Metrics and Performance Evaluation of AI-based Moving Object Detection: A Revisitation

Marc Van Droogenbroeck

December 2024

References / Credits

Four papers:

- *Summarizing the performances of a background subtraction algorithm measured on several videos*. In **IEEE Int. Conf. Image Process. (ICIP)**, 2020.

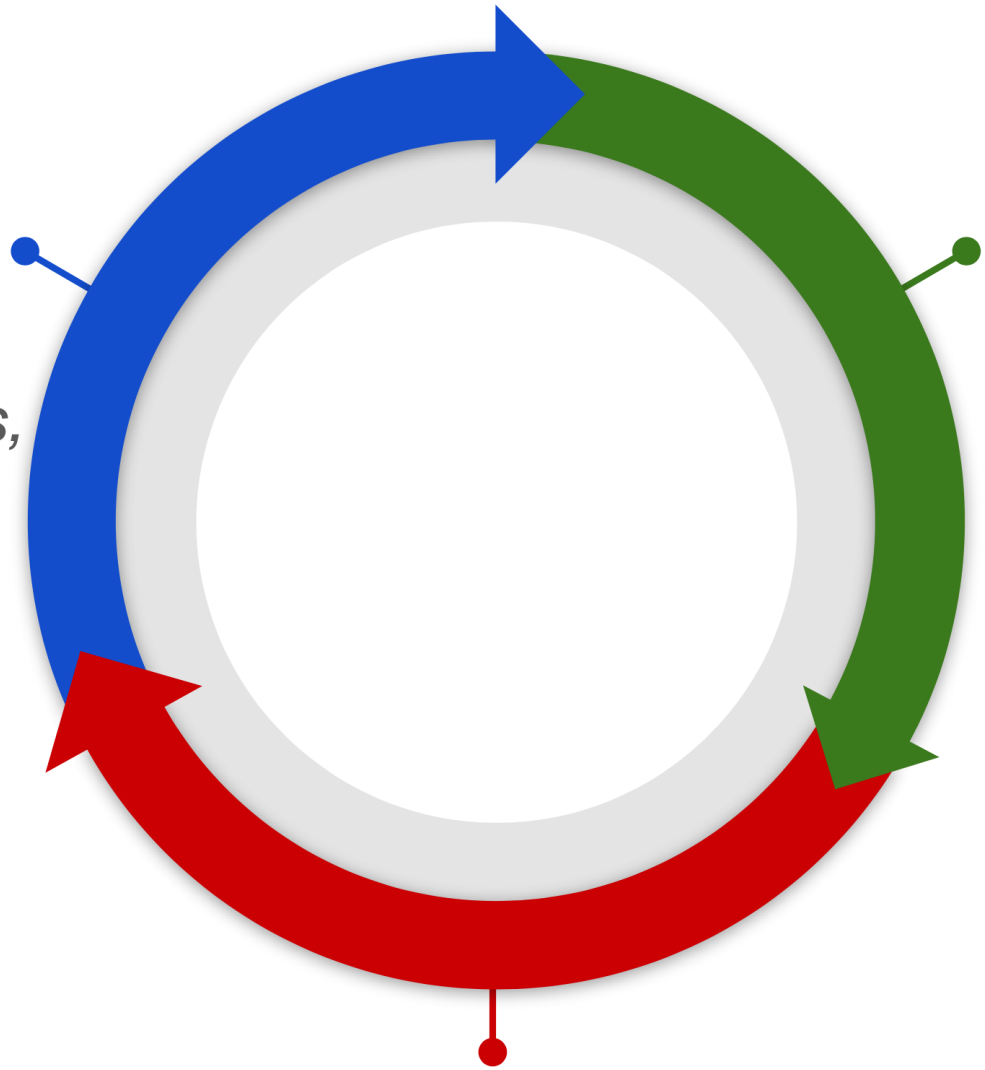
1. *Foundations of the Theory of Performance-Based Ranking*, **arXiv**, abs/2412.04227, December 2024.
2. *The Tile: A 2D Map of Ranking Scores for Two-Class Classification*, **arXiv**, abs/2412.04309, December 2024.
3. *A Hitchhiker's Guide to Understanding Performances of Two-Class Classifiers*, **arXiv**, abs/2412.04377, December 2024.

Co-authors:

Anaïs Halin, Sébastien Piérard, Anthony Cioppa, Adrien Deliège

Outline

Motivation: focus on the evaluation



Dataset

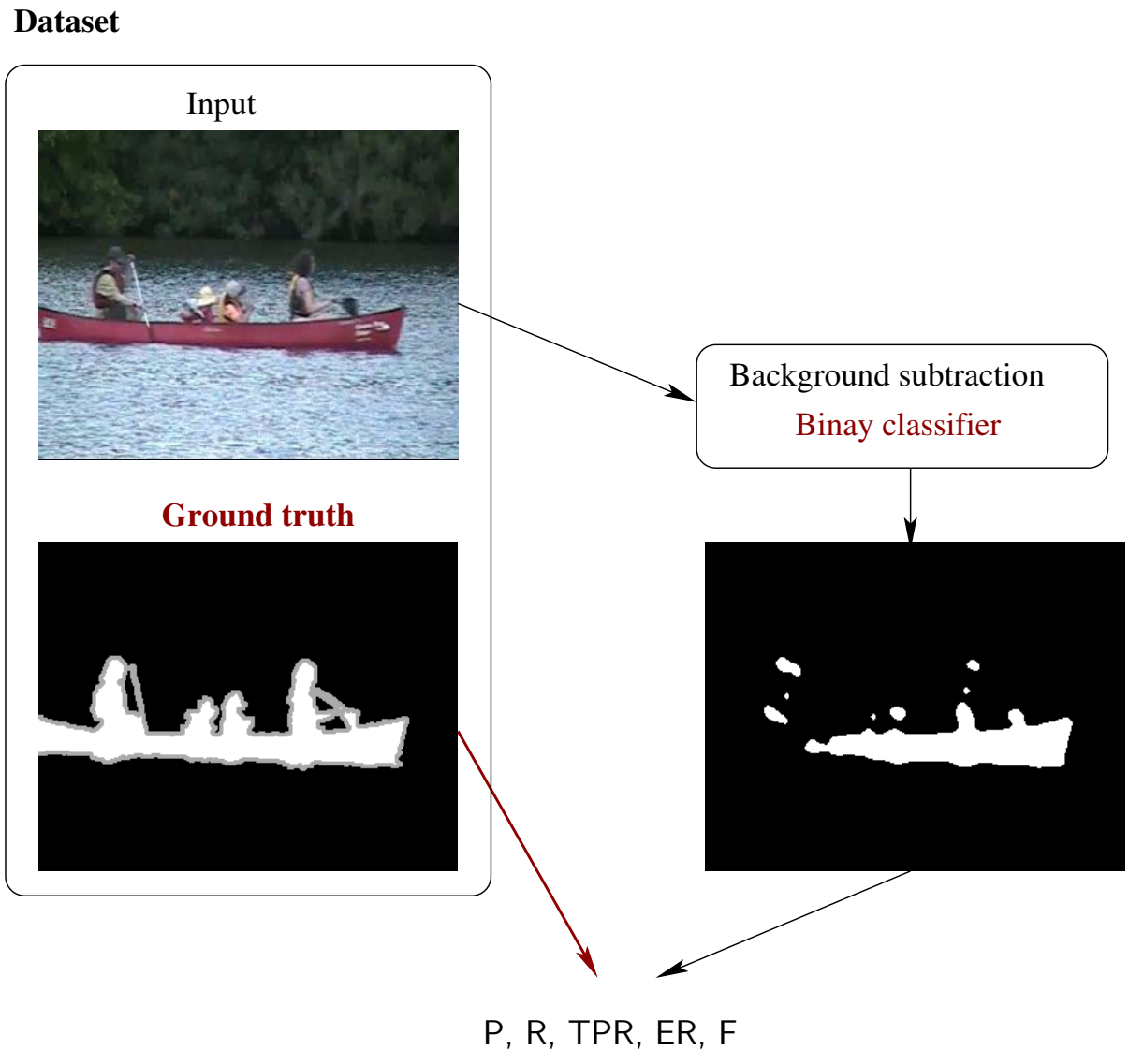
[CDNet: 11 categories, 53 videos]

Algorithms

[Task: motion detection by background subtraction]

Evaluation

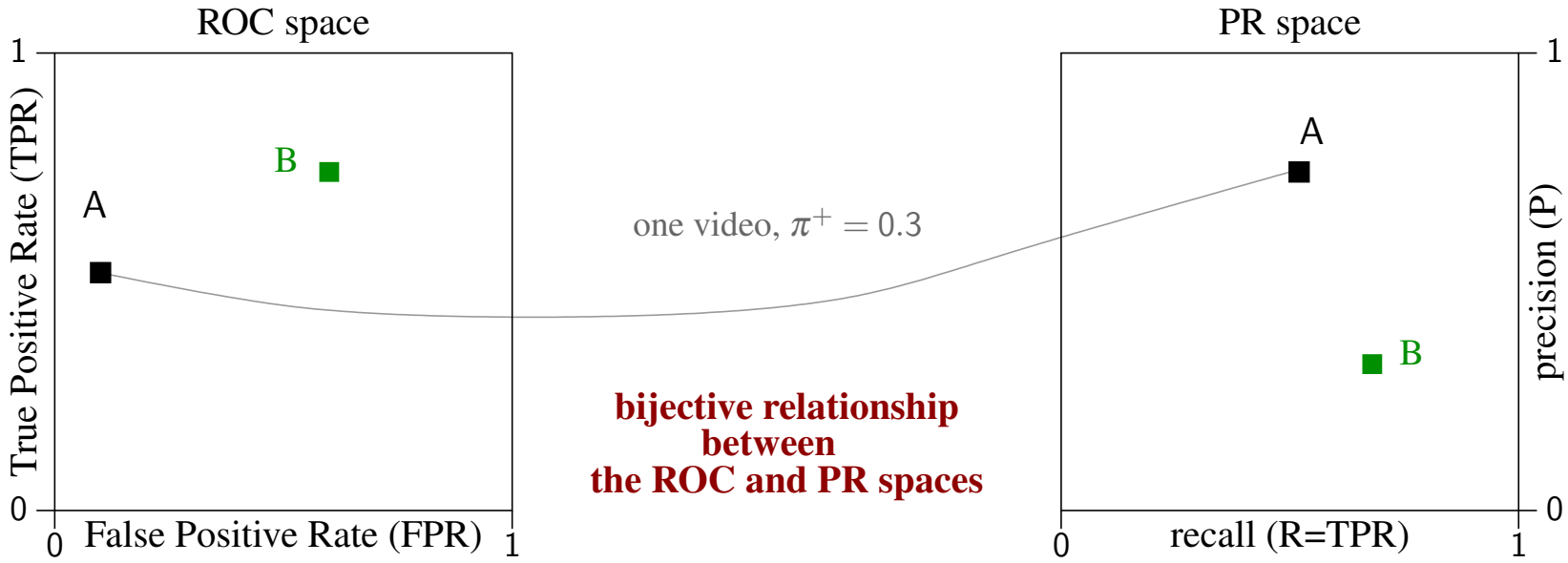
Evaluation of background subtraction algorithms: a solution by a series of scores



