Proposal for a Departmental Major in Data Science

Proposal Summary

This is a proposal for a 19-course major in Data Science offered by the Department of Statistics. This major will be available for all students at Northwestern who want to major in Data Science (including those who want to take Data Science as a second major).

Background and Rationale

Data Science is a newly emerged study program for integrated methods used to deal with complex and massive data encountered in every field nowadays. Data science concerns principles and methods applied in a broad array of activities spanning data collection, storage, integration, analysis, inference, reasoning, reporting and ethics. The breadth of these activities defines the interdisciplinary nature of data science programs, which typically entails statistics and computer science in the core curriculum, while embracing other domain sciences.

In the past two or three years, many data science programs have been established in peer institutions, mostly at the undergraduate (major or minor) or master’s level, with a few having PhD specialization. Currently at Northwestern, there is no Data Science major program. The Department of Statistics opened a Data Science minor program in the fall of 2019. Since its inception, the program has been exceedingly successful with an increase of enrollment from 76 in the fall of 2019 to 222 in the fall of 2020. The top six programs where minors come from are Economics (50), Statistics (39), Computer Science (35), Journalism (16), Mathematics (15) and Industrial Engineering (12). The demography of Data Science minors suggests that data science has become an increasingly popular direction to study across disciplines, while differing from any individual pre-existing program. Data scientists are in high demand in today’s job market. For past a few years, data scientist has been ranked very top in the list of 50 best jobs in America by Glassdoor. Thus we expect the Data Science major will become an equally popular program as the minor.

Our Data Science minor program was created to provide training in data science to students with an emphasis on data analytics complementary to their major focus. The major program in data science though is designed to train more specialized data scientists through a more rigorous and comprehensive curriculum that integrates math, statistics, computer science and other domain science courses. Data Science major graduates will possess strong understanding and skills in data management and integration, data structures and algorithms, data analytics and machine learning methods, data visualization and reporting, as well as solid foundation in math, statistics and computer science necessary in preparation for graduate study in data science or related fields.

Major Requirements

The proposed major in Data Science requires a curriculum of 8 related courses, 11 major courses and as tabulated in Appendix I.

The 8 related courses include 4 required courses in math, 1 in computer science, 2 electives from the approved list of technical electives or domain science courses, and 1 from a recommended list of ethics courses. The required math courses are: Single-Variable Differential and Integral Calculus (MATH220-1, 2), Multi-Variable Differential Calculus (MATH230-1), and Linear algebra
(MATH240), which will provide students with solid math training necessary for the proposed Data Science major curriculum. The Data Science major also requires COMP_SCI111 (Fundamentals of Computer Programming) to provide an introduction of computer programming which is essential to understand data structures and algorithms and database systems. The two additional electives may be chosen from an approved list of advanced technical electives or related domain science courses (see a tentative list in Appendix II). We welcome different programs to recommend suitable courses to be included into this list. These electives will strengthen students’ technical ability, expand their knowledge breadth in data science or engage students to apply their data science skills to a specialized area. Students in Data Science major are also required to take at least one course in ethics. A recommended list of the ethics courses can be found in Appendix III. Other courses may be substituted subject to the approval of the Department of Statistics Director of Undergraduate Studies.

The 11 required courses consist of 4 foundational courses, 6 data science core courses, and one data science capstone course, among which 4 will be new courses (indicated with * below). The 4 foundational courses include an intro-level statistics course (STAT202/210/232), Statistical Theory and Methods Series (STAT320-1, 2) and Topics in Mathematics for Data Science (*STAT228). The three statistics courses will provide an introduction to different statistical procedures, statistical inference methods and their underlying theory. STAT228 will provide further math training on important topics required by core courses including sequence and series, multi-variable integral calculus, and discrete mathematics.

