Department of Computer Science, University of Innsbruck - 9 March 2017

Adversarial Detection: theoretical foundations and Applications to Multimedia Forensics

Benedetta Tondi

University of Siena (Italy)

Summary

- Introduction to Adversarial Signal Processing
- **Adversarial Binary Detection**
- Contribution
- Theoretical analysis:
  - General framework for Binary Detection in the presence of adversary (some variants)
- Practical applications:
  - Multimedia Forensics
- Conclusions

Adversarial Signal Processing (AvdSP)

Motivations:

- Every digital system is exposed to *malicious* threats
- Security-oriented disciplines have to cope with the presence of adversaries
  - Watermarking - fingerprinting
  - Multimedia forensics
  - Spam filtering
  - intrusion detection
  - ….and many others

- Researchers have started looking for countermeasures, with *limited interaction*.

Adversarial Signal Processing (AvdSP)

- These fields face with similar problems
  - e.g. oracle attacks (in watermarking, in biometrics, in machine learning)
- ….and countermeasures are similar

Idea: a **unified framework**

A **unified view** would allow to:

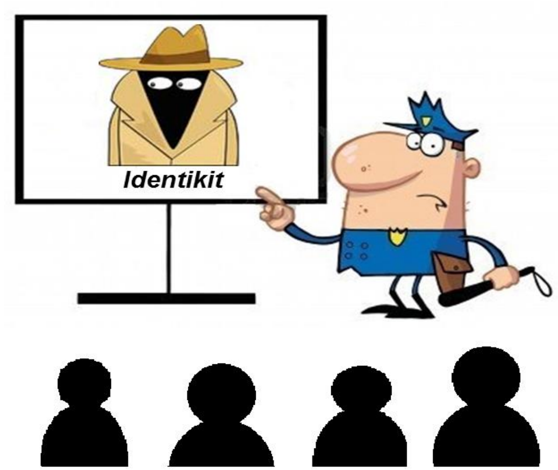
✓speed up the understanding of the security problems

✓work out effective and general solutions

Adversarial Signal Processing (AvdSP)

Purpose of AdvSP

Develop a general theory *of signal processing in the presence of an adversary.*



To do so, we need….

1. a model for the threat

2. a model for the interplay between Defender (D) and Attacker (A): a ***strategic interaction…..***

Tools: for modeling the D-A interplay (2.) - > ***Game Theory***

Binary Detection: a recurrent problem in SP

- Was a given image taken by a given camera ?
- Was this image resized/compressed twice ... ?

……an Attacker may aim at deleting the traces

Goal of the AdvBD: to study the *binary detection in the presence of adversary*

- Does this face/fingerprint/iris belong to Mr X ?

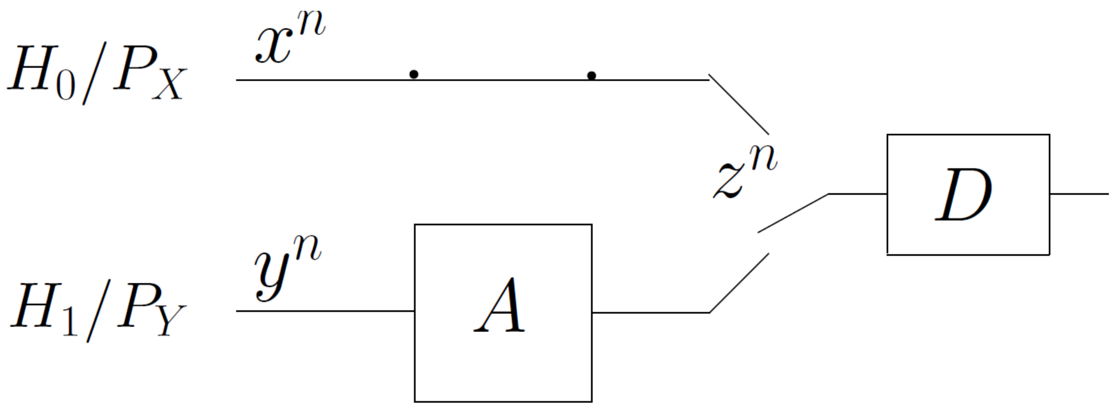
……an Attacker could build fake template

- Does an image contain a certain watermark ?

……an Attacker could either remove or inject illegally the watermark

Common element: the presence of an adversary aiming at making the test fail

Detection problem: basic setup



P_X and P_Y : pmf's of discrete memoryless sources X and Y

- **Goal of the Defender (D)**: decide if a sequence has been generated by P_X (under H_0) or P_Y (under H_1)
- **Goal of the Attacker (A)**: modify a sequence generated by P_Y so that it looks as if it were generated by P_X subject to a distortion constraint