Evaluation of background subtraction algorithms: a solution by using well-known evaluation spaces

ROC: Receiver Operating Characteristic, defined by (FPR, TPR)

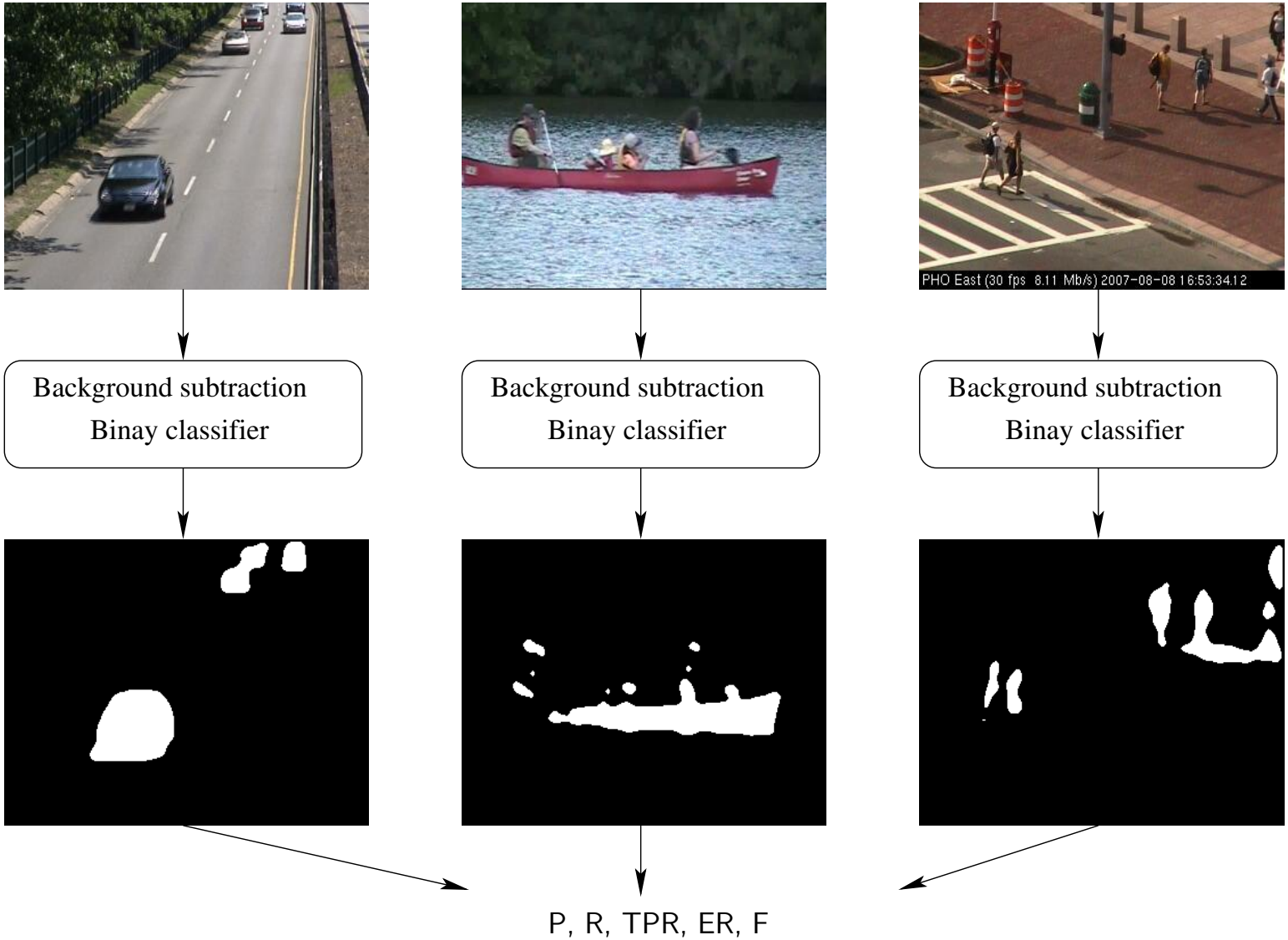
PR: Precision/Recall



Problems:

1. no clear ranking
2. family of classifier evaluated, not single ones!

Scoring multiple videos with a unique indicator



Series of difficulties

1. Which score(s) should we use to analyze/understand the performance of classifiers?
2. How can we determine if a newly designed classifier outperforms existing ones?
3. How do we decide which score to consider for ranking the classifiers (organization of challenges)?
4. How do we mix videos/domains/datasets?

