# Investigations of Student Difficulties in Data Science Instruction

**Year 0 - A TRIPODS+X:EDU Grant**


**Karl R. B. Schmitt[A], Katherine M. Kinnaird[C], Ruth E. H. Wertz[A], Bjorn Sandstede[B]**

VALPARAISO UNIVERSITY (A)

BROWN UNIVERSITY (B)

(C)


## Abstract


Web-browsing histories, online newspapers, streaming music, and stock prices all show that we live in an age of data. Extracting meaning from data is necessary in many fields to comprehend the information flow. This need has fueled rapid growth in data science education aiming to serve the next generation of policy makers, data science researchers, and global citizens. Initially, teaching practices have been drawn from data science's parent disciplines (e.g., computer science and mathematics). This work begins investigating data science education in its own right by aiming to identify preconceptions students may have when they first enter a data science classroom, and what other courses from related programs are shaping their preconceptions. The investigation will focus on topics identified in the National Academy of Sciences, Engineering, and Medicine (NASEM) Report: Data Science for Undergraduates: Opportunities and Options.


## Main Objectives of Project

1. Identify student misconceptions, difficulties, and non-expert thinking about data science concepts.
2. Document data science concepts that engineering (and other) students develop outside of data science courses.
   (a) Identify courses that establish this early data science knowledge.
   (b) Evaluate quality of previous knowledge, and provide formative assessment feedback for improving and refining prior instruction.
3. Identify disconnects between core elements of data science curricula and what early career data science practitioners use in their daily work

## Motivation

Protocol for Developing a Concept Inventory[1]

1. **Establish topics.** Determine topics important to faculty through self-reflection and discussion.
2. **Identify student thinking.** Observe and interview students to understand how their thinking deviates from expert thinking.
3. **Create open-ended survey questions.** Administer these questions to entire classes of students in order to further examine issues raised in the interviews.
4. **Create forced-answer test.** Establish distractors based on actual student responses obtained in the above steps. The test should contain less than 30 questions.
5. **Validate test questions through interviews.** Reach consensus among experts about correct responses. Ensure that students interpret the questions consistently and that maladaptive student thinking results in incorrect responses.
6. **Administer and statistically analyze.** Administer the inventory to several large classes, applying statistics to account for reliability and validity.

## Planned Work

### Course & Initial Misconception Identification Work

- Identify courses, especially those in engineering, that teach concepts in data science
- Collect syllabi and other course materials from identified courses
- Survey & interview faculty teaching courses with data concepts about student misconceptions
- Survey current engineering students to identify existing data science knowledge

### Student Thinking Work

- Develop protocol for conducting interviews about misconceptions and for recording student responses to the open-ended questions
- Collect reports from instructors and TAs of data science courses using the *reflection and difficulties protocol*
- Conduct ad-hoc interviews of students to probe for data science conceptual misconceptions
- Develop open-ended questions to probe student misconceptions/knowledge
- Conduct (and record) interviews regarding the developed open-ended questions with students in at least 4 classes to further probe misconceptions
- Provide a new undergraduate research opportunity in text mining and analysis

### Work Elucidating Prior Knowledge from Engineering students

- Conduct survey of engineering students on prior knowledge of data science concepts
- Use the student misconception protocol to conduct interviews with engineering students about their pre-conceptions about data science and to witness their responses to the open-ended questions
- Collect reports from instructors and TAs of engineering courses using the *reflection and difficulties protocol*
- Conduct ad-hoc follow-up interviews of engineering students to probe for data science conceptual knowledge *before* they take a formal data science course
- Conduct (and record) interviews regarding the developed open-ended questions with students in engineering classes to further probe misconceptions
- Provide a second undergraduate research opportunity in text mining and analysis

### Industry & Early Career Practioner Survey Work

- Develop a survey to discover gaps between what academic data science programs offer students and what graduates working in industry need on a daily basis
- Disseminate survey in collaboration with industrial data science practitioners
- Gather industrial data science job descriptions
- Create a summary report of the survey responses as well as the data science job descriptions
- Provide a third undergraduate research opportunity in text mining and analysis

### Work To Synthesize Efforts to Uncovering Misconceptions

- Develop reports highlighting:
  - Diversity of courses, including engineering, teaching data science concepts and ideas
  - Common student misconceptions taking after traditional data science courses
  - Common student misconceptions taking after engineering and before data science courses
- Create data science teaching resources, including
  - Proposed list of learning outcomes for data science curricula and programs
  - Example syllabi for traditional data science courses and courses adding elements of data science
  - Collection of existing materials that are particularly effective at teaching data science concepts that are most often misunderstood and/or left out of data science programs

