Proposal for a Major in Data Science

Executive Summary

Data Science is the interdisciplinary study of methods used to gain knowledge from diverse sets of data, which are frequently quite large and multi-dimensional. Students trained in Data Science will study a blend of topics from many subdomains of communications, philosophy, mathematics, computer science, and information science. This combination allows Data Science students to efficiently conduct computational analyses within their own knowledge domain, manage teams of more specialized individuals to answer far-ranging questions, and communicate technical findings to diverse audiences.

Data Science differs from applied mathematics, statistics, or computer and information science in that a data scientist has a breadth of experience across all of these fields but may not have as much knowledge as a specialist in any particular field.

Demand for a major in Data Science is high. Since the inception of the Data Science program in the fall of 2017, student enrollment has steadily risen and currently, as of January 2019, the program has approximately 60 students enrolled (34 self-designed majors and 26 minors). To meet the ever-increasing student demand, the Data Science Program steering committee concludes that a structured B.S. major in Data Science needs to be offered at William & Mary during the 2019/2020 academic year.

This document summarizes the necessary steps required to craft a degree program that the steering committee believes would be eligible for SCHEV-STEM certification, on the basis of on an analysis of peer and other institutions' implementations of Data Science curricula.

Justification for a Data Science Major at W&M

To quote President Rowe in a recent article in the Washington Post:

> "College graduates must have the essential capacities to be agile, multidimensional and adaptable critical thinkers, prepared for the rapid pace of change we know lies ahead — and for what we cannot yet know. The speed of technological transformation is indisputable. And that speed places a premium on intellectual agility, multidimensionality and the ability to navigate change ethically and in a data-informed and values-driven way" (https://www.washingtonpost.com/education/2018/11/09/problem-solvers-tomorrows-workforce/?utm_term=.d2f483e69b4a).

We concur with President Rowe, and argue that in the context of William & Mary, a major in Data Science is justified for three primary reasons. First, William &Mary is well positioned to lead the effort to train data scientists with a degree in the liberal arts, a unique and competitive niche that employers demand in response to the rapid increase in complex, data-driven problems. Second, there is clear student demand for a major. Third, a degree in Data Science is distinct from other degrees such as Computational and Applied Mathematics and Statistics (CAMS) or Business Data Analytics (BUAD). We elaborate on each of these points below.

William & Mary is poised to become a leader in training data scientists who have a background in the liberal arts; these are the exact kind of graduates that employers seek to meet growing demand for data scientists who can communicate effectively and think creatively about data-driven problems. As a liberal arts university, the Data Science program will integrate with all disciplines of Arts and Sciences and the Business School. Our comparative advantage--compared to other universities with data science programs--is our ability to train students with not only data science skills but also the communication and critical thinking skills needed to effectively define, discuss, and solve complicated problems related to the management, analysis, and dissemination of vast amounts of information.

Second, there is clear demand among William & Mary students for a formalized major. At present, the Data Science program offers a minor and self-designed major in Data Science. Since the inception of Data Science in the fall semester of 2017, student enrollment has steadily increased. As of January 2019, approximately 34 students have declared a self-designed major and 26 a minor. Given the ever-increasing student demand, the Data Science Steering committee believes a structured B.S. in Data Science is needed to meet student demand and to more effectively and efficiently administer the Data Science program.

Third, the proposed B.S. in Data Science differs from closely aligned programs in CAMS and BUAD in important ways. Requirements for majors in CAMS and BUAD only tangentially cover Data Science and do not include the breadth of courses required in a proposed B.S. in Data Science. As two illustrative examples, majors in CAMS' Applied Statistics track, the major most closely related to the Data Science major, are required to take three courses in micro- and macro-economics and five courses in fundamental mathematical statistics. For electives, students take two courses in advanced economics, and two to three courses in advanced mathematics and computer science. The BBA in Business Analytics with Data Science requires students to take at least 18 credits in business fundamentals, such as accounting and business management, 12 credits in business analytics, and one elective course from a pool of six courses, with only one choice in Data Science (BUAD 460: Big Data Analytics).

