DSCI 133. Introduction to Data Science and Engineering for Majors. 3 Units.
This course is an introduction to data science and analytics. In the first half of the course, students will develop a basic understanding of how to manipulate, analyze and visualize large data in a distributed computing environment, with an appreciation of open source development, security and privacy issues. Case studies and team project assignments in the second half of the course will be used to implement the ideas. Topics covered will include: Overview of large scale parallel and distributed (cloud) computing; file systems and file i/o; open source coding and distributed versioning, data query and retrieval; basic data analysis; visualization; data security, privacy and provenance. Prereq: ENGR 131 or EECS 132.

DSCI 134. Introduction to Applied Data Science. 3 Units.
This course is an introduction to data science and analytics. In the first half of the course, students will develop a basic understanding of how to manipulate, analyze and visualize large data in a distributed computing environment, with an appreciation of open source development, security and privacy issues. In the second half of the course, students will gain experience in data manipulation and analysis using scripted programming languages such as Python.

DSCI 234. Structured and Unstructured Data. 3 Units.
This course is an introduction to types of data and their representation, storage, processing and analysis. The course has three parts. In the first part of the course, students will develop a basic understanding and the ability to represent, store, process and analyze structured data. Structured data include catalogs, records, tables, logs, etc., with a fixed dimension and well-defined meaning for each data point. Suitable representation and storage mechanisms include lists and arrays. Relevant techniques include keys, hashes, stacks, queues and trees. In the second part of the course, students will develop a basic understanding and the ability to represent, store, process and analyze semi-structured data. Semi-structured data include texts, web pages and networks, without a dimension and structure, but with well-defined meaning for each data point. Suitable representation and storage mechanisms include trees, graphs and RDF triples. Relevant techniques include XML, YAML, JSON, parsing, annotation, language processing. In the third part of the course, students will develop a basic understanding and the ability to represent, store, process and analyze unstructured data. Unstructured data include images, video, and time series data, without neither a fixed dimension and structure, nor well-defined meaning for individual data points. Suitable representation and storage mechanisms include large matrices, EDF, DICOM. Relevant techniques include feature extraction, segmentation, clustering, rendering, indexing, and visualization. Prereq: DSCI 133.

DSCI 341. Introduction to Databases: DS Major. 3 Units.
Database management become a central component of a modern computing environment, and, as a result, knowledge about database systems has become an essential part of education in computer science and data science. This course is an introduction to the nature and purpose of database systems, fundamental concepts for designing, implementing and querying a database and database architectures. Weeks 1-6 provide an overview of basic database systems concepts including database design, database systems architecture, and database querying, using relational model and SQL as query language. Weeks 7-10 Objects, Semi structured data, XML and RDF basics. Weeks 11-14 provide an overview of more advanced topics including Database System Architectures (Parallel Databases and Distributed Databases), and Data Warehousing and Information Retrieval. Prereq: DSCI 234 or EECS 233.

DSCI 342. Introduction to Data Science Systems. 3 Units.
An introduction to the software and hardware architecture of data science systems, with an emphasis on Operating Systems and Computer Architecture that are relevant to Data Sciences systems. At the end of the course, the student should understand the principles and architecture of storage systems, file systems (especially, HDFS), memory hierarchy, and GPU. The student should have carried out projects in these areas, and should be able to critically compare various design decisions in terms of capability and performance. Prereq: DSCI 234.

DSCI 343. Introduction to Data Analysis. 3 Units.
In this class we will give a broad overview of data analysis techniques, covering techniques from data mining, machine learning and signal processing. Students will also learn about probabilistic representations, how to conduct an empirical study and support empirical hypotheses through statistical tests, and visualize the results. Course objectives: -Exposure to different analysis approaches. -Understand probabilistic representations and inference mechanisms. -Understand how to create empirical hypotheses and how to test them. Prereq: EECS 340 and DSCI 234.

DSCI 344. Scalable Parallel Data Analysis. 3 Units.
This course provides an introduction to scalable and parallel data analysis using the most common frameworks and programming tools in the age of big data. Covered topics include parallel programming models, parallel hardware architectures, multi-threaded, multi-core programming, cluster computing and GPU programming. The course is designed to provide a heavily hands-on experience with several programming assignments. Prereq: DSCI 342.

