PhD Program Overview
2022-2023
Welcome to the HDSI PhD Graduate Program!

The goal of the doctoral program is to create leaders in the field of Data Science who will lay the foundation and expand the boundaries of knowledge in the field. The doctoral program aims to provide a research-oriented education to students, teaching them knowledge, skills and awareness required to perform data driven research, and enabling them to, using this shared background, carry out research that expands the boundaries of knowledge in Data Science. The doctoral program spans from foundational aspects, including computational methods, machine learning, mathematical models and statistical analysis, to applications in data science.

**Research Rotation Program**

Research rotations provide the opportunity for first-year PhD students to obtain research experience in data analysis under the guidance of HDSI affiliated faculty members. Through the rotations, students can identify a faculty member under whose supervision their dissertation research will be completed.

A research rotation is a guided research experience lasting one quarter (10 weeks) obtained by registering for DSC 294 (4 units) with an instructor. All PhD students will participate in a *minimum of 2 research rotations during their first year*, and with a minimum of two different faculty members, and as much as three rotations including summer quarter. A student may rotate twice under the same faculty member as long as they rotate with at least two faculty members. The goal is to expose students to new methodological approaches or domain knowledge and help them find an advisor for their PhD research.

Please refer to the HDSI Faculty page for a list of HDSI faculty members and their research interests. Students should reach out to faculty members to learn more about potential rotation opportunities.

An important goal of HDSI is to foster interdisciplinary collaboration. Specifically, collaboration between method researchers and domain researchers. Domain researchers specialize in a particular scientific field and collect or generate large datasets. Domain experts can partner with HDSI faculty and jointly advise a PhD. student. The site matchpoint.com was designed to facilitate the creation of such collaborations.

Outcome of each rotation program: A written report, or curated datasets with jupyter notebooks. Students are expected to work 10 hours per week.

Research rotations must be completed by the end of the Fall Quarter of the second year with a signed commitment form from a faculty advisor. Those who fail to identify a research advisor shall be advised to leave the doctoral program with an optional assessment for completion of a terminal MS-DS degree.
Course Requirements – General Overview

There are three categories of courses: Foundation (group A), Core (group B), and Elective and Research requirements (group C) for the graduate program. These course requirements are intended to ensure that students are exposed to (1) fundamental concepts and tools (Foundation), (2) advanced, up-to-date views in topics central to Data Science for all students (the Core requirement), and (3) a deep, current view of their research or application are (the Elective requirement). Courses may not fulfill more than one requirement.

The doctoral program is structured as a total of 52 units in courses from these group A, B, and C as described in detail below. Out of the 52 units, 48 units (or 12 courses) must be taken for letter grade and at least 40 units must be using graduate-level courses.

The remaining 4 (= 52 – 48) units are for professional preparation, consisting of 1 unit of faculty research seminar, 2 units of TA/tutor training and 1 unit of survival skills course taken for a passing (satisfactory) grade. Finally, as mentioned earlier, out of the 12 regular courses, at most two can be upper-level undergraduate courses. 36 units or 9 courses must be completed within six quarters from the start of the degree program.

Course Requirements – Detailed Overview

Group A courses are introductory level graduate courses in the foundational areas of data science. Group B are core graduate level courses with prerequisites from Group A courses. Group C are advanced, specialized and free-standing courses, often part of the required courses in the Data Science specialization of the Graduate Program in other departments. In all three groups, required courses are indicated as such; they can not be substituted by other courses without exception approval from the graduate program committee.

