Mathematica Militaris

Volume 24 | Issue 1

Article 3

June 2019

Data Science at USNA

Will Traves
United States Naval Academy, traves@usna.edu

In the Cognitive Age, where leaders have to deal not only with incomplete data but also with analysis and decision making in a world that involves overwhelming data, the ability to evaluate information, reason strategically and ethically, and act decisively, will be essential elements of future success. These are skills that can be taught. These are talents that can be developed. The challenges and multi-disciplinary issues of our contemporary world can and should be specifically examined through our naval education programs.

– Education for Seapower
Final Report (Dec. 2018)[1]

I. Introduction

Data Science is an interdisciplinary field that promises to transform every aspect of our society, including business, medicine, education, and our military. The field is inspired by several other movements, such as data-driven decision making, big data, machine learning, and artificial intelligence; however, initiatives in Data Science are distinguished by seeking to make technical tools from these disciplines accessible to a broad audience. Much as the field of Cyber Science developed from computer security when society became more dependent on the internet, Data Science provides a framework for the pervasive and responsible use of data. The United States Naval Academy (USNA) has an active Data Science community consisting of a broad range of scholars from across campus. Our group meets in a weekly seminar, attended by anywhere from 4 to 40 faculty and midshipmen. Many of the faculty involved in this seminar will be funded by the Office of Naval Research to develop Data Science curricular materials this summer. After a short overview of Data Science and its importance to the Navy and Marine Corps, I'll survey some of the ways that civilian institutions are delivering Data Science curriculum, outline a vision for developing Data Science curriculum at USNA, and summarize some of the accomplishments and planned activities of the Data Science group at USNA. This article is intended to spur discussions about Data Science among the service academies by sharing what we are doing at USNA.[2]

II. An Overview of Data Science

Data Science is transforming the way that society and industry operate. In the internet era, massive data sets are commonplace, and sophisticated tools allow us to glean deep insight by mining this trove of information. For instance, deep learning models allow us to more accurately diagnose melanoma, the deadliest kind of skin cancer. Many businesses are using analysis of social media posts to better respond to customer demand and create a more personalized advertising experience. Social media posts can also help the government respond to natural disasters more effectively. And cities around the world (including New York and London) are using vast sensor networks to improve citizen well-being in cost effective ways. As these examples show, scientific, industrial, and military leaders will increasingly need to be savvy consumers of data, able to interpret and communicate the results of data analysis.

Data Science extends beyond the analysis of complex and massive data resources to include the acquisition, storage, curation, and visualization of massive datasets, methods to document data analysis for reproducibity, and ethical considerations[3] related to the proper use of data. Significant tools from mathematics, statistics, and computer science are required to work in Data Science, though undergraduate courses in Data Science can often teach this material in a streamlined manner, making advanced tools available even to beginning undergraduates.

The Department of the Navy (DON)'s 30 Year Research and Development Plan provides guidance on technologies needed to ensure the DON retains and increases its technological advantage in an increasingly dynamic environment, calling out data analytics, automated decision support, data fusion, machine learning, neuromorphic computing, and artificial intelligence among other technologies. The National Defense Strategy talks about the need to develop new technologies including big data analytics, artificial intelligence, and autonomy to ensure the U.S. will be able to fight and win the wars of the future. It is not only Navy and Marine Corps leaders who are interested in developing Data Science programs; the National Academy of Sciences [2] recently recommended that "academic institutions should encourage the development of a basic understanding of data science in all undergraduates."

III. What Civilian Institutions are Doing

Several top schools have a Data Science elective for freshmen, including UC Berkeley, Cornell, Carnegie Mellon University, and the University of Washington. UC Berkeley's program is particularly well developed and Cornell's program is modeled after it. UC Berkeley's introductory class, Data 8: Foundations of Data Science, is taken by over 1100 students each year. The course materials are available online and the instructors maintain a free online textbook for the course [1]. As well, UC Berkeley offers many courses in application areas that use Data Science for students enrolled in Data 8[4]. UC Berkeley also offers a follow-on technical course

[1] The report can be found at https://www.navy.mil/strategic/E4SFinalReport.pdf, last accessed May 6, 2019.

