Performance-Based Ranking of Building Damage Assessment Methods

Marc Van Droogenbroeck

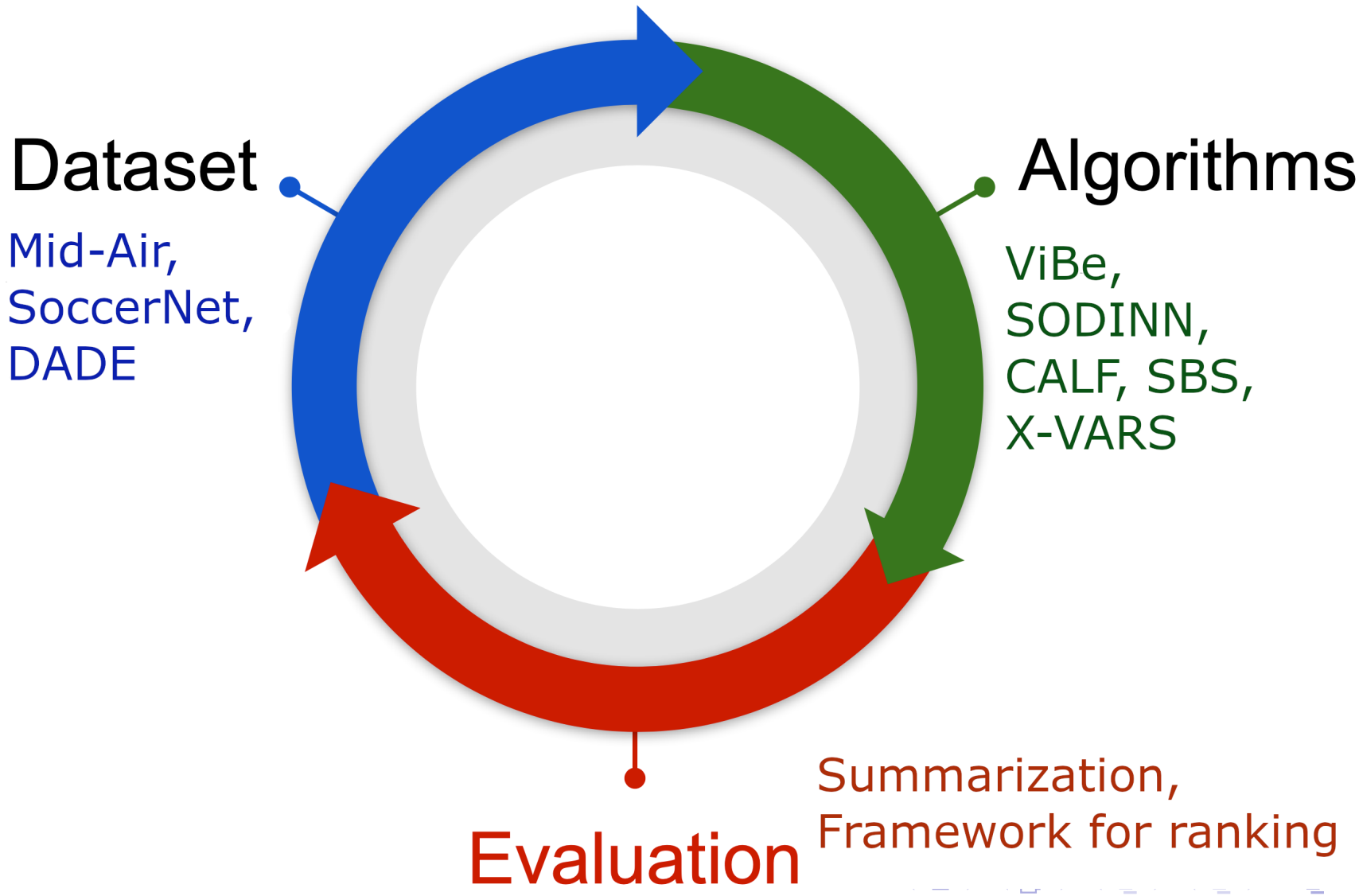
Department of Electrical Engineering and Computer Science
University of Liège, Belgium

November 2026



Vision and Image Understanding Laboratory (VIULab)

A glimpse of our activities at the VIULab (according to three pillars)



Main sources I

1. Several presentations by Dr. **Sébastien Piérard** and Marc Van Droogenbroeck
2. Sébastien Piérard and Marc Van Droogenbroeck. "Summarizing the performances of a background subtraction algorithm measured on several videos." In: *IEEE Int. Conf. Image Process. (ICIP)*. Abu Dhabi, United Arab. Emir., Oct. 2020, pp. 3234–3238. DOI: 10.1109/ICIP40778.2020.9190865
3. Sébastien Piérard et al. "Foundations of the Theory of Performance-Based Ranking." In: *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Nashville, TN, USA: IEEE, June 2025, pp. 14293–14302. DOI: 10.1109/cvpr52734.2025.01333
4. Sébastien Piérard et al. "The Tile: A 2D Map of Ranking Scores for Two-Class Classification." In: *arXiv* abs/2412.04309 (2024). DOI: 10.48550/arXiv.2412.04309. arXiv: 2412.04309

Main sources II

5. Sébastien Piérard et al. "A Methodology to Evaluate Strategies Predicting Rankings on Unseen Domains." In: *IEEE Int. Conf. Multimedia Expo Work. (ICMEW), Work. Big Surveill. Data Anal. Process. (big-surv)*. Nantes, France: IEEE, June 2025, pp. 1–6. DOI: 10.1109/icmew68306.2025.11152102
6. Sébastien Piérard, Adrien Deliège, and Marc Van Droogenbroeck. "Multi-domain performance analysis with scores tailored to user preferences." In: *arXiv* abs/2512.08715 (2025). DOI: 10.48550/arXiv.2512.08715. arXiv: 2512.08715
7. Sébastien Piérard, Adrien Deliège, and Marc Van Droogenbroeck. "What Is the Optimal Ranking Score Between Precision and Recall? We Can Always Find It and It Is Rarely F_1." In: *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Denver, CO, USA: IEEE, June 2026, pp. 9722–9731
8. Anastasia Zakharova et al. "Event Detection in Videos: A Framework for the Development of New Methods." In: *arXiv* abs/2607.04372 (2026). DOI: 10.48550/arXiv.2607.04372. arXiv: 2607.04372

Main sources III

9. Anthony Cioppa et al. "Revisiting the Task of Building Damage Assessment by Satellite Imaging." In: *arXiv* abs/2610.13857 (2026). DOI: 10.48550/arXiv.2610.13857. arXiv: 2610.13857
10. Sébastien Piérard, Adrien Deliège, and Marc Van Droogenbroeck. "The Optimal Ranking Tradeoffs Between Two Ranking Scores in Two-Class Classification." In: *arXiv* abs/2610.23294 (2026). DOI: 10.48550/arXiv.2610.23294. arXiv: 2610.23294

Outline

1. Introduction to performance evaluation
   - Motivation
   - Illustration: crisp two-class classification (binary classification)
   - More about the definition of a task
2. Evaluation of building damage assessment methods: a first glance
3. A theory for performance evaluation
4. Building an evaluation pipeline
5. Towards a standardized evaluation protocol

Scoring and comparing are very common tasks in daily life

Ask yourself: how many times a day/week do you

- rate/score/evaluate services?
- compare products?

There are three common evaluation "tasks" in AI:

- **Scoring** *one* method.
- **Comparing** *two* methods.
- **Ranking** *a series* of methods.

Scoring and comparing are very common tasks in daily life

Ask yourself: how many times a day/week do you

- rate/score/evaluate services?
- compare products?

There are three common evaluation "tasks" in AI:

- **Scoring** *one* method.
- **Comparing** *two* methods.
- **Ranking** *a series* of methods.

Why is benchmarking even more important in a world overwhelmed by AI?

AI has changed the development of algorithms over time:

1. Data play a larger role (more and curated data is better).
2. With LLMs, coding becomes easier; in particular, the development of variants can largely be automated.
3. Methods (and scientific papers) proliferate, with variable quality.
4. New tasks appear and proliferate, sometimes as variants of established tasks.

In this AI-dominated world, the evaluation of algorithms remains delicate and poorly understood.

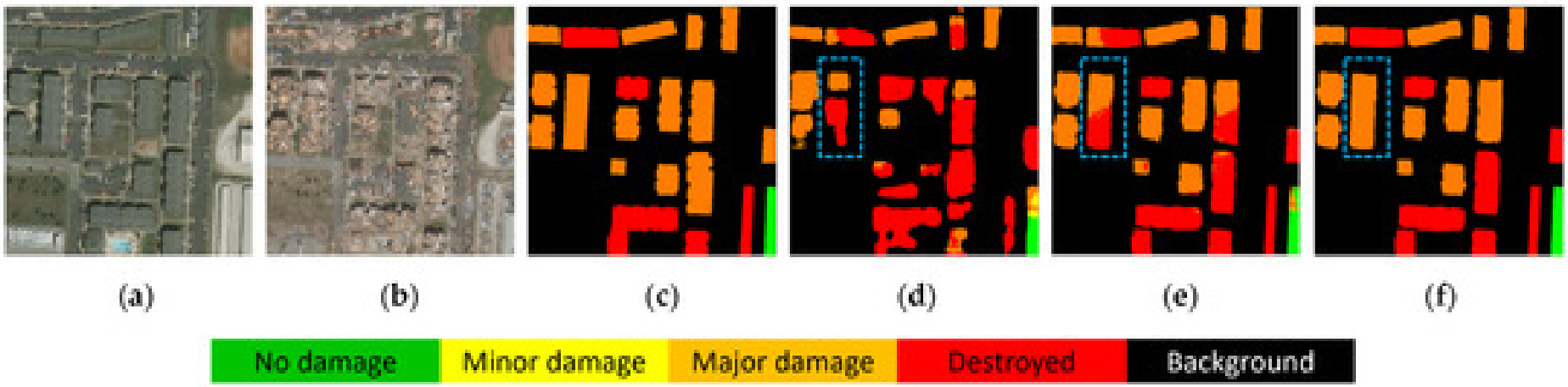
Aim of this talk

Explain current practices, propose a framework (one model among others), and bring partial solutions to the question of evaluation. The ultimate goal: work towards a complete evaluation "scenario" covering the three pillars (data, method, evaluation) for BDA [26].

There is a lot to say about how we conduct the evaluation

Illustration with two tasks:

- Tasks of Building Damage Assessment (BDA) by remote sensing



- Illegal Waste Dumping Detection (IWDD) contest: the difficulty of generalization

WACV 2026
Tucson, Arizona • Mar 6 - 10, 2026

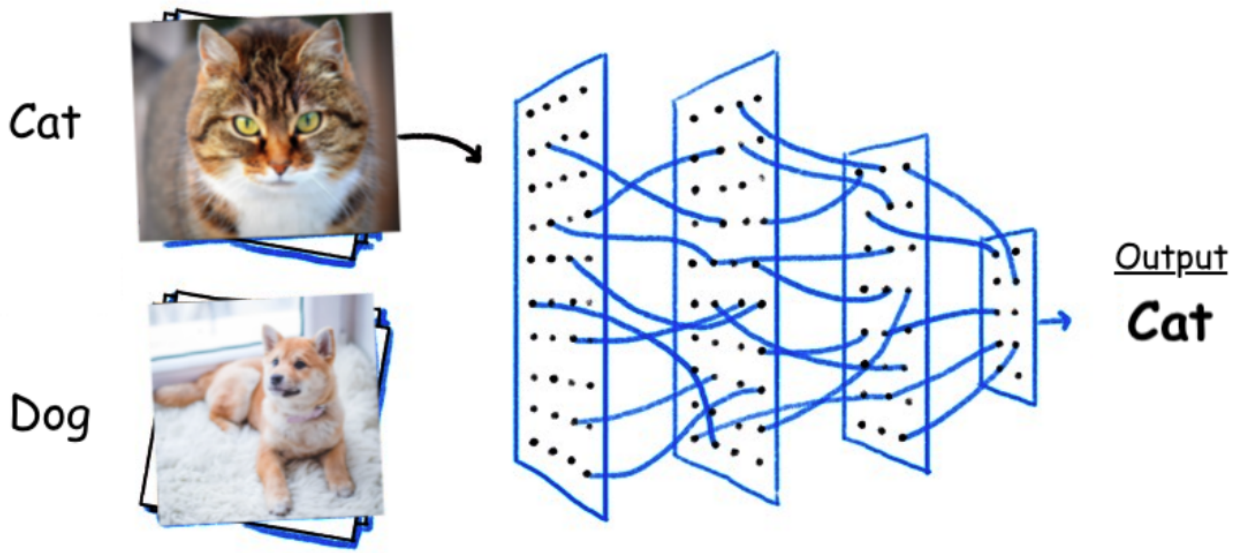
INTERNATIONAL CONTEST ON ILLEGAL WASTE DUMPING DETECTION

Smart City Surveillance

Outline

1. Introduction to performance evaluation
   - Motivation
   - Illustration: crisp two-class classification (binary classification)
   - More about the definition of a task
2. Evaluation of building damage assessment methods: a first glance
3. A theory for performance evaluation
4. Building an evaluation pipeline
5. Towards a standardized evaluation protocol

Our case study for illustration: crisp two-class classification (binary classification)



Notations. Let us consider that Cat is the negative class (-) and Dog is the positive class (+); note that these attributions are arbitrary!

There are two notions of classes:

[true class or ground truth] is denoted by $y \in \{c^-, c^+\}$.

[estimated class] is denoted by $\widehat{y} \in \{c^-, c^+\}$

Confusion matrix

Definition (Confusion matrix/contingency table)

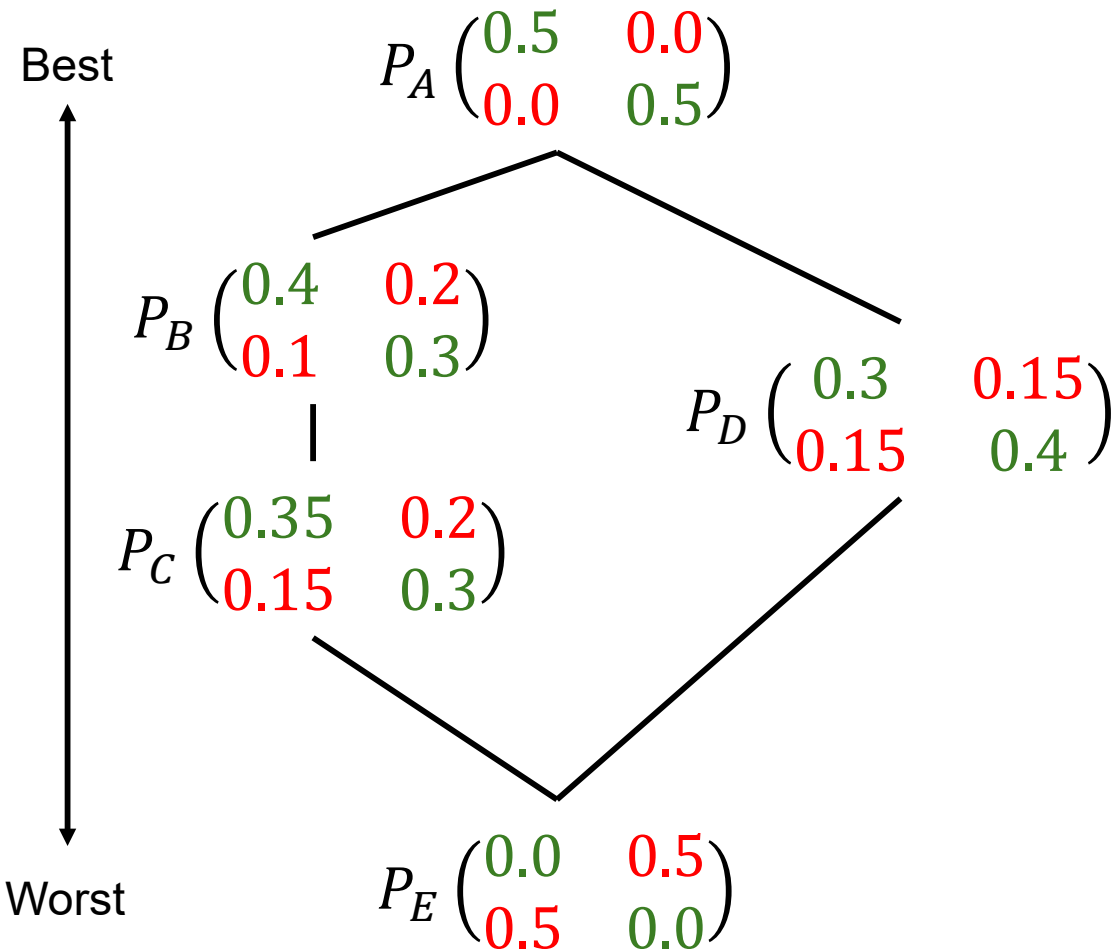
There are four possibilities, described by the confusion matrix:

		prediction $\widehat{y}$	
		c^+	c^-
real class y	c^+	True Positive (TP)	False Negative (FN)
	c^-	False Positive (FP)	True Negative (TN)

Remarks:

- There are four quantities: TP, FN, FP, TN.
- It is easy to extend a confusion matrix to more than two classes.
- The matrix defines a 3D manifold in a 4D space.

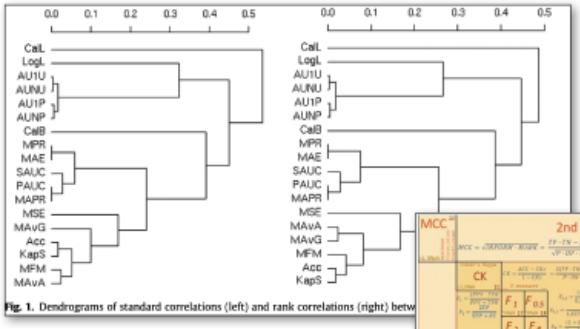
Is it difficult to rank classifiers?