DSCI 345. Files, Indexes and Access Structures for Big Data. 3 Units.
Database management become a central component of a modern computing environment, and, as a result, knowledge about database systems has become an essential part of education in computer science and data science. This course is an introduction to the nature and purpose of database systems, fundamental concepts for designing, implementing and querying a database and database architectures. Objectives: -An expert knowledge of basic data structures, basic searching, sorting, methods, algorithm techniques, (such as greedy and divide and conquer) -In-depth knowledge on Search and Index Structures for large, heterogeneous data including multidimensional data, high dimensional data and data in metric spaces (e.g., sequences, images), on different search methods (e.g. similarity searching, partial match, exact match), and on dimensionality reduction techniques. Prereq: DSCI 234 or EECS 233.
DSCI 351. Exploratory Data Science. 3 Units.
In this course, we will learn data science and analysis approaches to identify statistically significant relationships and better model and predict the behavior of these systems. We will assemble and explore real-world datasets, perform clustering and pair plot analyses to investigate correlations, and logistic regression will be employed to develop associated predictive models. Results will be interpreted, visualized and discussed. We will introduce basic elements of statistical analysis using R Project open source software for exploratory data analysis and model development. R is an open-source software project with broad abilities to access machine-readable open-data resources, data cleaning and munging functions, and a rich selection of statistical packages, used for data analytics, model development and prediction. This will include an introduction to R data types, reading and writing data, looping, plotting and regular expressions, so that one can start performing variable transformations for linear fitting and developing structural equation models, while exploring for statistically significant relationships. Offered as DSCI 351, DSCI 351M and DSCI 451. Prereq: (ENGR 131 or EECS 132 or DSCI 134) and (STAT 312R or STAT 201R or SYBB 310 or PQHS/EPBI 431).

DSCI 351M. Exploratory Data Science. 3 Units.
In this course, we will learn data science and analysis approaches to identify statistically significant relationships and better model and predict the behavior of these systems. We will assemble and explore real-world datasets, perform clustering and pair plot analyses to investigate correlations, and logistic regression will be employed to develop associated predictive models. Results will be interpreted, visualized and discussed. We will introduce basic elements of statistical analysis using R Project open source software for exploratory data analysis and model development. R is an open-source software project with broad abilities to access machine-readable open-data resources, data cleaning and munging functions, and a rich selection of statistical packages, used for data analytics, model development and prediction. This will include an introduction to R data types, reading and writing data, looping, plotting and regular expressions, so that one can start performing variable transformations for linear fitting and developing structural equation models, while exploring for statistically significant relationships. Offered as DSCI 351, DSCI 351M and DSCI 451. Prereq: (ENGR 131 or EECS 132 or DSCI 134) and (STAT 312R or STAT 201R or SYBB 310 or PQHS/EPBI 431).

DSCI 352. Applied Data Science Research. 3 Units.
This is a project based data science research class, in which project teams identify a research project under the guidance of a domain expert professor. The research is structured as a data analysis project including the 6 steps of developing a reproducible data science project, including: 1: Define the ADS question, 2: Identify, locate, and/or generate the data 3: Exploratory data analysis 4: Statistical modeling and prediction 5: Synthesizing the results in the domain context 6: Creation of reproducible research, Including code, datasets, documentation and reports. During the course special topic lectures will include Ethics, Privacy, Openness, Security, Ethics. Value. Offered as DSCI 352 and DSCI 452. Prereq: (DSCI 133 or DSCI 134 or ENGR 131 or EECS 132) and (STAT 312R or STAT 201R or SYBB 310 or PQHS/EPBI 431 or OPRE 207) and (DSCI 351 or (SYBB 311A and SYBB 311B and SYBB 311C and SYBB 311D) or SYBB 321 or MKMR 201).

