Education for Information 36 (2020) 109–117
DOI 10.3233/EFI-200372
IOS Press

Teaching undergraduate data science for information schools

Loni Hagen
School of Information, University of South Florida, FL, USA
E-mail: lonihagen@usf.edu

Using the Conway model of data science education as a guide, this paper introduces a model for undergraduate data science education for information schools. The core idea of the suggested model is that data science programs in information schools are unique due to their particular substantive expertise, which includes data management, information behavior, and ethics. This paper also suggests that, to create a data science program within an information school, it may be useful to expand curriculums by adding programming, statistics, and machine learning requirements.

Keywords: Data science, data science education, iSchools, information science, information science education, Conway, data science curriculum

1. Introduction

This paper proposes a model of data science education for information schools that is based on a program developed as part of an undergraduate information science degree. Data science is an interdisciplinary field that focuses on extracting useful patterns from large sets of data. As the field has expanded, the scope of "data science" has evolved into different areas of emphasis from the perspective of different disciplines. Computer scientists and statisticians seek to develop algorithms and software tools to process data efficiently in order to discover useful information, and to better represent the discovered information. On the other hand, social scientists extract useful patterns from human behavioral data to study and model behaviors and relationships among individuals within society. Information professionals are interested in a user's information behavior as well as collecting, organizing, disseminating, using, and preserving data and information. Despite these varying goals and focus areas across broad disciplines, there are common elements of data science, in particular the *use of large volumes of data and dependence on scientific approaches to make sense of it* (Stanford Data Science Initiative, 2019). To do data science, some skills and processes are also common and fundamental. I will discuss some common skills that should be taught for data science curriculums in information schools, and discuss the unique domain expertise that information science schools have, which distinguishes us from other disciplines providing data science education.

I conclude the paper by suggesting a model for undergraduate data science education in information schools. Just for clarification, this paper presumes that a goal of

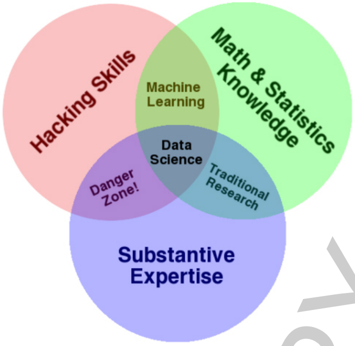

Fig. 1. The Data Science Venn Diagram (Conway, 2019).

data science is improving decision making, not only developing software products or algorithms.

2. Three fundamental requirements for data science

With minor adjustments, the model I'm advocating is based on the Conway Data Science Venn Diagram, according to which the three fundamental requirements for data science are "Hacking" (programming) skills, competence in math/statistics, and domain expertise (Conway, 2019).

First, programming enables data scientists to collect, to clean, and to manipulate massive quantities of data, which cannot be easily done using off-the-shelf software. These fundamental skills require a minimum of one year of learning and continuous practice. R and Python are predominant in data science education. I prefer to use R because of its strength in statistical analyses and visualization.

Second, in order to extract useful information from data, one needs to apply statistical and machine learning techniques. Although Conway highlights math and statistics, in this paper I modify this requirement to statistical and machine learning knowledge. I assume basic math is required for statistics and machine learning, but data science education in information schools does not require as much math as computer science and statistics departments require. Statistical tests support decision-making based on quantitative evidence found in data. When dealing with truly large and complex datasets, machine learning algorithms enable the discovery of potentially interesting patterns when we do not know a proper query. When statistical tests and machine learning are properly used, the extracted information is reliable and generalizable. Data scientists should have "the ability to assess which models are feasible, desirable, and practical in different settings" (Dhar, 2013, p. 69). Data scientists should also know how to interpret the results. Conway warns that "hackers" (programmers) without knowledge of math and statistics are in a "danger zone" because the products they create may not be based on a solid understanding of the process, and they may not be able to provide a reliable interpretation of the results.

Third, substantive domain expertise is crucial for generating project goals and interpreting the results. Substantive expertise, in combination with critical thinking skills, is necessary to draw useful insights from model outputs. Information science schools add value to data science education by equipping students with substantive expertise regarding information – how to collect, manage, disseminate, use, store, and retrieve information that meets a user's needs.

