Cybersecurity

Steganography, Steganalysis, Watermarking

Mauro Barni
University of Siena

Steganography and Steganalysis

Steganography: hidden communication

Steganography is the art-science of communicating hiding the existence of the communication

In contrast to cryptography, where the enemy is allowed to intercept and modify messages without being able to violate the security ensured by a cryptosystem, **the goal of steganography is to hide messages inside other harmless messages** in a way that does not allow the enemy to even detect the presence of a second secret message

Cryptography

Key: s=v, i=e, t=y …

Bob
See you at 8 p.m.

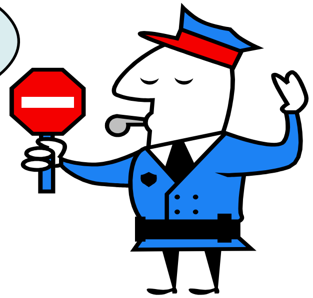
encryption

fiiteonagezto ziloorriadda

decryption

Alice
See you at 8 p.m.

Don't understand but suspicious



In some cases the very existence of a message is enough to raise a suspect

Steganography

Key: 6-3, 1-2, 2-2, 0, 4-5, 2-5, 11-1

Bob

See you at 8 p.m.

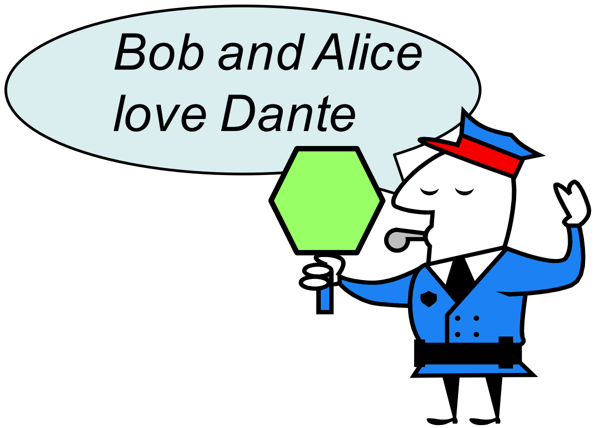
Steg

Nel mezzo del cammin di nostra vita
mi ritrovai per una selva oscura
ché la diritta via era smarrita

Decode

Alice

See you at 8 p.m.



Steganography hides the very presence of the message into an innocuous host

In a more flexible way

"My friend Bob: Until yesterday I was using binoculars for stargazing. Today I decided to try my new telescope. The galaxies in Leo and Ursa Major were unbelievable! Next, I plan to check out some nebulas and then prepare to take a few snapshots of the new comet. Although I am satisfied with the telescope, I think I need to purchase light pollution filters to block the xenon lights from a nearby highway to improve the quality of my pictures. Cheers, Alice."

Take initial letters:
mfbuyiwubfstidttmnttgilaumwuniptcosnatpttafsotncaiaswttitintplpftbtxlfanhtitqompca

Filter with p = 3.141592653689793…-> buubdlupnpsspx

Take the previous letter in the alphabet: ATTACK TOMORROW

Historical notes

- Steganography is as old as the humankind
- Herodotus:
  - Tatooing the head of a shaved slave
  - Writing on wood tablets then covered by wax
- Boccaccio: *Amorosa visione* (acrostic)
- Microdot technology: world war I and II
- Capt. Denton emprisoned by Vietnamese
- Korchnoi vs Karpov
- Invisible ink

Some examples: acrostic

News Eight Weather: Tonight increasing snow. Unexpected precipitation smothers eastern towns. Be extremely cautious and use snowtires especially heading east. The highway is, not knowingly, slippery. Highway evacuation is suspected. Police report emergency situations in downtown ending near Tuesday.

Newt is upset because he thinks he is President.

Some examples: word shifting

sentence 1:

We explore new steganographic and cryptographic algorithms and techniques throughout the world to produce wide variety and security in the electronic web called the Internet.

sentence 2:

We explore new steganographic and cryptographic algorithms and techniques throughout the world to produce wide variety and security in the electronic web called the Internet.