In contrast, the proposed B.S. in Data Science exposes students to a suite of courses in Data Science applications, including deliberation (considering the ethical, moral, and societal implications of data science) and communication, and courses that promote critical evaluation of how data can be used to solve novel problems. In particular, the proposed B.S. in Data Science requires students to take five courses with a focus on programing, modeling, machine learning, data visualization, database structures, and ethics in Data Science. Many of these courses are unique to the B.S. in Data Science, and not offered by any other programs or departments. Moreover, students will also take two courses in mathematical statistics and one in linear algebra. Last, again unique to the B.S. in Data Science, students are required to take three courses from one of the following tracks: algorithm track (Computer Science), spatial track (Geographical Information System), statistical mathematics track (Mathematics), or data application track (Interdisciplinary). The latter track will include new courses not offered in any other program/department at William & Mary. These courses, in part, will be designed by new TE/NTE faculty and will overlap with their areas of expertise in Data Science.

Summary of Proposed Curriculum (Total 41 credit hours)

We developed the B.S. in Data Science on the basis of feedback from the Data Science Steering committee, representing most of the disciplines within Arts and Science that have been affected by the data science revolution: Josh Burk (Psychological Sciences), Matthias Leu (Biology), Liz Losh (American Studies and English), Michael Lewis (Computer Science), Dan Parker (Linguistics), Dan Runfola (Applied Science), Jaime Settle (Government), and Leah Shaw (Mathematics).

We also sought input from the private sector, particularly from a senior data scientist at Intel (Leah Shaw) and attended a meeting with program directors at Automatic Data Processing and the Navy (Matthias Leu). We learned that training in statistics, mathematics (linear algebra), and computer science are a must. When choosing among applicants, Intel hires applicants who have the most extensive training in statistics. In our conversation with Automatic Data Processing and the Navy, we learned that they are looking for data scientists who, in addition to being quantitatively strong, can also explain and sell analyses to clientele in simple terms. We included this feedback in designing the curriculum for a B.S. in Data Science to ensure that graduates are competitive in the job market, and will further incorporate it into future decisions for the B.A. in Data Science.

The proposed B.S. in Data Science includes training in programming, statistics, data visualization, and deliberation, and a core set of courses that advance methodological skills. The minimum number of credit hours for the major is 41 with an additional three to six credit hours of prerequisites depending on students' prior preparation.

Prerequisites (maximum 6 credit hours)
MATH 111 and 112 or Math 131 or 132 Calculus I & II)

Core Coursework (13 credit hours)
DATA 141(Introductory Programming for Data Science)
DATA 146 (Reasoning under Uncertainty)
DATA 2XX (Ethics in Data Science / Deliberation Requirement)
DATA 2XX (Data Visualization)

Data Analytics (6 credit hours)
DATA 301 (Applied Machine Learning)
DATA 3XX (Databases and Structures)

Linear Algebra and Statistical Modeling (9 credit hours)
MATH 211 (Linear Algebra)
MATH 351 (Probability & Statistics for Scientists *or* MATH 451 [Probability])
MATH 352 (Statistical Data Analysis *or* MATH 452 [Mathematical Statistics])

Substantive (Choose 3 courses; total of 9 credit hours)
Three additional methods-oriented courses taken from aligned departments or programs, generally organized into tracks and approved by the relevant department or program. Example tracks might include: spatial data track (rooted in GIS), algorithms track (rooted in Computer

Science and Mathematics), statistical mathematics track (rooted in Mathematics) and data application track (Interdisciplinary).

Capstone (4 credit hours)
DATA 4XX (Senior Capstone, COLL 400)

Expanded Coursework Descriptions

This section provides more information on the proposed Data Science major curriculum, as well as details on what courses would be new or redesigned.

Core Coursework (13 credit hours)
Provides students with the basic knowledge required to succeed in upper level Data Science coursework, as well as fundamental basics (programming, statistics and modeling, data visualization) required to succeed in the field.

DATA 141 - Programming for Data Science: This is an existing course. Per an agreement between the Data Science program and Computer Science, this existing course is designed and offered by the Computer Science department. This is to facilitate students taking CSCI 141 *or* DATA 141 being able to continue along the DATA or CSCI track after they have had exposure to these disciplines.
Catalog Description: An introduction to computational problem solving in the context of data science and commonly used data analysis software.