The 6 data science core courses include the existing STAT301 Data Science Series (STAT301-1, 2, 3) and three new courses: Data Structures and Algorithms for Data Science (*STAT303), Information Management for Data Science (*STAT304) and Advanced Machine Learning for Data Science (*STAT336). The STAT301 Series provide a comprehensive introduction to computing for data science, data management, and the theory and application of a variety of data science techniques. *STAT336 will go further to cover advanced machine learning topics, which may include large-scale machine learning, distributed optimization algorithms, Bayesian machine learning, probabilistic graphical models, variational inference, probabilistic programming and etc. *STAT303 will cover abstract data type, data structures and algorithms, whereas *STAT304 will cover how to manage and process data through relational database. Together these core courses will provide a thorough training of different skills needed for a data scientist from data integration and management, computer algorithm design, data manipulation, to essential data analytics and machine learning methods.

The Data Science Project course (*STAT390, currently STAT359) provides a capstone experience to students who will apply their data science knowledge to a real-world problem.

**Learning Objectives**

The Data Science major program provides an integral training with a curriculum that comprises comprehensive courses from statistics, computer science and mathematics as well as a capstone experience specially crafted for students who aim for a career path or for advanced study in the rapidly growing data science field. Students will receive rigorous training in theory and methods and gain hands-on experience in computer programming and algorithms, relational database and advanced statistical and machine learning methods in data science. Students who complete the Data Science major will be able to:

1. Understand statistical theory and principles of statistical inference and apply them to data science.
2. Design efficient computer algorithms using tools including SQL language, R and/or Python programming.
3. Work independently and on teams to design and conduct analysis of complex, large data sets using statistical and machine learning methods to answer questions and solve problems in real-world contexts.
4. Communicate the results of data analysis in effective writing, presentation, and visualization.

Assessment plan

Students will be assessed whether they meet these learning goals as follows:

- Objective 1 can be assessed by examining students’ performance in homework assignments and exams in 202, 320, and their competence in conducting statistical inference and interpreting results in 301, 336 and 390. In 202 students will be assessed whether they can apply different statistical procedures in data analysis and interpret the results. In 320 series, they will be assessed whether they can understand the theory for justification of the optimal choice of different statistical inference methods. In 301, 336 and 390 they will be further assessed in this regard in the analysis of large scale data.

- Objective 2 can be assessed by examining students’ programing proficiency in 301, 304, 305, 336 and 390. In 301, 304 and 305 students’ programming skills will be assessed in a more content-specific way, i.e., whether they can write efficient codes (R/Python or SQL) to realize different types of procedures or analyses covered in the lectures. In 336 and 390, students will be assessed for their overall skills in programming to integrate knowledges learned from 301, 304 and 305 to design and conduct efficient and complex analyses in data science.

- Objectives 3 and 4 will be assessed by examining students’ performance in independent or group projects in 301, 336, and 390. In these courses, students will be required to complete projects of different scales either independently or in a team setting. They will be assessed for abilities from design of the project, data download and processing, proposing project aims or hypothesis, analysis of the data to presentation of the results (particularly in the capstone course).

- In addition to in-class assessment, the Department will indirectly assess these learning objectives in multiple ways. We will conduct student exit survey with articulated questions for their feedback about their accomplishments in these learning objectives. Each year, we will survey faculty on student learning and performance in the class. We will also frequently survey Data Science alumni in the future for their input on how the learning objectives help them succeed in their career and how the curriculum and learning objectives can be improved.
Appendix I: Data Science major Curriculum

Related Courses (8 units, required):

- Math : (4 units)
  1. MATH220-1: Single-Variable Differential Calculus
  2. MATH220-2: Single-Variable Integral Calculus
  3. MATH230-1: Multivariable Differential Calculus
  4. MATH240: Linear algebra
(Note: 1+2 or 3 or 4 can be replaced by their equivalent courses)

- Computer Science (1 unit)
  5. COMP_SCI111: Fundamentals of Computer Programming

- Technical electives and domain science electives (2 units, see Appendix II)
- Ethics (1 unit, see Appendix III for a list of recommended courses)

Required Major Courses (11 units):

