LIÈGE université
Sciences Appliquées

TRAIL
TRUSTED AI LABS
BY DIGITALWALLONIA4.AI / SPW-RECHERCHE

Novelties in Performance Analysis

PhD m'eating — Montefiore Institute — June 6th 2025

Sébastien Piérard and Anaïs Halin

Part 1:

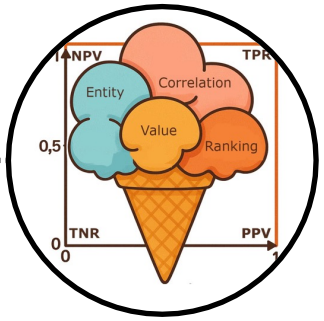
Preface

This presentation concerns original research carried out at the Montefiore Institute of the University of Liège. And, this is the bust of Georges Montefiore-Levi (1832-1906), sculpted by Thomas Vinçotte in 1904.

Menu

1
Presentation

Novelties in
Performance Analysis



2
Speakers

Sébastien Piérard

Anaïs Halin

3
Papers

Foundations of the Theory of Performance-Based Ranking

The Tile: A 2D Map of Ranking Scores for Two-Class Classification

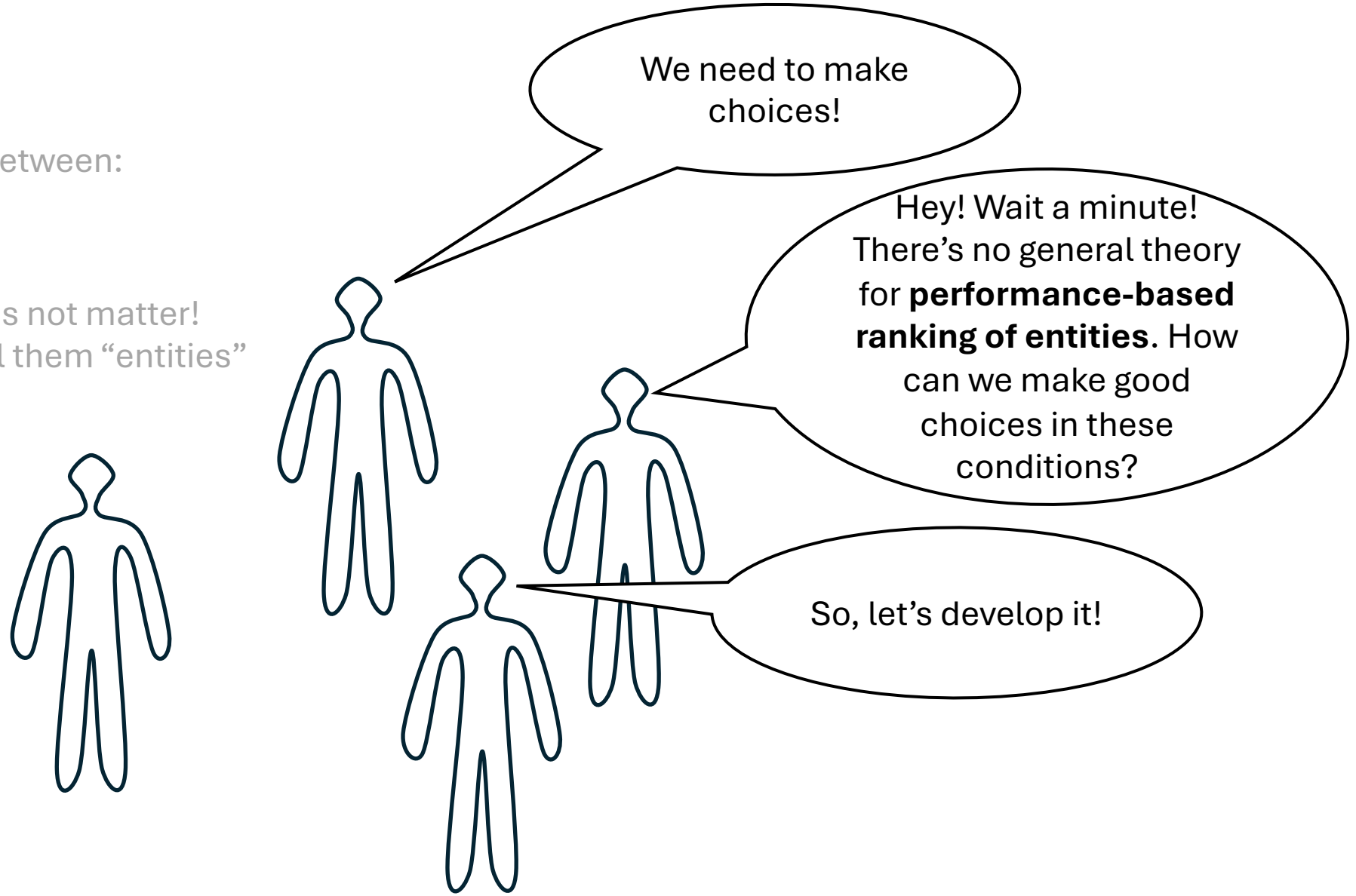
A Hitchhiker's Guide to Understanding Performances of Two-Class Classifiers

Everyday, millions of people need to make choices …

For engineers, the choices are between:

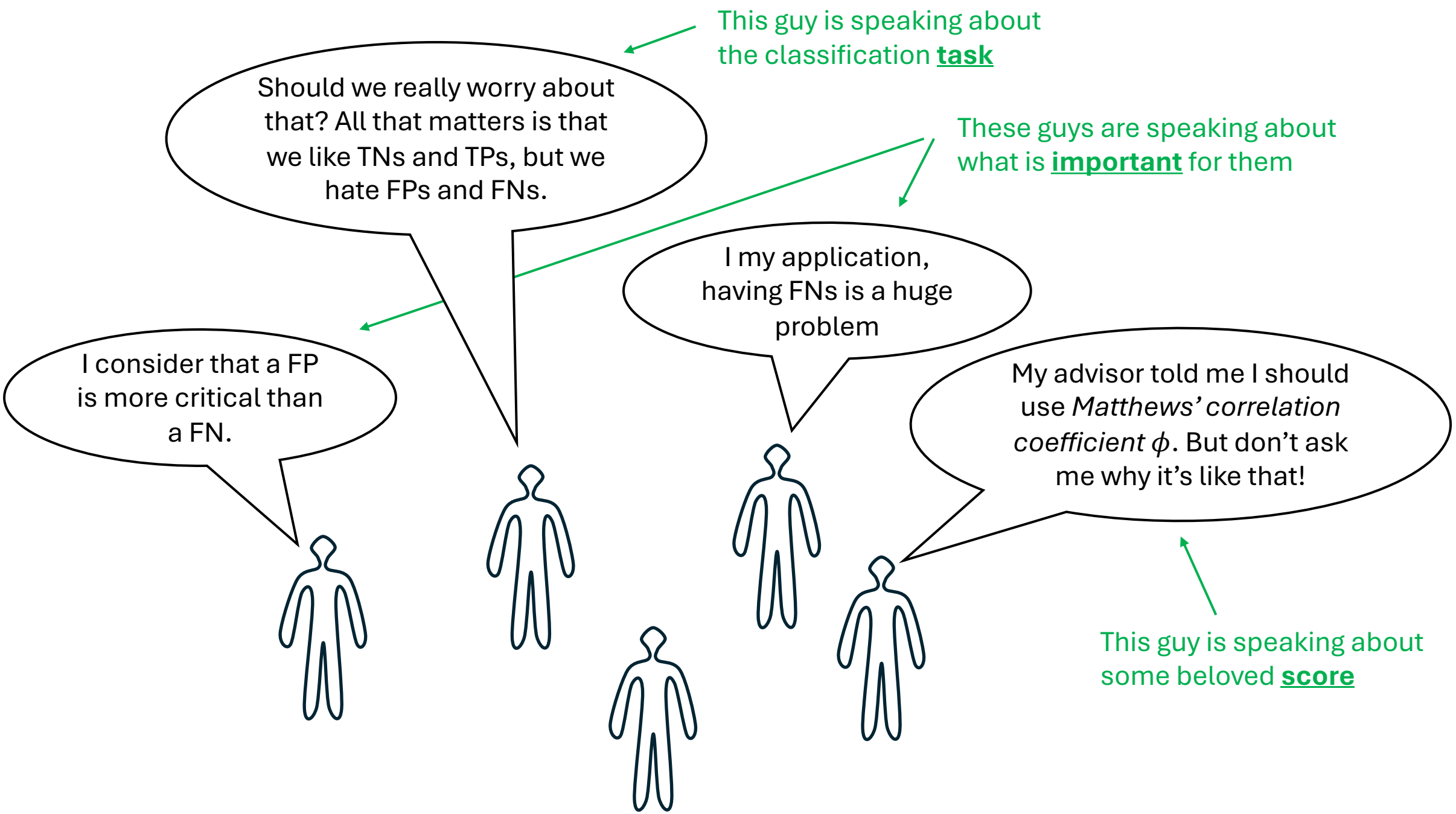
- devices;
- algorithms;
- methods;
- procedures;
- solutions;
- models;
- etc.

That does not matter!
Let's call them "entities"



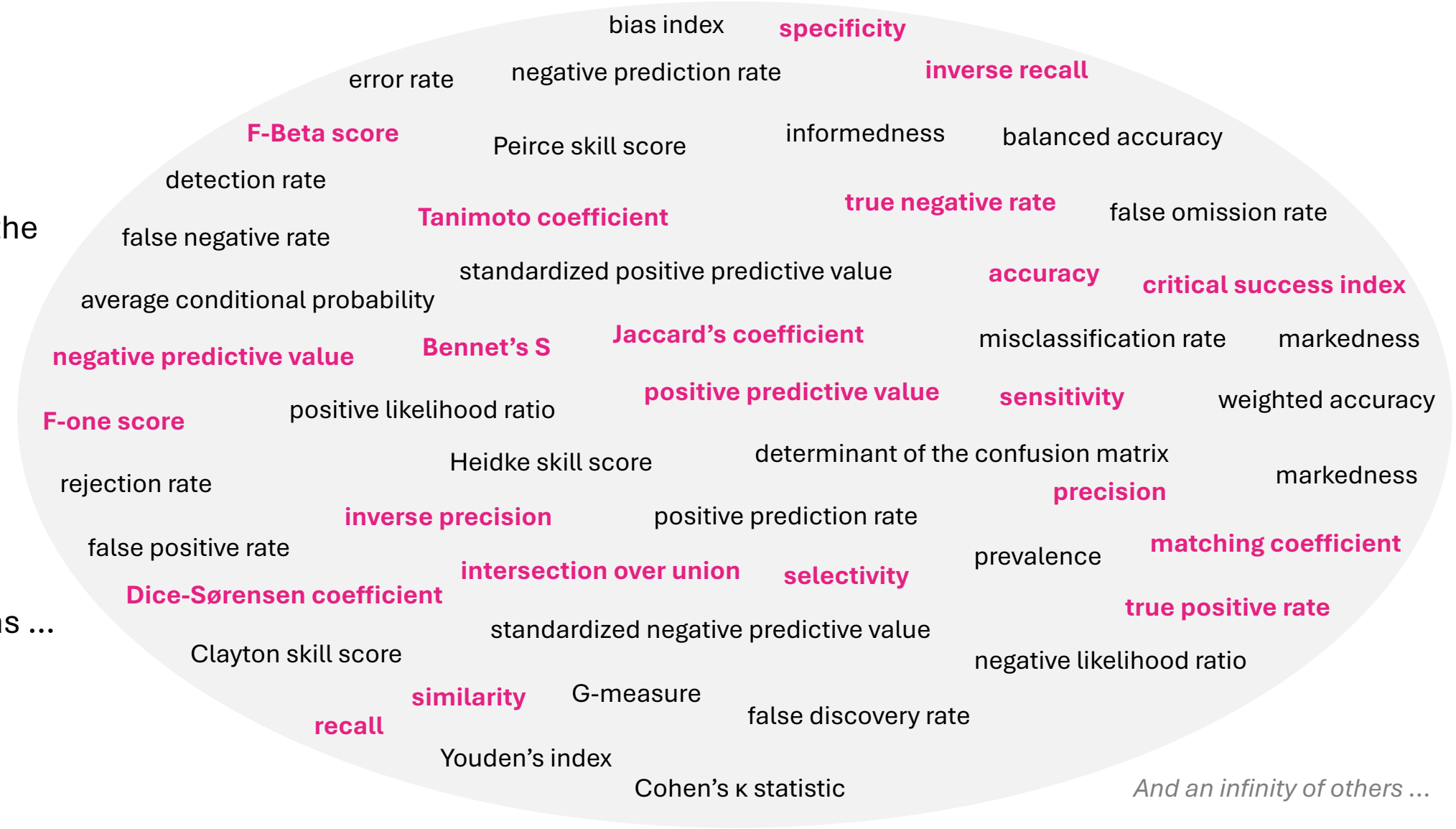
Example of problem: classification

- There is a set $\mathbb{C}$ of classes, $|\mathbb{C}| \geq 2$.
- Three types:
  1. In *soft* classification, we predict a confidence value for each class.
  2. In *binary* classification, we predict a Boolean value for each class.
  3. In *crisp* classification , we predict a class:
     - there is a ground-truth class Y;
     - there is a predicted (estimated) class $\hat{Y}$;
     - and we are *satisfied* if and only if $Y = \hat{Y}$.
- Convention in two-class classification:
  - $\mathbb{C} = \{c_-, c_+\}$: one class is called *negative* (c_-) and the other is called *positive* (c_+).
- We will speak a lot of two-class crisp classification:
  - A *true negative* (tn) occurs when $Y = c_-$ and $\hat{Y} = c_-$.
  - A *false positive* (fp) occurs when $Y = c_-$ and $\hat{Y} = c_+$.
  - A *false negative* (fn) occurs when $Y = c_+$ and $\hat{Y} = c_-$.
  - A *true positive* (tp) occurs when $Y = c_+$ and $\hat{Y} = c_+$.


This guy is speaking about the classification **task**
Should we really worry about that? All that matters is that we like TNs and TPs, but we hate FPs and FNs.
These guys are speaking about what is **important** for them
I my application, having FNs is a huge problem
I consider that a FP is more critical than a FN.
My advisor told me I should use *Matthews' correlation coefficient* ϕ. But don't ask me why it's like that!
This guy is speaking about some beloved **score**

Lost in an ocean of scores ?

This is a small part of the scores produced by the human imagination, during a few decades, and just for the two-class classification problems ...

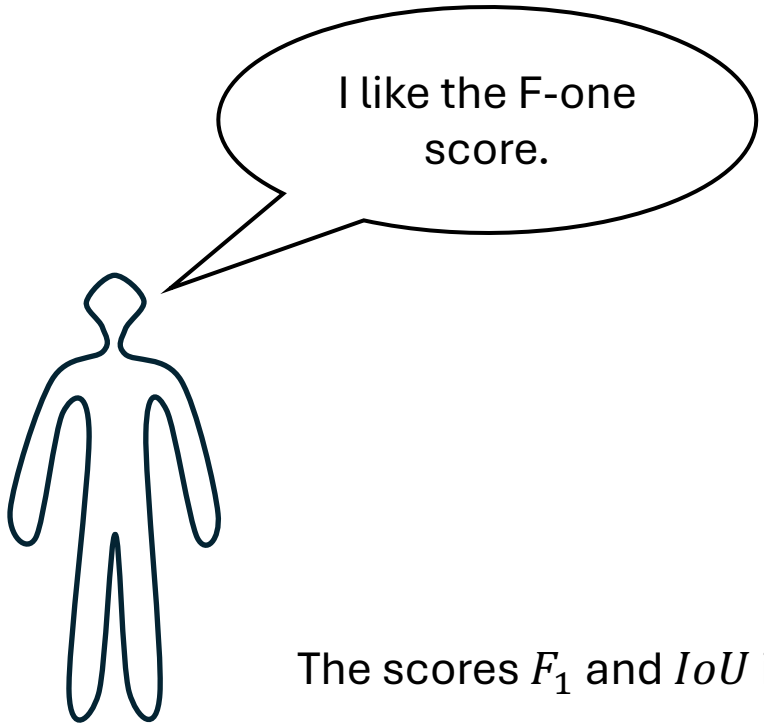


And an infinity of others ...

Some scores are equivalent: example

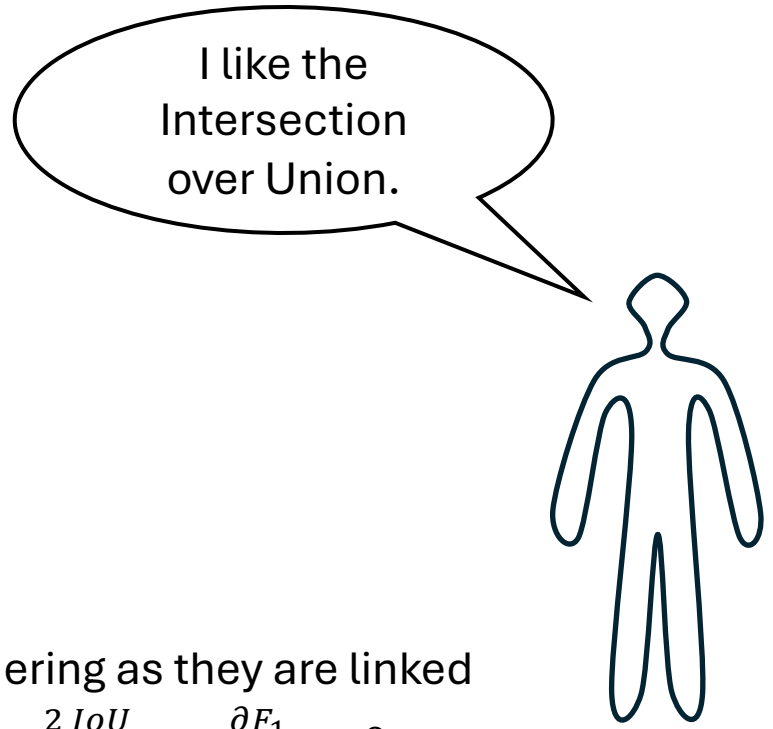
The *F-one score* is defined as the harmonic mean between the positive predictive value and the true positive rate:

$$F_1 = \frac{2\,P(\{tp\})}{P(\{fp, tp\}) + P(\{fn, tp\})}$$



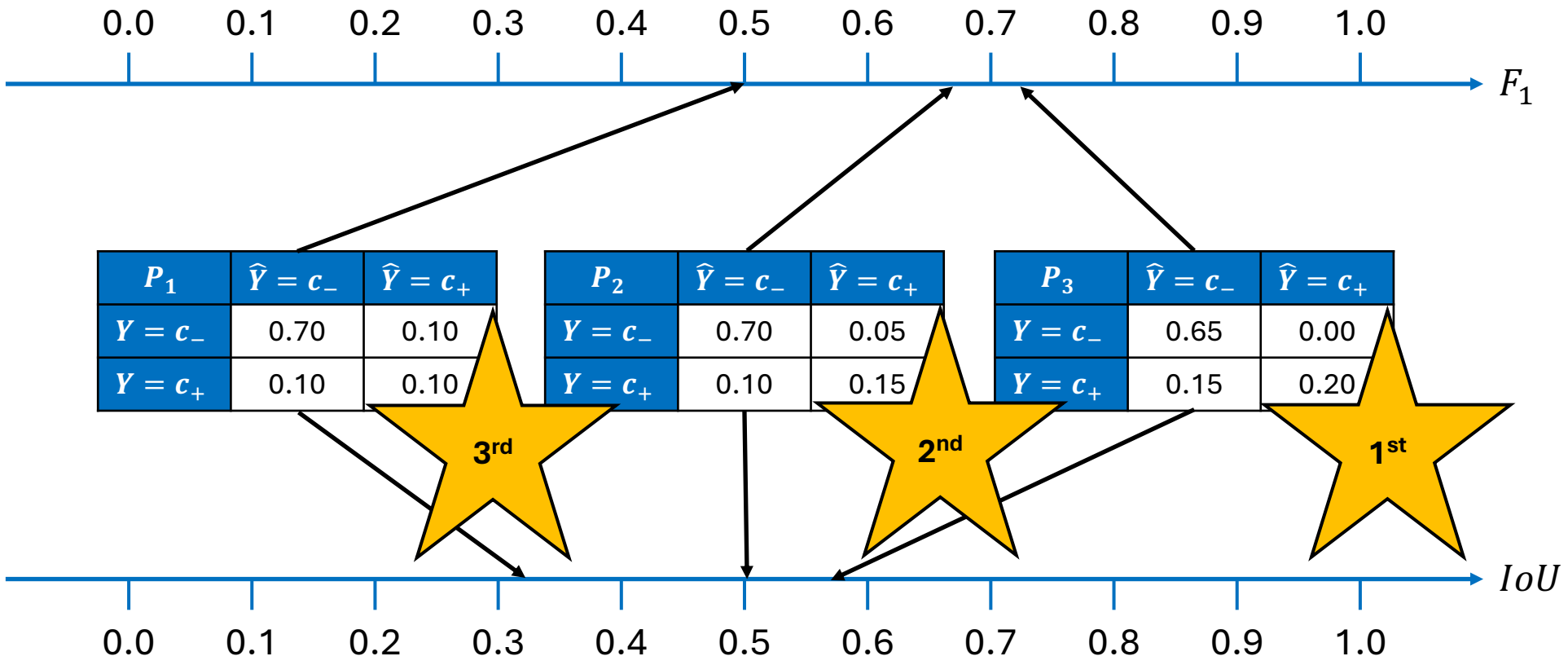
The *Intersection over Union* is defined as the ratio:

$$IoU = \frac{P\left((Y = c_+) \cap \left(\hat{Y} = c_+\right)\right)}{P\left((Y = c_+) \cup \left(\hat{Y} = c_+\right)\right)} = \frac{P(\{tp\})}{P(\{fp, fn, tp\})}$$



The scores F_1 and IoU induce the same performance ordering as they are linked by a strict monotonically increasing relationship: $F_1 = \frac{2\,IoU}{1 + IoU} \Rightarrow \frac{\partial F_1}{\partial IoU} > 0$.

