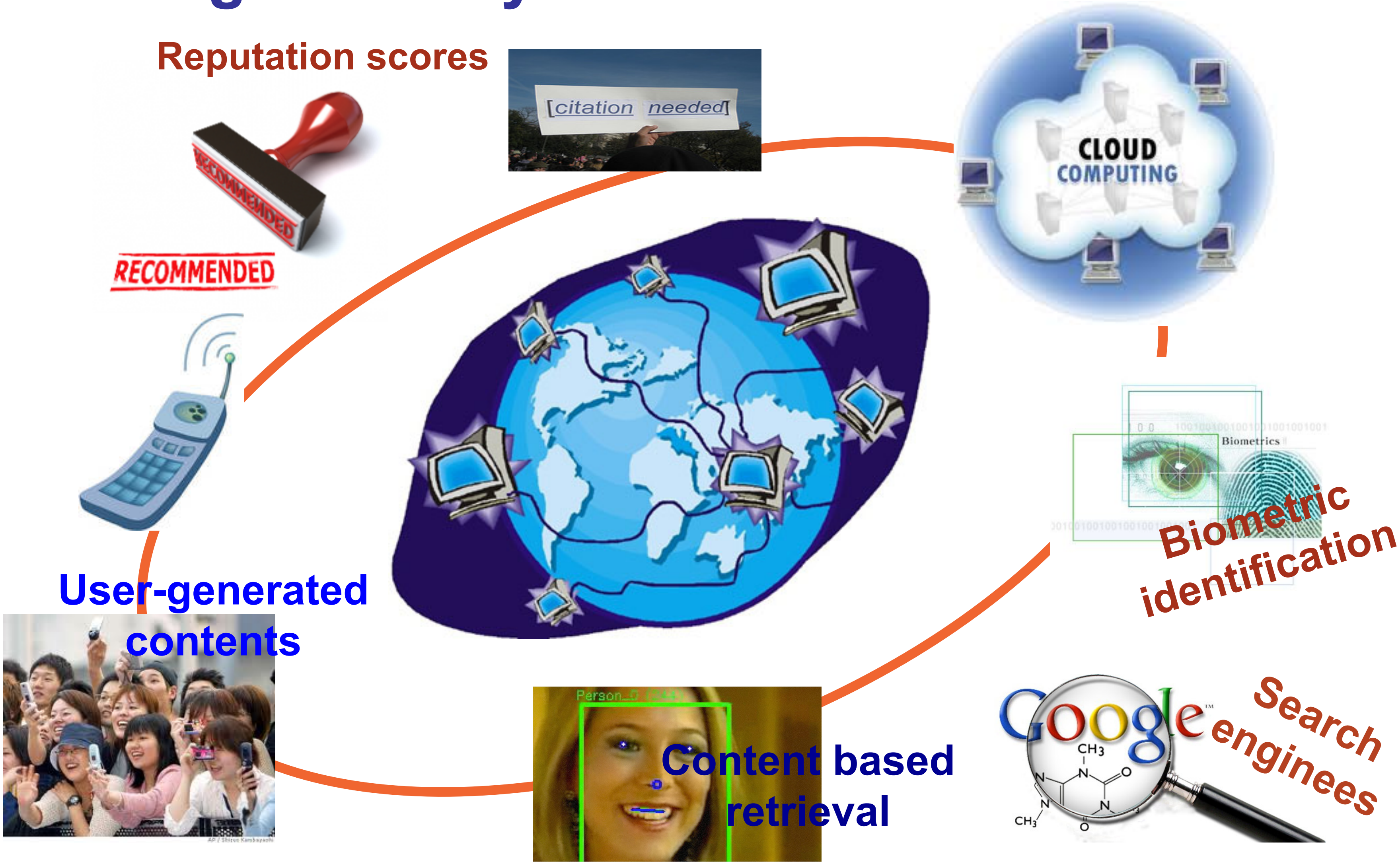
WIFS: Tenerife, Spain, December 3, 2012

Adversary-aware signal processing

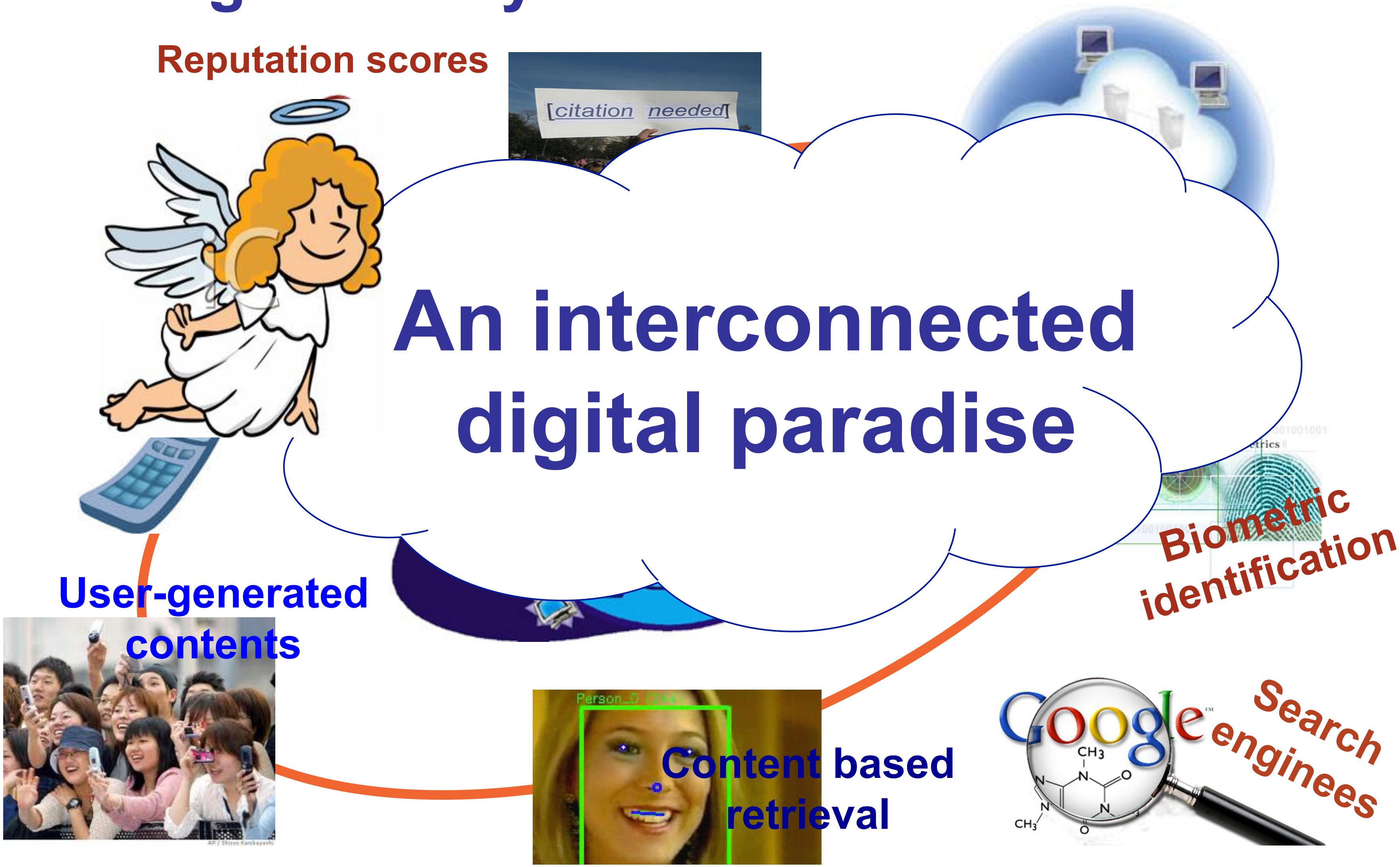
Mauro Barni

University of Siena

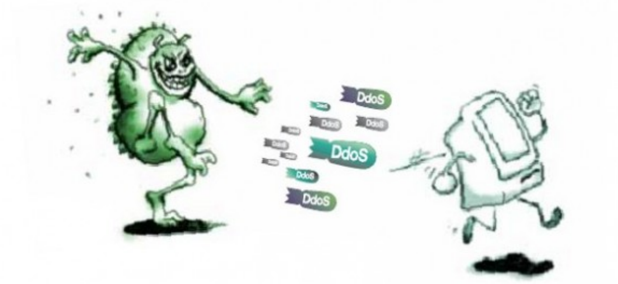
The digital ecosystem we live in

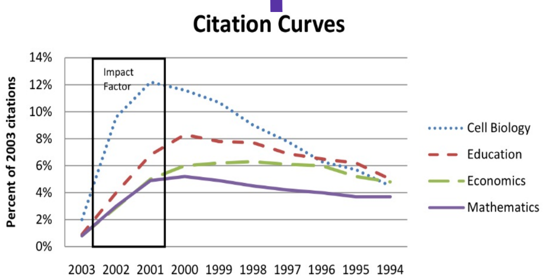


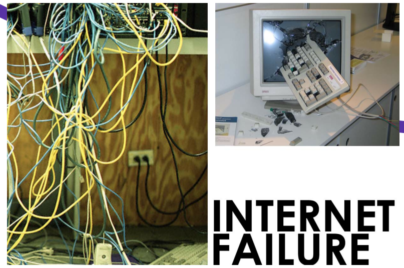
The digital ecosystem we live in



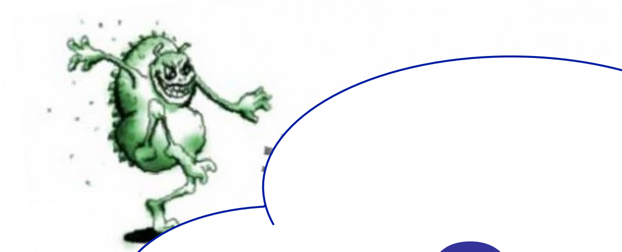
The digital ecosystem we live in



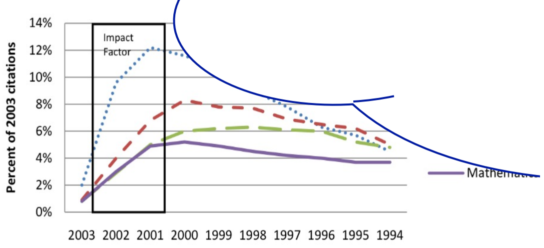


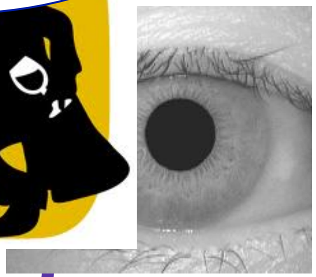


The digital ecosystem we live in



Or a battlefield ?





To the rescue

- Researchers with diverse background have started looking for countermeasures
  - Watermarking - fingerprinting
  - Multimedia forensics
  - Spam filtering
  - Secure classification/learning
  - Anti-spoofing biometrics
  - Network intrusion detection
  - Secure reputation systems
  - Protection against attacks to cognitive radio

  ... and many many others

To a closer look ...

- All these fields face with similar problems ...
- ... but interaction is very limited
- Same solutions are re-invented again and again
- Advances proceed at a slow pace

Even worse

- We miss a global view
- We do not understand the real essence of problems
- Solutions are less effective than possible
- Basic concepts are misunderstood
  - Often we do not even have proper security definitions

To a closer look ...