In the following section, I will discuss the substantive expertise that information schools offer in the context of data science education (i.e. the "Substantive expertise" requirement), then proceed to introduce specific education models to meet the other two requirements of the Conway model.

3. Substantive expertise provided by information schools

How do information schools' data science programs differ from other disciplines? I argue that the particular substantive expertise provided by information schools distinguishes our data science programs from those of other neighboring disciplines.

For example, while other disciplines require instruction in knowledge areas such as data management, data privacy, and data integrity, for information schools these are part of the core knowledge areas. These are the knowledge areas that information schools emphasize. For us, they are not merely tangential.

What distinguishes our data science programs from computer science and statistics is the value and the goal of our programs. The value information schools cherish is providing services to information users, and serving for the public good, democracy, and equality. This is in contrast to the value of data science expressed by mathematics and computer science disciplines, which include developing fundamental mathematical foundations and computational/statistical thinking (De Veaux et al., 2017). As a result, creating accurate and efficient algorithms and computational tools may be among the chief goals of these disciplines. In contrast, information science students may place more emphasis on the consequences of a computational tool in relation to social values such as the public good and equality.

In this section, I describe some of the substantive expertise provided by information schools, then demonstrate how the courses already provided at the University of South Florida (USF) School of Information fit in the substantive expertise requirements of the Conway model of data science.

First, information schools cherish values and ethics that provide a framework for information professionals' decisions on "conduct, policies, and services" (Rubin, 2015, p. 405). One can argue that data science programs in business and computer science may require students to take an ethics course as well. The major distinction is that information schools do not treat ethics as a compliance issue, which can be met by taking one such course. Instead, many information science courses include

Plan: This involves creating a data management plan, including how data will be managed during and after the project.
Collect: Data can be collected and organized in diverse ways from hand-written texts to sensors and automated data collection. Many tools and approaches are used for data collection such as spreadsheets and relational databases.
Assure: The quality of the data is assured through validation. The context of the data collection that might be related to the data quality should be described. Visual or statistical summaries should be provided. Missing values need to be identified, and the overall data quality should be communicated.
Describe: Data is described by adding information that is necessary to understand the "who, what, when, where, why and how of the dataset" using the appropriate metadata standards (Michener, 2018a, p. 71).
Preserve: Data is stored in an appropriate long-term archive or data center. Having some assistance from an archivist is recommended in order to identify the long-term value of the data, and to use standard terminology to describe the data (Strasser et al., n.d.)
Discover: Data discovery involves locating and identifying potentially useful data through data repositories, data directories, and data aggregators as well as Internet search engines.
Integrate: The majority of data science projects depend on data from diverse data sources. These sources need to be combined (or joined, or merged) to form a single set of data objects that are ready for analysis.
Analyze: Data is explored, analyzed, and visualized.

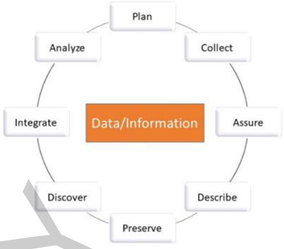

Fig. 2. Data life cycle based on DataONE (Michener & Jones, 2012).

consideration of ethics and human rights. For example, the USF School of Information provides a course called Information Policy and Ethics to teach information policy (rules and regulations) and ethical perspectives relevant to information and communication technology. In addition, many other courses cover ethics, policy, and values as part of the content: Introduction to Intelligence Studies covers ethics and policy issues, Information and Social Media covers social inclusion issues, Health Information Security and Privacy covers policies related with security and privacy, and Predictive Analytics discusses ethical issues arising from data collection and prediction results. These courses exemplify how values and ethics are embedded throughout our educational program.

