# Computer Science Defense

## *Attention-enhanced Deep Learning Models for Data Cleaning and Integration*

Jing Zhang
Emory University

**Abstract:** Data cleaning and integration is an essential process for ensuring the accuracy and consistency of data used in analytics and decision-making. Schema matching and entity matching tasks are crucial aspects of this process to merge data from various sources into a single, unified view. Schema matching seeks to identify and resolve semantic differences between two or more database schemas whereas entity matching seeks to detect the same real-world entities in different data sources. Given recent deep learning trends, pre-trained transformers have been proposed to automate both the schema matching and entity matching processes. However, existing models only utilize the special token representation (e.g., [CLS]) to predict matches and ignore rich and nuanced contextual information in the description, thereby yielding suboptimal matching performance. To improve performance, we propose the use of the attention mechanism to (1) learn the schema matches between source and target schemas using the attribute name and description, (2) leverage the individual token representations to fully capture the information present in the descriptions of the entities, and (3) jointly utilize the attribute descriptions and entity descriptions to perform both schema and entity matching.

Tuesday, March 28, 2023, 12:00 pm
Modern Language 219

Computer Science
Emory University