- All these fields face with similar problems ...
- ... but interaction is very limited

We keep patching techniques thought to work in the digital paradise while we should develop tools explicitly designed for the battlefield

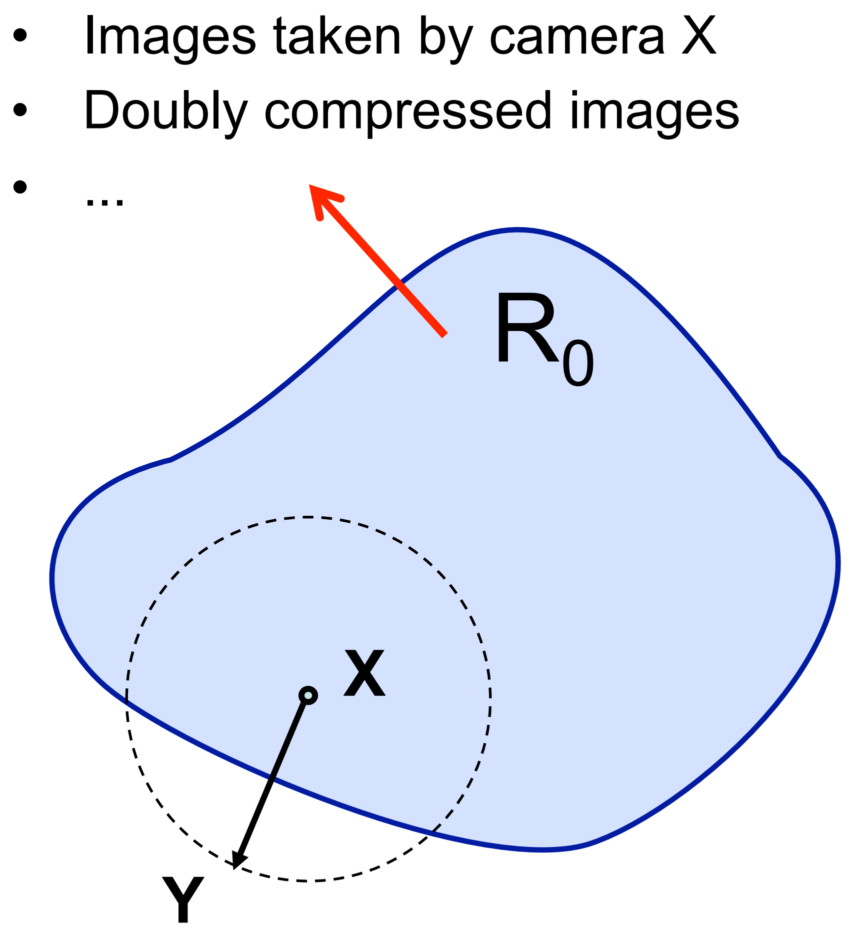
We do not understand the real essence of problems

- Solutions are less effective than possible
- Basic concepts are misunderstood
  - Often we do not even have proper security definitions

Binary decision: most recurrent problem

- Was a given image taken by a given camera ?
- Was this image resized/compressed twice ... ?
- Is this e-mail spam or not ?
- Does this face/fingerprint/iris belong to Mr X ?
- Is X a malevolent or fair user ?
  - Recommender systems, reputation handling
  - Cognitive radio
- Is traffic level indicating the presence of an anomaly/ intrusion ?
- Is this image a stego or a cover ?
- Does an image contain a certain watermark ?

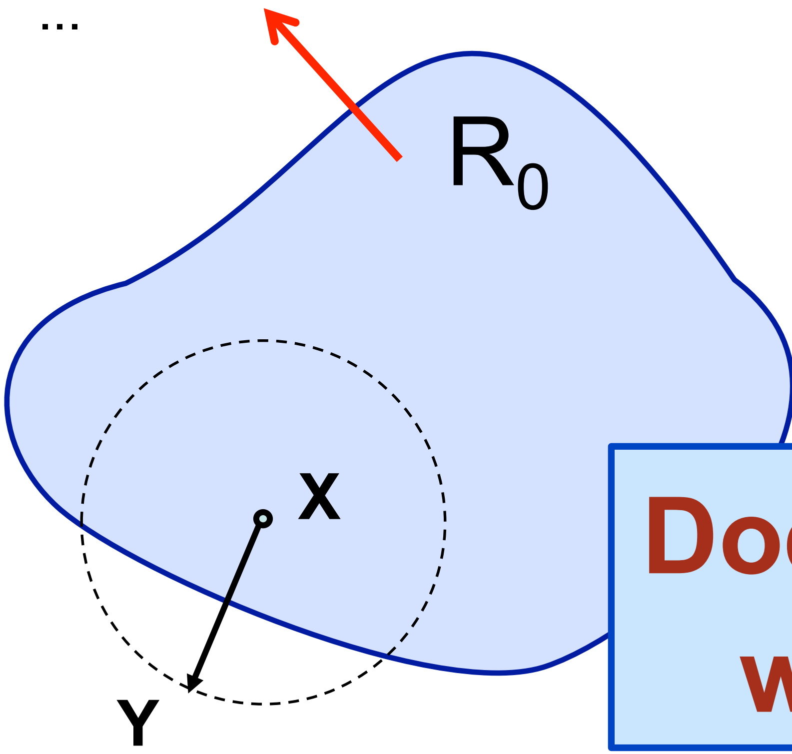
Attacks are also similar: the MF case



- Exit R_0 under a distortion constraint
- Exit R_0 with the minimum distortion
- If R_0 is known, then look for optimal solution
  - **Rarely done in MF**
- If R_0 is not known: oracle attacks are possible
  - **Gradient descent**
  - **Blind attacks (BNSA)**

Attacks are also similar: the MF case

- Images taken by camera X
- Doubly compressed images
- ...



