Program Guide, Master of Science in Data Science

The University of Michigan Master of Science in Data Science is a professional degree equipped with strong methodological training. The degree program will require at least 25 units of coursework. The specific requirements are expressed in terms of foundational training in Computing and Statistics, and 4 core areas in which proficiency is required. We expect most students admitted to the program to already meet the proficiency requirements in some core areas even if they have limited exposure to others, and possibly even without some of the required foundational training. The specific core areas in which an incoming student has proficiency will vary since we intend to admit students from a variety of backgrounds and then bring them to a required standard of proficiency by the time they graduate. Correspondingly, the number of units of coursework that students take will vary with their backgrounds.

The field of Data Science provides people with the right skills to start excellent career paths, ranging from business analysts across a variety of industries, to novel IT careers, to working on scientific research teams. For such careers, students require solid training in an array of computational and statistical analysis techniques shared across industries and disciplines. We characterize the required skills in two categories: statistical techniques and foundation, such as those taught by the Statistics and Biostatistics departments, and computational tools and principles, such as those taught by the Computer Science and Engineering Division and the School of Information. The design of the program is to require every student to receive balanced training in both "buckets." To create an academic plan that achieves this balance, and to foster a greater sense of shared community, we do not intend to offer any sub-plans or tracks within the proposed degree program. Rather, we expect graduates of this program to understand data representation and data analysis at an advanced level. With the MS in Data Science all students will be able to: identify relevant datasets and inferential questions of interest, apply and develop the appropriate statistical and computational tools to the data to answer questions posed by individuals, organizations or governmental agencies, design and evaluate analytical procedures appropriate to the data, implement these efficiently over large heterogeneous data sets in a multi-computer environment, and communicate their findings to end-users.

We define the program requirements in terms of expected skills at graduation and work backwards from there (rather than solely in terms of courses/activities required). In addition to graduate competence in at least some aspects of Data Science, we also require graduates from this program to develop skills in all Data Science core areas at least comparable to what advanced disciplinary bachelor's degree holders in each of sponsoring units would have. Students entering the program with an undergraduate degree in Data Science may already have the necessary skills in most core areas, and would then focus on developing deeper graduate level competence, in the fashion one would normally expect of most disciplinary MS programs. However, there are very few undergraduate Data Science programs today, and most admitted students may need to gain competence in several of these core areas during their MS program. This additional coursework will add time to graduation (see sample schedules below).

Curriculum and Requirements

The Master of Science Degree has the following requirements:

a. At least 25 units of graduate-level coursework.

b. At least 18 units of coursework at the advanced graduate level (500 level or above in LSA, UMSI, and CoE, and 600 level or above in SPH).

c. Demonstrated competence in a basic Computing Sequence, comprising the following three steps (i-iii), and in a basic Statistics Sequence (steps iv and v). (Note that steps i and ii in the Computing Sequence can be taken in any order, but step iii requires both previous steps to have been completed).

i. Discrete Mathematics. Satisfied through previous completion of a course equivalent to EECS 203, or by taking Math 465.
ii. Programming in a Full Programming Language, such as C++ or Java. Satisfied through previous completion of a course equivalent to EECS 280, or by taking EECS 402.
iii. Data Structures and Algorithms. Satisfied through previous completion of a course equivalent to EECS 281, or by taking EECS 403[1].
iv. Probability. Satisfied through completion of a course equivalent to MATH/STATS 425. STATS 510 or BIOSTAT 601 can be used to satisfy this requirement.
v. Statistical Inference. Satisfied through completion of a course equivalent to STATS 426. STATS 511 or BIOSTAT 602 can be used to satisfy this requirement.

Competence in the basic Computing Sequence and the basic Statistics sequence may be demonstrated by taking courses, such as any one of the classes listed for that area and obtaining a grade of at least B- or by having taken an equivalent class in a prior educational program, with equivalency determined through standard departmental procedures, or through passing tests explicitly devised for proving such competence.

Any courses taken to satisfy this requirement cannot be counted towards requirement (a) or (b) above.

d. Demonstrated expertise in each of four core areas, representing the breadth of Data Science. Knowledge in each core area may be demonstrated by taking courses, such as any one of the classes listed for that area and obtaining a grade of at least B- or by having taken an equivalent class in a prior educational program, with equivalency determined through standard departmental procedures. At least two out of four core courses must be advanced graduate level (500 level or above in LSA, UMSI, and CoE, and 600 level or above in SPH). The four core areas are divided into 2 groups of two areas each as given below.