DATA 146 - Reasoning under Uncertainty: This is an existing course. This course is offered on a temporary basis by the Computer Science department, and under this plan would transition to being offered by faculty formally reporting to the Data Science program.
Catalog Description: A computationally-oriented exploration of quantitative reasoning for situations in which complete information is not available. Topics will include an introduction to discrete probability theory, Monte Carlo simulations sampling theory, and elementary game theory.

DATA 2XX – Deliberation: This is a modified course, focused on ethical, societal, and/or policy implications of Data Science. A version of this course is currently taught as DATA 150 (Ethics in Data Science), and would be modified to accommodate upper-level and transfer students.

DATA 2XX - Data Visualization: This a new course focused on data visualization. No similar courses are currently offered at William & Mary, but are common to Data Science degree programs.

Data Analytics (6 credit hours)
Courses which provide students with key skills and critical thinking required for big data analytics.

DATA 301 - Applied Machine Learning: This is a modified course. DATA 301 is currently taught under the name “Data Driven Decision-making”, and under this proposal would be altered to focus on a coverage of applied machine learning topics.

DATA 3XX - Databases and Structures: This is new course preparing students for common database programs found in data analytics (i.e., SQL; Hadoop-based infrastructure) and provides an applications-based education on theoretical data structures. A similar course (CSCI 241 – Data Structure) is currently offered at William & Mary, but lacks the focus on applications and theory necessary for data structures specific to Data Science (especially Hadoop-like infrastructure that focuses on exceptionally large datasets).

Linear Algebra and Statistical Modeling (9 credit hours)
Courses which provide students with core statistical knowledge, focusing on providing more breadth than DATA 146 – Reasoning under Uncertainty.

MATH 211 – Linear Algebra: This an existing course offered by the Department of Mathematics.

MATH 351 or (MATH 451 with additional prerequisites) - Probability and Statistics for Scientists: This an existing course offered by the Department of Mathematics. A statistical and modeling background is critical for students to be prepared to engage in Data Science problems. Although DATA 146 (Reasoning under Uncertainty) will cover the basic concepts of parametric modeling, MATH 351 provides much more depth to students in mathematical statistics.

MATH 352 or (MATH 452 with additional prerequisites) - Statistical Data Analysis: This an existing course offered by the Department of Mathematics. The second component of mathematical statistics, again providing students with a more depth in statistical analyses.

Substantive Specialty Courses (pick 3 courses, 9 credit hours)
This pillar of the degree would focus on providing students with a substantive focus in data analytics. All such tracks would be built and approved in conjunction with the departments and programs that they might be housed within. A core data application track would also be offered for students that seek more depth within Data Science in particular. Examples of these tracks currently being pursued by students include a Spatial Data Track and an Algorithms Track. Other example tracks that are desired by students but would require negotiation with partner programs and departments include econometrics (Economics Track), survey design (Psychological Sciences/Government/Sociology Track), and bioinformatics (Biology Track).

We propose that students chose three courses in one of three tracks listed in Table 1. By no means is the list of tracks exhaustive and we foresee additional tracks to be added in the near future.