A motivating example from Image Forensics

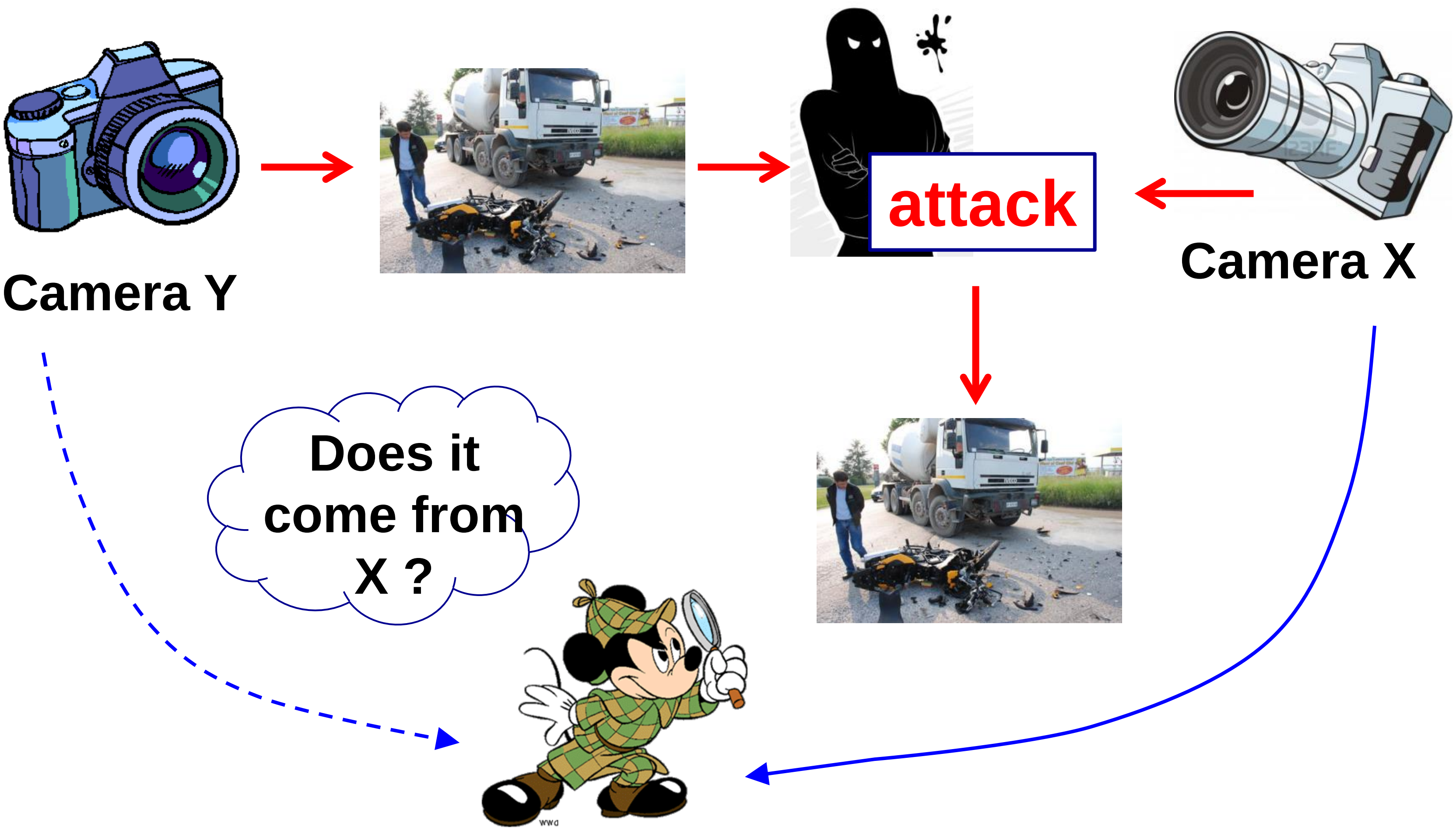
What is Multimedia Forensics ?

- Security-oriented discipline
- Goal: to retrieve information on the history of multimedia documents

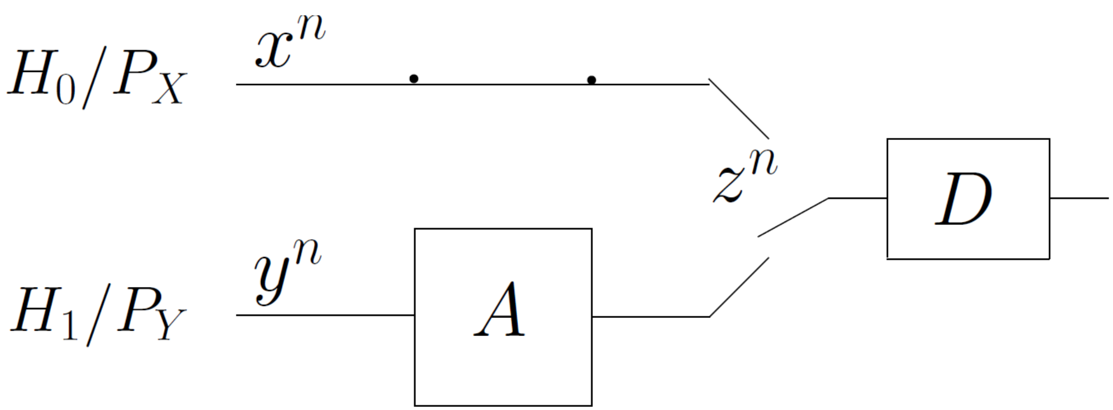
Forensic Analyst

Image Forensics: the media under analysis is an image

A motivating example from Image Forensics



Detection problem: basic setup



P_X and P_Y : pmf's of discrete memoryless sources X and Y

- **Goal of the Defender (D)**: decide if a sequence has been generated by P_X (under H_0) or P_Y (under H_1)
- **Goal of the Attacker (A)**: modify a sequence generated by P_Y so that it looks as if it were generated by P_X subject to a distortion constraint

Starting from this setup….

- We study the problem of the Adversarial Binary Detection in different scenarios depending on:
  - Threat model: attack under H_0 only or under both H_0 and H_1
  - Decision based on single or multiple observations
  - Knowledge available to Defender and Attacker (full or based on training data)
  - Possibility for the attacker of corrupting the training data

For the theoretical part….

What we will cover:

- **Detection games with known sources**

…..and

- Detection games with training data
- Detection games with corruption of the training

Theoretical analysis: Binary Detection Games

Game Theory in a nutshell

Two players, strategic game

$$G(S_1, S_2, u_1, u_2)$$

$S_1 = \{s_{1,1}, s_{1,2}, ..., s_{1,m_1}\}$ Set of strategies of Player 1

$S_2 = \{s_{2,1}, s_{2,2}, ..., s_{2,m_1}\}$ Set of strategies of Player 2

$u_1(s_{1,i}, s_{2,j})$ Payoff of Player 1 for a given profile $(s_{1,i}, s_{2,j})$

$u_2(s_{1,i}, s_{2,j})$ Payoff of Player 2 for a given profile $(s_{1,i}, s_{2,j})$

Competitive (zero-sum) game

$$u_1(\cdot,\cdot) = -u_2(\cdot,\cdot) = u$$

In game theory we are interested in the optimal choices of rationale players.

