A Portfolio of Data Analytics Classes at University of Oklahoma

Karen M. Leighly\textsuperscript{1}, Collin Dabbieri\textsuperscript{2}, Alexander Kerr\textsuperscript{1}, Donald Terndrup\textsuperscript{3}

\textsuperscript{1}The University of Oklahoma
\textsuperscript{2}Vanderbilt University
\textsuperscript{3}The Ohio State University
Genesis / Motivation

• Genesis
  • I wrote an NSF Astronomy and Astrophysics Grant proposal to develop an analysis methodology for broad absorption-line quasar spectra in 2014.
  • The methodology involved machine learning techniques.
  • The broader impacts focused on developing a graduate-level course on machine learning in astrophysics.

• The second step
  • I wrote an NSF renewal proposal in 2019.
  • We decided to add undergraduate data analytics classes to the portfolio
Machine Learning

• Astronomy and Physics graduate students and advanced undergraduates
• Used AstroML – accompanying materials for “Statistics, Data Mining and Machine Learning in Astronomy”
• Taught in 2015, 2017, 2020

- Introduction to Python
- Statistics Introduction / Review
- Markov Chain Monte Carlo
- Histograms and Kernel Density Estimation
- K-means Clustering / Gaussian Mixture Models
- Regression and Principal Components Analysis
- Classification / Neural Nets / Deep Learning
- Time Series and Spatial Analysis
Successes

• All students improved their Python programming skills
• Several students fully embraced machine learning techniques
  • Alex Kerr – Enhanced genetic algorithms with neural nets to find better molecular designs, used manifold learning and clustering techniques to identify topological quantum phase transitions
  • Collin Dabbieri – FeLoNET – convolutional neural net methodology to classify quasar spectra

• Several students who obtained data analysis jobs at Boeing, the FAA, and elsewhere cited this class as instrumental in their hiring.
Improvements for Next Time (?)

• Improved homework
  • Frequent (daily) “try this” exercises
  • Longer project-like problems (less recipe based) for 1-2 week HW assignments
Introduction to Research

- Sophomore and junior astronomy and astrophysics majors
- Developed during 2019 PICUP Summer Faculty Development Workshop
- Taught Spring 2020
Useful Reference

• National Academies Report published in 2018
• Useful for grant proposals, administrators, and convincing your colleagues this is a good idea.

“Data Science for Undergraduates Opportunities and Options”
Data Life Cycle - Course Design

• **Data Wrangling** – Accessing & cleaning data; preliminary data analysis
• **Visualization** – Graphical representation of data; characteristics of effective graphical displays
• **Statistical Thinking** – the recognition that all data is influenced by statistics and the effect of underlying assumptions (e.g., normal distribution)
• **Modeling** – fitting physical or empirical models to data; what constitutes a good fit
• **Computational Thinking** – expressing problems and their solutions in a way that a computer could execute
• **Communication Skills** – Sharing the results with your peers and the public; scientific and technical writing
• **(Attitudes towards Research and Science)**
| Lecture 1 - Introduction                  | Lecture 18 – Galaxy Evolution       |
| Lecture 2 – SciServer and Jupyter Notebook Intro | Lecture 19 – Introduction to Convolution |
| Lecture 3 – Python Fundamentals          | Lecture 20 – Reverberation Mapping  |
| Lecture 4 - Plotting                     | Lecture 21 – Sherpa (Spectral Fitting Software) |
| Lecture 5 - Functions                    | Lecture 22 – Black Hole Masses      |
| Lecture 6 – Loading Data                 | Lecture 23 – Astronomical Publications |
| Lecture 7 – Filter Photometry            | Lecture 24 – Writing a Paper        |
| Lecture 8 – Interpolation, Integration, Weighted Mean | Lecture 25 – Image Analysis        |
| Lecture 9 – Colors and Distances         | Lecture 26 – Radial Profile         |
| Lecture 10 - Uncertainty                 | Lecture 27 – More Radial Profile    |
| Lecture 11 - Histograms                  | Lecture 28 – Velocity Dispersion    |
| Lecture 12 - Errors                      | Lecture 29 – Cosmological Simulations |
| Lecture 13 – Galaxy Spectra              | Lecture 30 – Falling Sphere         |
| Lecture 14 – Cumulative Distributions    | Lecture 31 – Simple Hanging Harmonic Oscillator |
| Lecture 15 – Linear Least Squares        | Lecture 32 – Introduction to Rebound |
| Lecture 16 – The Hubble Law              | Lecture 33 – Jupiter Trojan Asteroids |
| Lecture 17 - SQL                         | Lecture 34 - Classes                |
Successes

• All students improved their Python programming skills
• Used the SciServer platform
  • Free access to python computation / Jupyter notebooks
  • Internal integration with SDSS SkyServer
  • Reliable
• Used nearly daily ”try this” exercises to explore concepts
• Converted to online (Zoom) delivery more or less seamlessly
Improvements

• Streamline content
• Reduce / eliminate quizzes and projects
An Ambitious Goal

• Collaborative research:
  • Don Terndrup - Ohio State University
  • Bruce Mason - University of Oklahoma
• Goal: **Develop data analytics pre-post assessment tests**
  • Like the ”Force Concept Inventory”
• **Challenging**, since they should cover the data life cycle, e.g.,
  • Statistical thinking
  • Computational thinking
  • Modeling
  • Visualization skills
• No progress yet – perhaps a draft version for Spring 2023?
Universal Challenge I

• **Range of backgrounds** – graduate class
  - Students with a background in Python can focus on the material.
  - Students without a background in Python struggle and have less time to learn the concepts

• **Range of skills** – undergraduate class
  - Some students already have hypothesis-testing skills.
  - Others struggle with the idea that they may have to try more than one method to successfully solve a problem.
Universal Challenge II

- *Can data analytics be taught?* I have my doubts.
- Students who want to work on my research project need to have excellent:
  - *Attention to detail*
  - *Trouble-shooting skills*
- Are these skills innate? Or developed by other creative activities?
Summary and Future

• We are developing a portfolio of data analytics courses at University of Oklahoma and Ohio State University.

• OU’s contribution:
  • *Machine Learning in Astrophysics* – graduate class taught Fall 2015, 2017, 2020
  • *Introduction to Research* – sophomore-level course taught Spring 2020 and to be taught Spring 2023

• Plans to develop assessment tools for the development of data analytics skills in the undergraduate classes at both OU and OSU.