An Undergraduate Program in Data Science

Jim Albert and Maria Rizzo
Department of Mathematics and Statistics
Bowling Green State University

August 2016
The Modern Statistician

- How to meet statistical needs with the collection of “Big Data”?
- Traditional curriculum – focuses on particular methodologies such as regression and ANOVA
- More common now for statisticians to work with researchers from other fields
- Apply flexible problem-solving to solve statistical problems
New ASA curriculum guidelines

- Focus on undergraduate programs in statistical science
- Increased importance of data science
- Real applications
- More diverse models and approaches
- Ability to communicate
Not just running statistical packages

Broader view of statistical computing – how to access and integrate large amounts of data?

Deeper understanding of computational reasoning (programming languages, writing algorithms)

Teach statistical computing in the context of statistical practice
Topics to Add to the Traditional Curriculum

- Exploratory data analysis
- Fundamentals of statistical graphics – creating new types of visualizations
- More general view of statistical modeling
- Simulation-based inference
- Statistical programming (using R)
How to Include More Data Science?

- A number of schools are offering “computing with data” courses
- Computational aspects of the data analysis cycle
- From data acquisition and cleaning to data organization and analysis and reporting
- What are the backgrounds of the students and what are their needs?
The Data Science Specialization

- BGSU received a five-year grant “EXTREEMS-QED: Computational Mathematics and Statistics Program” funded by National Science Foundation.

- Grant gives support for:
  - development of the program
  - scholarships to help recruit new majors
  - a new active learning space for the data science courses
Data Science Specialization Foundation

- Three calculus courses and linear algebra
- Calculus-based probability course
- Programming fundamentals (CS department)
- No early exposure to Statistics
<table>
<thead>
<tr>
<th>Year</th>
<th>Course 1</th>
<th>Course 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y1</td>
<td>Calculus 1</td>
<td>Calculus II</td>
</tr>
<tr>
<td></td>
<td>Seminar in Data Science</td>
<td>Intro to Programming</td>
</tr>
<tr>
<td>Y2</td>
<td>Calculus III</td>
<td>Linear Algebra</td>
</tr>
<tr>
<td></td>
<td>Computing with Data</td>
<td>Statistical Programming</td>
</tr>
<tr>
<td>Y3</td>
<td>Prob and Stat 1</td>
<td>Statistical Learning</td>
</tr>
<tr>
<td></td>
<td>Regression Analysis</td>
<td>Data Science Elective 1</td>
</tr>
<tr>
<td>Y4</td>
<td>Capstone Experience</td>
<td>Capstone Experience</td>
</tr>
<tr>
<td></td>
<td>Data Science Elective 2</td>
<td></td>
</tr>
</tbody>
</table>
Freshmen Level: Data Science Seminar

- One-hour seminar – goal is to introduce Data Science
- Invite speakers from a range of disciplines
- Topics
  - sports analytics
  - forensic sciences
  - GPS and spatial data
  - data visualization
  - geography
  - text mining
Teach in a computer lab – work within the RStudio IDE

Class work and assignments set up as Markdown files

Creating a notebook of documentation, R code, and output

Encourage group work with a number of projects
Catalog description:

Computational methods for collecting, manipulating, exploring, and graphing data. Basic principles of exploratory data analysis and statistical graphical methods. Methods for downloading and organizing large data sets. All of the computing methods will be illustrated using a high-level language such as R or Python.
Topics in Computing with Data

- Introduction to data science.
- Vectors, variable types, operations, data frames
- Basic data analysis methods (graphs and summaries for 1 and 2 vars)
- Basic data wrangling
- Creating csv files, uploading the files to a server
- Textual data / regular expressions
- Reading in HTML and XML pages
- Mining Twitter data
- Working with mapping data
In baseball, all pitches are measured by the PitchFX system.

Many variables collected on each pitch such as speed, break, location, name of pitcher and batter.

Assignment was to construct an interesting visualization of this data.
“Bigger Data” Project

Students work in groups on exploration of a larger dataset.

- Boston Marathon
- Airline Flight Delays
- Shotlink Golf
- NYC Taxi Cab
- Baltimore Crime
- Pronto Cycle
Catalog description:

Applying a statistics programming language to facilitate the exploration and visualization of data. Basic objects such as data frames, matrices, tables, and lists, and how to perform manipulations with these objects. Writing functions with looping and conditional structures. Use functions to perform simulation-based statistical algorithms. Understand and develop object-oriented programming. Develop manipulations with character data.
Topics in Statistical Programming

- Writing functions, vectorization
- Working with arrays
- Simulation experiments
- Benchmarking code, computational efficiency
- Regression, non-linear models, nonparametric
- Bootstrapping
- Working with financial data
- Integrating C++ code in R
Junior Level: Statistical Learning

- Survey of statistical methods for supervised and unsupervised learning tasks
- Linear regression, classification, nonparametric regression, tree-based methods, etc.
- Text: James et al (2013), *Introduction to Statistical Learning*
- Learn through a set of labs with interesting datasets
Senior Level: Capstone Experience

- Students works on a rich dataset in an application area
- Can be a project with a faculty advisor
- Participate in a REU (summer experience) or an Internship
- Required to write up analysis and present
Several students are working with local sports teams

A student is working with a state crime-lab on campus

Another student is working with the Office of Admissions
ASA Datafest

- 7 BGSU students participated in a Datafest at Miami University in April
- Intensive exploration of a rich dataset over a weekend
- Great experience – encourages work as a team
BGSU Students at ASA Datafest
Challenges and Opportunities

- Data science curriculum
- Recruitment
- Internships?
- Building connections with industry
- Building data science community at BGSU