Proposal for a Departmental Minor in Data Science

Proposal
This is a proposal for a six course Minor in Data Science, offered by the Department of Statistics. This Minor would be available to statistics majors and minors as well as those majoring in other WCAS departments and programs.

Rationale
Data science is an appellation usually applied to an assortment of computational and applied statistical techniques that have attained increasing prominence as part of the “big data revolution.” Other terms sometimes used for data science are predictive analytics and machine learning (including artificial neural networks or “deep learning”). Data science reflects an important new direction for scholarship in statistics and its applications. There is currently no formal program in WCAS that will explicitly prepare students and (importantly) signal to graduate programs or the labor market that they are prepared in this field. An undergraduate minor in data science would provide a coherent program of academic preparation in the area and, along with academic preparation in a discipline, provide a foundation for graduate work in data science or in computational approaches to other fields (computational biology, computational social sciences, or even computational approaches to the physical sciences or the humanities.)

Although the primary rationale for this minor program is academic, there are great professional opportunities in data science for individuals with various different levels of academic preparation, including those with Bachelor’s degrees. For example, some of the students who have taken our data science courses have obtained enviable positions with handsome salaries in data science positions after graduation with only undergraduate degrees. Individuals with a background in data science are in high demand and the area is very appealing to many of our undergraduates, both those intending to major in statistics and those intending to major in other fields. An undergraduate minor in data science would provide a signal to prospective employers that the individuals are prepared for further training or employment in this area.

Requirements
The minor would consist of six courses: one introductory course, four required courses, and one elective. The introductory course is one 200 level statistics course, that is, one of STAT 202 (Introduction to Statistics), STAT 210 (Introduction to Statistics for the Social Sciences), or STAT 232 (Applied Statistics), and is prerequisite for the other courses. In addition, we require all of STAT 301-1, 301-2, and 301-3 (Data Science), and STAT 302 (Data Visualization). The three course sequence STAT 301-1, 301-2, and 301-3 provides a comprehensive introduction to computing for data science, data management, and the theory and application of a variety of data science techniques. We have purposely decided to utilize a single analytic programming language/software throughout the series so that students will become extremely proficient in one of Data Science’s core analytic programming languages. STAT 301-1, -2, & -3 includes rather extensive project work involving not only data acquisition, management, and analysis, but also communication of results. STAT 302 introduces students to the logic, grammar, and application of static and dynamic graphical procedures for statistics and data science. The additional elective would be selected with the approval of the Director of Undergraduate Studies to broaden the students understanding of statistics and data science. It could be a course offered by the department of statistics but, with permission of the
department, could be a relevant course in another department. Some possibilities for the elective are (but need not be limited to):

1. A STAT 300 (other than 301, 302) or 400 level course
2. A course in the student’s major field related to data science
3. EECS396: Statistical machine learning
4. EECS396: Data science seminar
5. EECS394: Machine Learning
6. IEMS308: Data science and analytics
7. IEMS351: Optimization methods in data science
8. IEMS365: Analytics for social good

We encourage other WCAS departments to recommend suitable courses to be added to the list.

We believe that this set of courses will provide an adequate introduction to students who may want to specialize in the area, seek further academic preparation in the area, or seek employment as bachelor’s degree level data scientists.

Who Will Be Interested in Obtaining the Minor in Data Science?
We believe that WCAS undergraduates with many majors may be interested in the proposed minor. The most obvious are students majoring in a related mathematical science (mathematics, MMSS, or economics). Because the big data revolution is deeply affecting many fields, we believe that students majoring in the social and even the physical sciences may also be interested, as well as students already majoring in statistics. The Department originally proposed a certificate that would be available to statistics majors and minors. We believe that some statistics majors may be interested in some formal recognition of their preparation in data science. Such students could complete the statistics major and the data science minor by taking 9 statistics courses (plus required related courses in math) to complete the major and 6 additional courses to complete the minor. Because both the major and the minor require an introductory course such as STAT 202, students pursuing both a major and a minor will be allowed to substitute an additional 300 level statistics course for the 200 level course required for the major. Similarly non-STAT majors who have taken STAT 202 or its equivalent mandated by their own major programs can take an additional elective course with the approval of DUS to fulfill the six-credit requirement of the data science minor (exceptions can be made according to WCAS double-counting policy).

We also believe that there will be interest from students with majors outside the physical or even the social sciences. Because data science as it now exists is more computational than deeply mathematical, the courses required by the proposed minor reflect that computational rather than mathematical focus. This means that they are more accessible to students outside of the more mathematical fields of study than are many of the courses required for the statistics major. Thus the minor will be accessible even to students majoring in the humanities who may be interested. Several students with majors in the humanities who have taken STAT 301 in the past tell us that they took the course to give themselves additional skills that might be useful in their careers.

