BOSTON UNIVERSITY ACADEMY

Title: Accessible Machine Learning Through Data Democratization

Author: Justin Arhonson

Abstract:

Accurate interpretations of genetic variants can lead to quick interventions and life saving diagnoses. Unfortunately, many laboratories produce conflicting interpretations of genetic variants. Since both interpretations cannot be correct, some patients will receive improper care. Furthermore, some discovered variants have no definitive interpretations, limiting clinical treatment. The National Institutes of Health's (NIH) ClinVar documents each interpretation that participating laboratories submit to the database and releases them to the public. This provides a freely available, curated data source for both doctors and researchers that has the potential to fuel important and impactful analytics. The aim of XMLToML is to prepare the ClinVar documents for machine learning analyses. The value of the project is twofold – it both provides analytics on the ClinVar data and drastically shortens the time required to perform machine learning on other datasets processed by XMLToML. The generic nature of XMLToML can fast-track XML reformatting projects.