Data Science in the Humanities

A New Graduate Certificate

Much of the DH work carried out at Washington University under the sponsorship of the Humanities Digital Workshop (HDW) has focused on the analysis of large data sets. We now offer a graduate certificate in computational approaches to vital research in the humanities.

The curriculum addresses data management, statistics, text analysis, geospatial analysis, digital prosopography, and data visualization and information design. The certificate entails experience in digital project work, and features appreciable cross-disciplinary engagement. Our goal is to enrich the analytic skills that students can bring to bear on research in their home disciplines, *and* to enable them to contribute thoughtfully and resourcefully to research in other disciplines of the humanities.

Requirements for the certificate (15-units)

6 units from the 9-unit DASH Curriculum (See verso. Individual departments may determine that one or more of these courses might count as basic methods courses allocable toward satisfaction of home-disciplinary requirements.)

3 units of work on a faculty project in the HDW. Most students will earn these units by participating in the HDW summer workshop.

3 units of TA-ship in either DASH 1, DASH 2, DAMS + PROTA, or a new 300-level Introduction to Digital Humanities

3 units from the following:

CSE 104: Web Development (taken at the 400-level by special arrangement)
Art 437B: Information Design
a 400-level Digital Humanities course, home-based in a Humanities dept. (Such a course could count both towards the certificate and toward satisfaction of home-disciplinary requirements.)
3 more units from the DASH Curriculum
Independent project work substantively contributing to the student's dissertation

Roll-out

DASH 1 and DAMS will be offered in the Fall 2016; PROTA will be offered in Summer 2017.

For more information, contact Joseph F. Loewenstein, (jfloewen@wustl.edu)
Director, Humanities Digital Workshop

The DASH Curriculum

1) *DASH 1:* ***Statistics for Humanities Scholars*** (3 units)

A survey of statistical ideas and principles. The course will expose students to tools and techniques useful for quantitative research in the humanities, many of which will be addressed more extensively in other courses: tools for text-processing and information extraction, natural language processing techniques, clustering & classification, and graphics. The course will consider how to use qualitative data and media as input for modeling and will address the use of statistics and data visualization in academic and public discourse. By the end of the course students should be able to evaluate statistical arguments and visualizations in the humanities with appropriate appreciation and skepticism.

Details. Core topics include sampling, experimentation, chance phenomena, distributions, exploration of data, measures of central tendency and variability, and methods of statistical testing and inference. In the early weeks, students will develop some facility in the use of Excel; thereafter, students will learn how to use Python or R for statistical analyses.

2) *DAMS:* ***Data management skills for the humanities*** (1 unit, usually to be offered in the same semester as DASH 1)

The course will present basic data modeling concepts and will focus on their application to data clean-up and organization (text markup, Excel, and SQL). Aiming to give humanities students the tools they will need to assemble and manage large data sets relevant to their research, the course will teach fundamental skills in programming relevant to data management (using Python); it will also teach database design and querying (SQL).

Details. The course will cover a number of "basics": the difference between word processing files, plain text files, and structured XML; best practices for version control and software "hygiene"; methods for cleaning up data; regular expressions (and similar tools built into most word processors). It will proceed to data modeling: lists (Excel, Python); identifiers/keys and values (Excel, Python, SQL); tables/relations (SQL and/or data frames); joins (problem in Excel, solution in SQL, or data frames); hierarchies (problem in SQL/databases, solution in XML); and network graph structures (nodes and edges in CSV). It will entail basic scripting in Python, concentrating on using scripts to get data from the web, and the mastery of string handling.

3) *PROTA:* ***Programming for Text Analysis*** (2 units, offered as an independent sequel to DAMS. We will make an effort to schedule this during most summer terms.)

This course will cover the core data-scientific concepts required for analyzing large corpora of texts and will introduce basic programming together with text-analysis techniques relevant to the humanities. (There will be very slight overlap with the programming instruction in the statistics and data-management courses.)

Details. Students will learn to calculate basic corpus-statistics, and will develop facility with such techniques as tokenization, chunking, extraction of thematically significant words, stylometrics and authorship attribution. Later in the course, more advanced topics from natural language processing such as stemming, lemmatization, named-entity recognition, part-of-speech tagging will be introduced along with a survey of text-classification terminology.

4) *DASH 2:* ***Advanced Data Science for the Humanities*** (3 units; prerequisite, either DASH 1, DAMS, or PROTA)

This course will offer a broad survey of advanced data-analysis techniques widely used in digital humanities scholarship. It will present basic data-mining and machine-learning terminology and techniques, an overview of network analysis and visualization, and spatial analysis. Designed for students with some familiarity with programming, text-analysis, and statistics, the course will look at a wide range of information analysis, visualization, and, perhaps, sonification techniques in the context of qualitative humanistic data. Specific techniques and algorithms that are widely used in digital humanities literature such as principal component analysis, topic-modeling, and the use of force-directed networks will be covered in detail. The focus of the course will not be on a rigorous understanding of the mathematical foundations of these techniques but a broader survey that will allow students to engage critically with scholarship in the field and also to have a clear sense of what approaches might be applicable to their own work.

Details. As a pre-requisite, students should take one of the three courses listed above (in statistics, data management, or text analysis). Topics will include vector-spaces, data-mining and pattern identification using clustering and classification, cross-validation, the extraction and analysis of relationships with networks and basic graph-theoretic techniques, and a survey of spatial thinking and computational modeling of geospatial data in the humanities. Attention will be given to techniques linking the results of analyses to other resources, e.g. transforming recognized name-entities into triples, and mapping to shared, unified ontology schema. Other topics may be added.