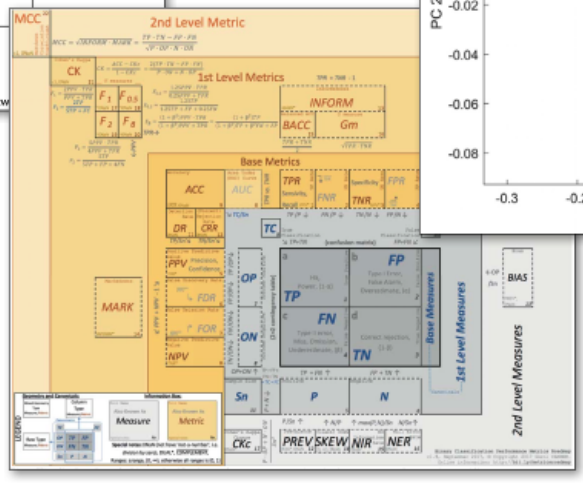
Strategies for evaluation are diverse and ... disappointing

Organizing scores: the disappointi… current practices

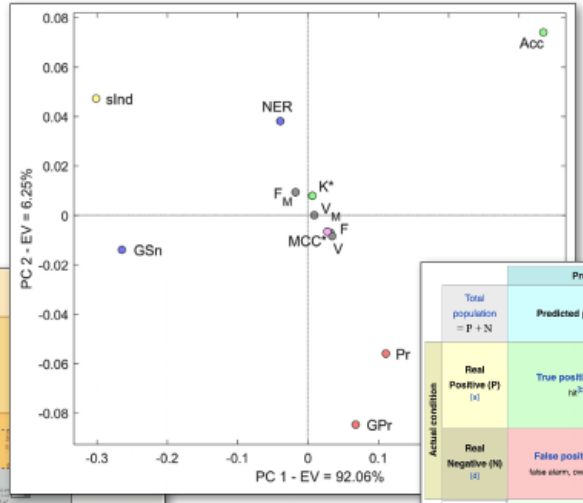
Without a family of scores, they are mixing apple and oranges …



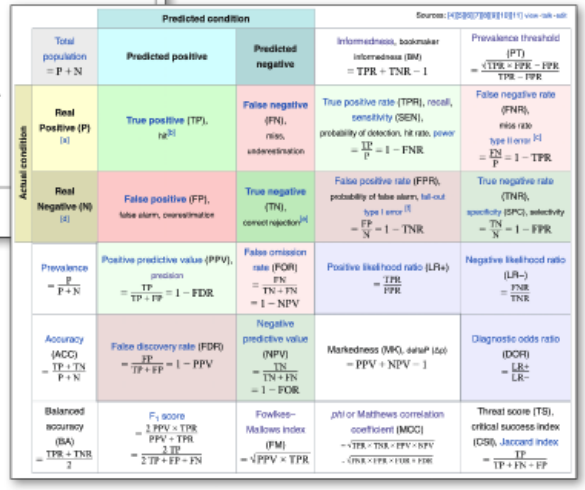
Ferri et al., 2009
Dendrograms



Canbek et al., 2017
"A chart like the periodic table of elements"



Ballabio et al., 2018
2D PCA of the values of 12 scores on 20 datasets



Wikipedia, 2025
A 2D Table

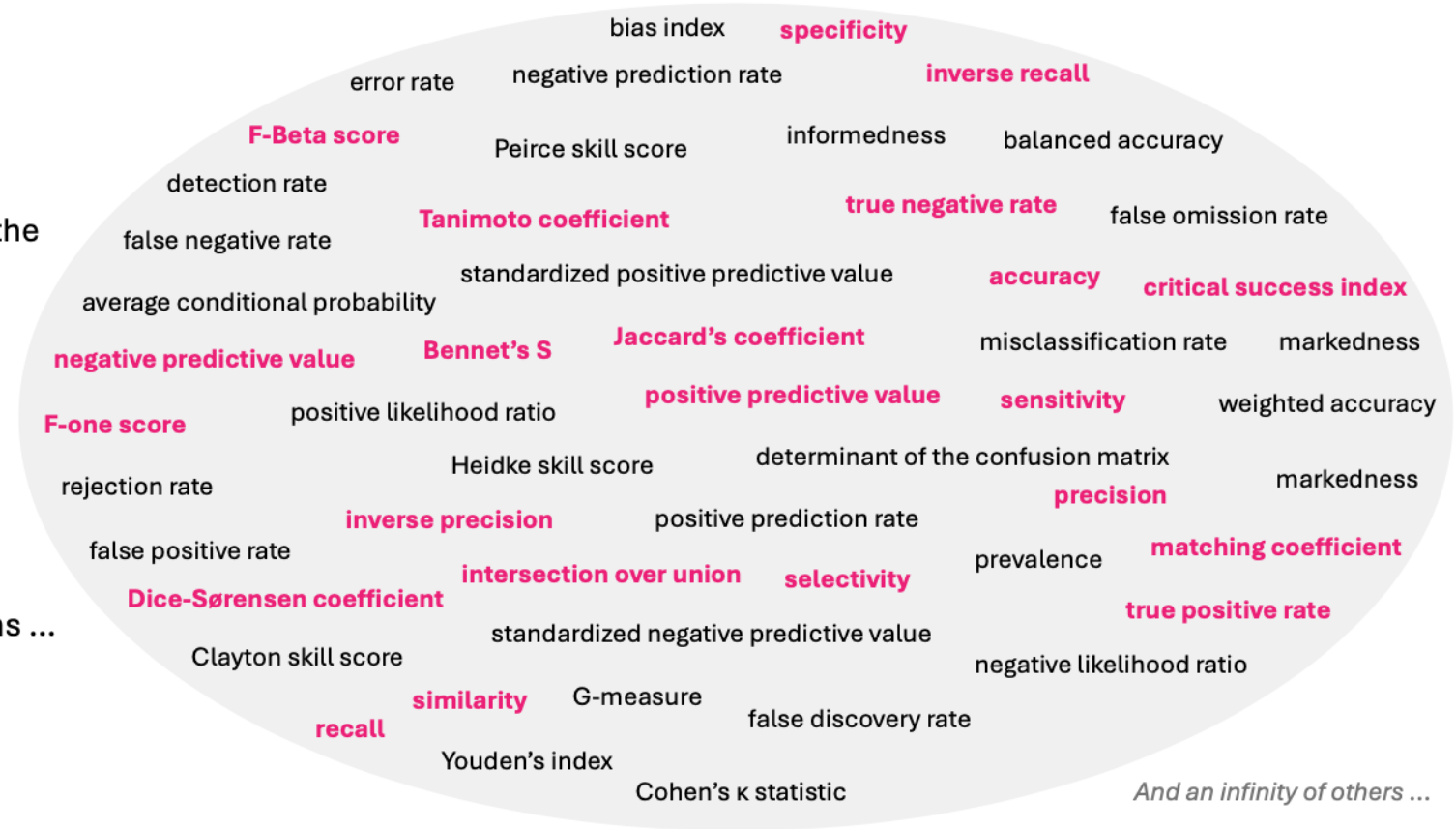
Why do they assume it is 2D?

In the end, a single score: all that work for this??? Really???

There are metrics and evaluation spaces I

There is an ocean of scores, all created from a 3D manifold. Hmmm?!

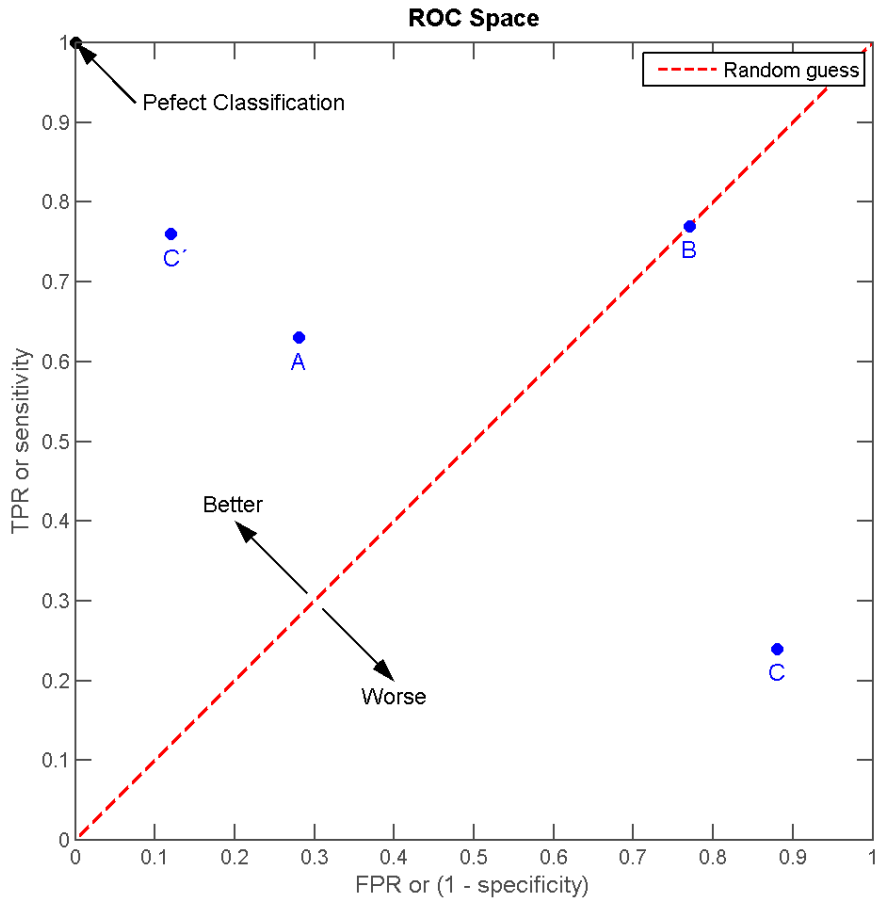
This is a small part of the scores produced by the human imagination, during a few decades, and just for the two-class classification problems ...



There are metrics and evaluation spaces II

Definition (Receiver Operating Characteristic, ROC)

The (FPR, TPR) pair defines the ROC evaluation space.



Outline

1. Introduction to performance evaluation
   - Motivation
   - Illustration: crisp two-class classification (binary classification)
   - More about the definition of a task
2. Evaluation of building damage assessment methods: a first glance
3. A theory for performance evaluation
4. Building an evaluation pipeline
5. Towards a standardized evaluation protocol

Formulation A: detection seen as a two-class classification

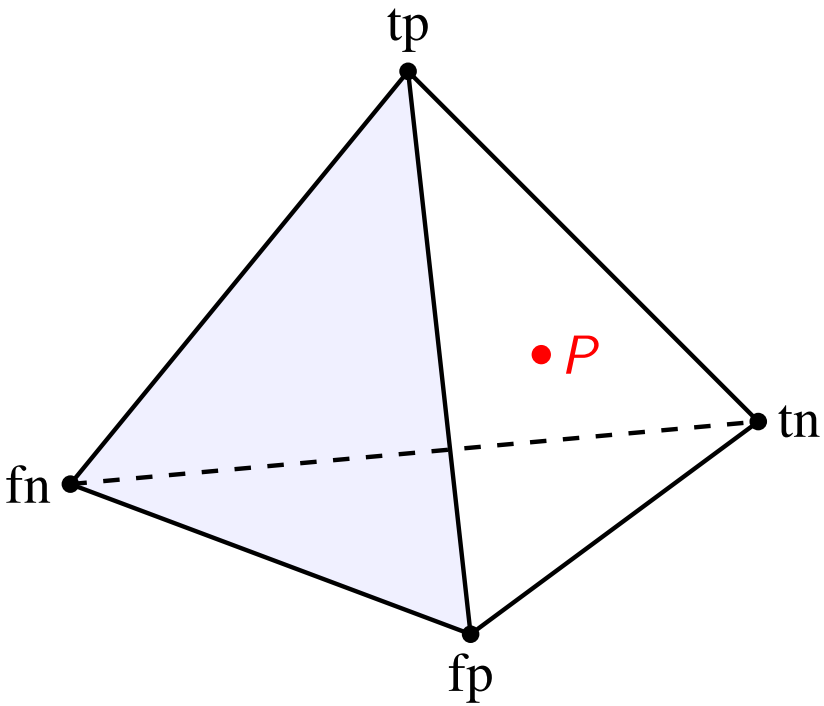
Formulation A (inspired by binary classification)

Apply the detector to all the N given samples and determine whether each result is positive or negative.

- Evaluation reuses everything from classification: confusion matrix, normalized version, derived scores.
- Normalized confusion matrix: $P(\{\mathrm{tn}\}), P(\{\mathrm{fp}\}), P(\{\mathrm{fn}\}), P(\{\mathrm{tp}\}) \geq 0$, with sum 1: a *probability simplex* (a 3D manifold in a 4D space).
- Examples: "Is there an exoplanet in this cube?", "Is there a human in this image?"

For detection, the true negatives are often ill-defined: how many "non-buildings" are there in an image?

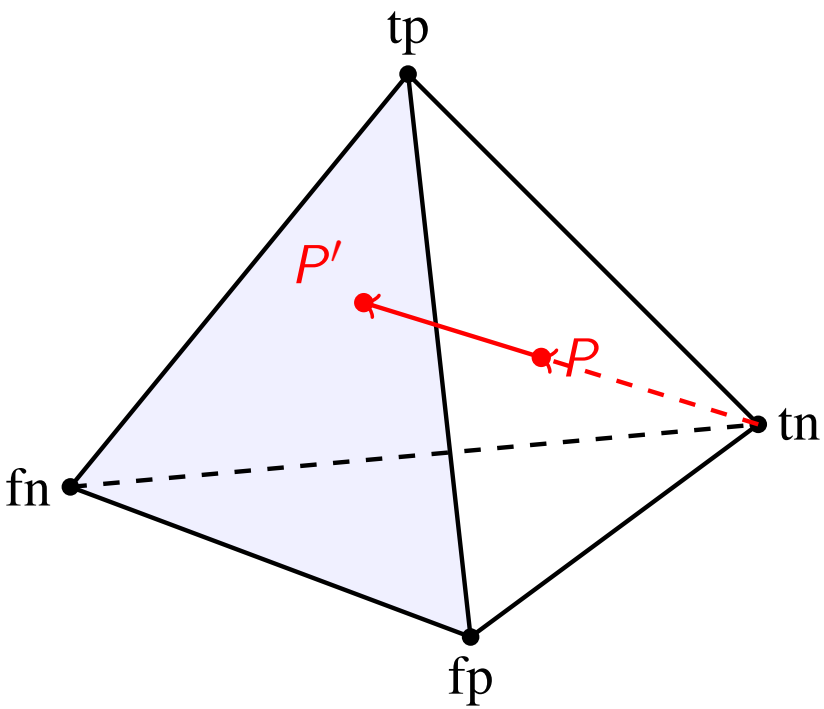
The performance tetrahedron (= a geometric representation of the performance space)



- All normalized confusion matrices form a *regular tetrahedron* (Bernoulli tetrahedron[a]).
- Vertices: the four outcomes tp, fn, fp, tn (top vertex: $P(\{\text{tp}\}) = 1$).
- A performance P is a point *inside*.
- A single score cannot locate P: 3 numbers are needed; ROC and PR spaces are 2D cuts.

[a] Gery Geenens. "An essay on copula modelling for discrete random vectors; or how to pour new wine into old bottles." In: *arXiv* abs/1901.08741 (Feb. 2019). arXiv: 1901.08741.

Ignoring true negatives (TNs): projection onto the (tp, fp, fn) face

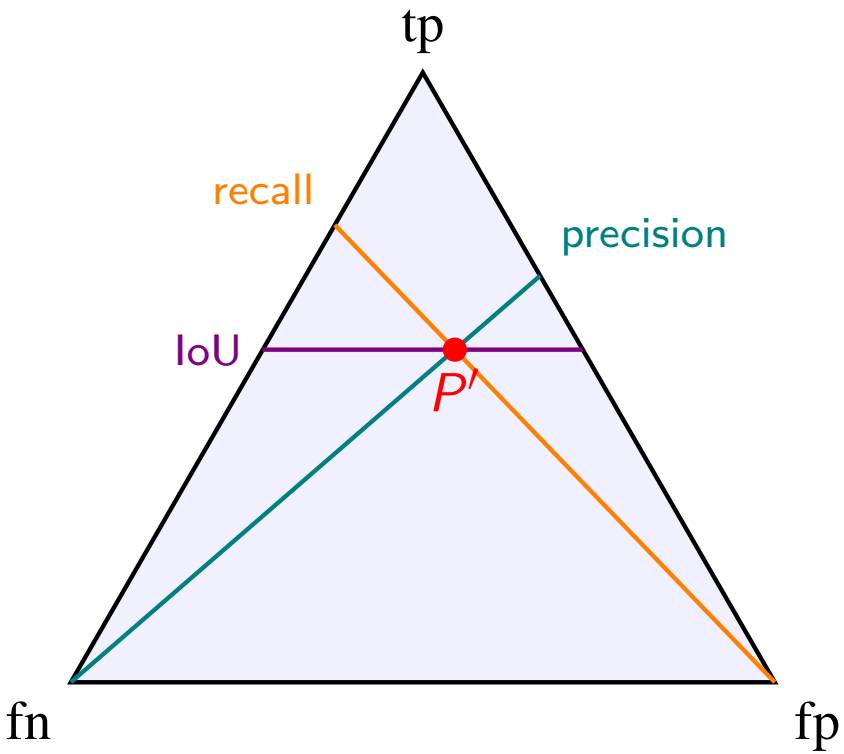


Central projection from the tn vertex:

$$P'(\{\omega\}) = \frac{P(\{\omega\})}{1 - P(\{\mathrm{tn}\})},\ \omega \neq \mathrm{tn}$$

- P' is the conditional distribution given $\{\mathrm{tp}, \mathrm{fp}, \mathrm{fn}\}$: the true negatives are *discarded*.
- Motivation:
  1. in detection, the number of true negatives is often meaningless, and
  2. the priors of a test set are unknown.

A new performance space: the detection triangle



iso-lines through P'

- The face (tp, fn, fp) is a performance space of its own: the *detection triangle* [25], free of true negatives.
- TN-free scores are read directly:
  - $\mathrm{IoU} = P'(\{\mathrm{tp}\})$: lines parallel to the fn–fp edge,
  - precision: lines through fn,
  - recall: lines through fp,
  - $F_1 = 2\mathrm{IoU}/(1+\mathrm{IoU})$.
- Every point is an achievable performance, unlike in the precision-recall (PR) space[a].