[2] I'd enjoy hearing about what other service academies are doing and I'd love to explore ways to engage in some joint activities. Please contact me if you'd like to be part of this conversation!

[3] Cathy O'Neil's book, Weapons of Math Destruction [3], describes many disastrous situations where opaque and biased algorithms operating at a huge scale have done real damage to our society, often to our most vulnerable citizens. For instance, predictive models for recidivism are being used in sentencing decisions even though those models effectively use racial data, a factor that is illegal in sentencing. As another example, China is widely deploying facial recognition technology and using it to assign 'social credit scores' to its citizens, reflecting how they behave at work and in public, a practice that could move China even further towards a police state.

[4] Connector courses are described at http://data.berkeley.edu/education/connectors. Examples include: Data Science for Smart Cities; Crime and Punishment; Social Networks; Web Data Visualization; Rediscovering Texts as Data; Time Series Analysis: Sea Level Rise and Coastal Flooding; Exploring Geospatial Data; Race, Policing, and Data Science; and How does History Count? Reading and Writing History in the Age of Big Data.

aimed at third-year students, Principles and Techniques of Data Science[5], that focuses on the Data Science life cycle: question formulation, data collection and cleaning, exploratory analysis and visualization, inference and prediction, and decision making. While their introductory course cannily avoids focusing on technology, the advanced class introduces students to state-of-the-art tools. Carnegie Mellon's statistics program builds on its introductory Reasoning with Data course with an advanced course called Methods for Statistics and Data Science. Their computer science program offers an independent course, Practical Data Science[6], focused on data collection and processing, data visualization and presentation, statistical model building using machine learning, and big data techniques for scaling these methods.

Many top schools also have majors in Data Science. These include: The University of Michigan, Virginia Tech, MIT, UC San Diego, and UC Irvine. The majors are usually housed in a business school, a mathematics or statistics department, or a computer science department. Regardless of where they are housed, many Data Science majors require multiple courses from each of these departments. As well, they typically include required or elective courses from outside the core departments, commonly from economics, business, psychology, oceanography, and biology. Many Data Science majors also require a hands-on practicum or capstone course.

Both EdX and Coursera have partnered with top schools (such as Stanford, Johns Hopkins, UC Berkeley, and UCSD) to offer Massive Open Online Courses (MOOCs), though typically not for credit. These can be valuable venues to introduce faculty to the Data Science enterprise. My own experience with MOOCs is that the student benefits proportionally to the time they spend and the skills they bring to the course; I wouldn't recommend relying on a MOOC to teach our midshipmen and cadets, though they can be great resources for instructors. Other for-profit firms such as Metis, General Assembly, and DataCamp offer intense (and sometimes residential) boot camps focused on Data Science, though these are often very expensive and aimed at persons with post baccalaureate degrees. Some non-profit firms – going under the umbrella of the Carpentries program[7] – offer workshops at local institutions given by highly-trained volunteers. Hosting local Software Carpentry and Data Carpentry programs may be a good opportunity for faculty development in Data Science.

A few schools are starting to offer summer programs in Data Science, often aimed at specific groups of students. The Iowa State University Midwest Big Data Summer School[8] is intended for research-bound undergraduates in engineering, mathematics, or computer science. The Microsoft Research Data Science Summer School in New York City[9] attracts underrepresented students from a broad spectrum of majors. The University of Chicago Data Science for Social Good program[10] is aimed at students interested in societal impact. The University of Michigan Big Data Summer Institute[11] focuses on health-care applications. These programs are often several months long and they seem particularly effective in launching undergraduates into Data Science careers and educational pathways. However, their length precludes midshipman and cadet participation. It is worth exploring ways that our service academies can offer a short-term summer Data Science experience aimed at students with significant military obligations.