Some scores are equivalent: example



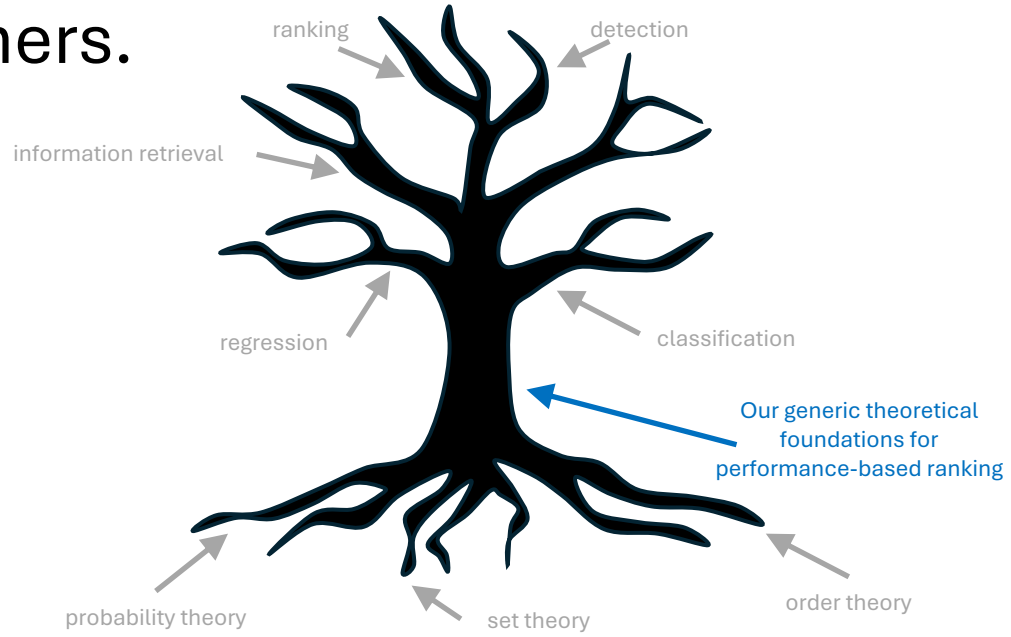
P_1	$\hat{Y} = c_-$	$\hat{Y} = c_+$
$Y = c_-$	0.70	0.10
$Y = c_+$	0.10	0.10

P_2	$\hat{Y} = c_-$	$\hat{Y} = c_+$
$Y = c_-$	0.70	0.05
$Y = c_+$	0.10	0.15

P_3	$\hat{Y} = c_-$	$\hat{Y} = c_+$
$Y = c_-$	0.65	0.00
$Y = c_+$	0.15	0.20

A few objectives of our research

1. Introducing the first generic *mathematical framework* to define, manipulate, analyze and compare performances.
2. Specify what a meaningful *performance-based ranking* is.
3. Study in depth the case of *two-class crisp classification*.
4. Develop *practical tools* for all researchers.



Part 2:

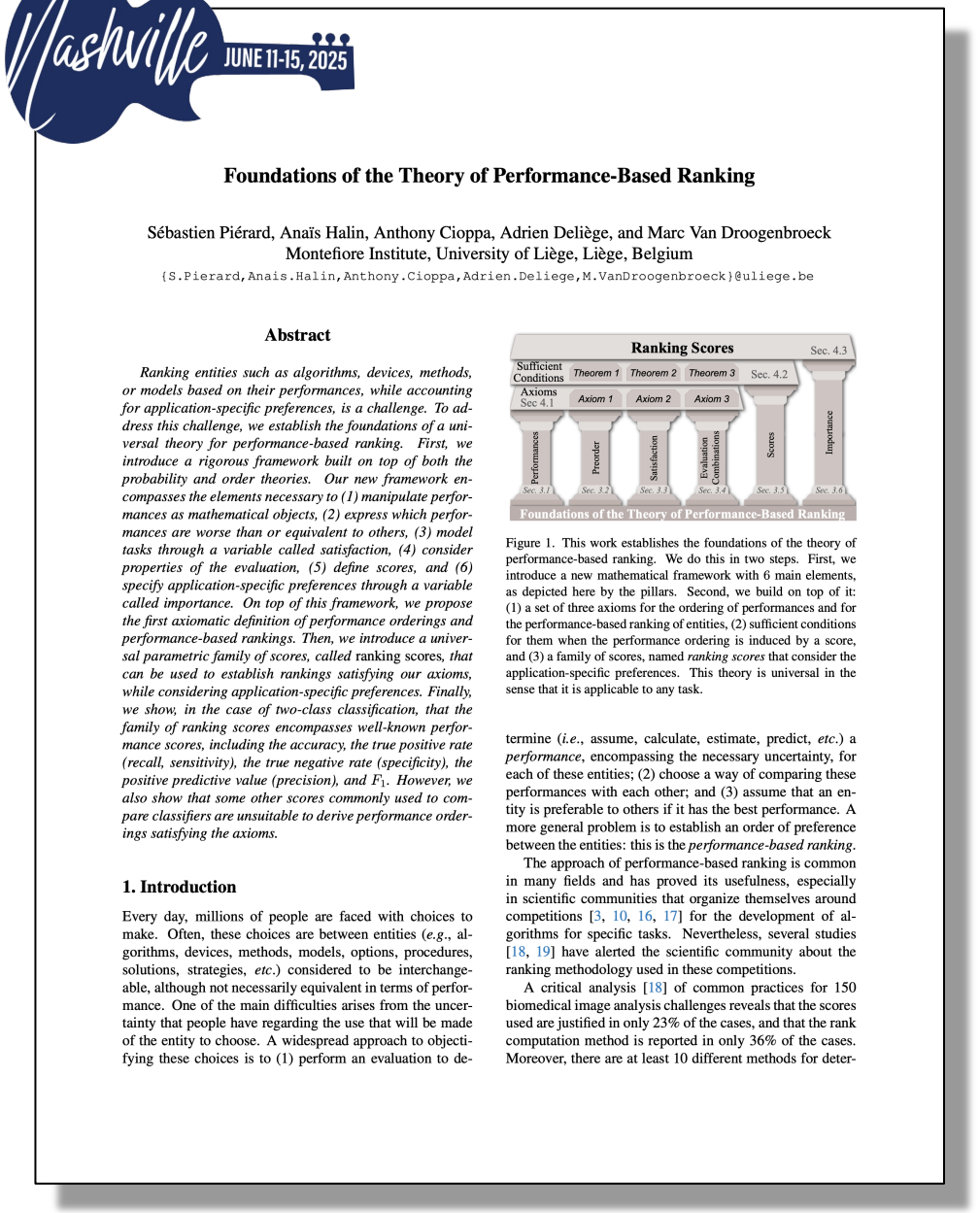
Our mathematical framework

This presentation concerns original research carried out at the Montefiore Institute of the University of Liège. And, this is the bust of Georges Montefiore-Levi (1832-1906), sculpted by Thomas Vinçotte in 1904.

Bibliography

This part of the presentation is based on the following paper:

Sébastien Piérard, Anaïs Halin, Anthony Cioppa, Adrien Deliège, and Marc Van Droogenbroeck. ***Foundations of the theory of performance-based ranking.*** In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, Tennessee, USA, June 2025.



Foundations of the Theory of Performance-Based Ranking

Sébastien Piérard, Anaïs Halin, Anthony Cioppa, Adrien Deliège, and Marc Van Droogenbroeck
Montefiore Institute, University of Liège, Liège, Belgium
{S.Pierard,Anais.Halin,Anthony.Cioppa,Adrien.Deliege,M.VanDroogenbroeck}@uliege.be

Abstract

Ranking entities such as algorithms, devices, methods, or models based on their performances, while accounting for application-specific preferences, is a challenge. To address this challenge, we establish the foundations of a universal theory for performance-based ranking. First, we introduce a rigorous framework built on top of both the probability and order theories. Our new framework encompasses the elements necessary to (1) manipulate performances as mathematical objects, (2) express which performances are worse than or equivalent to others, (3) model tasks through a variable called satisfaction, (4) consider properties of the evaluation, (5) define scores, and (6) specify application-specific preferences through a variable called importance. On top of this framework, we propose the first axiomatic definition of performance orderings and performance-based rankings. Then, we introduce a universal parametric family of scores, called ranking scores*, that can be used to establish rankings satisfying our axioms, while considering application-specific preferences. Finally, we show, in the case of two-class classification, that the family of ranking scores encompasses well-known performance scores, including the accuracy, the true positive rate (recall, sensitivity), the true negative rate (specificity), the positive predictive value (precision), and F_1. However, we also show that some other scores commonly used to compare classifiers are unsuitable to derive performance orderings satisfying the axioms.*

Figure 1. This work establishes the foundations of the theory of performance-based ranking. We do this in two steps. First, we introduce a new mathematical framework with 6 main elements, as depicted here by the pillars. Second, we build on top of it: (1) a set of three axioms for the ordering of performances and for the performance-based ranking of entities, (2) sufficient conditions for them when the performance ordering is induced by a score, and (3) a family of scores, named *ranking scores* that consider the application-specific preferences. This theory is universal in the sense that it is applicable to any task.

1. Introduction

Every day, millions of people are faced with choices to make. Often, these choices are between entities (*e.g.*, algorithms, devices, methods, models, options, procedures, solutions, strategies, *etc.*) considered to be interchangeable, although not necessarily equivalent in terms of performance. One of the main difficulties arises from the uncertainty that people have regarding the use that will be made of the entity to choose. A widespread approach to objectifying these choices is to (1) perform an evaluation to determine (*i.e.*, assume, calculate, estimate, predict, *etc.*) a *performance*, encompassing the necessary uncertainty, for each of these entities; (2) choose a way of comparing these performances with each other; and (3) assume that an entity is preferable to others if it has the best performance. A more general problem is to establish an order of preference between the entities: this is the *performance-based ranking*.

The approach of performance-based ranking is common in many fields and has proved its usefulness, especially in scientific communities that organize themselves around competitions [3, 10, 16, 17] for the development of algorithms for specific tasks. Nevertheless, several studies [18, 19] have alerted the scientific community about the ranking methodology used in these competitions.

A critical analysis [18] of common practices for 150 biomedical image analysis challenges reveals that the scores used are justified in only 23% of the cases, and that the rank computation method is reported in only 36% of the cases. Moreover, there are at least 10 different methods for deter-

The measurable space (Ω, Σ)

We build our framework on top of the *probability theory*, to address uncertainty.

Our advice:

- start by specifying explicitly a (thought) *random experiment* for your evaluations;
- the outcomes are the various cases that can happen during the *evaluation*;
- consider the set of possible *outcomes* for the sample space Ω;
- and, when Ω is finite, consider the σ-algebra 2^{Ω} for the event space Σ.

Several *choices* can be made for the outcomes. For example, in two-class crisp classification:

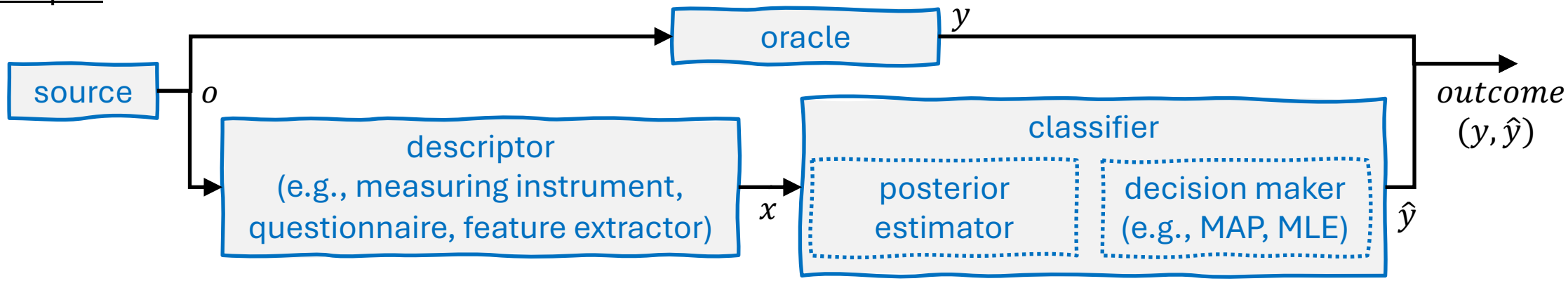
- $\Omega = \{correct\ result, incorrect\ result\}$
- $\Omega = \{correct\ result, type\ I\ error, type\ II\ error\}$
- $\Omega = \{true\ negative, false\ positive, false\ negative, true\ positive\}$

⚠ See slide about S:

- the evaluation cannot be *any* random experiment;
- we cannot choose anything we want as outcome!

The measurable space (Ω, Σ)

Example.



📣 We provide (in the supplementary material of our paper) a little catalog of other problems showcasing the use of our mathematical framework for *multi-class classification*, *regression*, *information retrieval*, *detection*, *clustering*, and *ranking* (ranking rankings, isn't this fun?).

The performances P

Performances are modeled as *probability measures* on (Ω, Σ).
The set of all possible performances, for the measurable space (Ω, Σ), is denoted by $\mathbb{P}_{(\Omega,\Sigma)}$.

Example: in classification, if one takes $\Omega = \mathbb{C}^2$ for all pairs of ground-truth and predicted classes, the performance object provides the same information as a *normalized confusion matrix*.

Interpretation: the performance of an entity (e.g., a classifier) represents the distribution of cases (outcomes) that can be encountered when evaluating this entity.

The performances P

Example.

- The probability of observing the outcome $\begin{matrix} tn=(c_-,c_-) \\ fp=(c_-,c_+) \\ fn=(c_+,c_-) \\ tp=(c_+,c_+) \end{matrix}$ is $\begin{matrix} \mathrm{P}(\{\mathrm{tn}\})=40\% \\ \mathrm{P}(\{\mathrm{fp}\})=30\% \\ \mathrm{P}(\{\mathrm{fn}\})=20\% \\ \mathrm{P}(\{\mathrm{tp}\})=10\% \end{matrix}$

P	$\hat{Y}=c_-$	$\hat{Y}=c_+$
$Y=c_-$	0.4	0.3
$Y=c_+$	0.2	0.1

- The probability of being right when predicting $\begin{matrix} c_- \\ c_+ \end{matrix}$ is $\begin{matrix} P(\{tn\}|\{tn,fn\})=2/3 \\ P(\{tp\}|\{fp,tp\})=1/4 \end{matrix}$

- The probability of taking the right decision when the ground truth is $\begin{matrix} c_- \\ c_+ \end{matrix}$ is $\begin{matrix} P(\{tn\}|\{tn,fp\})=4/7 \\ P(\{tp\}|\{fn,tp\})=1/3 \end{matrix}$

The performance orderings $\lesssim$

We build our framework on top of the *order theory*: the performance orderings $\lesssim$ are *preorders* on a given $\mathbb{P}_{(\Omega,\Sigma)}$ that are interpreted as *worse or equivalent than*.

📣 This prevents from comparing apples with oranges: there is no performance ordering that considers performances on (Ω_1, Σ_1) and performances on (Ω_2, Σ_2) with $(\Omega_1, \Sigma_1) \neq (\Omega_2, \Sigma_2)$.

📣 This is very practical: with a single homogeneous binary relation $\lesssim$, we have "mechanically" the others:

- *equivalent to*: $P_1 \sim P_2 \Leftrightarrow (P_1 \lesssim P_2)\ and\ (P_2 \lesssim P_1)$
- *worse than*: $P_1 < P_2 \Leftrightarrow (P_1 \lesssim P_2)\ and\ not\ (P_2 \lesssim P_1)$
- *better than*: $P_1 > P_2 \Leftrightarrow not\ (P_1 \lesssim P_2)\ and\ (P_2 \lesssim P_1)$
- *incomparable with*: $not\ (P_1 \lesssim P_2)\ and\ not\ (P_2 \lesssim P_1)$

One can prove that these binary relations have the properties needed to be interpreted like this, because we know that $\lesssim$ is a preorder.

📣 It is possible to induce a performance ordering $\lesssim_X$ from any score X as follows:

$$P_1 \lesssim_X P_2 \Leftrightarrow \begin{cases} X(P_1) \leq X(P_2)\ and\ P_1 \in dom(X)\ and\ P_2 \in dom(X) \\ or\ P_1 = P_2 \end{cases}$$

Modeling the evaluation by a function ϕ

Let us assume that some performances $P_1, P_2, \dots P_n$ are achievable.
In this case, are we sure that we are able to achieve a performance P?

Answering this question is the purpose of the idempotent function $\phi: 2^{\mathbb{P}(\Omega,\Sigma)} \to 2^{\mathbb{P}(\Omega,\Sigma)}$:

$$P \in \phi(\{P_1, P_2, \dots P_n\})\ ?$$

Example. When the entity is used only once during the realization of the random experiment, we can take

$$\phi = convex - hull$$

This is because, if we use a hybrid entity e that applies a non-deterministically chosen entity among $e_1, e_2, \dots e_n$, then we obtain a performance P that is convex combination of $P_1, P_2, \dots P_n$ with $P_i = eval(e_i)$; the combination weights being equal to the probabilities of choosing the respective entities.

Modeling the evaluation by a function ϕ

Question.

We have a classifier c_1 such that $eval(c_1) = P_1$.
We have a classifier c_2 such that $eval(c_2) = P_2$.
During the evaluation, the classifier is used only once.
Can we obtain the performance P?

P_1	$\hat{Y} = c_-$	$\hat{Y} = c_+$
$Y = c_-$	0.50	0.10
$Y = c_+$	0.30	0.10

P_2	$\hat{Y} = c_-$	$\hat{Y} = c_+$
$Y = c_-$	0.80	0.05
$Y = c_+$	0.00	0.15

P	$\hat{Y} = c_-$	$\hat{Y} = c_+$
$Y = c_-$	0.725	0.0625
$Y = c_+$	0.075	0.1375

Answer.

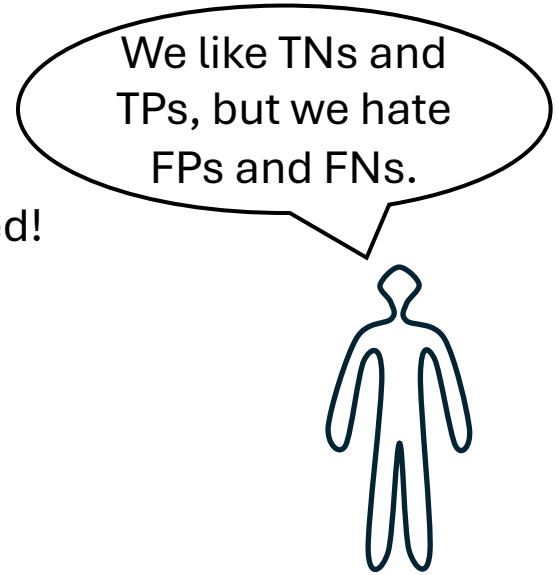
It turns out that $P = 0.25\ P_1 + 0.75\ P_2$. So, we can achieve the performance P by using the classifier c_1 with probability 0.25 and the classifier c_2 with probability 0.75.

