UCONN
UNIVERSITY OF CONNECTICUT

Office of the Provost
Anne D'Alleva, Ph.D.
Provost and Executive Vice President
for Academic Affairs

December 7, 2022

TO: Members of the Board of Trustees

FROM: Anne D'Alleva, Ph.D.
Provost and Executive Vice President for Academic Affairs

Anne D'Alleva

RE: Bachelor of Science in Statistical Data Science

RECOMMENDATION:

That the Board of Trustees approve a new undergraduate major in Bachelor of Science in Statistical Data Science in the College of Liberal Arts and Sciences.

BACKGROUND:

Data science and the analysis of quantitative data are rapidly growing fields that are revolutionizing science and society. The new Bachelor of Science in Statistical Data Science will provide undergraduate students with a major that prepares them for a range of careers in an area of rapid job growth. Data science-related jobs are anticipated to see significant growth in employment over the next decade or two. The U.S. Bureau of Labor Statistics anticipates an increase of about 40,000 data scientist positions between now and 2031, up about 36% (2021-2031), much faster than other occupations. This growth is faster than any other occupation requiring a college degree except for nurse practitioners. The median salary data scientists are currently $101,000. This data science major will provide skills that will make students competitive for other related high-growth occupations, like statisticians and web developers.

This undergraduate major will be offered for STEM students who need to acquire fundamental skills and competence for data-rich organizational contexts. The BS is housed in the Department of Statistics and has an advisory board which makes decisions about the curriculum, including approval and re-authorization (from time to time) of core skill and domain courses. We anticipate that the Bachelor of Science in Statistical Data Science will attract additional students who are interested in a data-focused interdisciplinary STEM education experience. Given the application process to the major and course requirements, we expect a cohort of 20-25 students per semester. Based on this distribution, we would estimate around 175 students in the program once it has been established. The New England Regional Tuition Break program may also bring an additional 40-70 students since most of the surrounding states do not have this type of undergraduate BS degree program. We estimate there will likely be a transition of around 50-75 students from the Mathematics-Statistics, Statistics, and Applied Mathematics programs into the new Statistical Data Science degree. We anticipate that this new degree will make UConn a destination for data science and will serve as a pipeline for the recently approved cross-college graduate program in data science.