[1] New course offering currently being planned.

Expertise in Data Management and Manipulation.

1. Databases: EECS 484 or EECS 584, AND
2. Data and Web Application: EECS 485 or EECS 486 or EECS 549/ SI 650 or SI 618.

Expertise in Data Science Techniques.

3. Regression: STATS 413 or STATS 513 or STATS 500 or BIOSTAT 650, AND
4. Data Mining/Statistical Learning/Machine Learning: STATS 415 or EECS 445 or SI 671 or BIOSTAT 616[2] or STATS 503 or EECS 545.

See Appendix A for a brief description of each of the courses listed above. All 400 level courses named above are accepted for graduate credit by Rackham.

e. Advanced Electives. At least one course from each of three buckets representing graduate-level disciplinary material. Moreover, among these at least two advanced graduate courses must be taken. These three buckets are: A) Principles of Data Science, B) Data Analysis, and C) Data Science Computation.

The courses currently in each bucket are listed in Appendix B. It is expected that these lists will change frequently, at the discretion of the program committee.

f. Integrative Capstone Experience. This is expected to be a capstone project, which must be pre-approved by the Student Affairs Committee. Capstone projects will earn credit through directed study courses. While the grade for such courses will be determined using the standard practices of the department offering the specific course, satisfaction of capstone requirements will be determined by the Student Affairs Committee based on a final report. The current list of approved courses for this purpose are listed as bucket D of advanced electives, in Appendix B.

g. Colloquium. Regular attendance of a weekly seminar on Data Science, to be offered as a 1 credit pass/fail course[3]. Students are required to pass this class in any one semester (attendance is taken). Students are encouraged to meet their colloquium requirement in the first semester of the program. We expect that many students will choose to attend the weekly Data Science seminar in additional semesters even if this is not explicitly required. This colloquial training will expose students to current DS developments beyond the boundaries of their coursework.

[2] New course under development.

[3] This seminar course, cross-listed across all four sponsoring units, is expected to have all approvals in place and launch in time for the arrival of the first cohort of students in this program. Initially, this course will piggyback on the existing MIDAS Colloquium Series, which brings internationally recognized data scientists to U-M.

Student Trajectories

Students with an undergraduate degree in Data Science would already have obtained a reasonable level of training towards the core skills: they would need only a limited number of additional courses to meet the core requirements, and hence can focus on more advanced graduate level coursework. In contrast, students with undergraduate degrees in other disciplines may be missing some aspects in one or both primary buckets (computational and statistical), and would need to take additional courses to cover the areas they are missing. Our program is designed such that a student with an undergraduate degree in Data Science from U-M can complete all requirements for an MS degree in one additional year. U-M students admitted to the Data Science MS may transfer to the master's up to half the required credits as long as: 1) credits are for graduate-level courses and meet the requirements of DS, and 2) have not been used to meet the requirements of the undergrad degree in any way. Students from outside U-M, following the above stipulations, may transfer up to 6 credits.

Students without sufficient training in some areas may need more than one year to complete requirements: possibly additional one or two semesters. Our expectation is that many students with undergraduate degrees in quantitative disciplines, such as Mathematics or Physics, will be able to complete all requirements in two years. A few sample trajectories are given below; please note that these are ***samples only***, and are not intended to be viewed as the only or even strongly suggested alternatives.

Student with BS in Data Science

A student entering the program with a Bachelor's degree in Data Science may already meet proficiency requirements in several of the core areas. For example, graduates of the U-M undergraduate program in Data Science may already meet proficiency requirements have expertise in several core areas, depending on exactly which electives they took. The MS program provides an opportunity for such a student to deepen their knowledge through advanced coursework requirement, while also attaining proficiency in any core areas not already covered. For example, consider a student missing only a background in statistical inference, but without advanced graduate coursework in any core area. Such a student may follow this trajectory, which requires 26 units in two terms:

FALL SEMESTER:

