AMERICAN STATISTICAL ASSOCIATION
Promoting the Practice and Profession of Statistics®

ASA Statement on *The Role of Statistics in Data Science and Artificial Intelligence*
August 4, 2023

Data science and artificial intelligence (AI) have, in recent years, captured the attention of a world audience for their spectacular contributions in a wide range of scholarly research and commercial endeavors. Whether it be the development of self-driving cars, machines to recognize speech and generate human-like text, or technology that can accurately detect cancer, the success of data science and AI is all around us and will continue to affect scientific innovation and how we live our lives. The ability to address challenging questions with complex data combined with thoughtful methods is largely the fruit of the innovative and entrepreneurial spirit that characterizes these burgeoning areas. Nonetheless, the interdisciplinary nature of data science and AI means a substantial collaborative effort is needed, and that statisticians—who themselves are data scientists—should be extensively involved in data science and AI initiatives to realize their full potential for productivity, innovation, and problem-solving.

In the past 20 years, data science and AI have rapidly evolved, fueled by the explosion of data and advancements in computing power. This evolution has been marked by the development of sophisticated machine learning (ML) algorithms, deep learning neural networks, and generative AI—including large language models—that have revolutionized industries such as health care, finance, and marketing and generated new areas of research in statistics and computing. The big tent of statistics has grown massively bigger, with the role of statisticians in this new and evolving world requiring adaptation. The boundaries between statistical and computational methods have become more blurred, given the advances in ML and deep learning algorithms, which often require a deep understanding of both statistical theory and computer science. This has led to a rethinking of the traditional role of statisticians in data science and a recognition that statisticians need to be equipped with a wider range of skills and expertise to remain effective and, indeed, relevant.

Data science and AI rely heavily on statistics, mathematics, and computer science to gain knowledge from data. These fields produce tools to interact with data, provide effective and meaningful summaries or inferences from data, and—in the case of AI—develop systems that can carry out tasks ordinarily requiring human intellectual processes. It is difficult to think of an area of science, industry, commerce, or government that in some way does not require such knowledge extraction from data on a regular basis. However, it is unrealistic to expect one person working in data science or AI to possibly be an expert in all relevant areas. A data scientist or AI scientist must fluently interact with the disciplines and areas of research the data

analyses attempt to address and must be able to work with those specializing in other areas within data science and AI in which they lack specialization. Data science and AI endeavors require maximum multifaceted collaboration to fully realize their potential.

Statistics plays a central role in data science and AI, especially in the areas of ML and deep learning. Framing questions statistically allows leveraging data resources to extract knowledge and obtain better answers. The central dogma of statistical inference, that there is a component of randomness in data, enables researchers to formulate questions in terms of underlying processes, quantify uncertainty in their answers, and separate signal from noise. A statistical framework allows researchers to distinguish between causation and correlation, and thus to identify interventions that will cause changes in outcomes. It also allows them to establish methods for prediction and estimation, to quantify their degree of certainty, and to do it all using algorithms that exhibit predictable and reproducible behavior. In this way, statistical methods aim to focus attention on findings that can be reproduced by other researchers with different data resources. Simply put, statistical methods enhance researchers' abilities to accumulate knowledge.

Contributing to the responsible development of data science and AI systems requires a sustained and substantial collaborative effort with researchers knowledgeable in areas not typically in the purview of statisticians, including those with expertise in data organization, distributed computation, and model lifecycle management. Statisticians must work with them, learn from them, and teach them. Engagement must occur at all levels—with individuals, groups of researchers, academic departments, and the profession as a whole. New problem-solving strategies are needed to develop end-to-end data science and ML operations pipelines, from raw data collection and management to model monitoring/retraining and governance to user-friendly implementations of principled statistical methods and the communication of substantive results. Statistical education and training must continue to evolve—the next generation of statistical professionals needs a broader skill set and must be more able to engage with software and ML/deep learning engineering experts. Effective teaching of data science and AI requires more than just a mechanical explanation of algorithms; it must be done in the context of the entire data science process. This includes understanding how to formulate a research question, collect and preprocess data, choose appropriate statistical methods and models, and interpret and communicate results in a meaningful way. While capacity is increasing within existing and innovative new degree programs, more is needed to meet the massive, expected demand. The next generation must include more researchers with skills that cross the traditional boundaries of statistics; there will be an ever-increasing demand for such multifaceted experts.

Working with statisticians, departments of statistics and data science, and other professional societies, the American Statistical Association (ASA) is well positioned to help formulate discussion about the role of statistics in data science and AI, navigate the way forward in this quickly evolving environment, and provide forums for communication and collaboration among data scientists and AI scientists, including statisticians and non-statisticians alike. The future of

data science and AI is uncertain, but one thing is clear: the field will undoubtedly continue to evolve rapidly and have a profound impact on society. As a result, the role of statisticians will also be subject to change and expansion, as they must adapt to new technologies and tools while continuing to provide expertise in traditional areas of statistics such as uncertainty quantification, sampling design, and causal inference. The need for interdisciplinary collaboration and a diverse range of skills will become increasingly important for statisticians to remain relevant in this dynamic and ever-changing field. The ASA aims to facilitate collaboration among all data-driven fields, and thus enable workers in these areas to achieve more than they could on their own.

For further insight about this topic, read an earlier statement and blog post about the role of statistics in data science.