Modeling the task as a random variable called *Satisfaction* (S)

📣 We have seen that performances are probability measures. But not all probability measures are performances! Besides, we do not aim at ordering any probability measures. The satisfaction is the key.

The measurable space (or the random experiment) should be such that one can specify the degree of satisfaction that is obtained by each sample (outcome).

$$S : \Omega \to \mathbb{R} : \omega \mapsto S(\omega)$$



The satisfaction is related to the task.
⚠️ Our advice: always specify explicitly the satisfaction, otherwise the task is undefined!

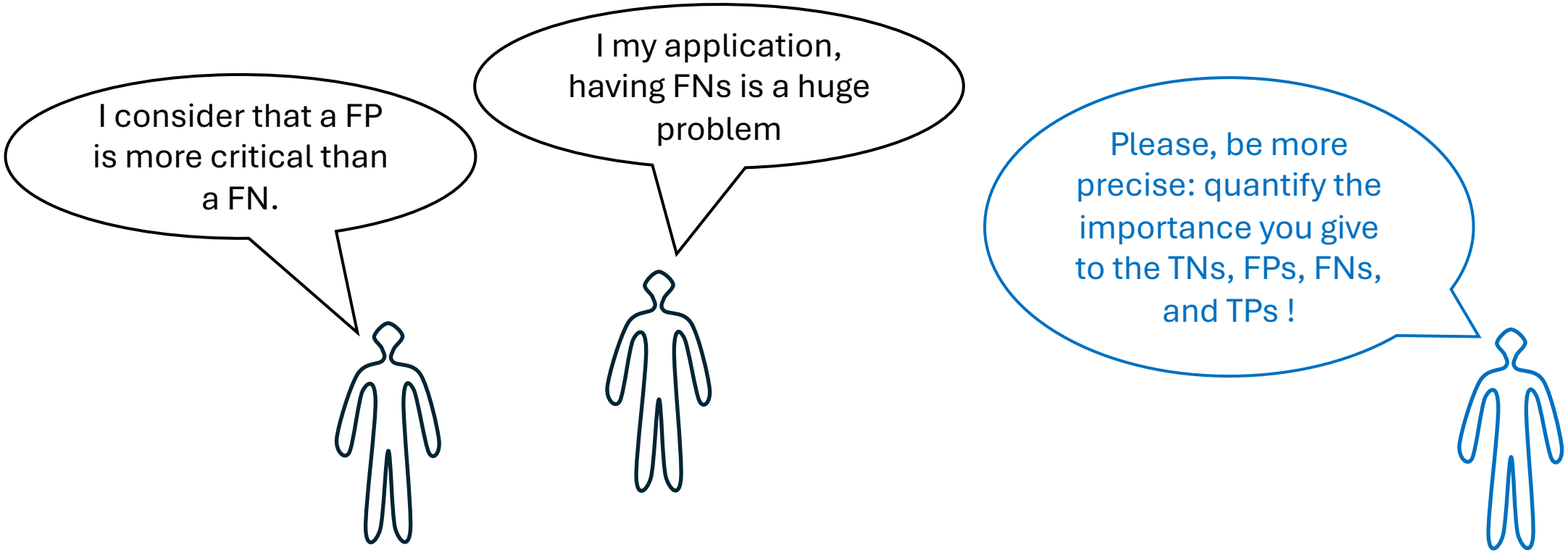
Example. In two-class crisp classification, by convention,

$$S(fp) = S(fn) = 0 \quad \text{and} \quad S(tn) = S(tp) = 1$$

📣 There is no constraint on satisfaction values: they can be negative, zero, or positive.

Modeling the application as a random variable called *Importance* (I)

The *importance* encodes some application-specific preferences.



📣 There is a constraint on importance values: they should be zero or positive.

The scores X are real functions of performances

$$X : dom(X) \to \mathbb{R} : P \mapsto X(P) \quad \text{with} \quad \mathrm{dom(X)} \subseteq \mathbb{P}_{(\Omega,\Sigma)}$$

⚠ It is essential to avoid confusions between the value taken by a score and the performance.

Somes scores have a probabilistic meaning, others not. Examples:

- the *expected satisfaction* is the score
$$P \mapsto E_P[S]$$
- in two-class classification, the *accuracy* is the score
$$A : \ P \mapsto A(P) = \ P(S = 1) = P(\{tn, tp\})$$
- in two-class classification, the *true positive rate* is the score
$$TPR : \ P \mapsto TPR(P) = \ P(S = 1|Y = c_+) = P(\{tp\}|\{fn, tp\})$$
- in two-class classification, the *positive predictive value* is the score
$$PPV : \ P \mapsto PPV(P) = \ P\left(S = 1\middle|\hat{Y} = c_+\right) = P(\{tp\}|\{fp, tp\})$$

Part 3:

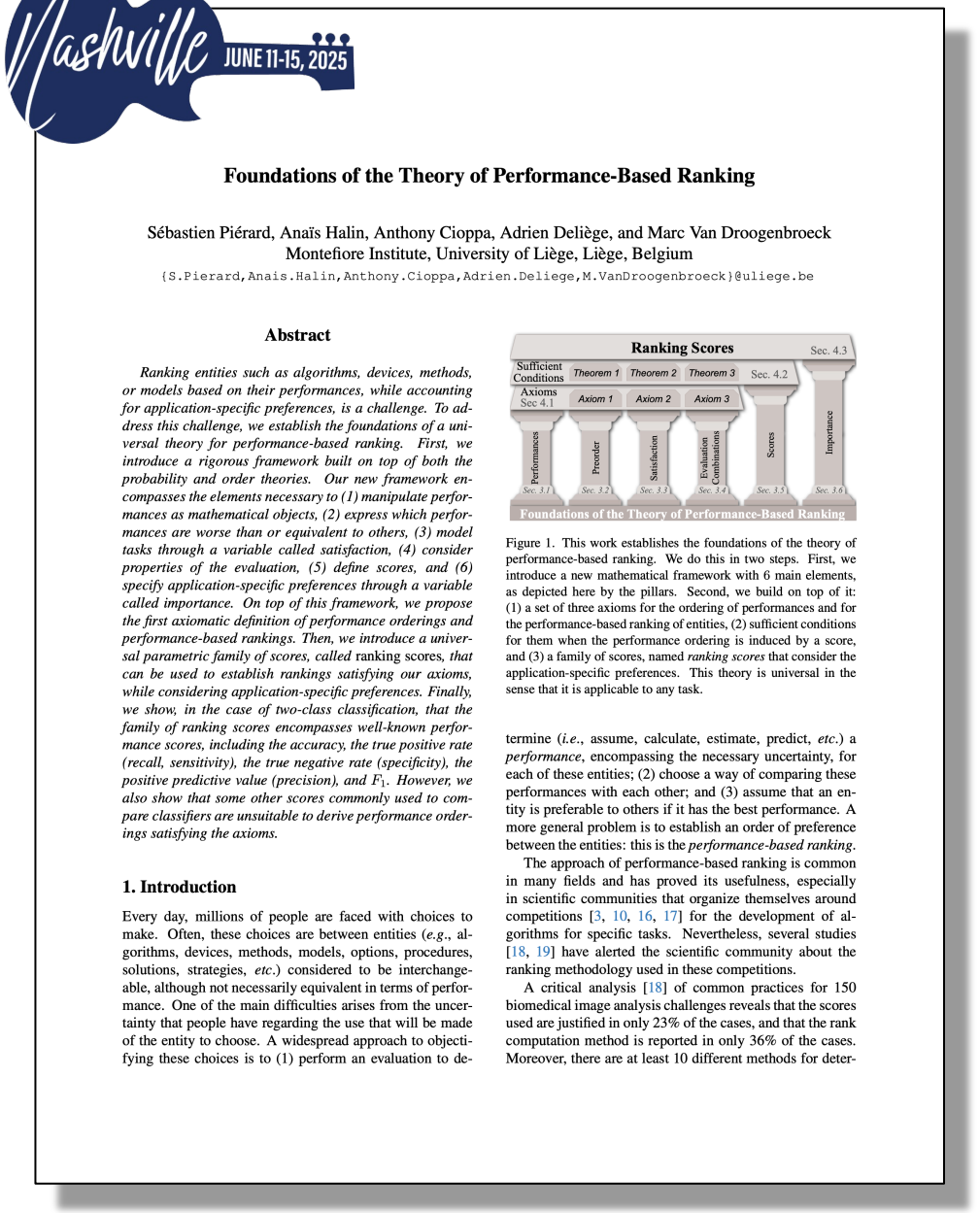
Our axiomatic definition of performance-based rankings

This presentation concerns original research carried out at the Montefiore Institute of the University of Liège. And, this is the bust of Georges Montefiore-Levi (1832-1906), sculpted by Thomas Vinçotte in 1904.

Bibliography

This part of the presentation is based on the following paper:

Sébastien Piérard, Anaïs Halin, Anthony Cioppa, Adrien Deliège, and Marc Van Droogenbroeck. ***Foundations of the theory of performance-based ranking.*** In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, Tennessee, USA, June 2025.



Foundations of the Theory of Performance-Based Ranking

Sébastien Piérard, Anaïs Halin, Anthony Cioppa, Adrien Deliège, and Marc Van Droogenbroeck
Montefiore Institute, University of Liège, Liège, Belgium
{S.Pierard,Anais.Halin,Anthony.Cioppa,Adrien.Deliege,M.VanDroogenbroeck}@uliege.be

Abstract

Ranking entities such as algorithms, devices, methods, or models based on their performances, while accounting for application-specific preferences, is a challenge. To address this challenge, we establish the foundations of a universal theory for performance-based ranking. First, we introduce a rigorous framework built on top of both the probability and order theories. Our new framework encompasses the elements necessary to (1) manipulate performances as mathematical objects, (2) express which performances are worse than or equivalent to others, (3) model tasks through a variable called satisfaction, (4) consider properties of the evaluation, (5) define scores, and (6) specify application-specific preferences through a variable called importance. On top of this framework, we propose the first axiomatic definition of performance orderings and performance-based rankings. Then, we introduce a universal parametric family of scores, called ranking scores*, that can be used to establish rankings satisfying our axioms, while considering application-specific preferences. Finally, we show, in the case of two-class classification, that the family of ranking scores encompasses well-known performance scores, including the accuracy, the true positive rate (recall, sensitivity), the true negative rate (specificity), the positive predictive value (precision), and F_1. However, we also show that some other scores commonly used to compare classifiers are unsuitable to derive performance orderings satisfying the axioms.*

Figure 1. This work establishes the foundations of the theory of performance-based ranking. We do this in two steps. First, we introduce a new mathematical framework with 6 main elements, as depicted here by the pillars. Second, we build on top of it: (1) a set of three axioms for the ordering of performances and for the performance-based ranking of entities, (2) sufficient conditions for them when the performance ordering is induced by a score, and (3) a family of scores, named *ranking scores* that consider the application-specific preferences. This theory is universal in the sense that it is applicable to any task.

1. Introduction

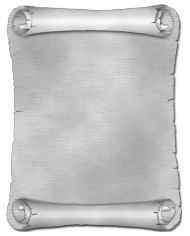
Every day, millions of people are faced with choices to make. Often, these choices are between entities (*e.g.*, algorithms, devices, methods, models, options, procedures, solutions, strategies, *etc.*) considered to be interchangeable, although not necessarily equivalent in terms of performance. One of the main difficulties arises from the uncertainty that people have regarding the use that will be made of the entity to choose. A widespread approach to objectifying these choices is to (1) perform an evaluation to determine (*i.e.*, assume, calculate, estimate, predict, *etc.*) a *performance*, encompassing the necessary uncertainty, for each of these entities; (2) choose a way of comparing these performances with each other; and (3) assume that an entity is preferable to others if it has the best performance. A more general problem is to establish an order of preference between the entities: this is the *performance-based ranking*.

The approach of performance-based ranking is common in many fields and has proved its usefulness, especially in scientific communities that organize themselves around competitions [3, 10, 16, 17] for the development of algorithms for specific tasks. Nevertheless, several studies [18, 19] have alerted the scientific community about the ranking methodology used in these competitions.

A critical analysis [18] of common practices for 150 biomedical image analysis challenges reveals that the scores used are justified in only 23% of the cases, and that the rank computation method is reported in only 36% of the cases. Moreover, there are at least 10 different methods for deter-

Axiom 1

Can we report perennial conclusions in our papers ?



Motivation: does it make sense ?

P_A	$\hat{Y} = c_-$	$\hat{Y} = c_+$
$Y = c_-$	0.759	0.141
$Y = c_+$	0.028	0.072

P_B	$\hat{Y} = c_-$	$\hat{Y} = c_+$
$Y = c_-$	0.532	0.368
$Y = c_+$	0.010	0.090

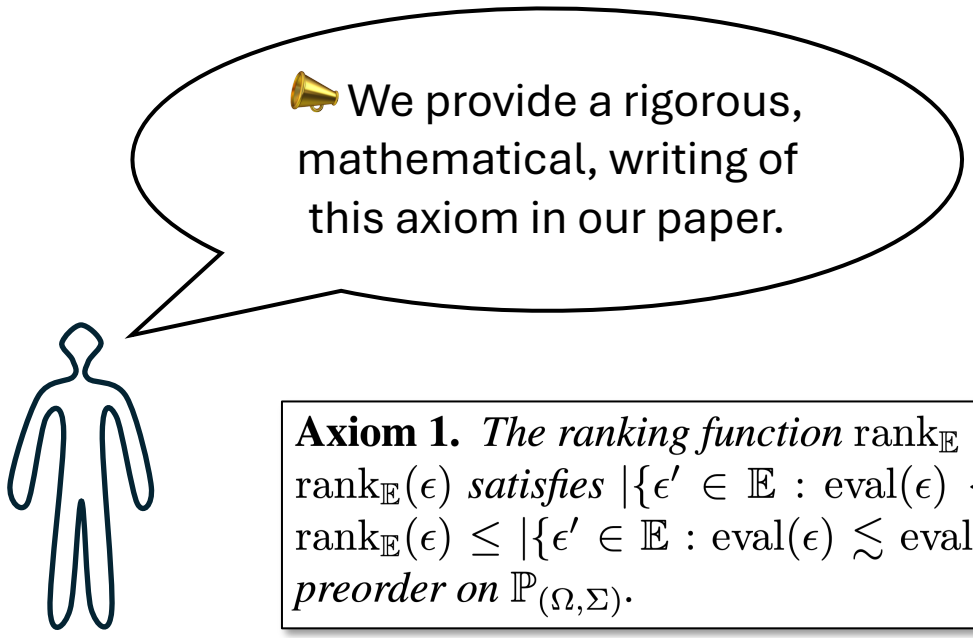
P_C	$\hat{Y} = c_-$	$\hat{Y} = c_+$
$Y = c_-$	0. 761	0.139
$Y = c_+$	0.019	0.081

Classifier	TPR	TNR	PPV	NPV	F-one	Mean Rank
Classifier A	0.7200 [2]	0.8433 [1]	0.3380 [1]	0.9644 [2]	0.4601 [1]	**[1.4]**
Classifier B	0.9000 [1]	0.5911 [2]	0.1965 [2]	0.9815 [1]	0.3226 [2]	**[1.6]**

Classifier	TPR	TNR	PPV	NPV	F-one	Mean Rank
Classifier C	0.8100 [2]	0.8456 [1]	0.3682 [1]	0.9756 [2]	0.5062 [1]	**[1.4]**
Classifier B	0.9000 [1]	0.5911 [3]	0.1965 [3]	0.9815 [1]	0.3226 [3]	**[2.2]**
Classifier A	0.7200 [3]	0.8433 [2]	0.3380 [2]	0.9644 [3]	0.4601 [2]	**[2.4]**

Axiom 1

Any performance-based ranking should be derived from a performance ordering $\lesssim$.



Axiom 1. *The ranking function* $\mathrm{rank}_{\mathbb{E}} : \mathbb{E} \to [1, |\mathbb{E}|] : \epsilon \mapsto \mathrm{rank}_{\mathbb{E}}(\epsilon)$ *satisfies* $|\{\epsilon' \in \mathbb{E} : \mathrm{eval}(\epsilon) < \mathrm{eval}(\epsilon')\}| + 1 \leq \mathrm{rank}_{\mathbb{E}}(\epsilon) \leq |\{\epsilon' \in \mathbb{E} : \mathrm{eval}(\epsilon) \lesssim \mathrm{eval}(\epsilon')\}|$, *where* $\lesssim$ *is a preorder on* $\mathbb{P}_{(\Omega,\Sigma)}$.

Axiom 2

Motivation: ensure these requirements, and more.

- Achieving the minimal satisfaction for sure is the worst we can do:

$$\left.\begin{aligned} s_{min} &= \min_{\omega \in \Omega} S(\omega) \\ P(S = s_{min}) &= 1 \end{aligned}\right\} \Rightarrow \nexists P' : P' < P$$

P	$\widehat{Y} = c_-$	$\widehat{Y} = c_+$
$Y = c_-$	0	…
$Y = c_+$	…	0

- Achieving the maximum satisfaction for sure is the best we can do:

$$\left.\begin{aligned} s_{max} &= \max_{\omega \in \Omega} S(\omega) \\ P(S = s_{max}) &= 1 \end{aligned}\right\} \Rightarrow \nexists P' : P' > P$$

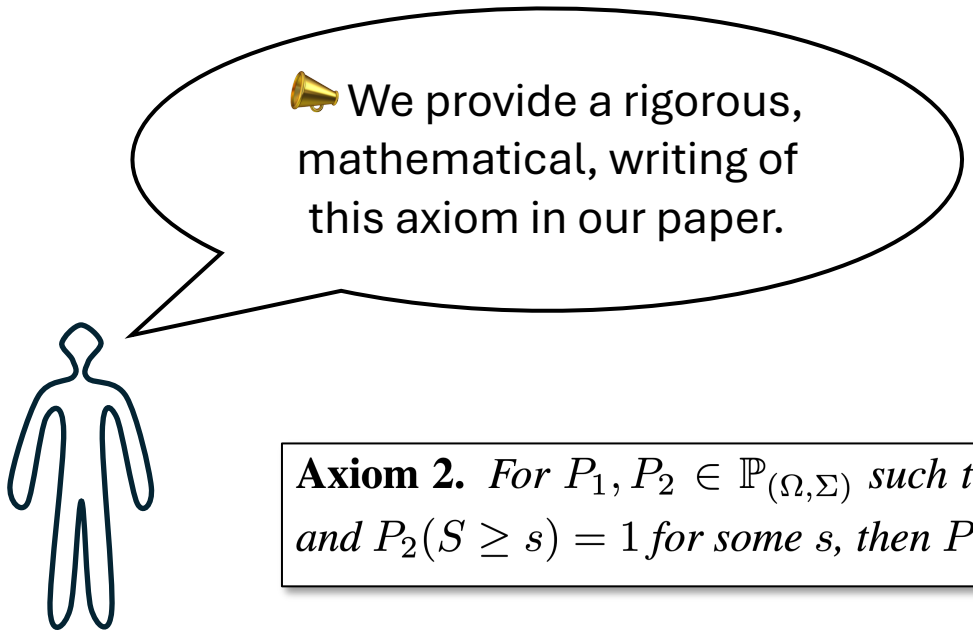
P	$\widehat{Y} = c_-$	$\widehat{Y} = c_+$
$Y = c_-$	…	0
$Y = c_+$	0	…

- If we achieve for sure the same satisfaction with two performances, then these performances are either equivalent or incomparable.