Outline

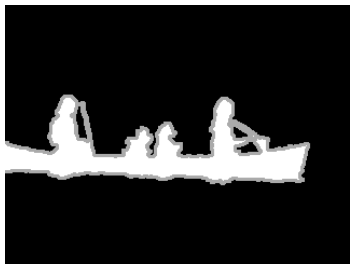
Components of a new framework for performance evaluation and ranking

Building a framework 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. A pre-order
5. Three axioms to define the notion of ranking

Step 1: A probabilistic (measurable) space (Ω, Σ); the confusion matrix

Ground truth



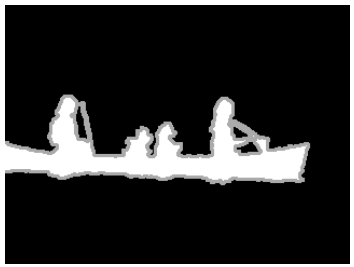
Output

		Predicted class $\hat{y}$	
		Positive	Negative
Actual class y	Positive	$\mathbf{p}(\mathrm{tp})$	$\mathbf{p}(\mathrm{fn})$
	Negative	$\mathbf{p}(\mathrm{fp})$	$\mathbf{p}(\mathrm{tn})$

The sample space contains 4 elements: $\mathrm{tn}, \mathrm{fn}, \mathrm{fp}, \mathrm{tn}$

Step 1: A probabilistic (measurable) space (Ω, Σ); the confusion matrix

Ground truth



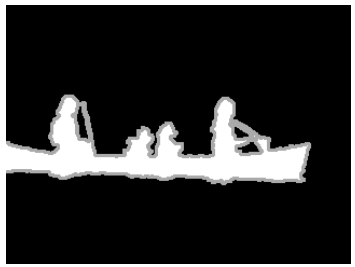
Output

		Predicted class $\hat{y}$	
		Positive	Negative
Actual class y	Positive	$\mathbf{p}(\mathrm{tp})$	$\mathbf{p}(\mathrm{fn})$
	Negative	$\mathbf{p}(\mathrm{fp})$	$\mathbf{p}(\mathrm{tn})$

The sample space contains 4 elements: $\mathrm{tn}, \mathrm{fn}, \mathrm{fp}, \mathrm{tn}$

Step 2: The Satisfaction variable (S) is used to specify a task

Ground truth

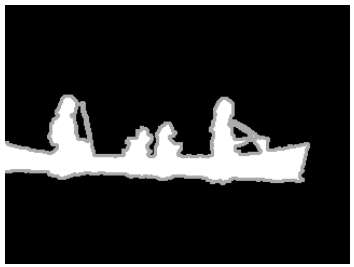


Output

		Predicted class $\hat{y}$	
		Positive	Negative
Actual class y	Positive	$\mathbf{p}(\mathrm{tp})$ $[S=1]$	$\mathbf{p}(\mathrm{fn})$ $[S=0]$
	Negative	$\mathbf{p}(\mathrm{fp})$ $[S=0]$	$\mathbf{p}(\mathrm{tn})$ $[S=1]$

Step 3: The Importance variable (I) is used to specify the needs of an application