IV. Data Science Curriculum at USNA

USNA's current Data Science activities are spread out across multiple departments and majors. In the fall, I taught Machine Learning and Artificial Intelligence (SM486[12]) to majors in the Mathematics department. This course was based on Andrew Ng's Machine Learning course on Coursera[13]. CAPT Dave Ruth taught an Introduction to Statistical Learning (SM486) in the spring. In the Computer Science department, Ric Crabbe offered a course in Artificial Intelligence (SI420[14]) in the fall and Gavin Taylor taught a course in Machine Learning and Data Science (SI486L[15]) in the spring. In the Weapons, Robotics and Control Engineering Department, Joel Esposito taught a 1-credit elective course, Deep Learning: The Truth Behind the Hype (ES481A), in the fall. As well, I taught an elective course, Introduction to Data Science (SM286A[16]), in the Spring. This last course was an inspirational survey course based on the Data8 curriculum at UC Berkeley.

This introductory class started with a quick bootcamp-style introduction to Python programming. Midshipmen used a streamlined version of the Pandas module to investigate complex data sets. Midshipmen built simulations and used resampling methods to develop intuition regarding probability and statistics. Subtle ideas in conditional probability, such as Bayes's Rule – which indicates how to revise our assumptions in the presence of new data – were much easier to understand when viewing the results of simulations. This led naturally to resampling techniques, such as bootstrapping, that can be used to assess whether data supports various arguments and claims. The reliance on bootstrapping techniques and simulation allowed us to avoid large amounts of statistical theory while still getting at many of the important ideas in statistics, such as hypothesis testing. Midshipmen in our class were fascinated by the wide variety of applications of machine learning and artificial intelligence. They were excited to build classifiers to detect breast cancer and regression models to predict housing prices near military bases. While artificial intelligence's upside has been widely touted, we also discussed

[5] The course website is http://www.ds100.org.

[6] The course website is http://www.datasciencecourse.org

[7] See http://carpentries.org.

[8] See http://mbds.cs.iastate.edu/2017.

[9] See https://ds3.research.microsoft.com.

[10] See https://dssg.uchicago.edu.

[11] See https://sph.umich.edu/bdsi.

[12] See the course website at https://sites.google.com/site/usnasa486ay19fall/.

[13] See https://www.coursera.org/learn/machine-learning.

[14] See the SI420 tab under the Courses tab at https://www.usna.edu/Users/cs/crabbe/.

[15] See https://www.usna.edu/Users/cs/taylor/courses/si486l/.

[16] See https://sites.google.com/site/sm286ausnaspring2019/.

privacy and security issues with machine learning. For instance we highlighted security challenges posed by personal devices[17] and Twitter posts. As well, midshipmen learned of predatory lenders that target troops with little financial experience[18].

Though our first attempt at an inspirational course (SM286A) was largely a success, there is considerable course development work remaining. The Office of Naval Research is funding further curriculum development in Data Science this summer. We used the online textbook [1], written by UC Berkeley faculty, in SM286A. Some aspects of the text are great: it is very readable, it is free (and free to modify), and it contains code snippets that the reader can modify and run in a browser – making learning an interactive experience. However, it would be good to rewrite portions of the text to raise its mathematical level; indeed, the text sometimes unnecessarily avoids mathematical notation and concepts that our midshipmen have the background to appreciate. All the examples in the text center on UC Berkeley and the Bay Area. It would be better to develop examples relevant to the Navy and Marine Corps (or, more broadly, the armed forces) for use in our course. A preliminary list of topics we intend to introduce includes: using predictive analytics to schedule repairs of complex machinery based on sensor-measured operating profiles rather than more conservative maintenance guidelines; classifying ships as friend or foe from image or other sensor data; using just-in-time supply chain management to support deployed troops and military hospitals; using predictive analytics to improve the Marine Corps's recruiting efforts; and modeling sea level rise and its impact on military facilities. Several of these topics have been addressed by midshipmen as part of their capstone or honors projects, and we intend to highlight their efforts to inspire new students. Midshipmen in our first course raved about the projects, which were assigned and completed as Python notebooks. They really enjoyed being able to write solutions that included text, formulas, and sophisticated graphics. I'd like to provide some instruction on how to get the most out of the markdown system, particularly as some of our Operations Research majors use this system to write their capstone projects. Finally, in order to easily scale the course to a larger audience, we'll need to implement an autograder to do the bulk of the grading. UC Berkeley developed such a solution (and offered to automatically grade our papers!) but they deploy their autograder in the cloud and USNA can't store grades in nonsecure settings, as grades are currently considered personally identifiable information (PII). We can implement a similar solution at USNA, but it will require close coordination with our IT staff.

