DATA SCIENCE AND ANALYTICS / APPLIED DATA SCIENCE

Bachelor of Science in Data Science and Analytics

Administered by the Department of Electrical Engineering and Computer Science (http://bulletin.case.edu/schoolofengineering/elecengcompsci/#undergraduatetext)
Glennan Building (7071)
http://engineering.case.edu/eecs/
Phone: 216.368.3800, Fax: 216.368.3209
Alexis R. Abramson, PhD, Interim Chair

The Bachelor of Science degree program in Data Science and Analytics provides our students with a broad foundation in the field and the instruction, skills, and experience needed to understand and handle large amounts of data that transform thinking about a collection of vast amounts of data into one that focuses on the data's conversion to actionable information. The degree program has a unique focus on real-world data and real-world applications.

This major is one of the first undergraduate programs nationwide with a unique curriculum that includes mathematical modeling, informatics, data analytics, visual analytics and project-based applications - all elements of the future emerging field of data science.

Minor in Applied Data Science (ADS)

Administered by the Department of Materials Science and Engineering (http://bulletin.case.edu/schoolofengineering/materialssciencesengineering/#undergraduatetext)
312 White Building (7204)
http://engineering.case.edu/emse/
Phone: 216.368.4230, Fax: 216.368.3209
Roger French, EMSE / CSE Faculty Director (ADS)

The Minor in Applied Data Science (https://case.edu/datascience) is based in the Case School of Engineering and includes faculty from schools across the university. The minor is directed to students studying in the domains of Engineering and Physical Sciences (including Energy and Manufacturing, Astronomy, Geology, Physics), Health (including Translational and Clinical), and Business (including Finance, Marketing, and Economics). Successful completion of the minor requirements leads to a "Minor in Applied Data Science" for the graduating student. The minor represents that the students have developed knowledge of the essential elements of Data Science and Analytics in the area of their major (their domain of expertise).

Bachelor of Science in Data Science and Analytics

In addition to engineering general education requirements (http://bulletin.case.edu/undergraduatetudies/csedegree) and university general education requirements (http://bulletin.case.edu/undergraduatestudies/degreeprograms), the major requires the following courses:

**Major Requirements**

<table>
<thead>
<tr>
<th>Course</th>
<th>Title</th>
<th>Credits</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHEM 111</td>
<td>Principles of Chemistry for Engineers</td>
<td>4</td>
</tr>
<tr>
<td>DSCI 133</td>
<td>Introduction to Data Science and Engineering for Majors</td>
<td>3</td>
</tr>
<tr>
<td>DSCI 234</td>
<td>Structured and Unstructured Data</td>
<td>3</td>
</tr>
<tr>
<td>DSCI 341</td>
<td>Introduction to Databases: DS Major</td>
<td>3</td>
</tr>
<tr>
<td>DSCI 342</td>
<td>Introduction to Data Science Systems</td>
<td>3</td>
</tr>
<tr>
<td>DSCI 343</td>
<td>Introduction to Data Analysis</td>
<td>3</td>
</tr>
<tr>
<td>DSCI 344</td>
<td>Scalable Parallel Data Analysis</td>
<td>3</td>
</tr>
<tr>
<td>DSCI 345</td>
<td>Files, Indexes and Access Structures for Big Data</td>
<td>3</td>
</tr>
<tr>
<td>EECS 132</td>
<td>Introduction to Programming in Java</td>
<td>3</td>
</tr>
<tr>
<td>EECS 302</td>
<td>Discrete Mathematics</td>
<td>3</td>
</tr>
<tr>
<td>EECS 340</td>
<td>Algorithms</td>
<td>3</td>
</tr>
<tr>
<td>EECS 393</td>
<td>Software Engineering</td>
<td>3</td>
</tr>
<tr>
<td>ENGL 398</td>
<td>Professional Communication for Engineers</td>
<td>2</td>
</tr>
<tr>
<td>ENGR 398</td>
<td>Professional Communication for Engineers</td>
<td>1</td>
</tr>
<tr>
<td>MATH 201</td>
<td>Introduction to Linear Algebra for Applications</td>
<td>3</td>
</tr>
<tr>
<td>MATH 121</td>
<td>Calculus for Science and Engineering I</td>
<td>4</td>
</tr>
<tr>
<td>MATH 122</td>
<td>Calculus for Science and Engineering II</td>
<td>4</td>
</tr>
<tr>
<td>MATH 223</td>
<td>Calculus for Science and Engineering III</td>
<td>3</td>
</tr>
<tr>
<td>MATH 224</td>
<td>Elementary Differential Equations</td>
<td>3</td>
</tr>
<tr>
<td>PHYS 121</td>
<td>General Physics I - Mechanics</td>
<td>4</td>
</tr>
<tr>
<td>PHYS 122</td>
<td>General Physics II - Electricity and Magnetism</td>
<td>4</td>
</tr>
</tbody>
</table>