UCONN

Proposal for
Bachelor of Science in Statistical Data Science

30.7001

College of Liberal Arts and Sciences

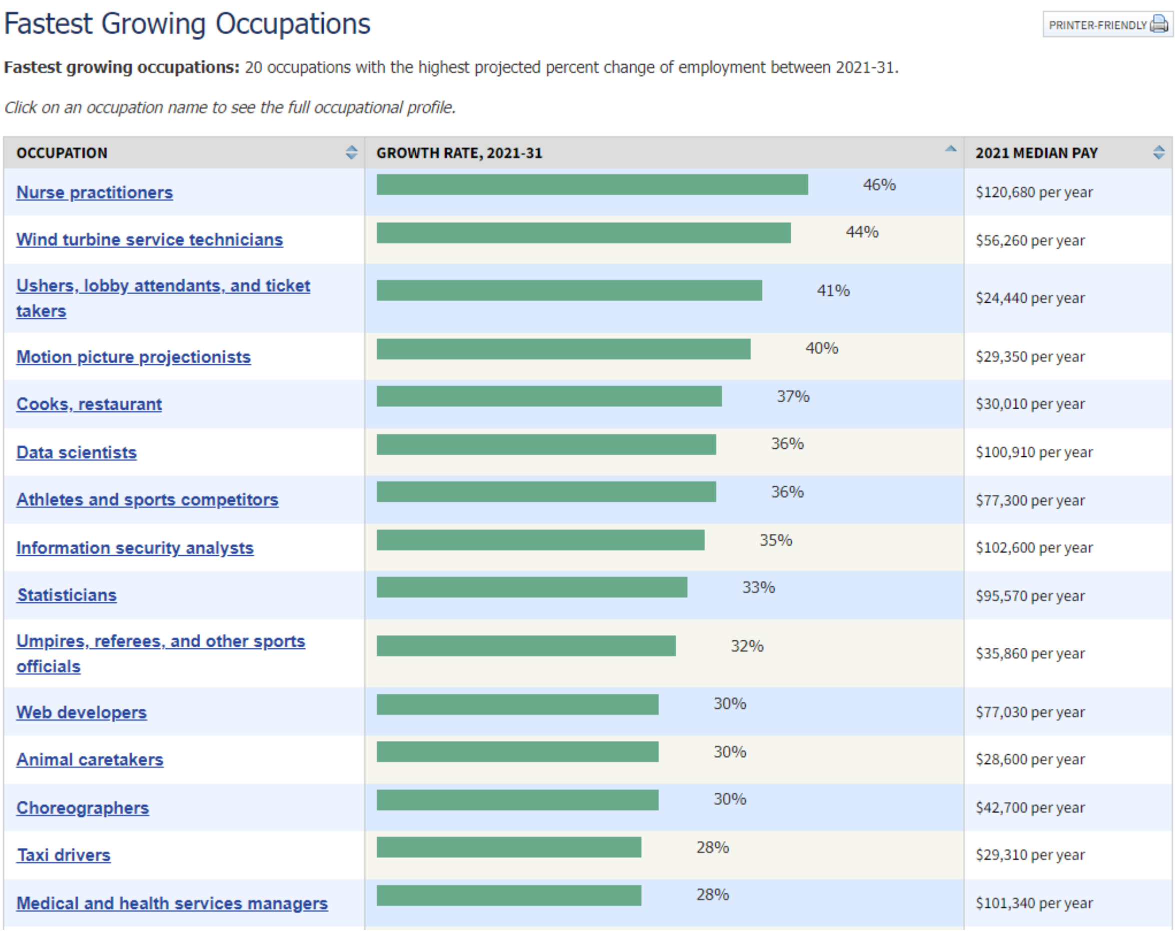
University of Connecticut

Introduction and Rationale

Data science and the analysis of quantitative data are rapidly growing fields that are revolutionizing science and society. Work is becoming increasingly more data-driven, and this affects the jobs that are available and the skills that are required. As data and data analysis tools become more widely available, more aspects of the economy, society, and daily life will become dependent on them. While today the term “data scientist” typically describes a knowledgeable worker who is principally occupied with analyzing complex and massive data resources, data science spans a much broader array of activities. These involve applying data science principles for data collection, storage, integration, analysis, inference, communication, and ethics. In future decades, undergraduates interested in many specialties will benefit from a fundamental awareness of and competence in data science.

The changing workplace requires more and more people with a basic understanding of data science and a substantial cadre of talented graduates with highly developed data science skills, acquired through substantial coursework and practice. Graduates of these types of programs can expect to find work in almost all occupational realms and will serve in a number of roles, including operating and designing the analytical systems, preparing data, coordinating analysis, visualizing output, and supporting data-driven decision making. Journalists, administrators in the public and private sector, artists, lawyers, teachers, and others will also increasingly need to understand and use data. Hence there is a great need to prepare students for the data-enriched world of the rest of this century.

Data science-related jobs are anticipated to see significant growth in employment over the next decade or two. The BLS anticipates an increase of about 40,000 “data scientist” positions between now and 2031, up about 36% (2021-2031) for the occupation of Data scientist”, much faster than other occupations, and faster than any other occupation requiring a college degree except for nurse practitioner (Figure 1). The median salary data scientists are currently well above average: $101,000.[1] Data science training provides skills that would make students competitive for other related high growth occupations, like statisticians and web developer. About 120 colleges and universities currently offer Data Science BA degrees, including many of UConn's peer and aspirant institutions: such as Boston University, Iowa State, Northeastern, Penn State, Purdue, Rutgers, SUNY-Albany, UC-Davis, UC-Irvine, University of Georgia, University of Iowa, UMass-Dartmouth, and URI.



Fastest Growing Occupations

Fastest growing occupations: 20 occupations with the highest projected percent change of employment between 2021-31.

Click on an occupation name to see the full occupational profile.