Some examples: word shifting

By overlapping S1 and S2, the following sentence results

We ***explore*** *new steganographic and cryptographic algorithms and techniques throughout* ***the world*** *to produce* ***wide*** *variety and security in the electronic* ***web*** *called the Internet.*

explore the world wide web

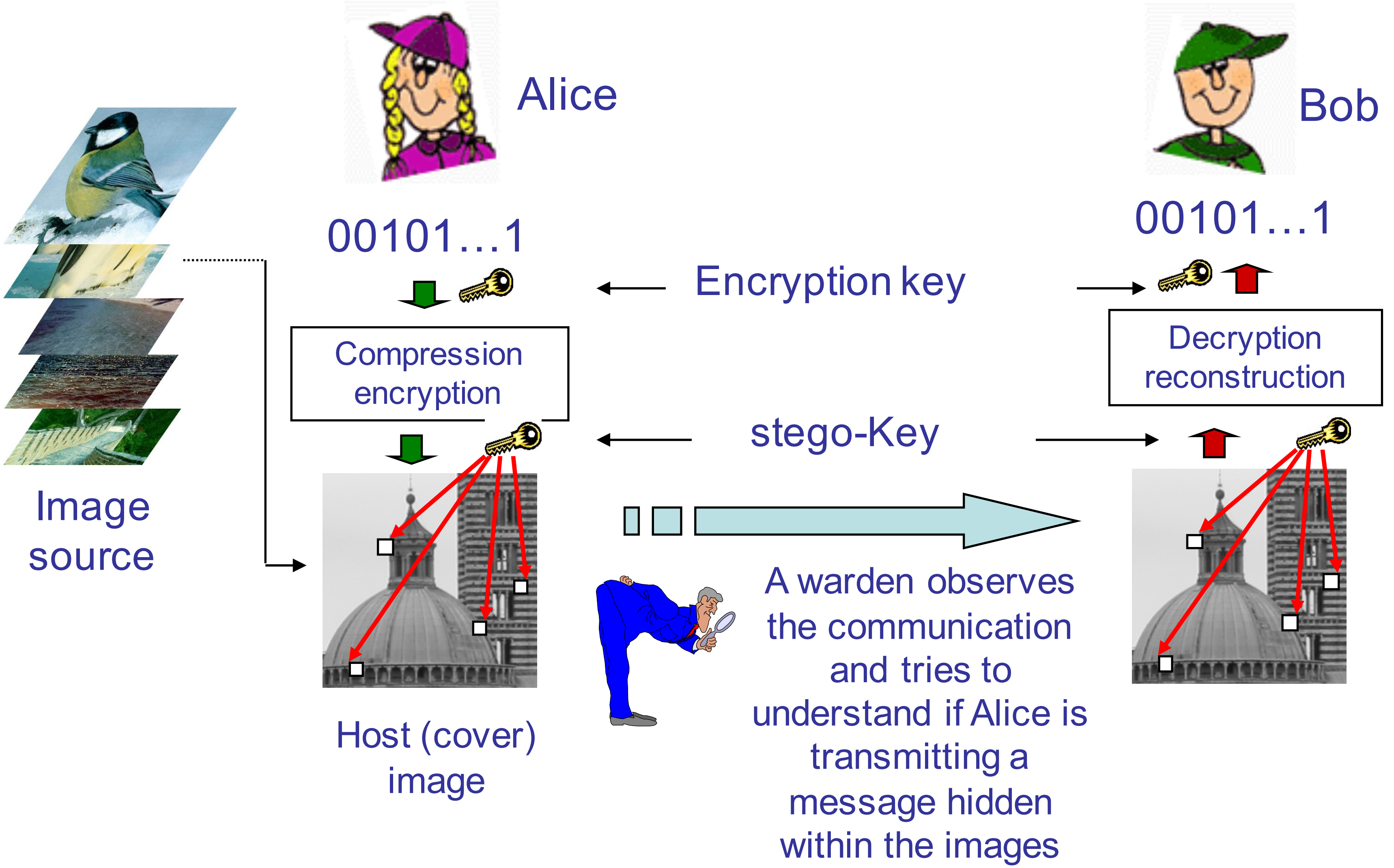
Steganography in the digital age

- Renewed interest starting from nineties
- Enabling technologies
  - Wide band communication channels
  - Diffusion of multimedia contents
  - Possibility of using automated steganographic techniques with high payloads
- Motivations
  - Espionage, terrorism
  - Dissidents, freedom of expressing own opinions against censorship
  - Privacy protection – avoid *big-brother* scenarios