Core courses provide our students with a strong background in signal processing, systems, and analytics. Students are required to develop depth in at least one of the following technical areas: signal processing, systems, and analytics. Each data science and analytics student must complete the following requirements:

**Technical Elective Requirement**

Each student must complete 8 courses (24 credit-hours) of approved technical electives. Technical electives shall be chosen to fulfill the probability/statistics elective (1 course), the computer and data security elective (1 course), the depth requirement (3 courses), and 3 courses otherwise chosen to increase the student’s understanding of data science and analytics. Technical electives not used to satisfy the probability/statistics elective, the computer and data security elective, or the depth requirement are more generally defined as any course related to the principles and practice of data science and analytics. This includes all DSCI courses at the 200 level and above and can include courses from other programs. All non-DSCI technical electives must be approved by the student’s academic advisor.
Depth Requirement
Each student must show a depth of competence in one technical area by taking at least three courses from one of the following three areas. Additional courses, beyond those that are listed, may be approved by the student’s academic advisor.

**Area I: Signal Processing**
- EECS 246 Signals and Systems 4
- EECS 313 Signal Processing 3
- STAT 332 Statistics for Signal Processing 3

**Area II: Systems**
- EECS 325 Computer Networks I 3
  or EECS 425 Computer Networks I
- EECS 338 Intro to Operating Systems and Concurrent Programming 4
- EECS 600 Special Topics (Cloud Computing) 1-18

**Area III: Analytics**
- DSCI 390 Machine Learning for Big Data 3
- DSCI 391 Data Mining for Big Data 3
- EECS 339 Web Data Mining 3
- EECS 346 Engineering Optimization 3
- EECS 440 Machine Learning 3
- EECS 442 Causal Learning from Data 3

**Computer and Data Security Elective Requirement**
- EECS 444 Computer Security 3
- MATH 408 Introduction to Cryptology 3

**Statistics Requirement**
- MATH 380 Introduction to Probability 3
- STAT 325 Data Analysis and Linear Models 3

**Design Requirement**
- DSCI 398 Engineering Projects I
- DSCI 399 Engineering Projects II

**Suggested Program of Study: Bachelor of Science in Data Science and Analytics**
The following is a suggested program of study. Current students should always consult their advisers and their individual graduation requirement plans as tracked in SIS (http://sis.case.edu).

**First Year**
- SAGES First Year Seminar* 4
- Principles of Chemistry for Engineers (CHEM 111) 4
- Calculus for Science and Engineering I (MATH 121) 4
- Introduction to Programming in Java (EECS 132) 3
- PHED (2 half semester courses)* 0
- SAGES University Seminar* 3
- General Physics I - Mechanics (PHYS 121) 4
- Calculus for Science and Engineering II (MATH 122) 4
- Introduction to Data Science and Engineering for Majors (DSCI 133) 3
- PHED (2 half semester courses)* 0
- Open Elective 3