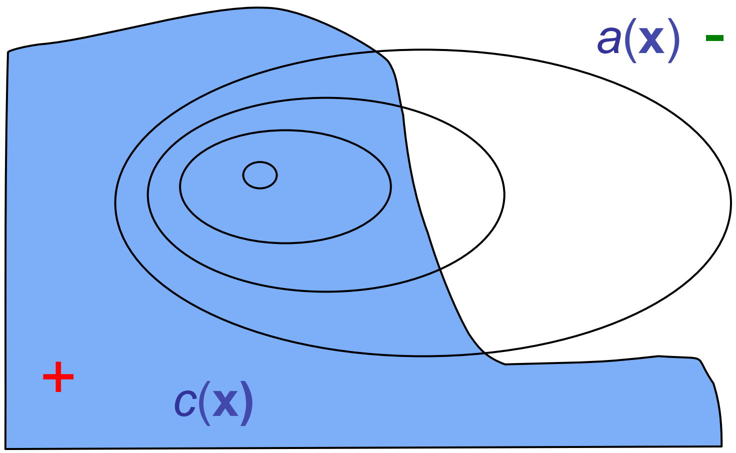
- Exit R_0 under a distortion constraint
- Exit R_0 with the minimum distortion
- If R_0 is known, then look for optimal solution
  - **Rarely done in MF**

Doesn't it resemble watermarking ?

SPAM filtering

Now consider the famous (not in our community) ACRE* attack:

Adversarial Classification Reverse Engineering



Adversary's Task:
Minimize $a(\mathbf{x})$ subject to $c(\mathbf{x})$ = -

The adversary does not know $c(\mathbf{x})$!

ACRE assumes that a(x) is known and that a polynomial number of queries to the decisor c(x) are possible

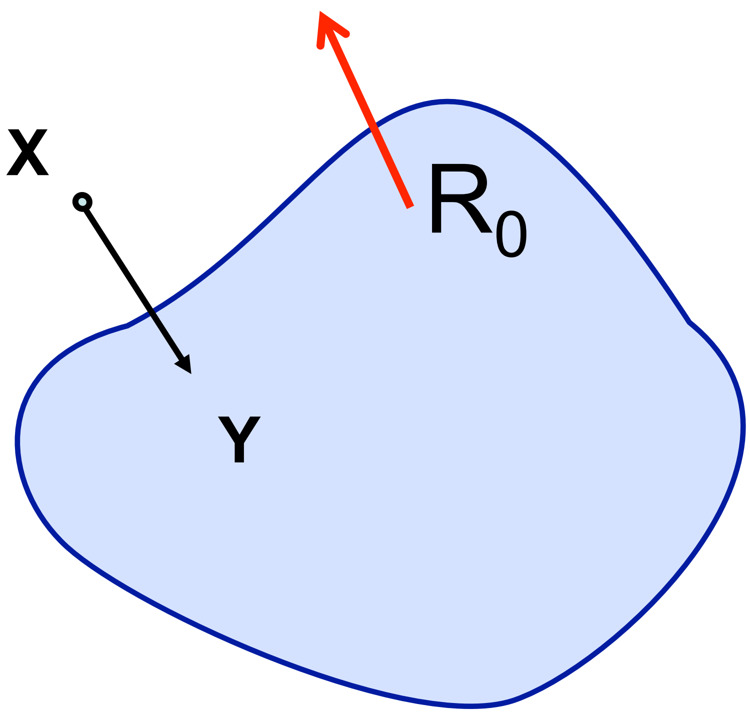
* D. Lowd, C. Meek. "Adversarial learning" Proc. of the 11th ACM SIGKDD Int. Conf. on Knowledge discovery in data mining. 2005.

Adversarial machine learning

- It may come as a surprise (it was surely a surprise to me), to know that a field named adversarial machine learning exists (since about 2004), studying problems very similar to those our community has been facing with in the same period
- A new twist is introduced: **the attacker may interfere with the learning phase**
- For a good introduction to this field I suggest:
  - M. Barreno, B. Nelson, A. D. Joseph, J.D. Tygar, “The security of machine learning”, Mach Learn (2010) 81: 121–148

Hill climbing in biometrics ...

- Region with valid biometric templates
- Verification requires that f(x) > T for some function f() and threshold T



- In masquerade attacks, the attacker aims at finding a valid biometric template
- Distortion is not an issue

- If R_0 is not known
  - **Gradient-based methods:** possible if f(x) is revealed
  - **Blind attacks:** not possible

And all the others ...

- **Reputation systems**: build a fair user profile starting from a malevolent scoring pattern
- **Cognitive radio**: provide fraudulent measurements by camouflaging them as trustworthy data
- **Traffic monitoring**: shape the request profile of a traffic monitoring system so to mimic innocuous requests
- **Fingerprinting**: modify multimedia documents so to pass copyright controls

Isn't a general theory of adversarial hypothesis testing advisable ?

Not only binary hypothesis testing

- From binary to multiple hypothesis: classification
- Pattern recognition
  - Biometrics (identification), speech recognition, machine vision, content-based image information retrieval, multimedia fingerprinting
- Adversarial learning
- Multiple players
  - Collaborative filtering, reputation systems, social aspects
- Communication-like scenarios
  - Watermarking, traitor-tracing
  - Communication in the presence of jamming

Where do we start from ?

