

# Adversarial Signal Processing

**Fernando Pérez-González**

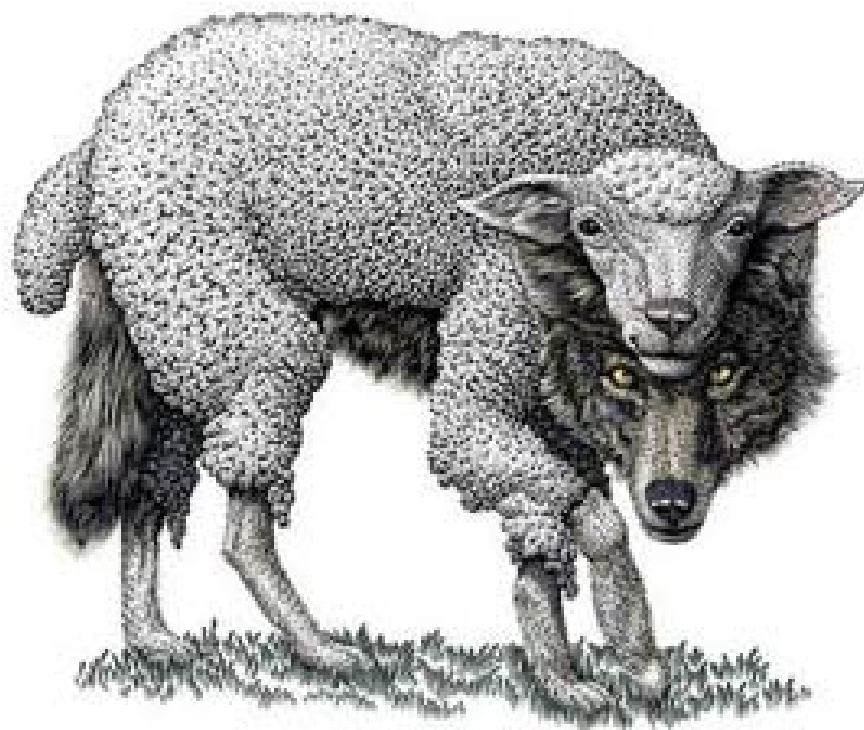
Signal Theory and Communications Department