- EECS 549: Information Retrieval (3 units) [Advanced Graduate Class in a Core Area]
- SI 649: Information Visualization (3 units) [DS Computation Elective]
- EECS 545 Machine Learning (4 units) [Advanced Graduate Class in a Core Area]
- BIOSTAT 682 or STATS 551: Bayesian Analysis (3 units) [Principles of DS Elective]
- Data Science Seminar (1 unit)

WINTER SEMESTER:

- STATS 511: Statistical Inference (3 units) [Step v of Statistics Sequence]

- STATS 531: Time Series Analysis (3 units) [DS Analysis Elective]
- SI 630: Natural Language Processing: Algorithms and People (3 units) [DS Computation Elective]
- BIOSTAT 698: Modern Statistical Methods in Epidemiologic Studies (3 units) [Integrative Capstone Experience]

Student with BS in Computer Science

A student entering the program with a Bachelor's degree in Computer Science would have expertise in some core areas, depending on the electives they have taken in their undergraduate program. Consider a student who has a satisfactory background in databases, data and web applications, and data mining/statistical learning/machine learning. Such a student may follow this trajectory, which requires 27 units in two terms:

FALL SEMESTER:

- MATH 425/ STATS 425 Probability (3 units) [Step iv of Statistics Sequence]
- SI 649: Information Visualization (3 units) [DS Computation Elective]
- BIOSTAT 650: Regression (4 units) [Core Area 3]
- SI 630: Natural Language Processing: Algorithms and People (3 units) [Free Elective]
- Data Science Seminar (1 unit)

WINTER SEMESTER:

- BIOSTAT 695: Categorical Data (3 units) [DS Analysis Elective]
- BIOSTATS 602: Biostatistical Inference (4 units) [Step v of Statistics Sequence]
- EECS 553: Theory and Practice of Data Compression (3 units) [Principles of DS Elective]
- EECS 599: Independent Study (3 units) [Integrative Capstone Experience]

Student with BS in Statistics or Biostatistics

A student entering the program with a Bachelor's degree in Statistics would likely have expertise in some core areas, depending on the electives they have taken in their undergraduate program. They may not have completed the full computing sequence requirement. Consider a student who has completed the first step of the computing sequence and has a satisfactory background in the areas of probability, statistical inference, regressions, and data mining/statistical learning/machine learning. Such a student may follow this trajectory, which requires 32 units in three terms:

FALL SEMESTER:

- STATS 510: Probability and Distribution Theory (3 units) [Principles of DS Elective]
- EECS 402: Programming for Scientists and Engineers (4 units) [Step ii of Computing Sequence]

- SI 618: Data Manipulation and Analysis (3 units) [DS Analysis Elective]
- Data Science Seminar (1 unit)

WINTER SEMESTER:

- SI 649: Information Visualization (3 units) [DS Computation Elective]
- EECS 403[4]: Data Structures and Algorithms (4 units) [Step iii of the Computing Sequence]
- EECS 545 Machine Learning (4 units) [Advanced Graduate Class in a Core Area]

FALL SEMESTER:

- EECS 484: Database Systems (4 units) [Core Area 1]
- BIOSTAT 699: Analysis of Biostatistical Investigations (4 units) [Integrative Capstone Experience]
- SI 650: Information Retrieval (3 units) [Core Area 2]

Student with BS in Information

A student entering the program with a Bachelor's degree in Information, may have gained considerable proficiency in several core areas, depending on the courses they selected during their undergraduate degree program. However, other core areas may be lacking. Consider such a student, who meets requirements for only step ii of the computing sequence and the areas of data and web applications, and regression, but none other. Such a student may follow this trajectory, which requires 34 units in three terms:

FALL SEMESTER:

- STATS 510: Probability and Distribution Theory (3 units) [Step iv of Statistics Sequence]
- MATH 465: Introduction to Combinatorics (3 units) [Step i of the Computing Sequence]
- SI 608 (Networks) (3 units) [DS Analysis Elective]
- Data Science Seminar (1 unit)

WINTER SEMESTER:

- EECS 403[20]: Data Structures and Algorithms (4 units) [Step iii of the Computing Sequence]
- STATS 511: Statistical Inference (3 units) [Step v of Statistics Sequence]
- SI 630: Natural Language Processing: Algorithms and People (3 units) [DS Computation Elective]
- BIOSTAT 617: Sample Design (3 units) [Principles of DS Elective]

FALL SEMESTER:

- EECS 484: Database Systems (4 units) [Core Area 1]

[4] New course, currently in planning, expected first offering in Winter 2019.

