TRAIL Summer Workshop' 25
Project Proposal

Full Name of Team Leader	Sébastien Piérard
Project Title	Let's Tile together: Implementation of a Python library for performance analysis
Profile of the Team Leader(s)	The team leader holds a master's degree in Computer Science Engineering. Over the past several years, he has led research in the field of performance analysis and evaluation, with a particular focus on computer vision methods. His expertise is demonstrated through multiple peer-reviewed publications on the topic (see Bibliographic References).
Abstract	In AI communities, as well as in many other research domains, rigorous evaluation of new methods is essential. A key component of the evaluation process is the possibility to compare performances of said methods in order to rank them. Until recently, doing so while accounting for application-specific preferences remained challenging. The literature often relies only on one or two standard scores to compare methods in a given domain, which fails to capture the nuances of application-specific requirements. However, an innovative visualization tool, named the Tile, has been recently developed by TRAIL researchers, based on strong theoretical foundations. For problems similar to two-class classification, the Tile organizes an infinity of performance scores (that have been demonstrated suitable for ranking) in a single 2D map. These scores, called ranking scores, include the accuracy, the true positive rate, the positive predictive value, Jaccard's coefficient, and all F-scores, to cite only a few. Alongside this tool, several interpretative flavors adapted to various user's needs have been proposed. These flavors map different values onto the Tile, such as scores, correlations, or rankings, allowing users to analyze, compare, and rank methods more effectively. This tool seeks to open new perspectives on how to perform informed method selection tailored to specific needs and applications. The aim of this project is to develop a first version of an open-source library that will be easy to use by all researchers wishing to produce Tiles for analyzing, comparing and ranking methods on the basis of their performances. The library will have the flexibility to handle a wide range of flavors. This project will be an opportunity for participants to discover the Tile in all its details and to share their talents in architecture design, software development, and documentation.

Project Objectives	During the Summer Workshop, our objective is to implement and release a software library dedicated to performance analysis and evaluation, which is a cornerstone in AI. To date, we have published several papers that lay the theoretical foundations of performance-based ranking [1], introduced a graphical tool for analyzing two-class classification performance called the "Tile" [2], provided a guide to understanding performances of two-class classifiers using the Tile and its various flavors [3], and proposed a methodology for evaluating strategies that predict rankings on unseen domains using the Tile with new flavors [4]. However, only limited code (a single Jupyter notebook accompanying [3]) has been made publicly available. As a result, the Tile is not widely and easily accessible to practitioners and researchers who could benefit from its use. During the workshop, we aim to address this gap by developing a clean, documented, easy-to-use software Python package that implements the core functionalities of the Tile. The library will be made open source, with the intention to submit a companion paper to, e.g., the Journal of Open Source Software (JOSS). We also plan to provide this work to the TRAIL Factory. By the end of the workshop, we expect to have a functional and documented software library, and a draft for publication.
Project Dataset	Not applicable. Yet, the data used in [3] and [4] is available and can be used to validate the library. This data will allow us to verify that the implemented functionalities reproduce the results and visualizations presented in the papers for the various Tile flavors. The data consists of a JSON file that, for each method being ranked, provides the corresponding confusion matrix values (true positives, true negatives, false positives, false negatives).
Background Information	Theoretical foundations for performance-based rankings, grounded in probability and order theories, have recently been introduced by Piérard et al. [1] through a rigorous axiomatic framework. Their first axiom states that any performance-based ranking should be derived from a preorder on performances, which ensures the stability of the rankings w.r.t. insertions and deletions of ranked entities. Their second axiom gives compatibility conditions between the preorders and the considered task, modeled by a random variable called satisfaction. Their third axiom gives compatibility conditions between the preorders and known properties about the evaluation (i.e., the mapping of the entities to their performances). These axioms are guardrails to guarantee meaningful rankings while leaving the flexibility to adjust the rankings w.r.t. application-specific preferences. Additionally, it introduces ranking scores that satisfy these axioms, parameterized by a random variable called importance, which allows for considering application-specific preferences.