Second, data management is a core competency area of information science education. Data management – standards of data collection, format, and quality – is one of the major "data challenges" identified by many private and public institutions (Sun & Medaglia, 2019). Outcomes from unverified data quality and unvalidated computational process may be misleading and can result in failure (Ault, 1987; Grimes, 2010; Janssen et al., 2017; Strong et al., 1997). Some scholars conceptualize data management following a data life cycle. According to DataONE (Strasser et al., n.d.), one of the leading data management models, the data life cycle has eight components: plan, collect, assure, describe, preserve, discover, integrate, and analyze (Fig. 2). The details of the components, which highlight the complexity of data management, are described in Fig. 2.

One may argue that business, computer science, and mathematics programs also provide some data management courses. Again, information schools are distinctive because of our values: we do these activities to serve information users, without ignoring those who are difficult to serve. We do not treat information as a commodity to sell, but we treat it as a service (Rubin, 2015). In addition, instead of providing one or two technical courses to cover data management, we provide a diverse array of courses to provide the competencies necessary for data management. For example, USF's

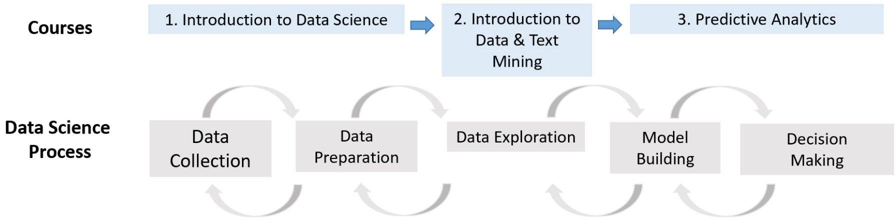

Fig. 3. Three sequential courses in relation to the process of data science.

School of Information provides courses to cover various aspects of data management, such as Information Architecture, Data Structures and Algorithms, and Introduction to Data Science, which teach the concepts and application of data structures as well as how to design and maintain digital information spaces for human access, navigation, and use. Other courses like Core IT Concepts for Information Professionals teach the basic concepts of computer hardware and software, which is crucial for designing and managing data. And Information Assurance and Security Management for IT teaches the foundations of physical, network, logical, and data security.

Third, information behavior is another essential area of expertise that enhances understanding of information behaviors of individuals and groups in context in order to seek, access, and retrieve information. USF's School of Information has a course named Information Behaviors that teaches individual and group information seeking, search, and retrieval behaviors in context.

These courses are only samples of courses that are relevant for "substantive expertise" provided by USF's School of Information. Each school will have their own flavor of substantive expertise. To that we add the two other requirements (programming and statistical & machine learning) to create the core of a data science program. These are discussed in Section 5. But first, in the following section, I will discuss the data science process.

4. The process of data science

All data science projects share a set of common processes: data collection, preparation, exploration, model building, and communicating results for decision making. These processes are not entirely linear, and each stage should be closely monitored by humans to ensure validity and integrity (see Fig. 3 for a visual representation of the process).

First, data science requires data collection. A variety of human behaviors are recorded through various data sources such as social media (e.g., Twitter, and Facebook), open government data repositories (e.g., data.gov, US Census Bureau), and other data repositories (e.g., Inter-university Consortium for Political and Social Research, Kaggle, the New York Times, Google Trends, and Amazon Web Services

Public datasets). While not all data are relevant, useful, or usable, a data scientist should be able to discover and collect data that are relevant and useful given the project's goals.

Second, data science involves time-intensive and work-intensive data preparation, which typically accounts for over 80% of the time involved in the data science process. Messy data need to be cleaned, normalized, and transformed so that a data scientist can trust that the data are useful for robust analysis.

Data collection and preparation processes are similar to the data management activities discussed in Section 3 (to collect, assure, describe, preserve, and discover). Data scientists do collect, integrate, and clean datasets, which does indeed consume much of our time, but our mental focus is on the exploration and model building process discussed below.

Third, data exploration refers to a process of understanding the relationships among variables in preparation for model building. Data exploration methods include such methods as data visualization and clustering, which can help in understanding the data from a high level. Data exploration itself can lead to the identification of useful patterns (such as anomalies) in the data.

The fourth stage is model planning and model building, where proper model building techniques are selected, and model performance and validity are measured by training and testing datasets.

Fifth, data scientists interpret the results, taking into consideration the statistical significance and validity of the results, and communicating them to decision makers.

