abundance of easily mined personal data is giving rise to numerous significant ethical concerns. These issues will be discussed as part of the UGR Project. Particular attention will be paid to how the vast quantities of personal data collected, often without our explicit awareness, can be used in both beneficial and detrimental ways.

**Safety Concerns**

There are no anticipated safety concerns.

**Materials Requested**

No special materials are specifically required to conduct this UGR Project. Ideally, any student participating in this project would have access to a computer with the appropriate software installed at all times. All of the software we are planning on using is free and open source. While the project could be conducted using campus computing resources, specific software packages would need to be installed. In the case of the R work the student will be doing, the student would optimally be able to install any R package required as the need for it arises. Since downloading software onto school hardware is often restricted to members of the IT department, it would be convenient if the student were to do the majority of his or her work on a personal computer on which he or she has administrative privileges. In the case of students who would like to participate in this research project, but who do not have personal computers of their own, it would be extremely helpful if South Seattle College could make "loaner" laptops available to students for use during the duration of the research project.

We propose that students involved in this research project should be funded to become student members of the Association for Computing Machinery and the Knowledge Discovery in Data special interest group. Since Data Science is at the early stages of development and is evolving rapidly, having access to authoritative and peer-reviewed scholarship and literature is essential for the students to stay current with the latest developments. The proposed student memberships will give them this access. Specifically, the following are included among the many advantages associated with the student memberships to ACM:

- access to the Learning Center, which provides online courses, books, and webinars
- e-mentoring via a mentor network
- access to educational software for data modeling
- electronic subscription to both the ACM’s student magazine and member magazine with articles of interest in all areas of computer science, including data science
- networking opportunities and access to the ACM career center
We also request funding for Python and R reference books that students can use during the project. These books will assist them in their data manipulation and data analysis tasks and can form the foundation of their personal “data scientist toolkits”. Since students will be required to present the results of their data, we will also be requesting funding to cover the costs associated with preparing research posters for their presentations. The supplies requested and estimated associated costs are provided in Appendix B of this document.

Location

The research will be conducted at South Seattle College and at the students' places of residence. To conduct this UGR project, students will simply need access to the internet and to a computer with the necessary software installed. Throughout the course of the research project, students will be encouraged to expand their knowledge of data science, including attending relevant lectures in the Seattle region. For example, the University of Washington eScience Institute, iSchool, Computer Science department, and Human Centered Design and Engineering department often sponsor events of interest to data scientists. In addition, Seattle has a vibrant community of people with interests in data science. Local Meetup groups such as Seattle Big Data Science and Data Science Dojo also sponsor talks that could be of interest to students involved in this project. Generally, a faculty advisor will try to attend any outside session that a student is interested in attending.

Faculty Advisor and Student Hours Estimates

We estimate that faculty advisors involved in this UGR Project will need to spend an average of 3 hours/week on it. Students will need to spend at least 5 hours/week on the project. The scope of the project and the potential for student learning will directly correlate to the magnitude of the time commitment the student is able to make. Faculty advisors will work with students to “benchmark” the projects, so that work is done consistently throughout the quarter(s) that each student is involved in the research. Students wanting to participate in only one quarter will be advised in the selection of simpler research topics, to ensure that they are able to complete an analysis in the time they are able to commit.

Background and Required Skills

Mathematics Background: Students interested in participating in this Data Science UGR Project should have completed math through college algebra (MATH 102). An introductory statistics class would be helpful, but will not be required, as the faculty advisors can provide assistance with necessary background for the basic statistical analyses planned for this project.

Computational Background: Students will be using computational tools to work with their dataset and will need to do some programming. Consequently, students should have taken at least one introductory programming class (either CSC 110 or CSC 142). Students will use either Python or R (or both) to analyze their datasets. Specific experience in these languages is not required, as the fundamentals of programming learned in one language are generally transferrable to others, but students should be ready to put in some effort to gain familiarity with at least one new programming language during the course of this project. The following online resources will be used, as needed, to help students get up to speed with the computational skills necessary for this project:
**Python programming:** For students with previous experience coding in a different language, the codecademy.com Python course, estimated to take 10-15 hours to complete, provides an introduction to Python’s syntax and features (http://www.codecademy.com/tracks/python).

**R programming:** Introduction to R (https://www.datacamp.com/courses/introduction-to-r) and the swirl R programming course (http://swirlstats.com/) are both short, introductory classes that can be completed within 2-3 days.

**Communication Skills:**

Students will be sharing the results of their research in both oral and written presentations. Accordingly, completion of college English classes (ENGL 098 or higher) is recommended.

