Appendix G

Summary

Data science is emerging as a field that is revolutionizing science and industries alike. Work across nearly all domains is becoming more data driven, affecting both the jobs that are available and the skills that are required. As more data and ways of analyzing them become available, more aspects of the economy, society, and daily life will become dependent on data. As a result, the National Academies of Sciences, Engineering, and Medicine were asked to set forth a vision for the emerging discipline of data science at the undergraduate level. To that end, the committee considered core underlying principles, intellectual content, and pedagogical issues specific to data science, including the essential concepts that distinguish it from neighboring disciplines. All of this was anchored in exploration related to applications of and careers in data science.

Today, the term “data scientist” typically describes a knowledge worker who is principally occupied with analyzing complex and massive data resources. However, data science spans a broader array of activities that involve applying principles for data collection, storage, integration, analysis, inference, communication, and ethics. In future decades, all undergraduates will benefit from a fundamental awareness of and competence in data science.

Recommendation 2.3: To prepare their graduates for this new data-driven era, academic institutions should encourage the development of a basic understanding of data science in all undergraduates.

The continued transformation of work requires both a larger population with a basic understanding of data science and a substantial cadre of talented graduates with highly developed data science skills and knowledge, acquired through substantial coursework and practice.

> **Recommendation 2.1: Academic institutions should embrace data science as a vital new field that requires specifically tailored instruction delivered through majors and minors in data science as well as the development of a cadre of faculty equipped to teach in this new field.**

The new majors and minors will initially combine ingredients from existing courses, in areas such as computer science, statistics, business analytics, information technology, optimization, applied mathematics, and numerical computing. Over time, as features of the new data-driven era take shape, academic programs will be compelled to develop new skill clusters, and a body of distinctive courses and instructional materials will emerge.

> **Recommendation 4.1: As data science programs develop, they should focus on attracting students with varied backgrounds and degrees of preparation and preparing them for success in a variety of careers.**

Graduates of these programs will work in virtually every job sector and will serve in a number of roles, including operating the systems on which analyses are run, preparing data for analysis, defining and coordinating the analysis, visualizing information, and supporting data-driven decision making to uncover the stories buried in the data. Others who use data science skills will be journalists, administrators, artists, lawyers, teachers, and other workers who need some ability to understand and use data. This need to prepare diverse students for various careers further increases the educational challenge.

A wide variety of instructional programs will be needed to prepare students for the data-enriched world of the coming years.

> **Recommendation 2.2: Academic institutions should provide and evolve a range of educational pathways to prepare students for an array of data science roles in the workplace.**

These include introductory courses, full degrees at both associate and bachelor levels, and a range of minors and certificates. The forms of these programs and their scope will vary depending on the culture of a given institution and the aims of its students.

Regardless of the type of program, certain elements need to be covered, though perhaps to varying degrees and with varying emphases. A key goal is to give all students the ability to make good judgments, use tools responsibly and effectively, and ultimately make good decisions using data. The committee defines this collection of abilities as "data acumen." To that end, students will need exposure to material from multiple disciplines—notably, mathematical, statistical, and computational foundations—and they will need training in data acquisition, modeling, management and curation, data visualization, workflow and reproducibility, communication and teamwork, domain-specific considerations, and ethical problem solving.

The committee underscores the centrality of studying the many ethical considerations that arise as workers engage in data science. These considerations include deciding what data to collect, obtaining permissions to use data, crediting the sources of data properly, validating the data's accuracy, taking steps to minimize bias, safeguarding the privacy of individuals referenced in the data, and using the data correctly and without alteration. It is important that students learn to recognize ethical issues and to apply a high ethical standard.[1]

Recommendation 2.4: Ethics is a topic that, given the nature of data science, students should learn and practice throughout their education. Academic institutions should ensure that ethics is woven into the data science curriculum from the beginning and throughout.

Recommendation 2.5: The data science community should adopt a code of ethics; such a code should be affirmed by members of professional societies, included in professional development programs and curricula, and conveyed through educational programs. The code should be reevaluated often in light of new developments.