University of Vigo - Spain

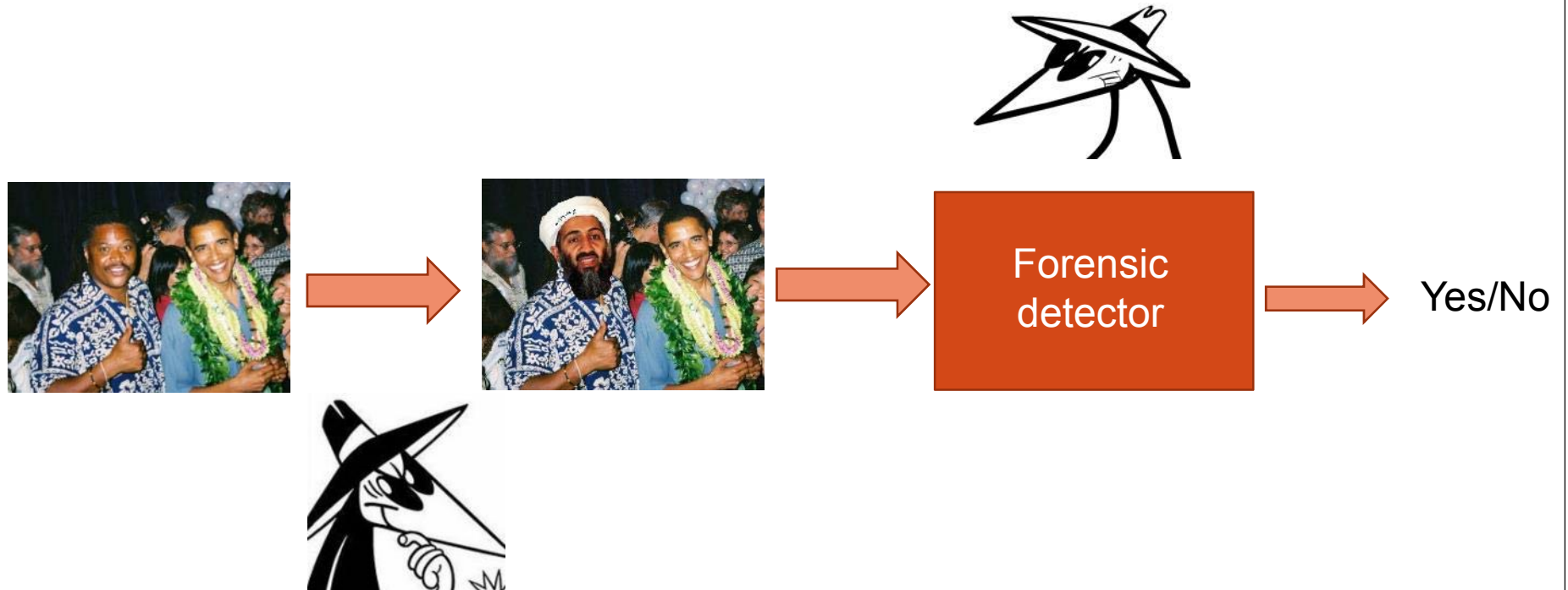
# Signal Processing's Dream



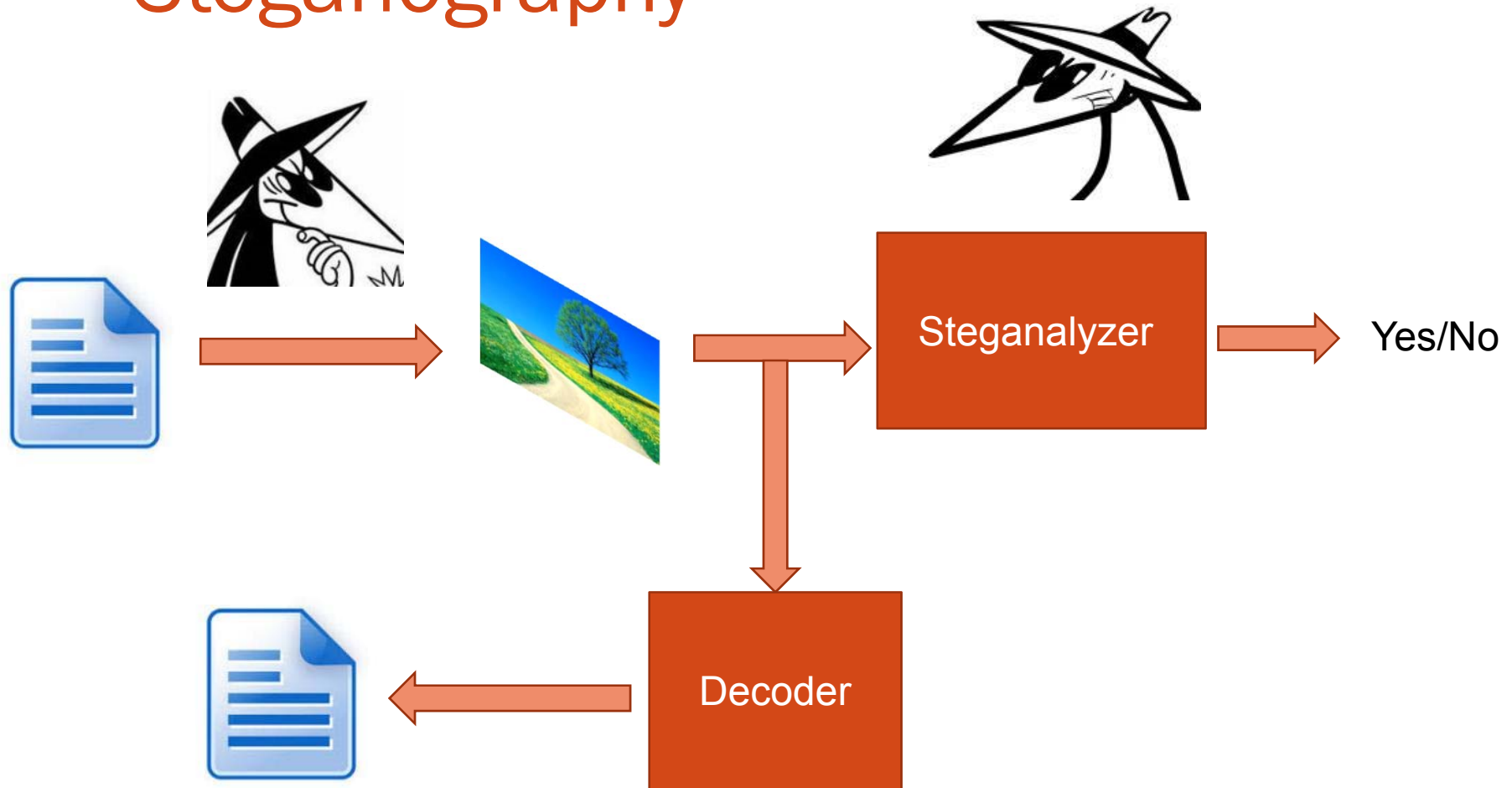
# Signal Processing's Nightmare



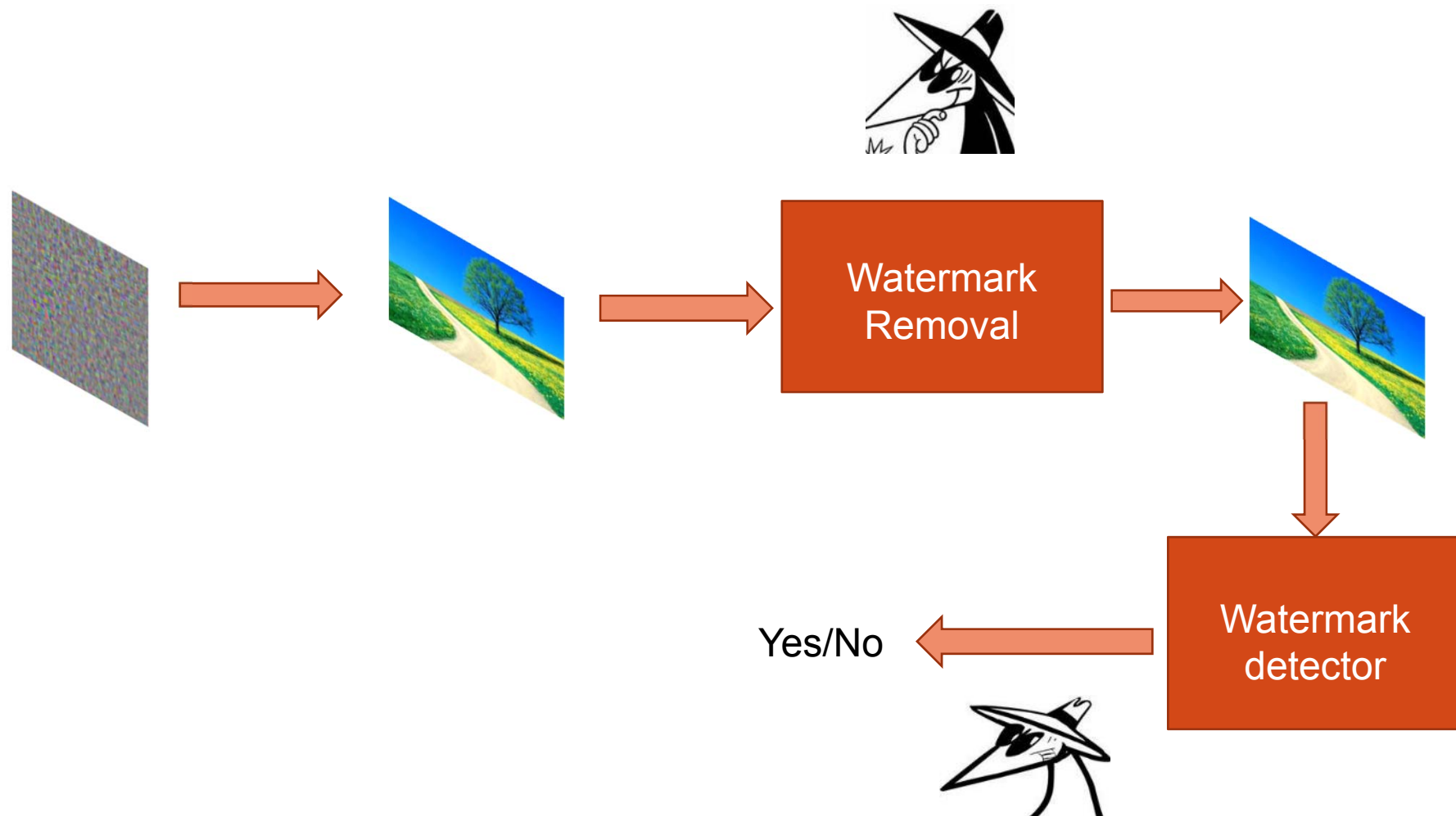
# Multimedia Forensics



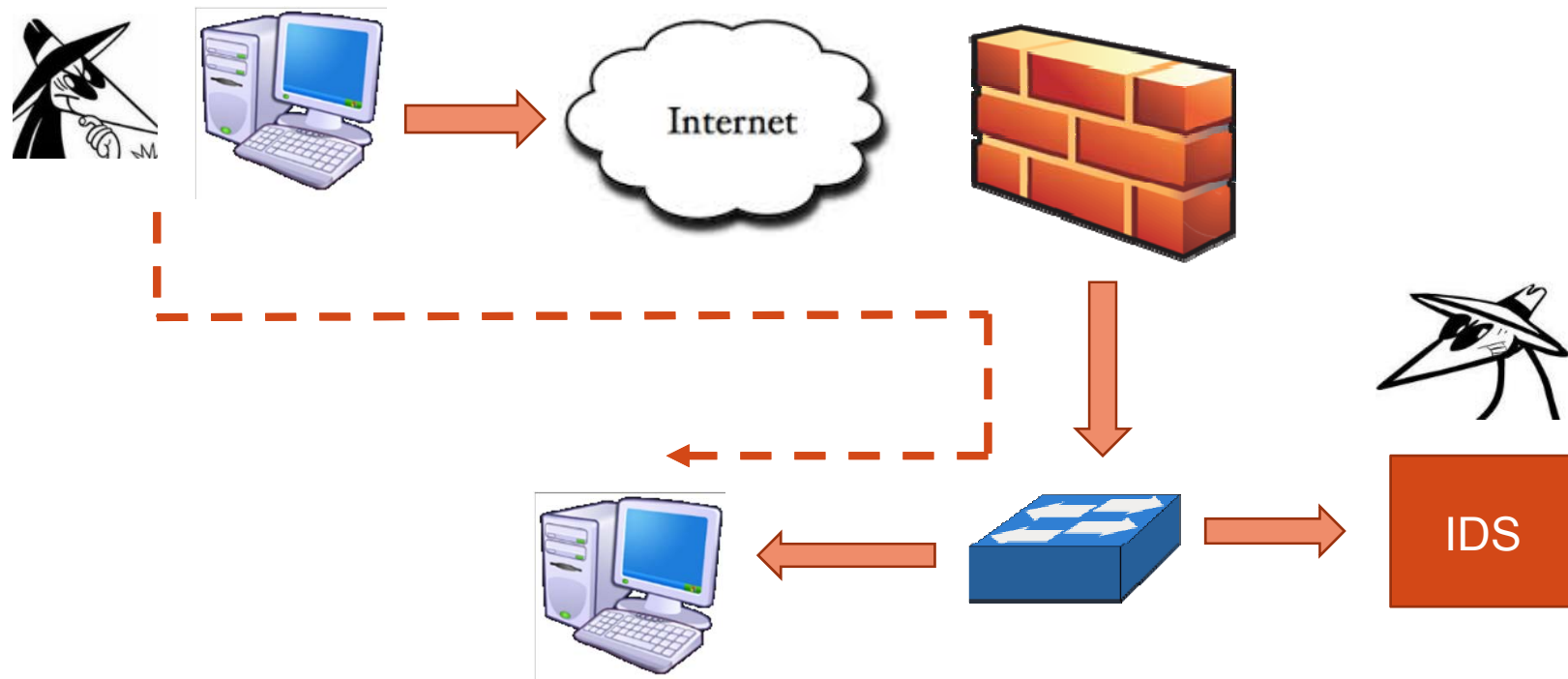
# Steganography



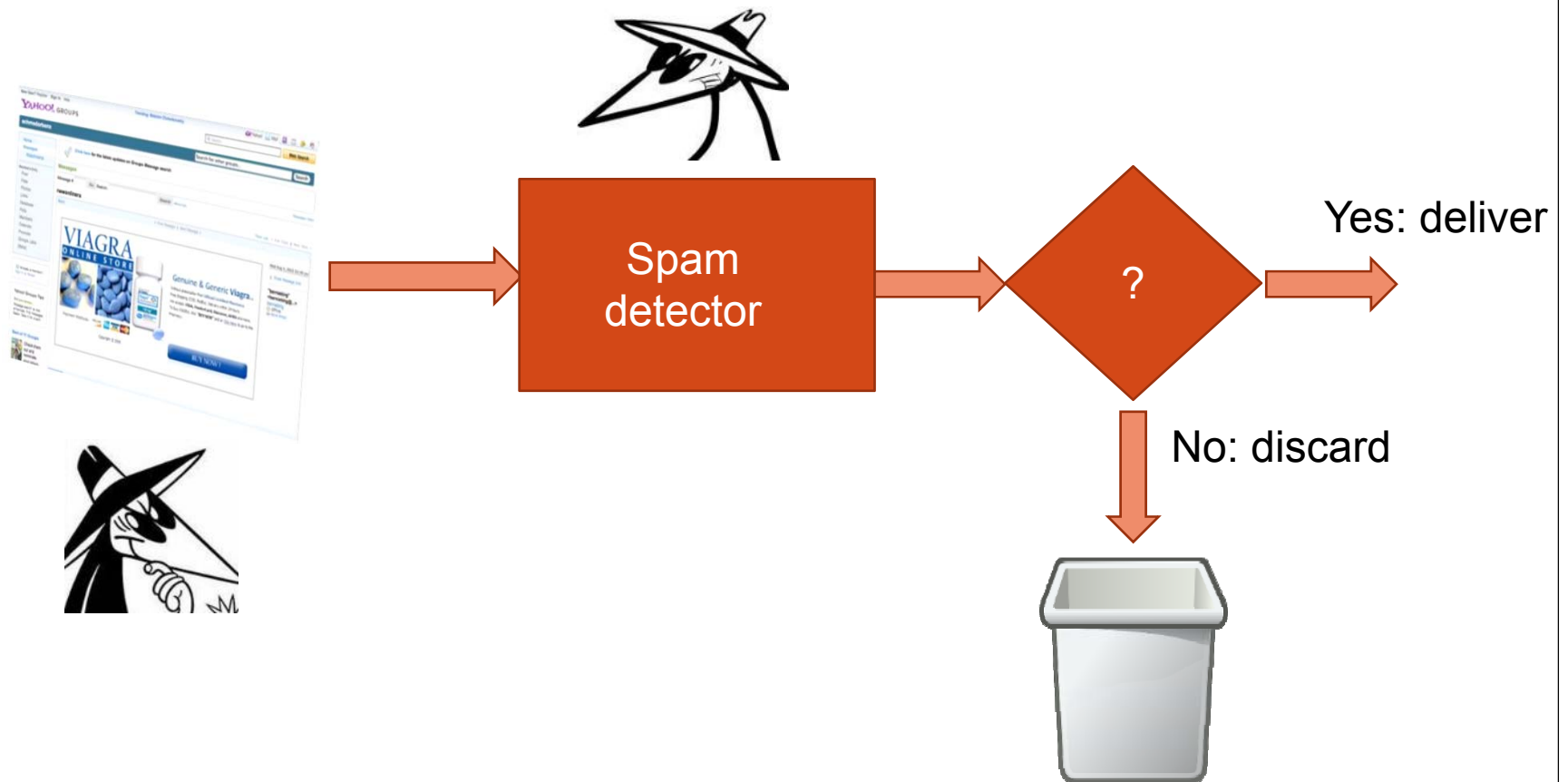
# Watermarking



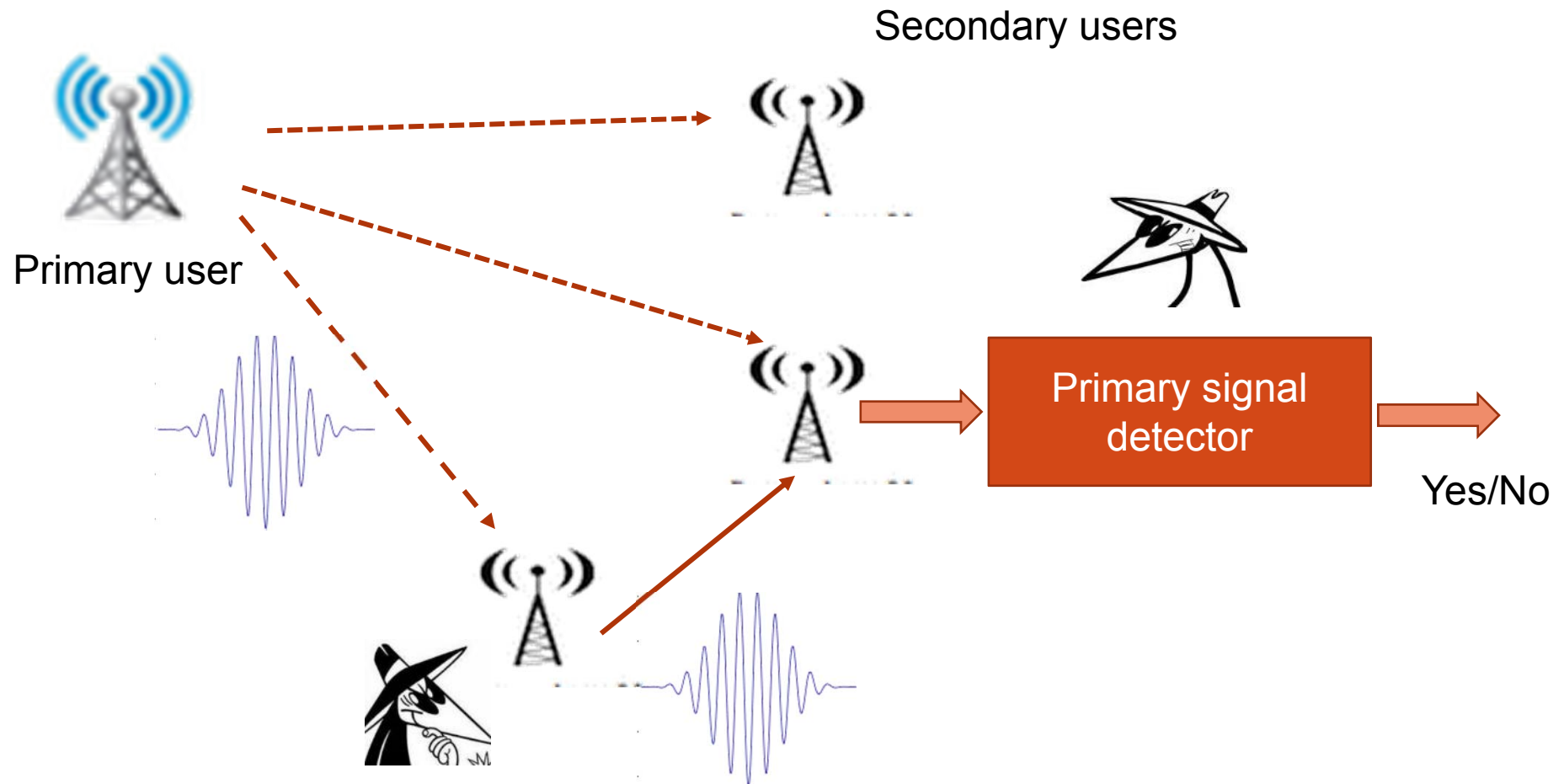
# Intrusion detection



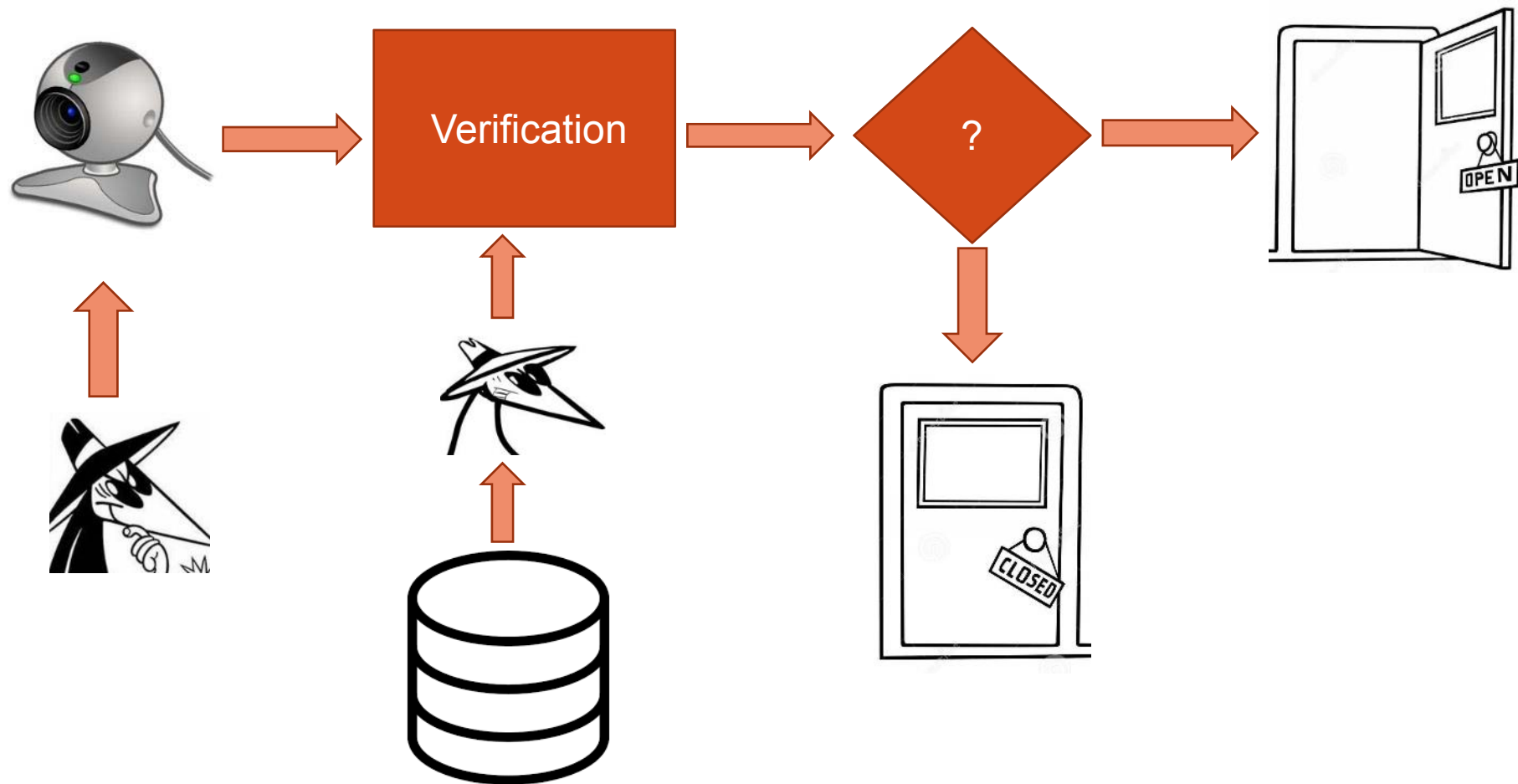
# Anti-spam filtering



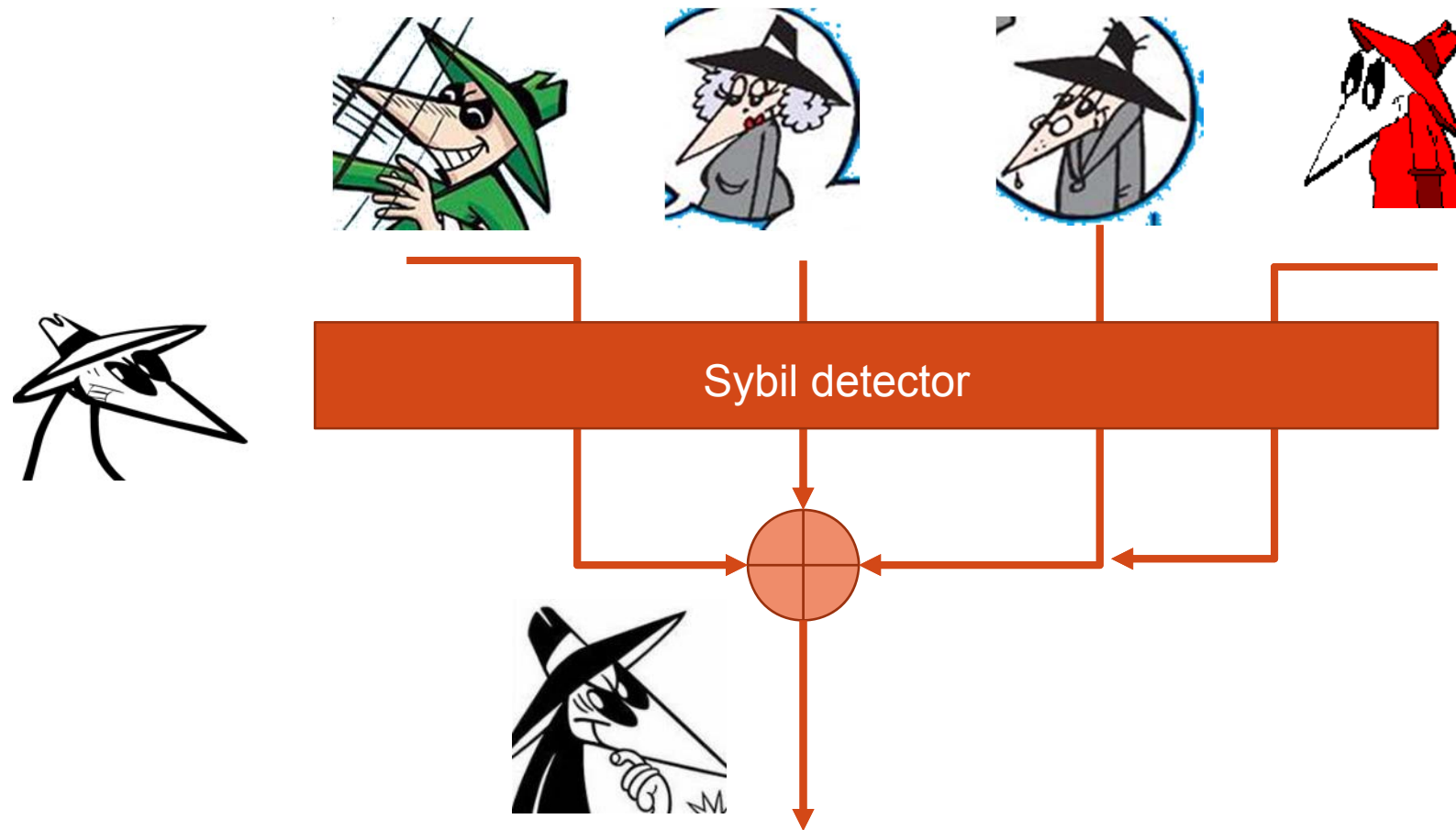
# Cognitive radio



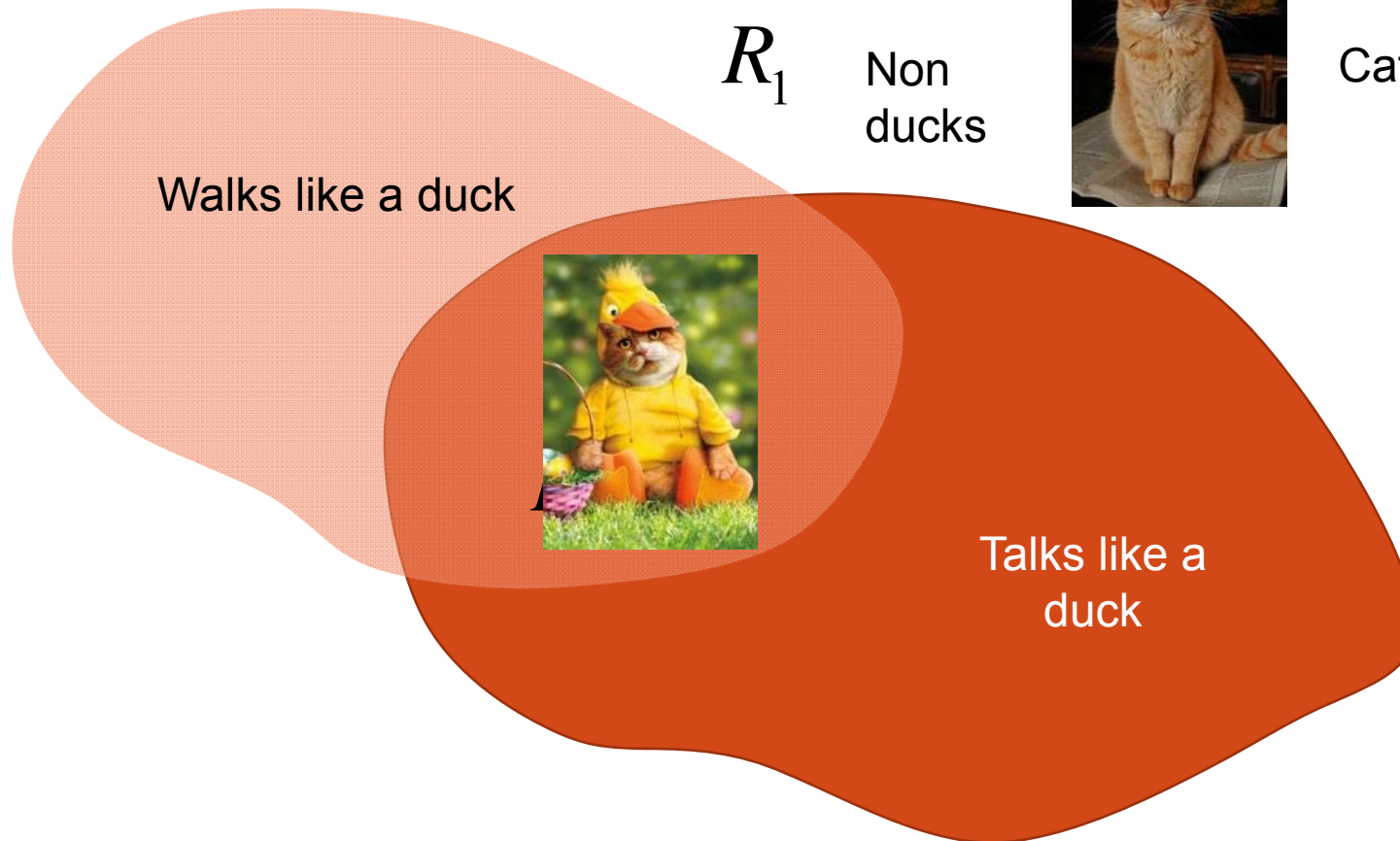
# Biometric identification/verification



# Reputation systems



# It walks like a duck, it talks like a duck...



It walks like a duck,  
it talks like a duck...



# Metrics

- False Positive Rate (FPR) or False Alarm Rate (FAR)

$$\int_{R_1} f(\mathbf{y} | H_0) d\mathbf{y}$$

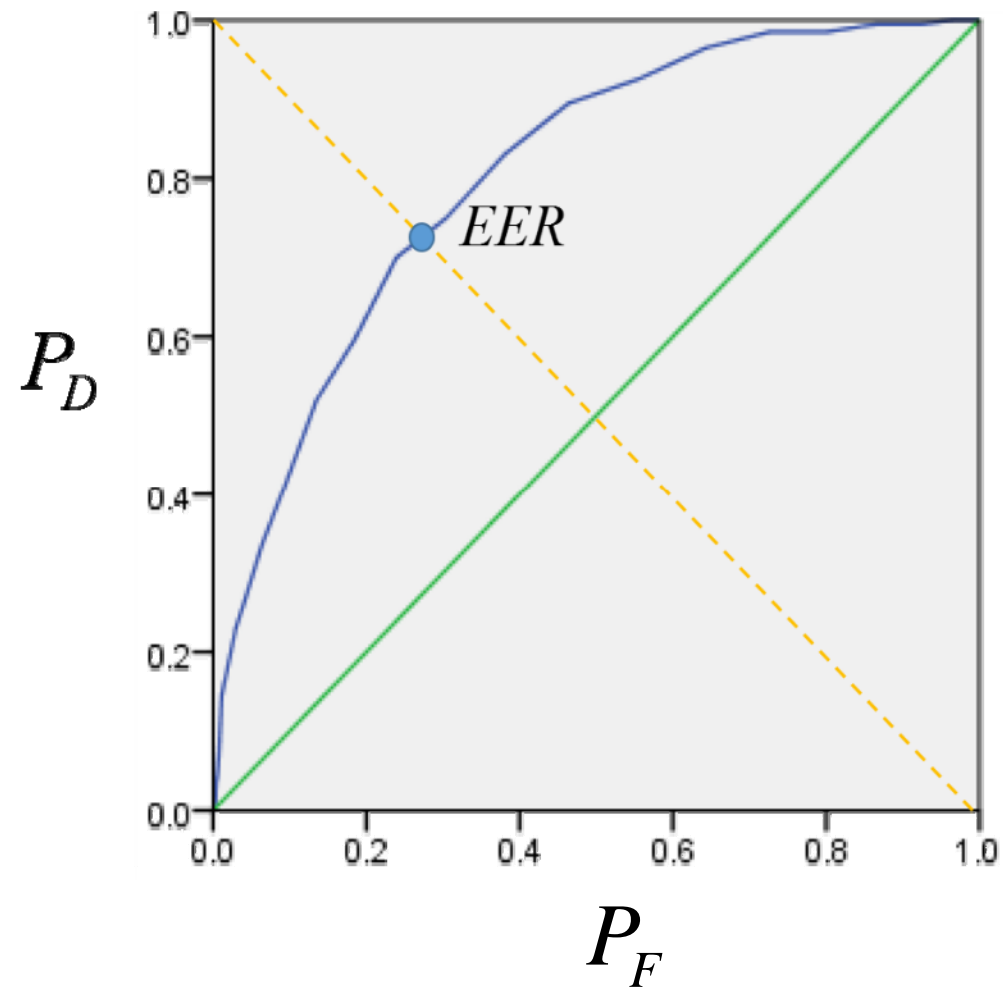
- False Negative Rate (FNR) or miss detection

$$\int_{R_0} f(\mathbf{y} | H_1) d\mathbf{y}$$

- True Positive Rate (TPR): 1-FNR [a.k.a. sensitivity or recall rate]
- True Negative Rate (TNR): 1-FPR [a.k.a. specificity]