Game Theory in a nutshell

Nash equilibrium

None of the players gets an advantage by changing his strategy (assuming the other does not change his own)

$$u_1((s_{1,i^*}, s_{2,j^*})) \geq u_1((s_{1,i}, s_{2,j^*})) \quad \forall s_{1,i} \in \mathcal{S}_1$$
$$u_2((s_{1,i^*}, s_{2,j^*})) \geq u_2((s_{1,i^*}, s_{2,j})) \quad \forall s_{2,j} \in \mathcal{S}_2$$

$(s^*_{1,i}, s^*_{2,j})$

Nash equilibrium

Dominated strategy

$$u_1(s_{1,k}, s_{2,j}) > u_1(s_{1,i}, s_{2,j}) \quad \forall s_{2,j} \in \mathcal{S}_2$$

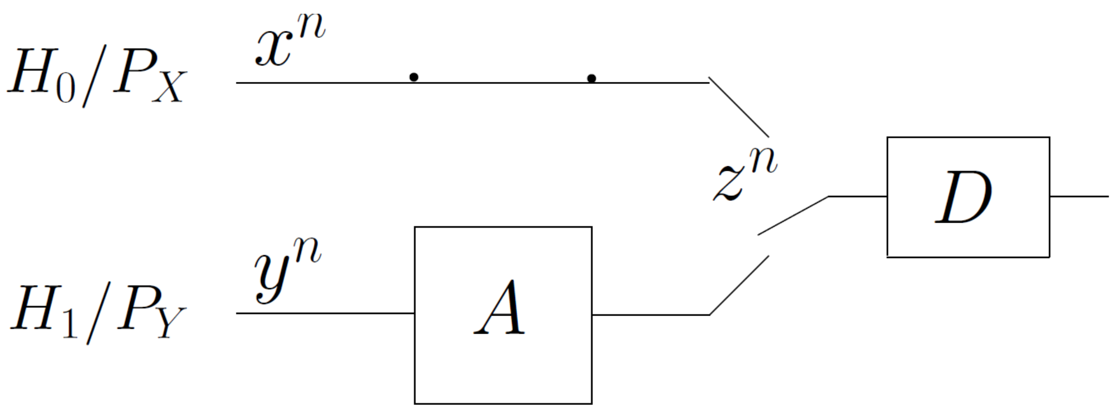
$s_{1,i}$ is strictly dominated by $s_{1,k}$

Rationalizable equilibrium

When the game can be solved through iterative elimination of strictly dominated strategies (Dominance solvability)

Detection games with known sources: DT_{ks}

Detection game with known sources* (DT_{ks})



- P_X and P_Y are known to A and D

Remarks

- A knows P_X. Worst case assumption
- D knows P_Y. Necessary for a valid game (relaxed later on)

*M. Barni, B, Tondi, "The Source Identification Game: an Information-Theoretic Perspective", *IEEE Trans. on Information Forensics and Security*, Vol. 8, No.3, March 2013

Strategies for the Defender (DT_{ks})

Set of *acceptance regions* of the test Λ^n...

- Neyman-Pearson (N-P) setup

> ***N-P setup**: D* puts a constraint on the *false positive* error probability **P_{FP}** (deciding H_1 when H_0 holds) and minimizes the *false negative* **P_{FN}** (deciding H_0 when H_1 holds)

Limitations

- D can rely *on first order statistics* only: $z^n \rightarrow P_{z^n}$
- *asymptotic analysis*

Empirical probability distribution or **type** of z^n

Then:

$$\mathcal{S}_D = \{\Lambda^n : P_{\mathrm{FP}} \leq 2^{-\lambda n}\}$$

Strategies for the Attacker (DT_{ks})

- Constraint on the *maximum (allowed) distortion* introduced

$$\mathcal{S}_A = \{g(\cdot) : d(y^n, g(y^n)) \leq nL\}$$

d(,) = distortion measure

L = maximum average per letter distortion

Remark:

- d(,) is permutation-invariant
- Note: considering determinitic functions is not a limitation (a posteriori)

The $\mathrm{DT}_{\mathrm{ks}}$ game

Set of strategies for D

$$\mathcal{S}_D = \{\Lambda^n : P_{\mathrm{FP}} \leq 2^{-\lambda n}\}$$

Λ^n defined by relying on P_{z^n} (first-order)

Set of strategies for A

$$\mathcal{S}_A = \{g(\cdot) : d(y^n, g(y^n)) \leq nL\}$$

$L,$ maximum average per letter distortion

Payoff (zero-sum game)

$$u(\Lambda^n, g) = -P_{\mathrm{FN}} = -\sum_{y^n : g(y^n) \in \Lambda^n} P_Y(y^n)$$

The DT_{ks} game: equilibrium point

Lemma (optimum defence strategy)

$$\Lambda^{n,*} = \left\{ P_{z^n} : \mathcal{D}(P_{z^n} || P_X) < \lambda - |\mathcal{X}| \frac{\log(n+1)}{n} \right\}$$

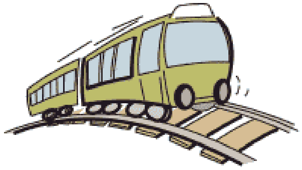
is a *dominant strategy* for the Defender.

K-L divergence

Remarks:

- regardless of the attacking strategy (the optimum strategy is *dominant!*)
- regardless of P_Y (the optimum strategy is *universal* w.r.t. Y)

Proof………[it relies on the *method of types*]



The DT$_{ks}$ game: equilibrium point

Optimum strategy for A

Given that D will play the dominant strategy, A must solve a minimization problem

$$g^*(y^n) = \arg \min_{z^n : d(z^n, y^n) \leq nL} \mathcal{D}(P_{z^n} || P_X)$$

Theorem (equilibrium point): the profile $(\Lambda^{n,*}, g^*)$ is the only **rationalizable equilibrium** of the game

The DT$_{ks}$ game: who wins?