[a] Kendrick Boyd et al. "Unachievable Region in Precision-Recall Space and Its Effect on Empirical Evaluation." In: *Int. Conf. Mach. Learn. (ICML)*. Edinburgh, UK, June 2012, pp. 639–646.

Reading the confusion matrix as answers to questions

Each row and each column of the confusion matrix answers a question:

- L_1 (row $+$): for a sample *known to be positive*, what does the detector predict? $\rightarrow$ TP or FN
- L_2 (row $-$): the same for a sample known to be negative $\rightarrow$ FP or TN
- C_1 (column $+$): for a sample *detected as positive*, is the detection correct? $\rightarrow$ TP or FP
- C_2 (column $-$): the same for a sample detected as negative $\rightarrow$ FN or TN

Why questions?

Questions have an obvious interpretation, unlike the confusion matrix. Questions L_1 and L_2 make sense when the ground truth is known (*training/development* mode), questions C_1 and C_2 when the detector is used (*test/operation* mode).

Source:[1]

[1] Marc Van Droogenbroeck. "Probabilistic Formulations and Evaluation Spaces for the Detection Task." Unpublished manuscript. 2026.

Formulation B: detection as answers to two questions

Formulation B (paradigm of answers to questions)

The detection task consists in being able to answer two questions, which only involve *positive samples* (L_1) or *positive detections* (C_1):

- L_1, asked *to the detector* about a sample known to be positive: "do you detect it?" Yes $\rightarrow$ agreed (TP); no $\rightarrow$ disagreed (FN).
- C_1, asked *to the ground truth* about a detection: "is this sample really positive?" Yes $\rightarrow$ true (TP$'$); no $\rightarrow$ false (FP).

Answers are counted in a *question/answer (Q&A) matrix*, a *pseudo* confusion matrix:

Questions	**Answers**	
L_1: sample known to be positive	TP (agreed)	FN (disagreed)
C_1: sample detected as positive	TP$'$ (true)	FP (false)

- No true negative is ever counted.
- TP and TP$'$ can differ: they come from two different questions (and possibly two different modes).

Formulation B: scores and evaluation space

For each question q, a "true positive rate"-like score can be defined:

$$S_q = \frac{\#(q^+)}{\#(q^+) + \#(q^-)}$$

- S_{L_1} is the true positive rate (recall) measured when the ground truth is known,
- S_{C_1} is the precision measured on the detections.
- The 2D space (S_{C_1}, S_{L_1}) resembles the precision-recall space, but the two modes are *disentangled*; according to the source, it has no unreachable region.
- The outcomes $\{\mathrm{TP}, \mathrm{FN}, \mathrm{TP}', \mathrm{FP}\}$ again form a probability simplex: all the tools of Formulation A (tetrahedron, projections, triangles) remain available.

Source:[2]

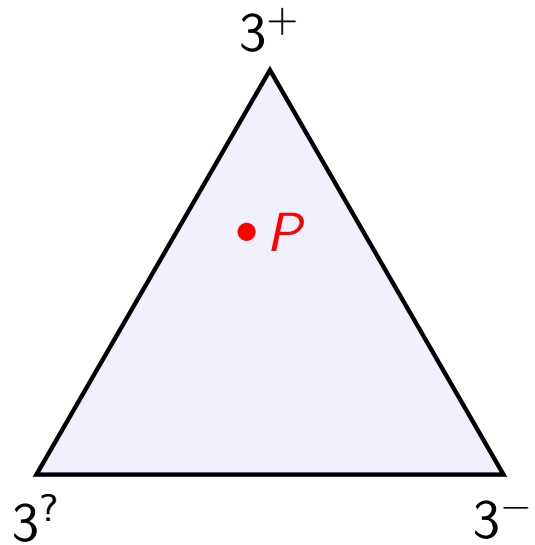
[2] Marc Van Droogenbroeck. "Probabilistic Formulations and Evaluation Spaces for the Detection Task." Unpublished manuscript. 2026.

Formulation C: doubt is allowed

Formulation C (answers to questions, with doubt)

For any sample detected as positive, is the detection "true" (3^+), "false" (3^-), or does the detector express a doubt, "don't know" ($3^?$)?

- Frequent in practice: a confidence score with two thresholds (high: true, low: false, in between: doubt).
- Three outcomes: the performance space is a triangle.
- Open point: satisfaction and importance of $3^?$.



Source:[3]

[3] Marc Van Droogenbroeck. "Probabilistic Formulations and Evaluation Spaces for the Detection Task." Unpublished manuscript. 2026.

Outline

1. Introduction to performance evaluation
2. Evaluation of building damage assessment methods: a first glance
3. A theory for performance evaluation
4. Building an evaluation pipeline
5. Towards a standardized evaluation protocol

How is the BDA task formulated?

Formulation	Semantic level	Description
Detection and binary classification	image crop	Predict whether the crop contains a damaged building (positive class) or not.
Multi-class classification	image crop	Multi-class classification extends the binary setting by assigning each image crop to one of several damage categories.
Semantic segmentation	pixel-level	Assigns each pixel to one of several semantic classes, such as background, undamaged building, damaged building, or destroyed building.
Detection and instance segmentation	pixel-level	Localize all buildings using bounding boxes, polygons, or any shapes, and assign each instance a damage label.

The formulation constrains the methods, the annotations, and the evaluation protocols.[4]

[4] Anthony Cioppa et al. "Revisiting the Task of Building Damage Assessment by Satellite Imaging." In: *arXiv* abs/2610.13857 (2026). DOI: 10.48550/arXiv.2610.13857. arXiv: 2610.13857.

Annotations: no common definition of damage classes

CEMS ODM/IWG-SEM [12]	**BRIGHT** [5]	**xBD** [9]
Destroyed	Destroyed	Destroyed
Damaged	Damaged	Major Damage and Minor Damage
Possibly damaged	— (merged into Intact)	—
No visible damage (IWG-SEM only)	Intact	No Damage
—	Background	—

- Approximate correspondences: definitions differ; BRIGHT merged "possibly damaged" into intact [5]; CEMS ODM has no class for undamaged buildings.
- "Possibly damaged" reflects uncertainty, not a physical state; manual annotations suffer from debris, misalignment, and geolocation errors.

Source:[5]

[5] Anthony Cioppa et al. "Revisiting the Task of Building Damage Assessment by Satellite Imaging." In: *arXiv* abs/2610.13857 (2026). DOI: 10.48550/arXiv.2610.13857. arXiv: 2610.13857.

Benchmarks: (almost) always a variant of F_1

Benchmark	Task	Evaluation strategy	Score definition
xBD [10]	Multi-class classification	Per-class F_1 and weighted F_1	$F_1 = \frac{2TP}{2TP+FP+FN}$
QuickQuake-Buildings [24]	Binary classification	Precision, recall, F_1, and AUROC	$P = \frac{TP}{TP+FP}$, $R = \frac{TP}{TP+FN}$, $F_1 = 2\frac{PR}{P+R}$
BRIGHT [5]	Classification, segmentation	Overall accuracy (OA), F_1 score split into F_1^{loc} and F_1^{cls}, and mean IoU (mIoU)	$OA = \frac{\sum TP_i}{\#elements}$, $IoU = \frac{TP}{TP+FP+FN}$
AI Model Challenge [11]	Detection (binary classification)	F_1 at the building level	$F_1 = 2\frac{PR}{P+R}$

Source:[6].

[6] Anthony Cioppa et al. "Revisiting the Task of Building Damage Assessment by Satellite Imaging." In: *arXiv* abs/2610.13857 (2026). DOI: 10.48550/arXiv.2610.13857. arXiv: 2610.13857.

Why is F_1 so popular in BDA?

- A per-class analysis is convenient. For multi-class classification, the accuracy (with the balanced accuracy) is the only ranking score in frequent use.
- Classes can often be collapsed into a two-class classification.
- F_1 ignores the true negatives which are rarely defined; its value does not depend on the (large, ill-defined) number of "non-damaged" pixels or buildings.
- F_1 is a compromise between precision and recall, like F_β. Note however that the ranking it induces can be closer to that of precision or recall [17].
- Tradition and comparability: the xView2 challenge on xBD scored methods with F_1 [10], and later works report the same score to compare with it.
- F_1 is the Dice coefficient, familiar in segmentation, where it is also used as a training loss [14].

A landscape of techniques with incomparable results

Model	Dataset, task	Metric	Setting	Value	Source
ResNet-50 baseline	xBD, localization and damage	xView2 weighted F_1 (pixels)	Supervised	0.2654	[10]
MambaBDA-Base	xBD, localization and damage	F_1^{overall}	Supervised	81.41	[4]
BDANet	xBD, localization and damage	F_1^{loc} / F_1^{dmg} / F_1^{overall}	Supervised	0.864 / 0.782 / 0.806	[23]
MTP (ViT-L+RVSA)	xBD, change detection	F_1^{overall}	Pre-trained on SAMRS, fine-tuned	77.62	[22]
ChangeMamba	BRIGHT, 14 events	mIoU (average)	Zero-shot / one-shot	39.4 / 43.2	[5]
DamFormer	BRIGHT, 5 events	mIoU	Post-event SAR / optical / both	65.6 / 69.8 / 70.8	[5]
U-Net vs. best foundation model	xView2 in PANGAEA	mIoU	Fine-tuning	58.7 vs. 60.7	[13]

Values in different rows are not comparable (datasets, tasks, metrics, settings).

Outline

1. Introduction to performance evaluation
2. Evaluation of building damage assessment methods: a first glance
3. A theory for performance evaluation
   - A probabilistic framework
   - Two random variables: satisfaction and importance
   - Three axioms to define the ranking operator
   - Ranking scores
4. Building an evaluation pipeline
5. Towards a standardized evaluation protocol

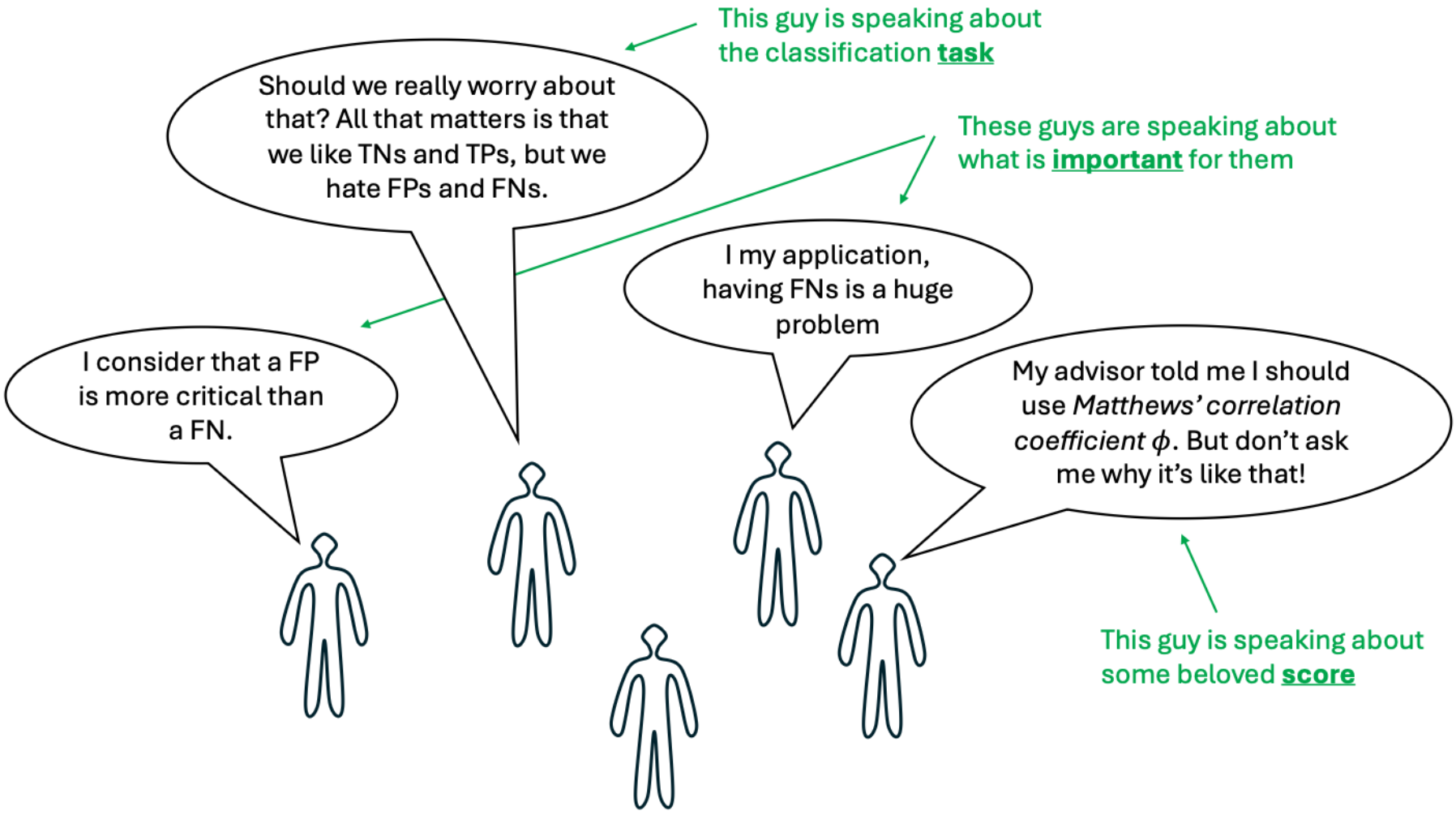
Four theoretical elements of an evaluation procedure

1. **Underlying statistical assumption.** The statistical approach (significance of results) is rarely used; BDA relies on a probabilistic, *frequentist approach*, with implicit assumptions of spatial stationarity and ergodicity.
2. **On the use of scores.** Scores serve to (1) optimize a method, (2) compare methods, (3) select the best one. A single widely adopted score (F_1 in BDA) leads to similar methods: we call this the *attractor effect*. Not every score is suitable for ranking [21]; averaging scores (e.g., mIoU) breaks this.
3. **Crisp vs. soft classifiers.** A crisp classifier gives no confidence; a soft one requires a threshold, and the two are difficult to compare.
4. **Generalization and transferability.** Leave-one-area-out evaluation is common, but predicting the rank of a method on new data is far from solved [20].

Source:[7]

[7] Anthony Cioppa et al. "Revisiting the Task of Building Damage Assessment by Satellite Imaging." In: *arXiv* abs/2610.13857 (2026). DOI: 10.48550/arXiv.2610.13857. arXiv: 2610.13857.

Discussions about a performance



Components of a new framework for performance evaluation and ranking

Building a framework[8] step by step:

1. A probabilistic (measurable) space (Ω, Σ)
2. A way to consider the specificities of a task $\rightarrow$ the concept of "Satisfaction"
3. A way to consider the needs of an application $\rightarrow$ the concept of "Importance"
4. Three axioms to define the notion of ranking

[8] Sébastien Piérard et al. "Foundations of the Theory of Performance-Based Ranking." In: *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Nashville, TN, USA: IEEE, June 2025, pp. 14293–14302. DOI: 10.1109/cvpr52734.2025.01333.

Outline

1. Introduction to performance evaluation
2. Evaluation of building damage assessment methods: a first glance
3. A theory for performance evaluation
   - A probabilistic framework
   - Two random variables: satisfaction and importance
   - Three axioms to define the ranking operator
   - Ranking scores
4. Building an evaluation pipeline
5. Towards a standardized evaluation protocol

Framework

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*;
- take the set of possible *outcomes* as the sample space Ω;
- when Ω is finite, take the σ-algebra $\Sigma = 2^{\Omega}$ as the event space.

Several *choices* are possible for the outcomes. In two-class crisp classification, for example:

- $\Omega = \{\text{correct}, \text{incorrect}\}$
- $\Omega = \{\text{correct}, \text{type I error}, \text{type II error}\}$
- $\Omega = \{\text{tn}, \text{fp}, \text{fn}, \text{tp}\}$

The evaluation cannot be *any* random experiment, and the outcomes cannot be chosen freely (see the satisfaction).

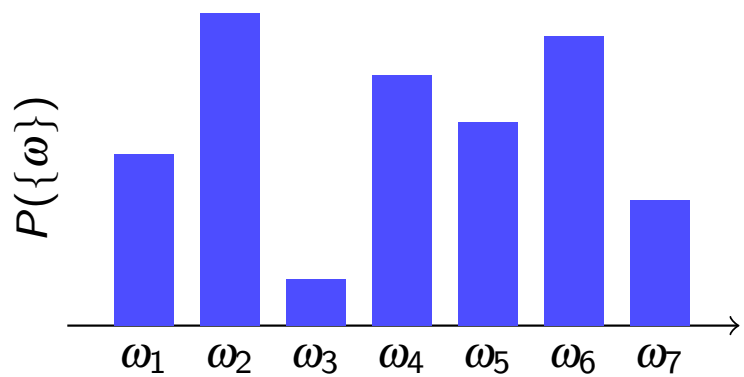
Model for a performance

- Performances are modeled as *probability measures* on (Ω, Σ); the set of all possible performances is denoted $\mathbb{P}_{(\Omega,\Sigma)}$; no notion of score at this stage.
- **Example.** In classification, with $\Omega = \mathbb{C}^2$ (all pairs of ground-truth and predicted classes), a performance carries the same information as a *normalized confusion matrix*.
- **Interpretation.** The performance of an entity (e.g., a classifier) is the distribution of the cases (outcomes) encountered when evaluating this entity.

two-class crisp classification

	$\hat{y} = c^-$	$\hat{y} = c^+$
$y = c^-$	0.759 tn	0.141 fp
$y = c^+$	0.028 fn	0.072 tp

any finite Ω



Outline

1. Introduction to performance evaluation
2. Evaluation of building damage assessment methods: a first glance
3. A theory for performance evaluation
   - A probabilistic framework
   - Two random variables: satisfaction and importance
   - Three axioms to define the ranking operator
   - Ranking scores
4. Building an evaluation pipeline
5. Towards a standardized evaluation protocol

A first random variable: Satisfaction

Performances are probability measures, but *not all probability measures are performances*, and we do not aim at ordering any probability measure. The satisfaction random variable is the key.