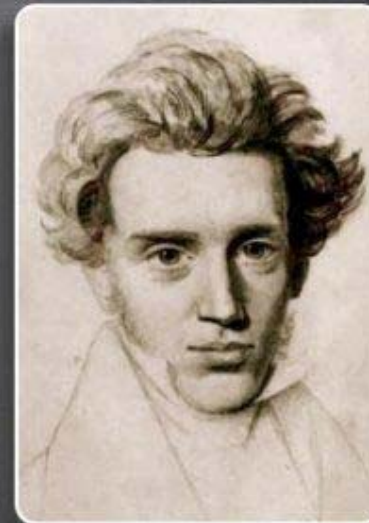
# Metrics

ROC



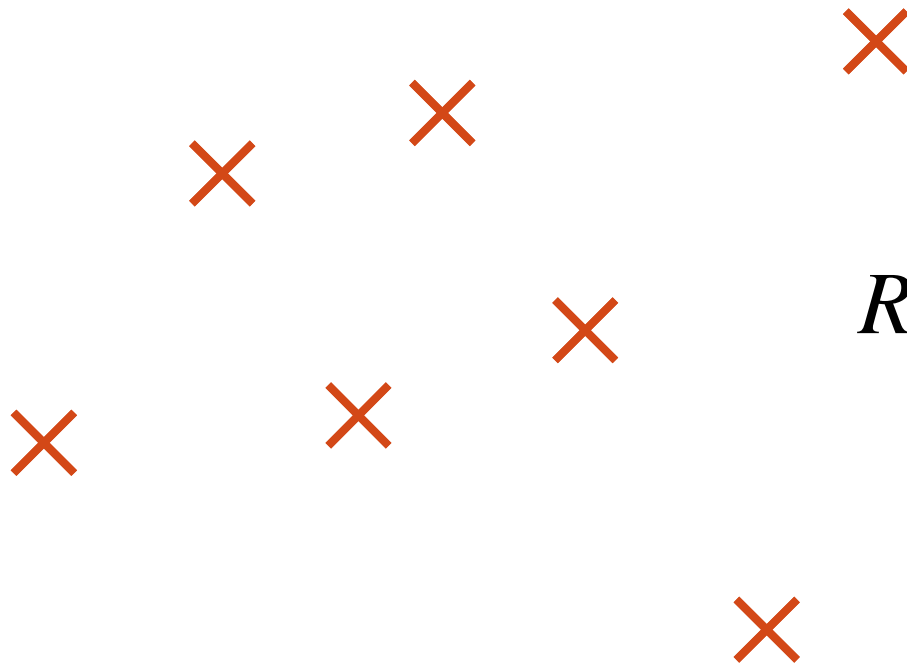
# Metrics

“THERE ARE TWO  
WAYS TO BE FOOLED.  
ONE IS TO BELIEVE  
WHAT ISN'T TRUE;  
THE OTHER IS TO  
REFUSE TO ACCEPT  
WHAT IS TRUE.”