Adv-SP and Game-Theory: a good fit

Vast amount of results to rely on

Clear definition of players

Clear definition of goals

Optimality criteria (equilibrium notion)

Modelling social interactions

Definition of possible moves

Several game structures are possible

Clear definition of constraints

Game Theory in a nutshell

Two-player game

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

$S_1 = \{s_{1,1}, s_{1,2} \dots s_{1,n1}\}$ Set of strategies available to first player

$S_2 = \{s_{2,1}, s_{2,2} \dots s_{n2}\}$ Set of strategies available to second player

$\mathrm{u}_1(s_{1,i}, s_{2,j})$ Payoff of first player for a given profile

$\mathrm{u}_2(s_{1,i}, s_{2,j})$ Payoff of second player for a given profile

Competitive (zero-sum) game

$\mathrm{u}_1(\cdot,\cdot) = -\mathrm{u}_2(\cdot,\cdot)$

Sequential vs strategic vs multiple moves games

Equilibrium

Optimal choices

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

Nash equilibrium

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

$$u_1(s_1^*, s_2^*) \geq u_1(s_1, s_2^*) \quad \forall s_1 \in S_1$$

$$u_2(s_1^*, s_2^*) \geq u_2(s_1^*, s_2) \quad \forall s_2 \in S_2$$

A possible GT-model for binary HT

Assumptions

- Two players: the **defender** (**D**) and the **attacker** (**A**)
- Two sources P_X and P_Y known to **D** and **A** (*relaxed later on*)
- Task of **D**: decide whether a given sequence has been drawn from P_X (H_0)
- Task of **A**: modify a sequence drawn from P_Y so that it looks as if it were drawn from P_X subject to a distortion constraint

Several variants

- Sequential vs strategic game
- **A** attacks both sequences drawn from P_X and P_Y
- **A** attacks any sequence without knowing which source produced them

A Neyman-Pearson version of the game

Strategies and payoff

$$S_D = \left\{ \Lambda_0 : P_X(x^n \notin \Lambda_0) \leq P_{fp} \right\} \qquad \Lambda_0 = \text{acceptance region for } \mathrm{H}_0$$

$$S_A = \left\{ f(y^n) : d(y^n, f(y^n)) \leq nD \right\} \quad D = \text{distortion constraint}$$

$$\mathrm{u}_D(\Lambda_0, f) = -P_{fn} = -\sum_{y^n : f(y^n) \in \Lambda_0} P_Y(y^n)$$

Variants

- Bayesian version: known a-priori probabilities, risk minimization
- Alternative strategy for **A**: induce an error, minimize distortion

Insights gained by GT modelling

- Attacking a fixed defender's strategy fails to recognize the game nature of the problem
  - Even if **D** moves first and **A** knows **D**'s move ...
  - ... **D** should choose its strategy knowing that **A** will attack it
  - Max-min problem
- Interesting questions for the sequential version of the game: what does **A** know about **D**'s move ?
- Strategic version of the game: look for Nash equilibrium
- Find the equilibrium point(s) would permit to:
  - Know optimum strategies
  - Compute the payoff at the equilibrium (who wins the game)
  - Benchmark practical solutions

Solving the binary-HT game*

Idea (1): asymptotic version of the game

$$S_D = \left\{ \Lambda_0 : P_{fp} \leq 2^{-\lambda n} \right\}$$

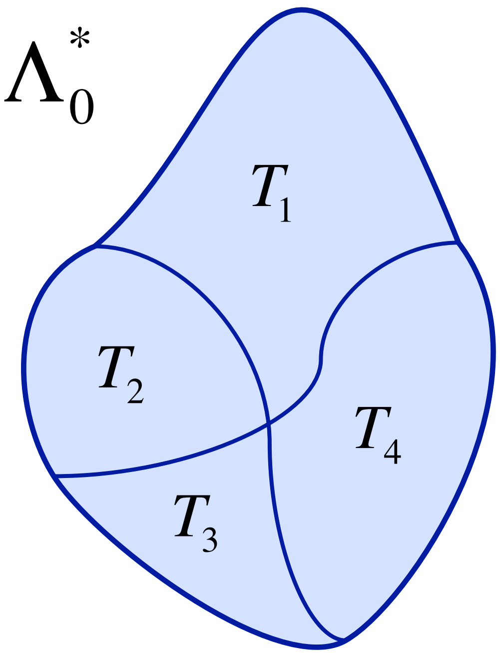
$$S_A = \left\{ f(y^n) : d(y^n, f(y^n)) \leq nD \right\}$$

$$\mathrm{u}_{\mathrm{D}}(\Lambda_0, f) = -P_{fn}$$

Idea (2): D relies only on first order statistics

- M. Barni, "A game theoretic approach to source identification with known statistics", Proc. ICASSP'12, IEEE Conf. on Acoustics, Speech and Signal Processing, Kyoto, 2012.

Solution based on method of types



A type class T is a set of sequences with the same empirical pdf (T)

First order statistics analysis -> Λ_0 is a union of type classes (or union of types)

The asymptotic probability of a type class T under a certain pdf P_X is

$$P_X(T) \approx 2^{-nD(T \| P_X)}$$

The optimum acceptance region contains only and all the type classes for which $D(T \| P_X) < \lambda$

First result

Nash equilibrium for the game

$$\Lambda_0^* = \left\{ x^n : D(P_{x^n} \| P_X) < \lambda - |\chi| \frac{\log(n+1)}{n} \right\} \quad regardless\ of\ P_Y$$

$$f^*(y^n) = \operatorname*{argmin}_{z^n : d(z^n, y^n) \leq nD} D(\hat{P}_{z^n} \| P_X)$$

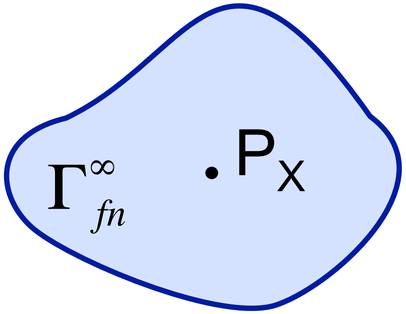
Remark: The optimum strategy of the **D** depends neither on P_Y nor on **A**'s strategy

Second result: who wins the game ?