The measurable space (or the random experiment) must be such that one can specify the degree of satisfaction obtained for each outcome:

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

- The satisfaction is related to the *task*.
- Always specify the satisfaction explicitly; otherwise, the task is undefined.
- **Example** (two-class crisp classification, by convention): $S(\mathrm{fp}) = S(\mathrm{fn}) = 0$ and $S(\mathrm{tn}) = S(\mathrm{tp}) = 1$, that is, "we like TNs and TPs, but we dislike FPs and FNs".
- There is no constraint on the satisfaction values: they can be negative, zero, or positive.

Source:[9]

[9] Sébastien Piérard et al. "Foundations of the Theory of Performance-Based Ranking." In: *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Nashville, TN, USA: IEEE, June 2025, pp. 14293–14302. DOI: 10.1109/cvpr52734.2025.01333.

A second random variable: Importance

The *importance* random variable $I : \Omega \to \mathbb{R}^+$ encodes application-specific preferences.

User 1 (insurance company)

"I consider that a FP is more critical than a FN."

User 2 (maintenance company)

"In my application, having FNs is a huge problem."

Evaluator

"Please, be more precise: quantify the importance you give to the TNs, FPs, FNs, and TPs!"

- The satisfaction is about the *task*; the importance is about the *application*.
- There is one constraint: importance values are zero or positive.

Ranking can be ... confusing

Mean rank over five usual scores (ranks in brackets). $(\mathrm{tn}, \mathrm{fp}, \mathrm{fn}, \mathrm{tp})$: A $= (.759, .141, .028, .072)$, B $= (.532, .368, .010, .090)$, C $= (.761, .139, .019, .081)$.

Classifier	TPR	TNR	PPV	NPV	F_1	**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_1	**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**

Does it make sense?

Adding C reverses the order of A and B ($1.4 < 1.6$ becomes $2.4 > 2.2$). We need a model of the ranking operator: three axioms[a].

[a] Sébastien Piérard et al. "Foundations of the Theory of Performance-Based Ranking." In: *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Nashville, TN, USA: IEEE, June 2025, pp. 14293–14302. DOI: 10.1109/cvpr52734.2025.01333.

Outline

1. Introduction to performance evaluation
2. Evaluation of building damage assessment methods: a first glance
3. A theory for performance evaluation
   - A probabilistic framework
   - Two random variables: satisfaction and importance
   - Three axioms to define the ranking operator
   - Ranking scores
4. Building an evaluation pipeline
5. Towards a standardized evaluation protocol

Axiom 1: rankings derive from a performance ordering

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

Axiom 1. The ranking function $\mathrm{rank}_{\mathbb{E}} : \mathbb{E} \to [1, |\mathbb{E}|] : \varepsilon \mapsto \mathrm{rank}_{\mathbb{E}}(\varepsilon)$ satisfies

$$|\{\varepsilon' \in \mathbb{E} : \mathrm{eval}(\varepsilon) < \mathrm{eval}(\varepsilon')\}| + 1 \leq \mathrm{rank}_{\mathbb{E}}(\varepsilon) \leq |\{\varepsilon' \in \mathbb{E} : \mathrm{eval}(\varepsilon) \lesssim \mathrm{eval}(\varepsilon')\}|$$

where $\lesssim$ is a preorder on $\mathbb{P}_{(\Omega,\Sigma)}$ and $\mathbb{E}$ is the set of entities to rank.

Consequence: the order between two entities never depends on the other entities in the ranking (unlike the mean rank of the previous slide).

Source:[10]

[10] Sébastien Piérard et al. "Foundations of the Theory of Performance-Based Ranking." In: *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Nashville, TN, USA: IEEE, June 2025, pp. 14293–14302. DOI: 10.1109/cvpr52734.2025.01333.

Axiom 2: the ordering is coherent with the satisfaction

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 and P_2 are incomparable.

Three intuitive implications:

- Achieving the minimal satisfaction for sure is the worst we can do: $P(S = s_{\text{min}}) = 1 \Rightarrow \nexists P' : P' < P$.
- Achieving the maximal satisfaction for sure is the best we can do: $P(S = s_{\text{max}}) = 1 \Rightarrow \nexists P' : P' > P$.
- Two performances achieving for sure the same satisfaction are either equivalent or incomparable.

P	$\hat{y} = c^-$	$\hat{y} = c^+$
$y = c^-$	0	$\cdot$
$y = c^+$	$\cdot$	0

worst

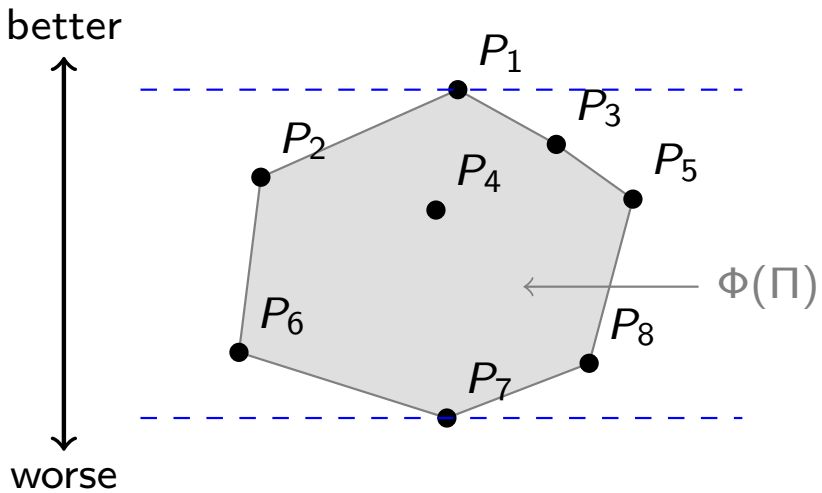
P	$\hat{y} = c^-$	$\hat{y} = c^+$
$y = c^-$	$\cdot$	0
$y = c^+$	0	$\cdot$

best

Source:[11]

[11] Sébastien Piérard et al. "Foundations of the Theory of Performance-Based Ranking." In: *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Nashville, TN, USA: IEEE, June 2025, pp. 14293–14302. DOI: 10.1109/cvpr52734.2025.01333.

Axiom 3: no gain from blind combinations



- Combine entities *blindly* (e.g., pick one at random, then run it). $\Phi(\Pi)$: performances reachable from the set Π (for random selection, the convex hull).

A blind combination should never be *better than the best* performance, nor *worse than the worst*.

Axiom 3 (simplified). Let P be comparable to every element of Π. If $P' \lesssim P$ for all $P' \in \Pi$, then $\tilde{P} \lesssim P$ for all $\tilde{P} \in \Phi(\Pi)$; if $P \lesssim P'$ for all $P' \in \Pi$, then $P \lesssim \tilde{P}$ for all $\tilde{P} \in \Phi(\Pi)$ (and the same with strict inequalities).

Source:[12]

[12] Sébastien Piérard et al. "Foundations of the Theory of Performance-Based Ranking." In: *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Nashville, TN, USA: IEEE, June 2025, pp. 14293–14302. DOI: 10.1109/cvpr52734.2025.01333.

Outline

1. Introduction to performance evaluation
2. Evaluation of building damage assessment methods: a first glance
3. A theory for performance evaluation
   - A probabilistic framework
   - Two random variables: satisfaction and importance
   - Three axioms to define the ranking operator
   - Ranking scores
4. Building an evaluation pipeline
5. Towards a standardized evaluation protocol

Do ranking scores exist? Yes!

We have:

- a sample space Ω
- $\mathbf{p}(\omega)$ is the probability of an outcome
- $S(\omega)$ is the satisfaction of that outcome
- $I(\omega)$ is the Importance of that outcome

then, members of the following family are all ranking scores

$$R(P) = \frac{E[IS]}{E[I]} = \frac{\sum_{\omega \in \Omega} I(\omega) S(\omega) \mathbf{p}(\omega)}{\sum_{\omega \in \Omega} I(\omega) \mathbf{p}(\omega)}$$

Note that

$$A = E[S]$$

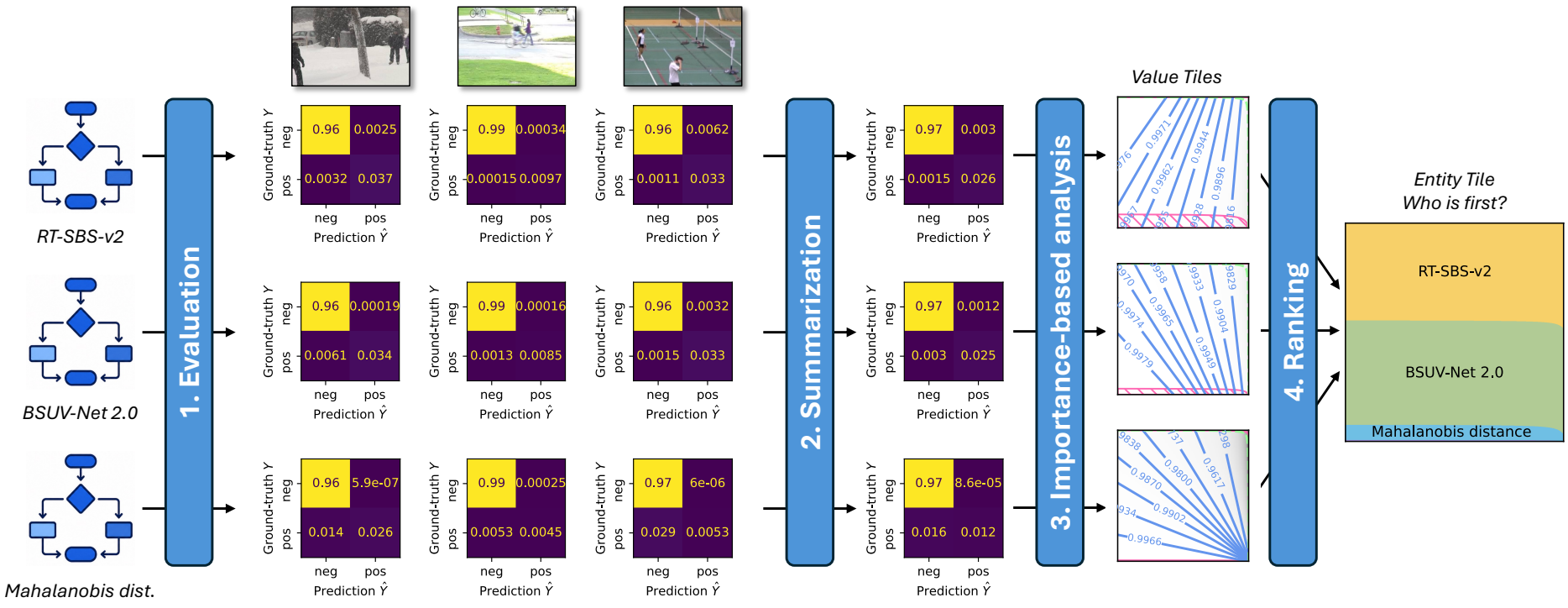
known as the *accuracy*, is also a ranking score!

Outline

1. Introduction to performance evaluation
2. Evaluation of building damage assessment methods: a first glance
3. A theory for performance evaluation
4. Building an evaluation pipeline
   - Step 1: Evaluation
   - Step 2: Summarization
   - Step 3: Importance-based analysis – the Tile
   - Understanding the Tile
   - Tile flavors
   - Step 4: Ranking
   - Taking the chance into account: Illegal Waste Dumping Detection (IWDD)
5. Towards a standardized evaluation protocol

A probabilistic performance evaluation pipeline

A pipeline with 4 steps



Outline

1. Introduction to performance evaluation
2. Evaluation of building damage assessment methods: a first glance
3. A theory for performance evaluation
4. Building an evaluation pipeline
   - Step 1: Evaluation
   - Step 2: Summarization
   - Step 3: Importance-based analysis – the Tile
   - Understanding the Tile
   - Tile flavors
   - Step 4: Ranking
   - Taking the chance into account: Illegal Waste Dumping Detection (IWDD)
5. Towards a standardized evaluation protocol

Step 1 (evaluation): from a method to its performance

Our illustrative use case: cast the detection problem as a *two-class crisp classification* problem.

- A performance P is a probability measure on (Ω, Σ) with $\Omega = \{\mathrm{tn}, \mathrm{fp}, \mathrm{fn}, \mathrm{tp}\}$ and $\Sigma = 2^{\Omega}$.
- *Scores* are real functions of the performance:
  - unconditional: priors $\pi^+(P) = P(\{\mathrm{fn}, \mathrm{tp}\})$, rate of positive predictions $\tau^+(P) = P(\{\mathrm{fp}, \mathrm{tp}\})$, accuracy $A(P) = P(\{\mathrm{tn}, \mathrm{tp}\})$;
  - conditional: precision $P(\{\mathrm{tp}\}|\{\mathrm{fp}, \mathrm{tp}\})$, recall $P(\{\mathrm{tp}\}|\{\mathrm{fn}, \mathrm{tp}\})$, TNR, NPV, IoU $= P(\{\mathrm{tp}\}|\{\mathrm{fp}, \mathrm{fn}, \mathrm{tp}\})$;
  - non-probabilistic but derived from probabilistic scores: F_β, balanced accuracy.
- These probabilistic readings (e.g., "precision is the probability that a positive decision is correct") are rigorous only when the performance is anchored in probability theory.

Source:[13]

[13] Anastasia Zakharova et al. "Event Detection in Videos: A Framework for the Development of New Methods." In: *arXiv* abs/2607.04372 (2026). DOI: 10.48550/arXiv.2607.04372. arXiv: 2607.04372.

A performance needs at least three numbers

- A performance can be represented by a 2×2 confusion matrix normalized to 1, or by triplets such as (π^+, τ^+, A) or $(\pi^+, \mathrm{TNR}, \mathrm{TPR})$.
- The set of all performances is a 3D simplex (a regular tetrahedron).
- By the "invariance of domain" theorem, no continuous bijection with a continuous inverse exists between $\mathbb{R}^n$ and $\mathbb{R}^m$ when $n \neq m$: *at least 3 scores* are needed.
- Not every triplet works: the often-reported (Pr, Re, F_1) is insufficient, since F_1 is a function of Pr and Re.

There is no such thing as a score that is better than another one in general.

The random evaluation experiment

An elementary randomized test, using the method, that leads to an outcome in Ω. It gives the precise meaning of tn, fp, fn, and tp.

Random Experiment (background subtraction)

Apply an oracle to label every pixel of every frame of a clip, and the method to be evaluated. Choose a frame and a pixel at random; compare the outputs y and $\hat{y}$ of the oracle and of the method: return tn, fp, fn, *or* tp.

Random Experiment (pedestrian detection)

Choose a frame at random; obtain the ground-truth boxes y and the predicted boxes $\hat{y}$. If both are empty, return tn. *Otherwise, apply a well-specified matching criterion and choose an element of $y \cup \hat{y}$ at random: unmatched prediction* $\to \mathrm{fp}$, *unmatched ground truth* $\to \mathrm{fn}$, *matched* $\to \mathrm{tp}$.

Source:[14]

[14] Anastasia Zakharova et al. "Event Detection in Videos: A Framework for the Development of New Methods." In: *arXiv* abs/2607.04372 (2026). DOI: 10.48550/arXiv.2607.04372. arXiv: 2607.04372.

A remote-sensing example, and why the experiment matters

Random Experiment (geographic feature detection)

From an aerial or satellite image, obtain the ground-truth map and the estimated map of a real-valued magnitude. Draw a threshold t from a chosen distribution T, and a pixel at random with values y and $\hat{y}$. Return tn *if* $\max(y, \hat{y}) < t$, tp *if* $t < \min(y, \hat{y})$, *otherwise* fp *if* $y < \hat{y}$ *and* fn *if not.*

- Under-specified events lead to ill-posed questions (Bertrand's paradox[15]): "how many people entered the scene?" has two answers, depending on whether a re-entering person counts twice.
- In practice, the method is run once and the probabilities are estimated by counting (frequentist interpretation); what matters is the *consistency* between the specified experiment and the implemented evaluation algorithm.

Source:[16]

[15] Joseph Bertrand. *Calcul des probabilités*. Gauthier-Villars et fils, 1889.

[16] Anastasia Zakharova et al. "Event Detection in Videos: A Framework for the Development of New Methods." In: *arXiv* abs/2607.04372 (2026). DOI: 10.48550/arXiv.2607.04372. arXiv: 2607.04372.

Outline

1. Introduction to performance evaluation
2. Evaluation of building damage assessment methods: a first glance
3. A theory for performance evaluation
4. Building an evaluation pipeline
   - Step 1: Evaluation
   - Step 2: Summarization
   - Step 3: Importance-based analysis – the Tile
   - Understanding the Tile
   - Tile flavors
   - Step 4: Ranking
   - Taking the chance into account: Illegal Waste Dumping Detection (IWDD)
5. Towards a standardized evaluation protocol

Averaging performances instead of score values

- Methods are evaluated on several sources (videos, scenes, areas, events): we need an *average* **performance**.
- Averaging *score values* is inadequate: the means are uninterpretable.
  - The mean values of a set of scores do not correspond to any possible performance.
  - The centroid in ROC does not correspond to the centroid in precision-recall space [7, 18].
  - The centroid in precision-recall space can fall in its unachievable region [3].

Solution: evaluate on a *hybrid source*. Drawing data from it amounts to drawing a source s with probability λ_s (arbitrary weights), and then drawing data from that source.

The summarized performance is the performance obtained on this mixture.

Source:[17]

[17] Anastasia Zakharova et al. "Event Detection in Videos: A Framework for the Development of New Methods." In: *arXiv* abs/2607.04372 (2026). DOI: 10.48550/arXiv.2607.04372. arXiv: 2607.04372.

What the evaluation is telling us

Some random evaluation experiments are *linear with respect to the data source*: the performance on a mixture is the convex combination of the performances.

$$P_{\text{summarized}} = \sum_s \lambda_s P_s \quad \text{with} \quad \lambda_s \geq 0, \sum_s \lambda_s = 1$$

- The summarized confusion matrix is the *weighted arithmetic mean* of the normalized confusion matrices.
- One can arithmetically average numerical representations involving only *linear* scores (unconditional probabilistic scores, expected values).
- Not $(\pi^+, \text{TNR}, \text{TPR})$: as priors differ between sources, TNR and TPR are not linear in the confusion matrix.
- A direct benefit of specifying the random evaluation experiment: we can derive which operations on performances are meaningful.