- EECS 545: Machine Learning (4 unites) [Core Area 4]
- SI 699-004 (3 units) Big Data Analytics [Integrative Capstone Experience]

Student with BS in Mathematics or Physics

A student entering the program with a Bachelor's degree in a quantitative discipline other than Statistics and Computer Science may not have gained proficiency in most core areas as part of their undergraduate training. They may also not have completed the Computing Sequence. However, they would have knowledge of pre-requisites, such as multi-variable calculus, linear algebra and undergraduate probability, so that they can gain the required proficiency level with one course per core area. Furthermore, we will require, as a condition of admission, that students meet at least some requirements. Consider such a student, who has completed the first two steps of the computing sequence, as well as the areas of probability and regression. Such a student may follow this trajectory, which requires 33 units in three terms:

FALL SEMESTER:

- EECS 403[20]: Data Structures and Algorithms (4 units) [Step iii of the Computing Sequence]
- SI 618: Data Manipulation (3 units) [Core Area 2]
- STATS 531: Times Series Analysis (3 units) [DS Analysis Elective]
- Data Science Seminar (1 unit)

WINTER SEMESTER:

- EECS 549/ SI 650: Information Retrieval (4 units) [DS Computation Elective]
- STATS 503: Statistical Learning and Multivariate Analysis (4 units) [Core Area 4]
- BIOSTAT 602: Statistical Inference (4 units) [Step v of Statistics Sequence]

FALL SEMESTER:

- EECS 484: Database Systems (4 units) [Core Area 1]
- STATS 551: Bayesian Analysis (3 units) [Principles of DS Elective]
- STATS 750: Independent Study (3 units) [Integrative Capstone Experience]

Appendix A: Descriptions of Core Courses

BIOSTAT 601: Probability and Distribution Theory (4). Fundamental probability and distribution theory needed for statistical inference. Probability, discrete and continuous distributions, expectation, generating functions, limit theorems, transformations, sampling theory. Satisfies statistical sequence requirement c-iv.

BIOSTAT 602: Biostatistical Inference (4). Fundamental theory that is the basis of inferential statistical procedures. Point and interval estimation, sufficient statistics, hypothesis testing, maximum likelihood estimates, confidence intervals, criteria for estimators, methods of constructing test and estimation procedures. Satisfies statistical sequence requirement c-v.

BIOSTAT 616: Machine learning for health sciences (3). This course will approach "machine learning" topics for large and high-dimensional data, traditionally in the overlapped domains of computer science and statistics, from the point of view of biostatistics. It will include an overview of graphical models; multiple comparisons and false discovery rate control; use of shrinkage, LASSO, and Bayesian variable selection in linear models; kernel methods; unsupervised learning methods including clustering, scaling, and principle components; model averaging and boosting; and computational algorithms and methods for complex systems. Satisfies core 4 requirement.

BIOSTAT 650: Applied Statistics I (3). The objective of this course is to help the student learn to plan strategies for linear regression analysis based on the scientific goals of a study and to implement these strategies. The student will learn to be aware of problems that arise in study design, power and data collection. The student will also learn to interpret the results of linear regression analysis and convert them into a language understandable to the broad scientific community. Satisfied core 3 requirement.

EECS 402: Programming for Scientists and Engineers (4). This course presents concepts and hands-on experience for designing and writing programs using one or more programming languages currently important in solving real-world problems. Intended for senior undergraduates and graduate students in science or engineering fields. Satisfies computational sequence requirement c-ii.

EECS 403: Data Structures and Algorithms for Scientists and Engineers (4). This course, currently under development, is expected to serve as an introduction to algorithm analysis and O-notation; Fundamental data structures including lists, stacks, queues, priority queues, hash tables, binary trees, search trees, balanced trees and graphs; searching and sorting algorithms; recursive algorithms; basic graph algorithms; introduction to greedy algorithms and divide and conquer strategy. Several programming assignments. Satisfies computational sequence requirement c-iii.

EECS 445: Intro. Machine Learning (4). Theory and implementation of state of the art machine learning algorithms for large-scale real-world applications. Topics include supervised learning (regression, classification, kernel methods, neural networks, and regularization) and unsupervised learning, (clustering, density estimation, and dimensionality and reduction). Satisfies core 4 requirement.

