SORS: "Explainability for Machine Learning Models: From Data Adaptability to User Perception"

Abstract: This presentation explores the generation of local explanations for already deployed machine learning models, aiming to identify optimal conditions for producing meaningful explanations. The primary goal is to develop methods for generating explanations faithful to the underlying model and comprehensible to the users. The presentation is divided into two parts. The first introduces a novel approach for evaluating the suitability of linear explanations to approximate a model. The second part focuses on user experiments to assess the impact of three explanation methods and two distinct representations. These experiments measure how users perceive their interaction with the model in terms of understanding and trust. This research aims to contribute to a better explanation generation, with potential implications for enhancing the transparency, trustworthiness, and usability of deployed AI systems.

Julien Delaunay

Short Bio: Julien Delaunay is currently pursuing his Ph.D. in Computer Science at Inria Rennes, France, with a focus on explainability, natural language processing, and human-computer interaction. He is actually about to defend his thesis on the 20th of December. He had the privilege of being supervised by Christine Largouët and Luis Galarraga. Additionally, he has also been a research visitor at Aalborg University and pursued studies at Sherbrooke University during his master's.

Speakers

Speaker: Julien Delaunay, PhD student at Inria Rennes. Visiting researcher at Aalborg University
Host: Davide Cirillo, Machine Learning for Biomedical Research, Life Sciences, BSC

Barcelona Supercomputing Center - Centro Nacional de Supercomputación