5. Programming, statistics, and machine learning: A case introduction

To illustrate how the fundamental skills and common processes of data science can fit into a data science curriculum, I here introduce three undergraduate classes that I developed for an undergraduate information science degree program (BSIS) with a concentration in Data Science and Analytics. I introduce these courses (all provided online) and the expected outcome goals of the courses. These courses are designed to cover the entire data science process (Fig. 3), and to meet the "Programming" and "Statistics & Machine Learning" requirements of the modified Conway Model as well. The courses and expected outcomes for each follow.

5.1. Introduction to data science

Students can define data science and identify multiple data formats. In addition, students are capable of collecting and cleaning various data formats (i.e., CSV, JSON, and SQL), preparing the data for robust analyses. Though not strictly required, students are expected to take at least one programming course (typically students take Object Oriented Programming before they take Introduction to Data Science).

5.2. Introduction to data and text mining

Students can extract actionable information from raw data using mining tools (e.g., association rules, support vector machine, sentiment analysis), present information using visualization tools (e.g., ggplot2, Shiny app), and draw insights from the extracted information. The required pre-requisite course is Introduction to Data Science.

5.3. Predictive analytics

Students can build/evaluate/validate models to make predictions regarding human behavior or their interactions within social systems, and can interpret the results critically, reflecting on the context of the data and considering potential limitations (such as data quality and algorithmic biases). The required pre-requisite course is Introduction to Data and Text Mining.

As illustrated in Fig. 3, the three courses are designed to cover the data science process in a sequential manner, thus the courses are sequential as well. Students can develop a high level of programming capability after finishing the three sequential courses. The level of programming complexity increases as the sequence progresses.

Learning to program involves learning by doing. Students have continuous practice opportunities for programming through the "DataCamp for the Classroom" service (https://www.datacamp.com/), which gives students in programming courses free access. In addition, students have weekly hands-on assignments that require applying the learned contents to real world datasets, such as local government and library datasets.

In the courses, I ask students to discuss analytics results in relation to the original data, data manipulation, and analyses processes. This gives students opportunities to think about the context of data, and the impact of each data science process on the outcome results.

Three other courses required for the concentration in Data Science and Analytics are Introduction to Visual Analytics, R Programming for Data Science and Advanced Statistics and Analytics. All of these courses directly enhance competencies in the "Programming" and "Statistics" requirements of the Conway Model as well.

6. Student population, and preliminary evaluation

An education model for data science (or any similar field of study) should take into consideration the student population, including the level of academic preparation and available job markets. Below, I provide a summary of our student population as background information on the target students of the BSIS program.

Preliminary results of a survey of students taking my Introduction to Data Science course indicate that about 40% of the students work full time, and the majority of their

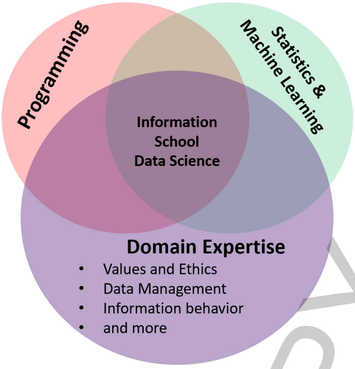

Fig. 4. A model for data science for information schools, a modified Conway model (2010).

jobs are related to information and data science (examples of employment include: a biotech company, an insurance company, a financial/data analytics department, and digital scholarship services for a university library). The majority of students have taken some kind of programming and statistics courses.

Data science education in my information school is still new, thus it is challenging to find relevant queries to search for the jobs that match with students' specialties. I am not able to systematically capture students' job placements, but I can report some anecdotal cases. The first graduate with a BSIS degree with a Data Science concentration works in a major insurance company as a risk analyst. Another student started working as a technical solutions analyst for a software development company. Some other students, who have full time jobs, have applied to Master's programs in data science. For information schools, vocational and professional training is one of our goals. So, capturing job placements after students' graduation is an important task that should be done by schools providing data science education.