EECS 484: Database Systems (4). Concepts and methods for the design, creation, query and management of large enterprise databases. Functions and characteristics of the leading database management systems. Query languages such as SQL, forms, embedded SQL, and application development tools. Database design, integrity, normalization, access methods, query optimization, transaction management and concurrency control and recovery. Satisfies core 1 requirement.

EECS 485: Web Database Systems (4). Design and use of databases in the Web context; data models, database design, replication issues, client/server systems, information retrieval, web server design; substantial project involving the development of a database backed web site. Satisfies core 2 requirement.

EECS 486: Information Retrieval and Web Search (4). This course will cover traditional material, as well as recent advances in Information Retrieval (IR), the study of indexing, processing, querying, and classifying data. Basic retrieval models, algorithms, and IR system implementations will be covered. While the course will primarily focus on IR techniques for textual data, it will also address IR for other media, including images/videos, music/audio files, and geospatial information. The course will also address topics in Web search, including Web crawling, link analysis, search engine development, social media, and crowdsourcing. Satisfies core 2 requirement.

EECS 545: Machine Learning (4). Survey of recent research on learning in artificial intelligence systems. Topics include learning based on examples, instructions, analogy, discovery, experimentation, observation, problem-solving and explanation. The cognitive aspects of learning will also be studied. Satisfies core 4 requirement.

EECS 584: Advanced Database Systems (4). Advanced topics and research issues in database management systems. Distributed databases, advanced query optimization, query processing, transaction processing, data models and architectures. Data management for emerging application areas, including bioinformatics, the internet, OLAP and data mining. A substantial course project allows in-depth exploration of topics of interest. Satisfies core 1 requirement.

MATH 425/STATS 425: Intro. Probability (3). Topics include the basic results and methods of both discrete and continuous probability theory: conditional probability, independent events, random variables, jointly distributed random variables, expectations, variances, covariances. Satisfies statistical sequence requirement c-iv.

MATH 465: Intro. Combinatorics (3). An introduction to combinatorics, covering basic counting techniques (inclusion-exclusion, permutations and combinations, generating functions) and fundamentals of graph theory (paths and cycles, trees, graph coloring). Additional topics may include partially ordered sets, recurrence relations, partitions, matching theory, and combinatorial algorithms. Satisfies computational sequence requirement c-i.

SI 618: Data Manipulation and Analysis (3). This course aims to help students get started with their own data harvesting, processing, aggregation, and analysis. Data analysis is crucial to evaluating and designing solutions and applications, as well as understanding user's information needs and use. In many cases the data we need to access is distributed online among many webpages, stored in a database, or available in a large text file. Often these data (e.g. web server logs) are too large to obtain and/or process manually. Instead, we need an automated way of gathering the data, parsing it, and summarizing it, before we can do more advanced analysis. Therefore, students in this course will learn to use Python and its modules to accomplish these tasks in a 'quick and easy' yet useful and repeatable way. Next, students will learn techniques of exploratory data analysis, using scripting, text parsing, structured query language, regular expressions, graphing, and clustering methods to explore data. Students will be able to make sense of and see patterns in otherwise intractable quantities of data. Satisfies core 2 requirement.

SI 630: Natural Language Processing (3) This course focuses on how to use machine learning techniques to understand, annotate, and generate the language we see in everyday situations. The techniques learned in this course can be applied to any kind of text and enable turning qualitative evaluation of text in a precise

quantitative measurement. Students will learn the linguistics fundamentals of natural language processing (NLP), with specific topics of part of speech tagging, syntax and parsing, lexical semantics, topic models, and machine translation. Additional advanced topics will include sentiment analysis, crowdsourcing, and deep learning for NLP. Satisfies core 2 requirement.

SI 649/EECS 548: Information Visualization (3). Introduction to information visualization. Topics include data and image models, multidimensional and multivariate data, design principles for visualization, hierarchical, network, textual and collaborative visualization, the visualization pipeline, data processing for visualization, visual representations, visualization system interaction design, and impact of perception. Emphasizes construction of systems using graphics application programming interfaces (APIs) and analysis tools. Satisfies core 2 requirement.