Theorem (asymptotic payoff at the equilibrium)

Given P$_X$, λ and L, it is possible to define a region Γ for which we have:

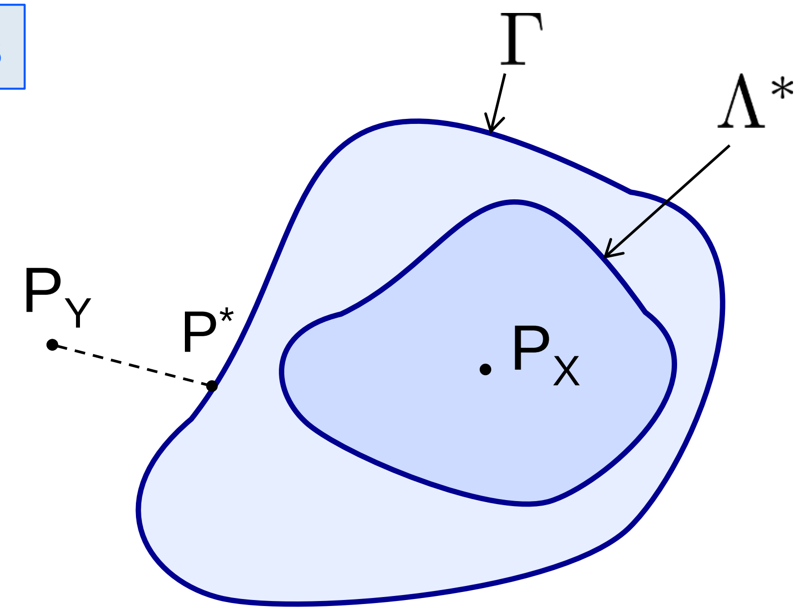
$$\begin{cases} P_Y \in \Gamma, & \text{then } P_{FN} \to 1 \\ P_Y \notin \Gamma, & \text{then } P_{FN} \to 0 \end{cases}$$

A wins

D wins

In the latter case we have:

$$\varepsilon = \min_{P \in \Gamma} \mathcal{D}(P || P_Y)$$



Proof: [it relies on a generalized Sanov's Theorem]….

The DT_{ks} game: who wins?

Theorem (asymptotic payoff at the equilibrium)

Given P_X, λ and L, it is possible to define a region Γ for which we have:

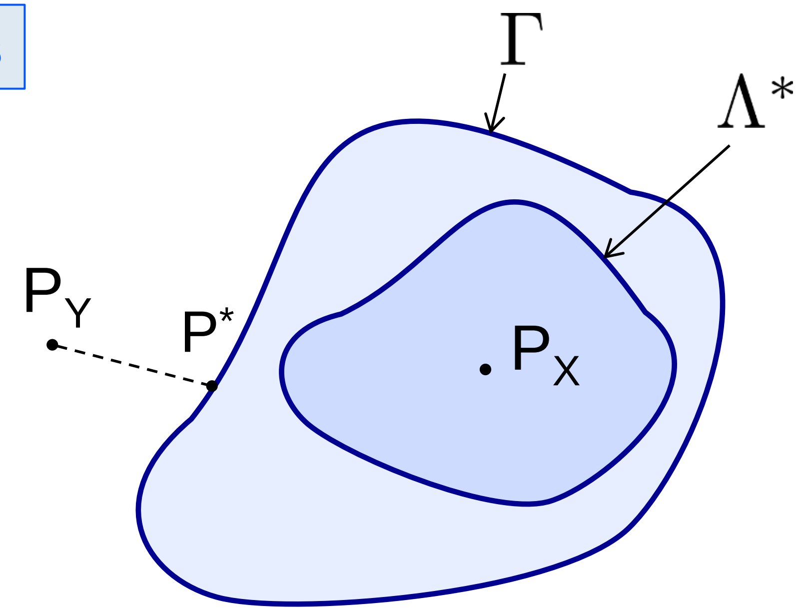
A wins

$$\begin{cases} P_Y \in \Gamma, & \text{then } P_{FN} \to 1 \\ P_Y \notin \Gamma, & \text{then } P_{FN} \to 0 \end{cases}$$

D wins

In the latter case we have:

$$\varepsilon = \min_{P \in \Gamma} \mathcal{D}(P||P_Y)$$



Γ -> ***indistinguishability region* of the test**

(set of the pmf's P that cannot be distinguished from P_X)

Ultimate achievable performance

- Drawback of the N-P setup -> asymmetric role of the error probabilities (λ is fixed)
- Case: λ -> 0 (Resembling Stein's lemma)
  - Best achievable performance for D
  - indistinguishability from P_X for a certain distortion level L

Ultimate achievable performance

Theorem (best achievable performance)

Given P_X and L, we can define

$$\Gamma(P_X, L, \lambda = 0) = \Gamma^{ind}$$

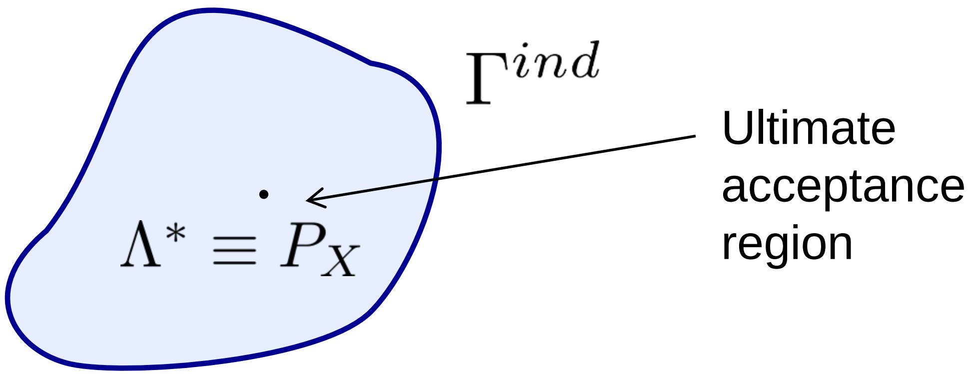
Ultimate (smaller) indistinguishability region