Axiom 2

The performance ordering $\lesssim$

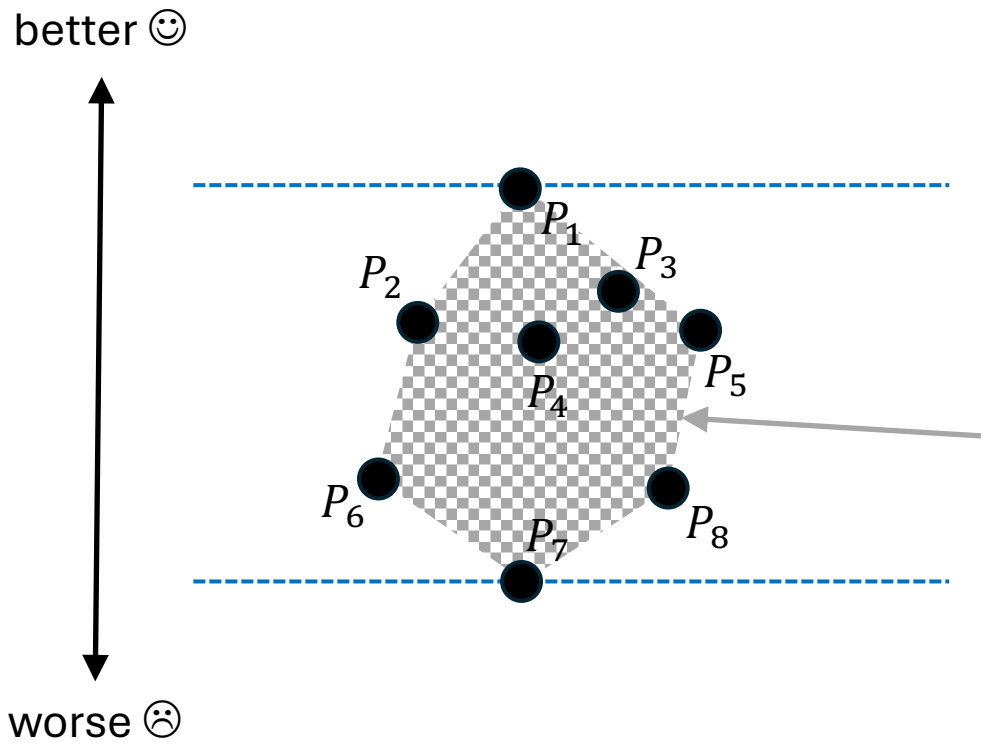
should be consistent with the task (modeled by S).



Axiom 2. *For $P_1, P_2 \in \mathbb{P}_{(\Omega,\Sigma)}$ such that $P_1(S \leq s) = 1$ and $P_2(S \geq s) = 1$ for some s, then $P_1 \lesssim P_2$ or $P_1 \nlesssim P_2$.*

Axiom 3

Motivation: magic doesn't exist!

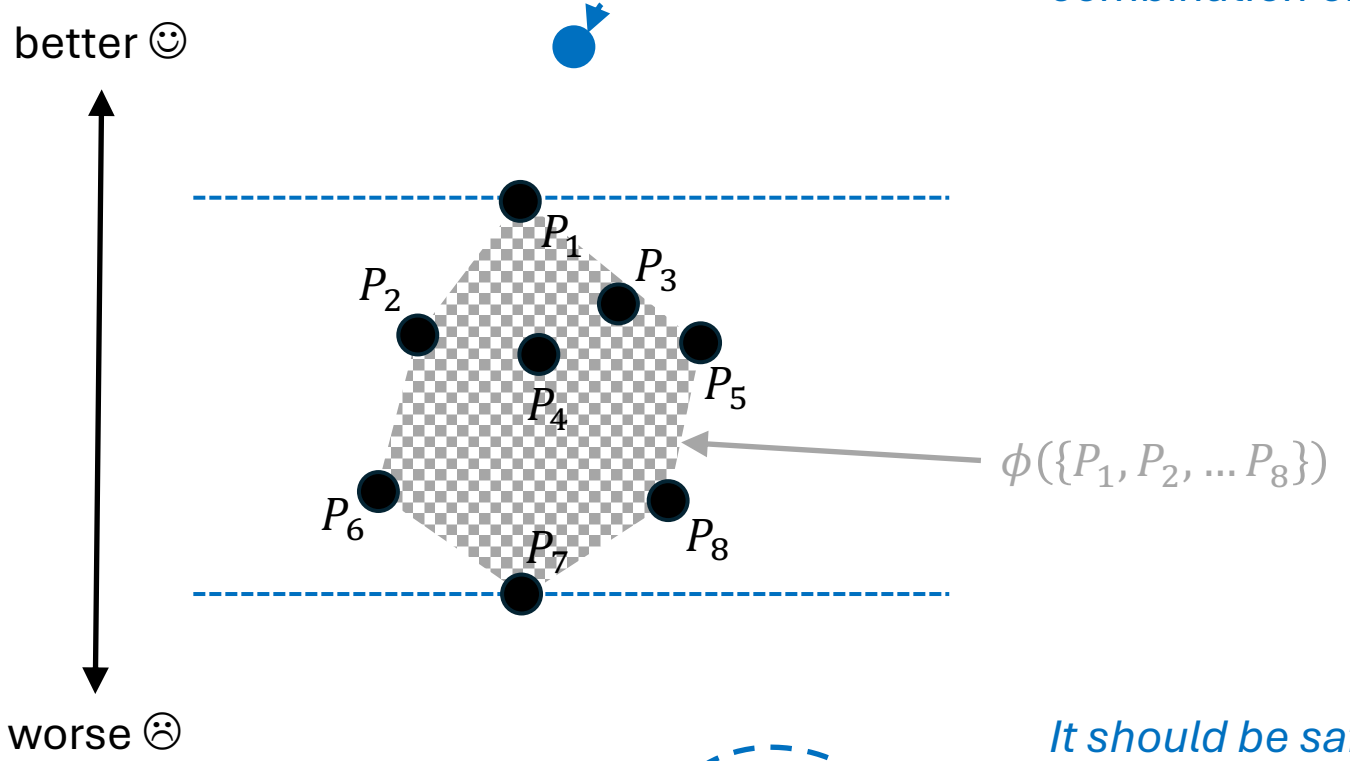


It should not be possible to obtain a performance better than the best one with a "blind" combination.

It should not be possible to obtain a performance worse than the worst one with a "blind" combination.

Axiom 3

It should be safe to say that this performance is better than all the ones that could be obtained with a "blind" combination of the ones depicted in black.

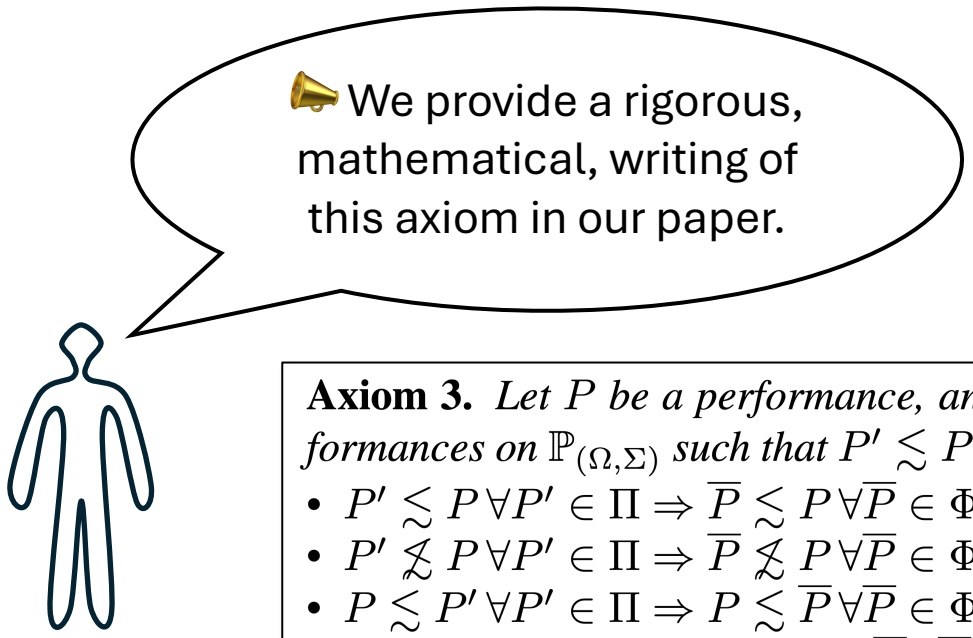


It should be safe to say that this performance is worse than all the ones that could be obtained with a "blind" combination of the ones depicted in black.

Axiom 3

The performance ordering $\lesssim$

should be consistent with the evaluation (modeled by ϕ).



Axiom 3. *Let P be a performance, and Π be a set of performances on $\mathbb{P}_{(\Omega,\Sigma)}$ such that $P' \lesssim P \vee P \lesssim P' \forall P' \in \Pi$.*

- $P' \lesssim P \forall P' \in \Pi \Rightarrow \overline{P} \lesssim P \forall \overline{P} \in \Phi(\Pi)$;
- $P' \not\lesssim P \forall P' \in \Pi \Rightarrow \overline{P} \not\lesssim P \forall \overline{P} \in \Phi(\Pi)$;
- $P \lesssim P' \forall P' \in \Pi \Rightarrow P \lesssim \overline{P} \forall \overline{P} \in \Phi(\Pi)$;
- *and* $P \not\lesssim P' \forall P' \in \Pi \Rightarrow P \not\lesssim \overline{P} \forall \overline{P} \in \Phi(\Pi)$.

Part 4:

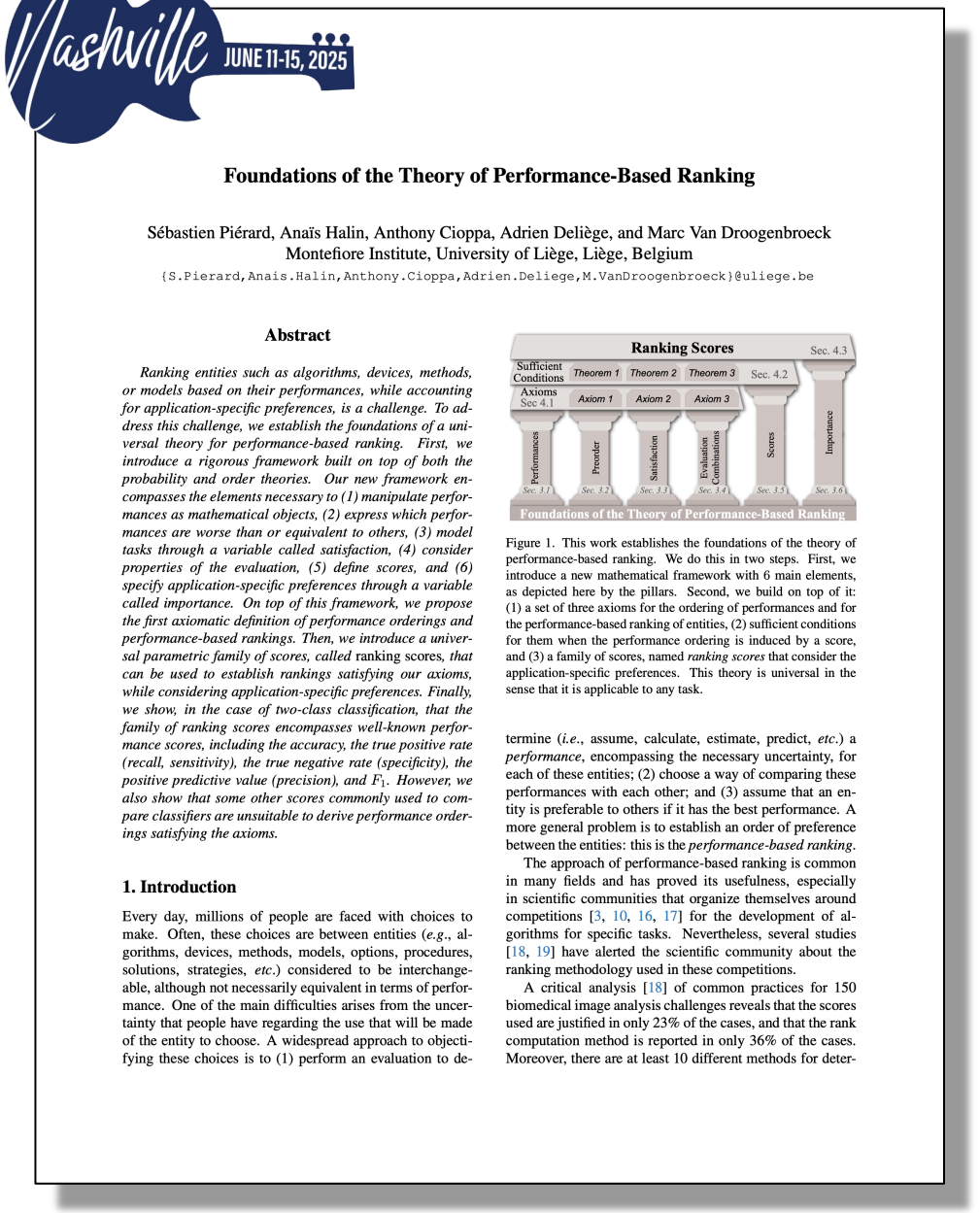
Our family of ranking scores

This presentation concerns original research carried out at the Montefiore Institute of the University of Liège. And, this is the bust of Georges Montefiore-Levi (1832-1906), sculpted by Thomas Vinçotte in 1904.

Bibliography

This part of the presentation is based on the following paper:

Sébastien Piérard, Anaïs Halin, Anthony Cioppa, Adrien Deliège, and Marc Van Droogenbroeck. ***Foundations of the theory of performance-based ranking.*** In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, Tennessee, USA, June 2025.



Foundations of the Theory of Performance-Based Ranking

Sébastien Piérard, Anaïs Halin, Anthony Cioppa, Adrien Deliège, and Marc Van Droogenbroeck
Montefiore Institute, University of Liège, Liège, Belgium
{S.Pierard,Anais.Halin,Anthony.Cioppa,Adrien.Deliege,M.VanDroogenbroeck}@uliege.be

Abstract

Ranking entities such as algorithms, devices, methods, or models based on their performances, while accounting for application-specific preferences, is a challenge. To address this challenge, we establish the foundations of a universal theory for performance-based ranking. First, we introduce a rigorous framework built on top of both the probability and order theories. Our new framework encompasses the elements necessary to (1) manipulate performances as mathematical objects, (2) express which performances are worse than or equivalent to others, (3) model tasks through a variable called satisfaction, (4) consider properties of the evaluation, (5) define scores, and (6) specify application-specific preferences through a variable called importance. On top of this framework, we propose the first axiomatic definition of performance orderings and performance-based rankings. Then, we introduce a universal parametric family of scores, called ranking scores*, that can be used to establish rankings satisfying our axioms, while considering application-specific preferences. Finally, we show, in the case of two-class classification, that the family of ranking scores encompasses well-known performance scores, including the accuracy, the true positive rate (recall, sensitivity), the true negative rate (specificity), the positive predictive value (precision), and F_1. However, we also show that some other scores commonly used to compare classifiers are unsuitable to derive performance orderings satisfying the axioms.*

Figure 1. This work establishes the foundations of the theory of performance-based ranking. We do this in two steps. First, we introduce a new mathematical framework with 6 main elements, as depicted here by the pillars. Second, we build on top of it: (1) a set of three axioms for the ordering of performances and for the performance-based ranking of entities, (2) sufficient conditions for them when the performance ordering is induced by a score, and (3) a family of scores, named *ranking scores* that consider the application-specific preferences. This theory is universal in the sense that it is applicable to any task.

1. Introduction

Every day, millions of people are faced with choices to make. Often, these choices are between entities (*e.g.*, algorithms, devices, methods, models, options, procedures, solutions, strategies, *etc.*) considered to be interchangeable, although not necessarily equivalent in terms of performance. One of the main difficulties arises from the uncertainty that people have regarding the use that will be made of the entity to choose. A widespread approach to objectifying these choices is to (1) perform an evaluation to determine (*i.e.*, assume, calculate, estimate, predict, *etc.*) a *performance*, encompassing the necessary uncertainty, for each of these entities; (2) choose a way of comparing these performances with each other; and (3) assume that an entity is preferable to others if it has the best performance. A more general problem is to establish an order of preference between the entities: this is the *performance-based ranking*.

The approach of performance-based ranking is common in many fields and has proved its usefulness, especially in scientific communities that organize themselves around competitions [3, 10, 16, 17] for the development of algorithms for specific tasks. Nevertheless, several studies [18, 19] have alerted the scientific community about the ranking methodology used in these competitions.

A critical analysis [18] of common practices for 150 biomedical image analysis challenges reveals that the scores used are justified in only 23% of the cases, and that the rank computation method is reported in only 36% of the cases. Moreover, there are at least 10 different methods for deter-

Our ranking scores

When $\phi = convex - hull$, the axioms are satisfied when the performance ordering $\lesssim$ is induced by any of the *ranking scores*:

$$R_I : P \mapsto R_I(P) = \frac{E_P[SI]}{E_P[I]} = \frac{\sum_{\omega\in\Omega} S(\omega)I(\omega)P(\{\omega\})}{\sum_{\omega\in\Omega} I(\omega)P(\{\omega\})} \qquad I \neq 0,\ I(\omega) \geq 0 \forall \omega \in \Omega$$

- 1st example. The *expected satisfaction*, $E_P[S]$, is a particular case of ranking score. In classification, it is known as the *accuracy*.

- 2nd example. When the *satisfaction* and the *importance* are binary, the ranking scores are the probabilities of being *satisfied* given that the encountered outcome is *important*, that is:

$$R_I(P) = P(\text{S=1}|\text{I=1})$$

Our canonical ranking scores

Let us consider the case of a binary satisfaction, i.e., $S(\omega) \in \{0,1\}$.

Property. If two ranking scores R_{I_1} and R_{I_2} are such that

$$\begin{cases} I_1(\omega) \propto I_2(\omega), \omega \in S^{-1}(0) \\ I_1(\omega) \propto I_2(\omega), \omega \in S^{-1}(1) \end{cases}$$

then the performance orderings induced by R_{I_1} and R_{I_2} are identical.

We get rid of this redundancy by introducing the *canonical ranking scores*:

$$\begin{cases} \sum_{\omega \in S^{-1}(0)} I(\omega) = 1 \\ \sum_{\omega \in S^{-1}(1)} I(\omega) = 1 \end{cases}$$

Part 5:

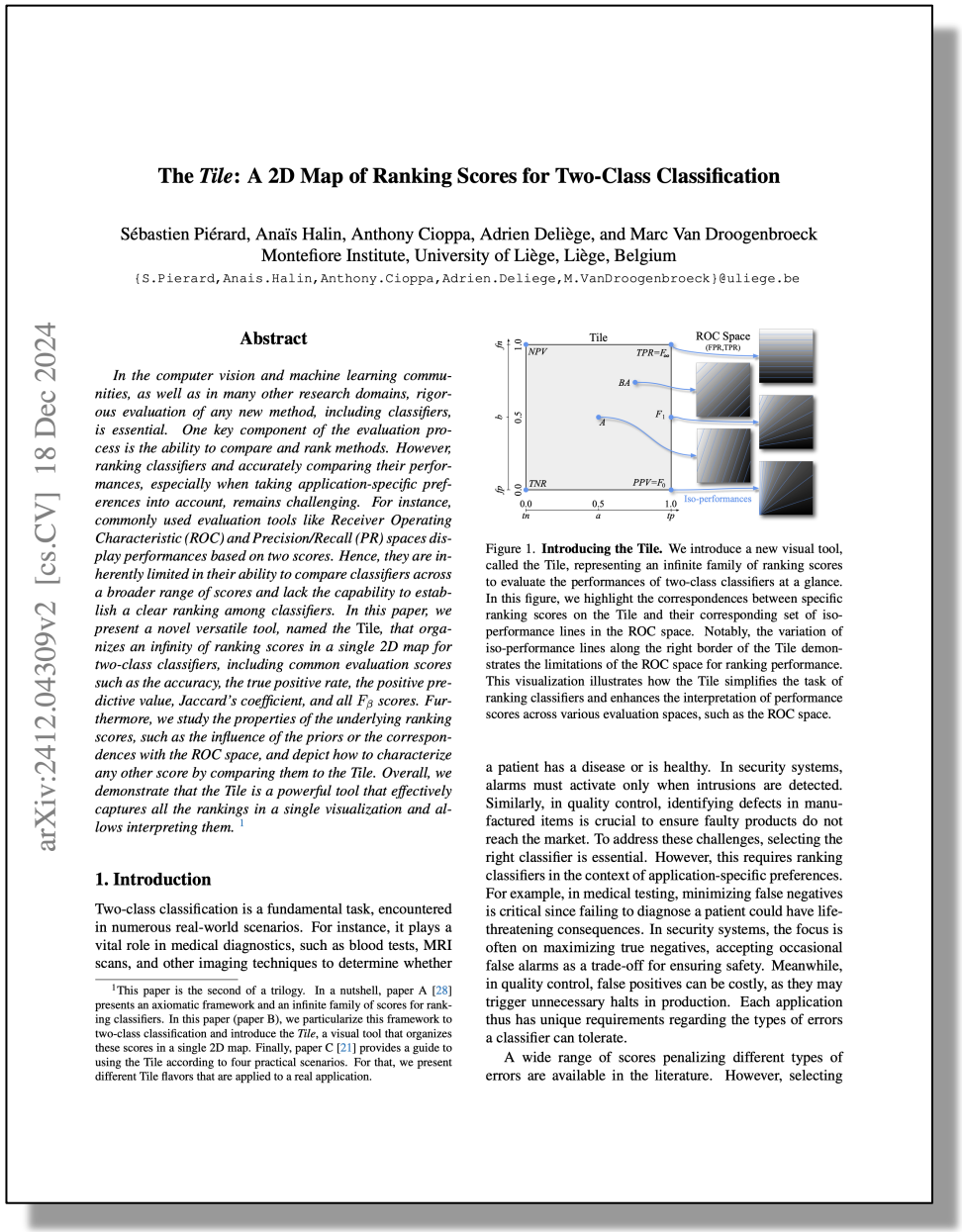
The Tile

This presentation concerns original research carried out at the Montefiore Institute of the University of Liège. And, this is the bust of Georges Montefiore-Levi (1832-1906), sculpted by Thomas Vinçotte in 1904.