OCCUPATION	GROWTH RATE, 2021-31	2021 MEDIAN PAY
Nurse practitioners	46%	$120,680 per year
Wind turbine service technicians	44%	$56,260 per year
Ushers, lobby attendants, and ticket takers	41%	$24,440 per year
Motion picture projectionists	40%	$29,350 per year
Cooks, restaurant	37%	$30,010 per year
Data scientists	36%	$100,910 per year
Athletes and sports competitors	36%	$77,300 per year
Information security analysts	35%	$102,600 per year
Statisticians	33%	$95,570 per year
Umpires, referees, and other sports officials	32%	$35,860 per year
Web developers	30%	$77,030 per year
Animal caretakers	30%	$28,600 per year
Choreographers	30%	$42,700 per year
Taxi drivers	28%	$29,310 per year
Medical and health services managers	28%	$101,340 per year

Figure 1: Fastest growing Occupations 2021-2021 and median incomes according to US Bureau of Labor Statistics. https://www.bls.gov/ooh/fastest-growing.htm

Enrollment Projections

Prospective students for this program will be STEM students who need to acquire fundamental skills and competence for data-rich organizational contexts and considering careers in these industries. Given the emerging nature of data science in institutions with similar profiles to UConn, we can only provide enrollment projections based on educated estimates and understanding of our existing institutional

[1] https://www.bls.gov/ooh/math/data-scientists.htm

context. We anticipate that the Bachelor of Science in Statistical Data Science will attract additional students who are interested in a data-focused interdisciplinary STEM education experience. Given the application process to the major and course requirements, we expect a cohort of 20-25 students per semester. Based on this distribution, we would estimate around 175 students in the program once it has been established. The New England Regional Tuition Break program may also bring an additional 40-70 students since most of the surrounding states do not have this type of undergraduate BS degree program. We further suspect that there may be a transition of 50-75 students from the Mathematics-Statistics, Statistics, and Applied Mathematics programs into the more modern Statistical Data Science degree. We anticipate that this new degree will make UConn a destination for Data Science and will serve as a pipeline for the recently approved cross-college graduate program in Data Science.

Required Resources

Specific courses and requirements are detailed below. The program that we have devised contains courses that are all taught in CLAS. The BS is housed in the Department of Statistics and has an advisory board which makes decisions about the curriculum, including approval and re-authorization (from time to time) of core skill and domain courses. Depending on the growth in the size of the major, additional sections (or more frequent offerings) of existing courses (e.g., STAT2255, STAT3255, STAT3215Q, STAT4255) may be needed in the future. Additional hires (both tenure- and non-tenure track) in Statistics are expected through the new MS in Data Science program; we expect these new hires will also be contributing to this undergraduate program, as well.

Justification

In March 2020, Dean Juli Wade called a meeting to discuss the creation of a new Undergraduate Data Science Program within CLAS, and asked interested CLAS Department Heads to appoint members to a CLAS Undergraduate Data Science Committee for this purpose. This committee, consisting of members of the Departments of Statistics, Political Science, Mathematics, Economics, Geography, Geosciences, Public Policy, Ecology and Evolutionary Biology, and Molecular Cell Biology, has since been meeting regularly to develop this major, as well as the BA in Applied Data Analysis. In addition, the Departments of Philosophy, Cognitive Sciences, Sociology, and Marine Science were involved in aspects of the curriculum. The curriculum for the degree was approved by the CLAS Courses and Curriculum Committee on 10/18/2022.

Analytic training includes courses where students will learn to:

-- formulate good questions and determine the types of data appropriate to answer those questions,
-- collect, retrieve, manipulate, store, analyze, and report on information *in an ethical manner,*
-- conduct work that is reproducible, and
-- make appropriate inferences from data analysis.

Students completing a BS in Statistical Data Science must attain competence in four core areas of *data science* as suggested in the 2018 National Academy of Science report *Data Science for Undergraduates: Opportunities and Options:*

A. Computer Programming, data generation, and analysis: Almost all data generation and analysis require the manipulation of large amounts of digitized information. Because most tasks associated with data analytics involve amounts of data that cannot be (re)processed by hand or involve processing data to be used by different analytical software and hardware and for different practical applications, degree recipients must have an elementary understanding of computer systems and languages, data structures and control, and algorithmic development and utilization. To be able to address specific, complex problems with students attaining the BA degree will learn to collect and manage the appropriate information such that it can be utilized effectively by individuals and organizations.