such that

$$\text{if } P_Y \in \Gamma^{ind},$$

$$P_{FN} \to 1, \quad \forall\lambda$$

A surely wins



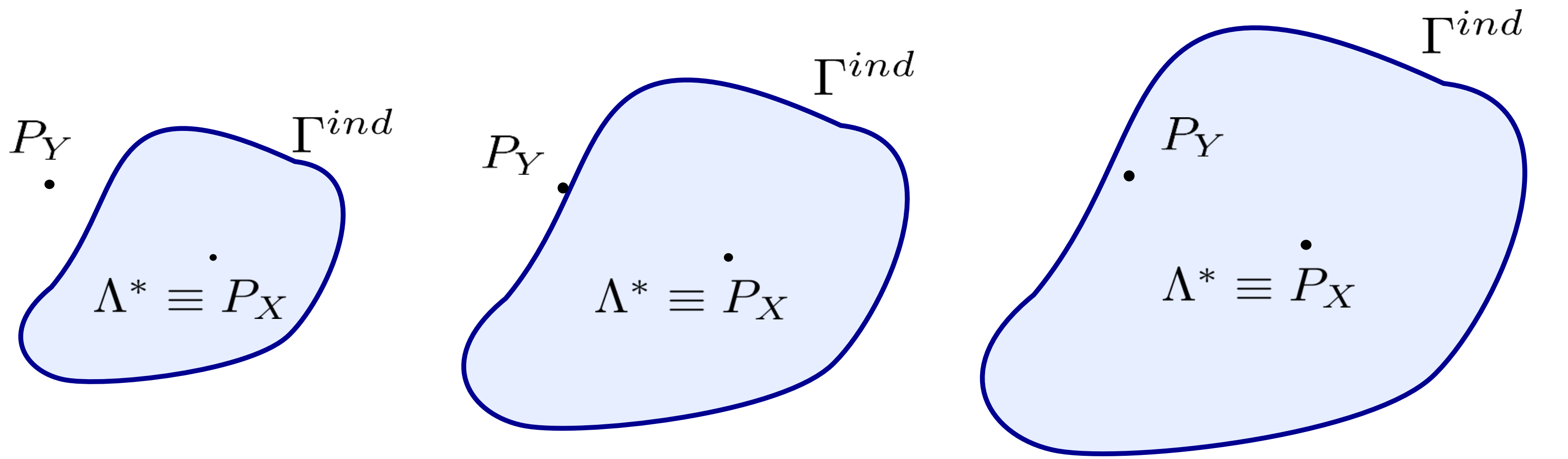
Proof: [it resembles the proof of Stein's Lemma]………

The Security Margin* (in the DT_{ks} setup)

Given Px and Py…..

L_{max} = maximum value of L for which P_X and P_Y can be distinguished

$\mathcal{SM}(P_Y, P_X)$ = L_{max} is the *Security Margin between P_X and P_Y*



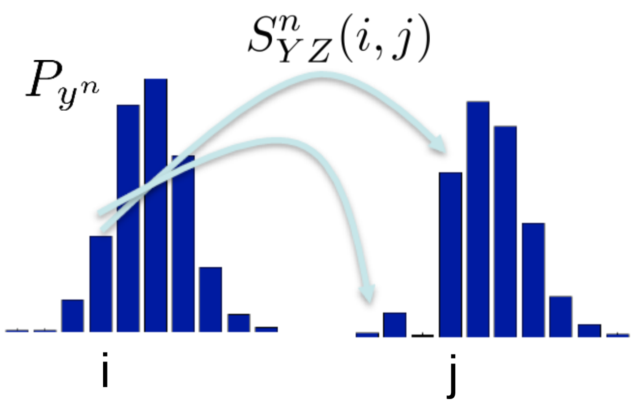
$L < \mathcal{SM}(P_Y, P_X)$ $\quad$ $L = \mathcal{SM}(P_Y, P_X)$ $\quad$ $L > \mathcal{SM}(P_Y, P_X)$

*M. Barni, B, Tondi, "Source Distinguishability Under Distortion-Limited Attack: An Optimal Transport Perspective", *IEEE Trans. on Information Forensics and Security*, Vol. 11, No.10, May 2016

SM and Optimal Transport

<u>Reformulation of the attack</u>

- Attack to the sequence y^n -> application of a *transportation map*



$$S^n_{YZ} = \{S_{YZ}(i,j), i,j \in \mathcal{X}\}$$

$$S^n_{YZ}(i,j) = n(i,j)/n$$

number of times symbol *i* in y^n is transformed into *j*

- E.g. additive distortion

per-letter distortion

$$d(y^n, z^n) = \sum_{i,j} n(i,j) d(i,j) \qquad \left(\frac{d(y^n, z^n)}{n} = \sum_{i,j} S^n_{YZ}(i,j) d(i,j) \right)$$

The distortion constraint defines the *admissible maps.*

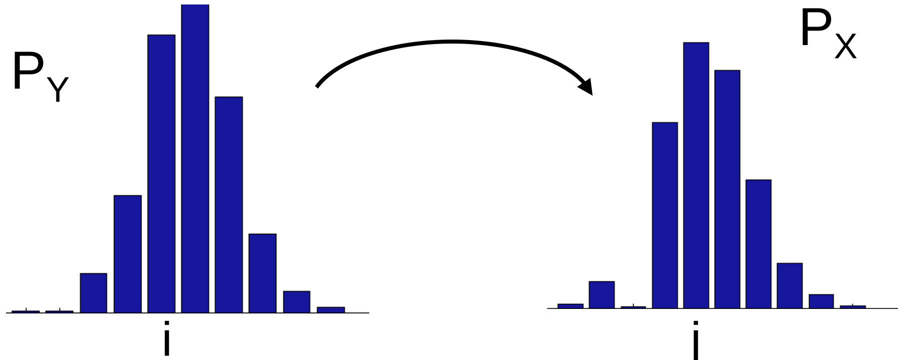
SM and Optimal Transport

Let us interpret P_Y and P_X as two different ways of piling up a certain amount of soil

Let d(i,j) be the cost of moving a unitary amount of soil from the i-th to the j-th bin

OT is concerned with finding the map which moves P_Y to P_X by *minimizing the cost of transportation*

The **Earth Mover Distance (EMD)** is the *minimum cost* necessary to transform P_Y into P_X



SM and Optimal Transport

Corollary (Security Margin in the DT_{ks} setup)