Bibliography

This part of the presentation is based on the following paper:

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, abs/2412.04309, 2024. URL https://doi.org/10.48550/arXiv.2412.04309.



The *Tile*: A 2D Map of Ranking Scores for Two-Class Classification

Sébastien Piérard, Anaïs Halin, Anthony Cioppa, Adrien Deliège, and Marc Van Droogenbroeck
Montefiore Institute, University of Liège, Liège, Belgium
{S.Pierard,Anais.Halin,Anthony.Cioppa,Adrien.Deliege,M.VanDroogenbroeck}@uliege.be

Abstract

In the computer vision and machine learning communities, as well as in many other research domains, rigorous evaluation of any new method, including classifiers, is essential. One key component of the evaluation process is the ability to compare and rank methods. However, ranking classifiers and accurately comparing their performances, especially when taking application-specific preferences into account, remains challenging. For instance, commonly used evaluation tools like Receiver Operating Characteristic (ROC) and Precision/Recall (PR) spaces display performances based on two scores. Hence, they are inherently limited in their ability to compare classifiers across a broader range of scores and lack the capability to establish a clear ranking among classifiers. In this paper, we present a novel versatile tool, named the Tile*, that organizes an infinity of ranking scores in a single 2D map for two-class classifiers, including common evaluation scores such as the accuracy, the true positive rate, the positive predictive value, Jaccard's coefficient, and all F_β scores. Furthermore, we study the properties of the underlying ranking scores, such as the influence of the priors or the correspondences with the ROC space, and depict how to characterize any other score by comparing them to the Tile. Overall, we demonstrate that the Tile is a powerful tool that effectively captures all the rankings in a single visualization and allows interpreting them.* [1]

Figure 1. **Introducing the Tile.** We introduce a new visual tool, called the Tile, representing an infinite family of ranking scores to evaluate the performances of two-class classifiers at a glance. In this figure, we highlight the correspondences between specific ranking scores on the Tile and their corresponding set of iso-performance lines in the ROC space. Notably, the variation of iso-performance lines along the right border of the Tile demonstrates the limitations of the ROC space for ranking performance. This visualization illustrates how the Tile simplifies the task of ranking classifiers and enhances the interpretation of performance scores across various evaluation spaces, such as the ROC space.

1. Introduction

Two-class classification is a fundamental task, encountered in numerous real-world scenarios. For instance, it plays a vital role in medical diagnostics, such as blood tests, MRI scans, and other imaging techniques to determine whether a patient has a disease or is healthy. In security systems, alarms must activate only when intrusions are detected. Similarly, in quality control, identifying defects in manufactured items is crucial to ensure faulty products do not reach the market. To address these challenges, selecting the right classifier is essential. However, this requires ranking classifiers in the context of application-specific preferences. For example, in medical testing, minimizing false negatives is critical since failing to diagnose a patient could have life-threatening consequences. In security systems, the focus is often on maximizing true negatives, accepting occasional false alarms as a trade-off for ensuring safety. Meanwhile, in quality control, false positives can be costly, as they may trigger unnecessary halts in production. Each application thus has unique requirements regarding the types of errors a classifier can tolerate.

A wide range of scores penalizing different types of errors are available in the literature. However, selecting

[1] This paper is the second of a trilogy. In a nutshell, paper A [28] presents an axiomatic framework and an infinite family of scores for ranking classifiers. In this paper (paper B), we particularize this framework to two-class classification and introduce the *Tile*, a visual tool that organizes these scores in a single 2D map. Finally, paper C [21] provides a guide to using the Tile according to four practical scenarios. For that, we present different Tile flavors that are applied to a real application.

Ranking scores for two-class classification

$$R_I : P \mapsto R_I(P) = \frac{E_P[SI]}{E_P[I]} = \frac{\sum_{\omega\in\Omega} S(\omega)I(\omega)P(\{\omega\})}{\sum_{\omega\in\Omega} I(\omega)P(\{\omega\})}$$

$\Omega = \{tn, fp, fn, tp\}$

$S = \mathbf{1}_{\{tn,tp\}}$

$$R_I : P \mapsto R_I(P) = \frac{I(tn)P(\{tn\}) + I(tp)P(\{tp\})}{I(tn)P(\{tn\}) + I(fp)P(\{fp\}) + I(fn)P(\{fn\}) + I(tp)P(\{tp\})}$$

📣 This continuum of scores contains some redundancy from the ranking point of view …

Ranking scores for two-class classification

🤯 If two ranking scores R_{I_1} and R_{I_2} are such that

$$\frac{I_1(tp)}{I_1(tn)+I_1(tp)} = \frac{I_2(tp)}{I_2(tn)+I_2(tp)} \quad \text{and} \quad \frac{I_1(fn)}{I_1(fp)+I_1(fn)} = \frac{I_2(fn)}{I_2(fp)+I_2(fn)}$$

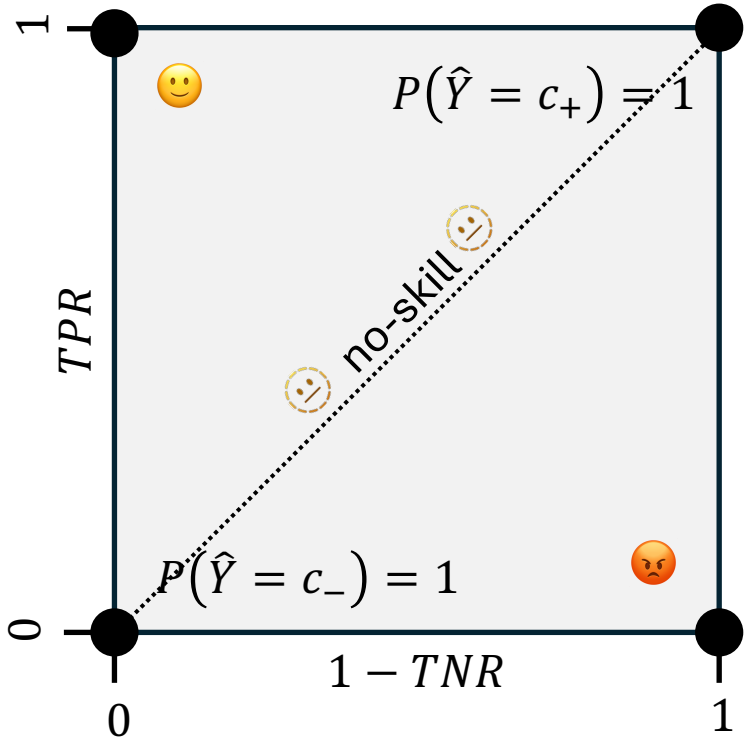
then the performance orderings induced by R_{I_1} and R_{I_2} are identical.

🥳 Great! This means that we can map the performance orderings induced by the ranking scores on the 2D square $[0,1]^2$ with the parameterization

$$\begin{cases} a(I) = \dfrac{I(tp)}{I(tn)+I(tp)} \\ b(I) = \dfrac{I(fn)}{I(fp)+I(fn)} \end{cases}$$

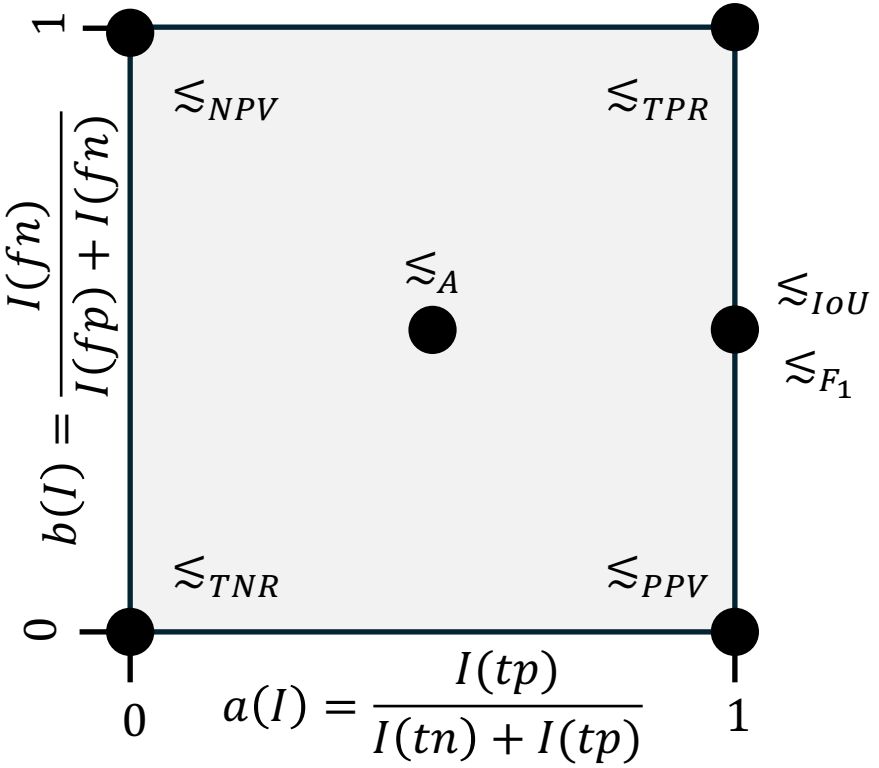
The Game-Changer

ROC space
Receiver Operating Characteristic
Since 1941



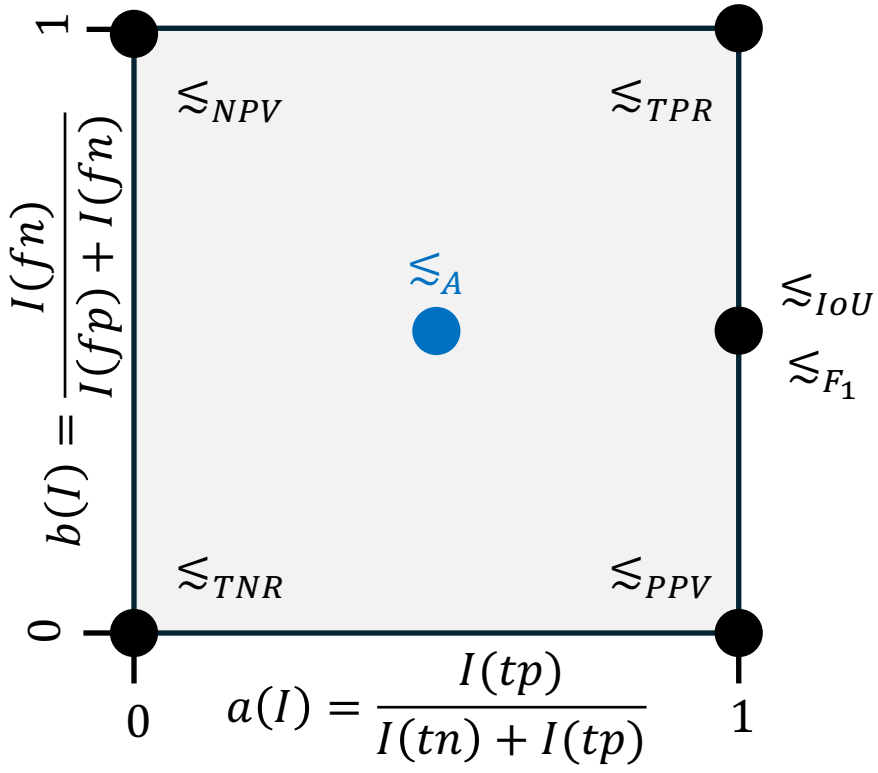
In this “space”, performances are projected as points.

83 years later, we innovate with
Our Tile
Since 2024



In this “space”, canonical ranking scores, performance orderings, and rankings are projected as points.

The Tile



Let us take:

- $I(tn) = 1$
- $I(fp) = 1$
- $I(fn) = 1$
- $I(tp) = 1$

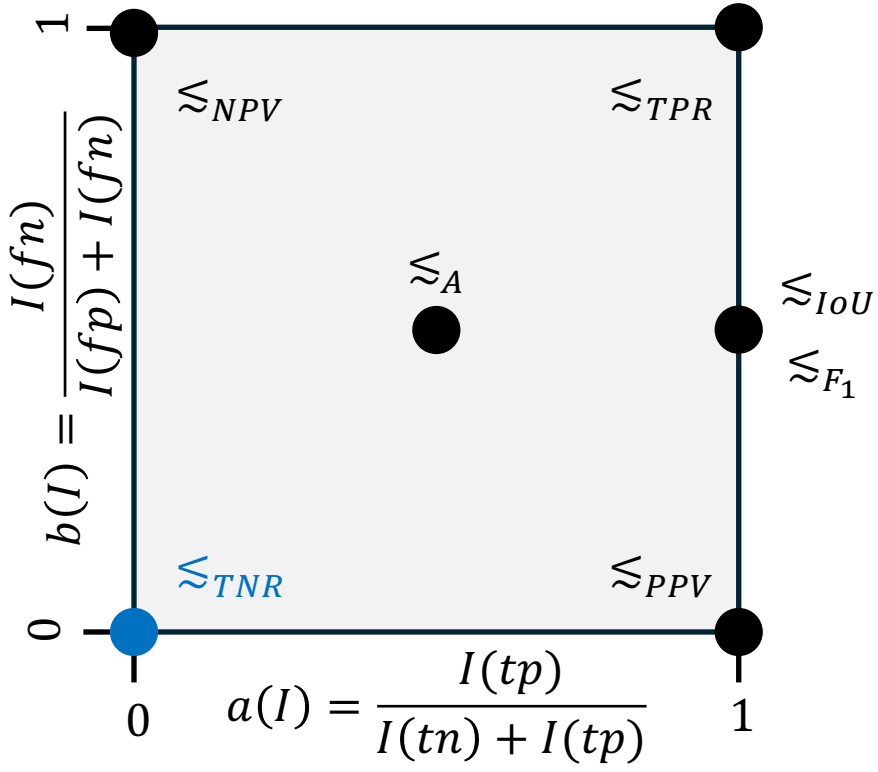
We obtain:

$$R_I(P) = \frac{P(\{tn\}) + P(\{tp\})}{P(\{tn\}) + P(\{fp\}) + P(\{fn\}) + P(\{tp\})} = \mathrm{A(P)}$$

📣 This is the *accuracy* $A: P \mapsto P(S = 1)$.

Synonym: matching coefficient

The Tile



Let us take:

- $I(tn) = 1$
- $I(fp) = 1$
- $I(fn) = 0$
- $I(tp) = 0$

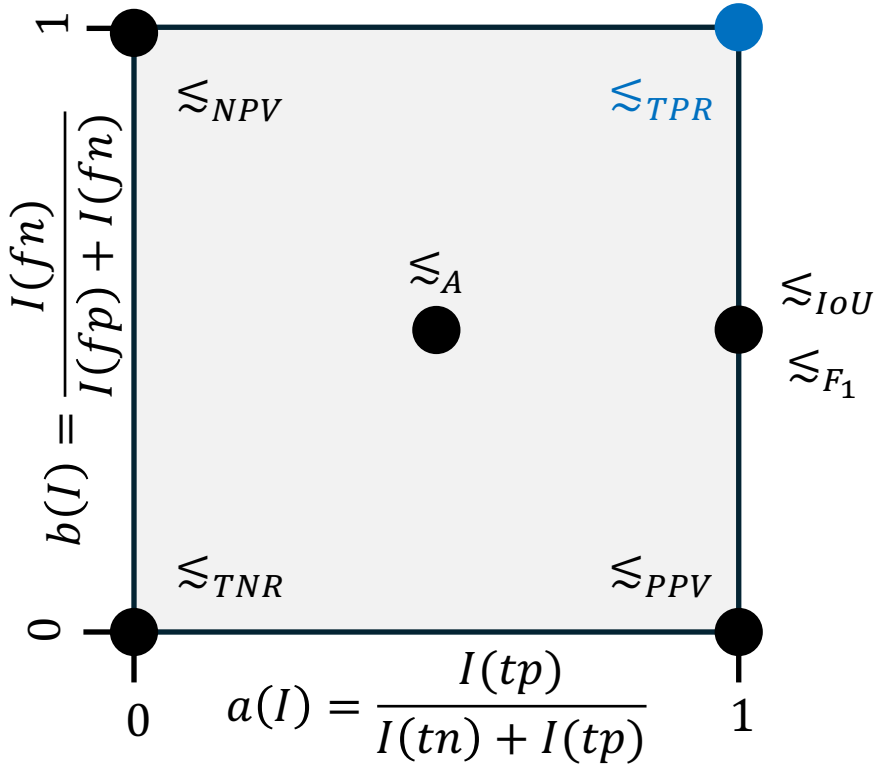
We obtain:

$$R_I(P) = \frac{P(\{tn\})}{P(\{tn\}) + P(\{fp\})} = \mathrm{TNR(P)}$$

📣 This is the *true negative rate* $TNR : P \mapsto P(S = 1|Y = c\_)$.

Synonyms: specificity, selectivity, inverse recall

The Tile



Let us take:

- $I(tn) = 0$
- $I(fp) = 0$
- $I(fn) = 1$
- $I(tp) = 1$

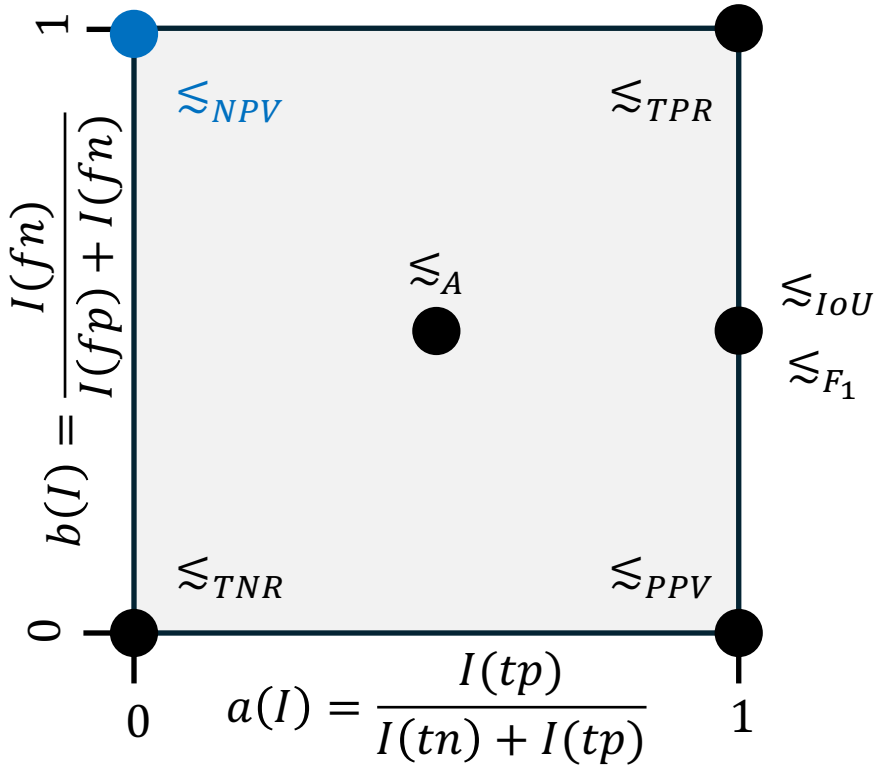
We obtain:

$$R_I(P) = \frac{P(\{tp\})}{P(\{fn\}) + P(\{tp\})} = \mathrm{TPR(P)}$$

📣 This is the *true positive rate* $\mathrm{TPR}: P \mapsto P(S = 1 | Y = c_+)$.

