ACM Taskforce Efforts on Computing Competencies for Undergraduate Data Science Curricula

Draft Report and Opportunity for Feedback

Paul Leidig
Grand Valley State University
leidig@gvsu.edu

Lillian (Boots) Cassel
Villanova University
lillian.cassel@villanova.edu

Outline

- The Data Science Curriculum Task Force Effort
  - Committee
  - Background
- The Second Draft Curriculum Report
  - Contents
  - Knowledge Areas and Competencies
- Community Engagement
  - Timeline
  - Discussion

ACM Data Science Task Force

- Andrea Danyluk, Co-chair, Williams College, USA
- Paul Leidig, Co-chair, Grand Valley State University, USA
- Scott Buck, Intel Corporation, USA
- Lillian Cassel, Villanova University, USA
- Maureen Doyle, Northern Kentucky University, USA
- Keegan Hines, Capital One, USA
- Tin Kam Ho, IBM, USA
- Andrew McGettrick, University of Strathclyde, UK
- Suzanne McIntosh, New York University, USA
- Jian Pei, Simon Fraser University, Canada
- Weining Qian, East China Normal University, China
- Karl Schmitt, Valparaiso University, USA
- Christian Servin, El Paso Community College, USA
- Hongzhi Wang, Harbin Institute of Technology, China

ACM Data Science Task Force Charter

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. The task force should seek to define what the computing contributions are to this new field, and should provide guidance for undergraduate data science programs of study.

To create a report, which may then be used to invite collaboration and coordination with other (non-computing) professional societies.

Background

- ACM Ed. Council summer meeting 2017
  - Build on the efforts of Boots Cassel & Heikki Topi, as well as other groups
  - Articulate importance of computing in the interdisciplinary data science space
  - Identify computing-based competencies for an undergraduate data science curriculum

Other Data Science Efforts

- EDISON Project (2017)
  - A competency-based framework to be used as guidance for educators, employers, etc.
  - Most similar to ACM effort; Europe focus.
- Park City Math Institute Report (2017)
  - Topics and learning outcomes for undergraduate data science curricula
  - Sample course outline
  - Sponsored by and published on *www.amstat.org*
- National Academies report (2018)
  - Higher level articulation of the importance of data science education

Computing Competencies for Undergraduate Data Science Curricula

Second Draft Report

http://dstf.acm.org/DSReportDraft2Full.pdf

Report Contents

Competency Framework

- Following ACM/IEEE-CS IT 2017; moving in the direction of CC 2020.
- Utilize a working definition of competency that connects knowledge, skills, and dispositions.
- Includes, but moves beyond articulation of topics and learning outcomes. [e.g., CS2013]

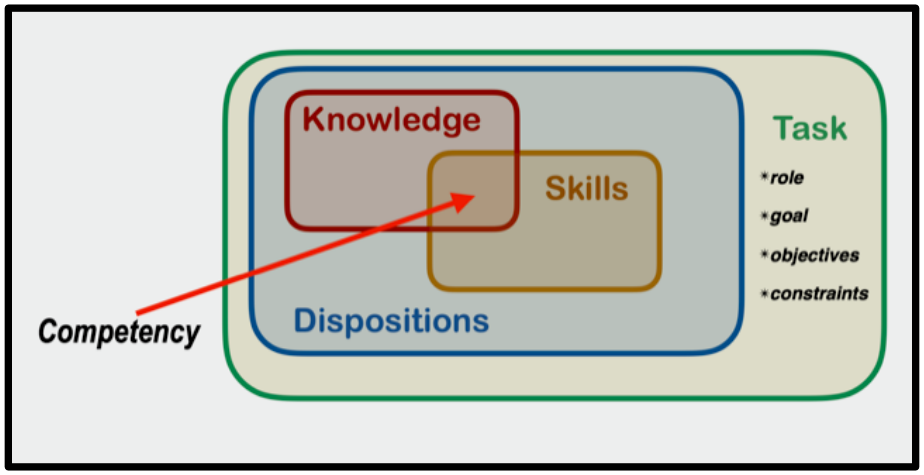
Competency = Knowledge + Skills + Dispositions

- Knowledge
  - Mastery of content
  - Transfer of learning
- Skills
  - Capabilities and strategies for higher-order thinking
  - Interactions with others and world around
- Dispositions
  - Personal qualities (socio-emotional skills, behaviors, attitudes) associated with success in college and career

From IT 2017; adapted from a publication by Council of Chief State School Officers (2013).