Distinguishable sources (in adversarial setting)

Given P_X λ and D, we can define a region Γ_{fn}^{∞} such that

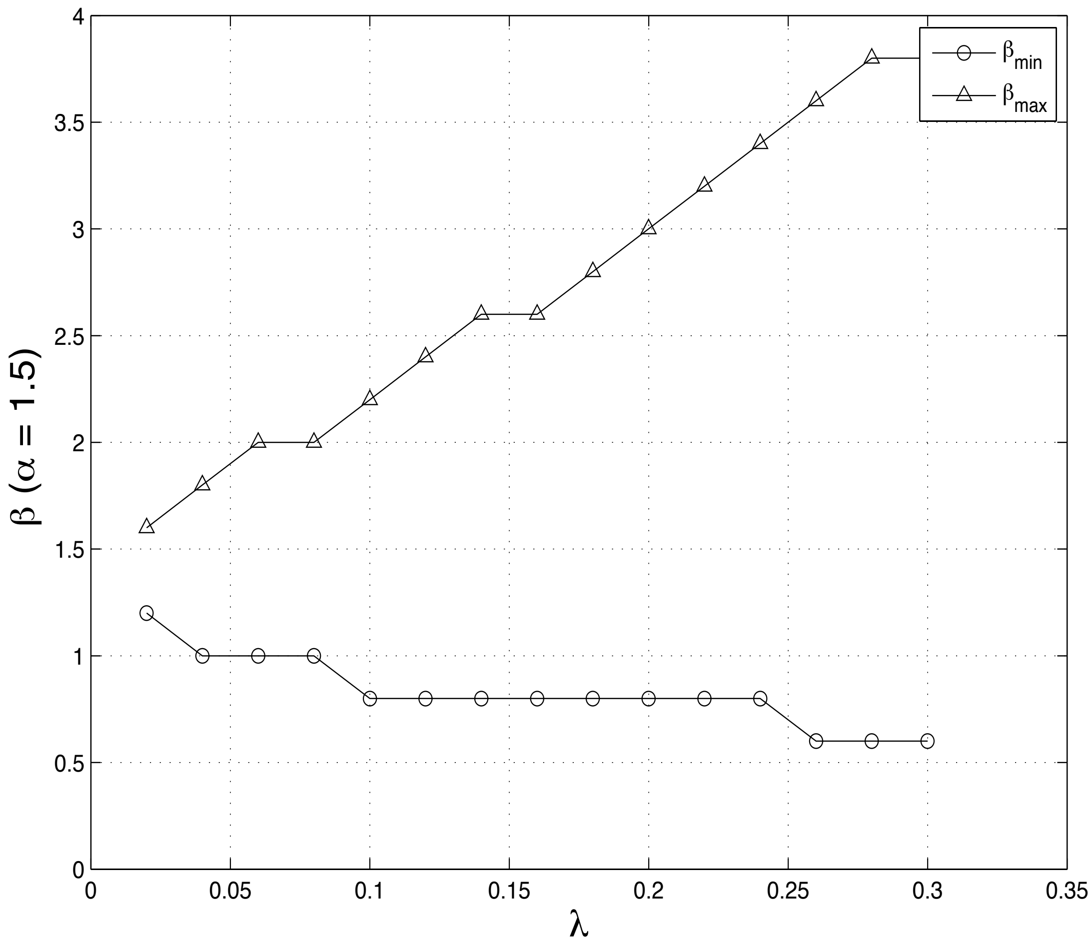
$$\begin{cases} \text{if } P_Y \in \Gamma_{fn}^{\infty} \text{ then } P_{fn} \rightarrow 1 \\ \text{if } P_Y \notin \Gamma_{fn}^{\infty} \text{ then } P_{fn} \rightarrow 0 \end{cases}$$



By letting λ -> 0 we obtain the region of distinguishable sources for a certain distortion level D.

Let D_{max} = maximum value of D for which P_X and P_Y are distinguishable, we can say that P_X and P_Y are distinguishable up to an attack of power D_{max}