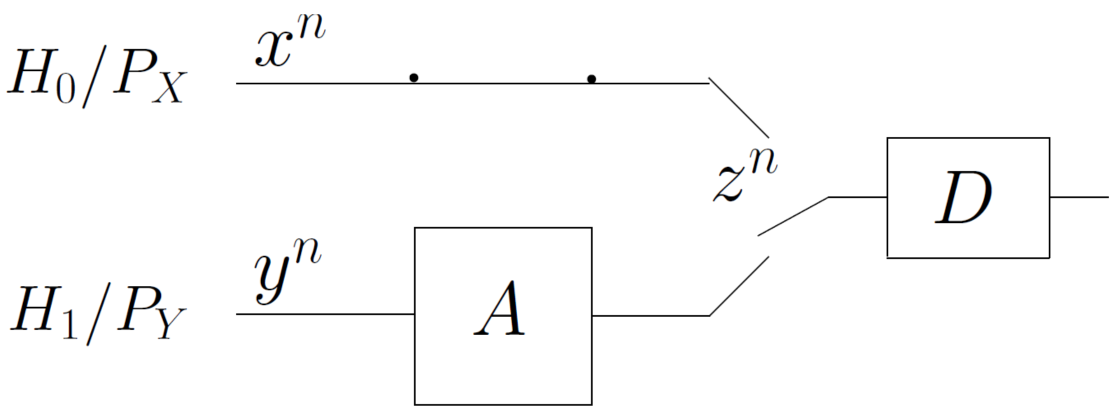
$$\mathcal{SM}(P_Y, P_X) = EMD(P_Y, P_X)$$

Remarks [on the Security Margin]:

- Characterize the *distinguishability* of sources under adversarial conditions
- Summarize the outcome of the game
- Has an efficient computation

Detection games with training data (DT_{tr})

Detection games with training data (DT_{tr})*

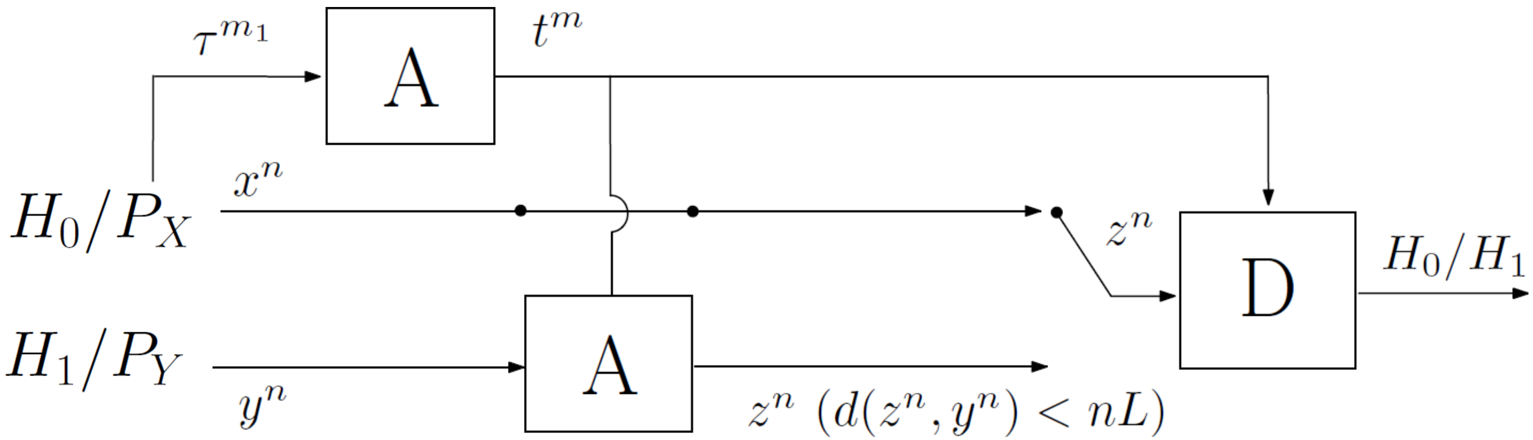


- P_X is *not* known to A and D
- D and A know *training sequences* t_D^N *and* t_A^K *generated by* P_X
- Versions: equal training sequences, independent training with N=K or N > K)
- Assumption: N (and K) is a function of *n* (interesting case: N = cn, c >0)

*M. Barni, B, Tondi, “Binary Hypothesis Testing Game with Training Data”, *IEEE Trans. on Information Theory*, Vol. 60, No.8, August 2014

Detection games with corrupted training ($DT_{c\text{-}tr}$)

Detection game with corrupted training ($DT_{c\text{-}tr}$)



- P_X is known to D by means of a training sequence of length *m*
- The training sequence observed by D is corrupted by A (α = percentage of corrupted samples).
- Cases:
  - *Addition* of fake samples: $m_1 = (1-\alpha)m$
  - *Replacement* of original samples with fake ones: $m_1 = m$

*M. Barni, B, Tondi, "Adversarial Source Identification Game with Corrupted Training", submitted to *IEEE Trans. on Information Theory*, on January 2017

Detection game with corrupted training ($DT_{c\text{-}tr}$)

- Same steps: definition and resolution the games (equilibrium point, payoff at the equilibrium)
- Source distinguishability:
  - Blinding corruption level α_b: the percentage of corrupted samples for which the two sources P_X and P_Y cannot be distinguished (L=0).
  - Security Margin (function of α) : maximum value of L for which P_X and P_Y can be distinguished for the given α

Applications to Image Forensics

Forensics and…. Counter-Forensics!

- MM Forensics: to retrieve information on the history of multimedia documents

- Goal of Counter-Forensics (C-F): to conceal the traces left by the processing (e.g., acquisition traces, double compression,…)

- Drawback of existing C-F approach: *tailored* to deceive a specific analyst, detectable in turn [….'cat&mouse' loop]
- **When designing a counter-forensic method, it is necessary to *simultaneously* consider the presence of an analyst who anticipate the attacker.**

From theory to practice

- ***Universal C-F attack*: optimum against *any* detector based on first order statistics (= image histogram)**
  - **Universal attack in the pixel domain**

    Application: for countering the detection of **manipulated images** (in the spatial domain):

