ACM Data Science Task Force Course Example

Introduction to Data Science
Rutgers University, New Brunswick, NJ
Ananda D. Gunawardena
andyguna@gmail.com

Link to complete course material
http://andyguna.com/DSforCS
Provided under a Creative Commons – Attribution and ShareAlike License

Knowledge Area	**Total Number of Contact Hours (56 hours)**
Analysis and Presentation (AP)	*6 hours*
Data Acquisition, Management, and Governance (DAMG)	10 hours
Data Mining (DM)	8 hours
Machine Learning (ML)	12 hours
Big Data Systems (BDS)	6 hours
Programming, Data Structures, and Algorithms (PDA)	8 hours
Computing and Computer Fundamentals (CCF)	6 hours

Where does the course fit in your undergraduate Data Science curriculum?

This is a senior level computer science elective course reserved for CS majors in their final year of study. The prerequisites for the course include a course in discrete structures, a sophomore level required class for CS majors. The data science course is taken by over 300 students each year with a waitlist of 100+ students. This is the second course in data science in the curriculum. A freshman course in data literacy is preceded by this course. The course teaches data science techniques in the first half and introductory machine learning (ML) and deep learning (DL) concepts in the second half. The course is intended for CS majors who have not taken an undergraduate ML class (which is not available at our institution)

Is this course from or used in other curricula/majors?

This is a new course introduced in 2018, explicitly designed for senior computer science majors who have not taken an undergraduate course in machine learning. The course is expected to be part of a data science track that Rutgers University is currently developing for CS majors. The two courses currently in the curriculum are Data Literacy (Freshman) and Introduction to Data Science (Senior) and the data science track within the CS department to be supported by at least 2 other data science courses at Sophomore and Junior levels. Although the course is currently named “introduction to data science”, it is a very challenging course with students requiring a good understanding of machine learning, linear algebra concepts, statistics and probability concepts and their applications.

What is covered in the course?

This course covers topics needed to solve problems involving data, which includes preparation (collection and integration), characterization and presentation (information visualization), analysis (machine learning and data mining), and products (applications). An optional module on Human Factors in Data Science was added as part of a grant project with Nvidia to address bias in ML and Data Science. A more detailed collection of topics includes

- Python for Data Science
- Data Munging and Cleaning
- Data Modeling and Visualization
- Statistical Analysis, Linear Algebra
- Linear and Logistic Regression
- Bayesian Classification
- Deep Learning
- Big data platforms and technology
- Map Reduce, Tensor Flow
- Human Factors in Data Science

What is the format of the course?

This is a face-to-face class with a total of 56 contact hours (14 lab hours and 42 lecture hours). The course meets twice a week for an 80-minute lecture followed by section recitations conducted by graduate student teaching assistants (TA). The lecture provides fundamental concepts with examples and applications while the weekly labs provide hands-on activity with Jupyter labs. Recitations are designed to take students through some examples to help them complete labs. Students are expected to complete a practical lab (with real data sets) almost each week.

How are students assessed?

The major component of student assessment is 7-8 well-designed Jupyter Labs that are expected to take about a week to complete. Students also complete a mid-semester group project using data sets from social media sites like twitter or Facebook. The mid-semester project is expected to take about 2-3 weeks. There is a written midterm exam and a written final exam. In addition to that, students take 8-10 online quizzes that test their knowledge in weekly material. Students also receive a participation grade that is based on their activities on CUvids (https://cuvids.io/app/course/37/). Students watch curated videos to supplement lectures and attempt self-assessments that are part of the video.

Course tools and materials

There is no assigned textbook for the course. Students are provided with links to free online textbook chapters as needed. Students use Python to complete assignments in a Jupyter notebook framework. Students are provided large public data sets from many sources including US Government Data (data.gov) and open datasets from NY City (https://opendata.cityofnewyork.us/). Students also are given the opportunity to subscribe to curated supplemental video collection that is specifically aggregated for the course. The videos come with synchronized transcripts, table of content-based navigation, in-video quizzes, and dashboard for tracking student activities. Students pay $10/semester to access all content and platform resources (https://cuvids.io/app/course/37/) such as self-assessments, dashboards and recommendations. The course typically uses Canvas to manage online weekly quizzes that are auto graded. Students use videos and in-video quizzes to prepare for the quizzes.

Why do you teach the course this way?

This is a new course designed for CS majors in their final year of study. The course becomes a place to bring in many skills from courses they have taken throughout the CS undergraduate curriculum. The skills include programming, design and analysis of algorithms, systems, databases, linear algebra, probability and statistics and calculus. Furthermore, students are exposed to ML concepts at the undergraduate level for the first time. The course was first developed in 2018 and is frequently undergoing changes. Students consider the course to be very time consuming and challenging, yet highly beneficial due to the applied nature of its labs that deals with real world data sets and solving specific problems in a domain. The course emphasizes more of the "science" part of Data Science by exploring how to combine theoretical foundations to find better solutions to problems. The course also emphasizes the role of a data scientist in a traditional application development ecosystem where a data scientist, ML engineer and software engineers must work in collaboration to develop products to promote business goals of an organization. Students have been able to secure employment in companies like Google and Amazon, mostly because of taking this data science course.

Body of Knowledge coverage

KA	Sub-domain	Competencies Covered	Hours
AP	• Foundational considerations • Visualization	• Be able to identify situations that require the applications of data science • Learn how to do exploratory data analysis (EDA) using the appropriate visualization techniques and tools.	*6 hours*
DAG	• Data acquisition • Information extraction • Working with various types of data • Data integration • Data transformation • Data cleaning	• Be able to find the large data sets and use programming techniques to clean the data for visualization tasks and training models. • Learn various data types including nominal, ordinal, interval and ratio data and their applications. • Learn regular expressions and other text-based processing to make data more valid and apply various techniques to find outliers. • Be able to learn techniques to convert data to information	10 hours
DM	• Data preparation • Information extraction • Cluster analysis • Classification and regression • Pattern mining • Mining web data • Information retrieval	• Learn how to prepare data for ML tasks • Be able to generate clusters from unlabelled data • Be able to understand linear, multi and logistic regression and their applications in high dimensions • Be able to extract/scrape data from public webpages and prepare them for training tasks • Be able to identify patterns to reduce the scope of data collection	8 hours

ML	• General • Supervised learning • Unsupervised learning • Deep learning	• Learn general ideas of ML and applications • Be able to identify tasks for supervised and unsupervised ML • Be able to describe the functions of a deep neural network • Be able to train a ML model from given data and refine components to reduce training time and/or error rate.	12 hours
BDS	• Problems of scale • Big data computing architectures	• Survey about big data computing architectures such as Rapids GPU's, MongoDB, Hadoop and Apache Spark.	6 hours
PDA	• Programming • Data structures • Algorithms • complexity analysis • Numerical computing	• Learn to develop python libraries and reuse existing libraries build practical applications • Learn to analyse the system and memory requirements for running a big data job and find ways to make the process more efficient. • Use linear algebra packages to do large scale numerical computing using SVD and PCA	8 hours
CCF	• Storage systems fundamentals • The web and web programming	• Learn to work with large data sets that are spread across multiple data bases • Learn to build small web applications for collecting data sets for preliminary data analysis	6 hours

Additional topics

Human Factors in Data Science. An optional lecture and assignments to address bias in ML training and effect on data science-based decisions. This topic is part of a project to address inclusivity in the course in collaboration with a grant from Nvidia and Gates Foundation.

Other comments

None