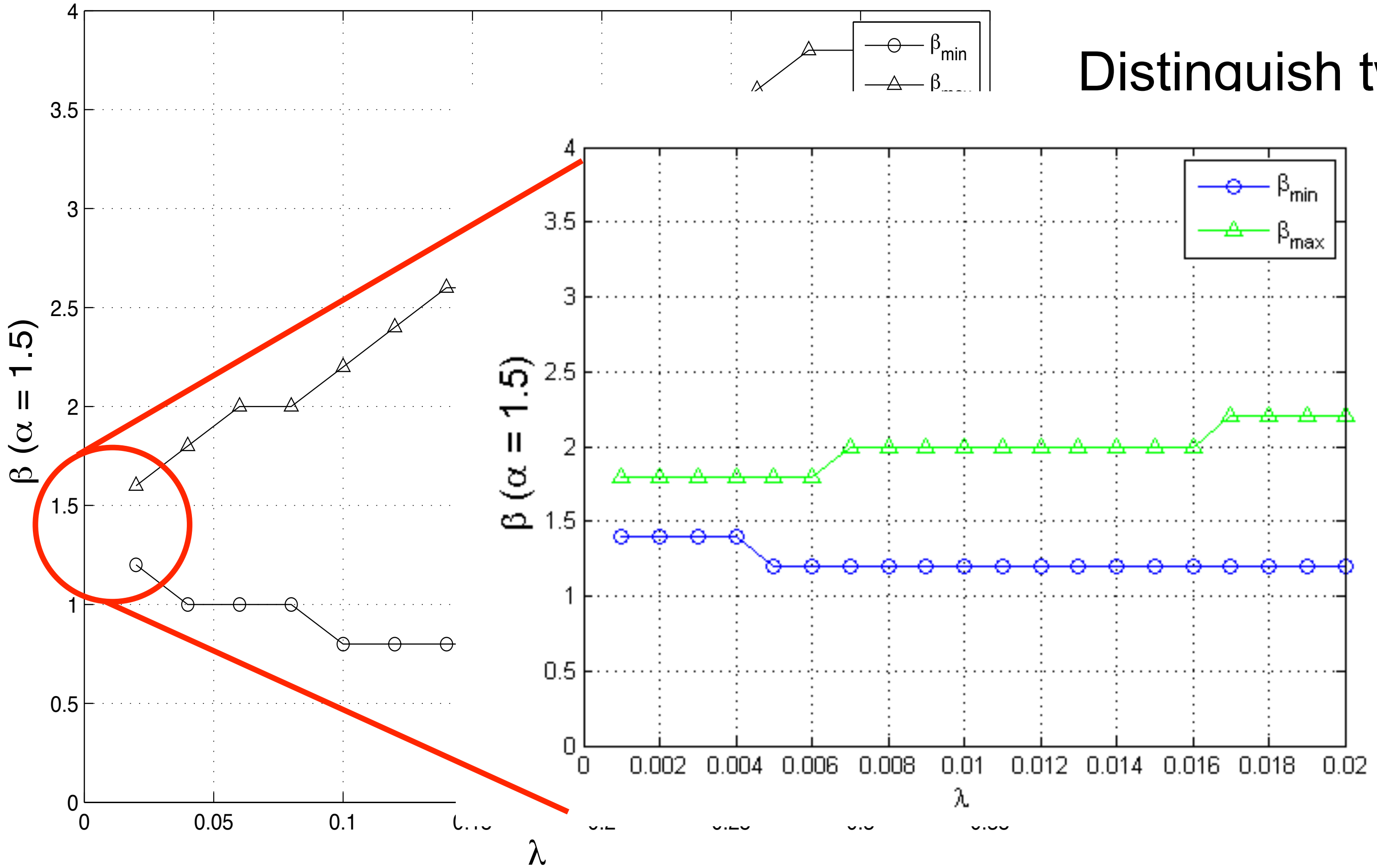
A numerical example



Distinguish two exponential sources with different decay factors (α, β)

When λ approaches 0, the distinguishable and non-distinguishable pdf's are determined

A numerical example



Distinguish two
urces
ecay

ches 0,
able and
able
mined

Security vs robustness

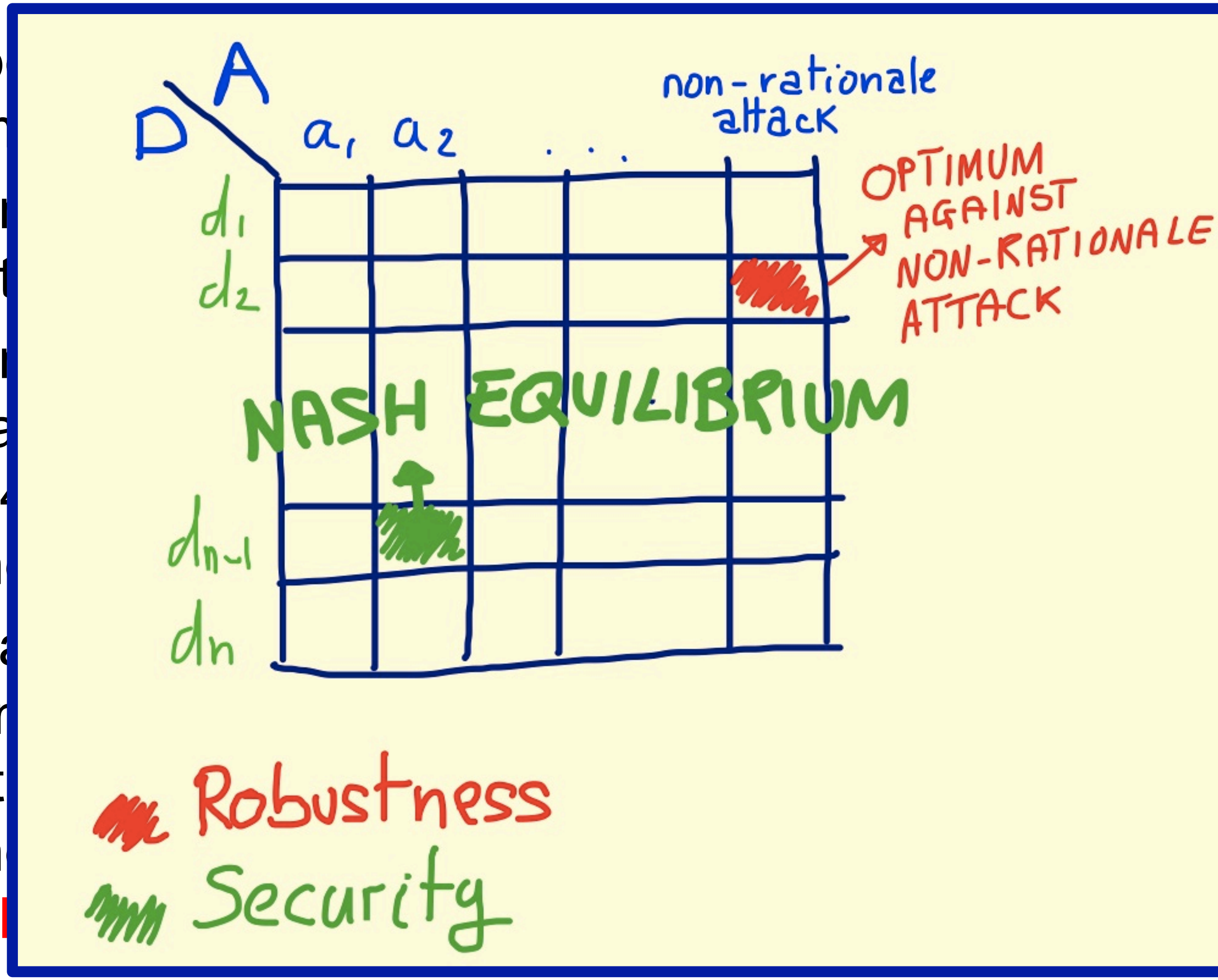
- The capability of distinguishing two sources in the presence of a non-rationale attack, e.g. noise additions, should be regarded as **robustness**
- The capability of distinguishing two sources at the Nash equilibrium of a game, should be regarded as **security** against a certain type of adversary (e.g, an adversary with unlimited computing power) and under certain conditions (game structure)