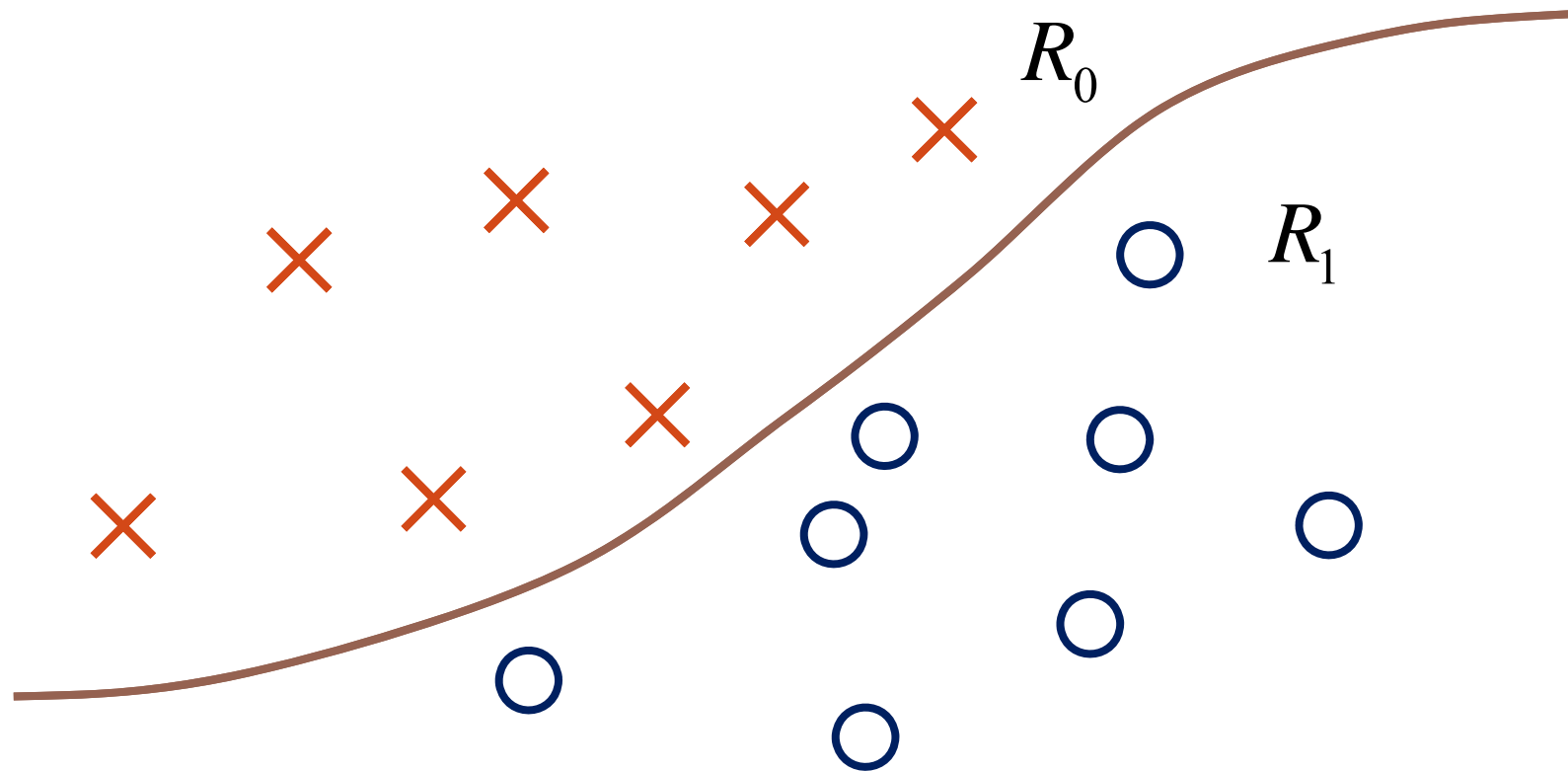
- SOREN KIERKEGAARD

# Signature-based

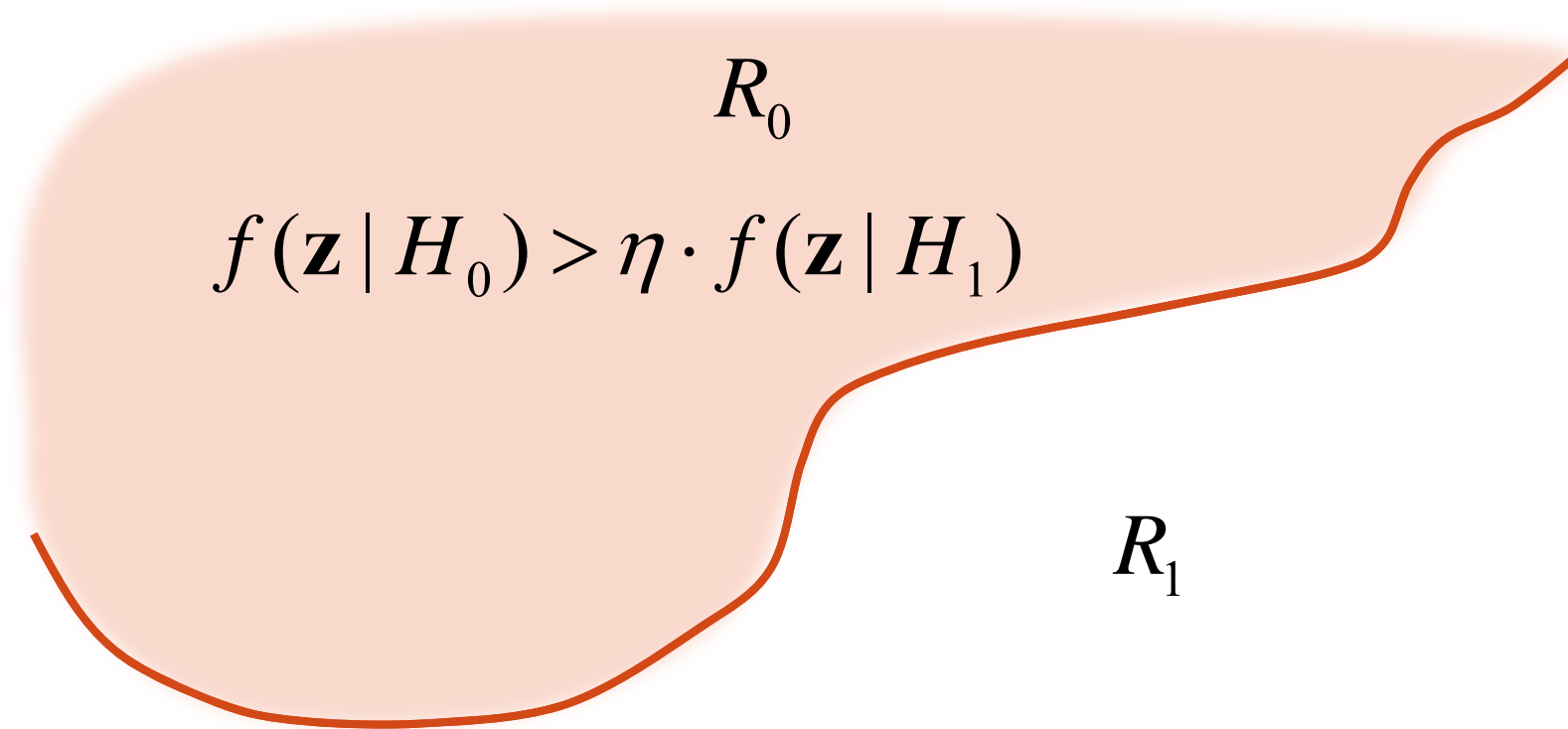


$$R_0 = \bigcup \mathbf{x}_i$$

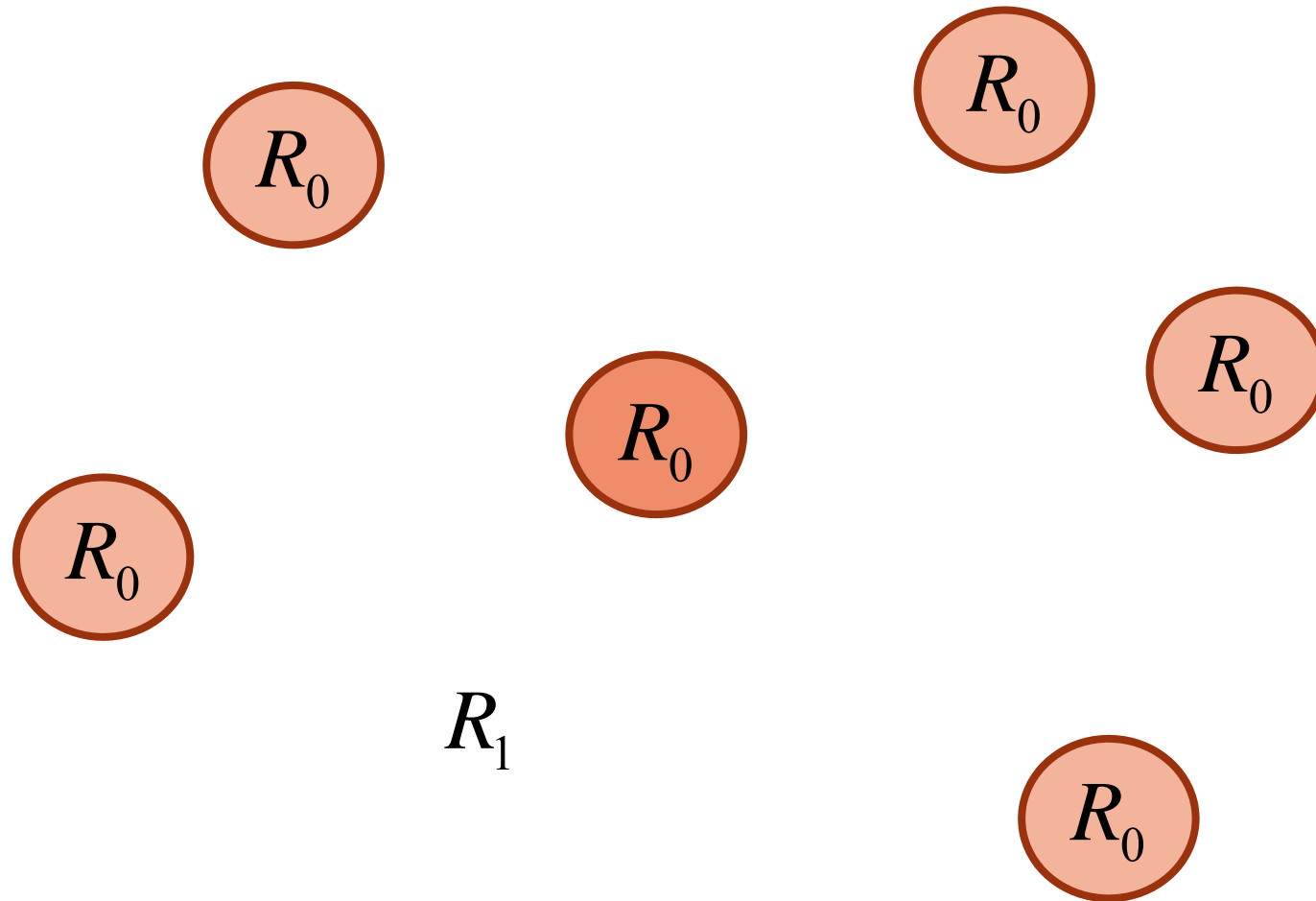
# Supervised learning based



# Neyman-Pearson based

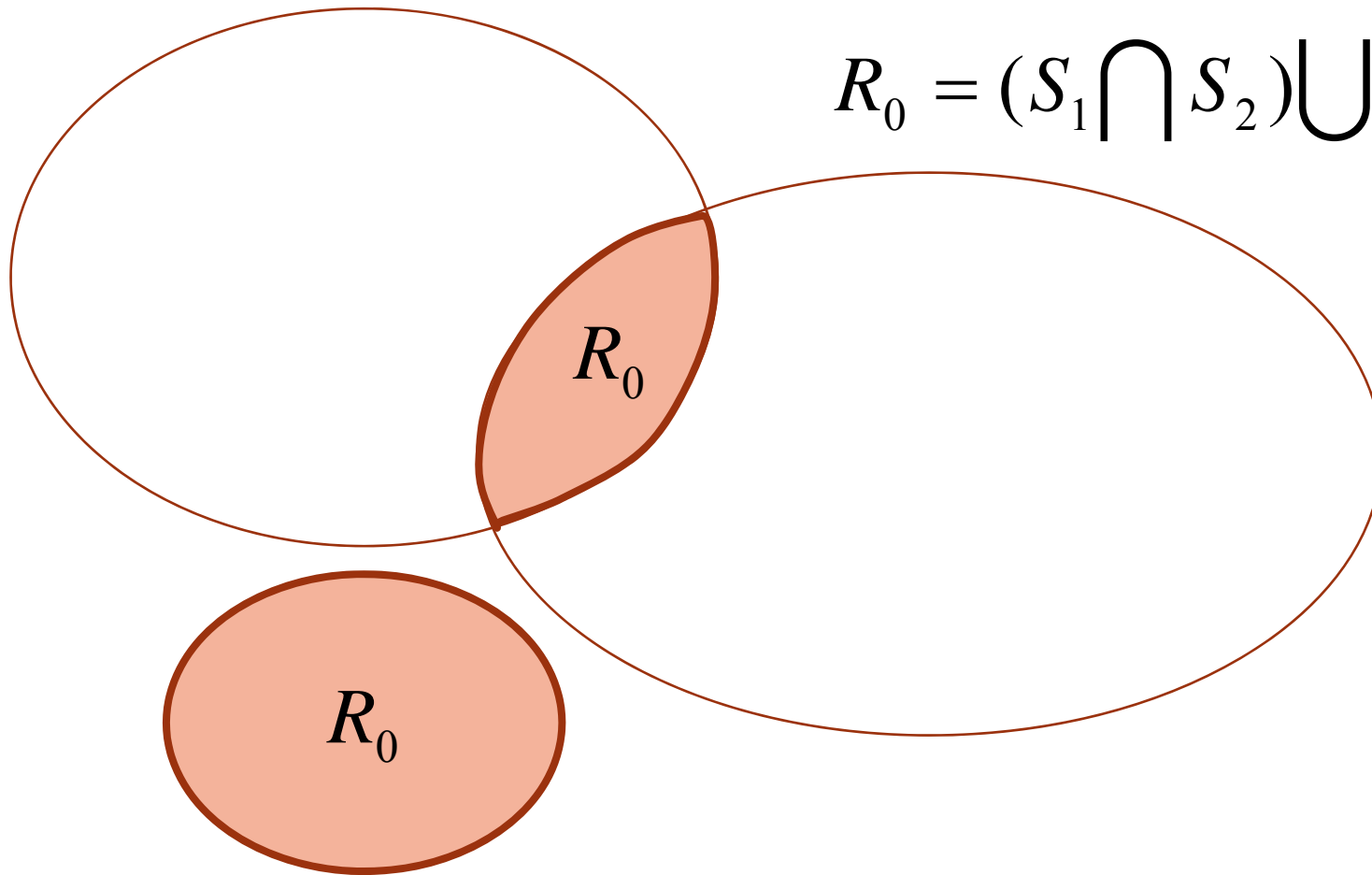


# Multiple/binary hypothesis based

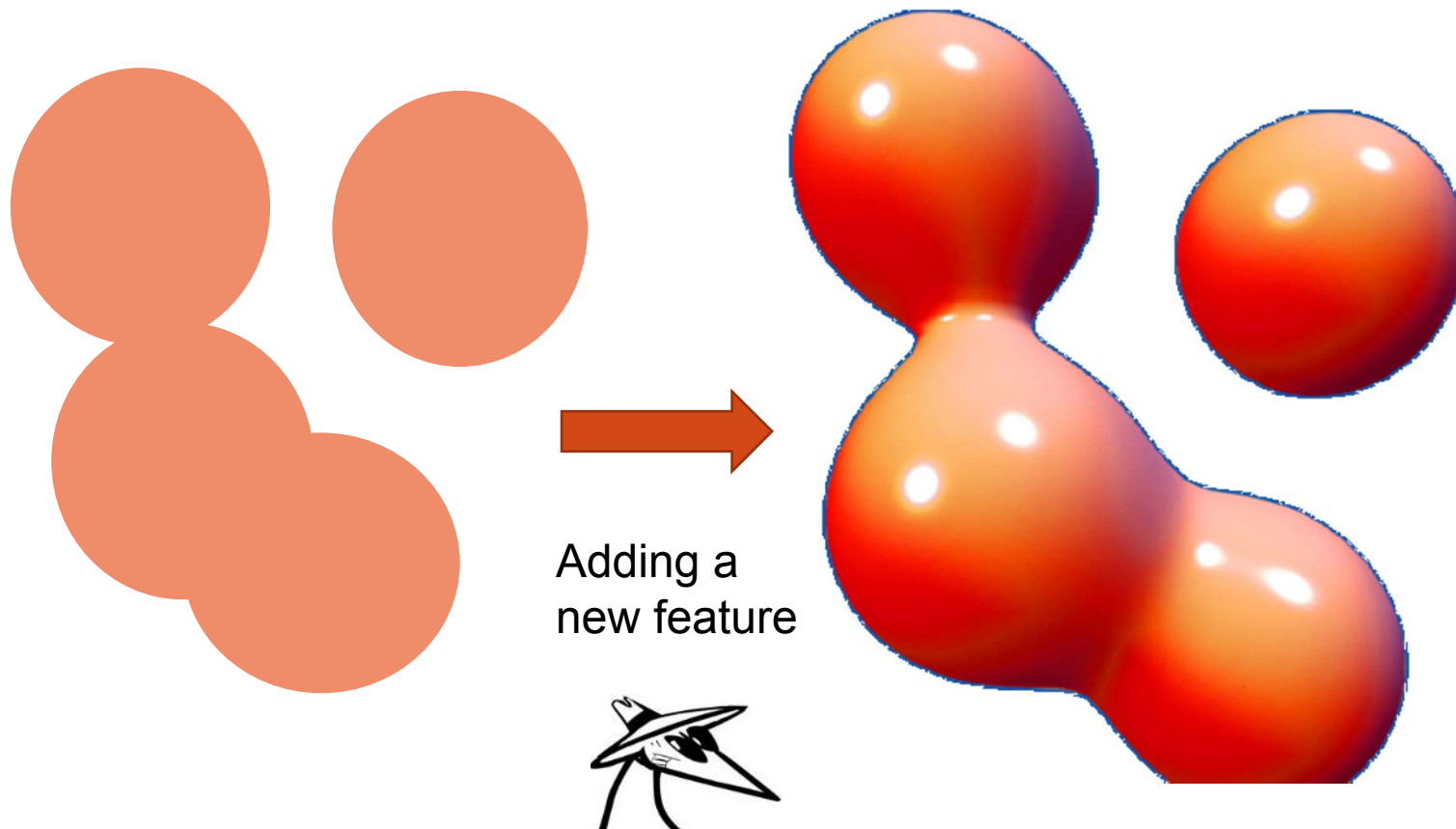


## Rule/property based

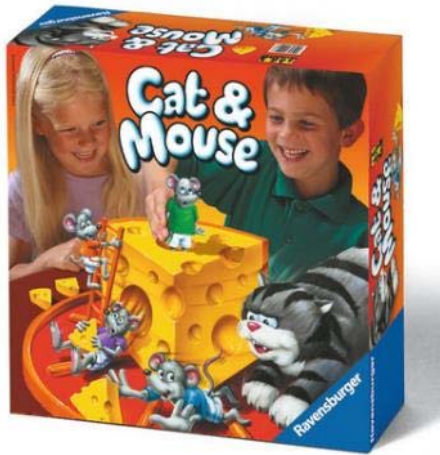
$$R_0 = (S_1 \cap S_2) \cup S_3$$



# Increasing the dimensionality



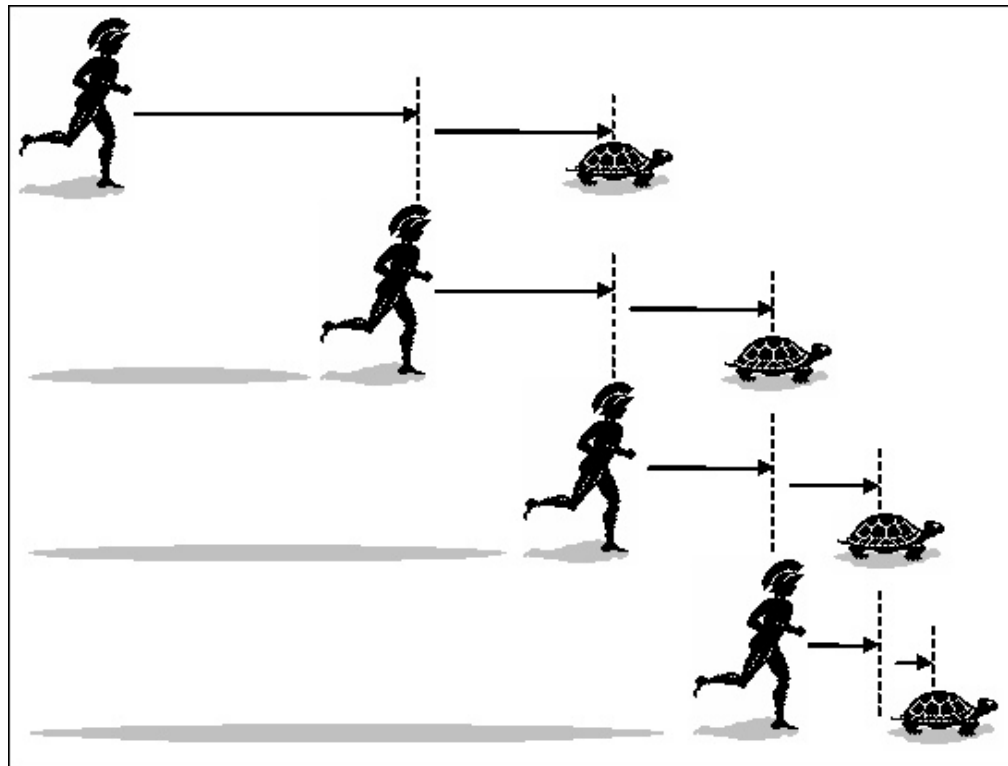
# 20 years of research in steganography



- Cat and mouse game between defender and adversary.
- Methods more complicated ever: more features, higher-order dependences...
- Optimality not guaranteed in any sense:
  1. No Nash equilibria proven.
  2. Possible evolution due to specific paths taken => suboptimality even if convergence.

# 20 years of research in steganography

- Improvements by both adversaries occur at an ever slower pace.



# Sundews (*Drosera*)



# Orchids



# Mimicry



# Batesian mimicry



# A game-theoretic approach

## [Barni 13]

- Two-player game
- Set of actions for player 1:  $S_1 = \{s_{11}, s_{12}, \dots, s_{1N}\}$
- Set of actions for player 2:  $S_2 = \{s_{21}, s_{22}, \dots, s_{2M}\}$
- Payoff for player 1:  $u_1(s_{1i}, s_{2j})$
- Payoff for player 2:  $u_2(s_{1i}, s_{2j})$
- Zero-sum game:  $u_1 = -u_2$
- Non-sequential game.

# A game-theoretic approach

- Nash equilibrium: none of the players improves his payoff with a different action (if the other players also stay the same).

$$u_1(s_1^*, s_2^*) \geq u_1(s_{1i}, s_{2j}), \text{ for all } i, j$$

$$u_2(s_1^*, s_2^*) \geq u_2(s_{1i}, s_{2j}), \text{ for all } i, j$$

# The source identification game

## [Barni 13]

- **Two players:** Defender (D) and Adversary (A). Binary hypothesis testing setup.
- **Adversary:** Generates an i.i.d. sequence according to a distribution  $P_Y$  and modifies the samples so they look like produced by Defender (with a distortion constraint).
- **Defender:** Generates an i.i.d. sequence according to a distribution  $P_X$  and constructs a detector that bounds the probability of false positive. Free to choose the decision region.
- **Payoff:** The probability of false negative (for A), minus this probability (for D). Zero-sum game.

# The source identification game

- There is a (distortion) constraint on the changes that the adversary may do to his sequence.
- Asymptotic version of the game  $n \rightarrow \infty$ : allow any defender region  $R_0$  s.t.  $P_{FP} \leq 2^{-n\lambda}$
- And allow any attacker modification  $\varphi(\mathbf{y}_n)$  to the sequence  $\mathbf{y}_n$  generated according to  $P_Y$

$$d(\varphi(\mathbf{y}_n), \mathbf{y}_n) \leq nD$$

- Both  $P_X$  and  $P_Y$  are known to both players.

# The source identification game

- It turns out that the Nash equilibrium for the game is such that

$$R_0^* = \{\mathbf{x}_n : D(P(\mathbf{x}_n) \| P_X) \leq \lambda - |\mathcal{X}| \frac{\log(n+1)}{n}\}$$

Regardless of what the adversary does or what  $P_Y$  is !!!

- For the adversary, the optimal strategy is

$$\varphi^*(\mathbf{y}_n) = \arg \min_{\mathbf{z}_n : d(\mathbf{z}_n, \mathbf{y}_n) \leq nD} D(P(\mathbf{z}_n) \| \mathbf{x}_n)$$

- In both cases, “closeness” is measured using the Kullback-Leibler distance:

$$D(P \| Q) = \sum_{\mathcal{X}} P(i) \log \frac{P(i)}{Q(i)}$$

# So where do we stand?

- Two main limitations:
- Sources are i.i.d.
- Optimality is shown only in the asymptotic case.
- But it supports the use of the Kullback-Leibler distance as a good strategy for the Defender.
- The Attacker, however, still needs to solve an optimization problem: find the closest sequence (in KLD) satisfying a distortion constraint.

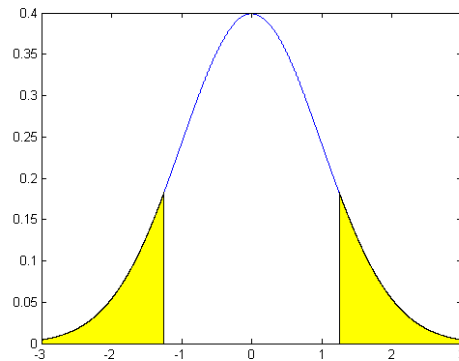
# You better know your adversary

- Kullback-Leiber distance between two (continuous) Gaussians with different means and identical (and known) variance

$$D(P \parallel P_X) = \frac{(\mu - \mu_X)^2}{2\sigma^2}$$

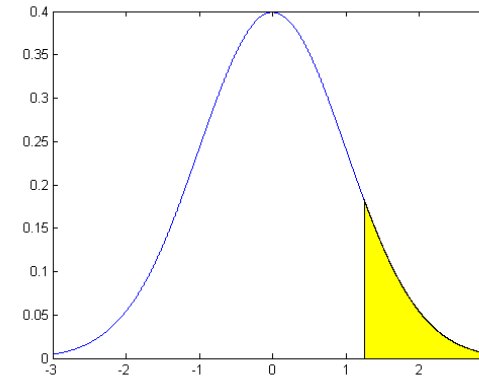
- The asymptotically optimal test imposes a bound on the distance between the means. This implies a symmetric  $R_0^*$  about  $\mu$
- However, if it is known that  $P_Y$  is Gaussian with positive mean  $\mu_Y > \mu$  then  $R_0$  will try to “avoid” positive sequences and will be a non-symmetric region.

# You better know your adversary



$$\mu = 0$$

Two - tailed test

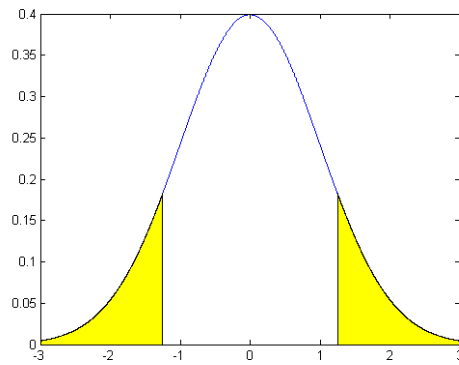


$$\mu = 0$$

One - tailed test

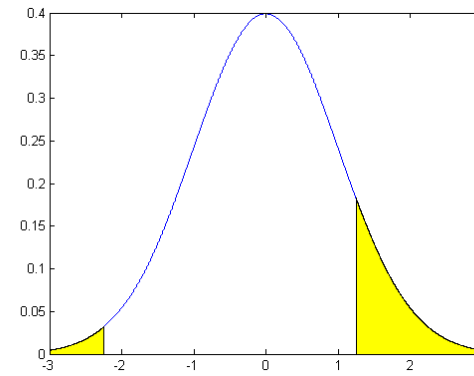
- For approximately the same Probability of false negatives, the probability of false positive is half in the one-tailed test.
- However, asymptotically with  $n$  both tests decrease the probability of false positive at the same (exponential) rate.

# You better know your adversary



$$\mu = 0$$

Two - tailed test



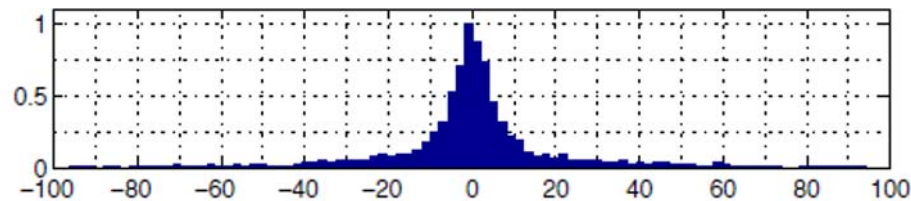
$$\mu = 0$$

Non - symmetric test

- The test based on the KLD guarantees that the region  $R_1^*$  contains any other  $R_1$  with the same asymptotic decrease (as it happens in a non-symmetric test).

# Attacks to histogram-based detectors

- In many “anomaly detection” instances, there is some statistical property that **normal data** satisfy.
- Image proc.: Generalized gaussian distribution in DCT domain



- TCP networks: Exponential interarrival time in non-congested networks.

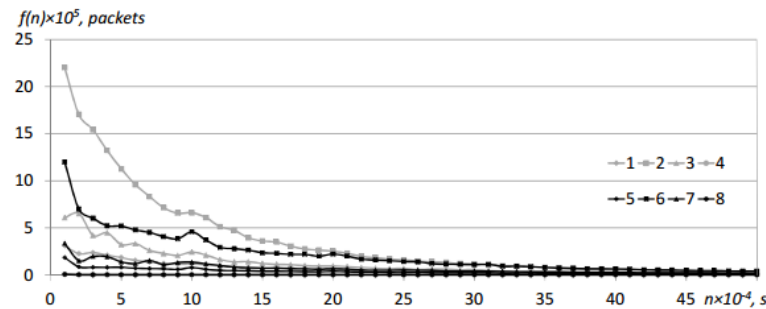
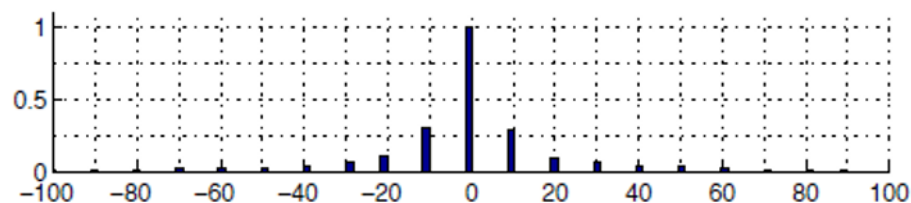


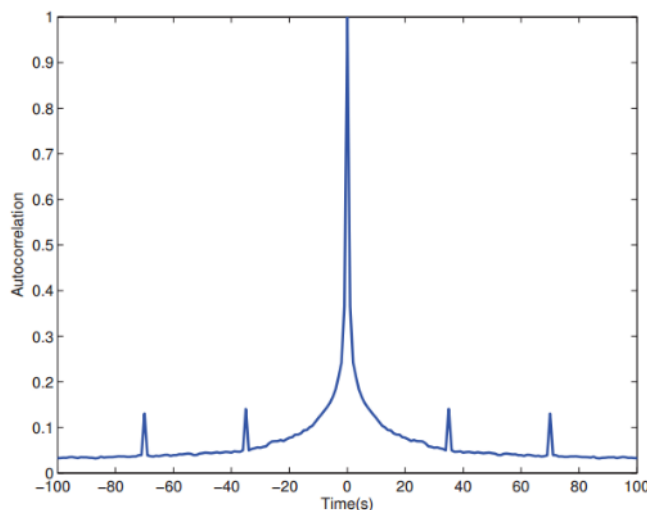
Fig. 2. Distribution of TCP packet inter-arrival time.

# Attacks to histogram-based detectors

- In some other cases, **anomalous data** have a known distribution.
- Image proc.: Comb distribution in DCT domain after compression.



- Spiked autocorrelation in a watermarked flow to a hidden server in Tor.