This summer we expect to develop a follow-on class, Practical Data Science. This would be a blend of UC Berkeley's Principles and Techniques of Data Science and Carnegie Mellon University's Practical Machine Learning. As in the introductory course, we'd like to focus on problems faced by the Navy and Marine Corps, highlighting potential capstone research projects and internships for midshipmen. It is our long-term ambition that this class and the inspirational survey course will find their way into USNA's core curriculum, though finding space for additional core material is a difficult task.

We expect to continue to offer technical courses related to Data Science. One interesting development is that in the Fall Professor Evelyn Lunasin and I will teach the Intermediate Linear Algebra course, required of all Mathematics and Applied Mathematics majors, with an emphasis on Data Science topics, including neural networks, low rank approximation, and principal components analysis. The course will use Gil Strang's new book, Linear Algebra and Learning from Data [4].

We also hope to develop lesson plans for core courses this summer. Since Data Science is such a pervasive topic, it can be woven into the curriculum of core classes already required of all midshipmen at USNA. For instance, we hope to discuss data privacy issues in the era of big data and machine learning systems in our Cyber Security class. Ways that machine learning allow advocacy groups and political candidates to target voters with distinct (and sometimes contradictory) messages on social media could be explored in our U.S. Government and Constitutional Development class. We could include material on gradient descent for fitting machine learning models in our Multivariable Calculus class and we could add projects on Data Science to our Introduction to Applied Mathematics class.

Setting up connector courses is a key plank in our vision to spread Data Science education across the Academy. Courses in the Humanities and Social Sciences that tackle large-scale problems using quantitative methods are important opportunities to increase data literacy among midshipmen and these courses can serve as interdisciplinary research incubators, raising important new research questions. UC Berkeley is somewhat further along than USNA in this regard and already has a stable of connector courses that require their Data Science class as a prerequisite. For several years the Defense Information Assurance Program (funded by the National Security Agency) funded faculty to develop course materials involving high performance computing. This project has been highly successful, and we'd like to mimic it for Data Science, funding course development for lesson plans and connector courses. An example of a connector course is the Climate and Migration course planned by Brad Barrett (Climate Science), Sharika Crawford (Latin American History), and Sylvia Peart (Applied Linguistics). They have been working to explore the links between, and consequences of, climate shocks and human migration. Climate shocks can

[17] For details, see https://www.theguardian.com/world/2018/jan/28/fitness-tracking-app-gives-away-location-of-secret-us-army-bases, a story in the Guardian from January 28, 2018.

[18] On Sept. 6, 2018 the Federal Trade Commission settled a complaint against two companies, Sunkey Publishing and Fanmail.com, for over $12 million. They operated websites like army.com and armyenlist.com. The websites prompted consumers to submit their information to learn more about joining the armed forces. The companies then sold the information as marketing leads to for-profit post-secondary schools for $15 to $40 per lead.

be short term (hurricanes, flooding) or long term (drought); either can influence human migration patterns. Working with two publicly available datasets[19] our team is developing new Data Science methods to examine the links between climate shocks and migration between the U.S. and Mexico and the dry corridor of Central America. This research effort has already produced four midshipmen capstone projects, two peer-reviewed publications, several guest lectures, one evening town-hall forum, and three conference presentations. The aim of the connector course will be to build on this success and spread knowledge of Data Science methods to a wider group of midshipmen. I've chosen this example to explain the kind of material that might appear in a connector course because the content aligns very well with the DoD strategic plan to build and train a Navy and Marine Corps with appropriate skills, regional expertise, and cultural capabilities to meet current and future conflicts. Moreover, the course studies particularly sensitive regions, where climate shock events in the near future could destabilize already unstable states, driving cross-border migration and creating a U.S. national security problem.