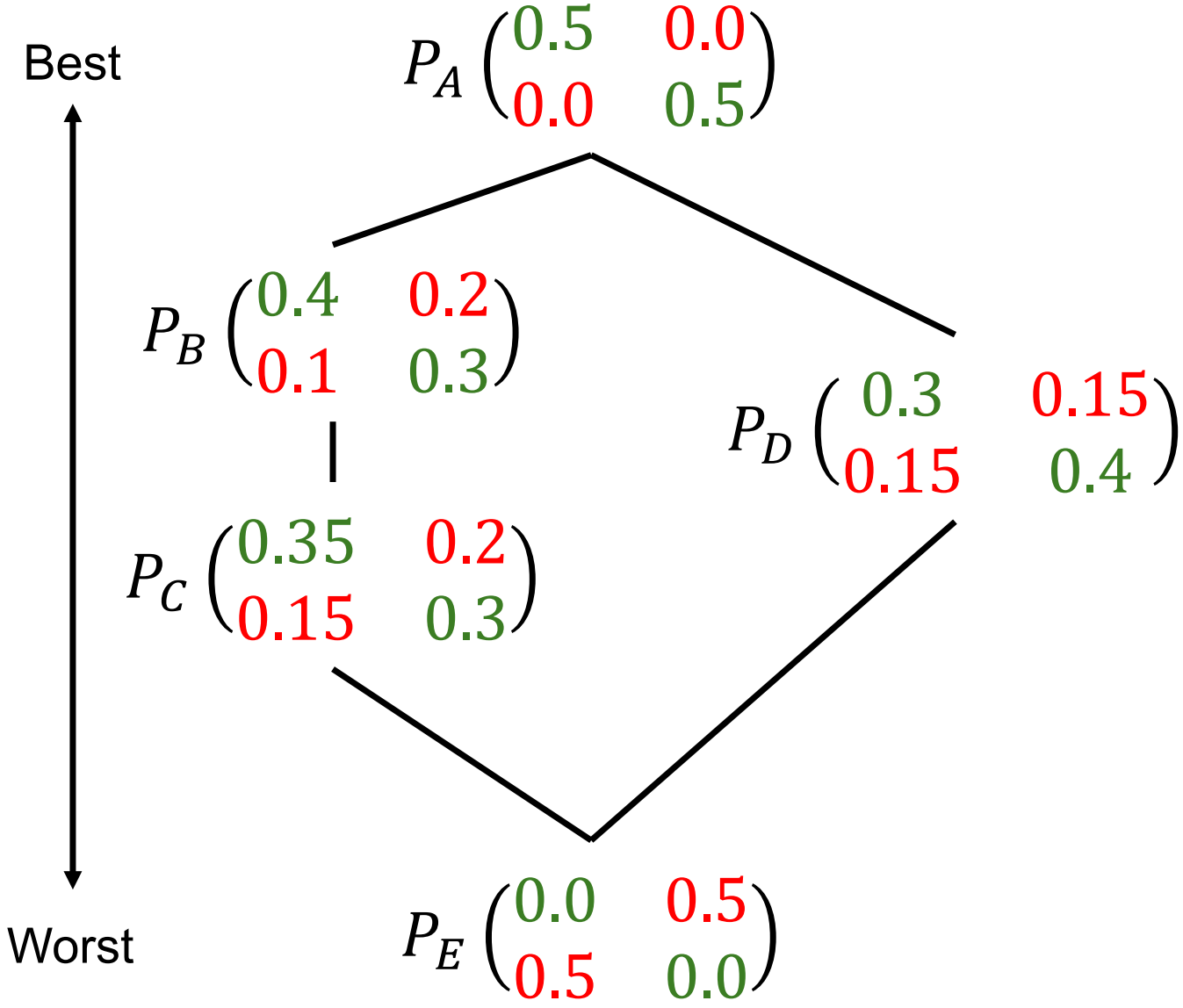
Ground truth



Output

		Predicted class $\hat{y}$	
		Positive	Negative
Actual class y	Positive	$\mathbf{p}(\mathrm{tp})$ $[I = a]$	$\mathbf{p}(\mathrm{fn})$ $[I = b]$
	Negative	$\mathbf{p}(\mathrm{fp})$ $[I = 1 - b]$	$\mathbf{p}(\mathrm{tn})$ $[I = 1 - a]$

Step 4: We need a pre-order



Step 5: Three axioms to define the notion of ranking

[Axiom 1] Adding a new method does not change the relative rank of the other methods.

[Axiom 2] The ranking should reflect the task, *i.e.* that we should be able to say which cases are satisfactory or not.

[Axiom 3] A blind combination of methods cannot lead to a better performance than the best method.

Note that the ranking method of CDNet2014 is not compliant with these axioms!

There is an infinite family of ranking scores to evaluate a performance P that satisfy those axioms

We have:

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

then, the family of ranking scores is

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

Particularization to two-class (binary) classification/segmentation

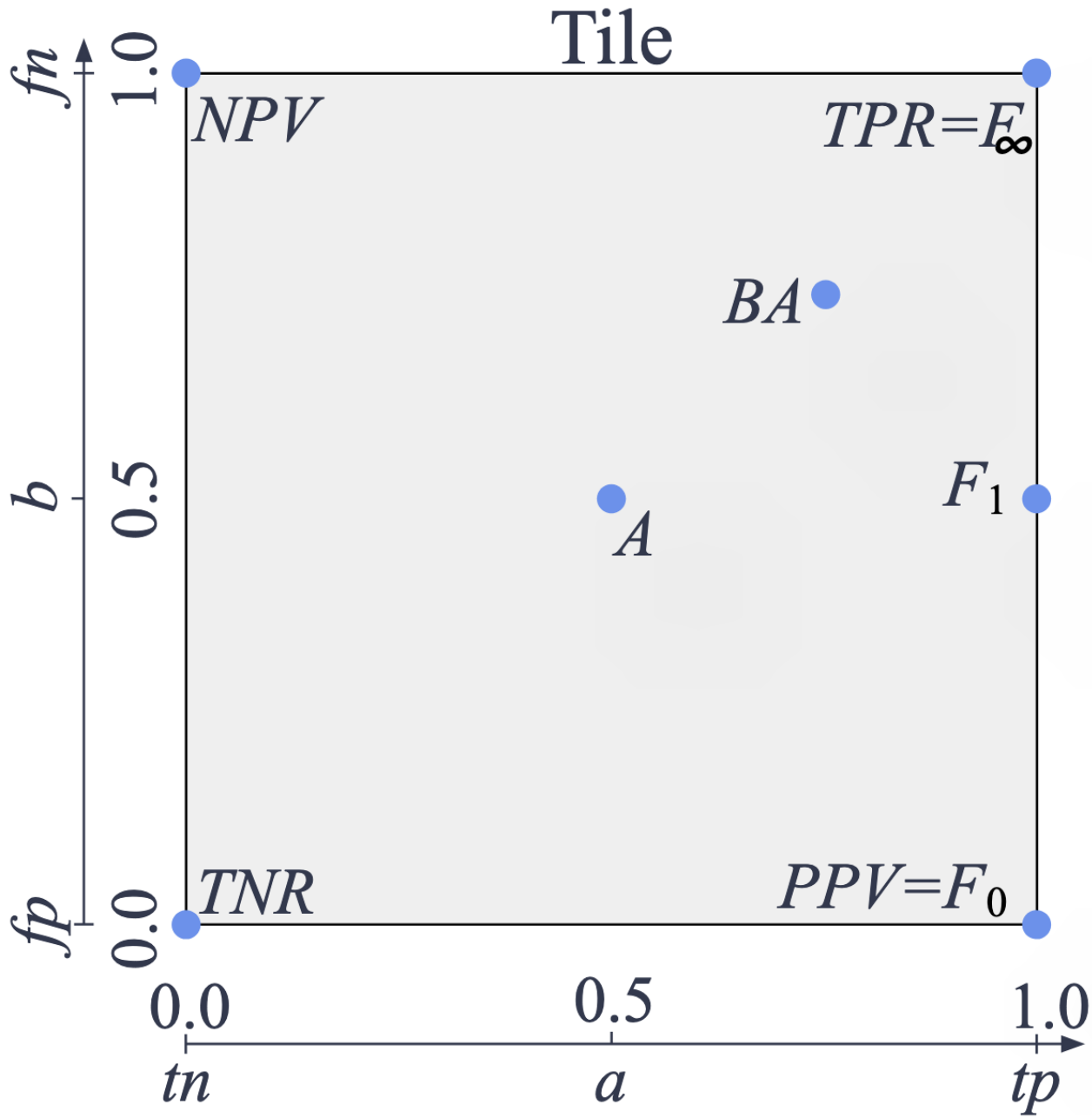
For a two-class (binary) classification/segmentation:

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

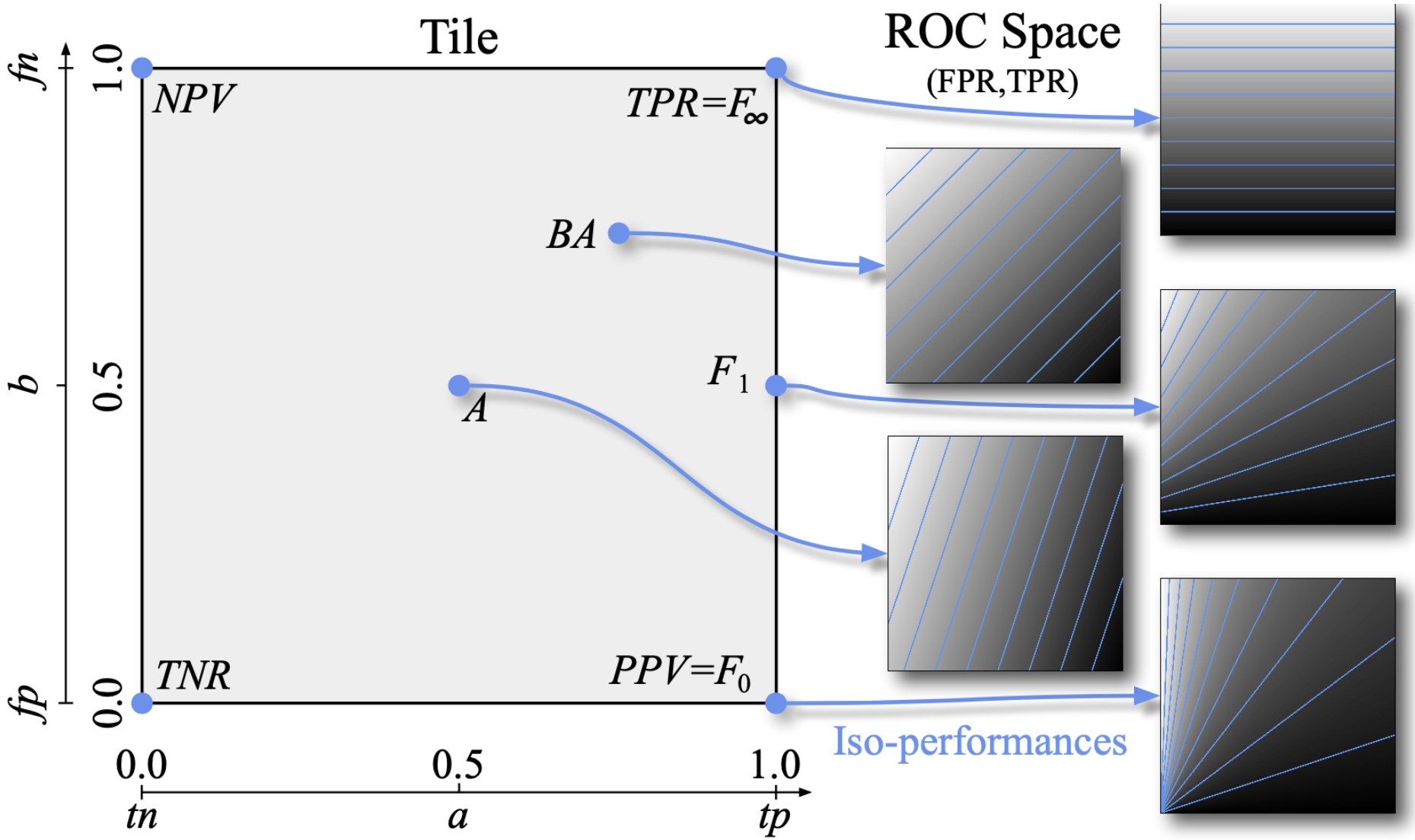
Therefore:

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

The Tile



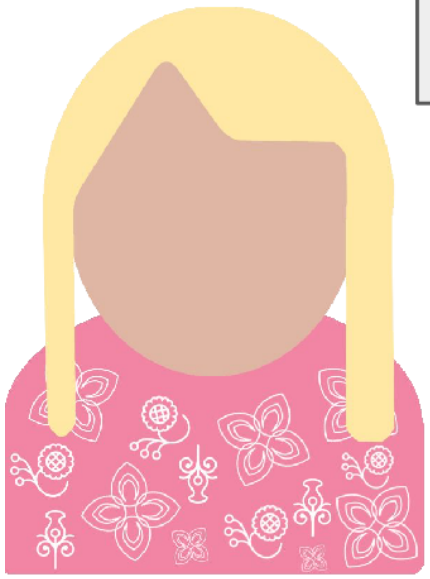
Link between the Tile and the ROC space

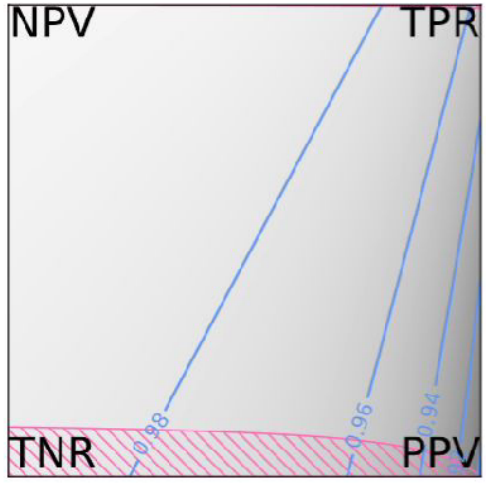


The Value Tile (Show me the value of all scores at once!)