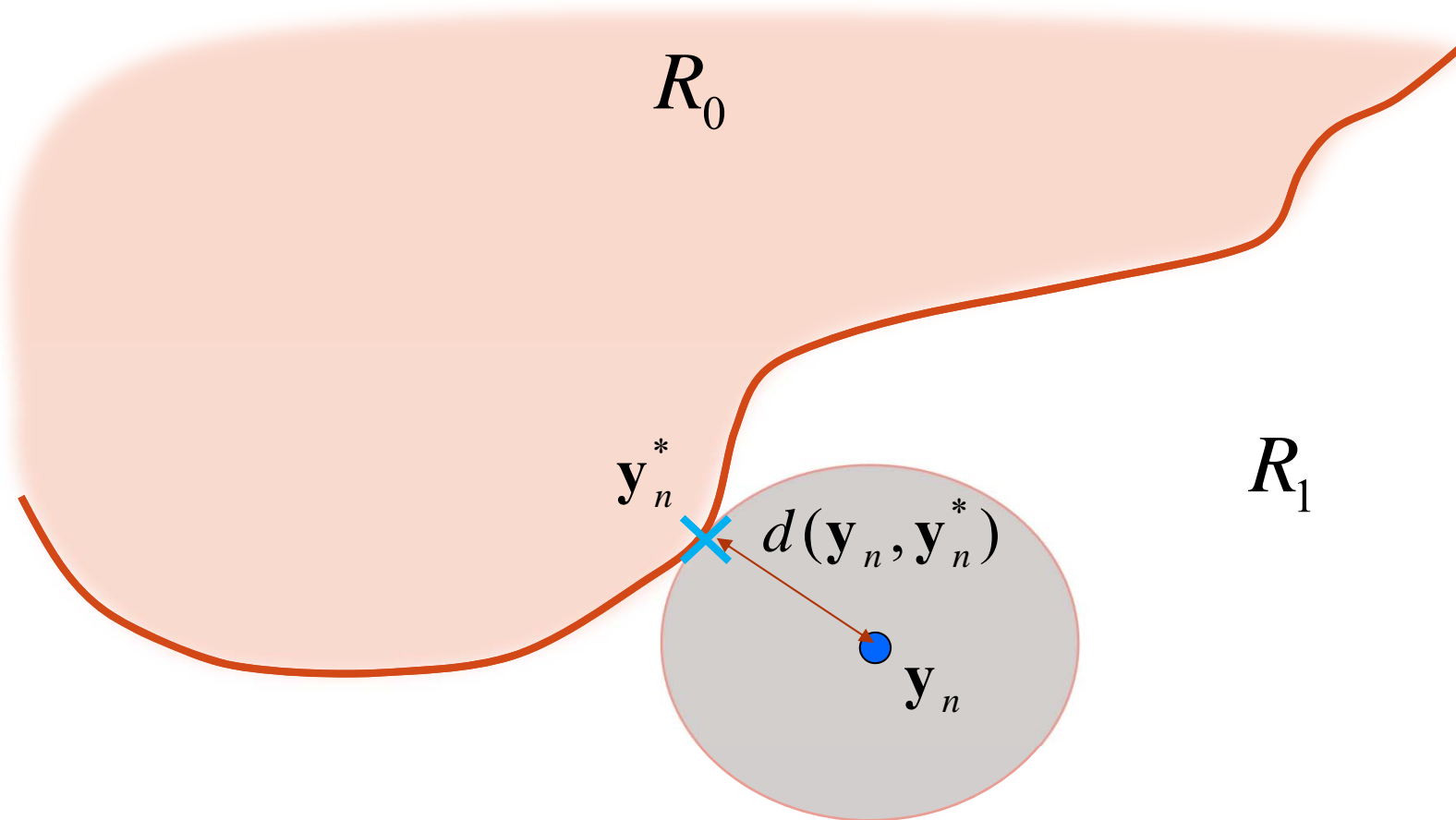


# Attacks to histogram-based detectors

- But the optimal asymptotic detector for i.i.d. sources is based on the histogram!!
- In all these cases, the acceptance region for the defender  $R_0$  is based on the histogram.
- Given a distortion metric and a vector  $\mathbf{y}_n$ , can we solve the adversary's optimization for 1D histogram-based detectors?

$$\mathbf{y}_n^* = \arg \min_{\mathbf{z}_n \in R_0} d(\mathbf{y}_n, \mathbf{z}_n)$$

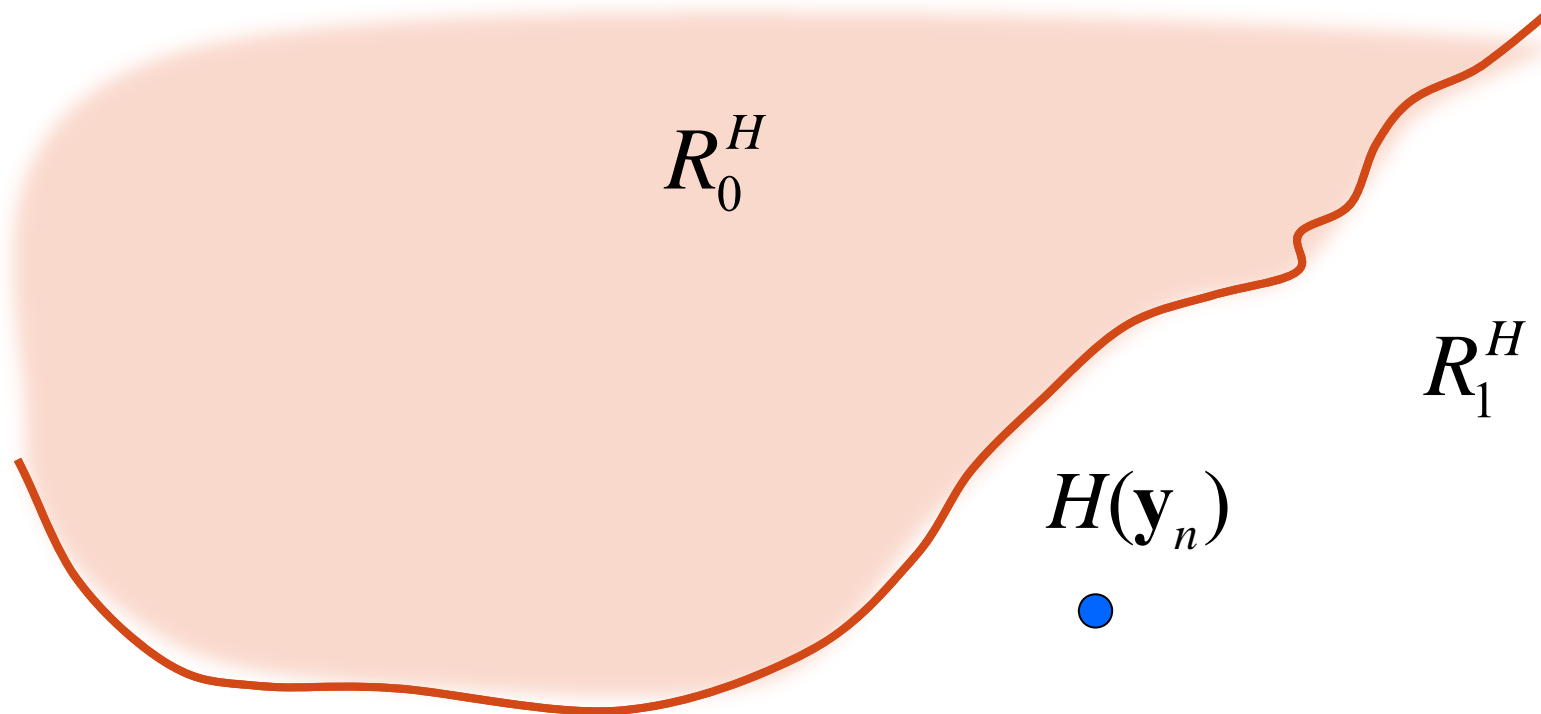
# Attacks to histogram-based detectors



# Attacks to histogram-based detectors

[Comesaña, Pérez-González, 13]

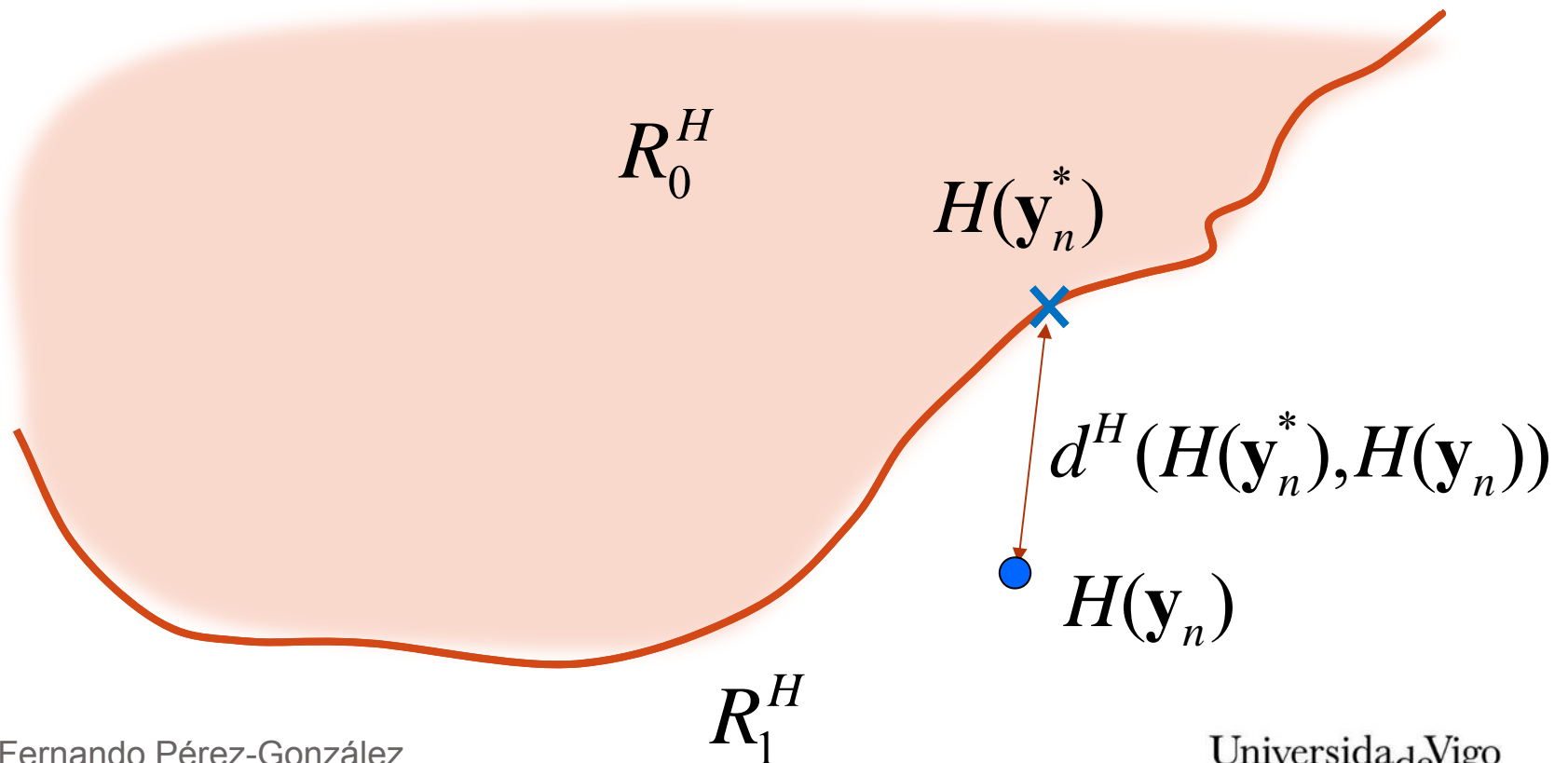
- If we use a Euclidean distance, then the problem can be solved searching along the boundary of the decision region IN THE HISTOGRAM DOMAIN.



# Attacks to histogram-based detectors

## [Comesaña & Pérez-González 13]

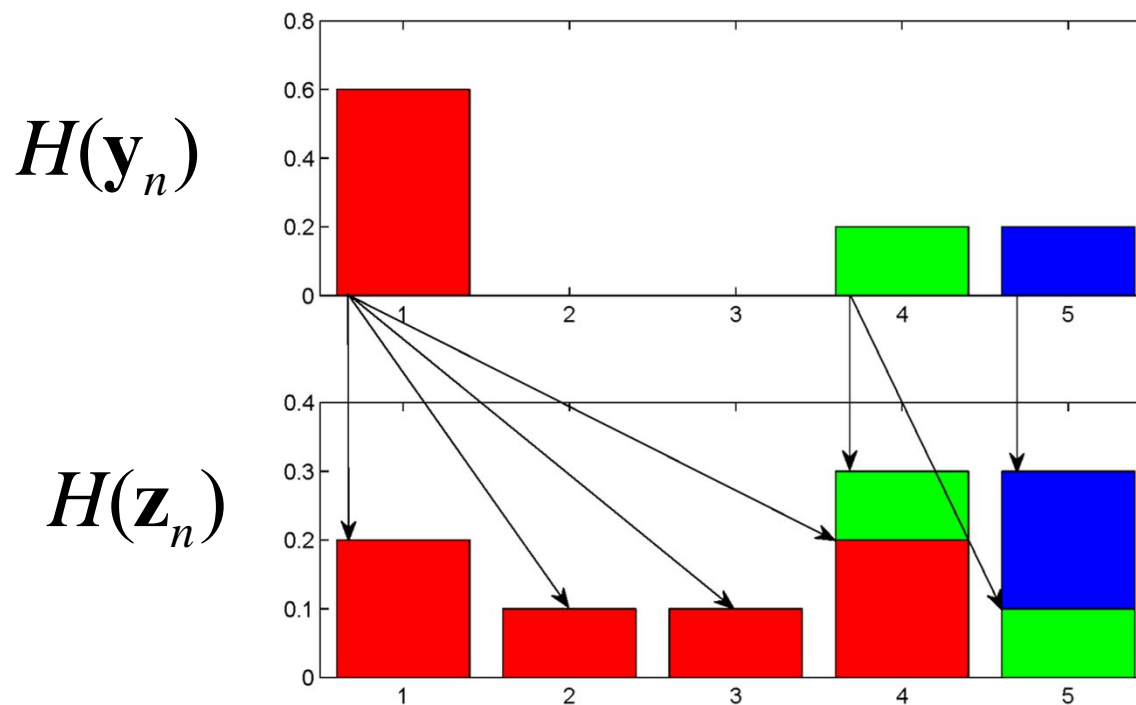
- But we must use a different “distance” between histograms, or better yet, between their cumulative distributions.



# Attacks to histogram-based detectors

## [Comesaña & Pérez-González 13]

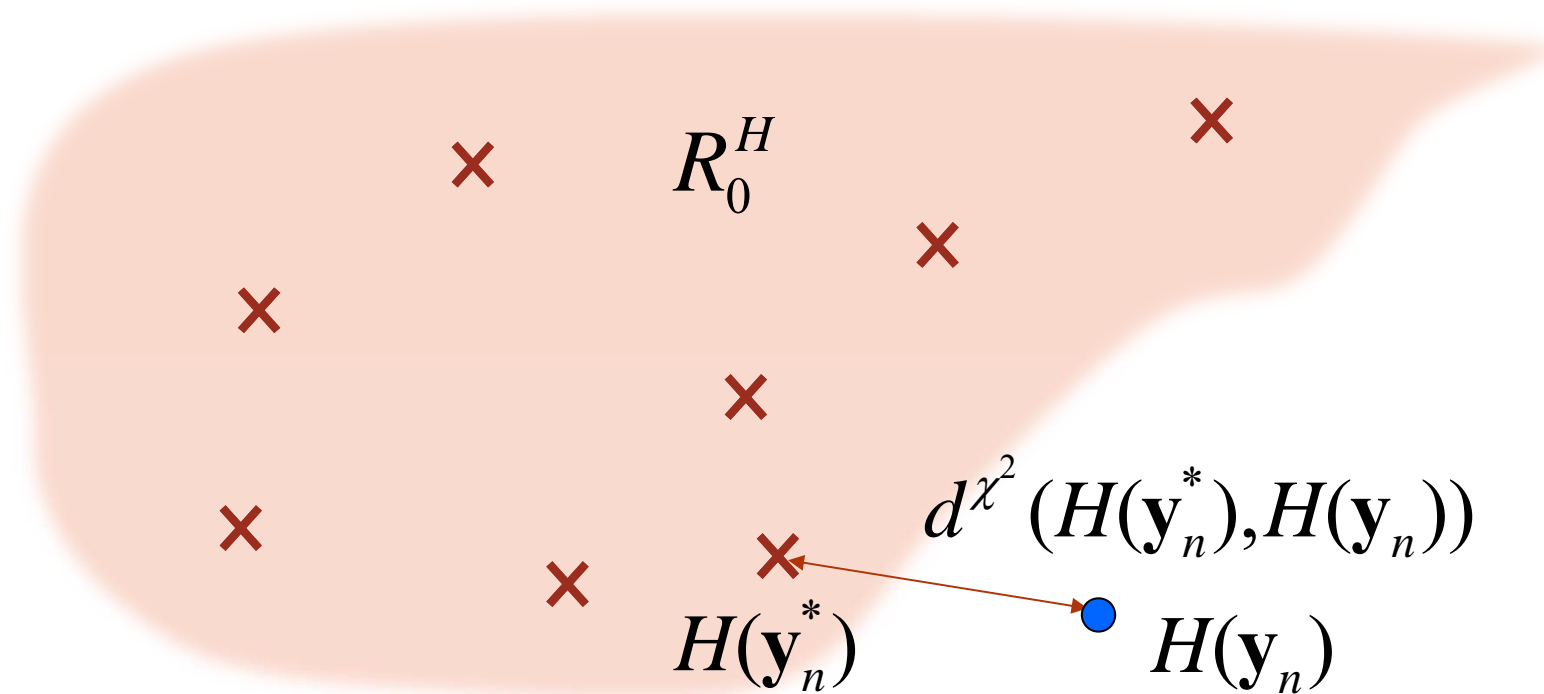
- This “distance” between histograms comes from the use of transportation theory. We seek the histogram in  $R_0$  that is “cheapest” in terms of transportation of probability masses.



# Attacks to histogram-based detectors

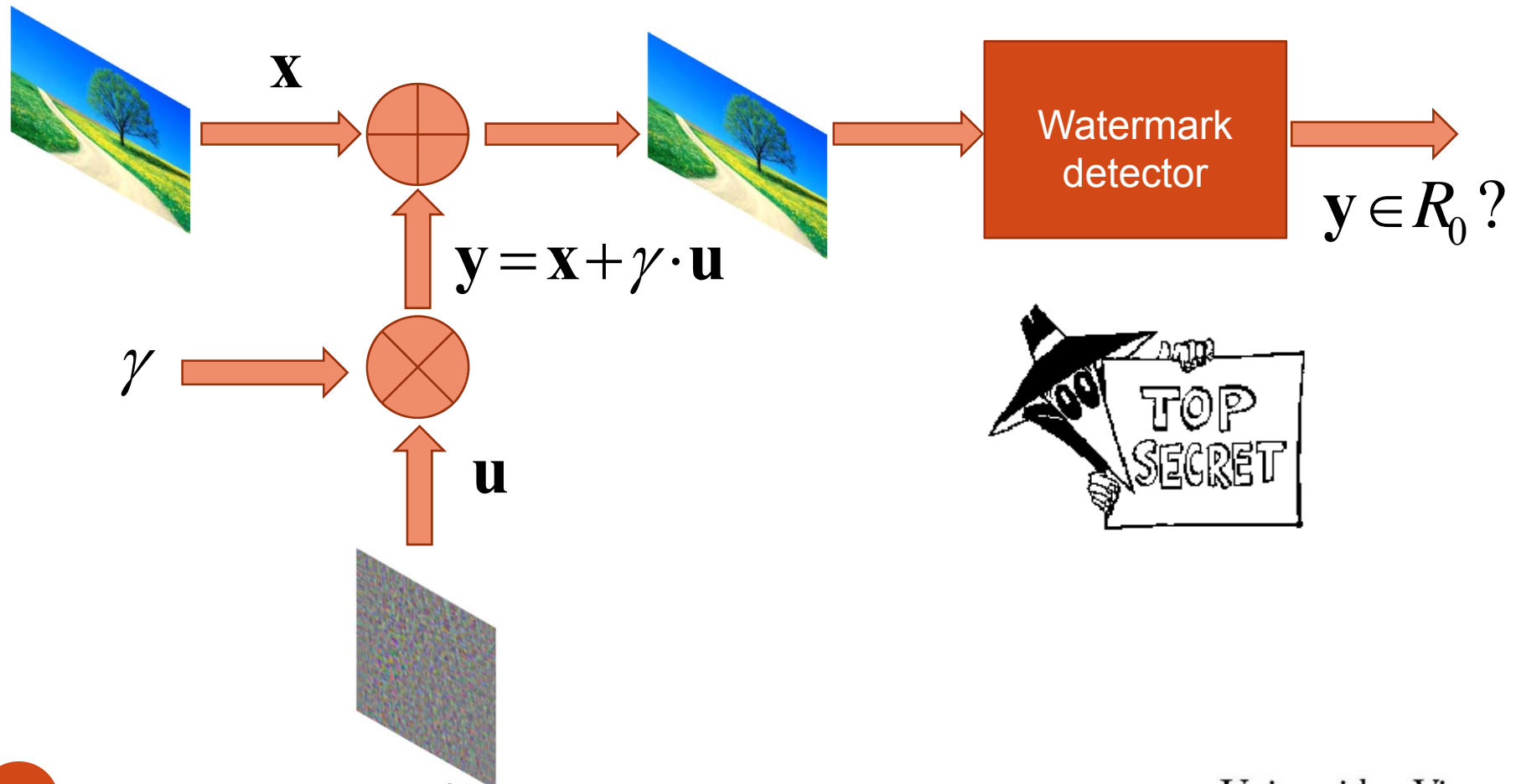
[Barni et al. 12]

- Describe the set  $R_0^H$  by enumeration. Find the closest representative in chi-squared distance. Then, use transportation theory with an adapted perceptual measure to do the pixel remapping.