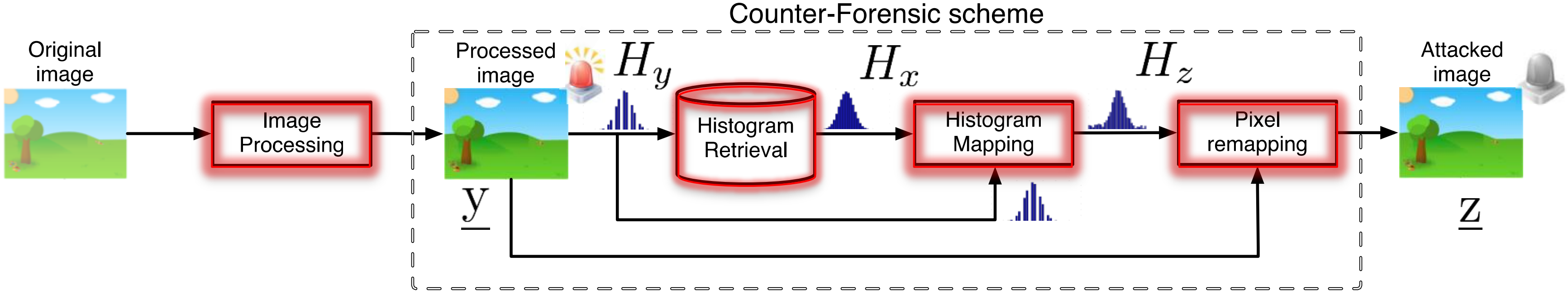
    » *Contrast-ehancement, color-adjustment*

    [Theoretical modeling: DT_{tr} game]
  - **Universal attack in the frequency (DCT) domain**

    Application: for countering the detection of **multiple JPEG compressed images** (*tellltale of manipulation*!)

    [Theoretical modeling: DT game based on multiple observations]

Universal attack in the pixel domain

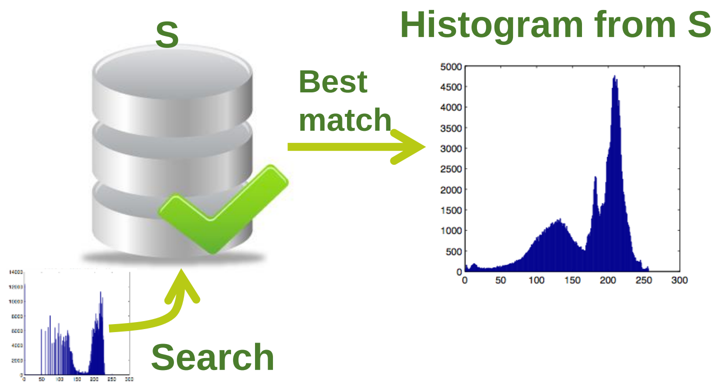


- The A *processes* an image.
- Then:
  - Searches a DB for the closest untouched histogram
  - Computes a transformation map from one histogram to the other subject to a distortion constraint
  - Applies the transformation into the image, minimizing perceptual distortion

*M. Barni, M. Fontani, B. Tondi, "A Universal Attack Against Histogram-Based Image Forensics", International Journal of Digital Crime and Forensics (IJDCF), IGI Global, USA, Vol. 5, no. 3, 2013.

Histogram Retrieval phase

- Given H_y, the A searches for the nearest target histogram H_x^* in a database S of untouched histograms



- The search is carried out by performing

h is the *optimum test function* from the theory (DT_{tr} game)

$$\min_{H_x \in S} h(\nu_x, \nu_y)$$

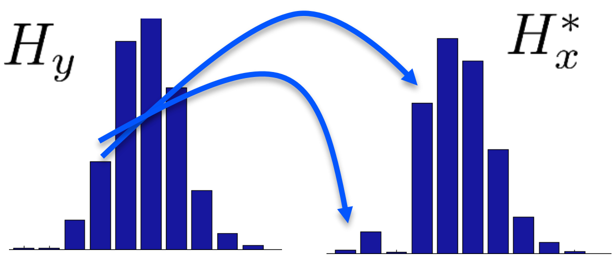
generalized K-L divergence

Normalized histograms H_x and H_y

Histogram Mapping phase

- Given H_x^* , the A has to find the best transportation map from H_y namely $N^* = \{n^*(i,j)\}_{i,j=1,\ldots,255}$
- Distortion constraint?
- …on the absolute pixel distortion (maximum distance)

$$\max_i |\underline{y}(i) - \underline{z}(i)| \leq L$$



- Then, A has to solve

$$\min_{\nu_z} h(\nu_z, \nu_x^*)$$

$$\begin{cases} \sum_j n(i,j) = n\nu_y(i) \; \forall i \\ n(i,j) = 0, \;\; \forall (i,j) \in \mathcal{I} \times \mathcal{I} : |i-j| > L \\ n(i,j) \geq 0 \quad \forall i,j \\ n(i,j) \; \in \mathbb{N}. \end{cases}$$

Convex MINLP

[Complexity: $\sim 2L|\mathcal{I}|$]