- Foundational courses (4 units)
  1. STAT202/210/232: Intro to Statistics
  2. *STAT228: Topics in Mathematics for Data Science
  3. STAT320-1: Statistical Theory and Methods I
  4. STAT320-2: Statistical Theory and Methods II
(Note: students who haven’t taken MATH226 + MATH 230-2 must take *STAT228 before or when concurrently take STAT320-1)

- Data science core courses (6 units)
  5. *STAT303: Data Structures and Algorithms for Data Science
  6. *STAT304: Information Management for Data Science
  7. STAT301-1: Data Science I
  8. STAT301-2: Data Science II
  9. STAT301-3: Data Science III
  10. *STAT336: Advanced Machine Learning for Data Science

- Capstone experience course (1 unit)
  11. *STAT390: Data Science Project (currently STAT359)

(Note: *STAT303, 304 can be replaced by COMP_SCI214 and COMP_SCI217 respectively.)
Appendix II : Tentative List of Advanced Technical Electives and Domain Applications
(subject to further revision)

Advanced Technical Electives

**Computer Science**
COMP_SCI 325-1 Artificial Intelligence Programming
COMP_SCI 336 Design & Analysis of Algorithms
COMP_SCI 339 Intro to Databases
COMP_SCI 348 Intro to Artificial Intelligence
COMP_SCI 352 Machine Perception of Music & Audio
COMP_SCI 376 Game Design and Development
COMP_SCI 377 Game Development Studio

**Economics**
ECON380-1,2 Game theory,
ECON381-1,2 Econometrics
ECON382: Applied Econometrics

**Engineering Sciences and Applied Mathematics**
ES_APPM 346 Modeling and Computation in Science & Engineering
ES_APPM 370 Introduction to Computational Neuroscience

**Industrial Engineering and Management Sciences**
IEMS 336 Stochastic models
IEMS 341 Social network
IEMS 351 Optimization Methods In Data Science
IEMS 313 Foundation of optimization

**Mathematics**
MATH 308-0 Graph Theory
MATH 306-0 Combinatorics & Discrete Mathematics
MATH 366-0 Mathematical Models in Finance
MATH 390-2,3 Probability and Stochastic Processes

**Statistics**
STAT 302 Data Visualization
STAT 320-3 Statistical Theory and Methods III
STAT 328 Causal Inference
STAT 342 Statistical Data Mining
STAT 344 Statistical Computing
STAT 348 Applied Multivariate Analysis
STAT 350 Regression Analysis
STAT 351 Design and Analysis of Experiments
STAT 352 Nonparametric Statistical Methods
STAT 354 Applied Time Series Modeling and Forecasting
STAT 356 Hierarchical Linear Models
STAT 365 Introduction to the Analysis of Financial Data
STAT 357 Elementary Bayesian Statistics (currently 359)
Domain Science Electives

Anthropology
ANTHRO 362 Advanced Methods in Quantitative Analysis
ANTHRO 324 Archaeological Survey Methods
ANTHRO 322 Introduction to Archaeology Research Design and Methods
ANTHRO 389 Ethnographic Methods and Analysis

Biomedical Engineering
BMD_ENG 311 Computational Genomics

Chemical and Biological Engineering
CHEM_ENG 379 Computational Biology: Principles & Applications

Engineering Sciences and Applied Mathematics
ES_APPM 375-1,2 Quantitative Biology I: Experiments, Data, Models, and Analysis

Environmental sciences:
(These courses are also cross-listed in other departments EARTH)
ENVR_SCI 390-0-01 GIS Level 1
ENVR_SCI 390-0-02 GIS Level 2
ENVR_SCI 390-0-03 – Topics: R Data Science

Earth and Planetary Science
EARTH 327 Geophysical Time Series Analysis
EARTH 353 Mathematical Inverse Methods in Earth and Environmental Sciences
EARTH 362 Data Analysis for Earth and Planetary Sciences