Group A: Preparatory Courses

There are five important knowledge and skills necessary for understanding (and advancing) core data science. It is, therefore, important that all our entering students either have background preparation or have courses available in the program to ensure a successful completion of the stipulated doctoral degree program. A student can receive credit towards the PhD degree for a maximum of three courses from the list of courses below:

1. DSC 200: Data Science Programming.
2. DSC 202: Data Management for Data Science
3. DSC 210: Numerical Linear Algebra
4. DSC 211: Introduction to Optimization
5. DSC 212: Probability and Statistics for Data Science
**Group B: Core Courses**

Four core courses are required for all PhD students, including those with a Bachelors in Data Science. The four required courses are:

1. DSC 240: Machine Learning
2. DSC 260: Data Ethics and Fairness
3. (*)DSC 241: Statistical Models (or MATH 282B)
4. (*)DSC 204A: Scalable Data Systems (or CSE 202);
   or DSC 206: Algorithms for Data Science

(* Depending on academic preparation, a PhD student can take an advanced course on Applied Statistics, such as MATH 282B instead of DSC 241. Similarly, instead of DSC204A, a student can take a course on Algorithms, such as DSC 206: Algorithms for Data Science, or CSE 202: Design and Analysis of Algorithms.

In addition, a doctoral student must select at least 2 out of the following 8 core courses

1. DSC 203: Data Visualization and Scalable Visual Analytics
2. DSC 204B: Big Data Analytics and Applications
3. DSC 215: Statistical critical thinking
4. DSC 242: High-dimensional Probability and Statistics
5. DSC 243: Advanced Optimization
6. DSC 244: Large-Scale Statistical Analysis
7. DSC 245: Introduction to Causal Inference
8. DSC 250: Advanced Data Mining
9. DSC 261: Responsible Data Science

Thus, doctoral students are required to take a minimum of 6 courses for letter-grade credit from Group B courses. Students can take more than 6 courses from this group to satisfy letter grade course requirements except (satisfactory completion of professional preparation) teaching, survival skills and research seminar courses. Students who satisfy all letter-grade course requirements are expected to enroll into individual research (DSC 298) in a section offered by the faculty advisor to meet residency requirements and maintain graduate student standing during the period of dissertation research.
Group C: Professional Preparation and Elective Courses

Group C courses aim to provide either practical experiences in chosen specialization areas, or advanced training for students preparing for doctoral programs. The courses include required professional preparation courses: 2 unit TA/tutor training (DSC 599), 1 unit of academic survival skills (DSC 295) and 1 unit faculty research seminar (DSC 293), all of which must be completed with a Satisfactory (S) grade using the S/U option.

Professional Preparation Courses

1. DSC 599: TA/Tutor Training
2. DSC 293: Faculty Research Seminar
3. DSC 294: Research Rotation
4. DSC 295: Academia Survival Skills

General elective and Specialization Courses

Courses here aim to provide advanced training for students in the doctoral programs, or practical experiences in chosen specialization areas. Students can choose from the following elective or specialization tracks. Additional elective courses will be offered based on faculty interest and availability. In addition to the elective courses listed below, these include advanced courses in core Data Science subjects listed under Group B above (not counted towards the min-6 group B course requirements), or offered as research topics (DSC 291), or they can be graduate courses in other departments subject to approval by the student’s HDSI academic advisor.

Elective Course List:
CSE 234, MATH 181 A-B-C, MATH 284, MATH 285, MATH 287A-B, COGS 243.

Preliminary Assessment Examination

The goal of the preliminary assessment examination is to assess students’ preparation for pursuing a PhD in data science, in terms of core knowledge and readiness for conducting research. The preliminary assessment is an advisory examination.

There is an oral presentation that needs to be completed before the end of Fall quarter of the second year. The student will propose the topic of the presentation (e.g., the outcome of a research rotation or a literature survey). The Graduate Program Committee will set up a committee consisting of two members (students can suggest committee members but there is no
guarantee that only those suggested faculty members will be chosen for the committee. The oral presentation from the student will be followed by a Q&A session by the committee members. The committee will assess both the oral presentation as well as the students academic performance so far (especially in the required core courses). Students who did not get a satisfactory evaluation will receive a recommendation from the Graduate Program Committee regarding ways to remedy the lacking preparation (e.g., suggestion of courses to be taken), or an opportunity to receive a terminal MS in Data Science degree provided the student can meet the degree requirements of the MS program.