7. Conclusion

I conclude by suggesting a conceptual model for undergraduate data science education in information schools using a slightly revised version of Conway's Venn diagram. Information schools provide substantive expertise: values and ethics, information management, information behavior, etc. In order to create a data science program within an information school, we may need to expand curriculums to add programming, statistics, and machine learning requirements.

Programming is a fundamental skill for data science, and should be introduced throughout the program so that students can develop sufficient proficiency in at least one programming language, preferably R or Python. Statistics and machine learning are also fundamental, and so should be taught independently and/or integrated into substantive expertise courses.

The updated Venn diagram (in Fig. 4) reflects the suggested contents of data science education in information schools. As reflected in the size of the circles in Fig. 4, data science in information schools may involve less programming, statistics, and machine learning compared to the disciplines of computer science and statistics, but adds specific kinds of domain expertise. The substantive expertise dimension can be adjusted based on the departments' specific strengths.

Acknowledgments

This study is supported by the National Research Foundation of Korea Grant funded by the Korean Government (NRF-2017S1A3A2066084).

References

Ault, M. R. (1987). Combating the Garbage-In, Gospel-Out Syndrome. *Radiation Protection Management*. https://www.researchgate.net/publication/268357767_Combating_the_Garbage-In_Gospel-Out_Syndrome.

Conway, D. (2019). *The Data Science Venn Diagram*. Drew Conway. http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram.

De Veaux, R. D., Agarwal, M., Averett, M., Baumer, B. S., Bray, A., Bressoud, T. C., Bryant, L., Cheng, L. Z., Francis, A., Gould, R., Kim, A. Y., Kretchmar, M., Lu, Q., Moskol, A., Nolan, D., Pelayo, R., Raleigh, S., Sethi, R. J., Sondjaja, M., et al. (2017). Curriculum Guidelines for Undergraduate Programs in Data Science. *Annual Review of Statistics and Its Application*, *4*(1), 15-30. 10.1146/annurev-statistics-060116-053930.

Dhar, V. (2013). Data science and prediction. *Communications of the ACM*, *56*(12), 64-73. 10.1145/2500499.

Grimes, D. A. (2010). Epidemiologic Research Using Administrative Databases: Garbage In, Garbage Out. *Obstetrics & Gynecology*, *116*(5), 1018-1019. 10.1097/AOG.0b013e3181f98300.

Janssen, M., Voort van der, H., & Wahyudi, A. (2017). Factors influencing big data decision-making quality. *Journal of Business Research*, *70*, 338-345. 10.1016/j.jbusres.2016.08.007.

Michener, W. K. (2018a). 5. Creating and Managing Metadata. In F. Recknagel & W. K. Michener (Eds.), *Ecological Informatics: Data Management and Knowledge Discovery* (pp. 71-88). Springer International Publishing. 10.1007/978-3-319-59928-1_5.

Michener, W. K. (2018b). 7. Data Discovery. In F. Recknagel & W. K. Michener (Eds.), *Ecological Informatics: Data Management and Knowledge Discovery* (pp. 115-128). Springer International Publishing. 10.1007/978-3-319-59928-1_7.

Michener, W. K., & Jones, M. B. (2012). Ecoinformatics: Supporting ecology as a data-intensive science. *Trends in Ecology & Evolution*, *27*(2), 85-93. 10.1016/j.tree.2011.11.016.

Rubin, E. (2015). *Foundations of Library and Information Science: Fourth Edition*. American Library Association. http://ebookcentral.proquest.com/lib/usf/detail.action?docID=5185108.

Stanford Data Science Initiative. (2019). *Data Science for Humanity*. Stanford University. https://sdsi.stanford.edu/about/data-science-humanity.

Strasser, C., Cook, R., Michener, W., & Budden, A. (n.d.). *Primer on Data Management: What you always wanted to know*. 11.

Strong, D. M., Lee, Y. W., & Wang, R. Y. (1997). Data quality in context. *Communications of the ACM*, *40*(5), 103-110.

Sun, T. Q., & Medaglia, R. (2019). Mapping the challenges of Artificial Intelligence in the public sector: Evidence from public healthcare. *Government Information Quarterly*, *36*(2), 368-383. 10.1016/j.giq.2018.09.008.