Academic institutions are stepping up to these educational challenges with a variety of programs and educational pathways. Several 4-year undergraduate institutions offer data science majors and/or minors—serving not only those students pursuing data science as a career but also those students who want to acquire data skills while majoring in another field. Two-year institutions are starting to introduce associate's degrees and certificates in data science to prepare students to transfer to 4-year

[1] For information about community efforts toward more transparent data-driven decision making for social good, see http://datafordemocracy.org, accessed March 12, 2018.

programs or to give them skills to compete in the workforce. Summer programs enable undergraduate students to build up data science skills rapidly. Boot camps and intensive training programs that aim to refresh or retool postgraduate students with the skills required of the growing data science workforce are now appearing. Massive open online courses in data science are proliferating and serve as stand-alone points of entry for all kinds of students (and flexible opportunities for professional development for instructors).

These pioneering examples of programs show what is possible, but there are significant challenges to developing data science programs more broadly and pervasively. The popularity of data science courses and programs will affect the entire academic institution by influencing enrollment, budgets, classroom allocation, computing resources, and scheduling. Institutions may need to consider how to create incentives for faculty in multiple departments and fields to collaborate to develop and deliver curricula that best meet students' needs. Today, there is a shortage of faculty in this rapidly evolving area. Enlisting and training existing faculty will be essential in the short term, and developing new faculty will be important in the long term. These challenges, among others, will need to be addressed to ensure the success of undergraduate data science students.

> **Recommendation 5.1: Because these are early days for undergraduate data science education, academic institutions should be prepared to evolve programs over time. They should create and maintain the flexibility and incentives to facilitate the sharing of courses, materials, and faculty among departments and programs.**

The evolution of data science programs will be affected by a broad range of factors, including their initial home and structure, the needs and interests of students, and institutional culture. Although new programs could be launched by combining existing courses and materials, over time new classes and materials will need to be developed. Institutions will need to think through the pathways students are taking into data science and how to create bridges and remove barriers. Academic and career advising will be vital parts of data science programs; the advising programs will themselves need to evolve as the field and the market for graduates mature.

Data science itself provides tools to continuously evaluate and improve data science education. Evaluation should include assessment of student learning and assessment of how well a program is meeting the needs of the market it aims to serve. Evaluation can be used to shape a program at a given institution, showing what is working and where improvement is needed. It can also be used comparatively to detect approaches, classes, or curricula that may be of value to other campuses or contexts.

> **Recommendation 5.3: Academic institutions should ensure that programs are continuously evaluated and should work together to develop professional approaches to evaluation. This should include developing and sharing measurement and evaluation frameworks, data sets, and a culture of evolution guided by high-quality evaluation. Efforts should be made to establish relationships with sector-specific professional societies to help align education evaluation with market impacts.**

Much of the necessary data for evaluation could come from institutions' administrative records. These records, used in conjunction with other data sources such as economic information and survey data, could enable effective transformation and generalization of programs and might even inform a cohesive national approach to undergraduate data science education.

In many fields, professional societies play a role in creating and nurturing community, in facilitating the sharing of resources and results, and in convening groups to set standards or determine best practices. Such capabilities are valuable to data science as well. However, it may be difficult for a single existing society to represent all the interests of the data science community. A structured collaboration of existing professional societies might work better, with potential development of subsocieties devoted to data science elements in any of many preexisting societies.

> **Recommendation 5.4: Existing professional societies should coordinate to enable regular convening sessions on data science among their members. Peer review and discussion are essential to share ideas, best practices, and data.**

Conferences, workshops, training sessions, and other networking opportunities would benefit the joint communities. Other opportunities for the collaborating societies would be collecting materials; convening discussions around critical topics such as curriculum, evaluation, and ensuring broad participation; and potentially creating publication venues for the broad community. *As data science continues to evolve, it is essential that academic institutions and other stakeholders take steps to prepare students for a data-enabled world. The time to act is now.*