I want to analyze the performances of my new method and compare it to the state of the art!

Use the *Value Tile*!

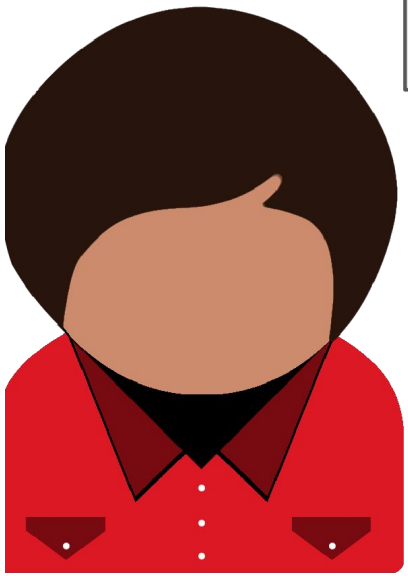


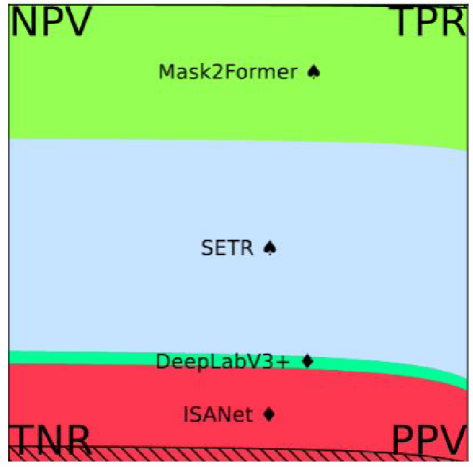


The Entity Tile (Who's first?)

I want to select the most appropriate method considering my application requirements!

Use the *Entity Tile*!





Outline

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

Obviously,

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

The M4CD algorithm of CDNet 2014 topically illustrates the problem

$$\bar{F} = 0.69 \quad \neq \quad \bar{\bar{F}} = 2\frac{\bar{P} \times \bar{R}}{\bar{P} + \bar{R}} = 0.75$$

In fact, the arithmetic mean has severe drawbacks:

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

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

The solution: use the summarization formulas

A probabilistic model for summarization

Definition (Parametric random experiment for several videos)

First, draw one video V at random in the set $\mathbb{V}$, following an arbitrarily chosen distribution $p(V)$. Then, draw one pixel at random from V and observe the ground-truth class Y and the predicted class $\hat{Y}$ for this pixel.

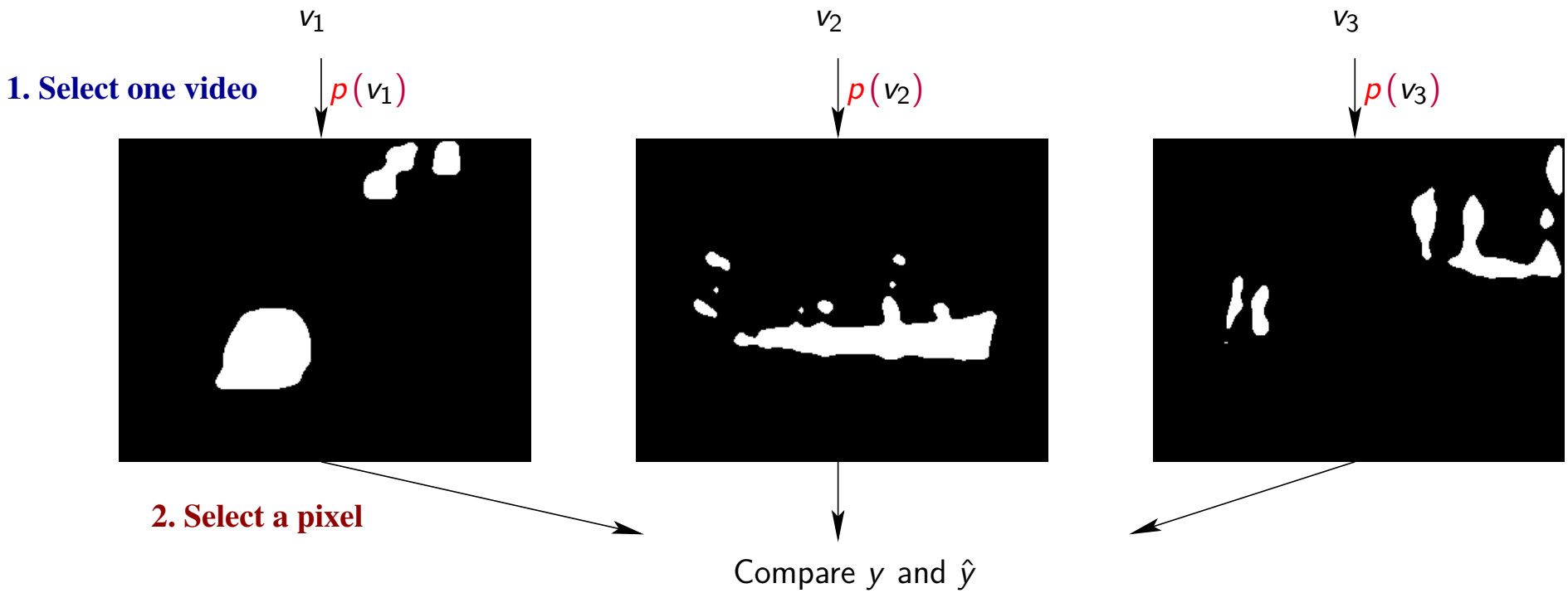


Figure: A probabilistic model for summarization: $\Delta = (V, Y, \hat{Y})$.