Table 1

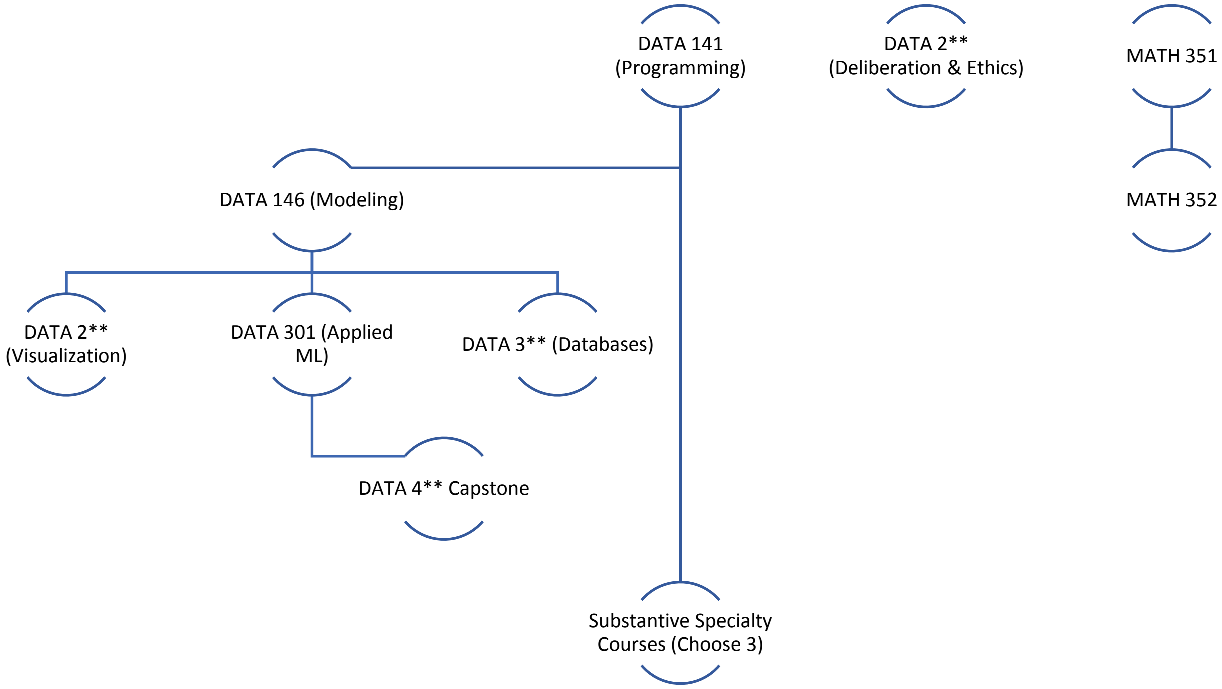
Examples of tracks currently pursued by Data Science students (Spatial Data and Algorithms Track) and a new track (Data Application track) to be implemented by the fall 2019. Some of the courses listed here would be new. Data Science affiliated faculty would be given the opportunity to design advanced elective courses on the basis of their interests, such as Bioinformatics, Data Driven Journalism, Computational Linguistics, Computational Geography etc.

Spatial Data Track	*Data Application Track*	*Algorithms Track*
GIS 201 (Intro. to GIS)	DATA 400-level courses being taught in the specialty areas of the new TE & NTE hires	CSCI 241 (Data Structures)
GIS 400 (Remote Sensing)		CSCI 243 (Discrete Structures of Computer Science)
GIS 405 (Geovis. & Spatial Design)	*Examples:*	
GIS 420 (Advanced GIS)	DATA 440 (Data Driven Journalism)	CSCI 303 (Algorithms)
	GOVT 391 (Topics in Government)	
GOVT 391 (GIS for Social Science)	LING 380 (Computational Methods in Language Science)	
BIOL 445 (GIS for Biologists)	Data Science for Humanities	

Senior Capstone (4 credit hours)

DATA 4XX - Senior Capstone; COLL 400: This is a new course. We would offer three sections within an academic year, for students whose primary interests are in ALV, NQR, and CSI. These would be seminar-style courses where the focus is on each student developing a capstone project or portfolio that applies their Data Science skills to a substantive problem of their choice. This would be taught as a hybrid course that integrates elements of a seminar on a particular topic an honors colloquium, and a studio art senior capstone experience. To fulfill the COLL 400 requirement, students would present their research as either a presentation, poster, or video, to a diverse audience.

Trajectory for a B.S. in Data Science



1. In the first three semesters, a student should take DATA 141, DATA 146, and Calculus I & II, or MATH 211 if Calculus II was completed prior to enrolling at William & Mary. Ideally, a student would also take either/both the deliberation (currently DATA 150 or PHIL 215, 303, 312 or 3XX, a new course in Data Ethics) requirement or the visualization requirement (DATA 2XX).

2. In semesters four to six, a student will take DATA 301, DATA 3XX (Databases and Structures), complete the deliberation and visualization requirements, MATH 351, MATH352, and at least two of their three required Substantive Specialty Courses.

3. In a student's final year, they would complete any missing requirements (due to study abroad, etc.) and take their capstone seminar.