**Second Year**
- SAGES University Seminar* 3
- General Physics II - Electricity and Magnetism (PHYS 122) 4
- Calculus for Science and Engineering III (MATH 223) 3
- Structured and Unstructured Data (DSCI 234) 3
- Discrete Mathematics (EECS 302) 3
- Introduction to Databases: DS Major (DSCI 341) 3
- Elementary Differential Equations (MATH 224) 3
- Algorithms (EECS 340) 3
- Breadth elective** 3
- Probability/Statistics Elective* 3

**Third Year**
- Introduction to Data Science Systems (DSCI 342) 3
- Software Engineering (EECS 393) 3
- Breadth elective** 3
- Introduction to Data Analysis (DSCI 343) 3
- Introduction to Linear Algebra for Applications (MATH 201) 3
- Professional Communication for Engineers (ENGL 398) 2
- Professional Communication for Engineers (ENGR 398) 1
- Scalable Parallel Data Analysis (DSCI 344) 3
- Computer and Data Security Elective* 3
- Files, Indexes and Access Structures for Big Data (DSCI 345) 3
- Technical Elective** 3

**Fourth Year**
- Technical Elective** 3
- Technical Elective** 3
- DSCI 398 Senior Project I 4
- Technical Elective** 3
- Breadth elective** 3
- Breadth elective** 3
- DSCI Technical elective** 3
- DSCI 399 Senior Project II 4
- Technical Elective** 3
- Open elective 3

Total Units in Sequence: 125
The ADS minor curriculum
The curriculum is based on five 3-credit courses, with one class chosen from each of Levels 1 through Level 5, which cover the spectrum of learning needed to achieve domain area expertise in data science and analytics. The courses are chosen to be both cross-cutting, i.e., intermixing students from across the university in the fundamental concepts such as scripting and statistics (Levels 1, 2, and 4), and domain-focused (Levels 3 and 5). For the Level 4 undergraduate research course, the research topic will be approved by the minor advisor, and will also be a 3-credit project. This will provide minor students both the domain focused learning they need, and a broadening perspective on applications, methods, and uses of ADS in other domains.

Courses Counted Toward Minor Requirements
Established courses included in the Minor are found in Case School of Engineering (Materials Science, Electrical Engineering and Computer Science, Manufacturing), College of Art & Science (Mathematics, Astronomy, Philosophy); School of Medicine, School of Nursing, and Weatherhead School of Management (Marketing, Finance, Operations, and Economics).

The courses that meet the requirements for the Minor can also be taken by students to meet requirements in Major programs, and therefore serve a dual purpose in our academic offerings. However, each program, department, and school may have its own criteria on whether a given course could be "double counted" towards major and minor requirements.

Level 5:
<table>
<thead>
<tr>
<th>Course</th>
<th>Title</th>
<th>Credits</th>
</tr>
</thead>
<tbody>
<tr>
<td>BAFI 361</td>
<td>Empirical Analysis in Finance</td>
<td>3</td>
</tr>
<tr>
<td>ECON 327</td>
<td>Advanced Econometrics</td>
<td>3</td>
</tr>
<tr>
<td>MKMR 308</td>
<td>Measuring Marketing Performance</td>
<td>3</td>
</tr>
<tr>
<td>MKMR 310</td>
<td>Marketing Analytics</td>
<td>3</td>
</tr>
<tr>
<td>SYBB 459</td>
<td>Bioinformatics for Systems Biology</td>
<td>3</td>
</tr>
</tbody>
</table>

Level 4:
<table>
<thead>
<tr>
<th>Course</th>
<th>Title</th>
<th>Credits</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASTR 369</td>
<td>Undergraduate Research</td>
<td>1-3</td>
</tr>
<tr>
<td>DSCI 352</td>
<td>Applied Data Science Research</td>
<td>3</td>
</tr>
<tr>
<td>EMSE 325</td>
<td>Undergraduate Research in Materials Science and Engineering</td>
<td>1-3</td>
</tr>
<tr>
<td>SYBB 387</td>
<td>Undergraduate Research in Systems Biology</td>
<td>1-3</td>
</tr>
</tbody>
</table>

