Teaching Data Science for Classics

This paper will present the design, implementation, and lessons learned of a joint Classics and Computer Science class offered in the fall 2016. The class features two instructors, namely one Classicist and one Computer Scientist. The objective is to equip Classics students with basic awareness and skills in using digital research methods and in applying that knowledge to understanding and interrogating the numerical claims produced by data scientists. Additionally, the course aims to provide experience with a variety of computational tools to approach Classics research questions, topics, and datasets. After this course, students should be familiar with the steps involved in working with data. They will be prepared to ask questions of numerical claims they encounter in publications and online. Where did the source data come from? How was it processed? Why was it visualized in a certain way and how might the visualization help or hinder appropriate interpretation? Students will also be prepared to engage in collaborative research with data scientists.

In order to reach these objectives, each data science module covered during the semester is conducted with Classics research questions and uses Classics data. Our main data sources include the Perseus Digital Library (www.perseus.tufts.edu), the Lexicon of Greek Personal Names (www.lgpn.ox.ac.uk) and the Pleiades gazetteer (www.pleiades.stoa.org). The data science techniques covered target general data literacy: how to read a data analysis, infographics, interactive graphs, and how to interpret numerical claims. This involves becoming familiar with the process of data analysis, including question/hypothesis formulation and the type of analysis to be performed (descriptive, inferential, exploratory, predictive, causal). Tidying data (reshaping, pivoting, melting, splitting, casting) and munging (fusing, normalizing, coping with missing values, etc) as well as analyzing (visualizing, inferring, predicting) are covered in detail.

Basic statistics such as distribution, mean and standard deviation, outliers, and confidence intervals are also given attention. Finally, a variety of data visualization types are explored (line plots, scatter, bubble, area, bar, pie, box) as well as predictive modeling. Because Classics deals so much with text, several weeks are dedicated to natural language processing, with the *Iliad* as the text to be analyzed.

Since this class is designed as an introductory course for first year students, no prior knowledge of either Classics or Computer Science is assumed. The Classics material to be covered was chosen for its relative accessibility, and only basic instruction in reading the Greek alphabet needs to be offered in order for students to be able to process the data. The side-by-side translations and vocabulary tools offered by the Perseus Digital Library also greatly facilitate the accessibility of the Greek and Latin materials. Similarly, no programming or data science skills were assumed, and the instructors chose to use the KNIME platform to perform data analysis. KNIME (www.knime.org) is an open source, modular platform for data analysis. Users can ingest data under multiple formats and manipulate it in the form of workflows. For instance, a student could ingest the information about all doctors listed in LGPN, visualize their chronological and geographical distribution, and then perform a statistical analysis of the number of doctors vs. other professions in the ancient world, while taking corrective measures for possible error or bias stemming from the loss or incompleteness of data.

Since the class aims to teach students how to question and analyze data claims, all instruction is directed at hands-on work. In each lecture, the instructors perform a joint analysis of a given data set to demonstrate data science techniques as well as interpretative methods. Weekly labs let students manipulate data for themselves and formulate research questions. Finally, a semester-long team project offers students an opportunity to experience all the stages

of a data science project, from data gathering and hypothesis formulation to data manipulation and interpretation. This project forms a tangible outcome to the class and can serve as the basis for student theses or other research.

From a pedagogical standpoint, this class is both a challenge and an opportunity for the instructors to move beyond the typical limitations of both Classics and Computer Science classes. From a curricular standpoint, offering this class as an introduction to both Classics and data science to first year students highlights the increasing place of digital methods in Classics and the current tendencies in our field. Finally, this course will serve as the mandatory introductory course in the new M.A. program in Digital Humanities starting September 2017. The innovative character of the course and its deep engagement with source materials in the Humanities speak to the general orientations of the program and the skills its graduates are intended to have acquired upon completion.