SI 650/EECS 549: Information Retrieval (3). Information retrieval studies the interaction between users and large information environments. An in-depth survey of the field from classic concepts to state-of-the-art applications such as crawlers and spiders. Topics include information need, documents and queries, indexing and searching, retrieval evaluation, multimedia and hypertext search, Web search, and bibliographical databases. Satisfies core 2 requirement.

SI 671: Data Mining: Methods and Applications (3). Automatic, robust, and intelligent data mining techniques have become essential tools to handle heterogeneous, noisy, nontraditional, and large-scale data sets. This is a doctoral seminar course of advanced topics in data mining, the state-of-the-art methods to analyze different genres of information, and the applications to many real world problems. The course will highlight the practical applications of data mining instead of the theoretical foundations of machine learning and statistical computing. The course materials will focus on how the information in different real world problems can be represented as particular genres, or formats of data, and how the basic mining tasks of each genre of data can be accomplished using the state-of-the-art techniques. To this end, the course is not only suitable for doctoral students who are doing research in data mining related fields, but also for those who are consumers of data mining techniques in their own disciplines, such as natural language processing, networks science, human computer interaction, economics, social computing, sociology, business intelligence, and biomedical informatics, etc. Satisfies core 4 requirement.

STATS 413: Applied Regression Analysis (4). The following topics will be covered: a) models and methods of inference for simple and multiple regression, regression splines; b) diagnostics, multicollinearity, influence, outliers, transformation, model selection, and dimension reduction; c) principal component regression, ridge and robust regression, non-linear regression, non-parametric regression, and Lasso; d) generalized linear models, binary and Poisson regression. Satisfies core3 requirement.

STATS 415: Data Mining (4). This course covers the principles of data mining, exploratory analysis and visualization of complex data sets, and predictive modeling. Topics include: a) techniques and algorithms for exploratory data analysis and for discovering associations, patterns, changes, and anomalies in large data sets; and b) modern methods for multivariate analysis and statistical learning in regression, classification, and clustering. The presentation balances statistical concepts (such as model bias and over-fitting data, and interpreting results) and computational issues (including algorithmic complexity and strategies for efficient implementation). Satisfies core 4 requirement.

STATS 426: Intro. to Theoretical Statistics (3). This course covers the basic ideas of statistical inference, including sampling distributions, estimation, confidence intervals, hypothesis testing, regression, analysis of variance, nonparametric testing, and Bayesian inference. Satisfies statistical sequence requirement c-v.

STATS 500: Statistical Learning I: Regression (3). Linear models: definitions, fitting, identifiability, collinearity, Gauss-Markov theorem, variable selection, transformation, diagnostics, outliers and influential observations. ANOVA and ANCOVA. Common Designs. Applications and real data analysis are stressed, with students using the computer to perform statistical analyses. Satisfies core 3 requirement.

STATS 503: Statistical Learning II: Multivariate Analysis (4). The course covers methods for modern multivariate data analysis and statistical learning, including both their theoretical foundations and practical applications. Topics include principal component analysis and other dimension reduction techniques, classification (discriminant analysis, decision trees, nearest neighbor classifiers, logistic partitioning methods, model-based methods), and categorical data analysis. There will be a significant data analysis component. Satisfies core 4 requirement.

STATS 510: Probability and Distribution Theory (3). Essential concepts of probability and distribution theory that are important for statistical inference including: random variables, probability, conditional probability, distribution functions, independence, modeling dependence, transformations, quantiles, order statistics, laws of large numbers, central limit theorem, and sampling distributions. Satisfies statistical sequence requirement c-iv.

STATS 511: Statistical Inference (3). This is a graduate-level introductory course to key concepts, methods and theory in statistical inference. The topics covered will include univariate and multivariate families of distributions, likelihood principle, point estimation, confidence regions, hypothesis tests, large sample properties, and other selected topics in contemporary methods. Satisfies statistical sequence requirement c-v.

STATS 513: Regression and Data Analysis (3). The course is designed for graduate students interested in quantitative research to learn about regression models for data analysis. Topics include estimation and inference, diagnostics, model selection, and interpretation of results associated with linear models and general linear models. Additional topics may vary with the instructor. Satisfies core 3 requirement.

Appendix B: Lists of Approved Electives

Bucket A (Principles)
Courses in the first bucket cover principles and foundations of data science.