The implication for scores

For a collection $\mathbb{V}$ of videos with weights λ_v:

$$\pi^+(P_{\mathbb{V}}) = \sum_{v \in \mathbb{V}} \lambda_v \pi^+(P_v), \qquad \mathrm{TPR}(P_{\mathbb{V}}) = \frac{\sum_{v \in \mathbb{V}} \lambda_v \, \pi^+(P_v) \, \mathrm{TPR}(P_v)}{\pi^+(P_{\mathbb{V}})}$$

- A *prior-weighted* mean of TPR, not an arithmetic mean.
- Formulas exist for conditional scores [18] and for ranking scores [15], but they are not intuitive: average the performances first, then compute the score.

Toy example (two videos, $\lambda_1 = \lambda_2 = \frac{1}{2}$)

Video 1: $\pi^+ = 0.01$, $\mathrm{TPR} = 0.2$. Video 2: $\pi^+ = 0.5$, $\mathrm{TPR} = 0.9$.
Arithmetic mean of the TPRs: 0.55.
Summarized TPR: $\frac{1}{2}(0.01 \times 0.2 + 0.5 \times 0.9)/0.255 \approx 0.89$.

Source:[18]

[18] Anastasia Zakharova et al. "Event Detection in Videos: A Framework for the Development of New Methods." In: *arXiv* abs/2607.04372 (2026). DOI: 10.48550/arXiv.2607.04372. arXiv: 2607.04372.

We should not use the mean for scoring multiple videos/domains/datasets!

Obviously,

$$\text{for any video } i,\ \mathrm{F}_i = 2\frac{\mathrm{P}_i \times \mathrm{R}_i}{\mathrm{P}_i + \mathrm{R}_i} \quad \text{but} \quad \bar{\mathrm{F}} \neq 2\frac{\bar{\mathrm{P}} \times \bar{\mathrm{R}}}{\bar{\mathrm{P}} + \bar{\mathrm{R}}} = \bar{\bar{\mathrm{F}}}$$

In fact, the arithmetic mean has severe drawbacks:

- it breaks the intrinsic relationships between probabilistic indicators.
- because of these inconsistencies, we might have that

$$\bar{\mathrm{F}}_1 < \bar{\mathrm{F}}_2 \quad \text{while} \quad \bar{\bar{\mathrm{F}}}_1 > \bar{\bar{\mathrm{F}}}_2$$

The solution: use the summarization formulas

Summarization formulas and properties

Formulas (for unconditional and conditional probabilities, respectively)

$$\mathrm{R}_{\mathscr{A}}(\mathbb{V}) = \sum_{v \in \mathbb{V}} p(V = v)\, \mathrm{R}_{\mathscr{A}}(v)$$

$$\mathrm{R}_{\mathscr{A}|\mathscr{B}}(\mathbb{V}) = \sum_{v \in \mathbb{V}} p(V = v | \Delta \in \mathscr{B})\, \mathrm{R}_{\mathscr{A}|\mathscr{B}}(v)$$

Examples:

$$\mathrm{TPR}(\mathbb{V}) = \frac{1}{\pi^+(\mathbb{V})} \sum_{v \in \mathbb{V}} p(V = v)\, \pi^+(v)\, \mathrm{TPR}(v)$$

$$\mathrm{PPV}(\mathbb{V}) = \frac{1}{\tau^+(\mathbb{V})} \sum_{v \in \mathbb{V}} p(V = v)\, \tau^+(v)\, \mathrm{PPV}(v)$$

where $\pi^+(v)$ and $\tau^+(v)$ are the positive prior and the rate of positive predictions, respectively.

Outline

1. Introduction to performance evaluation
2. Evaluation of building damage assessment methods: a first glance
3. A theory for performance evaluation
4. Building an evaluation pipeline
   - Step 1: Evaluation
   - Step 2: Summarization
   - Step 3: Importance-based analysis – the Tile
   - Understanding the Tile
   - Tile flavors
   - Step 4: Ranking
   - Taking the chance into account: Illegal Waste Dumping Detection (IWDD)
5. Towards a standardized evaluation protocol

Towards canonical ranking scores

For a two-class (binary) classification/segmentation:

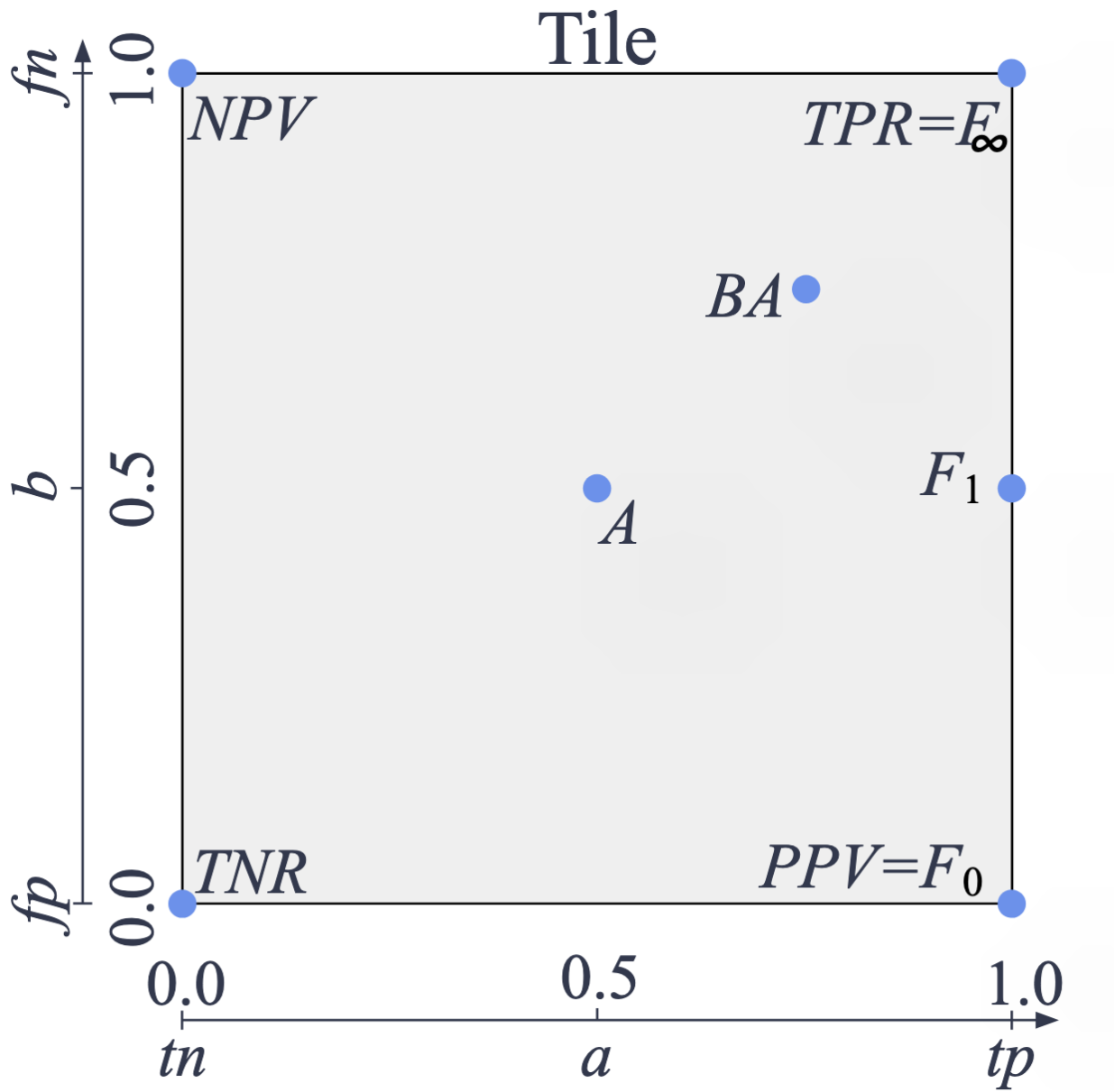
- the sample space contains 4 elements: $\mathrm{tp}, \mathrm{fn}, \mathrm{fp}, \mathrm{tn}$
- $S(\mathrm{tp}) = S(\mathrm{tn}) = 1$ and $S(\mathrm{fp}) = S(\mathrm{fn}) = 0$
- $I(\mathrm{tp}) = a$, $I(\mathrm{fn}) = b$, $I(\mathrm{fp}) = 1 - b$, and $I(\mathrm{tn}) = 1 - a$

We can define a 2D representation of the family of ranking scores by using two parameters (a, b), called the *Tile*, defined as follows[19],

$$R_{a,b}(P) = \frac{\sum_{\omega \in \Omega} I(\omega) S(\omega) \mathbf{p}(\omega)}{\sum_{\omega \in \Omega} I(\omega) \mathbf{p}(\omega)}$$
$$= \frac{a\mathbf{p}(\mathrm{tp}) + (1-a)\mathbf{p}(\mathrm{tn})}{a\mathbf{p}(\mathrm{tp}) + b\mathbf{p}(\mathrm{fn}) + (1-b)\mathbf{p}(\mathrm{fp}) + (1-a)\mathbf{p}(\mathrm{tn})}$$

[19] Sébastien Piérard et al. "The Tile: A 2D Map of Ranking Scores for Two-Class Classification." In: *arXiv* abs/2412.04309 (2024). DOI: 10.48550/arXiv.2412.04309. arXiv: 2412.04309.

The *Tile* (the first space of ranking scores)



Well-known scores are points of the Tile

Canonical ranking scores: $I(\mathrm{tp}) = a$, $I(\mathrm{tn}) = 1 - a$, $I(\mathrm{fn}) = b$, $I(\mathrm{fp}) = 1 - b$, with $(a, b) \in [0, 1]^2$.

Score	(a, b)	**Reading**
Accuracy A	$(\frac{1}{2}, \frac{1}{2})$	center of the Tile
Precision (PPV)	$(1, 0)$	corner
NPV	$(0, 1)$	corner
TPR (recall)	$(1, 1)$	corner
TNR (specificity)	$(0, 0)$	corner
F_β	$(1, \frac{\beta^2}{1+\beta^2})$	right edge ($a = 1$: no true negative)
Balanced accuracy	(π^-, π^-)	depends on the priors

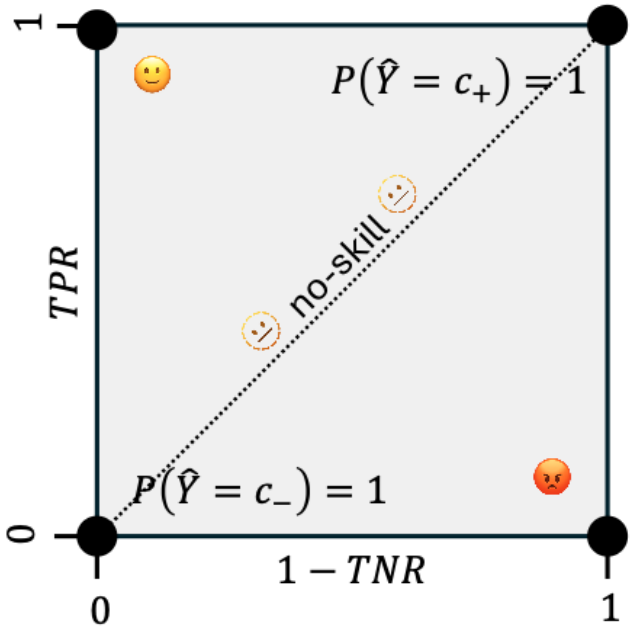
- All TN-free scores (precision, recall, F_β, IoU) lie on the right edge $a = 1$.

Outline

1. Introduction to performance evaluation
2. Evaluation of building damage assessment methods: a first glance
3. A theory for performance evaluation
4. Building an evaluation pipeline
   - Step 1: Evaluation
   - Step 2: Summarization
   - Step 3: Importance-based analysis – the Tile
   - Understanding the Tile
   - Tile flavors
   - Step 4: Ranking
   - Taking the chance into account: Illegal Waste Dumping Detection (IWDD)
5. Towards a standardized evaluation protocol

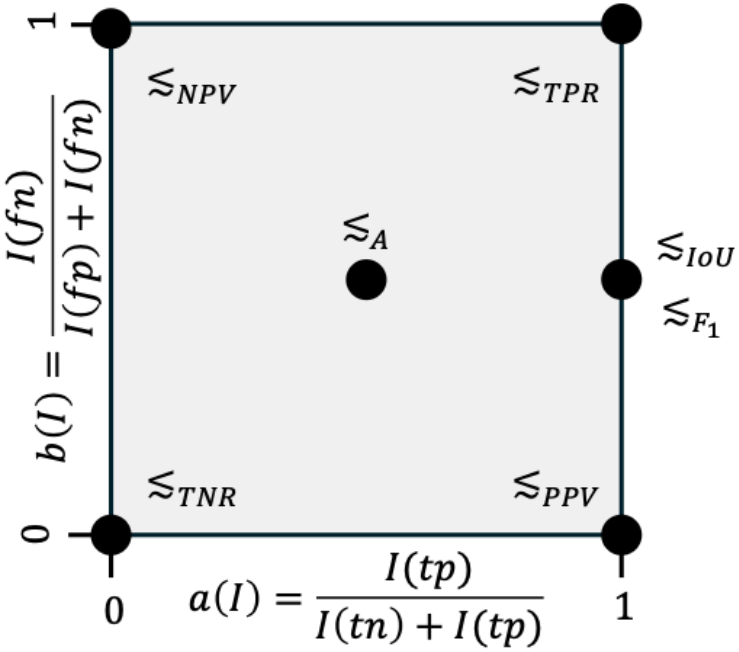
ROC vs Tile

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



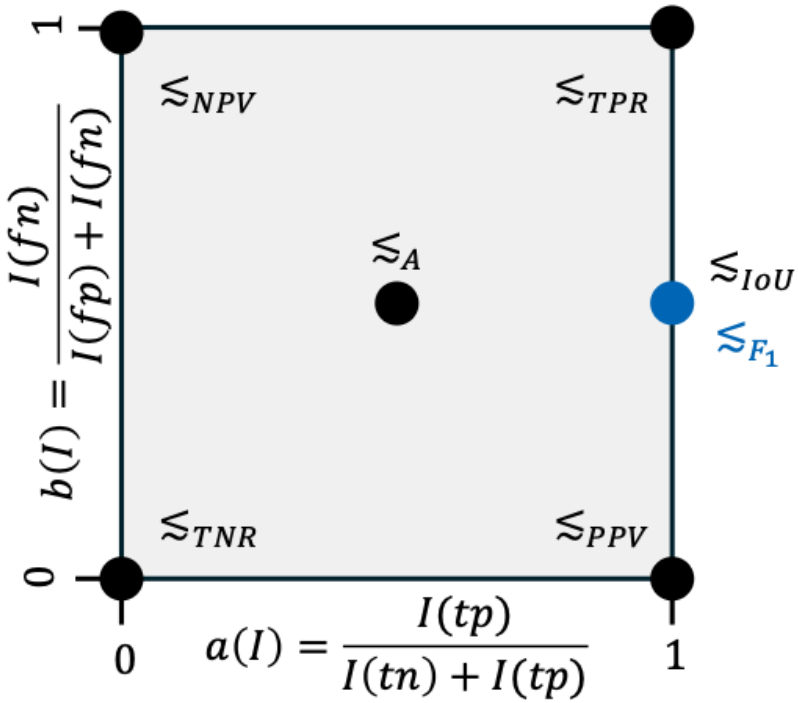
In this "space", canonical ranking scores, performance orderings, and rankings are projected as points.

F_1

Let us take

- $I(\mathrm{tn}) = 0,\ I(\mathrm{fp}) = 1$
- $I(\mathrm{fn}) = 1,\ I(\mathrm{tp}) = 2$

Then $a(I) = \frac{I(\mathrm{tp})}{I(\mathrm{tn})+I(\mathrm{tp})} = 1$ and $b(I) = \frac{I(\mathrm{fn})}{I(\mathrm{fp})+I(\mathrm{fn})} = \frac{1}{2}$.



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

This is the F_1 score

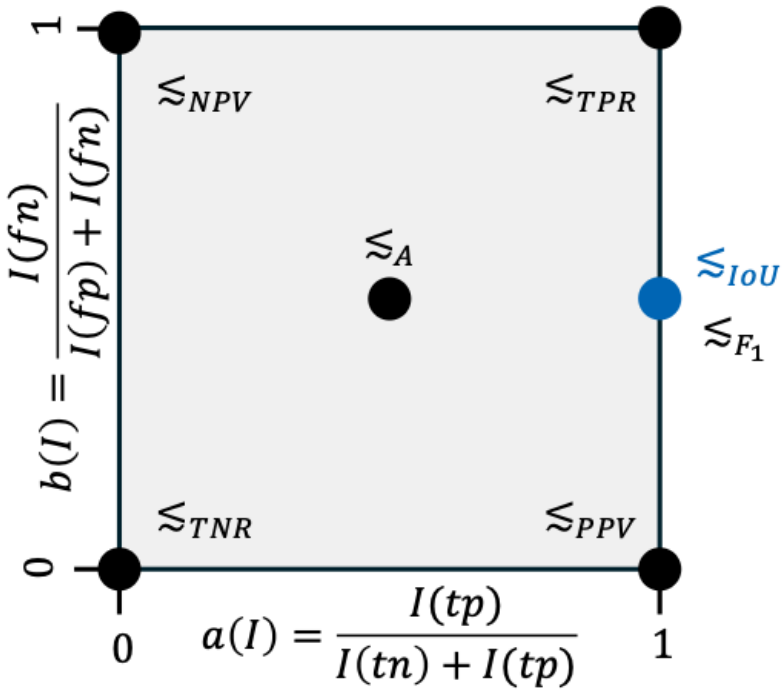
$F_1 = \left(\frac{1}{2}\mathrm{PPV}^{-1} + \frac{1}{2}\mathrm{TPR}^{-1}\right)^{-1}$. Synonym: Dice-Sørensen coefficient [19].