$|\mathcal{I}| = 255$

Pixel remapping phase

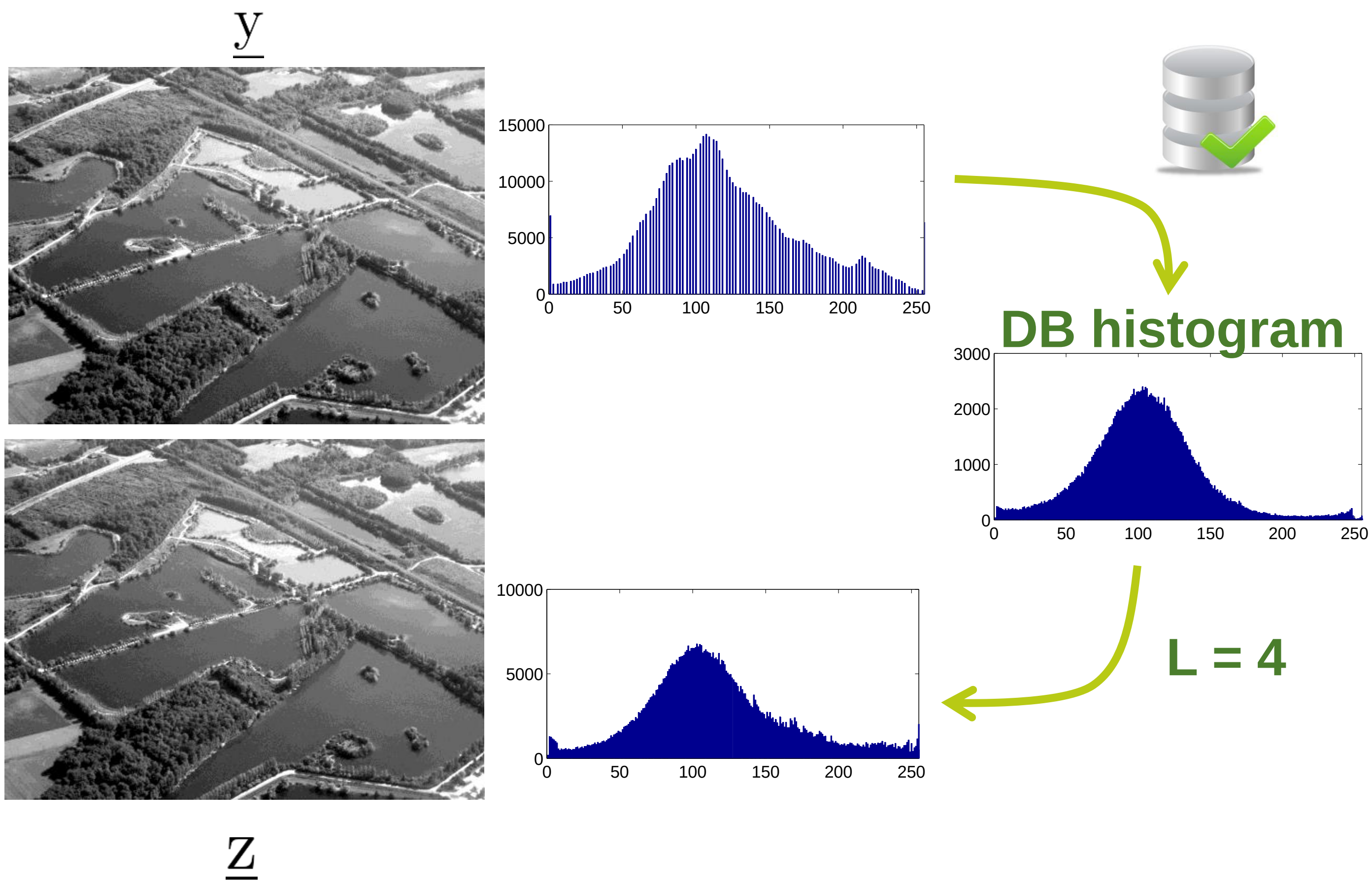
- Having H_z, the A modifies the image to produce the attacked image $\underline{z}$
- The mapping implementation exploits the peculiarity of the Human Visual Sysemt (HVS)

$$N^* = \{n^*(i,j)\}_{i,j=1,\dots,255}$$

- Note: this phase does not have impact on the results of the forensic analysis

Application: contrast enhancement

- An example:



UNIVERSITÀ DI SIENA

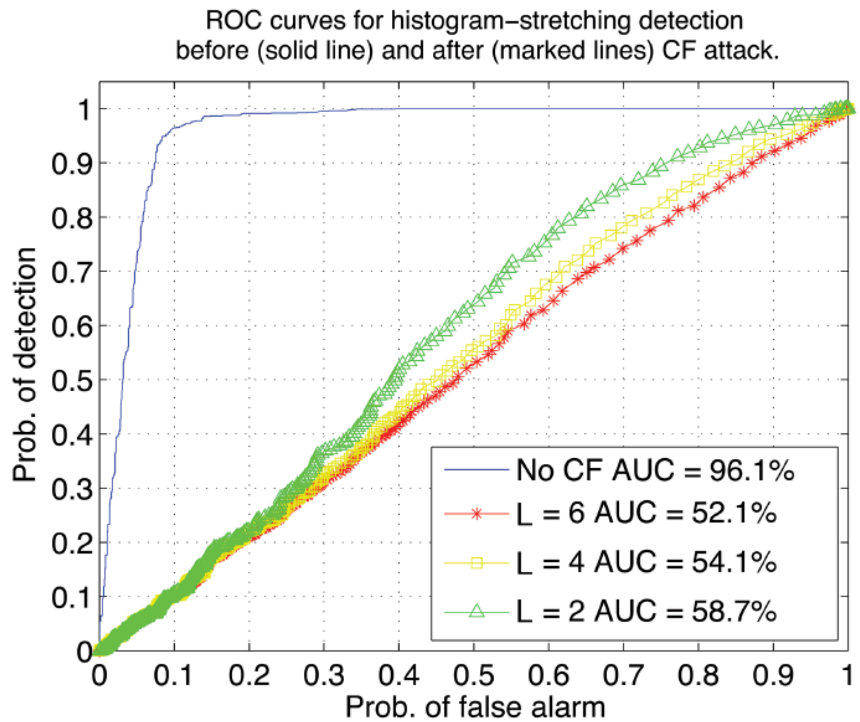
Experimental results: contrast ehancement

Setup:

- DB of 25000 images (MIRFLICKR)
- Test on 1338 images (UCID)

Attack:

- L = 2,4,6



(a)

Detector:

Matthew C. Stamm and K. J. Ray Liu. Blind forensics of contrast enhancement in digital images. In *Proc. of ICIP 2008, IEEE Int. Conference on Image Processing*, pages 3112–3115, 2008.

	PSNR			**SSIM**			**AUC**
L	mean	min	95th perc	mean	min	95th perc	
2	44.8	43.3	45.6	0.994	0.977	0.998	0.587
4	39.2	37.3	40.4	0.981	0.938	0.993	0.541
6	36.1	34.1	37.6	0.964	0.908	0.989	0.521

Conclusions

Summing up:

- Theoretical framework for the study of various versions of the *binary detection problem in the presence of adversary* and applications to problems of MM-Forensics

Future (on-going) work:

- Extension to
  - higher-order statistics (adversary-aware data driven classification)
  - sources with memory
  - continuous sources
- Multiple-hypothesis testing or classification
- Application of the universal attack to other fields (not only MM-F)

Thank you for the attention