B. Data analysis: This skill includes a core set of features consistent with probability and statistics in quantitative data analysis: e.g., sampling, randomness, experimental/observational research designs, parametric/non-parametric estimation and inference up through at least multiple linear regression.

C. Data visualization: Visualization refers generally to the effective presentation and communication of data in a manner that can stand alone as a communication tool or that complements the narrative text. As part of data visualization training, students will learn modern visualization standards and how to use computer visualization tools. They will also learn to effectively communicate to different audiences and avoid engaging in the misrepresentation of data.

D. Ethics of data collection and use: The ethical challenges of collecting and using data to inform decision-making are enormous. This is particularly the case when much of the data used concerns observations about behaviors or characteristics collected without the full knowledge of those being observed. The very power of data science makes it important that all parts of the data science curriculum educate students about the ethical use of data science tools.

In addition to these skill requirements, students must learn about a specific substantive domain area.

After completing the skill and domain area training, students will conduct a final research project which applies all the core data science skills to a practical problem in or related to their domain area.

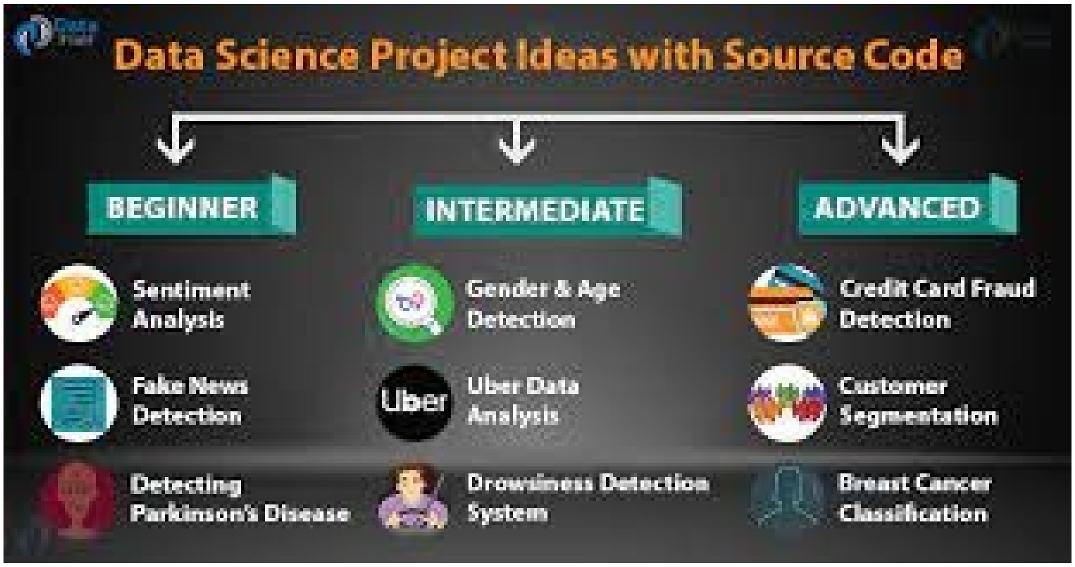

Data Science Project Ideas with Source Code
BEGINNER
INTERMEDIATE
ADVANCED
Sentiment Analysis
Gender & Age Detection
Credit Card Fraud Detection
Fake News Detection
Uber
Uber Data Analysis
Customer Segmentation
Detecting Parkinson's Disease
Drowsiness Detection System
Breast Cancer Classification

Figure 2: Examples of Data Science final projects ((https://data-flair.training/blogs/data-science-project-ideas/)

Bachelor of Science Degree Curriculum

The BS in Statistical Data Science requires 36 credits, with one or more courses in each of the core areas below, a nine-credit domain sequence, STAT3255 (Introduction to Data Science), and STAT 4915 (capstone)†. To satisfy the information literacy competency and writing in the major requirement, Statistical Data Science majors must also take STAT4916W†.