IoU

Let us take

- $I(\mathrm{tn}) = 0,\ I(\mathrm{fp}) = 1$
- $I(\mathrm{fn}) = 1,\ I(\mathrm{tp}) = 1$

Then $a(I) = 1$ and $b(I) = \frac{1}{2}$: the *same point* as F_1, hence the same ordering ($F_1 = 2\mathrm{IoU}/(1+\mathrm{IoU})$).



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

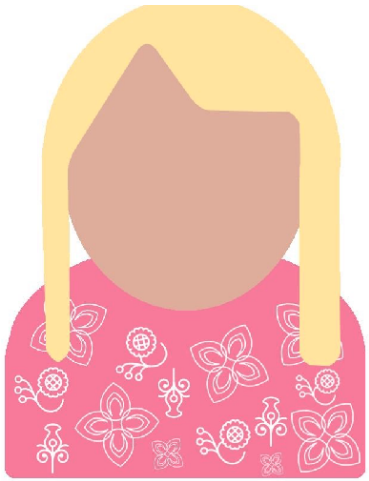
This is the Intersection over Union

$\mathrm{IoU} : P \mapsto P(S = 1 | Y = c^+ \vee \hat{Y} = c^+)$. Synonyms: Jaccard's coefficient, Tanimoto coefficient, similarity, critical success index, G-measure (in some works) [19].

Outline

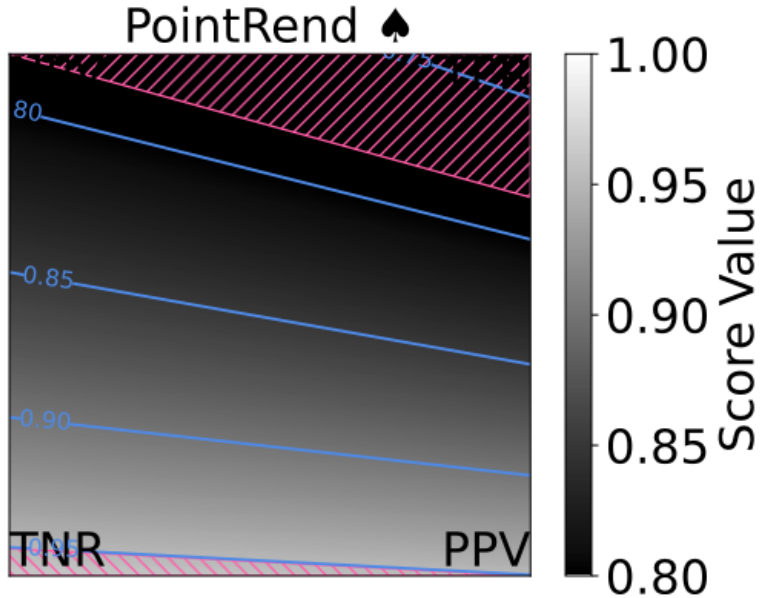
1. Introduction to performance evaluation
2. Evaluation of building damage assessment methods: a first glance
3. A theory for performance evaluation
4. Building an evaluation pipeline
   - Step 1: Evaluation
   - Step 2: Summarization
   - Step 3: Importance-based analysis – the Tile
   - Understanding the Tile
   - Tile flavors
   - Step 4: Ranking
   - Taking the chance into account: Illegal Waste Dumping Detection (IWDD)
5. Towards a standardized evaluation protocol

Value Tile



For a method designer

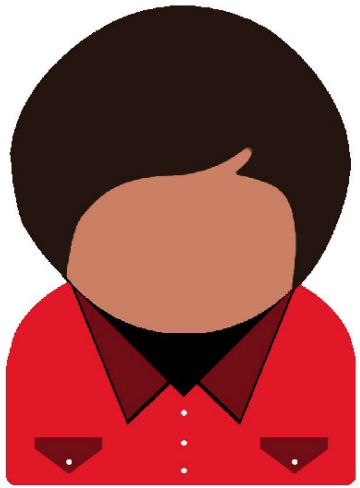
"What are the strengths and weaknesses of the method I'm developing?"



Suggested interpretation. The values taken by the canonical ranking scores show that the strength of the method is when the importance is put on the TNs and FPs, and its weakness when the importance is put on the TPs and FNs. Hatched areas: criteria for which a no-skill classifier (always predicting the same class) performs better[20].

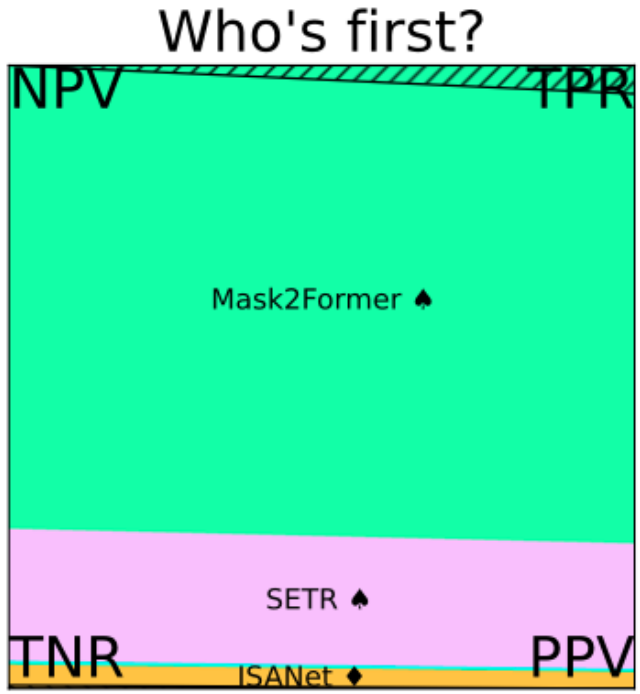
[20] Anastasia Zakharova et al. "Event Detection in Videos: A Framework for the Development of New Methods." In: *arXiv* abs/2607.04372 (2026). DOI: 10.48550/arXiv.2607.04372. arXiv: 2607.04372.

Entity Tile



For an app developer

"Show me the most appropriate method w.r.t. the application-specific preferences!"



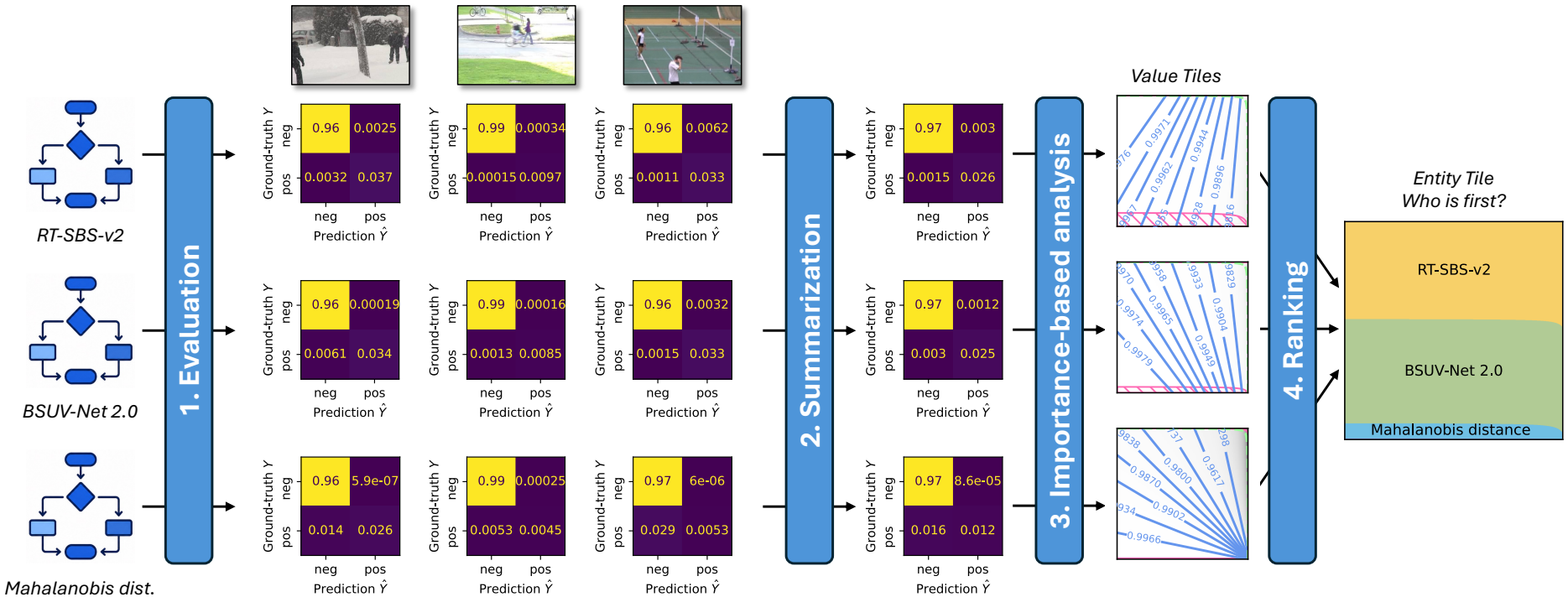
Suggested interpretation. Mask2Former is the best method for most criteria. Only when much more importance is given to the FPs than to the FNs (bottom of the Tile) does SETR come first, and ISANet at the very bottom edge. Hatched area: criteria for which a no-skill classifier performs better[21].

[21] Anastasia Zakharova et al. "Event Detection in Videos: A Framework for the Development of New Methods." In: *arXiv* abs/2607.04372 (2026). DOI: 10.48550/arXiv.2607.04372. arXiv: 2607.04372.

Outline

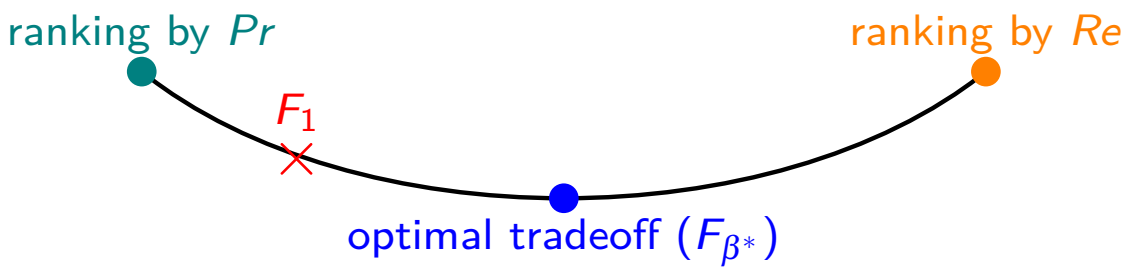
1. Introduction to performance evaluation
2. Evaluation of building damage assessment methods: a first glance
3. A theory for performance evaluation
4. Building an evaluation pipeline
   - Step 1: Evaluation
   - Step 2: Summarization
   - Step 3: Importance-based analysis – the Tile
   - Understanding the Tile
   - Tile flavors
   - Step 4: Ranking
   - Taking the chance into account: Illegal Waste Dumping Detection (IWDD)
5. Towards a standardized evaluation protocol

A probabilistic performance evaluation pipeline



Tradeoffs between score values vs. tradeoffs between rankings

- **Values.** F_1, harmonic mean of precision and recall, has a value *between* theirs.
- **Ranks.** Only the *ordering* matters: F_1 and IoU differ in value but induce *identical* rankings ($F_1 = 2\mathrm{IoU}/(1+\mathrm{IoU})$ is increasing).
- An intermediate value is no guarantee of an intermediate ranking.

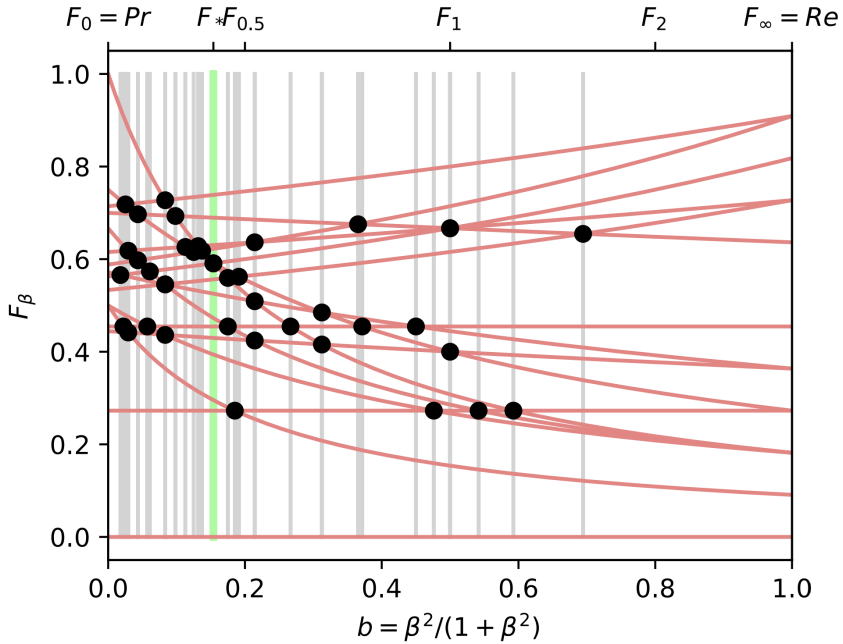


- The rankings induced by the F_β scores form a path between the rankings by *Pr* and *Re*; the optimal tradeoff (Kendall rank correlation) lies at mid-path.
- Its β depends on the compared performances: it is *rarely* $\beta = 1$.

~~Source:~~[22]

[22] Sébastien Piérard, Adrien Deliège, and Marc Van Droogenbroeck. "What Is the Optimal Ranking Score Between Precision and Recall? We Can Always Find It and It Is Rarely F_1." In: *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Denver, CO, USA: IEEE, June 2026, pp. 9722–9731.

How the ranking changes with β: an example



- CADA-RRE challenge: the curves F_β of 16 performances against $b = \beta^2/(1+\beta^2)$, from Pr ($b = 0$) to Re ($b = 1$).
- Each black dot is a swap in the ranking (gray lines): 44 different rankings.
- Only a handful of swaps separate F_1 ($b = \frac{1}{2}$) from Re: the ranking by F_1 is close to the ranking by recall, and identical to it for $\beta \geq 1.508$.

Different importances generally lead to different rankings: there is no universal ranking.

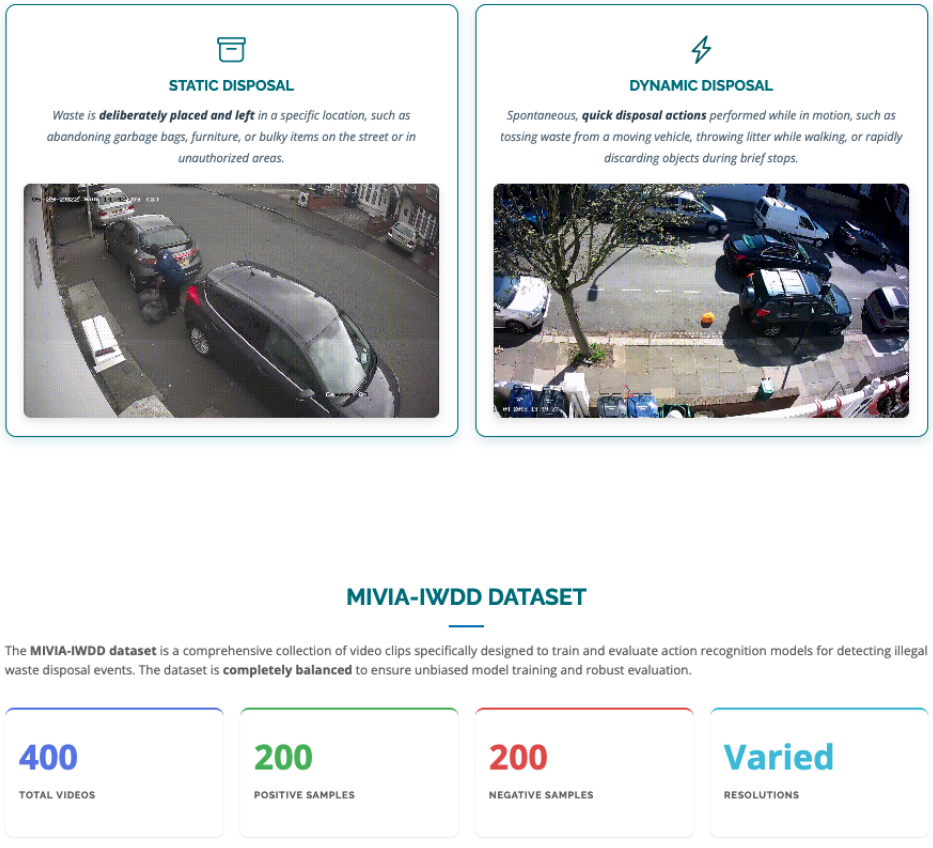
Source:[23]

[23] astien Piérard, Adrien Deliège, and Marc Van Droogenbroeck. "What Is the Optimal Ranking Score Between Precision and Recall? We Can Always Find It and It Is Rarely F_1." In: *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*.

