# Crowdsourcing Classroom Observations to Identify Misconceptions in Data Science

**Prof. Ruth E. H. Wertz, Valparaiso University**

Dr. Wertz is an Assistant Professor of General Engineering at Valparaiso University, located in Valparaiso Indiana. She has earned a B.S. in Civil Engineering from Trine University, a M.S. in Civil Engineering from Purdue University, and a Ph.D. in Engineering Education also from Purdue University.

**Prof. Karl RB Schmitt, Valparaiso University**

Karl is an Assistant Professor and Director of Data Sciences at Valparaiso University (Valpo) in Indiana. He is housed in the Mathematics and Statistics Department with an affiliate appointment to the Computing and Information Sciences Department. He's run the Masters in Analytics and Modeling program since 2014, and is the founding director of Valpo's Bachelor's in Science in Data Science.

Karl specializes in data science as applied to networks and graphs. He's done work with applying network algorithms to improve genome assembly and published fundamental work in understanding K-Dense graphs. He's also very interested in finding ways to connect data science with social good, especially through the classroom and experiential learning.

Karl's teaching includes Optimization, Data Mining, Multivariable Calculus and Differential Equations. He's also designed and implemented an Introduction to Data Science course targeted at students with minimal programming experience that centers around a data-driven service learning project.

**Dr. Linda Clark, Brown University**
**Prof. Bjorn Sandstede, Data Science Initiative, Brown University**
**Dr. Katherine M. Kinnaird, Smith College**

# [WIP] Crowdsourcing Classroom Observations to Identify Misconceptions in Data Science

## Project Overview

Societal needs for converting the vast amounts of data into meaningful information drive the current demand for the field of data science. As a developing field, consensus on curricular content and learning objectives has yet to be reached, particularly weighing the disciplinary contributions of computer science, mathematics, statistics, and other domain knowledge areas. This need fueled the rapid growth of data science education training the next generation of data-centric workers. Initially, data science teaching practices drew from data science's parent disciplines (e.g., computer science, mathematics and statistics). However, because little consensus exists on the appropriate blend of these fields, pedagogical practices need to be critically evaluated for their effectiveness in the new context of data science education.

The *Investigations of Student Difficulties in Data Science Instruction* project addresses the early development of concept inventory topics which present the most difficulty for students to learn within data science. In particular, this project addresses three primary research objectives: (1) identify student misconceptions in data science courses; (2) document students' prior knowledge and identify courses that teach early data science concepts; and (3) confirm expert identification of foundational data science concepts, and their importance for introductory-level data science curricula.

During the first year of this grant we progressed on items (1) and (3). For objective (1) we developed and launched a pilot assessment, the difficulty protocol, for identifying student difficulties within data science courses. The difficulty protocol includes weekly reflective responses from faculty, teaching assistants, and students engaged in data science courses offered at the three participating institutions. For objective (3) we collected approximately 200 survey responses validating important data science concepts from the existing body of knowledge presented by the Edison Project [1]. Faculty and industry practitioners from data science and closely related fields comprised the survey respondents.

Preliminary results of our overall efforts will be presented at the ASEE National Conference and Exposition in the NSF Grantees poster session, however this paper focuses on the development and analysis of the difficulty protocol. As a work in progress, our data, analyses, and findings are not yet final and are subject to change as we progress through additional phases of the project.

## Theories of Difficulty and Misconceptions

As an emerging field, data science presents complex questions around appropriate educational objectives, resources, and curricular norms. Given the ubiquity of data science it is critical that current industrial and academic definitions of what data science is, and what it means to practice data science, be examined and embedded in early curricular offerings. Notably ahead of efforts in the United States, the European Union established and publicly discussed data science curricula

through the Edison Project [1]. The Edison Project includes a complete undergraduate curriculum including course learning objectives, goals, and more. Developed in consultation with European industrial interests and having undergone several revisions, the Edison Project developed a data science body of knowledge which serves as a foundation for a concept inventory in the field. However, there is still sparse literature of theories or investigations around learning difficulties within data science.

The identification of misconceptions and difficult concepts within data science allows researchers to begin carefully examining teaching practices in data science. For example, Freeman et al. [2] and other work by Handelsman et al. [3,4] demonstrate the positive impacts of active learning methods on student's grades in STEM classes, drawing extra attention to the significantly lower failure rates in STEM courses that use active learning methods. Similarly, opinion pieces like "The worst way to teach" by David M. Bressoud [5], advocate for educators choosing active learning methods in their classrooms over the traditional lecture-based methods. Much of this scholarship of teaching and learning work was spurred by having an effective method for consistently assessing student learning and earlier documentation of student misconceptions.

**Crowdsourcing A Difficulty Protocol**

A significant part of this research program investigates student difficulty within and misconceptions of data science in actual classrooms. Typically the best way to understand student behavior in a classroom is through direct observation, but gleaning the nuances of their behavior requires extended observations. As such, an extended, direct observation is impractical. Extensive work shows that student self-reports alone can be unreliable. Students may under- or over-report their degree of misunderstanding based on any number of external factors, or they may legitimately not know the degree of their misunderstandings relative to certain topics. Instead of relying only on student self-observations, this study uses a triangulated approach incorporating instructors, teaching assistants, and students each completing a weekly reflection. The reflection asks about the difficulties or misunderstandings experienced in the classroom during the past week. The protocol consists of five items that are tailored to the instructor, TA, and student audiences, but generally consist of:

Commented [1]: do you have any citations for self-reports?

1. What topics did you cover this week? [Open Answer]
2. What kinds of activities did your students focus on this week out of class? [Select All That Apply]
3. What questions did students raise this week? [Open Answer]
4. By observation, what concepts or processes did students struggle with? [Open Answer]
5. Which student questions were surprising to you? [Open Answer]

In the fall 2019 term, Smith College piloted the difficulty protocol with one faculty member and five students in an introductory data science course. Even in the relatively small dataset, patterns are beginning to emerge where students report difficulty around selection and context and the instructor reports center more on presumed difficulty with high-level concepts and specific application functionality. Further analysis will be presented at the conclusion of the spring semester once additional data has been collected and analyzed.

**Acknowledgements**
This material is based upon work supported by the National Science Foundation (NSF) under Grant No 1839357, 1839270, 1839259. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.

**References**

[1] EDISON: Building the data science profession; Edison Project.

[2] S. Freeman, S. L. Eddy, M. McDonough, M. K. Smith, N. Okoroafor, H. Jordt, and M. P. Wenderoth, Active learning increases student performance in science, engineering, and mathematics,
Proceedings of the National Academy of Sciences, 111 (2014), pp. 8410{8415.

[3] J. Handelsman, D. Ebert-May, R. Beichner, P. Bruns, A. Chang, R. DeHaan, J. Gentile, S. Lauffer, J. Stewart, S. M. Tilghman, and W. B. Wood, Scientific Teaching, Science, 304
(2004), pp. 521{522.

[4] J. Handelsman, S. Miller, and C. Pfund, Scientific Teaching, Macmillan, 2007. Google-Books-
ID: suf0MvxqoLQC.

[5] D. M. Bressoud, The worst way to teach, MAA Launchings, July, (2011).