Synonyms: sensitivity, recall

The Tile



Let us take:

- $I(tn) = 1$
- $I(fp) = 0$
- $I(fn) = 1$
- $I(tp) = 0$

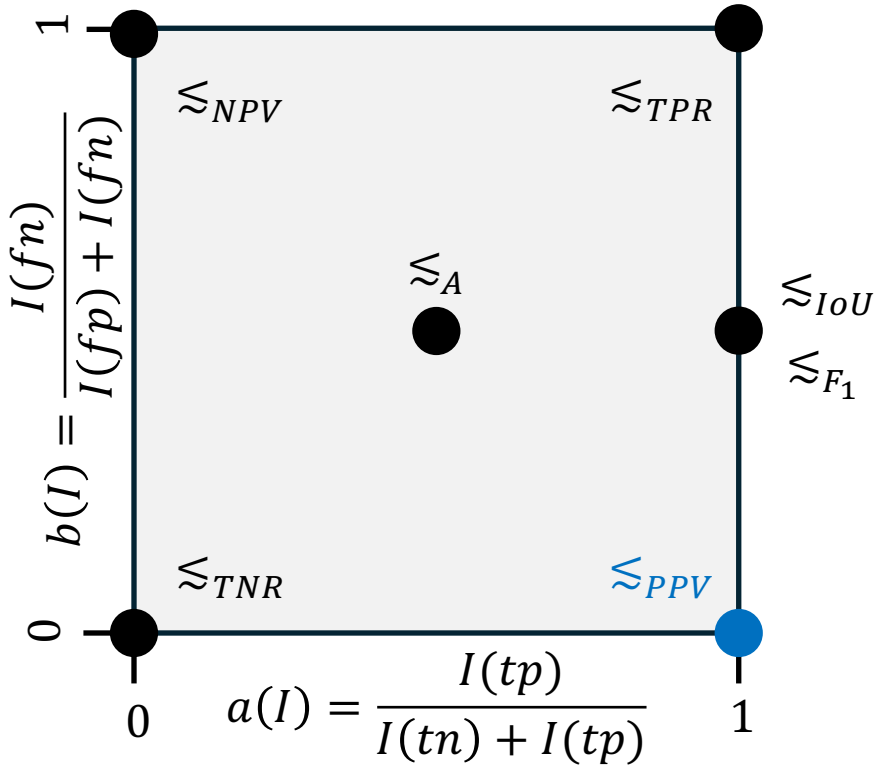
We obtain:

$$R_I(P) = \frac{P(\{tn\})}{P(\{tn\}) + P(\{fn\})} = \mathrm{NPV(P)}$$

📣 This is the *negative predictive value* $\mathrm{NPV} : P \mapsto P(S = 1 | \hat{Y} = c\_)$.

Synonym: inverse precision

The Tile



Let us take:

- $I(tn) = 0$
- $I(fp) = 1$
- $I(fn) = 0$
- $I(tp) = 1$

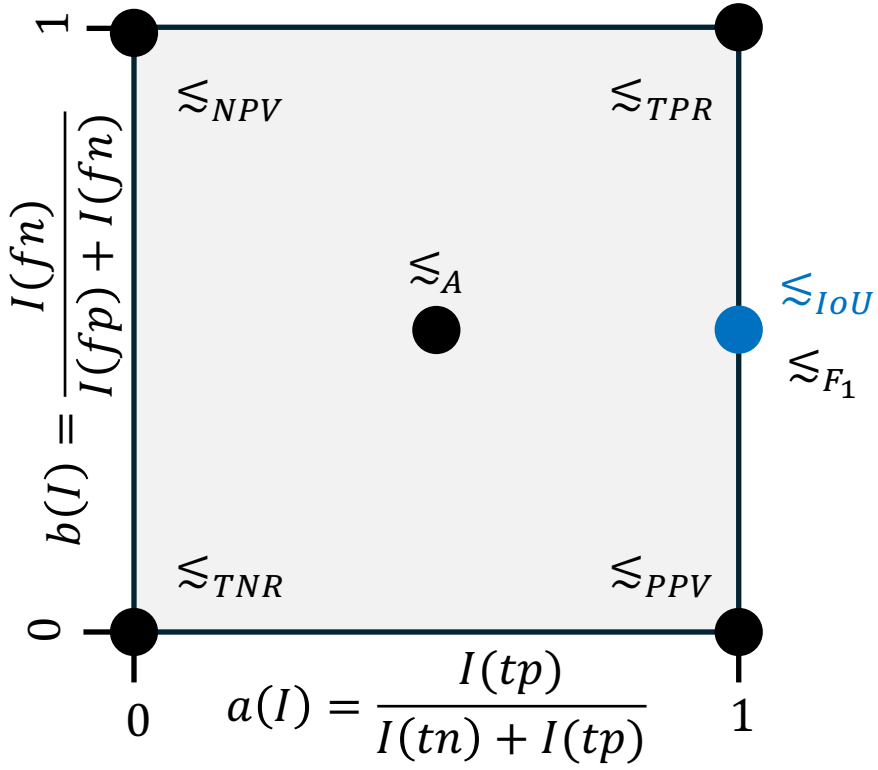
We obtain:

$$R_I(P) = \frac{P(\{tp\})}{P(\{fp\}) + P(\{tp\})} = \mathrm{PPV(P)}$$

📣 This is the *positive predictive value* $\mathrm{PPV} : P \mapsto P(S = 1 | \hat{Y} = c_+)$.

Synonym: precision

The Tile



Let us take:

- $I(tn) = 0$
- $I(fp) = 1$
- $I(fn) = 1$
- $I(tp) = 1$

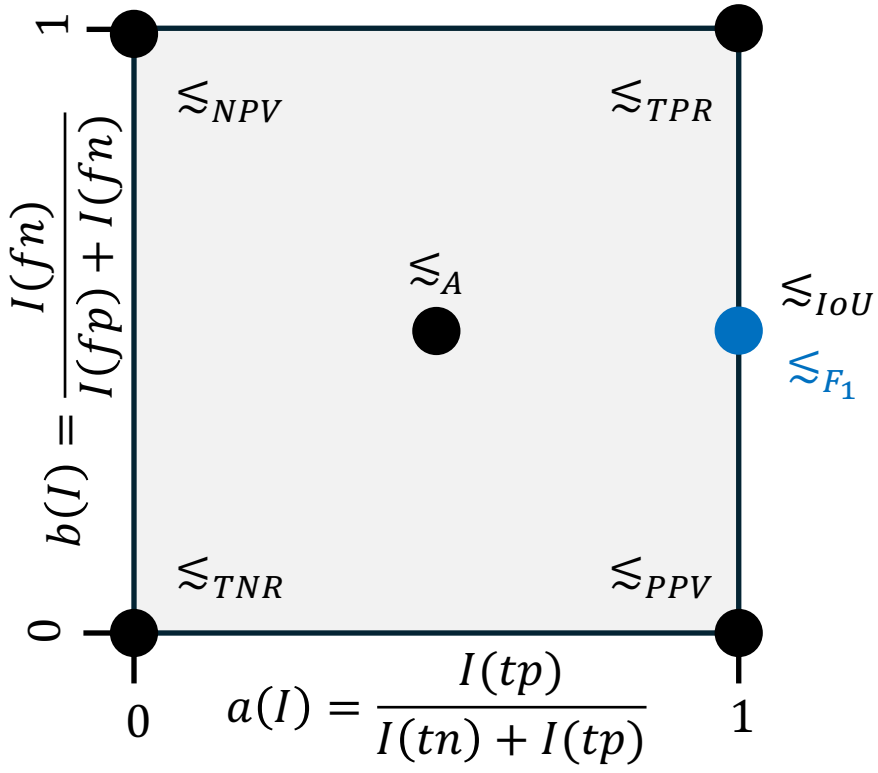
We obtain:

$$R_I(P) = \frac{P(\{tp\})}{P(\{fp\}) + P(\{fn\}) + P(\{tp\})} = \mathrm{IoU(P)}$$

📣 This is the *Intersection over Union* $\mathrm{IoU} : \mathrm{P} \mapsto P(S = 1 | Y = c_+ \vee \hat{Y} = c_+)$.

Synonyms: Jaccard's coef., Tanimoto coef., similarity, critical success index, G-measure

The Tile



Let us take:

- $I(tn) = 0$
- $I(fp) = 1$
- $I(fn) = 1$
- $I(tp) = 2$

We obtain:

$$R_I(P) = \frac{2P(\{tp\})}{P(\{fp\}) + P(\{fn\}) + 2P(\{tp\})} = \mathrm{F}_1(\mathrm{P})$$

📣 This is the *F-one score* $F_1 = ({}^1\!/_2\, PPV^{-1} + {}^1\!/_2\, TPR^{-1})^{-1}$.

Synonym: Dice-Sørensen coefficient

How do ranking scores rank-correlate with each other ?

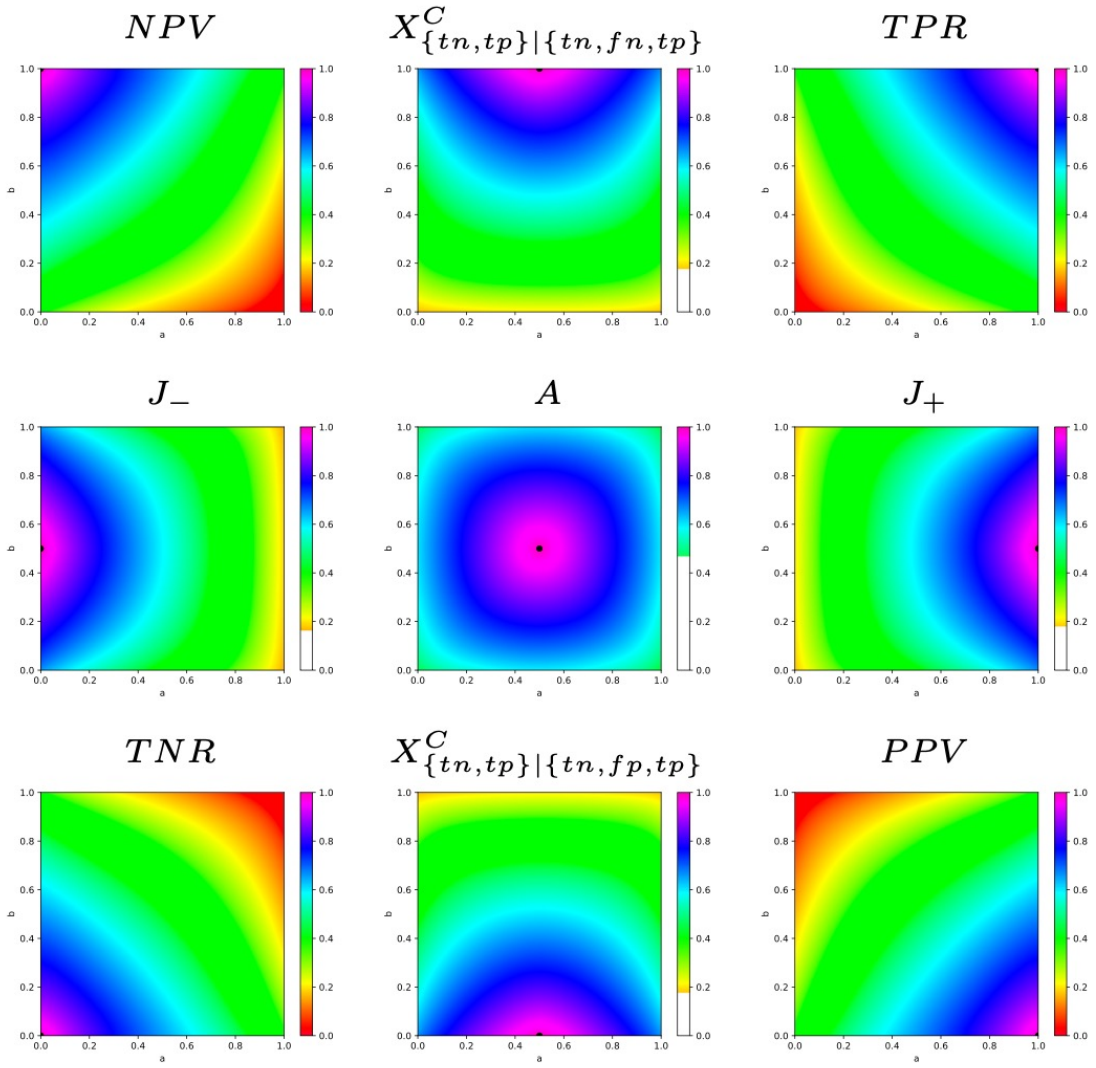
Example:

- uniform distribution of performances;
- Kendall's τ.

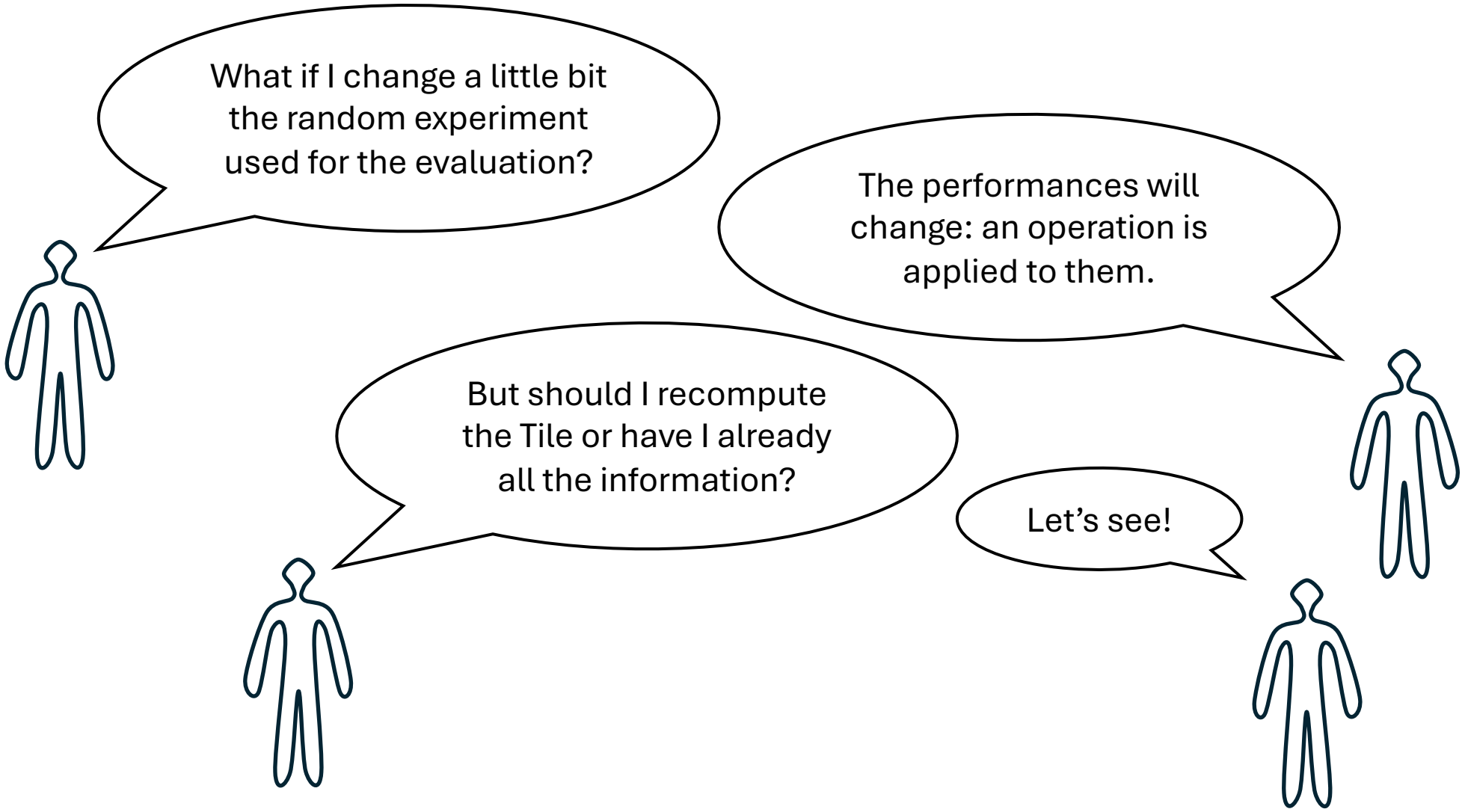
📣 We see that all performance orderings induced by ranking scores are different:

we need a 2D space to organize them.

📣 This is the case for most distributions of performances.



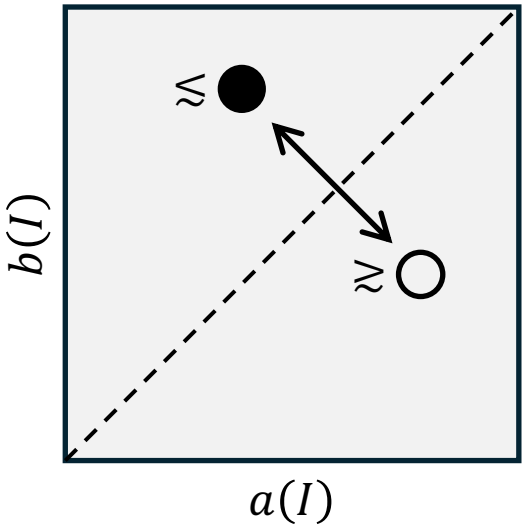
Operations on performances



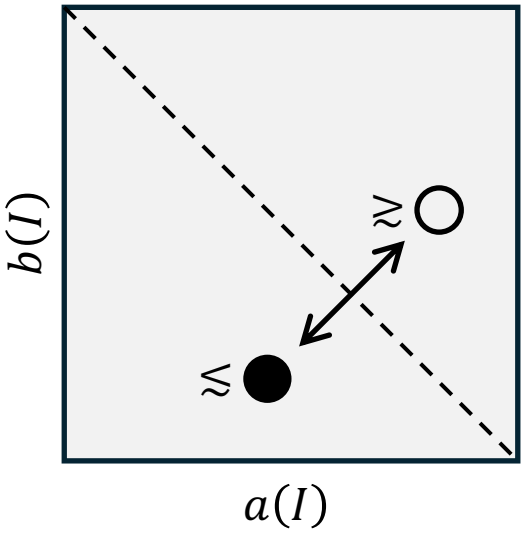
Operations on performances

When these operations are applied to all compared performances, there is no need to recompute the Tiles: applying geometric transforms suffice.