**Presenting Research Results**

Students will be strongly encouraged to prepare a research poster describing their projects and presenting at research symposia either through the Ready, Set, Transfer (RST) program or the UGR program at South Seattle College (or both) or at the undergraduate research symposium at the University of Washington. Students will also be encouraged to consider presenting their research to interested student groups on campus – these groups are anticipated to differ based on the choice of research question.
Appendix A: Sample Research Question Summaries

Research Question 1 (Medicine): Can gene expression data be used to gain insights into inducing terminal differentiation in leukemia cells?

Background: Microarray data provides information about gene expression. When data is collected under different conditions, differences in the patterns in gene expression can be used to characterize a given condition of interest. In a rare subtype of acute myelogenous leukemia (AML), treatment with a compound capable of inducing terminal differentiation of the leukemic cells results in clinical remission. Identifying compounds that are able to reproduce this effect in more common forms of AML would be of great benefit. Using publically available gene expression data from patients with AML, it should be possible to analyze the gene expression patterns to gain insights into the differentiation status of the leukemic cells after treatment with various chemical compounds. Such information could potentially help in formulating therapies for AML.

Possible questions to investigate include:

• Can we identify specific patterns of gene expression that correlate with a terminally differentiated cell state, thus allowing more efficient screening of compounds potentially capable of inducing leukemic cell differentiation?
• Can we formulate hypotheses as to possible mechanisms by which terminal differentiation is induced by identifying genes which are consistently upregulated or downregulated?

Research Question 2 (Social Media): Can sentiment analysis of Yelp reviews be used to predict star ratings?

Background: Sentiment analysis, also known as opinion mining, refers to the use of text analysis to draw conclusion regarding the emotional state of the author of a given sample of text at the time the text was written. Sentiment analysis scores words based on how commonly they are associated with negative or positive sentiments. This data is especially useful to businesses in evaluating their online reputations as determined by online reviews and references in social media sites.

Possible questions to investigate include:

• Can we reliably analyze reviews on Yelp and predict how positive or negative the review is?
• Can we create a reliable scoring system for words commonly found in Yelp reviews?
Appendix B: Supplies List

6 copies of each of the following books (4 student copies, 2 mentor copies):

Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython by Wes McKinney -- $25.24 on Amazon

http://www.amazon.com/Python-Data-Analysis-Wrangling-IPython/dp/1449319793/ref=sr_1_2?ie=UTF8&qid=1416156721&sr=8-2&keywords=data+science+python

R for Everyone: Advanced Analytics and Graphics (Addison-Wesley Data & Analytics Series) by Jared Lander -- $26.67 on Amazon

http://www.amazon.com/Everyone-Advanced-Analytics-Graphics-Addison-Wesley/dp/0321888030/ref=sr_1_1?ie=UTF8&qid=1416157088&sr=8-1&keywords=r+for+everyone

SUBTOTAL: 311.46 + shipping costs

4 Student Memberships to ACM and SIGKDD

ACM (Association for Computing Machinery)
The world's largest educational and scientific computing society, delivers resources that advance computing as a science and a profession. ACM provides the computing field's premier Digital Library and serves its members and the computing profession with leading-edge publications, conferences, and career resources.
$19/year (student membership)

SIGKDD - Knowledge Discovery in Data
SIGKDD's primary mission is to provide the premier forum for advancement, education, and adoption of the "science" of knowledge discovery and data mining from all types of data stored in computers and networks of computers. SIGKDD promotes basic research and development in KDD, adoption of "standards" in the market in terms of terminology, evaluation, methodology, and interdisciplinary education among KDD researchers, practitioners, and users.
$15/year on top of ACM Membership (student membership)

SUBTOTAL: $136

Printing

Printing costs to print two professional style posters (3' x 4', color posters).
If printed outside ~$50/poster (printing may be available at South Seattle College at a lower cost)
Poster tubes to carry/protect posters (~ $10 or less)

SUBTOTAL: $120 (or less)

APPROXIMATE TOTAL: 567.46 (total for 4 students involved in 2 projects)
References and Resources Cited


Online sources of publically available datasets:

HealthData.gov: http://www.healthdata.gov/dataset/search
Centers for Disease Control: http://www.cdc.gov/nchs/data_access/sets/available_data.htm
National Cancer Institute: http://epi.grants.cancer.gov/dbc/
Broad Institute (MIT): http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi
databaseSports.com: http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi
The Data Page (NYU): http://pages.stern.nyu.edu/~adamodar/New_Home_Page/data.html
Yelp: https://www.yelp.com/academic_dataset

Software Tools:

The R Programming Language: http://www.r-project.org/
GitHub and git: http://www.git-scm.com/, https://github.com/

Online tutorials:

Python: http://www.codecademy.com/tracks/python
GitHub/git: https://www.kernel.org/pub/software/scm/git/docs/v1.5.0.7/core-tutorial.html