STRATEGIC TASK FORCE AREA REPORT

Data Science Education Task Force

19.January.2018

ILLINOIS

Table of Contents

I. TASK FORCE CHARGE

In Spring 2017, at the initiative of the Department of Statistics, the College of Liberal Arts and Sciences submitted a proposal on data science education to the campus's "Investment for Growth" call. Interim Provost John Wilkin responded that the reviewers supported the concept, but indicated that "the proposed offerings should be situated in the context of curriculum planning currently underway between LAS, the College of Business, College of Engineering and School of Information Sciences." He asked that these units collaborate to develop a coordinated proposal. The deans of those units established this Data Science Education (DSEd) task force in October, with the following charge.

> Dear Colleagues:
>
> Data science plays an increasingly important role in our research enterprise and in our society. There is an urgent need for a collaborative approach to **data science education** at Illinois that is commensurate with the importance of data science, its potential to benefit our students, and the broad ambitions of the University of Illinois.
>
> The Department of Statistics and the College of Liberal Arts and Sciences recently submitted a proposal to "Jump Start Data Science Education" to the campus' Investment for Growth program. Interim Provost Wilkin responded that the campus would invest in data education, and asked us to submit a proposal that coordinates the efforts of our colleges. We ask that you serve as a task force to develop such a proposal. Professor Matthew Ando has agreed to chair the task force.
>
> As you develop this proposal, keep in mind the diversity of scholarship on our campus, of our students, and of opportunities in data science. Data science education at Illinois should be both broad and deep, and it should respect the remarkable reach of data science into many disciplines. We believe that every Illinois undergraduate should have the opportunity to receive some training in data science. We should produce leaders in data science and leaders who know data science.
>
> The need for the University of Illinois to have a rich menu of educational opportunities in data science is pressing. We invite you to make interim reports and recommendations as your work progresses, and we ask that you complete your work by December 15.
>
> Sincerely,
>
> Feng Sheng Hu, Dean, College of Liberal Arts and Sciences
> Jeff Brown, Dean, College of Business
> Andreas Cangellaris, Dean, College of Engineering

Allen Renear, Dean, School of Information Sciences

Soon thereafter, Chancellor Jones and Interim Provost Wilkin established the Data Science Strategy Task Force as part of the university's strategic planning process. The membership of the DSEd task force was included in the larger task force, and they charged the two task forces to connect.

This report describes the DSEd task force's thinking in response to the strategic questions raised by the Chancellor's and Provost's charge to the Data Science Strategy Task Force. Because of the larger task force, this task force focused primarily on undergraduate programs. We will provide additional information about implementation to our deans in a follow-up report.

II. TASK FORCE MEMBERSHIP

Professor Matthew Ando, College of Liberal Arts and Sciences, *Chair*
Professor Roy Campbell, College of Engineering
Professor Jana Diesner, School of Information Sciences
Professor Jessen Hobson, College of Business
Professor Douglas Simpson, College of Liberal Arts and Sciences
Professor Ranamath Subramanyam, College of Business
Professor Ted Underwood, School of Information Sciences and College of Liberal Arts and Sciences
Professor Venugopal Veeravalli, College of Engineering

III. EXECUTIVE SUMMARY

The university has an exciting opportunity to build on its strengths and multiple interesting offerings in data science, to develop a broad-based, comprehensive, and inclusive program of data science education. The availability of massive amounts of data is transforming many fields of endeavor, and data science provides students and faculty with new problems and new approaches to old problems. Data science education will enhance our students' experience while they are here and create opportunities for them after they graduate. Broad-based and inclusive data science education means several things:

- *Every Illinois student should have the opportunity to have a meaningful exposure to data science.* There should be multiple on-ramps to data science education at Illinois, including for students from underserved populations and students who do not intend to major in STEM fields.
- Data science education at Illinois should be flexible to encompass the widely varying needs of different domains in which data science is used.
- Data science education at Illinois should be a collaborative undertaking.

In order to meet the varying needs of our students and faculty, a comprehensive and broad-based program of data science education should include the following elements, enabling students, faculty, and programs to engage with data science according to their needs:

- A data science major
- A family of data-science enhanced majors in other fields, in the manner of the innovative CS+X majors
- A data science minor or certificate
- Introductory courses to enable students to get started in data science

It will be important to support the incorporation of data science ideas into courses across the academy, including supporting faculty for whom data science is not an area of research expertise.

The university's broad-based approach to data science should have the flexibility to recognize and collaborate with discipline-specific data science education programs emerging in the curricula of various colleges.

IV. THE NEED FOR BROAD-BASED DATA SCIENCE EDUCATION

The need to jump start data science education at Illinois is clear. We are already doing a lot, and indeed what we are doing has brought national recognition to our university. Existing courses and programs in data science have grown dramatically in recent years.

