

Artificial Intelligence Competencies for Data Science Undergraduate Curricula

Andrea Danyluk
Williams College
47 Lab Campus Drive
Williamstown, MA 01267
andrea@cs.williams.edu

Scott Buck
Intel Corporation
5000 W. Chandler Blvd., m/s CH7-230
Chandler, AZ 85226
scott.buck@intel.com

Abstract

In August 2017, the ACM Education Council initiated a task force to add to the broad, interdisciplinary conversation on data science, with an articulation of the role of computing discipline-specific contributions to this emerging field. Specifically, the task force is seeking to define what the computing contributions are to this new field, in order to provide guidance for computer science or similar departments offering data science programs of study at the undergraduate level. The ACM Data Science Task Force has completed the initial draft of a curricular report. The computing-knowledge areas identified in the report are drawn from across computing disciplines and include several sub-areas of AI. This short paper describes the overall project, highlights AI-relevant areas, and seeks to open a dialog about the AI competencies that are to be considered central to a data science undergraduate curriculum.

The ACM Data Science Task Force was created by the ACM Education Council and tasked with articulating the role of computing discipline-specific contributions to this emerging field. The task force is developing recommendations for curricula at the undergraduate level, and has released a first draft report for public discussion and comment. Here we give a broad overview of this effort, describe the draft report and its contents, and call for a discussion with the AI education community. AI and Data Science are by no means synonymous, but there are many areas of AI that are directly relevant to Data Science, including machine learning, computer vision, intelligent interfaces, and speech and natural language processing. Our goal is to present the intersections identified as being most central to Data Science, and solicit feedback from the AI community.

Overview of the Report Draft

The ACM, in partnership with other professional societies, has been producing curricular recommendations for well over 60 years (ACM 2018). The product of the ACM Data Science Task Force will not be a *complete* volume of curricular recommendations in the sense of other ACM curricular volumes. Instead, it is intended to be part white paper and part curricular volume: a thorough articulation of the

computing discipline-specific contributions to undergraduate data science curricula, as well as a call to create a joint task force of the ACM with other (non-computing) societies to produce a complete volume that is global in scope. While there have been several important curricular efforts surrounding data science, they have either been regional [e.g., the EDISON project in Europe (EDISON 2018)]; focused on core curricula without a full articulation of possible electives and connections to other disciplines [e.g., (DeVeaux et al. 2017)], or higher level statements on the importance of data science [e.g., (National Academies of Sciences, Engineering, and Medicine 2018)]. The draft of the ACM Data Science Task Force report contains an introduction to the purpose and processes of the task force, a summary of two surveys fielded by the committee, a review of other data science curricular reports and projects, itemization of computing discipline-specific Knowledge Areas, and data science-relevant competencies in each of these areas.

ACM Data Science Survey

In late spring / summer 2018, the ACM Data Science Task Force fielded two surveys to gather the computing community's perspective on the relative importance of computing concepts for data science. The academic survey asked academics whether their institution had any sort of data science program at the undergraduate level, asked what type of program was offered, in what department(s) it was housed, and what computing areas were required, elective, or not present in the program. It also allowed respondents to add to the list of computing areas. Finally, the survey asked participants whether their data science program had a "data science in context" requirement – i.e., a requirement that students apply data science to another area. The industry survey roughly mirrored the academic one; however, the primary question was whether the company looked for job applicants with data science experience and what computing experience they required or preferred those applicants to have.

Other Data Science Reports and Projects

As noted above, important work has already been done in the data science curriculum space. For example, in October 2015, the National Science Foundation (NSF) sponsored a workshop with representatives of many perspectives on data science (Cassel and Topi 2015). The EDISON Data Science

Project was launched in 2015 with the purpose of “accelerating the creation of the Data Science profession” (Manieri et al. 2015). The EDISON consortium, composed of partners from across Europe, recently made available Release 2 of their Data Science Competence Framework (EDISON 2018). The Park City Math Institute brought together an interdisciplinary faculty group to devise curriculum guidelines for undergraduate programs in data science (DeVeaux et al. 2017). And recently, the National Academies of Sciences, Engineering, and Medicine released an extensive report on Data Sciences for Undergraduates (National Academies of Sciences, Engineering, and Medicine 2018).