	In [2], a spatial organization of the ranking scores in a 2D square map, called the Tile, is proposed for the particular case of two-class classification. The paper also studies properties of the Tile from a theoretical perspective. A hitchhiker's guide to understand the performances of two-class classifiers is provided in [3]. It presents four scenarios showcasing different user profiles: a theoretical analyst, a method designer, a benchmarker, and an application developer. For each profile, an interpretative flavor of the Tile is introduced. This guide is illustrated by ranking and analyzing the performances of 74 semantic segmentation models through the lens of the four scenarios. Finally, [4] presents an original methodology to evaluate strategies predicting rankings in a new domain based on assessments on known domains, without having to carry out new and costly evaluations. This is performed in a leave-one-domain-out fashion, for a variety of application-specific preferences. Its use is illustrated with 30 strategies to predict the rankings of 40 entities (unsupervised background subtraction methods) on 53 domains (videos). Furthermore, a new Tile flavor is introduced to depict the performance of strategies predicting rankings of two-class classifiers on unseen domains, and to compare them.
Bibliographic References	[1] Sébastien Piérard, Anaïs Halin, Anthony Cioppa, Adrien Deliège, and Marc Van Droogenbroeck, "Foundations of the theory of performance-based ranking," in IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, June 2025. [2] Sébastien Piérard, Anaïs Halin, Anthony Cioppa, Adrien Deliège, and Marc Van Droogenbroeck, "The Tile: A 2D map of ranking scores for two-class classification," arXiv, vol. abs/2412.04309, 2024. [3] Anaïs Halin*, Sébastien Piérard*, Anthony Cioppa, and Marc Van Droogenbroeck, "A hitchhiker's guide to understanding performances of two-class classifiers," arXiv, vol. abs/2412.04377, 2024. (* authors contributed equally) [4] Sébastien Piérard, Adrien Deliège, Anaïs Halin, and Marc Van Droogenbroeck, "A Methodology to Evaluate Strategies Predicting Rankings on Unseen Domain", in IEEE International Conference on Multimedia and Expo Workshops (ICMEW), Nantes, France. June 2025.
Detailed Work Plan	Our project is organized into the following phases. 0. Introduction to the Tile and its flavors This preliminary phase aims at presenting the Tile, its flavors, and the theoretical foundations of performance-based ranking to the members of the project. No background in performance analysis is required to join this project. 1. Definition of use cases and Tile flavors The first step involves identifying the use cases and corresponding Tile flavors that the library should support. As a starting point, we will rely on the Tile flavors introduced in [3] and [4], as well as additional flavors from ongoing research works (e.g., flavors for multi-

	domain or multi-class scenarios). Importantly, we welcome and encourage new use cases and corresponding tile flavors proposed by other participants during the workshop. 2. Design of software architecture Next, we will define the software architecture of the library. Our goal is to make the architecture as modular and generic as possible, allowing for the straightforward integration of new Tile flavors. The library should be easy to use and offer a variety of possibilities. 3. Implementation and documentation The third step consists in the implementation of the library in Python, using an object-oriented approach and relying on standard libraries such as Matplotlib, SciPy, and NumPy. In parallel, we will develop a comprehensive documentation to ensure accessibility and ease of adoption. 4. Validation and testing We will then test the library and verify that it reproduces the results presented in [3] and [4]. 5. Presentation and publication Finally, we will prepare a presentation, push code on GitHub, create a package for PyPI, and prepare a submission to a journal such as JOSS.
Other Remarks	We welcome people from all backgrounds and expertise to join the project, whether you are passionate about development and coding, interested in the evaluation of AI methods, or simply keen to explore performance analysis. Everyone can connect to this work in one way or another, regardless of research interests. We value the diversity of skills and perspectives, and look forward to collaborating with those who are willing to contribute in any capacity.

Fig. 1 (source: [4]): Two equivalent readings of the Tile: a map of application-specific importances (left) and a map of scores to induce meaningful performance-based rankings (right).

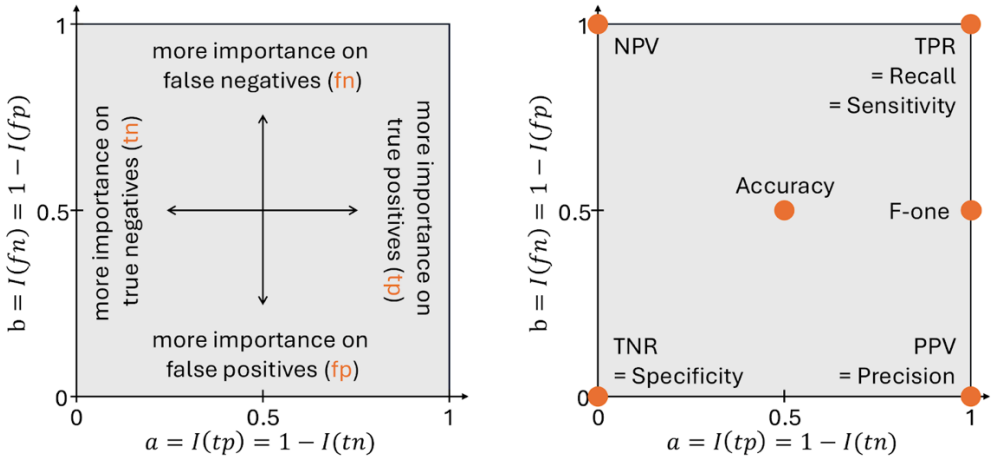

Fig. 2 (source: [3]): Example of Tile flavors corresponding to four user profiles.