Abstract: Enterprise data is usually scattered across departments and geographic regions, and it is often dirty and inconsistent. Data scientists spend most of their time finding, preparing, integrating, and cleaning relevant datasets. I will describe our current efforts to ease the pain of data scientists in our Data Civilizer project. Key components of Data Civilizer include data discovery, data cleaning, data transformation, and entity resolution and consolidation. I'll first briefly explain how through data profiling, indexing, and semantic elicitation, we built a data discovery component. I'll then talk about the deep learning-based entity resolution component. Finally, I'll describe how to detect a special type of data errors, namely disguised missing values, which turned out to be quite frequent in various proprietary and open data.

Bio: Mourad Ouzzani is a Principal Scientist with the Qatar Computing Research Institute at Hamad Bin Khalifa University, Qatar Foundation. Before joining QCRI, he was a research associate professor at Purdue University. Since joining in 2011, Mourad has played a key role in establishing the data analytics group within QCRI. Mourad's research interests lie in the fields of data management and analytics with a focus on data integration and data cleaning. He is the project lead of Rayyan, the systematic reviews web and mobile app, that now serves more than 18k users worldwide. His work has led to numerous publications in top tier venues including PVLDB, TKDE, SIGMOD, and ICDE. Mourad has been PI or CoPI in more than 15 grant proposals funded by NSF, NIH, DHS, as well as other funding agencies. Mourad received two Seeds of Success Awards from Purdue University. He hold a PhD from Virginia Tech, and a BSc and MSc from USTHB, Algiers, all in computer science.