The goal of the ACM Data Science Task Force is not to duplicate these important efforts but to further articulate the contributions of the computing discipline to data science, as well as to provide additional detail on computing discipline-specific competencies. The task force and associated advisory committee are global committees, with representatives from four continents, as well as from academia and industry.

Knowledge Areas

Following the structure of ACM curricular volumes, the task force is organizing the body of computing knowledge into Knowledge Areas (KAs). In the draft report, the computing discipline-specific KAs for undergraduate data science are:

- Programming, algorithmic design, and development;
- Data acquirement and governance;
- Data management, storage, retrieval;
- Data privacy, security, integrity;
- Machine learning;
- Data mining;
- Systems considerations for data mining “in the large”: big data, distributed systems, parallel computing, HPC;
- Professionalism: communication, collaboration, ethics, evidence-based decision making and problem solving;
- Analysis and presentation: visualization, HCI.

Competencies

For each KA, the draft report provides the scope of material that should be covered in an undergraduate data science curriculum, as well as the competencies to be achieved by students. As an example, we provide the competencies identified for machine learning (ML): (i) Compare and contrast broad classes of learning approaches, with a focus on inputs, outputs, and ranges of problem types to which they can be applied; (ii) Select and apply a broad range of ML tools/implementations to real data; (iii) Derive a (current) learning algorithm from first principles and/or justify a (current) learning algorithm from a mathematical, statistical, or information-theoretic perspective; (iv) Express formally the representational power of models learned by an algorithm, and relate that to issues such as expressiveness and overfitting; (v) Exhibit knowledge of methods to mitigate the effects of overfitting and curse of dimensionality in the context of ML algorithms; (vi) Provide an appropriate performance

metric for evaluating ML algorithms/tools for a given problem; (vii) Apply appropriate empirical evaluation methodology to assess the performance of a ML algorithm/tool for a problem; (viii) Apply appropriate empirical evaluation methodology to compare ML algorithms/tools to each other; (ix) Implement ML programs from their algorithmic specifications; (x) Be aware of problems related to algorithmic and data bias, as well as privacy and integrity of data; (xi) Consider and evaluate the possible effects – both positive and negative – of decisions arising from ML conclusions; (xii) Compare differences in interpretability of learned models.

Relationship between AI and Data Science Education

AI and Data Science education are by no means synonymous, but there are very clear overlaps both in topics and competencies. Obvious intersections include machine learning, data preparation (to the extent that it includes feature extraction, for example), and HCI (if it includes intelligent interfaces). One could also argue that there are important AI areas that are not currently represented in the list above but that should be, such as certain areas of natural language processing or ontologies and the semantic web. The goal of this paper is to focus on the intersections already identified, encourage discussion with the AI education community, and solicit input on any areas that the task force has missed.

References

- ACM. 2018. Curricula Recommendations. <https://www.acm.org/education/curricula-recommendations>. Accessed: 2018-09-05.
- Cassel, B., and Topi, H. 2015. Strengthening data science education through collaboration. Technical report.
- DeVeaux, R.; Agarwal, M.; Averett, M.; Baumer, B.; Bray, A.; Bressoud, T.; Bryant, L.; Cheng, L.; Francis, A.; Gould, R.; Kim, A.; Kretchmar, M.; Lu, Q.; Moskol, A.; Nolan, D.; Pelayo, R.; Raleigh, S.; Sethi, R.; Sondjaja, M.; Tiruvilumala, N.; Uhlig, P.; Washington, T.; Wesley, C.; White, D.; and Ye, P. 2017. Curriculum guidelines for undergraduate programs in data science. *Annual Review of Statistics and Its Application* 4:15–30.
- EDISON. 2018. The EDISON Data Science Competence Framework. <http://edison-project.eu/edison/edison-data-science-framework-edsf>. Accessed: 2018-08-30.
- Manieri, A.; Brewer, S.; Riestra, R.; Demchenko, Y.; Hemmje, M.; Wiktorski, T.; Ferrari, T.; and Frey, J. 2015. Data Science Professional Uncovered: How the EDISON Project will Contribute to a Widely Accepted Profile for Data Scientists. In *2015 IEEE 7th International Conference on Cloud Computing Technology and Science (CloudCom)*, 588–593.
- National Academies of Sciences, Engineering, and Medicine. 2018. *Data Science for Undergraduates: Opportunities and Options*. Washington, DC: The National Academies Press.