Global Health Studies
GBL_HLTH 320 Qualitative Research Methods in Global Health

Journalism
JOUR 377 Data Analysis and Visualization
JOUR 390 Special Topics (Data Journalism)

Linguistics
LING 334 Introduction to Computational Linguistics
LING 336 Words, Networks and the Internet
LING 330 Research Methods in Linguistics

Molecular BioSciences
BIOL_SCI 323 Bioinformatics: Sequence and Structure Analysis
BIOL_SCI 341 Population Genetics

Political Science:
POLI_SCI 390 Methods of Political Inference
POLI_SCI 312 Statistical Research Methods
Psychology
PSYCH 380 Advanced Statistics and Experimental Design
PSYCH 387 Consumer Psychology and Marketing Research

SEPS
SOC_POL 330 Economics of Social Policy
SOC_POL 351 Special Topics (Quantitative Tools)

Sociology
SOCIOL 303 Analysis and Interpretation of Social Data
SOCIOL 329 Field Research and Methods of Data Collection

Communication
COMM_ST 371 Cultural Analytics
Appendix III: List of Recommended Ethics Courses for the Major

AF_AM_ST 215-0 Introduction to Black Social & Political Life
AF_AM_ST 220-0 Civil Rights and Black Liberation
AF_AM_ST 319-0 Race, Ethnicity and the American Constitution
COMP_LIT 207-0 Introduction to Critical Theory
ENTREP 360-0 Leadership, Ethics, and You
GBL_HLTH 302-0 Global Bioethics
GBL_HLTH 324-0 Volunteerism and the Ethics of Help
GNDR_ST 233-0 Gender, Politics, and Philosophy
HUM 220-0 Health, Biomedicine, Culture, and Society
HUM 325-5 Humanities in the Digital Age
IMC 310-0 Integrated Marketing and Communication Law, Policy and Ethics
ISEN 230-0 Climate Change and Sustainability: Ethical Dimensions
JOUR 303-0 Framed: Media and the Marginalized
JOUR 370-0 Media Law & Ethics
LATINO 342-0 Latina and Latino Social Movements
LATINO 392-0 Topics in Latina and Latino Social and Political Issues
LING 222-0 Language, Politics, and Identity
PERF_ST 306-0 Performance and Race
PHIL 220-0 Introduction to Critical Theory
PHIL 221-0 Gender, Politics, & Philosophy
PHIL 224 Philosophy, Race and Racism
PHIL 240-0 Freedom and Responsibility
PHIL 262 Ethical Problems and Public Issues
PHIL 268 Ethics and the Environment
PHIL 269 Bioethics
PHIL 269-0 Bioethics
PHIL 270 Climate Change: Economic and Ethical Dimensions
PHIL 270-0 Climate Change and Sustainability: Economic and Ethical Dimensions
PHIL 273-2 The Brady Scholars Program: The Moral Life
PHIL 273-2 The Moral Life (note: part of a sequence)
PHIL 273-3 The Brady Scholars Program: The Good Society
PHIL 273-3 The Good Society (note: part of a sequence)
PHIL 363-0 Kant's Moral Theory
PHIL 364-0 Business and Professional Ethics
POLI_SCI 302-0 Subjects, Citizens, Revolutionaries: Early Modern Political Thought
POLI_SCI 303-0 Modernity and Its Discontents
POLI_SCI 304-0 Human Rights Between East and West
POLI_SCI 307-0 Deportation Law and Politics
POLI_SCI 309-0 Political Theories of the Rule of Law
POLI_SCI 347-0 Ethics in International Relations
POLI_SCI 382-0 Politics of Religious Diversity
RELIGION 373 Religion and Bioethics
SESP 303-0 Designing for Social Change
SLAVIC 222-0 Language, Politics, & Identity
SLAVIC 260 Economics and Humanities: Understanding Choice
SOCIOL 220-0 Health, Biomedicine, Culture, and Society
SOCIOL 321 Numbers, Identity and Modernity
SOCIOL/HUM 220 Health, Biomedicine, Culture and Society