The core area requirements are:

1. *Programming and data management:* 1 course (3 credits): STAT 2255
2. *Basic Data Analysis:* 2 courses (6 credits): STAT 3025Q or STAT 3375Q* or MATH3160; STAT 3215Q
3. *Data Ethics:* 1 course (3 credits): PHIL 3202
4. *Data Visualization:* 1 course (at least 3 credits): STAT 3675Q* or GEOG 3510 or EEB 4100**
5. *Advanced analysis:* 2 courses (6 credits): MATH 2210Q; STAT 4255

†Students completing a Biological Data Science domain may take any of the following to meet the capstone and W requirement: (i) STAT 4915 / STAT4916W, (ii) EEB 4896W, or (iii) MCB 4897W. Credits in EEB 4896W cannot simultaneously count towards both an Honors thesis in EEB and a Data Science capstone.
*Students completing a Statistics domain must take STAT3375Q and STAT3675Q to meet these requirements.
** Recommended for students completing the Biological Data Science Domain.

To complete the nine-credit domain sequence, students must take at least nine credits from one of the following groups:

Advanced Statistics: STAT 3445 and two of the following: STAT 3515Q, STAT 4625, STAT 4825, STAT 4845, STAT 4190***
*** At least and no more than 3 credits of STAT4190 may count towards the major and must be pre-approved by the Department of Statistics for adequate data science content.

American Political Institutions: three of the following: POLS 3600, POLS 3601, POLS 3603WQ, POLS 3604, POLS 3606

American Political Representation: three of the following: POLS 2607, POLS 3608W, POLS 3612, POLS 3617, POLS 3618, POLS 3625

Biological Data Science: three of the following: EEB 3899‡, EEB 5050, EEB 5300, EEB 5348, EEB 5349, MCB 3637, MCB 4008, MCB 4009, MCB 4014, MCB 5430, MCB 5472, MCB 5631, MCB 4896‡
Students can choose any three courses‡ from the list above based on availability, however, interested students might consider choosing subsets of courses from the list above that align with established sub-areas:

- Genome sequencing and analysis: EEB 5300, MCB 3637, MCB 5430

- Phylogenetics and evolution: EEB 5348, EEB 5349, MCB 3421, MCB 5472
- Ecological analyses: EEB 5050, EEB 5348, MCB 5631
- Molecular structure and function: MCB 4008, MCB 4009, MCB 4014

‡ Only 3 credits of either EEB 3899 or MCB 4896 can count towards the major, and these credits cannot simultaneously count towards another major or degree.

Financial Analysis: three of the following: ECON 3313, ECON 3315, ECON 3413, ECON 4323

Marine Science: three of the following: MARN 3001, MARN 3002, MARN 3014, MARN 4001, MARN 4210Q

Population Dynamics: SOCI 2110(W), SOCI 2651(W), SOC 2660(W), SOCI 2820(W), SOCI 2901(W), SOCI 3971(W)

(Domain areas may be added by petitioning the advisory board.)

Explanation for core courses

STAT 2255 (Statistical Programming) addresses *programming and data management.* STAT 2255 introduces statistical programming via Python including data types, control flow, object-oriented programming, and graphical user interface-driven applications such as Jupyter notebooks. The emphasis of the course is on algorithmic thinking, efficient implementation of different data structures, control and data abstraction, file processing, and data analysis and visualization. The Python programming language is used because it is both accessible to beginners and widely used in real-world scientific programming. The concepts and skills are general, however, and will be helpful in mastering other programming languages as well.

STAT3215Q (Applied Linear Regression in Data Science) addresses *basic data analysis* as it covers simple linear regression and correlation analysis, multiple linear regression, analysis of variance, goodness of fit, comparing regression models through partial and sequential F tests, dummy variables, regression assumptions and diagnostics, model selection and penalized regression, prediction and model validation, principles of design of experiments, one-way and two-way analysis of variance. Beyond STAT3215Q, the Bachelor of Science *basic data analysis core* requires an additional three credits of a statistical methods course via **STAT3025Q** (Statistical Methods) or **STAT3375Q** (Introduction to Mathematical Statistics I) or **MATH3160** (Probability). This minimally includes basic probability distributions, point and interval estimation, tests of hypotheses, correlation and regression, analysis of variance, experimental design, and non-parametric procedures. (Additionally, STAT1000Q/STAT1100Q or equivalent is a prerequisite for entry into the major, and covers sampling, randomness, and experimental/observational research designs, among other topics.)