**Research Qualifying Examination and Advancing to Candidacy**

A research qualifying examination (UQE) is conducted by the dissertation committee consisting of four or more members approved by the graduate division as per senate regulation 715(D). One senate faculty member must have a primary appointment in the department outside of HDSI. Faculty with 25% or less partial appointment in HDSI may be considered for meeting this requirement on an exceptional basis upon approval from the graduate division.

The goal of UQE is to assess the ability of the candidate to perform independent critical research as evidenced by a presentation and writing a technical report at the level of a peer-reviewed journal or conference publication. The examination is taken after the student and his or her adviser have identified a topic for the dissertation and an initial demonstration of feasible progress has been made. The candidate is expected to describe his or her accomplishments to date as well as future work. The research qualifying examination must be completed no later than fourth year or 12 quarters from the start of the degree program; the UQE is tantamount to the advancement to PhD candidacy exam.

A petition to the Graduate Committee is required for students who take UQE after the required 12 quarters deadline. Students who fail the research qualifying examination may file a petition to retake it; if the petition is approved, they will be allowed to retake it one (and only one) more time. Students who fail UQE may also petition to transition to a MS in Data Science track.

**Dissertation Defense Examination and Thesis Requirements**

Students must successfully complete a final dissertation defense oral presentation and examination to the Dissertation Committee consisting of four or more members approved by the graduate division as per senate regulation 715(D). The primary Thesis Adviser, who will chair the Dissertation Committee, must be a senate faculty member with an appointment of 0% or more at HDSI. One senate faculty member in the Dissertation Committee must have a primary appointment in a department outside of HDSI. Partially appointed faculty in HDSI (at 25% or less) are acceptable in meeting this outside-department requirement as long as their main (lead) department is not HDSI.

A dissertation in the scope of Data Science is required of every candidate for the PhD degree. HDSI PhD program thesis requirements must meet Regulation 715(D) requirements. The final
form of the dissertation document must comply with published guidelines by the Graduate Division.

The dissertation topic will be selected by the student, under the advice and guidance of Thesis Adviser and the Dissertation Committee. The dissertation must contain an original contribution of quality that would be acceptable for publication in the academic literature that either extends the theory or methodology of data science, or uses data science methods to solve a scientific problem in applied disciplines.

The entire dissertation committee will conduct a final oral examination, which will deal primarily with questions arising out of the relationship of the dissertation to the field of Data Science. The final examination will be conducted in two parts. The first part consists of a presentation by the candidate followed by a brief period of questions pertaining to the presentation; this part of the examination is open to the public. The second part of the examination will immediately follow the first part; this is a closed session between the student and the committee and will consist of a period of questioning by the committee members.

**Special Requirements: Generalization, Reproducibility and Responsibility**
A candidate for doctoral degree in data science is expected to demonstrate evidence of generalization skills as well as evidence of reproducibility in research results. Evidence of generalization skills may be in the form of — but not limited to — generalization of results arrived at across domains, or across applications within a domain, generalization of applicability of method(s) proposed, or generalization of thesis conclusions rooted in formal or mathematical proof or quantitative reasoning supported by robust statistical measures. Reproducibility requirement may be satisfied by additional supplementary material consisting of code and data repository. The dissertation will also be reviewed for responsible use of data.

**Special Requirements: Professional Training and Communications**
All graduate students in the doctoral program are required to complete at least one quarter of experience in the classroom as teaching assistants regardless of their eventual career goals. Effective communications and ability to explain deep technical subjects is considered a key measure of a well-rounded doctoral education. Thus, PhD students are also required to take a 1-unit DSC 295 (Academia Survival Skills) course for a Satisfactory grade.

**Obtaining an MS in Data Science**
PhD students may obtain an MS Degree in Data Science along the way or a terminal MS degree, provided they complete the requirements for the MS degree.