Why Do We Think There is Demand for the Minor?
The original impetus for creating STAT 301-1, 301-2, and 301-3 was a committee of undergraduate students who came to us and asked for such a course. We began cautiously, teaching STAT 301 with a purposefully small enrollment (and without advertising) while we developed the courses. We believe that there will be substantial demand if we raise the profile of these courses and the entire
program. Note that the Department set the enrollment cap for STAT 301-1, 301-2, and 301-3 at 50 students and it has filled (even without the minor).

**Will this Minor Require the Creation and Staffing of New Courses?**
All of the courses required for the minor program have been created and previously taught. STAT 301 has been taught on a limited basis for the last two years (with limited enrollments largely to work out the design of the courses). STAT 302 was previously taught as a topics course which proved highly popular, drawing 40 students, and we made arrangements to offer it on a regular basis.

**Will this Minor Require Additional Faculty?**
Because it involves courses that are already being taught, it will not require new faculty if we keep the enrollment limited to one section of each required 300 level course. The recent addition of faculty members has made staffing of these courses easier. Kuyper, Hedges, W. Jiang and Liu are all possible instructors for these courses.

**Why is a New Minor Needed?**
Statistics already offers a minor, which focuses more broadly on the entire field of the theory and applications of statistics. The point of the Data Science Minor is to certify that students have competence in a specific area (data science) that is essentially a subset of statistics, not the entire field.

**Does This Minor Overlap with Other Programs at Northwestern?**
The minor most closely resembles the Master of Science in Analytics (MSIA) program offered by the McCormick School of Engineering and the related Master’s Degree in Predictive Analytics offered through the Northwestern School of Professional Studies. The minor differs from the MSIA in that the MSIA is a full-time, 15 month graduate program and the School of Professional Studies program is an online program of similar structure and intensity. These two programs differ from our Data Science Minor program in that they are intensive graduate programs with a highly vocational focus. While our minor program certainly provides skills that will make those who complete the minor attractive to potential employers, it is primarily an academic program, not a vocational program. We note that WCAS students are not allowed in the MSIA program. The most obvious overlap is with the statistics major and minor. However, we do not expect that the minor will reduce the number of statistics majors or minors. We expect that it may actually encourage more students to pursue the statistics major.

**Learning Goals**
The Data Science Minor proposed by the Department of Statistics is designed to offer comprehensive preparation in aspects of data science that will prepare students for employment, graduate study in data science or (when combined with appropriate disciplinary training) graduate study in data- or computation-intensive programs in science or social science.

When students complete the minor program they should be able to:
1. Given a dataset, select appropriate simple descriptive and inferential statistical techniques to describe univariate data, detect bivariate relations, and explain these analyses to non-experts.
2. Create datasets from multiple input files (matching data elements from different files as appropriate), automate creation of datasets (e.g., by web scraping or other search strategies), and
manipulate datasets to create alternative data structures, using a suitable programming environment such as R.

3a. Given a regression (numerical response) prediction problem and a corresponding dataset, they should be able to select appropriate statistical learning methods for supervised learning, apply those methods to the data, evaluate the models chosen, and select a preferred prediction model.

3b. They should also be able to explain, to both experts in data science and to individuals with substantive knowledge of the data but who are not experts in data science, their formulation of the problem, how they chose the methods they employed, why they chose their preferred model, how to use that model, and how well it might be expected to work.

4a. Given a classification (categorical response) prediction problem and a corresponding dataset, they should be able to select appropriate statistical learning methods for supervised learning, apply those methods to the data, evaluate the models chosen, and select a preferred prediction model.

4b. They should also be able to explain, to both experts in data science and to individuals with substantive knowledge of the data but who are not experts in data science, their formulation of the problem, how they chose the methods they employed, why they chose their preferred model, how to use that model, and how well it might be expected to work.

5. Given a statistical model, select several alternative graphical representations of the relationships embodied in that model, use a suitable graphical environment (such as R, ggplot, or shiny) to build static or dynamic representations of those relationships, and explain those graphical representations to individuals with substantive knowledge of the data who are not experts in data science.

Assessment of Whether the Students Meet These Learning Goals

Each of the required courses in this minor emphasizes project-based learning where students learn by carrying out data science projects. Each course has articulated learning objectives and a plan to assess those learning objectives through student work. Thus the assessment of whether the minor program meets its learning goals will draw heavily on the assessment in the classes that make up the minor.

Learning goal 1 will largely be evaluated through assignments in the 200 level statistics course students will take as part of the program. Learning goal 2 will largely be evaluated through assignments in STAT 301-1. Learning goals 3a and 4a will largely be evaluated through class assignments in STAT 301-2. Learning goals 3b and 4b will primarily be evaluated through class assignments in STAT 301-3, which emphasize larger projects and communication of results. Learning goal 5 will primarily be evaluated through classroom assignments in STAT 302.