Outline

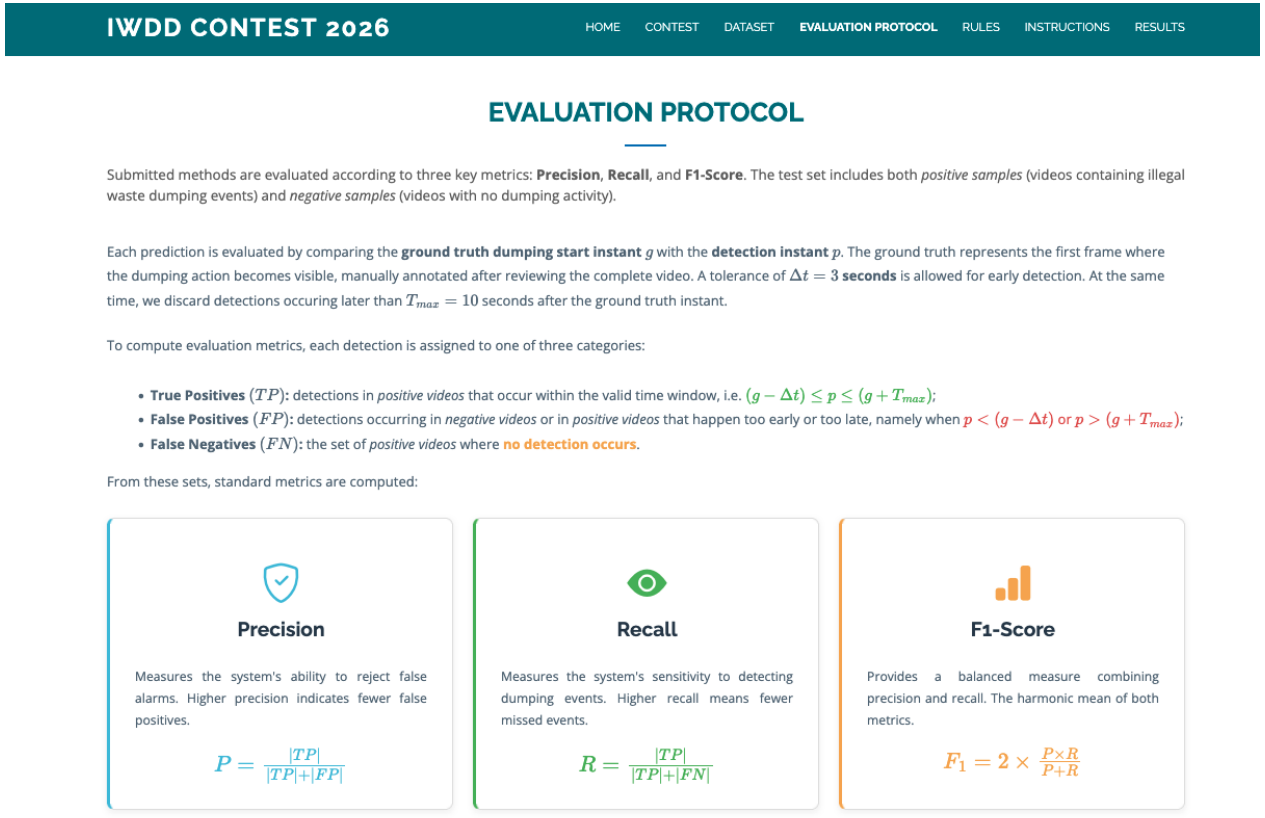
1. Introduction to performance evaluation
2. Evaluation of building damage assessment methods: a first glance
3. A theory for performance evaluation
4. Building an evaluation pipeline
   - Step 1: Evaluation
   - Step 2: Summarization
   - Step 3: Importance-based analysis – the Tile
   - Understanding the Tile
   - Tile flavors
   - Step 4: Ranking
   - Taking the chance into account: Illegal Waste Dumping Detection (IWDD)
5. Towards a standardized evaluation protocol

International contest on Illegal Waste Dumping Detection (IWDD)



- Test set: 200 videos
- Formulated as a detection task: F_1 is used to rank methods.

IWDD: metrics



IWDD CONTEST 2026

HOME CONTEST DATASET **EVALUATION PROTOCOL** RULES INSTRUCTIONS RESULTS

EVALUATION PROTOCOL

Submitted methods are evaluated according to three key metrics: **Precision**, **Recall**, and **F1-Score**. The test set includes both *positive samples* (videos containing illegal waste dumping events) and *negative samples* (videos with no dumping activity).

Each prediction is evaluated by comparing the **ground truth dumping start instant** g with the **detection instant** p. The ground truth represents the first frame where the dumping action becomes visible, manually annotated after reviewing the complete video. A tolerance of $\Delta t = 3$ **seconds** is allowed for early detection. At the same time, we discard detections occuring later than $T_{max} = 10$ seconds after the ground truth instant.

To compute evaluation metrics, each detection is assigned to one of three categories:

- **True Positives** (TP)**:** detections in *positive videos* that occur within the valid time window, i.e. $(g - \Delta t) \leq p \leq (g + T_{max})$;
- **False Positives** (FP)**:** detections occurring in *negative videos* or in *positive videos* that happen too early or too late, namely when $p < (g - \Delta t)$ or $p > (g + T_{max})$;
- **False Negatives** (FN)**:** the set of *positive videos* where **no detection occurs**.

From these sets, standard metrics are computed:

Precision

Measures the system's ability to reject false alarms. Higher precision indicates fewer false positives.

$$P = \frac{|TP|}{|TP|+|FP|}$$

Recall

Measures the system's sensitivity to detecting dumping events. Higher recall means fewer missed events.

$$R = \frac{|TP|}{|TP|+|FN|}$$

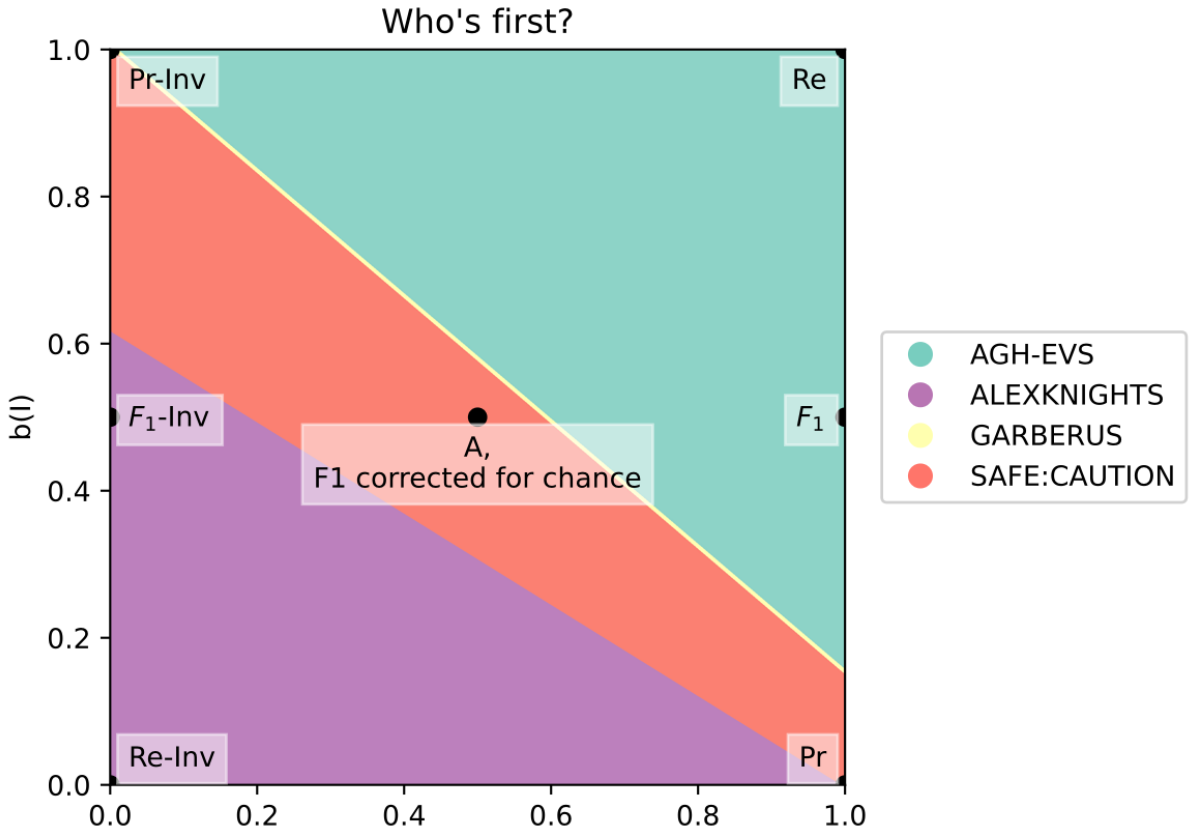
F1-Score

Provides a balanced measure combining precision and recall. The harmonic mean of both metrics.

$$F_1 = 2 \times \frac{P \times R}{P + R}$$

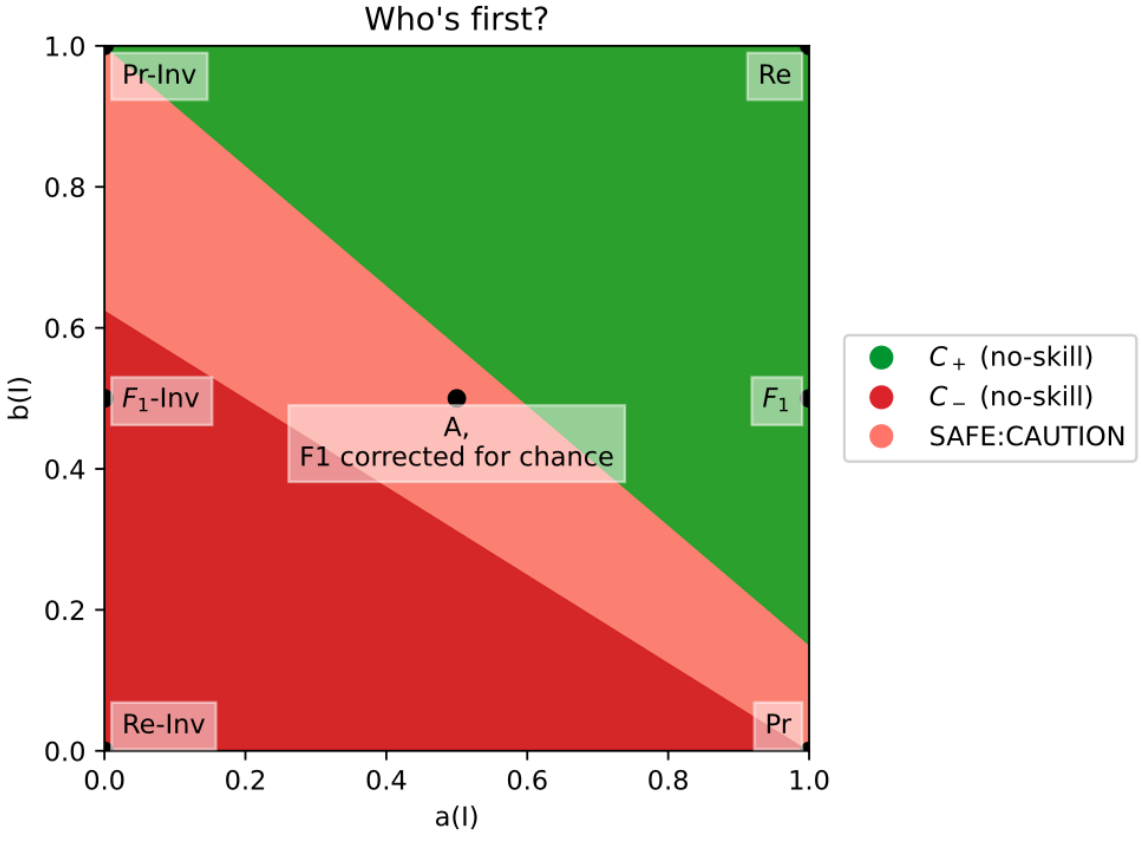
Task: clearly defined!

Evaluation on the test set



The Entity Tile shows that four methods cover the Tile.

On the private set



The Entity Tile shows that no method performs better than the positive no-skill classifier for the F_1 score.[a]

[a] Thierry Bouwmans et al. "Illegal waste dumping detection." In: *IEEE/CVF Winter Conf. Appl. Comput. Vis. Work. (WACVW)*. Tucson, AZ, USA, Mar. 2026, pp. 539–548.

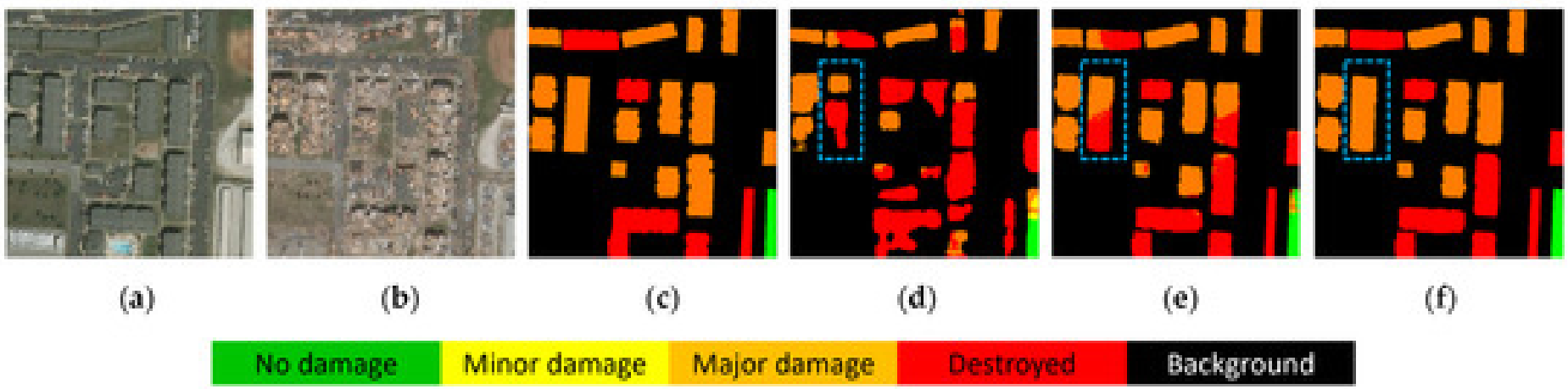
Outline

1. Introduction to performance evaluation
2. Evaluation of building damage assessment methods: a first glance
3. A theory for performance evaluation
4. Building an evaluation pipeline
5. Towards a standardized evaluation protocol

BDA methods: task definition

Tasks

Several tasks are implemented: *detection*, *classification*, *semantic segmentation*, *instance segmentation*, ..., but they are rarely defined explicitly!

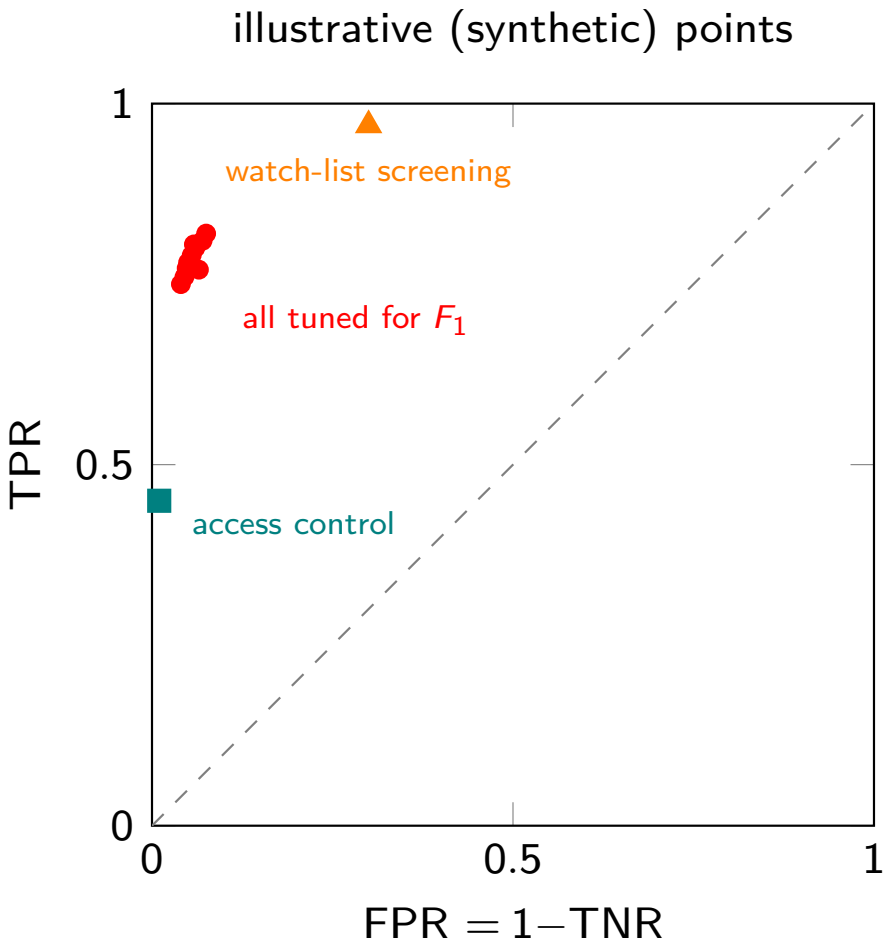


Why a single score is not enough

Performance evaluation in BDA is often based on a single score (F_1, mIoU, accuracy, ...). Shortcomings:

1. Ranking requires a score that satisfies precise axioms [21]: not all scores are ranking scores, and some only under constant priors (hard to guarantee on new images).
2. The absolute value of a score is only meaningful near its upper limit; arguments valid for values do not transfer to rankings. The ranking induced by F_1 is often closer to that of precision or recall than to a compromise [17].
3. A single criterion does not express the full potential of a method, and predicting its rank on a new dataset remains far from solved [20].
4. A single score impoverishes research: similar methods, incremental rather than disruptive improvements.
5. A single score carries the risk of overfitting on the score rather than on the task.

Note to challenge organizers: avoid the attractor effect



- When every method is optimized for the same single score (typically F_1), the field is pulled toward a narrow region of the performance space.
- This suppresses diversity and penalizes applications with other preferences.
- Same face matcher, opposite needs: *watch-list screening* must avoid FNs, *access control* must avoid FPs.
- In BDA, the AI Model Challenge for Earthquake Response recognized its top four teams "because of the closeness of results" [11].