# What if adversary doesn't know $R_0$ ?

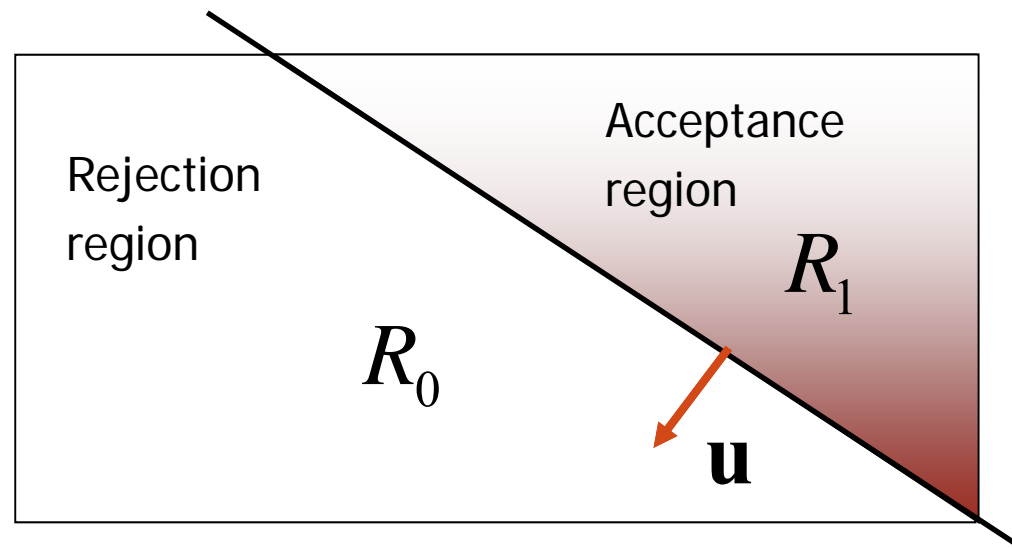
- **Example:** Spread-spectrum watermarking



# Spread-spectrum watermarking

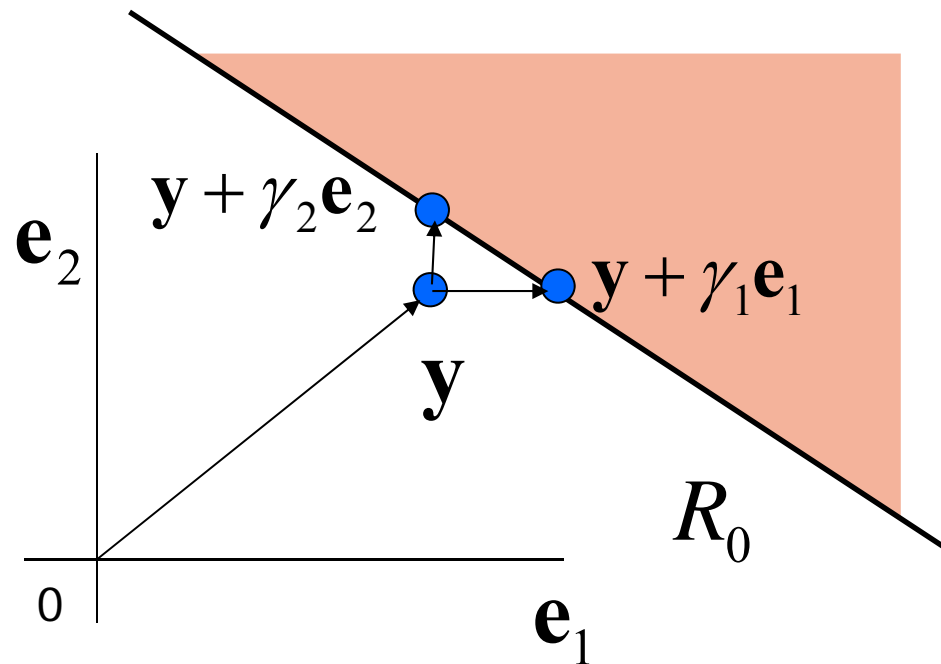
- Optimal (Neyman-Pearson) detector for Gaussian hosts is the cross-correlation.

$$\|\mathbf{y}\|^2 - \|\mathbf{y} - \gamma \cdot \mathbf{u}\|^2 \geq \eta \Leftrightarrow R_1 = \{\mathbf{y} : \mathbf{y}^T \mathbf{u} \geq \lambda\}$$



# Boundary estimation attacks

[El Choubassi & Moulin 05]



- Step 1: Generate an image close to the boundary.
- Step 2: Find values  $\gamma_i$ ,  $i=1, \dots, N$  using line search.
- Step 3: Solve the linear system

$$\mathbf{1} \mathbf{y}^T \mathbf{u} + \text{diag}\{\boldsymbol{\gamma}\} \mathbf{u} = \lambda \mathbf{1}$$

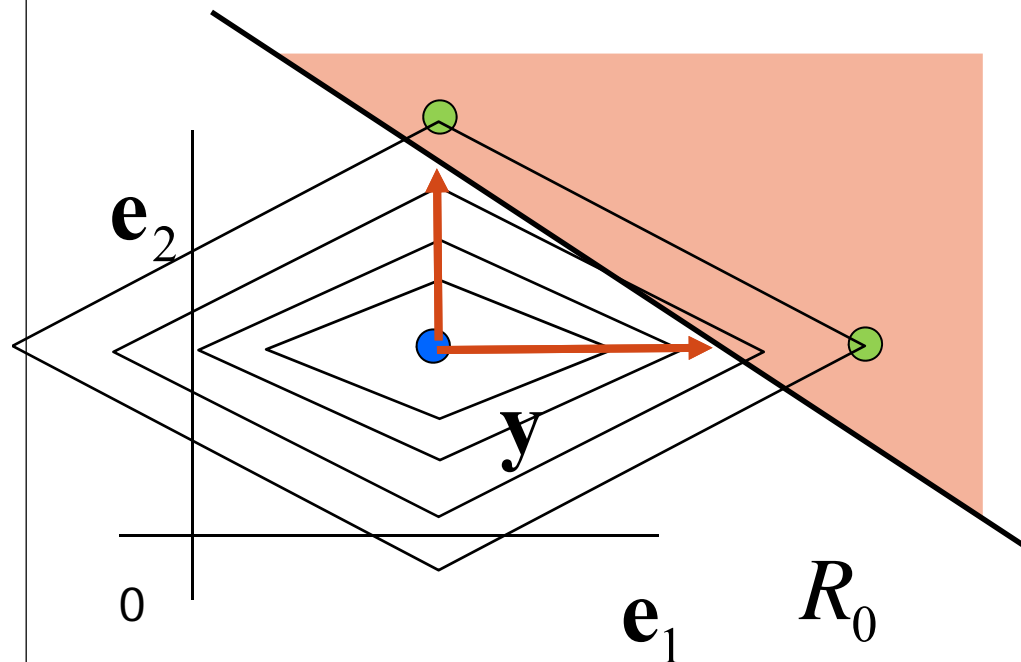
## Boundary estimation attacks (2)

### [El Choubassi & Moulin 05]

- The previous attack can be extended to more complicated decision regions under the following conditions:
  - Knowledge of the detection function (but not the secret parameters) is required.
  - The decision statistic is twice differentiable.
  - The gradient  $\nabla(\partial R_0)$  is invertible.
- This is applicable to find out the secret parameters for polynomial and  $l_p$ -norm-based (if the shape parameter  $p > 1$ ) detectors.

# Adversarial Classification Reverse Engineering

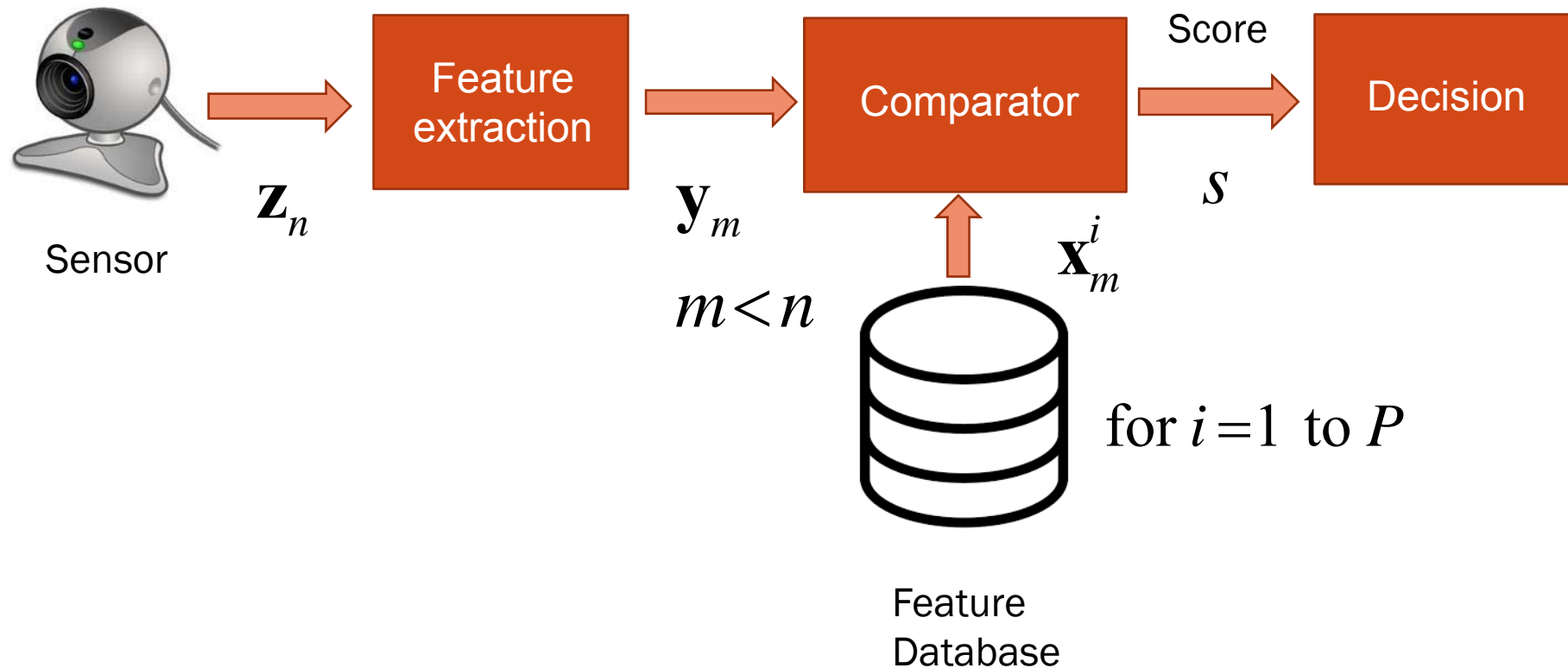
[Lowd & Meek 05]



- Use weighted  $l_1$  cost to find closest point to a linear classifier in **antispam filtering**.
- Assumes features are known and coincide with cost coordinates.
- First, find the sign of cost weights.
- Then, do line search to learn the boundary.

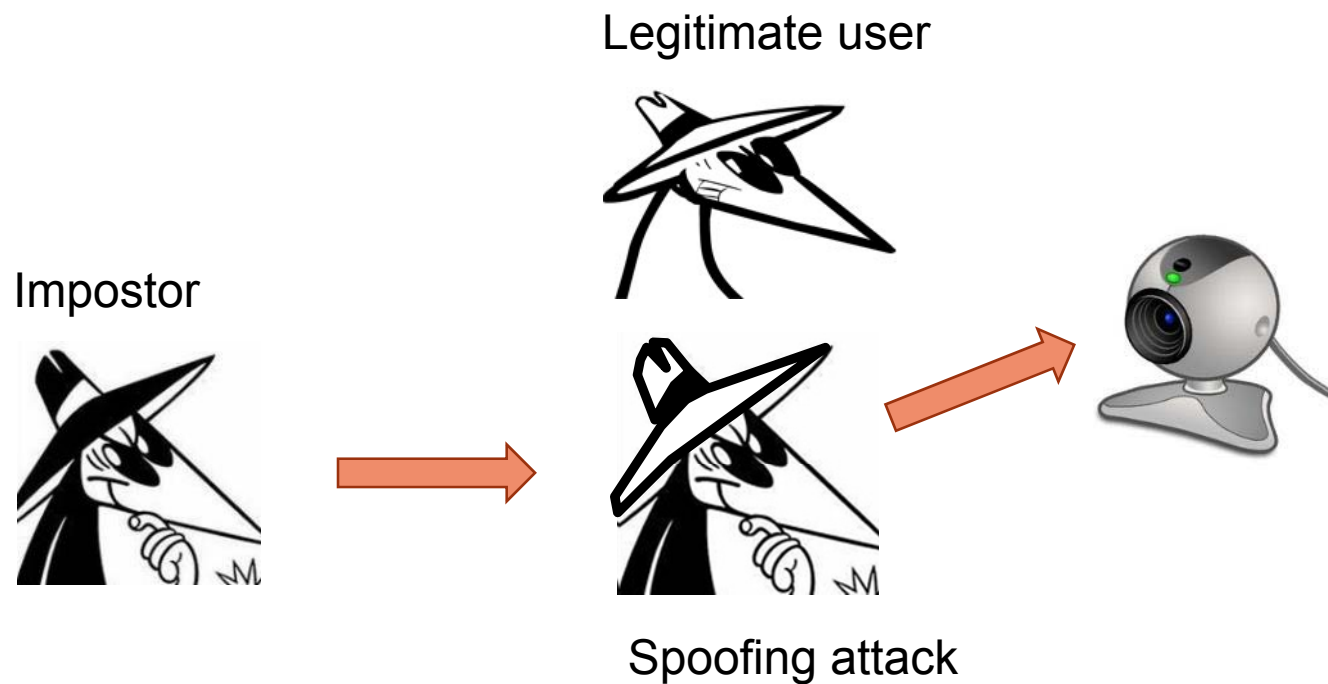
# What if adversary doesn't know $R_0$ ?

- **Example:** Biometric identification

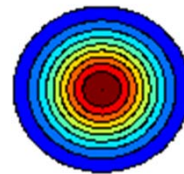
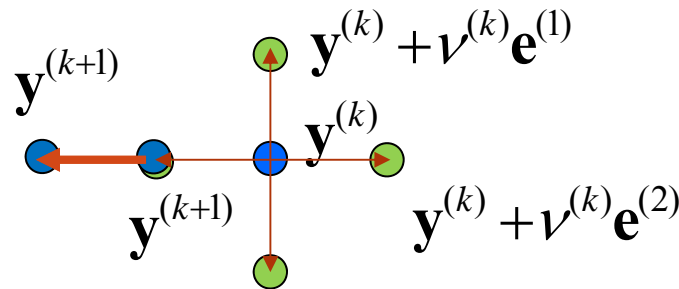


# Biometric identification

- Indirect attacks: within the digital boundaries.
- Spoofing attacks: at the sensor level.

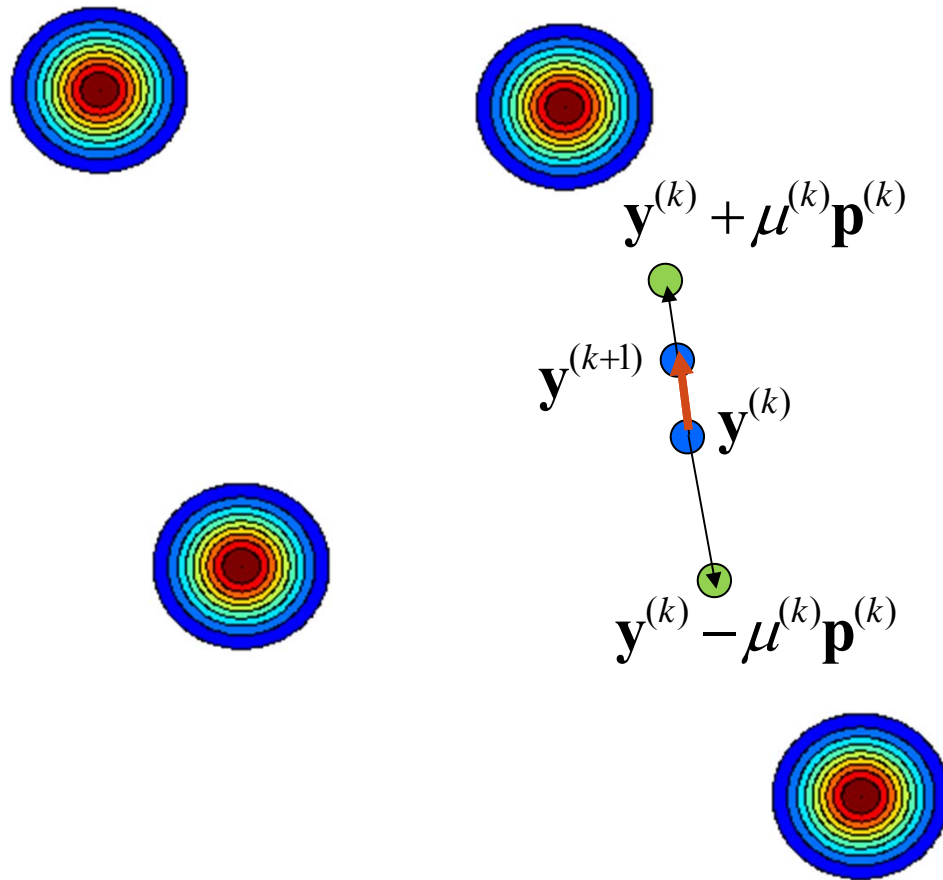


# Hill-climbing: Hook-Jeeves algorithm



- Step 1: Select a starting vector  $\mathbf{y}^{(0)}$  from known priors
- Step 2: For each canonical vector  $\mathbf{e}^{(i)}, i = 1, \dots, n$ 
  - Evaluate score at  $\mathbf{y}^{(k)} \pm v^{(k)} \mathbf{e}^{(i)}$
- Step 3: Take the maximum as  $\mathbf{y}^{(k+1)}$
- Step 4: If  $s(\mathbf{y}^{(k+1)}) > s(\mathbf{y}^{(k)})$  then explore further in the direction  $(\mathbf{y}^{(k+1)} - \mathbf{y}^{(k)})$  and replace  $\mathbf{y}^{(k+1)}$  if improvement.
- Else,  $v^{(k+1)} = v^{(k)} / 2$
- Step 5: Go back to 2