BIOSTAT 601 (Probability and Distribution Theory)
BIOSTAT 602 (Biostatistical Inference)
BIOSTAT 617 (Sample Design)
BIOSTAT 680 (Stochastic Processes)
BIOSTAT 682 (Bayesian Analysis)
EECS 501 (Probability and Random Processes)
EECS 502 (Stochastic Processes)
EECS 551 (Matrix Methods for Signal Processing, Data Analysis and Machine Learning)
EECS 553 (Theory and Practice of Data Compression)
EECS 564 (Estimation, Filtering, and Detection)
EECS 548 (Information Visualization) = SI 649
STATS 451 (Introduction to Bayesian Data Analysis)
STATS 470 (Introduction to Design of Experiments)
STATS 510 (Probability and Distribution Theory)
STATS 511 (Statistical Inference)
STATS 551 (Bayesian Modeling and Computation)
STATS 570 (Design of Experiments)

Bucket B (Data Analysis)
Courses in the second bucket are primarily concerned with statistical tools for analyzing data

BIOSTAT 645 (Time series)
BIOSTAT 651 (Generalized Linear Models)
BIOSTAT 653 (Longitudinal Analysis)
BIOSTAT 665 (Population Genetics)
BIOSTAT 666 (Statistical Models and Numerical Methods in Human Genetics)
BIOSTAT 675 (Survival Analysis)
BIOSTAT 685 (Non-parametric statistics)
BIOSTAT 695 (Categorical Data)
BIOSTAT 696 (Spatial statistics)
EECS 556 (Image Processing)
EECS 559 (Advanced Signal Processing)
EECS 659 (Adaptive Signal Processing)
STATS 414 (Topics in Applied Data Analysis)
STATS 449 (Applied Survival Analysis)
STATS 501 (Statistical Analysis of Correlated Data)
STATS 503 (Statistical Learning II: Multivariate Analysis) [If not used to satisfy core]
STATS 509 (Statistics for Financial Data)
STATS 531 (Analysis of Time Series)
STATS 600 (Linear Models)
STATS 601 (Analysis of Multivariate and Categorical Data)
STATS 605 (Advanced Topics in Modeling and Data Analysis)
STATS 700 (Topics in Applied Statistics)

Bucket C (Computation)
Courses in the third bucket are primarily involved with computational aspects of implementing data science methods.

BIOSTAT 615 (Statistical Computing)
EECS 481 (Software Engineering)
EECS 485 (Web Systems) [If not used to satisfy core]
EECS 486 (Information Retrieval and Web Search) [If EECS486/EECS549/SI650 not used to satisfy core]
EECS 493 (User Interface Development)
EECS 504 (Computer Vision)
EECS 542 (Advanced Topics in Computer Vision)
EECS 549 (Information Retrieval) = SI 650 [If EECS486/EECS549/SI650 not used to satisfy core]
EECS 586 (Design and Analysis of Algorithms)
EECS 587 (Parallel Computing)
EECS 592 (Artificial Intelligence)
EECS 595 (Natural Language Processing) = SI 561
EECS 597 (Language and Information) = SI 760
SI 561 (Natural Language Processing) = EECS 595
SI 608 (Networks)
SI 630 (Natural Language Processing (Algorithms and People)
SI 650 (Information Retrieval) = EECS 549 [If EECS486/EECS549/SI650 not used to satisfy core]
SI 671 (Data Mining: Methods and Applications)
STATS 406 (Computational Methods in Statistics and Data Science)
STATS 506 (Computational Methods and Tools in Statistics)
STATS 606 (Statistical Computing)
STATS 607 (Programming and Numerical Methods in Statistics)
STATS 608 (Monte Carlo Methods and Optimization Methods in Statistics)

Bucket D (Capstone Integrative Experience)
Courses in the fourth bucket will give students experience in applying data science methodology

BIOSTAT 610 (Independent Study)
BIOSTAT 698 (Modern Statistical Methods in Epidemiologic Studies)
BIOSTAT 699 (Analysis of Biostatistical Investigations)
EECS 599 (Independent Study)
SI 699-004 (Big Data Analytics)
SI 699-00X (Computational Social Science)
STATS 504 (Statistical Consulting)
STATS 750 (Independent Study)