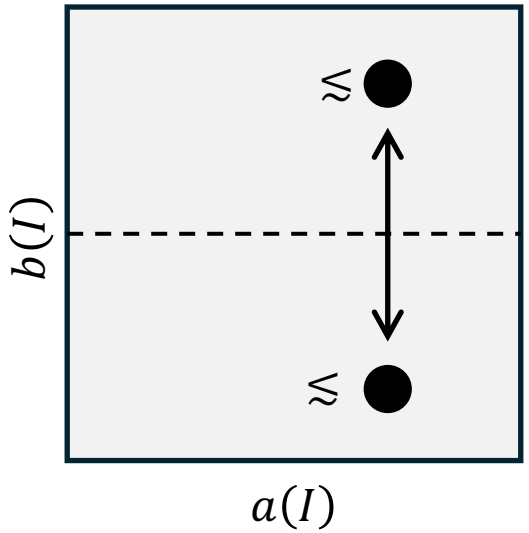
Change the predicted class
$\hat{Y} = c_- \leftrightarrow \hat{Y} = c_+$



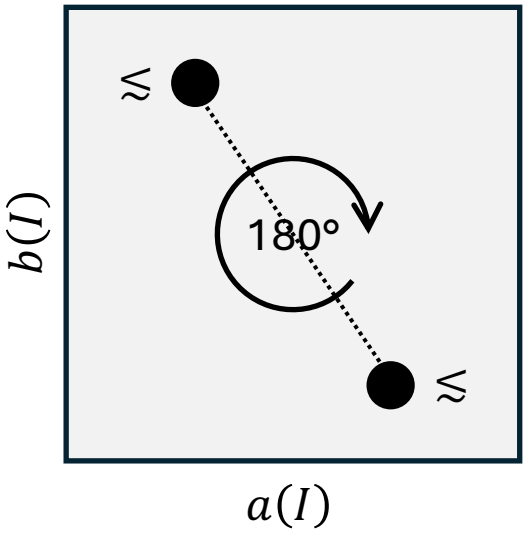
Change the ground-truth class
$Y = c_- \leftrightarrow Y = c_+$



Swap the oracle and classifier outputs
$Y \leftrightarrow \hat{Y}$

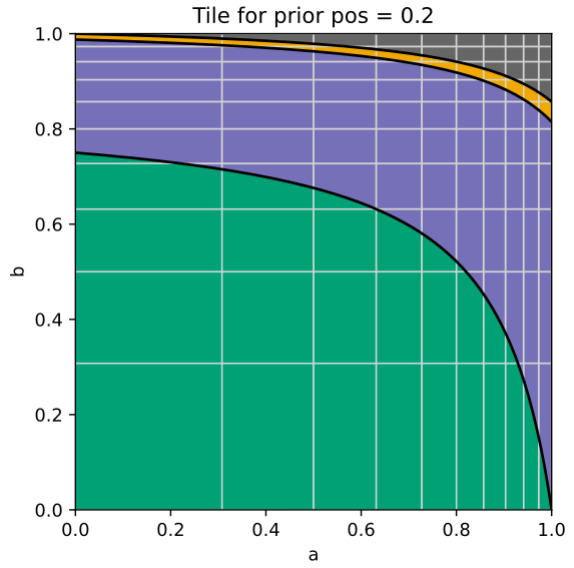


Swap the classes
$c_- \leftrightarrow c_+$

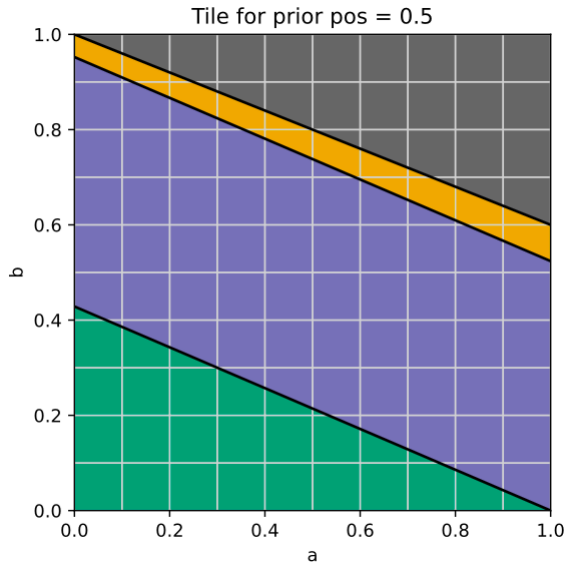


Operations on performances

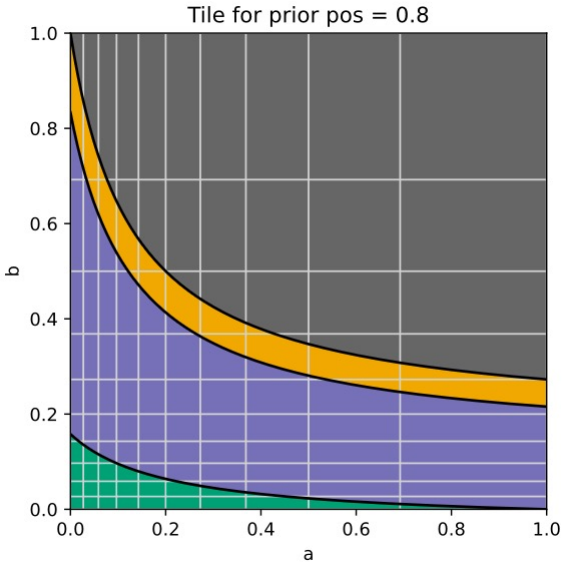
A *shift of class priors* moves the performance orderings (and thus the rankings) on the Tile.



	performances			
	P_- ●	P_1 ●	P_2 ●	P_+ ●
PTN	0.80	0.56	0.40	0.00
PFP	0.00	0.24	0.40	0.80
PFN	0.20	0.06	0.04	0.00
PTP	0.00	0.14	0.16	0.20



	performances			
	P_- ●	P_1 ●	P_2 ●	P_+ ●
PTN	0.50	0.35	0.25	0.00
PFP	0.00	0.15	0.25	0.50
PFN	0.50	0.15	0.10	0.00
PTP	0.00	0.35	0.40	0.50



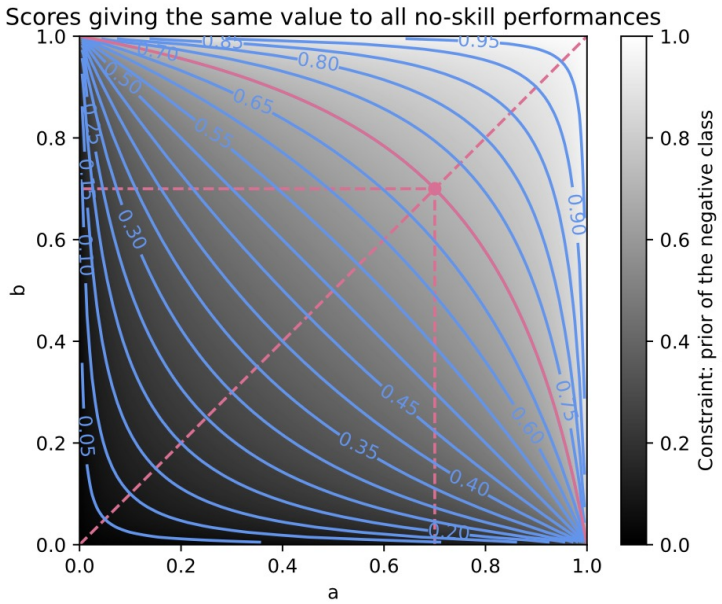
	performances			
	P_- ●	P_1 ●	P_2 ●	P_+ ●
PTN	0.20	0.14	0.10	0.00
PFP	0.00	0.06	0.10	0.20
PFN	0.80	0.24	0.16	0.00
PTP	0.00	0.56	0.64	0.80

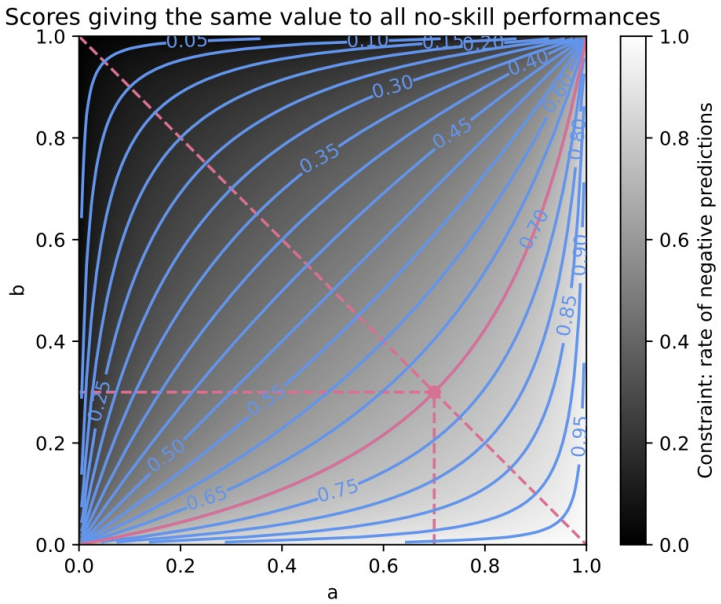
About no-skill performances

ℹ️ *No-skill performances* P are such that $Y \perp_P \hat{Y}$

$$\Leftrightarrow P\left(Y = c_1, \hat{Y} = c_2\right) = P(Y = c_1)\, P\left(\hat{Y} = c_2\right) \forall\, c_1, c_2 \in \mathbb{C}$$

📣 Some ranking scores put all no-skill performances on an equal footing (they become equivalent) if one restricts the compared performances to *some given class priors* or to *some given prediction rates*: see the curves below.



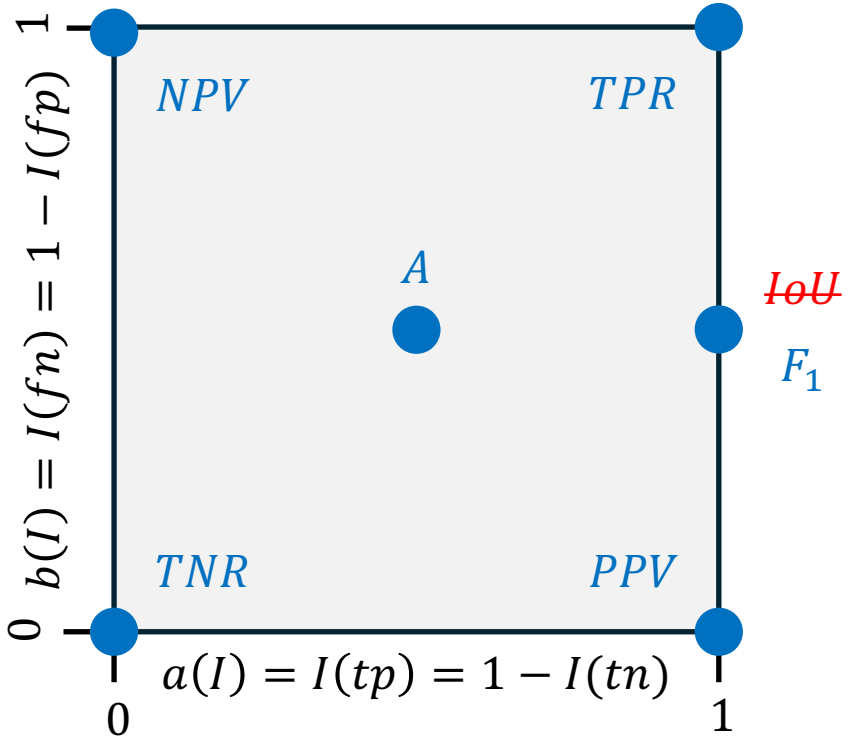


Canonical ranking scores

Objective: having a sole score in each point of the Tile.

Definition: the canonical ranking scores are the ranking scores

$$R_I \text{ with } \begin{cases} I(tn) + I(tp) = 1 \\ I(fp) + I(fn) = 1 \end{cases}$$



Part 6:

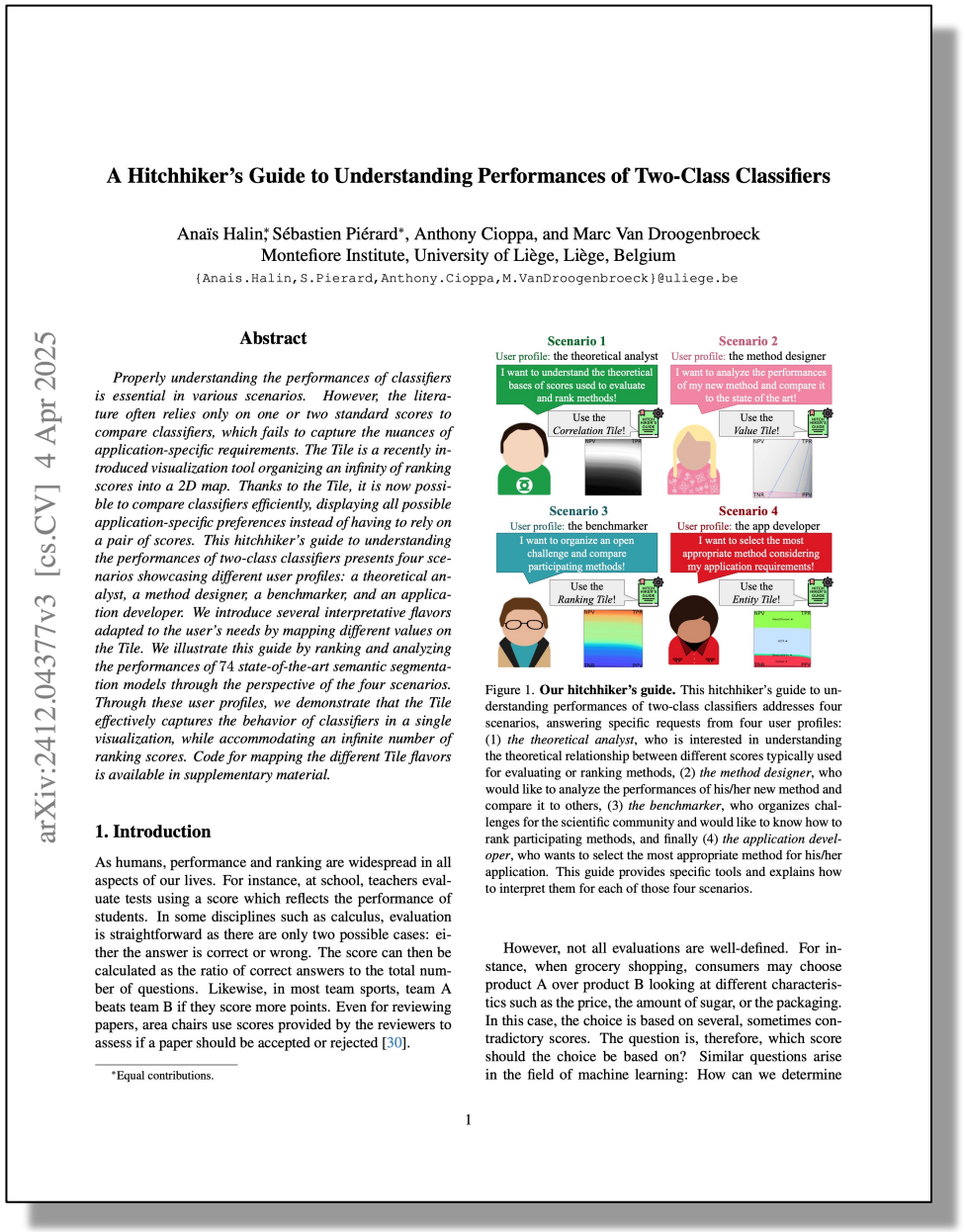
The Flavors

This presentation concerns original research carried out at the Montefiore Institute of the University of Liège. And, this is the bust of Georges Montefiore-Levi (1832-1906), sculpted by Thomas Vinçotte in 1904.

Bibliography

This part of the presentation is based on the following paper:

Anaïs Halin, Sébastien Piérard, Anthony Cioppa, and Marc Van Droogenbroeck. ***A hitchhiker's guide to understanding performances of two-class classifiers.*** arXiv, abs/2412.04377, 2024. URL https://doi.org/10.48550/arXiv.2412.04377.



A Hitchhiker's Guide to Understanding Performances of Two-Class Classifiers

Anaïs Halin*, Sébastien Piérard*, Anthony Cioppa, and Marc Van Droogenbroeck
Montefiore Institute, University of Liège, Liège, Belgium
{Anais.Halin,S.Pierard,Anthony.Cioppa,M.VanDroogenbroeck}@uliege.be

Abstract

Properly understanding the performances of classifiers is essential in various scenarios. However, the literature often relies only on one or two standard scores to compare classifiers, which fails to capture the nuances of application-specific requirements. The Tile is a recently introduced visualization tool organizing an infinity of ranking scores into a 2D map. Thanks to the Tile, it is now possible to compare classifiers efficiently, displaying all possible application-specific preferences instead of having to rely on a pair of scores. This hitchhiker's guide to understanding the performances of two-class classifiers presents four scenarios showcasing different user profiles: a theoretical analyst, a method designer, a benchmarker, and an application developer. We introduce several interpretative flavors adapted to the user's needs by mapping different values on the Tile. We illustrate this guide by ranking and analyzing the performances of 74 state-of-the-art semantic segmentation models through the perspective of the four scenarios. Through these user profiles, we demonstrate that the Tile effectively captures the behavior of classifiers in a single visualization, while accommodating an infinite number of ranking scores. Code for mapping the different Tile flavors is available in supplementary material.

1. Introduction

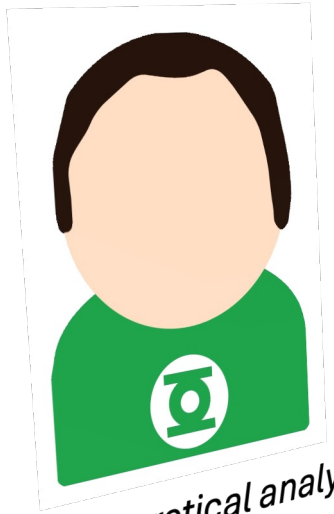
As humans, performance and ranking are widespread in all aspects of our lives. For instance, at school, teachers evaluate tests using a score which reflects the performance of students. In some disciplines such as calculus, evaluation is straightforward as there are only two possible cases: either the answer is correct or wrong. The score can then be calculated as the ratio of correct answers to the total number of questions. Likewise, in most team sports, team A beats team B if they score more points. Even for reviewing papers, area chairs use scores provided by the reviewers to assess if a paper should be accepted or rejected [30].

*Equal contributions.

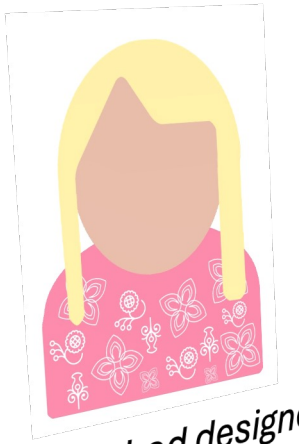
Figure 1. **Our hitchhiker's guide.** This hitchhiker's guide to understanding performances of two-class classifiers addresses four scenarios, answering specific requests from four user profiles: (1) *the theoretical analyst*, who is interested in understanding the theoretical relationship between different scores typically used for evaluating or ranking methods, (2) *the method designer*, who would like to analyze the performances of his/her new method and compare it to others, (3) *the benchmarker*, who organizes challenges for the scientific community and would like to know how to rank participating methods, and finally (4) *the application developer*, who wants to select the most appropriate method for his/her application. This guide provides specific tools and explains how to interpret them for each of those four scenarios.