As we studied data science education nationally, we frequently came across indications of Illinois's influence. We learned that our Statistics & Computer Science major was one starting point for the development of Michigan's undergraduate data science major. Multiple Illinois scholars were involved in the preparation of the National Academies' report *Envisioning the data science discipline: the undergraduate perspective*, and that report cites the CS+X program and other work at Illinois as leading contributions to data science education.

The growth of enrollment of data science programs here and elsewhere is phenomenal. For example:

- The Statistics & CS major has grown from 20 to 200 undergraduate majors in five years
- Enrollment in Statistics 200 has grown from 300 students in 2012 to 800 students in 2017. Enrollment in Statistics 212 has grown from 90 students in 2012 to 530 students in 2017.
- Enrollment in INFO 490 "Introduction to Data Science" was 100, 200, and then 300 in the Fall 2015, 2016, and 2017.

- Berkeley's introductory course Data 8 now enrolls 1,000 students each semester, in its sixth semester of its existence.

Some of the most active work in data science education has been carried out by data science researchers, who needed to prepare their students to work in the field. The breadth and intensity of these efforts is striking, involving multiple faculty, departments, and colleges. Collectively they demonstrate the strong commitment of our data science researchers to data science education. As a result of their initiative, our efforts have been more focused on advanced undergraduate students and graduate students. What is needed is to jump start a broad-based, inclusive program of undergraduate data science education that is accessible to all undergraduate students.

V. KEY COMPONENTS OF DATA SCIENCE EDUCATION

The task force considered a variety of sources in its review of the landscape of data science education: in addition to our own data science offerings, we studied among other sources data science programs at Berkeley and Michigan, and we reviewed two national position papers on data science education, one a consensus report from the National Academies of Science, Engineering, and Medicine, and one the report of the Park City Math Institute Summer Undergraduate Faculty Program ([1,2]).

Data science is rapidly evolving, but there is a consensus about key components and features of a data science curriculum. The foundational ideas come from statistics, computer science, and mathematics, but compared to traditional statistical and mathematical thinking, modern data science involves a greater emphasis on computational thinking and an immersive approach to complex data. One of the hallmarks of data science is that it has arisen simultaneously from the need to handle massive amounts data in multiple contexts: thus, it is important for data scientists to be able to solve problems in specific and different situations, to write and communicate, and to work well in multidisciplinary teams. Data science plays a large and rapidly increasing role in decision-making with vast consequences, and so data science education should involve a serious engagement with the ethical and social dimensions of data and data-driven decision-making. The following list of components of a data-science curriculum, from the National Academies' report, is typical of the sources and reflects the consensus of the task force.

- Mathematical foundations
- Computational thinking
- Statistical thinking
- Data management
- Data description and curation
- Data modeling
- Ethical problem solving

- Communication and reproducibility
- Domain-specific considerations

VI. PRINCIPLES FOR DATA SCIENCE EDUCATION AT ILLINOIS

Data science education at Illinois should enable and encourage broad participation. From the National Academies' report [2, p 4-1]:

> Data science programs have the potential to attract broad participation, including diverse members from different disciplines (including the humanities, social sciences, and the arts) and from populations that are underrepresented in other similar science, technology, engineering, and mathematics (STEM) fields… Part of this potential comes from the various compelling application areas of data science, including digital humanities, computational social science, public policy, and many others. There are also numerous skill sets that are currently captured under a *data scientist* label that span multiple training and education levels.

The principle of inclusion and collaboration extends through many aspects of education, especially the students we serve and the disciplines we engage.

- Every Illinois student should have the opportunity to have a meaningful exposure to data science. There should be multiple on-ramps to data science education at Illinois, including for students from underserved populations and students who do not intend to major in STEM fields.
- Data science education at Illinois should be flexible enough to encompass the widely varying needs of different domains in which data science is used.
- Data science education at Illinois should be a collaborative undertaking.
- Data science education at Illinois should encourage students learn to communicate and to work well in collaborative multidisciplinary settings.

An important ingredient is providing excellent on-ramps for data science. Berkeley has had tremendous success with a large introductory data science course, designed to be widely accessible. The course has a large and growing menu of "connector" courses, either building out core data science knowledge or providing domain-specific experience. With the rapidly growing use of data across our campus, an introductory course could jump-start a major transformation of the undergraduate experience.

VII. PROSPECTS FOR DATA SCIENCE EDUCATION AT ILLINOIS

A full set of offerings in undergraduate data science education would include a full-fledged data science major, a family of data-science enhanced majors, a minor/certificates/badges, and introductory courses. The seeds of many of these activities have already been planted.