STAT3255 (Introduction to Data Science) addresses *all core areas of data science* by introducing data science for effectively storing, processing, analyzing, and making inferences from data. Topics include project management, data preparation, data visualization, statistical models, machine learning, distributed computing, and ethics. It also provides training in the ability to formulate good questions;

assess which kinds of data are appropriate to answer those questions; conduct ethical data collection, manipulation, and analysis that is reproducible; and make appropriate inferences based on the data.

To meet the core area of *data visualization*, students must take at least three credits of **GEOG3510** (Cartographic Techniques) or **STAT3675Q** (Statistical Computing) or **EEB 4100** (Big Data Science for Biologists). GEOG3510 covers methods for representing geographic data in tables, graphs, and maps emphasizing proper application, integration, and interpretation of methods in data visualization. STAT3675Q, while arguably also a *programming* course, covers dynamic reports, and both basic and advanced graphics (with ggplot2) in the R programming language with RStudio. EEB 4100 emphasizes data creation, integration, curation, manipulation, and visualization through interaction with real data from molecular biology, ecology, agriculture, evolutionary biology, and systems biology. The course uses both R/RStudio and Python/Jupyter Notebook.

The core area of *ethics* will be addressed in both **PHIL 3202** (Data Ethics), and also **STAT 3255** (Introduction to Data Science). PHIL 3202 will introduce students to issues of ethics and equity in the contemporary practice of data science. The ability to collect, store, process, and analyze ever greater amounts of data offers great opportunities, as well as potential perils. Topics to be covered will include systematic approaches to assessing ethical issues; privacy and confidentiality; defining research and the responsibilities associated with conducting ethical research; implicit and structural biases in data collection and analysis. STAT 3255 further covers the American Statistical Association's *Ethical Guidelines for Statistical Practice*, designed to help statistical practitioners make decisions ethically.

Both **MATH2210Q** (Applied Linear Algebra) and **STAT4255** (Introduction to Statistical Learning) are *advanced analysis* courses. MATH2210Q is an introduction to the techniques of linear algebra with elementary applications, covering systems of equations, matrices, determinants, linear transformations on vector spaces, characteristic values and vectors, from a computational point of view. STAT4255 covers modern statistical learning (also called "machine learning") methods arising frequently in data science and predictive modeling with real-world applications, including linear and logistic regression, generalized additive models, decision trees, boosting, support vector machines, and neural networks (deep learning).

The capstone course **STAT 4915** (Data Science in Action) will allow students to combine their domain knowledge with the core areas of data science in a final culminating research project. Biological Data Science students may alternatively opt to take either **EEB 4896W** or **MCB 4897W** to satisfy the capstone (and W) requirement. EEB 4896W (Senior Research Thesis in Ecology and Evolutionary Biology; 3 credits) and MCB 4897W (Research Thesis in MCB; 3 credits) will provide a capstone writing synthesis describing a unique data analysis project completed by each student, designed in collaboration with their supervising EEB or MCB faculty member (i.e., their Thesis Mentor). Projects will include computational analyses of big datasets, including problem-specific programming (e.g., using shell, R, and/or Python), statistical analysis, and data visualization.

Students meet the university "writing in the major" requirement through **STAT4916W** (Writing in Data Science), co-requisite with the STAT4915 capstone course. Students in the Biological Data Science domain may alternatively opt to take either **EEB 4896W** or **MCB 4897W** to satisfy the W (and capstone) requirement.

Information literacy involves a general understanding of and competency in three integrally related processes:

- Information development and structure – an understanding of how information is created, disseminated and organized;
- Information access – an understanding of information communication processes and a facility with the tools required to tap into these processes;
- Information evaluation and integration – an ability to evaluate, synthesize and incorporate information into written, oral, or media presentations.

In addition to the basic competency achieved in ENGL 1007, ENGL 1010, ENGL 1011, ENGL 2011 or equivalent, students will receive instruction on how to conduct an effective search for information on the web for applicable topics in the required capstone and W courses.