Competencies

CC2020



Competencies = Knowledge (K) + Skills (S) + Dispositions (D)

- **Knowledge** - a fact/idea that enables satisfactory performance of relevant tasks
- **Skill** - a degree of mastery in applying a fact/idea to achieve a valued outcome
- **Dispositions** - values and motivation that moderates skilled behavior to influence a quality of professional performance

The Body of Knowledge:
Computing Competencies for Data Science

- **Analysis and Presentation (AP)**
- **Artificial Intelligence (AI)**
- **Big Data Systems (BDS)**
- **Computing and Computer Fundamentals (CCF)**
- **Data Acquisition, Management, and Governance (DG)**
- **Data Mining (DM)**
- **Data Privacy, Security, Integrity, and Analysis for Security (DPSIA)**
- **Machine Learning (ML)**
- **Professionalism (PR)**
- **Programming, Data Structures, and Algorithms (PDA)**
- **Software Development and Maintenance (SDM)**

Levels of expectation

- Like CS2013, the report lists competencies in levels of expectation:
  - T1 (Tier 1) denotes an item that all Data Science graduates should have mastered
  - T2 (Tier 2) denotes an item that most Data Science graduates would be expected to have mastered. Any given Data Science graduate would be expected to have mastered a majority of T2 items.
  - E (Elective) signifies an item that, although important, could reasonably be regarded as forming part of an elective

Current KA Structure

Data Acquisition, Management, and Governance

There can be no analysis of data without the data itself. A data scientist must understand the source and quality of their data, as well as understand appropriate processes for acquiring and maintaining high quality data.

Scope	Competencies
Acquiring data from physical world and extracting data to a form suitable for analysis.	Construct and tune the data acquisition and governance process according to the requirements of an application, including the selection of data sources, data acquisition equipment, and data preparation algorithms.

Disposition: An ability to assess the trade-off between accuracy and efficiency in data acquisition

Additional Examples of Competencies

Computing Fundamentals: Algorithms

Scope	Competencies
Comparison of well-known algorithms' complexity, including machine learning and statistics techniques	Provide the big-Oh time and space complexity for a given procedure.

Additional Examples of Competencies

Computing Fundamentals: Software Engineering

Scope	Competencies
Software engineering principles, including design, implementation and testing of programs.	Implement a small software project that uses a defined coding standard.

Additional Examples of Competencies

Data Privacy, Security, Integrity: Privacy

Scope	Competencies
Technologies to safeguard data privacy.	Evaluate common practices and technologies and identify the tools that reduce the risk of data breaches while safeguarding data privacy.

Additional Examples of Competencies

Machine Learning

Scope	Competencies
Problems related to model expressivity as well as availability of data, and techniques for mitigating their effects. E.g., problem of overfitting and regularization techniques for mitigating effects of overfitting; curse of dimensionality and feature selection/weighting/ reformulation techniques for mitigating effects.	Exhibit knowledge of methods to mitigate the effects of overfitting and curse of dimensionality in the context of machine learning algorithms.

Additional Examples of Competencies

Professionalism: Teamwork

Scope	Competencies
Team selection, the need to complement abilities and skills of team members on techniques for mitigating effects.	Document and justify the considerations involved is selecting a team to undertake a specific data science investigation

Timeline

- Early 2019:
  - Draft report out for comment
  - Outreach and gathering of feedback
  - Note: Initial comment period ends March 31
- Spring 2019 (f2f @ SIGCSE)
  - Begin work on next phase, including new KAs and competency details
- Spring 2019+: Outreach, presentations and information gathering
- Summer 2019:
  - Next draft to SIGCSE Education Advisory Committee (formerly Education Council)
- Fall 2019:
  - Second Draft report out for comment
- Fall 2020
  - Release final report
  - Call for interdisciplinary joint taskforce to create full curricula guidelines

Call for Example Courses

- http://dstf.acm.org/callForExamples.html

Example Feedback Received

- Recognize and promote Data Science as its own discipline
- There is a general desire for more specificity in competencies and knowledge/skill areas
- Recognize there is a strong desire for a data science minor/certificate
- Communication skills need to be emphasized more
- Professionalism needs to address new forms of professional and ethical responsibility that come with handling other people's data
- Professionalism needs explicit experiences that build skills in cross-disciplinary teamwork
- Big Data Systems should also include data representation as well as processing

Example Feedback (continued)

- Feedback suggesting topics are too general (e.g. general artificial intelligence) or not specific enough (e.g. need for automated reasoning systems and probability-based models)
- Data science does not need broad computing coverage on topics such as operating systems, networking, storage systems, etc.
- Numerous suggestions regarding specific elements that should be T1, T2, or not included. We recognize the many views of what constitutes data science.
- We recognize the need for Statistics and Mathematics Competencies, knowledge and skill level articulations. We hope that a future effort that includes appropriate representation from multiple societies will join in this effort.

Discussion

- http://dstf.acm.org
- Comments on the report?
  - Positive and negative reactions
  - Thoughts on the Knowledge Areas and Competencies?
  - Thoughts on the proposed expansion of the report?
- Additional comments?