Level 3:
<table>
<thead>
<tr>
<th>Course</th>
<th>Title</th>
<th>Credits</th>
</tr>
</thead>
<tbody>
<tr>
<td>DSCI 351</td>
<td>Exploratory Data Science</td>
<td>3</td>
</tr>
<tr>
<td>MKMR 201</td>
<td>Marketing Management</td>
<td>3</td>
</tr>
<tr>
<td>SYBB 311A</td>
<td>Survey of Bioinformatics: Technologies in Bioinformatics</td>
<td>1</td>
</tr>
<tr>
<td>SYBB 311B</td>
<td>Survey of Bioinformatics: Data Integration in Bioinformatics</td>
<td>1</td>
</tr>
<tr>
<td>SYBB 311C</td>
<td>Survey of Bioinformatics: Translational Bioinformatics</td>
<td>1</td>
</tr>
<tr>
<td>SYBB 421</td>
<td>Fundamentals of Clinical Information Systems</td>
<td>3</td>
</tr>
<tr>
<td>SYBB 412</td>
<td>Survey of Bioinformatics: Programming for Bioinformatics</td>
<td>3</td>
</tr>
</tbody>
</table>

Level 2:
<table>
<thead>
<tr>
<th>Course</th>
<th>Title</th>
<th>Credits</th>
</tr>
</thead>
<tbody>
<tr>
<td>PQHS 431</td>
<td>Statistical Methods I</td>
<td>3</td>
</tr>
<tr>
<td>OPRE 207</td>
<td>Statistics for Business and Management Science I</td>
<td>3</td>
</tr>
<tr>
<td>STAT 201R</td>
<td>Basic Statistics for Social and Life Sciences Using R Programming</td>
<td>3</td>
</tr>
</tbody>
</table>
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 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 330. Cognition and Computation. 3 Units.
An introduction to (1) theories of the relationship between cognition and computation; (2) computational models of human cognition (e.g. models of decision-making or concept creation); and (3) computational tools for the study of human cognition. All three dimensions involve data science: theories are tested against archives of brain imaging data; models are derived from and tested against datasets of e.g., financial decisions (markets), legal rulings and findings (juries, judges, courts), legislative actions, and healthcare decisions; computational tools aggregate data and operate upon it analytically, for search, recognition, tagging, machine learning, statistical description, and hypothesis testing. Offered as COGS 330, COGS 430, DSCI 330 and DSCI 430.

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: -Expose students 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 has 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. The M section of DSCI 351 is for students focusing on Materials Data Science. 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 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 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. The M section of DSCI 351 is for students focusing on Materials Data Science. 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 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 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. The M section of DSCI 352 is for students focusing on Materials Data Science. Offered as DSCI 352, DSCI 352M 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 352M. 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. The M section of DSCI 352 is for students focusing on Materials Data Science. Offered as DSCI 352, DSCI 352M 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 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 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 430. Cognition and Computation. 3 Units.
An introduction to (1) theories of the relationship between cognition and computation; (2) computational models of human cognition (e.g. models of decision-making or concept creation); and (3) computational tools for the study of human cognition. All three dimensions involve data science: theories are tested against archives of brain imaging data; models are derived from and tested against datasets of e.g., financial decisions (markets), legal rulings and findings (juries, judges, courts), legislative actions, and healthcare decisions; computational tools aggregate data and operate upon it analytically, for search, recognition, tagging, machine learning, statistical description, and hypothesis testing. Offered as COGS 330, COGS 430, DSCI 330 and DSCI 430.

DSCI 451. Exploratory Data Science. 3 Units.
In this course, we will learn data science and analysis approaches to identify statistically significance 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 assembly 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, loops, 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. The M section of DSCI 351 is for students focusing on Materials Data Science. 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. The M section of DSCI 352 is for students focusing on Materials Data Science. Offered as DSCI 352, DSCI 352M 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.