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.

Summarization with the Tile

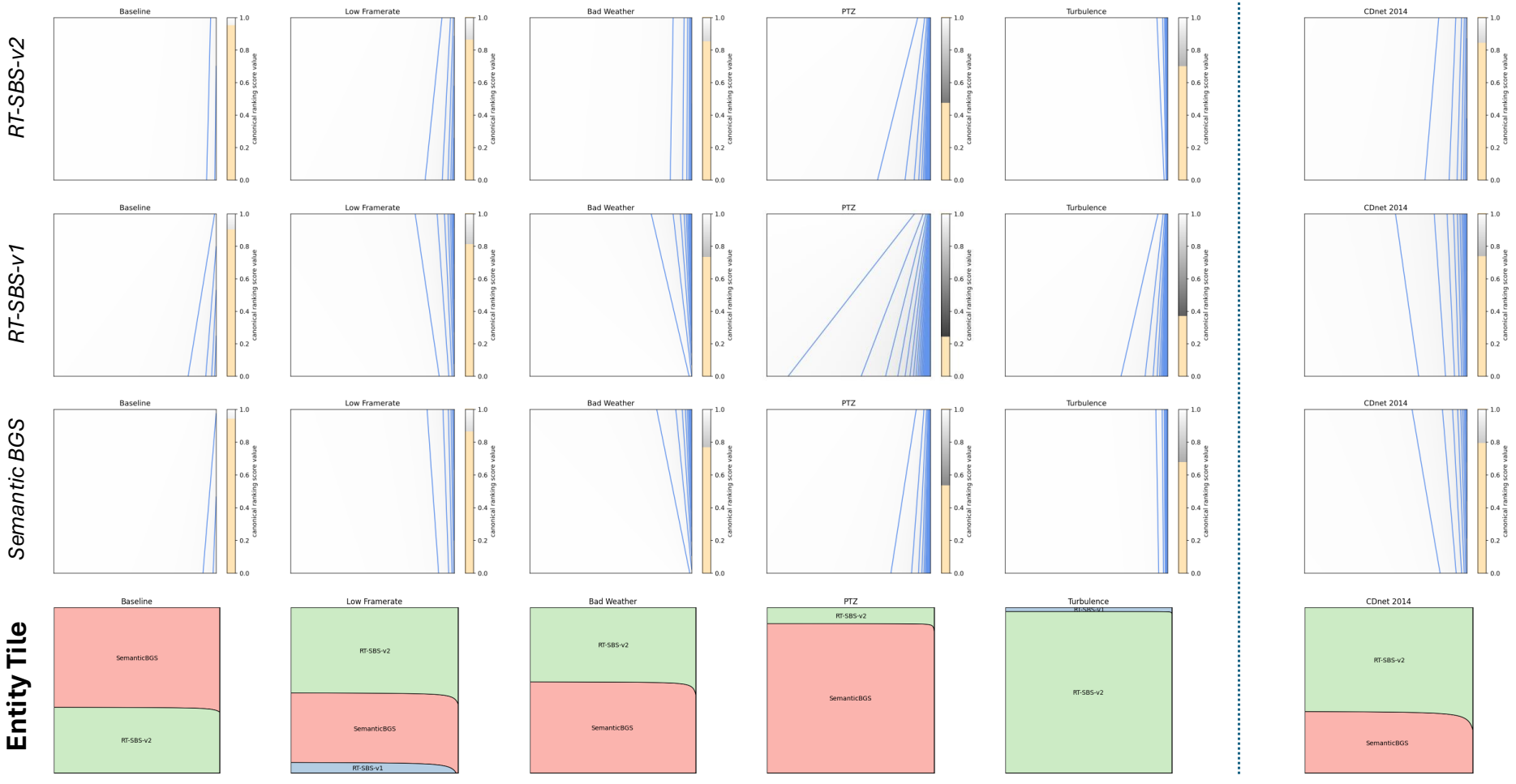
Procedure:

1. Calculate the values of the four corners for each Tile.
2. Use the summarization formulas to derive their values when combining multiple sources.
3. Based on the four re-calculated corners, proceed to the interpolation (as usual).

Multi-video/domain evaluation

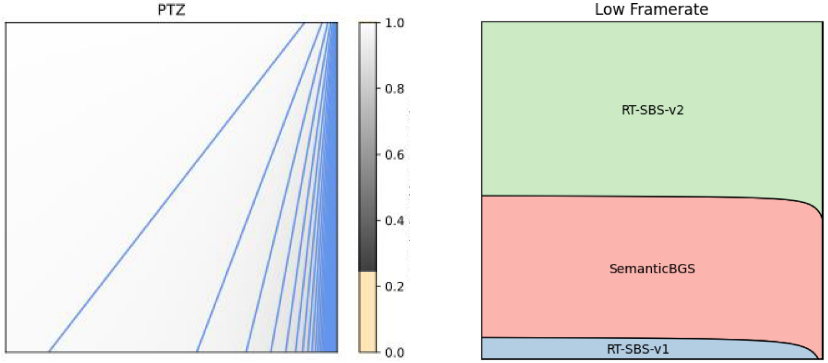
5 of the 11 categories

summarization



Take-home messages

- Proper evaluation requires to be precise about:
  - the task (via the Satisfaction)
  - the application (via the Importance)
  - the weights of included videos
- The Tile is a tool to capture, at a glance, a large family of known scores. It comes with different flavors:



- When organizing a challenge, do not take a unique ranking score to drive the developments of a community.