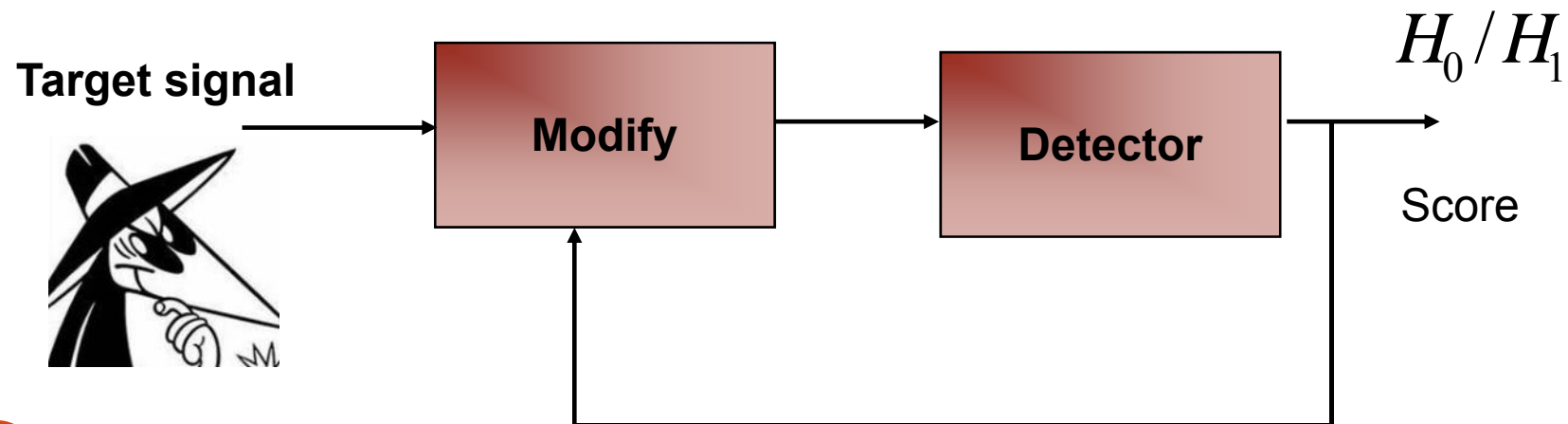
# Hill-climbing: Simultaneous perturbation stochastic approximation [Spall 98]



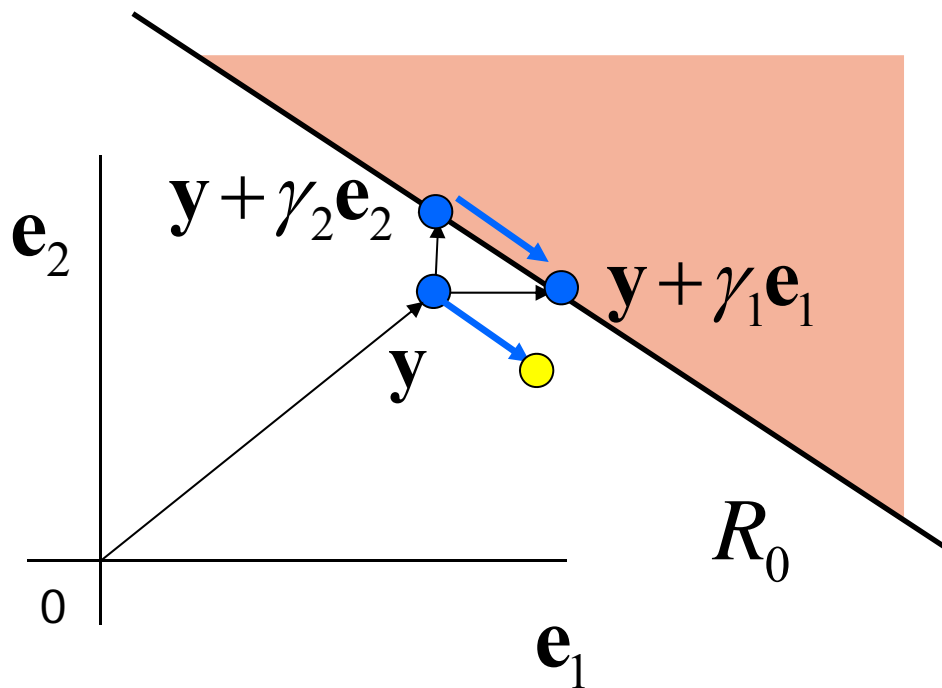
- Step 1: Select a starting vector  $\mathbf{y}^{(0)}$  from known priors
- Step 2: Generate random perturbation vector  $\mathbf{p}^{(k)} \in \{\pm 1\}^n$
- Step 3: Evaluate the gradient  $\hat{\nabla}_{\mathbf{p}^{(k)}} s(\mathbf{y}^{(k)})$  in the direction  $\mathbf{p}^{(k)}$
- Step 4: Update
$$\mathbf{y}^{(k+1)} = \mathbf{y}^{(k)} - \nu^{(k)} \hat{\nabla}_{\mathbf{p}^{(k)}} s(\mathbf{y}^{(k)})$$
- Step 5: Update  $\nu^{(k)}$  and  $\mu^{(k)}$  decreasingly.
- Step 6: Go back to 2.

# Oracle attacks

- Suitable when the detection function is unknown to the adversary
  - Learning-based with unknown training set (e.g., antispam filters)
  - Rule-based with unknown rules (e.g., anomaly-based detection)
  - Unknown template-based (e.g., biometric identification)
  - Key-dependent (e.g., watermarking)
- Typically, require a very large number of queries



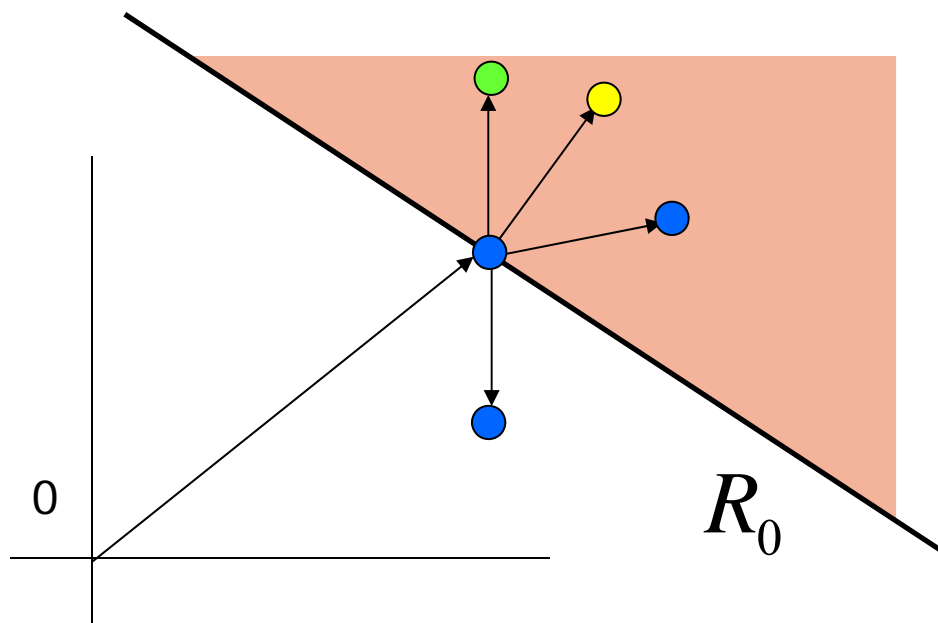
## Sensitivity attacks [Cox & Linnartz 97]



Estimation of the tangent is not so evident in higher dimensions.

- Step 1: Generate an image  $y$  close to the boundary (in the rejection region).
- Step 2: For the  $j$ -th dimension.
  - Find  $\gamma_j$  such that  $y + \gamma_j e_j$  is on the boundary.
- Step 3: Estimate the tangent.
- Step 4: Move along the tangent and evaluate perceptual quality.
- Step 5: Go back to Step 2.

## A variant of the sensitivity attack [Kalker 98]



- Step 1: Generate an image at the boundary.
- Step 2: Add a random perturbation.
  - If the answer is in  $R_0$ , change the sign.
- Step 3: Average the answers.

The result is an estimate of the projection vector.

# Countermeasures

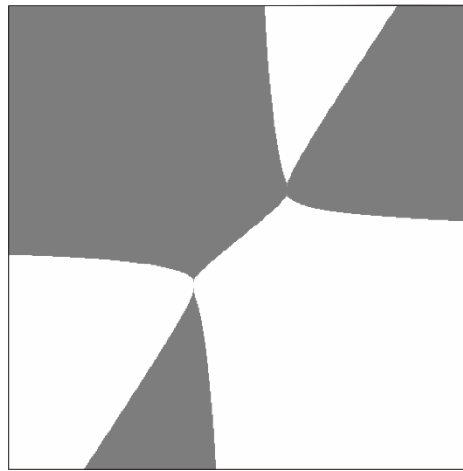
- One obvious countermeasure is to complicate the decision region, BUT we still want good detection performance!
- Several available solutions:
  - Based on  $l_p$ -norms.
  - Based on polynomial functions.
  - Based on “fractalizing” the boundary.
- Another solution is to “randomize” the boundary of the decision región.

# $l_p$ -norm based detection [Hernandez & Pérez-González 98]

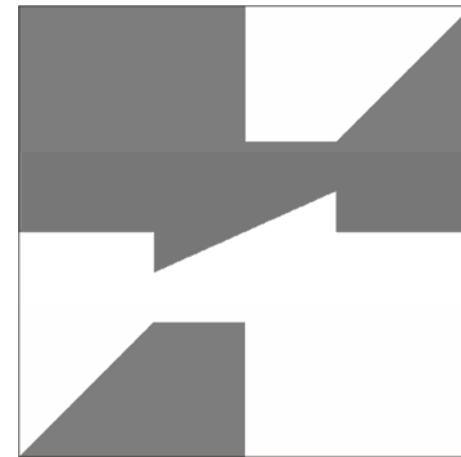
- Arises naturally from considering the host image following a generalized-gaussian distribution with shape parameter  $p$ .
- Can be implemented privately.



$p=2$ . Gaussian



$p=1/2$



$p=1$ . Laplacian

# Polymial detectors [Furon et. al 02]

- JANIS: Just Another N-order Side-Informed Scheme.
- Based on the following detection function

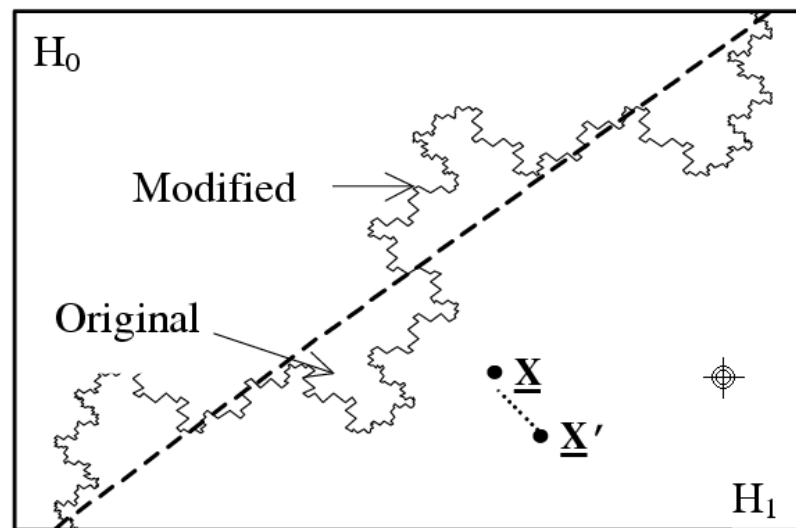
$$D_K(\mathbf{y}) = \sum_{k=1}^{N/n} \prod_{j=0}^{n-1} u[i_{j,k}] \cdot y[i_{j,k}]$$

where the indices  $i_{j,k}$  denote a pseudorandom ordering (also key-dependent).

- The watermark is obtained as  $\mathbf{w} = \gamma \nabla D_K(\mathbf{y})$
- For  $n=1$ , the classical correlation detector is recovered.

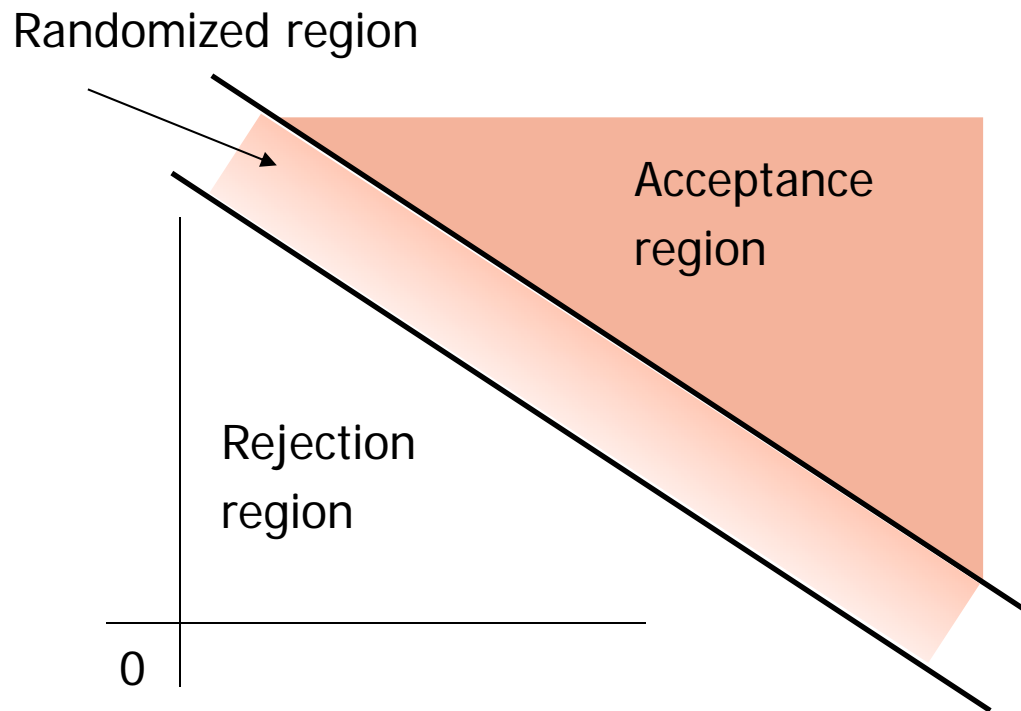
# Fractal-boundary detectors

- Perturb the decision boundary by a Peano curve, which is kept secret.
- Then, the watermarked signal is also perturbed to preserve the distance to the decision boundary. This adds some degradation.



Taken from A.H. Tewfik and M.F. Mansour,  
“Secure Watermark detection with  
nonparametric decision boundaries”,  
ICASSP 2002. © IEEE

# Randomized-boundary detectors



- Idea: provide less information at the boundary by randomizing it.

- Rule:

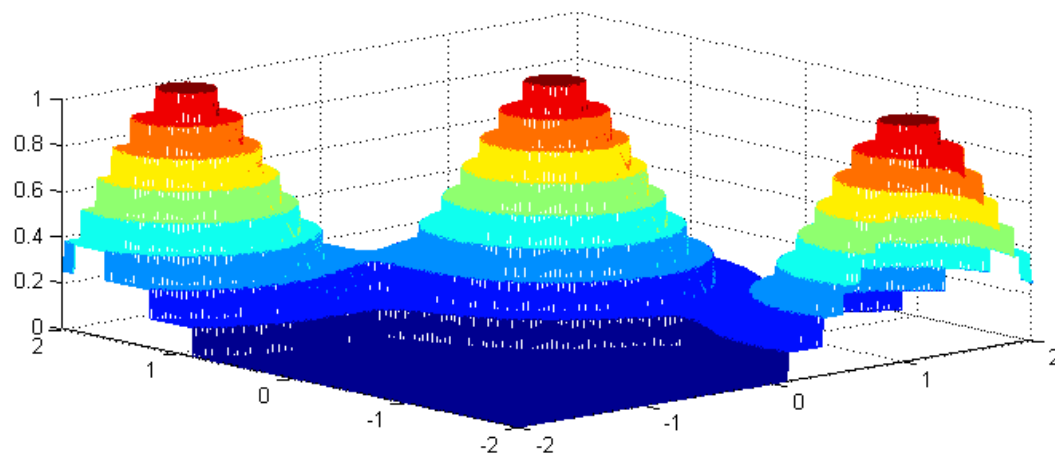
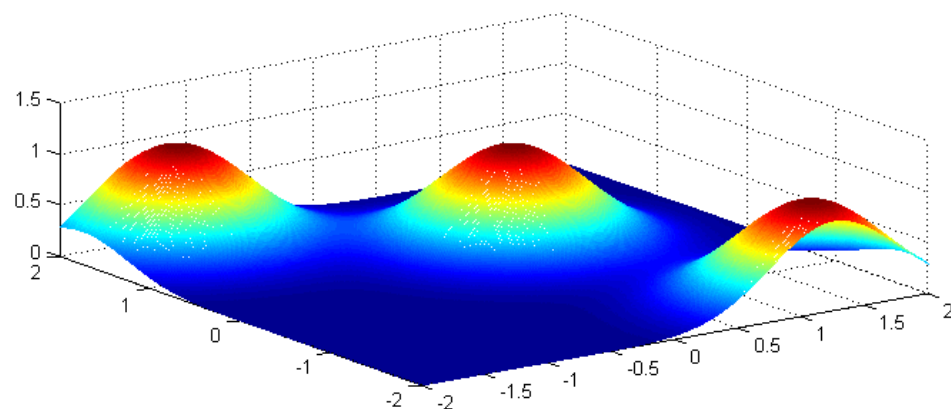
Decide  $H_0$  if  $\mathbf{y}\mathbf{u}^T > \eta_1$

Decide  $H_1$  if  $\mathbf{y}\mathbf{u}^T < \eta_2$

Decide  $H_1$  with prob  
 $p(\mathbf{y}\mathbf{u}^T)$  if  $\eta_1 > \mathbf{y}\mathbf{u}^T > \eta_2$

The internal behavior may be deterministic/purely random for a given  $\mathbf{y}\mathbf{u}^T$

# Score quantization



# The Blind-Newton Sensitivity Attack

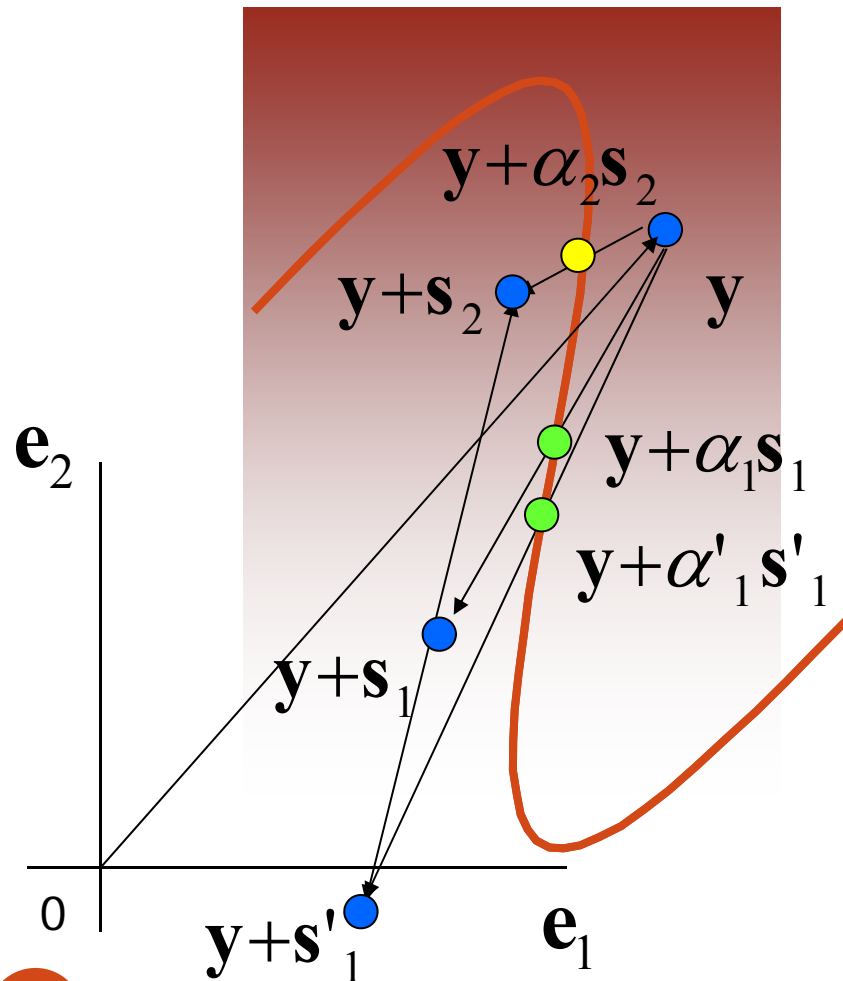
## [Comesaña et al. 05]

*P. Comesaña, L. Pérez-Freire, F. Pérez-González, "The return of the sensitivity attack", IWDW'05.*

- In many cases, the “security by obscurity” principle applies: we don’t know what is in the black box.
- The principles of the sensitivity attack can be extended to devise a blind descent algorithm (Newton-like).
- The objective function to be minimized is the Euclidean distance to the available watermarked image  $\mathbf{y}$ , i.e., we seek the perturbation  $\mathbf{t}$  such that  $f(\mathbf{t}) = \|(\mathbf{y} + \mathbf{t}) - \mathbf{y}\|^2$  is minimum and yield  $H_0$ .
- This is done by moving along the boundary of the decision region.