However, not all evaluations are well-defined. For instance, when grocery shopping, consumers may choose product A over product B looking at different characteristics such as the price, the amount of sugar, or the packaging. In this case, the choice is based on several, sometimes contradictory scores. The question is, therefore, which score should the choice be based on? Similar questions arise in the field of machine learning: How can we determine

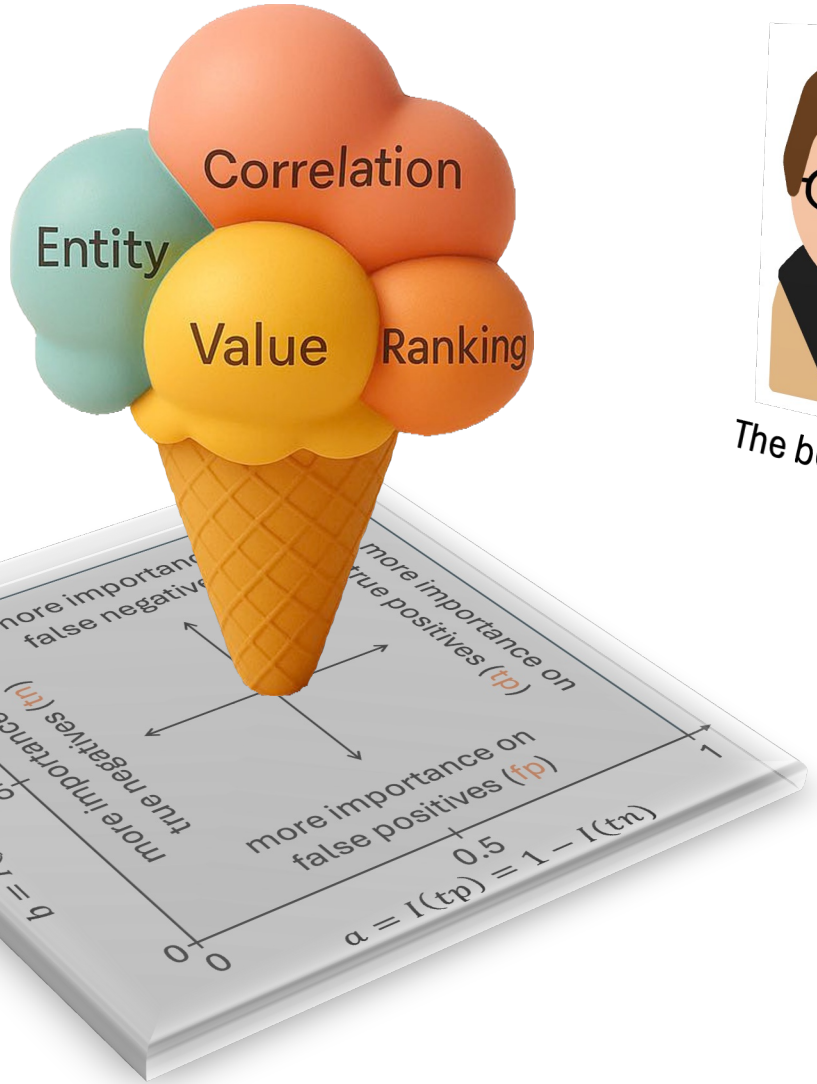
The Tile and its Flavors

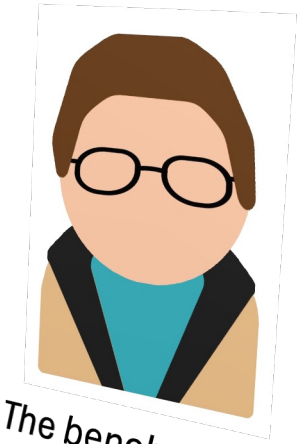


The theoretical analyst

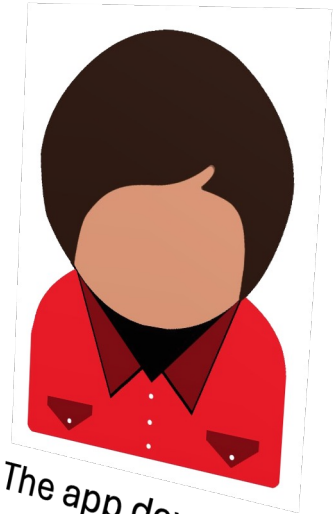


The method designer





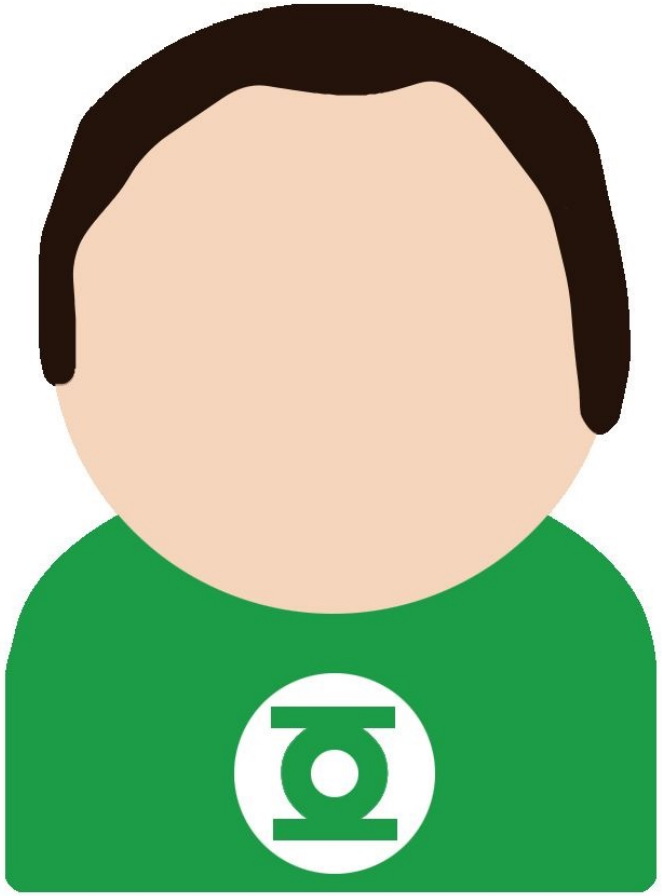
The benchmarker



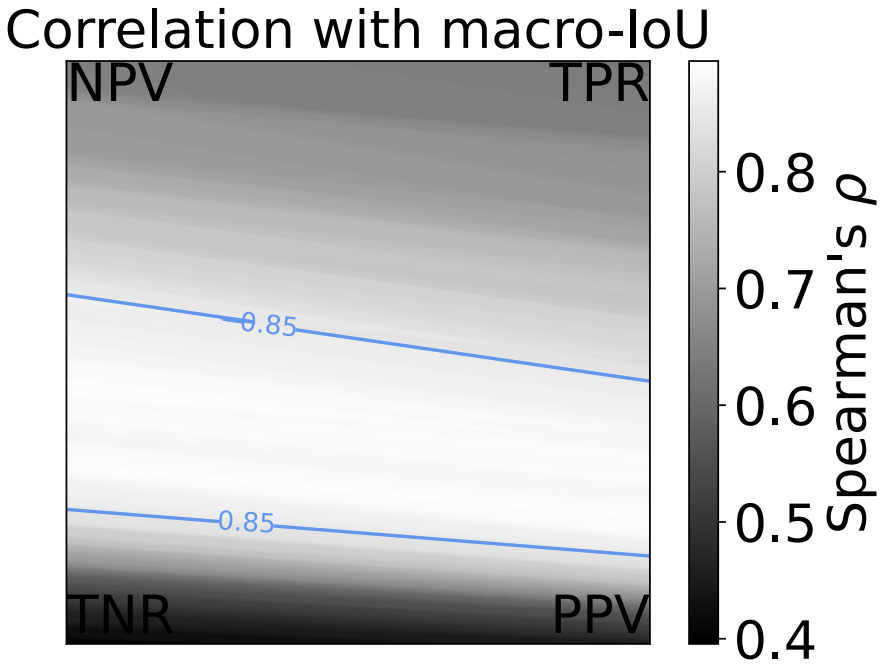
The app developer

Example of Flavor: *Correlation Tiles*

« What is the behavior of this particular score? »

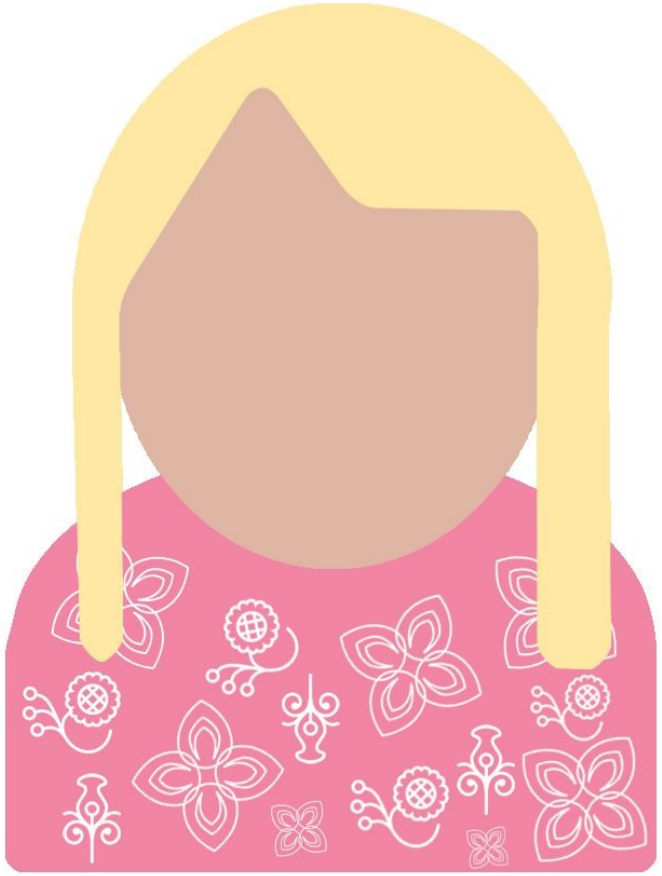


For a theoretical analyst



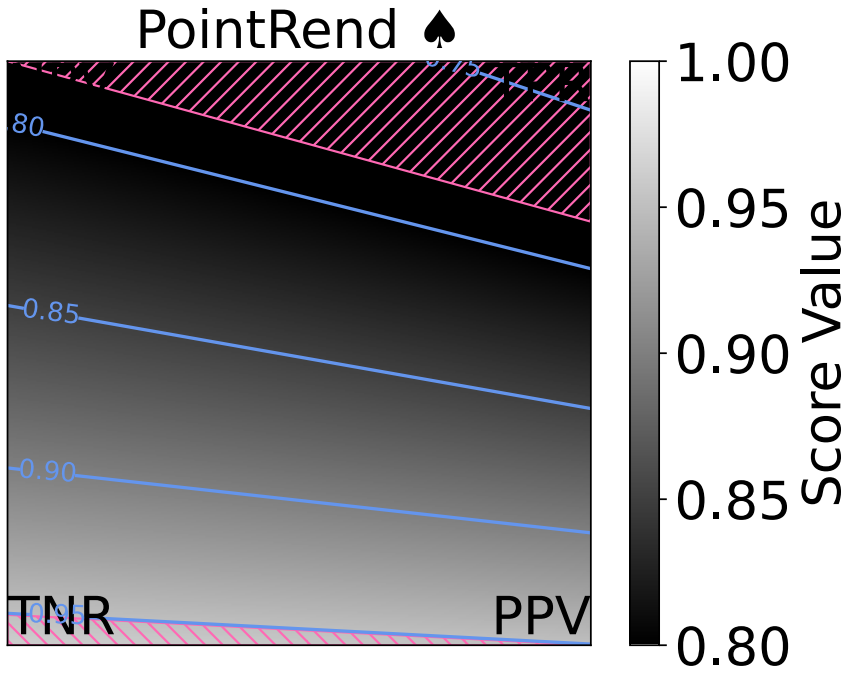
Suggested interpretation. The score is not perfectly rank-correlated with any of the ranking scores. This is usually the sign that it cannot be used to establish meaningful performance orderings. However, it has a correlation of about 0.85 with the accuracy (center of the Tile), which is quite high.

Example of Flavor: *Value Tiles*



For a method designer

« What are the strengths and weaknesses of the method I'm developing? »

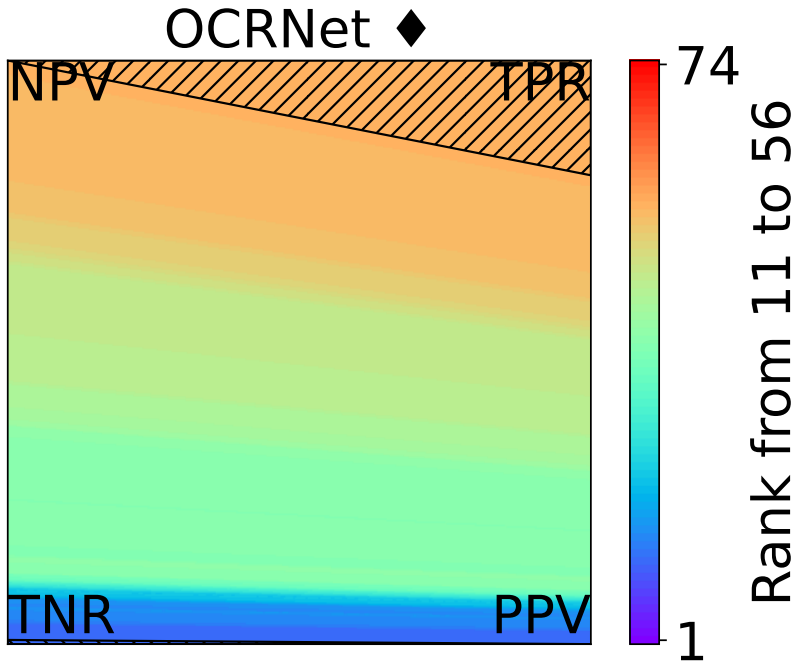


Suggested interpretation. The values taken by the canonical ranking scores show that the strength of the method is when importance is put on the TNs and FPs and its weakness is when the importance is put on the TPs and FNs. Moreover, the region of the Tile where no-skill classifiers perform better is huge.

Example of Flavor: *Ranking Tiles*

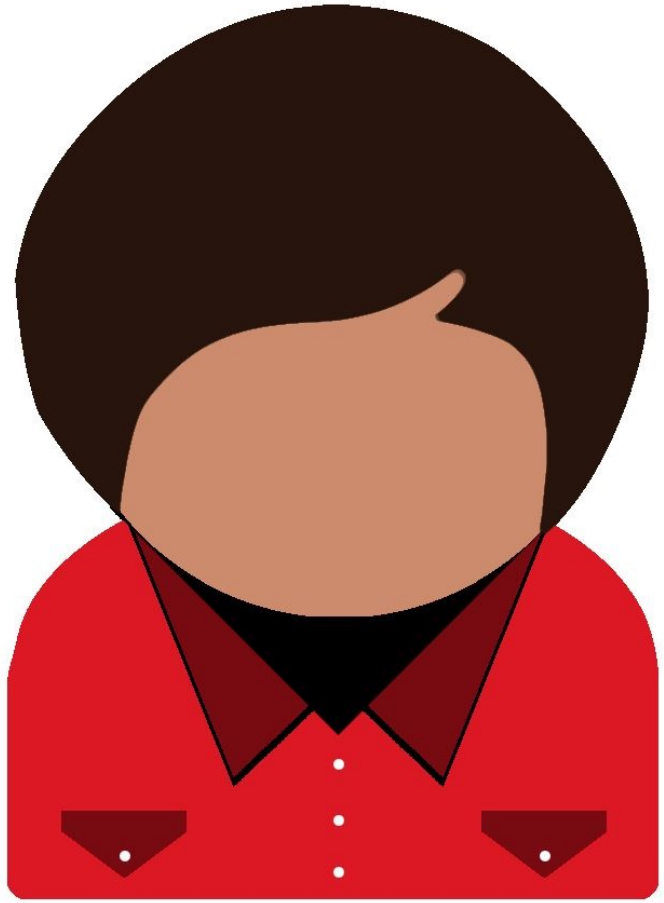
For a benchmarker

« How does this particular method rank w.r.t. the application specific preferences? »



Suggested interpretation. The method is ranked 11th to 56th , among 74, depending on the application-specific preferences. It performs better than most methods in the bottom of the Tile, where more importance is put on FPs than on FNs.

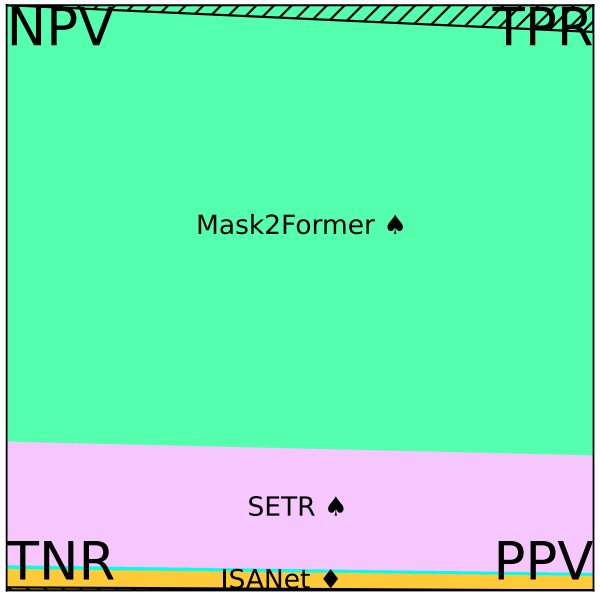
Example of Flavor: *Entity Tiles*



For an app developer

« Show me the most appropriate method w.r.t. the application specific preferences! »

Who's first?



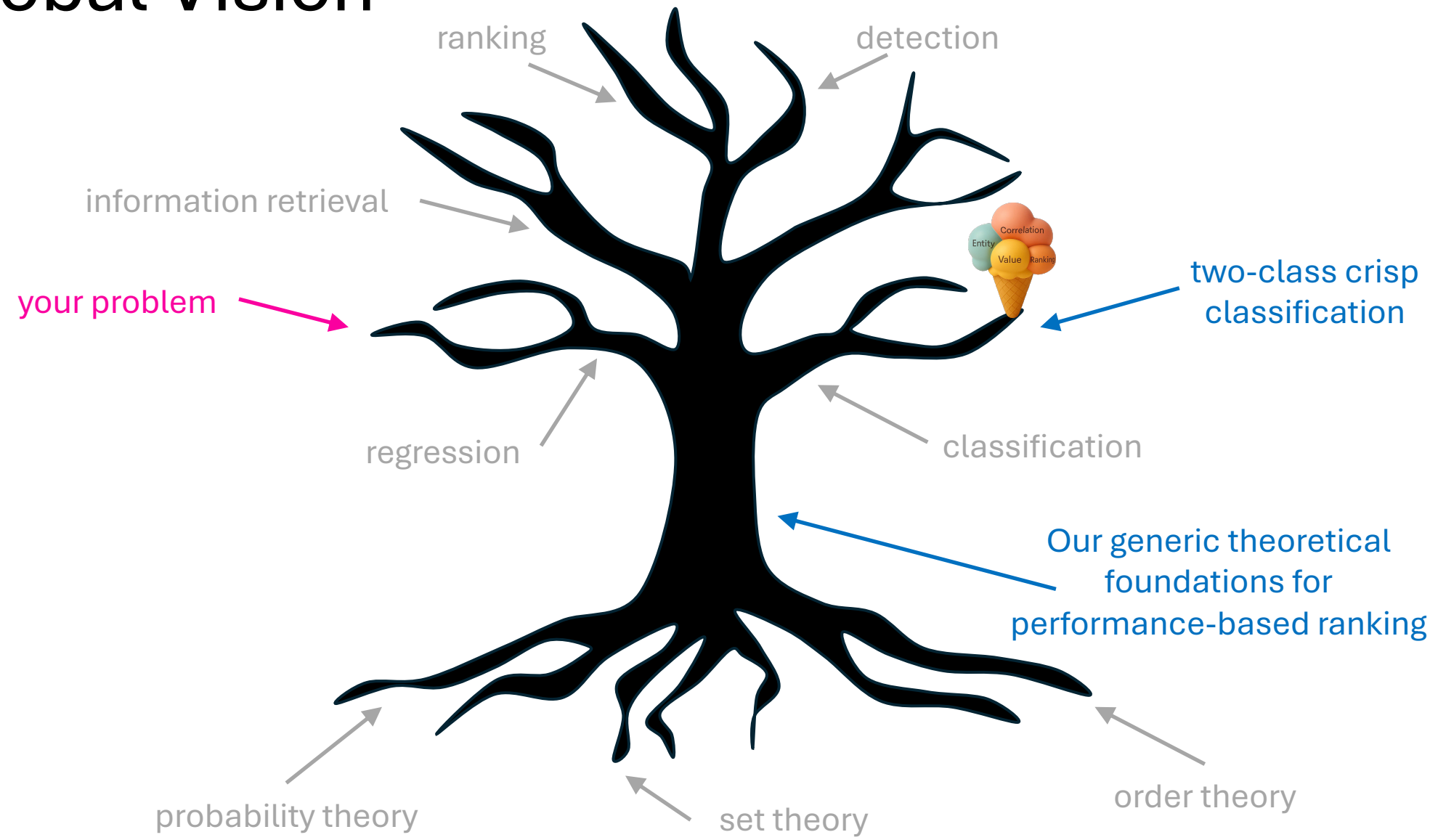
Suggested interpretation. Unless you give much more importance to the FPs than to the FNs (bottom of the tile), Mask2Former and SETR are the methods you should implement.

Part 7:

Postface

This presentation concerns original research carried out at the Montefiore Institute of the University of Liège. And, this is the bust of Georges Montefiore-Levi (1832-1906), sculpted by Thomas Vinçotte in 1904.

Global Vision



TRΔIL summer workshop ‘25, London, UK


Do you want to get involved in this project?

- Become a member of Trusted AI Labs (TRΔIL);
- join us in London from August 25th to September 5th;
- for the 6th summer workshop on artificial intelligence;
- and get involved in project n°3, *Let’s Tile Together!*

Registration deadline: June 20th

https://trail.ac/en/trail-summer-workshops/

The objectives:

1. a first version of an open-source library to produce Tiles;
2. easy to use by all researchers;
3. flexible to handle a wide range of flavors;
4. ready to be published on GitHub, on the TRAIL factory, and on PyPI;
5. and ready to be submitted to a journal (e.g. JOSS).

Bibliography

1) Sébastien Piérard, Anaïs Halin, Anthony Cioppa, Adrien Deliège, and Marc Van Droogenbroeck. ***Foundations of the theory of performance-based ranking.*** In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, Tennessee, 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, abs/2412.04309, 2024. URL https://doi.org/10.48550/arXiv.2412.04309.

3) Anaïs Halin, Sébastien Piérard, Anthony Cioppa, and Marc Van Droogenbroeck. ***A hitchhiker's guide to understanding performances of two-class classifiers.*** arXiv, abs/2412.04377, 2024. URL https://doi.org/10.48550/arXiv.2412.04377.

And more to come, stay tuned!