V. Data Science Activities at USNA

A growing community of faculty and midshipmen at USNA are working in Data Science, Machine Learning and Artificial Intelligence. We have a weekly interdisciplinary lunchtime seminar in which faculty and midshipmen meet to learn about cutting-edge techniques and applications. We've competed in machine learning contests, such as IEEE's Large Scale Semantic 3D Reconstruction Challenge[20] (classifying objects using LIDAR data from airborne drones) and sent over 20 midshipmen to participate in an eight day Data Science contest at the University of Maryland[21]. Faculty attended DoD conferences like the Naval Research and Development Enterprise Data Science and Analytics Workshop in November and reported back to our seminar[22]. We are helping to organize (and hosting at the Naval Academy) the summit event for the USN Data Sustainment Challenge in June[23]. In the future we hope to hold a hackathon (a workshop designed to identify and solve problems) at USNA, possibly one focused on improving USNA security. We already do some consulting with Navy and Marine Corps units – mostly through capstone projects in Engineering, Operations Research[24], Applied Mathematics, Computer Science and Cyber Science – and hope to further develop our consulting activities in the future.

Acknowledgments

I'd like to thank all the members of the USNA Data Science community. I've tried to include ideas discussed with many faculty at USNA, including: John Burkhart, Brad Barrett, Chris Brown, Rob Correll, Ric Crabbe, Sharika Crawford, Joel Esposito, CAPT Jay Foraker, Gary Fowler, Peter Guth, Manuel Jusino, CDR Gary Lazzaro, Evelyn Lunasin, Luke McDowell, Kevin McIlhany, Brice Merlin Nguelifack, CAPT (ret.) Rob Niewoehner, Sylvia Peart, LT Erick Roman-Perez, CAPT Dave Ruth, Gavin Taylor, Nelson Uhan, Naomi Utgoff, LCDR Jonathan Vance, Doug Vanderwerken, Max Wakefield, Deb Wheeler, and Vrej Zarikian. I'm particularly grateful for midshipmen that faithfully attended our seminar: Jacob Klosowski, Steel Templin, and Michael Wallace. Our Academic Dean, Andy Phillips, and our Director of Research, Reza Malek-Madani have been, and continue to be, wonderfully supportive of our Data Science efforts.

References

[1] Ani Adhikari and John DeNero (with contributions by David Wagner and Henry Milner). Computational and Inferential Thinking: The Foundations of Data Science. Last accessed from https://www.inferentialthinking.com/ on 9 MAY 2019.

[2] National Academies of Sciences, Engineering, and Medicine. 2018. Data Science for Undergraduates: Opportunities and Options. Washington, DC: The National Academies Press. Last accessed from http://sites.nationalacademies.org/cstb/currentprojects/cstb_175246 on 9 MAY 2019.

[3] Cathy O'Neil. Weapons of Math Destruction. Crown Publishers, New York, NY, 2016.

[4] Gilbert Strang. Linear Algebra and Learning from Data. Wellesley-Cambridge Press, Wellesley, MA, 2019.

[19] The Mexican Migration Project data contains detailed responses to interview questions from 1982 to the present and the Climate Hazards Group InfraRed Precipitation with Station (CHIRPS) data is a high-resolution geospatial data set that crosses space (5 km horizontal grid spacing) and time.

[20] Details at http://www.grss-ieee.org/community/technical-committees/data-fusion/data-fusion-contest/.

[21] See http://datachallenge.ischool.umd.edu/.

[22] We intend to send faculty to other conferences and workshops, like the Naval Applications in Machine Learning Workshop and the ONR Applied Artificial Intelligence Summit.

[23] See https://www.navy.mil/submit/display.asp?story_id=107956 for the announcement of the competition.

[24] One project is described at https://www.usna.edu/NewsCenter/2014/05/naval-academy-operations-research-team-wins-flightline-of-the-future-competition.php.