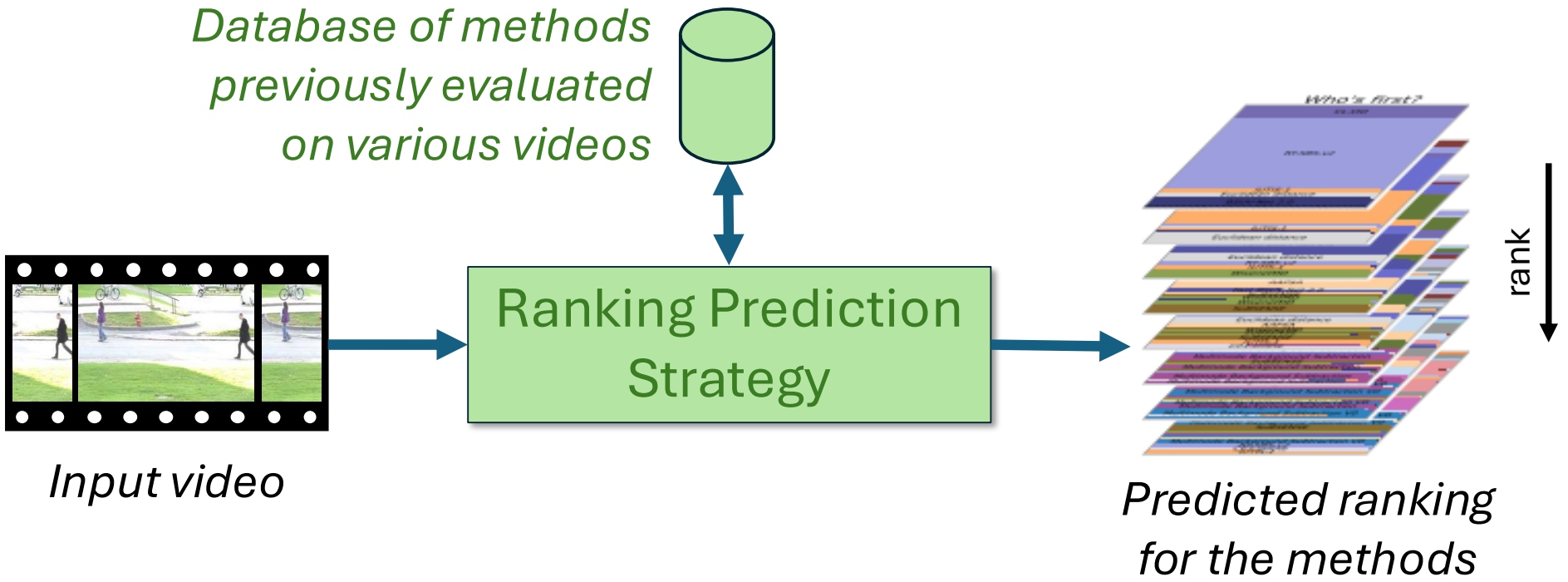
Remedy: diversify the evaluation scores, e.g., adopt the Tile as a reporting standard[24].

[24] Anastasia Zakharova et al. "Event Detection in Videos: A Framework for the Development of New Methods." In: *arXiv* abs/2607.04372 (2026). DOI: 10.48550/arXiv.2607.04372. arXiv: 2607.04372.

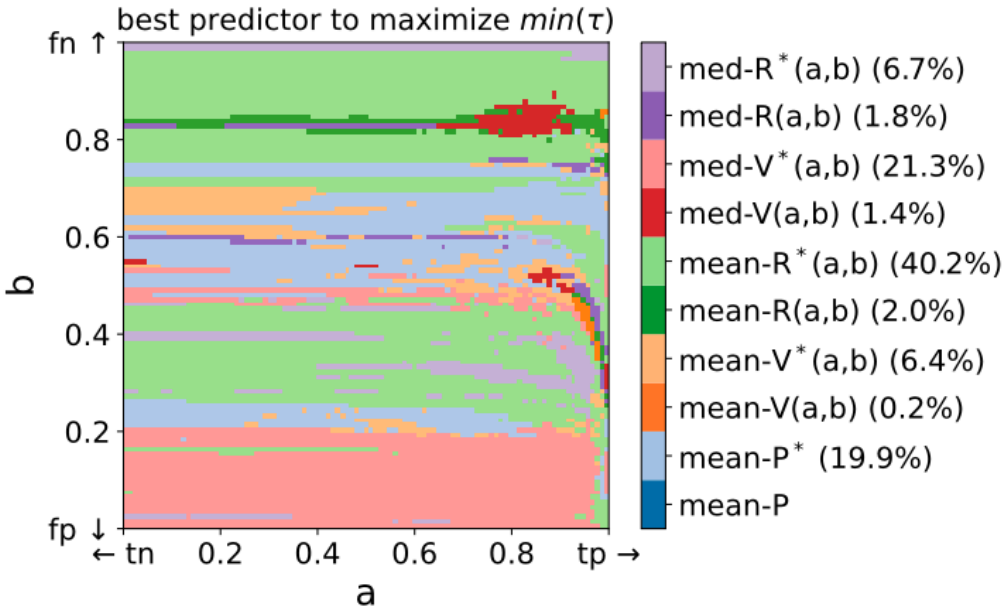
Generalization to unseen data is still an open problem: Predicting rankings (I)

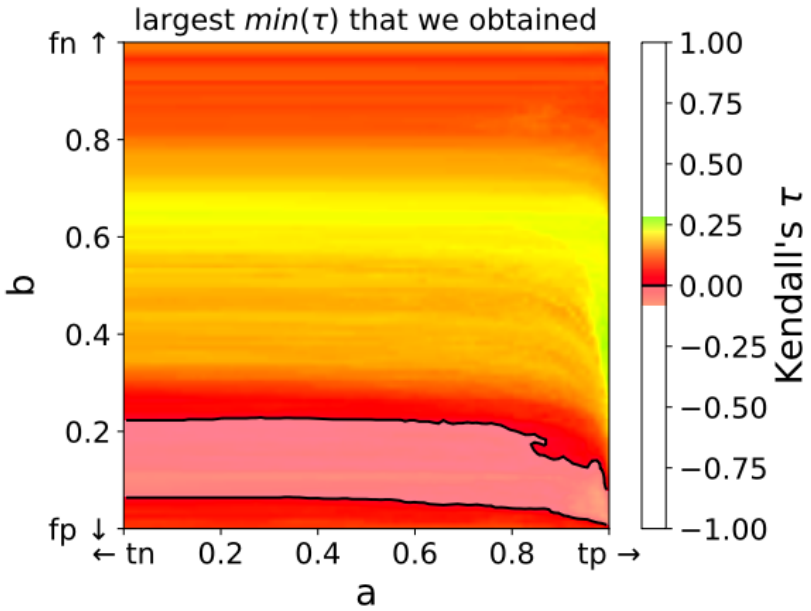
Question

Is it possible to predict the rankings of computer vision methods on any new domain (video) based on a database storing the performances of these methods, previously evaluated on other domains (videos)?



Predicting rankings (II)





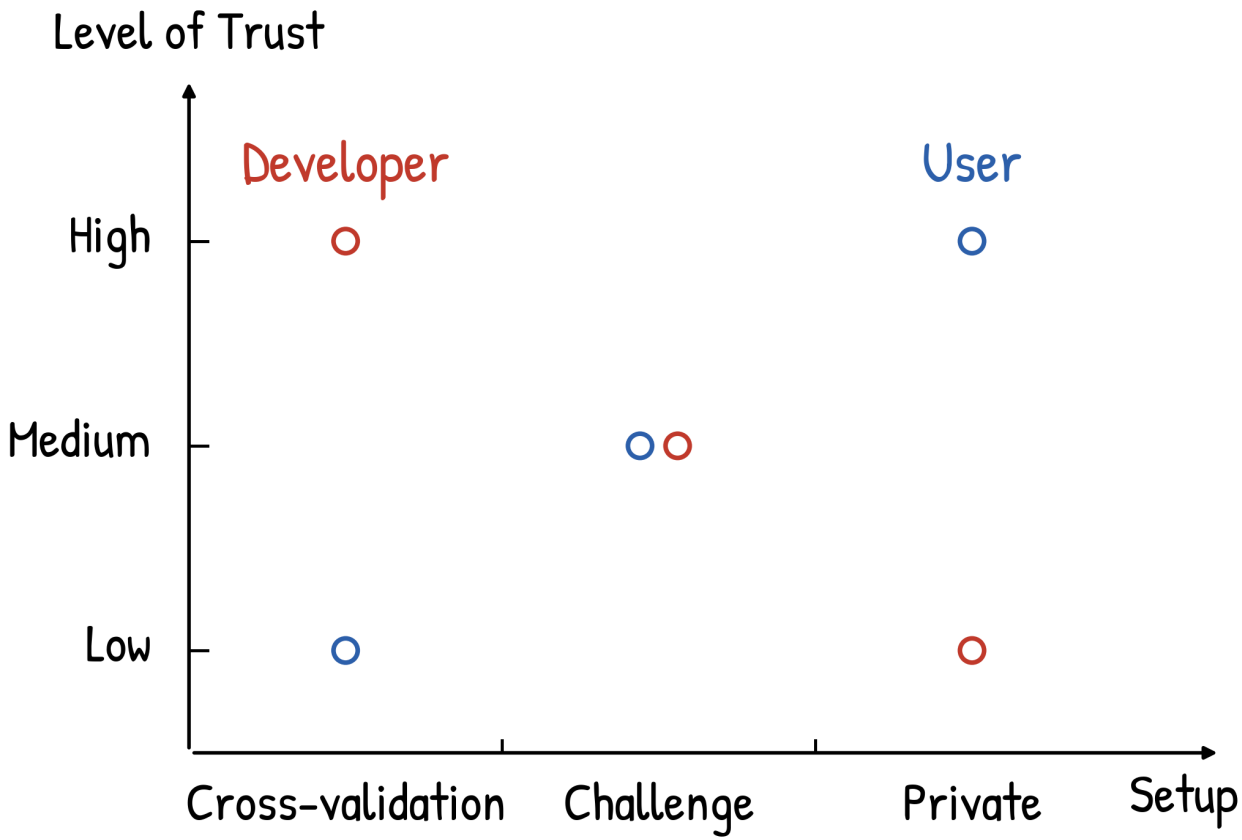
Unfortunately, predicting a ranking on a new domain is not easy/solved.[a]

[a] Sébastien Piérard et al. "A Methodology to Evaluate Strategies Predicting Rankings on Unseen Domains." In: *IEEE Int. Conf. Multimedia Expo Work. (ICMEW), Work. Big Surveill. Data Anal. Process. (big-surv)*. Nantes, France: IEEE, June 2025, pp. 1–6. DOI: 10.1109/icmew68306.2025.11152102.

Principles for a standardized and informative evaluation protocol

1. Compare comparable methods, that is, methods implementing the same scenario.
2. Emphasize the comparison of methods through *ranking*: the *stability* of the ranking takes precedence over score values.
3. Adopt a *multiple-score* approach, e.g., the Tile [19] when the task is similar to crisp classification (ecological argument).
4. Consider *distinct application cases* (best ratio of correct detections, fewest false positives, ...); their definition remains the prerogative of the end user.
5. Use a theoretically sound strategy for *multi-domain* evaluation [15].
6. Allow an estimation of the performance in a *transferability* mode, to anticipate the performance in a new situation (problem far from being solved).

Private set: building trust



Levels of trust between a developer and a user for three setups (Source:[25]).

[25] Anthony Cioppa et al. "Revisiting the Task of Building Damage Assessment by Satellite Imaging." In: *arXiv* abs/2610.13857 (2026). DOI: 10.48550/arXiv.2610.13857. arXiv: 2610.13857.

Key messages: evaluation is far from being solved

1. Define the task precisely: random evaluation experiment, outcomes, satisfaction.
2. Do not average scores blindly: average performances (summarization).
3. Interpreting a single score is debatable. Prefer
    1. ranking with ranking scores, and multiple scores (use Tile flavors),
    2. comparing a new method to baselines and no-skill classifiers.
4. Report full confusion matrices (performances); avoid the attractor effect of a single benchmark score.
5. Define precise application scenarios (data, method, evaluation, hardware), and report hardware indicators next to task scores.

- Predicting a performance or a rank on new data does not work well yet.
- New benchmarking strategies are needed, including setups in which private data are never shared.

References I

[1] Joseph Bertrand. *Calcul des probabilités*. Gauthier-Villars et fils, 1889.

[2] Thierry Bouwmans et al. "Illegal waste dumping detection." In: *IEEE/CVF Winter Conf. Appl. Comput. Vis. Work. (WACVW)*. Tucson, AZ, USA, Mar. 2026, pp. 539–548.

[3] Kendrick Boyd et al. "Unachievable Region in Precision-Recall Space and Its Effect on Empirical Evaluation." In: *Int. Conf. Mach. Learn. (ICML)*. Edinburgh, UK, June 2012, pp. 639–646.

[4] Hongruixuan Chen et al. "ChangeMamba: Remote Sensing Change Detection With Spatiotemporal State Space Model." In: *IEEE Trans. Geosci. Remote Sens.* 62 (2024), pp. 1–20. DOI: 10.1109/tgrs.2024.3417253.

[5] Hongruixuan Chen et al. "BRIGHT: a globally distributed multimodal building damage assessment dataset with very-high-resolution for all-weather disaster response." In: *Earth System Science Data* 17.11 (Nov. 2025), pp. 6217–6253. DOI: 10.5194/essd-17-6217-2025.

[6] Anthony Cioppa et al. "Revisiting the Task of Building Damage Assessment by Satellite Imaging." In: *arXiv* abs/2610.13857 (2026). DOI: 10.48550/arXiv.2610.13857. arXiv: 2610.13857.

[7] Jesse Davis and Mark Goadrich. "The Relationship Between Precision-Recall and ROC Curves." In: *Int. Conf. Mach. Learn. (ICML)*. Pittsburgh, Pennsylvania: ML Res. Press, June 2006, pp. 233–240. DOI: 10.1145/1143844.1143874.

References II

[8] Gery Geenens. "An essay on copula modelling for discrete random vectors; or how to pour new wine into old bottles." In: *arXiv* abs/1901.08741 (Feb. 2019). arXiv: 1901.08741.

[9] Ritwik Gupta et al. "Creating xBD: A Dataset for Assessing Building Damage from Satellite Imagery." In: *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW)*. Long Beach, CA, USA: Carnegie Mellon University, June 2019, pp. 10–17.

[10] Ritwik Gupta et al. "xBD: A Dataset for Assessing Building Damage from Satellite Imagery." In: *arXiv* abs/1911.09296 (2019). DOI: 10.48550/arXiv.1911.09296. arXiv: 1911.09296.

[11] International Charter: Space and Major Disasters and ESA Phi-lab. *AI Model Challenge for Earthquake Response Dataset – Final Results Announced*. https://disasterscharter.org/news/ai-model-challenge-for-earthquake-response-final-results-announced. Compiled from Charter archive imagery (KARI, USGS/Maxar, CNES/Airbus, Blacksky, CNSA). Oct. 2025.

[12] International Working Group on Satellite-based Emergency Mapping (IWG-SEM). *Emergency Mapping Guidelines: Building Damage Assessment Chapter*. Tech. rep. Version 1.0. UN-SPIDER, Sept. 2018.

[13] Valerio Marsocci et al. "PANGAEA: A Global and Inclusive Benchmark for Geospatial Foundation Models." In: *arXiv* abs/2412.04204 (2024). DOI: 10.48550/arXiv.2412.04204. arXiv: 2412.04204.

References III

[14] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation." In: *Int. Conf. 3D Vis. (3DV)*. Stanford, CA, USA: IEEE, Oct. 2016, pp. 565–571. DOI: 10.1109/3dv.2016.79.

[15] Sébastien Piérard, Adrien Deliège, and Marc Van Droogenbroeck. "Multi-domain performance analysis with scores tailored to user preferences." In: *arXiv* abs/2512.08715 (2025). DOI: 10.48550/arXiv.2512.08715. arXiv: 2512.08715.

[16] Sébastien Piérard, Adrien Deliège, and Marc Van Droogenbroeck. "The Optimal Ranking Tradeoffs Between Two Ranking Scores in Two-Class Classification." In: *arXiv* abs/2610.23294 (2026). DOI: 10.48550/arXiv.2610.23294. arXiv: 2610.23294.

[17] Sébastien Piérard, Adrien Deliège, and Marc Van Droogenbroeck. "What Is the Optimal Ranking Score Between Precision and Recall? We Can Always Find It and It Is Rarely F_1." In: *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Denver, CO, USA: IEEE, June 2026, pp. 9722–9731.

[18] Sébastien Piérard and Marc Van Droogenbroeck. "Summarizing the performances of a background subtraction algorithm measured on several videos." In: *IEEE Int. Conf. Image Process. (ICIP)*. Abu Dhabi, United Arab. Emir., Oct. 2020, pp. 3234–3238. DOI: 10.1109/ICIP40778.2020.9190865.

[19] Sébastien Piérard et al. "The Tile: A 2D Map of Ranking Scores for Two-Class Classification." In: *arXiv* abs/2412.04309 (2024). DOI: 10.48550/arXiv.2412.04309. arXiv: 2412.04309.

References IV

[20] Sébastien Piérard et al. "A Methodology to Evaluate Strategies Predicting Rankings on Unseen Domains." In: *IEEE Int. Conf. Multimedia Expo Work. (ICMEW), Work. Big Surveill. Data Anal. Process. (big-surv)*. Nantes, France: IEEE, June 2025, pp. 1–6. DOI: 10.1109/icmew68306.2025.11152102.

[21] Sébastien Piérard et al. "Foundations of the Theory of Performance-Based Ranking." In: *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Nashville, TN, USA: IEEE, June 2025, pp. 14293–14302. DOI: 10.1109/cvpr52734.2025.01333.

[22] Gilbert Rotich and Sudeep Sarkar. "Evaluating the Robustness of Foundation Models for Satellite Imagery." In: *IEEE Access* 13 (2025), pp. 107720–107735. DOI: 10.1109/access.2025.3581067.

[23] Yu Shen et al. "BDANet: Multiscale Convolutional Neural Network With Cross-Directional Attention for Building Damage Assessment From Satellite Images." In: *IEEE Trans. Geosci. Remote Sens.* 60 (2022), pp. 1–14. DOI: 10.1109/tgrs.2021.3080580.

[24] Yao Sun, Yi Wang, and Michael Eineder. "QuickQuakeBuildings: Post-Earthquake SAR-Optical Dataset for Quick Damaged-Building Detection." In: *IEEE Geosci. Remote Sens. Lett.* 21 (2024), pp. 1–5. DOI: 10.1109/lgrs.2024.3406966.

[25] Marc Van Droogenbroeck. "Probabilistic Formulations and Evaluation Spaces for the Detection Task." Unpublished manuscript. 2026.

References V

[26] Anastasia Zakharova et al. "Event Detection in Videos: A Framework for the Development of New Methods." In: *arXiv* abs/2607.04372 (2026). DOI: 10.48550/arXiv.2607.04372. arXiv: 2607.04372.