- The existing Stat+CS provides the basis for strong data science major; as already mentioned, it was a reference point for Michigan's Data Science-Engineering degree. There is room to enhance its focus on data science. There is room to offer a concentration in this degree analogous to Michigan's Data Science-LSA degree.

- LAS and the Department of Statistics have already initiated a Certificate in data science. We can build on this certificate as such or as a minor, to collect and showcase data science offerings across the university and to offer students a coherent, flexible foundation in data science.

- There is an enormous opportunity to build a family of data science-enriched majors. The CS+X program has already demonstrated both the feasibility and the power of this approach. Because of the way massive data access is changing so many fields, dual degrees in many, many areas will offer compelling opportunities for education and scholarship.

- The Department of Statistics has a strong track record of developing accessible introductory courses. The approach of the current Stat 200/212 is more traditionally statistical than Berkeley's Data 8, but it has similarly enjoyed phenomenal growth, and we have a good foundation on which to build introductory data science courses. As one step in this direction, the College of Business is already introducing data science into its first-year undergraduate curriculum. We see tremendous potential in making similar opportunities available to all Illinois undergraduates.

- We see considerable potential to build an introductory course from a collection of mostly independent modules, whose core content could be available online. Individual modules could be (re)deployed in courses that need one or two ideas or data science methods. Additional modules could provide introductory material depending on students' preparation.

- We anticipate that many core courses in data science will be offered by Statistics, Computer Science, Information Science, and Mathematics. But we emphasize that many colleges and departments are developing data science courses and curricula. It will be good for students and for the university for us to collaborate in offering and recognizing data science education across the university.

Even with these strong elements, there remains a great deal to do to build a coherent, well-connected, and well-documented whole. From this effort we anticipate a tremendous benefit to the university, to our students, and to our academic programs. The committee identified some key steps for jump-starting data science education at Illinois.

- There is an urgent need for faculty to respond to current student demand and to accommodate future demand. This problem really needs a jump start: the phenomenal growth in students' interest in data science has exceeded the rate at which the standard mechanisms have enabled us to add capacity.

- Many of the core elements listed above, while an excellent starting point, need to be redesigned with a focus on data science. (The Park City volume [1] anticipates this situation, describing how to build a data science curriculum starting from existing traditional courses)

- We will need personnel who can advise and support faculty as they incorporate data into their courses.

This last point deserves emphasis. In the real world, data is messy. Wrangling messy data from multiple sources is an important part of a data scientist's job. When we are teaching that part of the job, messiness is a feature, but messiness is an impediment to the widespread incorporation of data across the curriculum. Supporting the use of data and data science across the curriculum is one of our principles of inclusion. We must be prepared to support faculty, data scientists or not, who wish to use data and data science methods in their courses.

VIII. THE ROLE OF AN INSTITUTE

The opportunity and challenge to work with massive amounts of data have arisen across the academy, and data science is an inherently interdisciplinary undertaking. There are faculty using cutting-edge data science in their research in almost every college. As these faculty turn to data science education, they encounter many common challenges. Many of them develop solutions which could be widely adopted. A data science institute could play a powerful enabling role for data science education, by hosting a community of scholars committed to data science and data science education.

- It would provide a community where new ideas can be identified, explored, and shared, and then incorporated into curricula.
- It would provide a place where best practices, data sets, and technical solutions can be organized, curated, and distributed.
- It would be a valuable resource not only for data science researchers but also for other faculty who wish to incorporate data and data science ideas into their courses

IX. OUTREACH AND IMPACT

A comprehensive broad-based program of data science education will enhance our students' experience both while they are here and after they leave. By producing scholars who are knowledgeable about data across our curricula, we will enhance the use and understanding of data across the state and beyond.

Two other impacts are worth mentioning. First, by focusing our attention on data science in early undergraduate education, we will develop expertise and capacity in data science education that can contribute to K-12 education both locally and statewide. The addition of statistics to high school curricula has broadened participation in science and mathematics. Data science has the potential to do the same. We have an opportunity to increase participation in the data science economy by helping our K-12 partners figure out their students' pathways to undergraduate data science. Second, there are great opportunities to assist local communities by helping them to understand data available to them. An important part of data science education is working with data, and there is ample evidence that student projects can have important impacts. One sees in the recent projects of CS 205 "Data driven discovery" the potential of students to study our own institution. The potential impact of connecting energetic students to data problems arising in the community is truly exciting.

REFERENCES

[1] R. De Veaux et al, "Curriculum guidelines for undergraduate programs in data science", *Annual Review of Statistics and its applications,* 2017, 4:2.1—2.16.

[2] "Envisioning the data science discipline: the undergraduate perspective (Interim Report)", Consensus study report, National Academies of Science, Engineering, and Medicine, 2017. http://nap.edu/24886