DSCI 353. Data Science: Statistical Learning, Modeling and Prediction. 3 Units.
In this course, we will use an open data science tool chain to develop reproducible data analyses useful for inference, modeling and prediction of the behavior of complex systems. In addition to the standard data cleaning, assembly and exploratory data analysis steps essential to all data analyses, we will identify statistically significant relationships from datasets derived from population samples, and infer the reliability of these findings. We will use regression methods to model a number of both real-world and lab-based systems producing predictive models applicable in comparable populations. We will assemble and explore real-world datasets, use pair-wise plots to explore correlations, perform clustering, self-similarity, and logistic regression develop both fixed-effect and mixed-effect predictive models. We will introduce machine-learning approaches for classification and tree-based methods. Results will be interpreted, visualized and discussed. We will introduce the basic elements of data science and analytics using R Project open source software. R is an open-source software project with broad abilities to access machine-readable open-data resources, data cleaning and assembly functions, and a rich selection of statistical packages, used for data analytics, model development, prediction, inference and clustering. For students with little prior R experience, we'll introduce resources to learn R data types, reading and writing data, looping, plotting and regular expressions. With this background, it becomes possible to start performing variable transformations for linear regression fitting and developing structural equation models, fixed-effects and mixed-effects models along with other statistical learning techniques, while exploring for statistically significant relationships. The class will be structured to have a balance of theory and practice. We'll split class into Foundation and Practicum a) Foundation: lectures, presentations, discussion b) Practicum: coding, demonstrations and hands-on data science work. Offered as DSCI 353, DSCI 353M and DSCI 453.
DSCI 353M. Data Science: Statistical Learning, Modeling and Prediction. 3 Units.
In this course, we will use an open data science tool chain to develop reproducible data analyses useful for inference, modeling and prediction of the behavior of complex systems. In addition to the standard data cleaning, assembly and exploratory data analysis steps essential to all data analyses, we will identify statistically significant relationships from datasets derived from population samples, and infer the reliability of these findings. We will use regression methods to model a number of both real-world and lab-based systems producing predictive models applicable in comparable populations. We will assemble and explore real-world datasets, use pair-wise plots to explore correlations, perform clustering, self-similarity, and logistic regression develop both fixed-effect and mixed-effect predictive models. We will introduce machine-learning approaches for classification and tree-based methods. Results will be interpreted, visualized and discussed. We will introduce the basic elements of data science and analytics using R Project open source software. R is an open-source software project with broad abilities to access machine-readable open-data resources, data cleaning and assembly functions, and a rich selection of statistical packages, used for data analytics, model development, prediction, inference and clustering. For students with little prior R experience, we’ll introduce resources to learn R data types, reading and writing data, looping, plotting and regular expressions. With this background, it becomes possible to start performing variable transformations for linear regression fitting and developing structural equation models, fixed-effects and mixed-effects models along with other statistical learning techniques, while exploring for statistically significant relationships. The course will be structured to have a balance of theory and practice. We’ll split class into Foundation and Practicum: a) Foundation: lectures, presentations, discussion b) Practicum: coding, demonstrations and hands-on data science work. Offered as DSCI 353, DSCI 353M and DSCI 453.

DSCI 390. Machine Learning for Big Data. 3 Units.
Machine learning is a sub-field of Artificial Intelligence that is concerned with the design and analysis of algorithms that "learn" and improve with experience. While the broad aim behind research in this area is to build systems that can simulate or even improve on certain aspects of human intelligence, algorithms developed in this area have become very useful in analyzing and predicting the behavior of complex systems. Machine learning algorithms have been used to guide diagnostic systems in medicine, recommend interesting products to customers in e-commerce, play games at human championship levels, and solve many other very complex problems. This course is an introduction to algorithms for machine learning and their implementation in the context of big data. We will study different learning settings, the different algorithms that have been developed for these settings, and learn about how to implement these algorithms and evaluate their behavior in practice. We will also discuss dealing with noise, missing values, scalability properties and talk about tools and libraries available for these methods. At the end of the course, you should be able to: --Understand when to use machine learning algorithms; --Understand, represent and formulate the learning problem; --Apply the appropriate algorithm(s) or tools, with an understanding of the tradeoffs involved including scalability and robustness; --Correctly evaluate the behavior of the algorithm when solving the problem. Prereq: DSCI 234 and DSCI 343.