An oversimplified example

- Suppose we want to distinguish two Bernoulli sources with parameters *p* and *q* (say p = 0.9, q = 0.5)
- As long as p ≠ q you can do that and both error probabilities tend to 0 exponentially fast
- Assume that the source output passes through a BSC with error probability (Hamming distortion) = 0.2. At the output we have: p' = 0.74, q' = 0.5, then the two sources continue to be distinguishable: **the test is robust for D = 0.2**
- If an attacker is allowed to modify the output sequences with a maximum average distortion = 0.2, then he can act in such a way that p' = 0.7, q' = 0.7. Then the two sources can not be distinguished: **the test is secure only against attacks for which D < (p-q)/2**.

An oversimplified example

- Supp… with paran…
- As lo… bilities tend t…
- Assu… with error proba… e have: p' = 0.74… distin…
- If an… es with a maxin… such a way t… ot be distin… **s for whic…**



Binary HT with training sequences

Given two sources P_X and P_Y and two pairs of training sequences $\left(t^N_{X,A}, t^N_{Y,A}\right), \left(t^N_{X,D}, t^N_{Y,D}\right)$

Given a test sequence x^n decide if x^n was drawn from P_X or P_Y

Strategies and payoff

$$S_D = \left\{ \Lambda_0 : \max_{P_X} P_X(x^n \notin \Lambda_0) \leq P_{fp} \right\} \qquad \Lambda_0 = \text{acceptance region for } \mathrm{H}_0$$

$$S_A = \left\{ f(y^n) : d(y^n, f(y^n)) \leq nD \right\} \quad D = \text{distortion constraint}$$

$$\mathrm{u}_D(\Lambda_0, f) = -P_{fn} = -\sum_{y^n : f(y^n) \in \Lambda_0} P_Y(y^n)$$

Variants

$$\left(t^N_{X,A}, t^N_{Y,A}\right) = \left(t^N_{X,D}, t^N_{Y,D}\right)$$

$$\left(t^N_{X,A}, t^N_{Y,A}\right) \text{ independent from } \left(t^N_{X,D}, t^N_{Y,D}\right)$$

Binary HT with training sequences

Given two sources P_X and P_Y and two pairs of training sequences $\left(t^N_{X,A}, t^N_{Y,A}\right), \left(t^N_{X,D}, t^N_{Y,D}\right)$

Given a test sequence x^n decide if x^n was drawn from P_X or P_Y

For a discussion on the optimal strategies of the Binary-HT game with training sequences visit my poster on wednesday

Strategie

$S_D = \left\{ \Lambda_0 : \right.$ … I_0

$S_A = \left\{ f(y^n \right.$ …

$u_D(\Lambda_0, f)$ … $y^n : f(y^n) \in \Lambda_0$

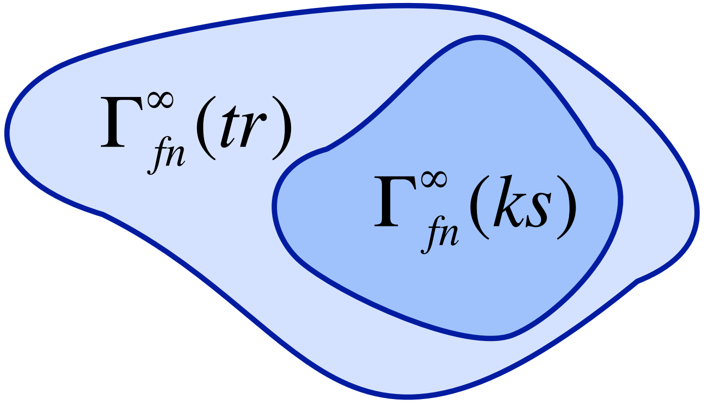
Variants

$\left(t^N_{X,A}, t^N_{Y,A}\right) = \left(t^N_{X,D}, t^N_{Y,D}\right)$

$\left(t^N_{X,A}, t^N_{Y,A}\right)$ independent from $\left(t^N_{X,D}, t^N_{Y,D}\right)$

New research cues

- Can we define (in-)distinguishability as for known sources ?
- Yes we can, and ...



- By assuming that **A** can alter (part of) the training sequence used by **D** we move towards the **adversarial learning** scenario
- By assuming independent training sequences we move towards **key-based security**

How can we improve security ?

- What can the defender do to improve security ?
  Complexity enters the picture
- The defender may move to higher order statistics
- The game theoretic analysis still works, and the attacker may still go for the optimum attack ... but
- The optimum attack could be **computationally unfeasible**

- **Robustness could be lost to gain security !!!**

From theory to practice: oracle attacks

- A similar trajectory could be followed to cope with oracle attacks (similarity with watermarking)
- By complicating the decision boundary, exerting an optimal attack may become too complex
  - Insertion of local traps
  - Randomization, fractal boundaries
  - **Again we exchange robustness for security**
- Try to discover if an attacker is at work and switch from robustness to security ... **Yet another game** !!!!

In summary ... there's a lot to work on

- **Theory**
  - A whole theory to develop
  - Depart from binary HT to more complicated and realistic scenarios
- **Practice**
  - Stop with cat and mouse loop
  - Develop adversary-aware forensics tools
  - Security against computationally bounded attackers
- **Sinergy**
  - Go beyond MF, steganography, watermarking
  - Exploit synergies with contiguous fields

I look forward to seeing you working on Adv-SP

Thank you for your attention