# Blind-Newton Sensitivity Attack

[Comesaña et al. 05]



- Step 1: Get perturbation  $s$  and find  $\alpha$  such that  $y+\alpha s$  is on the boundary.
- Step 2: Numerically evaluate gradient of  $\|s\|^2$  and possibly Hessian on the boundary.
- Step 3: Update
 
$$s_{k+1} = s_k - \varepsilon_k [\nabla^2 f(s)]^{-1} \cdot \nabla f(s)$$
 where  $f$  is the objective function defined ONLY on the boundary.
- Step 4: Go back to 1.

# BNSA Against Spread Spectrum



Positive  
detection



Negative detection:  
additive noise

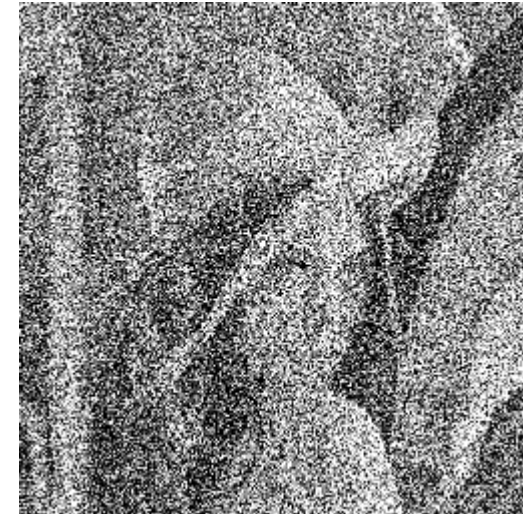


Negative detection:  
BNSA, 1 iteration

# BNSA Against JANIS



Positive  
detection

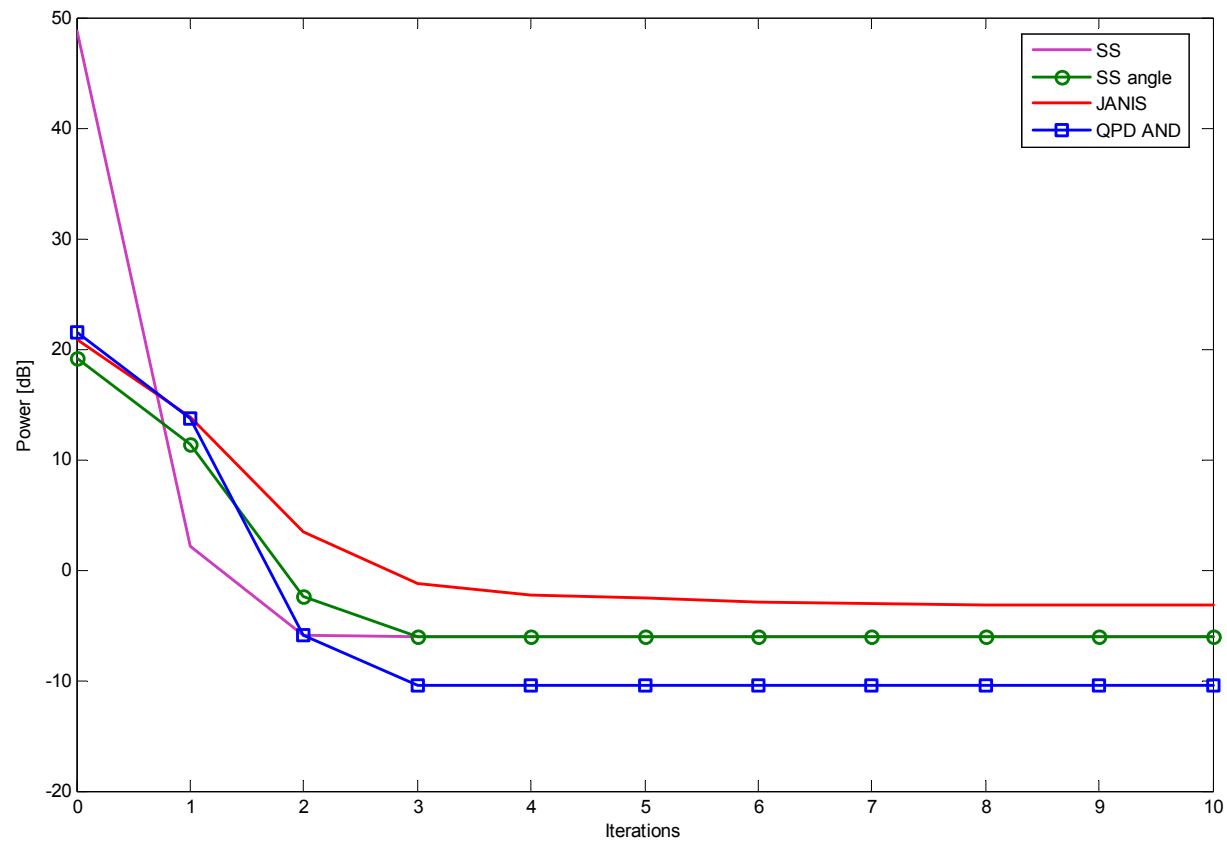


Negative detection:  
additive noise



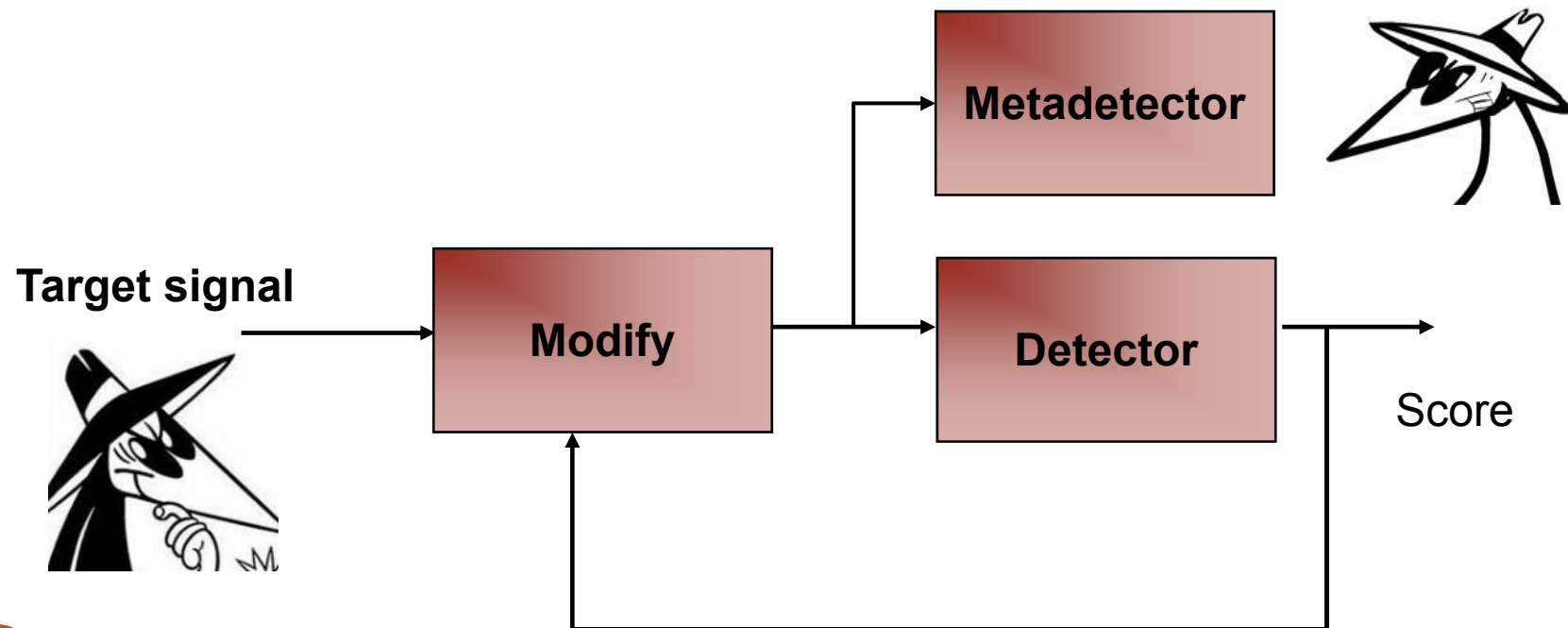
Negative detection:  
BNSA, 1 iteration

# BNSA attack power vs. iterations



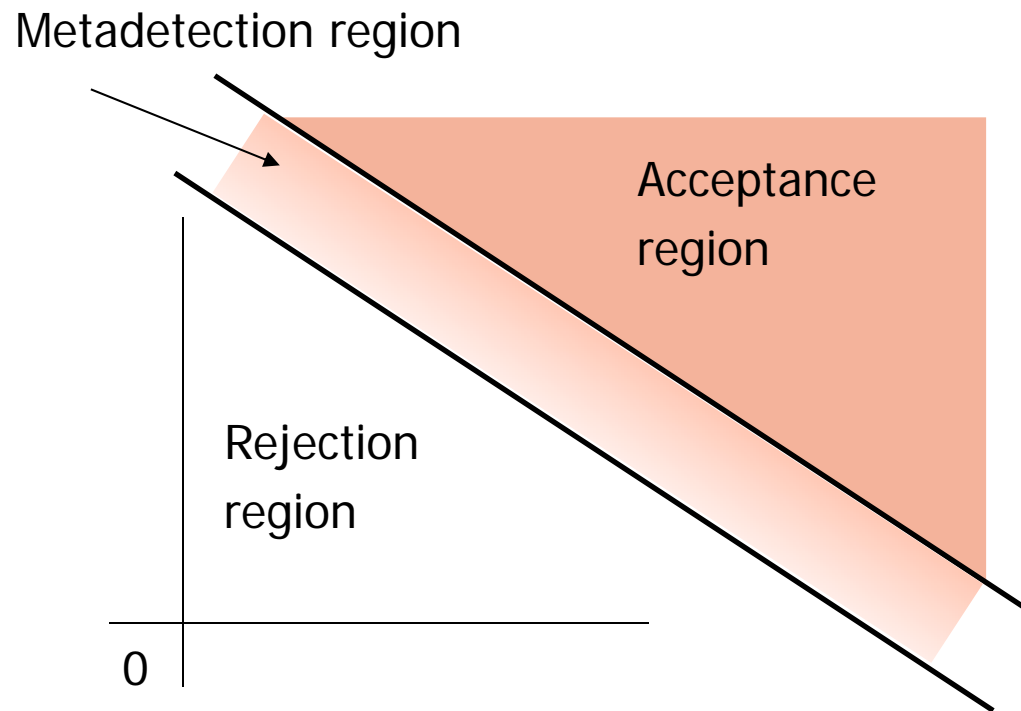
# Metadetection

- Detect anomalous behavior in the set of queries.
- Most Oracle-based attacks induce distinctive patterns that can be (meta)detected.
- In large-dimensional spaces normal queries will look random; targeted attacks will not.



# Closeness to the boundary detector

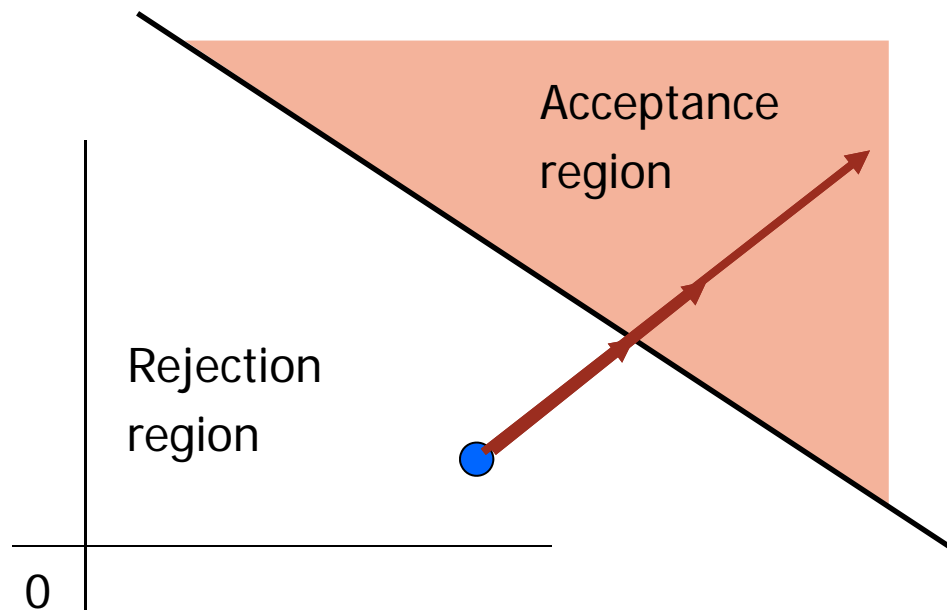
[Tondi et al. 14]



- Idea: measure whether the attacker is sending queries that are close to the boundary.
- Even if attack queries are disguised among normal ones, they can be statistically detected.
- In fact, as long as the number of malicious queries increases as the log of the number of dimensions, the attack can be detected.

# Line-search detector

[Tondi et al. 14]



- Most attacks do binary searches to locate points on the boundary.
- The idea is to detect queries forming a line.
- In a large dimensional space, the likelihood that even three queries are close to a line is very small.

## Conclusion

- It's hard to avoid the cat and mouse game.
- In some specific cases game-theory can provide (Nash) equilibria. Can we enlarge the subset of problems where solutions are known?
- Can we at least detect that there is a catversary?... However, doesn't this lead to a metadetection game?
- How about setting traps to the adversary? Some of the existing ones seem to be effective...
- Adversaries can partially influence the learning process (adversarial machine learning).
- How do we factor in bounded computational resources for the parties.



## Conclusion

- And adversarial detection is just the beginning....



# Bibliography

- Javier Galbally, Sebastien Marcel, Julian Fierrez. Biometric Antispoofing Methods: A Survey in Face Recognition. IEEE Access, vol. 2, pp. 1530-1552, 2014.
- M. El Choubassi and P. Moulin, “Noniterative algorithms for sensitivity analysis attacks,” IEEE Trans. on Information Forensics and Security, vol. 2, no. 2, pp. 113–126, 2007.
- P. Comesaña and F. Pérez-González, “Breaking the BOWS watermarking system: key guessing and sensitivity attacks,” EURASIP Journal on Information Security, vol. 2007, 2007
- M. F. Mansour and A. H. Tewfik, “LMS-based attack on watermark public detectors,” in IEEE International Conference on Image Processing, ICIP’02, September 2002, vol. 3, pp. 649–652.
- J.R. Troncoso-Pastoriza and F. Pérez-González, “Zero-knowledge watermark detector robust to sensitivity attacks,” in Proceedings of the 8th workshop on Multimedia and security. ACM, 2006, pp. 97–107.
- J-P. M. G. Linnartz and M. Van Dijk, “Analysis of the sensitivity attack against electronic watermarks in images,” in Proc. Information Hiding, LNCS Vol. 1525, 1998, pp.

# Bibliography

- P. Comesaña, L. Pérez-Freire, and F. Pérez-González, “Blind Newton sensitivity attack,” in IEE Proc. on Information Security. IET, 2006, vol. 153, pp. 115–125.
- R. Venkatesan and M. H. Jakubowski, “Randomized detection for spread-spectrum watermarking: defending against sensitivity and other attacks,” in Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing, 2005.
- M. Barni and B. Tondi, “The source identification game: an information-theoretic perspective,” IEEE Trans. Information Forensics and Security, vol. 8, no. 3, pp. 450–463, March 2013.
- D. Lowd and C. Meek, “Adversarial learning,” in Proc. of 13th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining European, 2005, pp. 641–647.
- M. C. Stamm, W. S. Lin, and K. J. R. Liu, “Forensics vs anti-forensics: a decision and game theoretic framework,” in ICASSP 2012, IEEE International Conference on Acoustics, Speech and Signal Processing, Kyoto, Japan, 25-30 March 2012.
- I. J. Cox and J. P. M. G. Linnartz, “Public watermarks and resistance to tampering,” in IEEE International Conference on Image Processing ICIP’97, , vol. 3, pp. 3–6., Santa Barbara, California, USA, October 1997

# Bibliography

- A. Adler, “Biometric system security,” in Handbook of biometrics, A.K. Jain, P. Flynn, and A.A. Ross, Eds., pp. 381–402. Springer, 2008
- M. Barni, M. Fontani, and B. Tondi, “A universal technique to hide traces of histogram-based image manipulations,” in Proc. of ACM Workshop on Multimedia and Security, Coventry, UK, September 2012, pp. 97–104
- P. Comesaña and F. Pérez-González, “Optimal counterforensics for histogram-based forensics,” in Proc. IEEE ICASSP, Vancouver, Canada, May 2013, pp. 3048–3052
- M. Barni and F. Perez-González, “Coping with the enemy: Advances in adversary-aware signal processing,” in Proc. IEEE ICASSP, Vancouver, Canada, May 2013, pp. 8682–8686.
- E. Maiorana, G.E. Hine, and P. Campisi, “Hill-Climbing Attacks on Multibiometrics Recognition Systems”, IEEE Trans. On Information Forensics and Security, vol. 10, no.5, pp. 900-915, 2015.
- M. Barni, P. Comesana Alfaro, F. Perez-Gonzalez, B. Tondi. "Are you threatening me?: towards smart detectors in watermarking". IS&T/SPIE Electronic Imaging 2014, San Francisco, California, February 2014.

# Bibliography

- T. Furon, B. Macq, N. Hurley and G. Silvestre, “JANIS: Just Another N-order side-informed watermarking scheme”, in Proc. of IEEE International Conference on Image Processing, October 2002.
- J. Hernández and F. Pérez-González, “Statistical Analysis of Watermarking Schemes for Copyright Protection of Images”, Proceedings of the IEEE, vol. 87, pp. 1142-1166, July 1999.
- T. Kalker, “Watermark estimation through detector observations”, Benelux Signal Proc. Symp. 1998.
- J. C. Spall, “Implementation of the simultaneous perturbation algorithm for stochastic optimization,” *IEEE Trans. Aerosp. Electron. Syst.*, vol. 34, no. 3, pp. 817–823, Jul. 1998.
- R. Hooke and T. A. Jeeves, ““Direct search’ solution of numerical and statistical problems,” *J. ACM*, vol. 8, no. 2, pp. 212–229, 1961.