1. DISCUSSANT SUMMARY: TOPIC 3 – THE INFLUENCE OF DATA SCIENCE ON THE SCHOOL CURRICULUM AND INTRODUCTORY STATISTICS COURSES

Discussant: Adam Molnar & Chair: Nicholas J. Horton
Oklahoma State University, USA & Amherst College, USA
adam.molnar@okstate.edu

PRESENTERS

	Title	Presenter/Co-Author(s)
Long paper	*Statistical thinking for the era of big data and artificial intelligence: Toward understanding sustainability trends and issues for the future society*	Orlando González (Thailand)
Long paper	*International Data Science in Schools Project*	Neil Sheldon (UK)
Long paper	*Data science education in secondary school: How to develop statistical reasoning when exploring data using CODAP and Jupyter Notebooks*	Daniel Frischemeier (Germany)
Long paper	*Data science education in secondary schools: Teaching and learning decision trees with CODAP and Jupyter Notebooks as an example of integrating machine learning into statistics education*	Yannik Fleischer (Germany) / Rolf Biehler (Germany)

PRELIMINARY RESEARCH QUESTIONS

- How do we prepare people to cope with the complexity of big data?
- What knowledge, skills and dispositions are required in data science to develop data acumen?
- What is the role of statistics, computation and domain knowledge in data science curriculum?
- What are the ways to engage students in studying data science?
- What are the challenges for integration of data science in the school curriculum/undergraduate statistics courses or designing a data science curriculum at school level/undergraduate statistics?
- What are the ways to support teachers/instructors implementing aspects of data science in schools/at the tertiary level?

KEY DISCUSSION THEMES

Topic 3 contained two papers with a top-down approach and two papers with a bottom-up approach. In session 1, González and Sheldon started from the big picture and created curriculum guidelines for data science. Both reimagined the PPDAC cycle (Wild & Pfannkuch, 1999) by increasing the role of data exploration and not starting with a fixed question. Discussants noted that data exploration could lead to fewer naïve questions such as “is group A better than group B?” and more causal questions about patterns and associations. It was pointed out that these proposals were essentially exploratory data analysis, a well-established idea for many years (Tukey, 1977), yet rarely included in school curriculum.

The bottom-up Paderborn project, presented in session 2 by Frischemeier, Fleischer, and Biehler, implemented prediction-based data science curriculum in an elective Grade 12 computer science course. The PPDAC cycle was used in the first data science unit, although more consideration was given to the use of computer tools. The second unit contrasted two prediction algorithms, explainable decision trees and unexplainable neural networks, while predicting if German high school students played online games. In design interactions with computer science and mathematics teachers, computer science teachers gave more positive feedback.

Both sessions included discussion about academic culture. One discussion contrasted the two cultures of Breiman (2001), inferential understanding (more common in statistics) against predictive efficiency (more common in data science). Although Breiman divided the topics, multiple commenters believed that data science curriculum should include both prediction and understanding. Data science includes more prediction than statistics, however, as illustrated in the Paderborn project.

The computer science teacher comment illustrated another comparison, between two cultures with stronger division – structure-based mathematics certainty against domain-based statistics uncertainty. These cultures have different patterns of thought. Participants discussed this difference in Topic 3 and elsewhere during the conference. Multiple commenters wondered if mathematics teachers were best suited to implement data science, because of mathematics' emphasis on structure over data, although others argued that the most common statistics home of mathematics remained best.

Shorter discussions occurred about aspects of preparing people to deal with big data. Participants contrasted data literacy for all informed citizens, data skills for scientists in other fields, and data science as a field of study. Low-cost accessible technologies such as CODAP, Juypter, iNZight, Python, and R were mentioned. Quantitative content was included, such as classification vs. regression, transparent vs. nontransparent machine learning, and cost functions. Participants also brought up less quantitative aspects of data science in schools, such as helping students pose good questions, critical thought about procedures, and teacher support in less well-off areas.

REFERENCES

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science, 16*, 199–231.

Tukey, J. W. (1977). *Exploratory Data Analysis.* New York City, NY: Pearson.

Wild, C. J., & Pfannkuch, M. (1999). Statistical Thinking in Empirical Enquiry. *International Statistical Review*, *67*(3), 223–265.