Proposal: Create a Graduate Certificate in Data Science

The proposal is to create a new professional Graduate Certificate degree, consisting of four new courses as described below. The program would be administered through OES.

Expected background/Admission criteria: Applicants are expected to have:
- an undergraduate degree (engineering/computer science preferred, but not required)
- prior quantitative coursework (i.e., calculus, linear algebra, etc.)
- prior introductory to computer programming coursework (i.e. Python, Java, C++, etc.) or significant programming/systems administration experience

Logistics: Each course will be offered as about 13-14 2.5 hour classes. One of the classes could be final exam. The take-home work would need to be minimal given the target audience. The goal is to start in Fall 2016, with two of the courses offered then, and two offered in Spring 2017. Some of the administrative stuff would be handled by OES; we will need to hire a coordinator to handle the remaining stuff (e.g., admissions, advertising).

Who will teach these: For the first 1 or 2 offerings, tenure-track faculty would be expected to teach these in order to develop the courses. After that, at least some of the teaching is expected to be done by adjuncts (e.g., folks from the industry, our ex-students, etc).

Course 1: Introduction to Data Science
Goals: Broad overview of what data science means, some cases studies, some topics like data cleaning, information visualization, that are not covered elsewhere.

Tentative syllabus (assuming 2.5 hour classes, with approximately 13 total)
- Introduction (1-1.5 lectures)
  - What is Data Science? Who is a data scientist? The scope of Data Science: the open data movement, science, business, government, education, sport
  - Typical workflow of a data scientist / life-cycle of data
  - Best practices: organizing projects, managing collaborations and expectations
  - Primer on scripting language (Python or R) used for labs
- Computational and Statistical Data Analysis (2 lectures)
  - Basic Statistics, Probability Distributions, Hypothesis Testing, Fallacies
  - Labs on the basic statistics
- Some basic ML/Stats algorithms (2 lectures)
  - Linear regression, clustering, classification
- Data preparation (4 lectures)
  - Issues with data: data quality, metrics, missing data
  - Combining with data publicly available datasets
  - Data scraping, ETL, Data wrangling/cleaning approaches
  - Tools: Unix-based tools, Openrefine, R/Python-based tools
- Communicating with Data (3 lectures)
  - Writing with and about data: communicating the result of a data analysis
  - Visualization III: information visualization, visualizing for an audience
Putting it together: interactive data dissemination on the web (d3.js and related technologies)

Case Studies (3 lectures)
- Applications in Social Sciences, Business, etc.

Course 2: Big Data Systems; Cloud Computing
Overview of data management systems, how to use the different categories of systems, when to use them, what are the main pros and cons, best practices for data modeling.

Tentative syllabus (assuming 2.5 hour classes, with approximately 13 total)
- Background overview: structure of a modern OS, Cloud computing, Data Centers, Performance issues in cloud computing (2 lectures)
- Data models and importance of modeling, Normal Forms (1 lecture)
- Relational databases + SQL (3 lectures)
- Parallel Databases (1 lecture)
- Map-Reduce Framework (4 lectures)
  - Fundamentals
  - Writing different algorithms in MR
  - Hadoop + Spark
- NoSQL Systems (3 lectures)
  - Basics, How and where to use them
  - Key-value stores: Cassandra, HBase
  - Document stores: MongoDB, Couchbase
  - Graph Databases: Neo4j, OrientDB, AllegroGraph
- Batch graph analytics systems (1 lecture)
  - GraphX, Giraph, GraphLab
- Data streaming systems (1 lecture)
  - Storm, Spark Streaming

Course 3: Machine Learning and Data Mining
Overview of topics in Machine Learning and Data Mining, also covering topics in how to do those at scale.

Tentative syllabus (assuming 2.5 hour classes, with approximately 13 total)
- Introduction and roadmap; Primer on software package (1 lecture)
- Linear models: More detailed treatment of regression, least-squares (2 lectures)
- Classification, Decision trees (1 lecture)
- Support Vector Machines; Kernel methods (1 lectures)
- Clustering, K-Means (2 lectures)
- Bayesian networks, graphical models, Hidden Markov Models (2 lectures)
- Dimensionality Reduction Methods (2 lectures)
- Association Rule Mining (1 lecture)
- Recommendation Systems (1 lecture)
- Mining Social Networks (1 lecture)
- Machine learning at scale (2 lectures)
Packages like GraphLab, MLib, etc.

Note: Labs interspersed with all the topics

Course 4: Algorithms for Data Science

Broad overview of algorithmic techniques useful in Data Science.

Tentative syllabus (assuming 2.5 hour classes, with approximately 13 total)

- Random sampling (2 lectures)
  - Markov/Chernoff bounds, estimating sample sizes
  - Different techniques for sampling, reservoir sampling etc.
- Graph Algorithms and Network Science (4-5 lectures)
  - Shortest paths, matching, flows, etc.
  - Community detection, centrality measures, other network science-y stuff
- Data streams (4-5 lectures)
  - Approximate selection, Frequency moments, Clustering
  - Incremental algorithms for some categories of problems
- Linear and convex programming (2 lectures)
- Basics of Matrix Algebra (2 lectures)
  - Eigenvalues, singular values, PCA, etc.
- Optimization algorithms (2 lectures)
  - Gradient descent, Stochastic gradient descent, etc.