DSCI 391. Data Mining for Big Data. 3 Units.
With the unprecedented rate at which data is being collected today in almost all fields of human endeavor, there is an emerging economic and scientific need to extract useful information from it. Data mining is the process of automatic discovery of patterns, changes, associations and anomalies in massive databases, and is a highly interdisciplinary field representing the confluence of several disciplines, including database systems, data warehousing, machine learning, statistics, algorithms, data visualization, and high-performance computing. This course is an introduction to the commonly used data mining techniques. In the first part of the course, students will develop a basic understanding of the basic concepts in data mining such as frequent pattern mining, association rule mining, basic techniques for data preprocessing such as normalization, regression, and classic matrix decomposition methods such as SVD, LU, and QR decompositions. In the second part of the course, students will develop a basic understanding of classification and clustering and be able to apply classic methods such as k-means, hierarchical clustering methods, nearest neighbor methods, association based classifiers. In the third part of the course, students will have a chance to study more advanced data mining applications such as feature selection in high-dimensional data, dimension reduction, and mining biological datasets. Prereq: DSCI 234 and DSCI 343.

DSCI 451. Exploratory Data Science. 3 Units.
In this course, we will learn data science and analysis approaches to identify statistically significant relationships and better model and predict the behavior of these systems. We will assemble and explore real-world datasets, perform clustering and pair plot analyses to investigate correlations, and logistic regression will be employed to develop associated predictive models. Results will be interpreted, visualized and discussed. We will introduce basic elements of statistical analysis using R Project open source software for exploratory data analysis and model development. R is an open-source software project with broad abilities to access machine-readable open-data resources, data cleaning and munging functions, and a rich selection of statistical packages, used for data analytics, model development and prediction. This will include an introduction to R data types, reading and writing data, looping, plotting and regular expressions, so that one can start performing variable transformations for linear fitting and developing structural equation models, while exploring for statistically significant relationships. Offered as DSCI 351, DSCI 351M and DSCI 451.

DSCI 452. Applied Data Science Research. 3 Units.
This is a project based data science research class, in which project teams identify a research project under the guidance of a domain expert professor. The research is structured as a data analysis project including the 6 steps of developing a reproducible data science project, including 1: Define the ADS question, 2: Identify, locate, and/or generate the data 3: Exploratory data analysis 4: Statistical modeling and prediction 5: Synthesizing the results in the domain context 6: Creation of reproducible research, including code, datasets, documentation and reports. During the course special topic lectures will include Ethics, Privacy, Openness, Security, Ethics. Value. Offered as DSCI 352 and DSCI 452.
DSCI 453. Data Science: Statistical Learning, Modeling and Prediction. 3 Units.
In this course, we will use an open data science tool chain to develop reproducible data analyses useful for inference, modeling and prediction of the behavior of complex systems. In addition to the standard data cleaning, assembly and exploratory data analysis steps essential to all data analyses, we will identify statistically significant relationships from datasets derived from population samples, and infer the reliability of these findings. We will use regression methods to model a number of both real-world and lab-based systems producing predictive models applicable in comparable populations. We will assemble and explore real-world datasets, use pair-wise plots to explore correlations, perform clustering, self-similarity, and logistic regression develop both fixed-effect and mixed-effect predictive models. We will introduce machine-learning approaches for classification and tree-based methods. Results will be interpreted, visualized and discussed. We will introduce the basic elements of data science and analytics using R Project open source software. R is an open-source software project with broad abilities to access machine-readable open-data resources, data cleaning and assembly functions, and a rich selection of statistical packages, used for data analytics, model development, prediction, inference and clustering. For students with little prior R experience, we’ll introduce resources to learn R data types, reading and writing data, looping, plotting and regular expressions. With this background, it becomes possible to start performing variable transformations for linear regression fitting and developing structural equation models, fixed-effects and mixed-effects models along with other statistical learning techniques, while exploring for statistically significant relationships. The class will be structured to have a balance of theory and practice. We’ll split class into Foundation and Practicum a) Foundation: lectures, presentations, discussion b) Practicum: coding, demonstrations and hands-on data science work. Offered as DSCI 353, DSCI 353M and DSCI 453.