## Timeline of Activities

| Year | Spring AY | Summer | Fall AY |
|---|---|---|---|
| Year 1 2019 | • Practitioner/Industry survey<br>• Prototyping of reflection & difficulty protocol in PI classes | • Finalize reflection & difficulty protocol<br>• Text analysis/data mining research on Practitioner survey | • TA and Instructor data collection training sessions<br>• Distributed, *in vivo* data collection<br>• Ad-hoc follow-up student interviews |
| Year 2 2020 | • Distributed, *in vivo* data collection<br>• Ad-hoc follow-up student interviews | • Development of open-ended questions<br>• UG research opportunity in text mining on interview notes | • Distributed, *in vivo* data collection<br>• Collecting student thinking on open-ended questions<br>• Digitizing of responses |
| Year 3 2021 | • Distributed, *in vivo* data collection<br>• Collecting student thinking on open-ended questions<br>• Digitizing of responses | • Final reports and dissemination produced<br>• UG research opportunity in text mining on interview notes | • Distribute reports through the TRIPODS, BDH, Project NExT, and other professional networks<br>• Present at major disciplinary, topical, and regional conferences. |

## Evaluation Plan

| Tasks | Outcomes | Methods |
|---|---|---|
| Conduct Process Evaluation | • Improve communication & collaboration<br>• Assess progress in meeting goals<br>• Integrate results of annual evaluations into subsequent activities for continuous feedback and improvement | • Quarterly teleconferences with PIs and Annual meeting with PIs<br>• Ad hoc email communications<br>• Sharing of documents and archiving data |
| Create Logic Model | • Ensure that key project indicators and outcomes are being assessed<br>• Support project team in obtaining appropriate approvals from IRB | • Conduct annual mapping from evaluation results to Logic Model indicators for NSF Report<br>• Update Logic Model annually |
| Develop & Administer Interview Protocol | • Evaluate PI & stakeholder engagement, attitudes, and feedback on progress and approaches | • Develop interview protocol for PIs and stakeholders<br>• Design methods, informed consent, etc.<br>• Conduct interviews<br>• Provide interview analysis summaries |
| Develop Dissemination Evaluation Tools | • Assess effectiveness of dissemination efforts in engaging educators, increase understanding, and catalyze change in data science instruction | • Develop presentation evaluation tools<br>• Develop professional development evaluation tool<br>• Report outcomes & impact in final summative report(s) |
| Evaluate Annual PI Meetings | • Assess engagement of stakeholders<br>• Collect insights into behavioral change that is possible or desired<br>• Provide project feedback to make changes and improvements to process and goals. | • Participate remotely with web conferencing<br>• Analyze audio transcripts<br>• Provide feedback in annual report |
| Provide Annual Evaluation Report | • Provide annual formative assessment and insights<br>• Assess progress in meeting goals<br>• Provide course corrections & recommendations | • Conduct formative assessment in first two years and provide annual reports<br>• Conduct summative assessment in year 3 and provide final summative report |

The project will be externally evaluated by Peak Research (LoC 2). The evaluation will utilize mixed-methods design, including telephone interviews, behavioral observation, and survey research. Both formative and summative evaluation will be conducted.

## Broader Impacts

NASEM's report *Data Science for Undergraduates: Opportunities and Options* recommends that "...institutions should ... work together to develop professional approaches to evaluation...shar[e] measurement and evaluation frameworks, data sets, and culture ... [and] align educational evaluation with market impacts" (Rec. 5.3) [2].

### Summary of Broader Impacts

- Connecting academic lens of data science to industry lens of data science through an explicit collection of disconnects between the two
- Gathering common misconceptions will help focus the teaching of various data science concepts by illuminating which concepts are currently the most misunderstood by students
- Better understanding of common misconceptions will facilitate diversity, inclusion, and accessibility in the discipline
- Providing a necessary step towards creating a standardized assessment tools in data science could quantify the value of various data science educational programs
- A clearly defined scope for data science will support the sweeping creation of educational undergraduate opportunities (courses, programs, etc.) with potential impacts in graduate education
- Grant activities will establish new interdisciplinary collaborations and consensus for data science research (especially educational research and engineering education research)

## Intellectual Merit

An extensive and repeated study of student misconceptions within data science is an important next step in data science education research. As a core topic in the scholarship of teaching and learning, this knowledge will help educators develop more effective teaching materials to remedy common misconceptions. This follows the National Research Council's (NRC) recommendation that "more studies measure outcomes other than test scores and course performance" [3].

### Summary of Intellectual Merits

- Formal documentation and study of student misconceptions within data science
- Investigation of inter- and intra- disciplinary transfer of data skills
- Design and testing of a distributed, *in vivo* qualitative data collection method

## References


1. Adams, W.K. and Wieman, C.E. 2011. Development and Validation of Instruments to Measure Learning of ExpertLike Thinking. International Journal of Science Education. 33, 9 (Jun. 2011), 12891312. DOI:https://doi.org/10.1080/09500693.2010.512369.
2. Committee on Envisioning the Data Science Discipline: The Undergraduate Perspective et al. 2018. Data Science for Undergraduates: Opportunities and Options. National Academies Press.
3. Singer, S.R. et al. 2012. Discipline-based education research: Understanding and improving learning in undergraduate science and engineering. National Academies Press.