Steganalysis

- Complementary motivations pushed researchers to study steganalysis
  - Techniques to reveal the presence of hidden messages (possibly without decyphering them)
- Motivations
  - Intelligence, police
  - Control of public opinion
- Regardless of motivations, the study of steganalysis is necessary to determine the security of steganographic techniques

A rigorous framework: the prisoner problem



Opposite requirements

In steganography designers must face with 2 opposite requirements

(statistical) Invisibility

Capacity (payload)

Perceptual invisibility

The hidden message must remain invisible even after the applications of signal processing tecniques

message

Cover image

LSB of the green channel (original)

LSB of the green channel (stego-image)

The invisibility requirement

- More than perceptual invisibility, we require **statistical invisibility**
- Assumptions on warden behavior
  - Active, **passive**
  - **Kerckhoff's principle**:
    - The warden knows the steganographic algorithm
    - The warden knows the statistics of the image source used by Alice
- Invisibility alone is not sufficient
  - Real life is always more complex than mathematical models (as cryptographers learnt quite soon)

A first choice: hiding domain

- Pixel domain steganography
  - Easy to use
  - High capacity
  - Simple analysis of perceptual visibility
- Compressed domain steganography (JPEG)
  - The message is conveyed by (block) DCT coefficients
  - Wide diffusion of JPEG images
  - Lower security (due to the availability of good statistical models to describe DCT coefficients)
  - Example: F5, OutGuess, Jsteg (most of them are available on the internet)

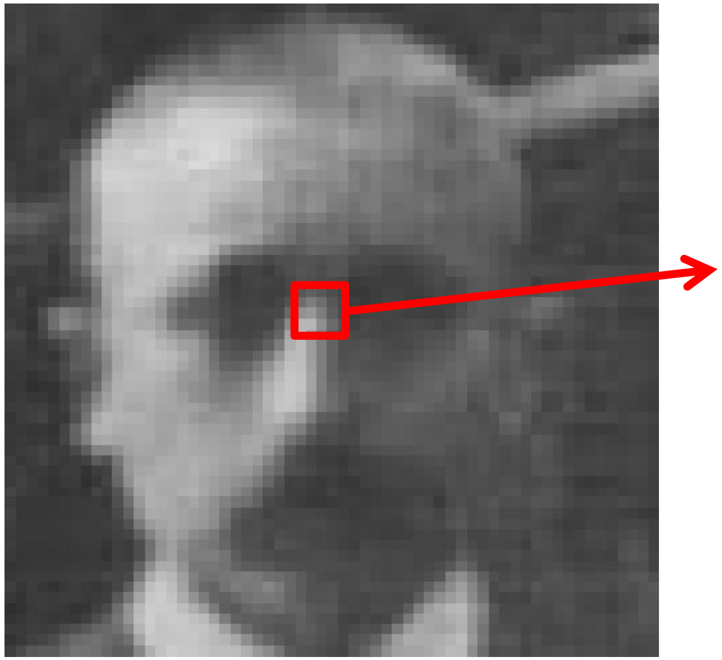
Pixel domain

- The stego-message is hidden in the array of integer numbers a digital image consists of

100	102	104	156	157	190	201	201
100	102	130	120	123	191	199	199
103	105	127	118	125	190	190	188
110	112	112	116	123	131	190	189
101	102	106	102	120	130	191	199
101	104	107	109	134	135	199	220

Frequency domain

- In some cases, for instance with JPEG images, the stego message is hidden into the (block) DCT coefficents of the images



-16	90	37	-17	-1	-2	-2	-1
63	10	-46	-14	12	0	0	2
-2	-9	-5	12	4	-5	-2	1
1	-3	-2	0	-3	-1	1	1
0	-2	-1	-1	0	1	1	-1
0	0	0	0	-1	0	0	0
0	-1	0	0	0	0	0	0
0	-1	0	1	0	0	0	0

Three classes of steganographic algorithms

- Steganography by cover selection
- Steganography by cover synthesis
- Steganography by cover modification

Steganography by cover selection

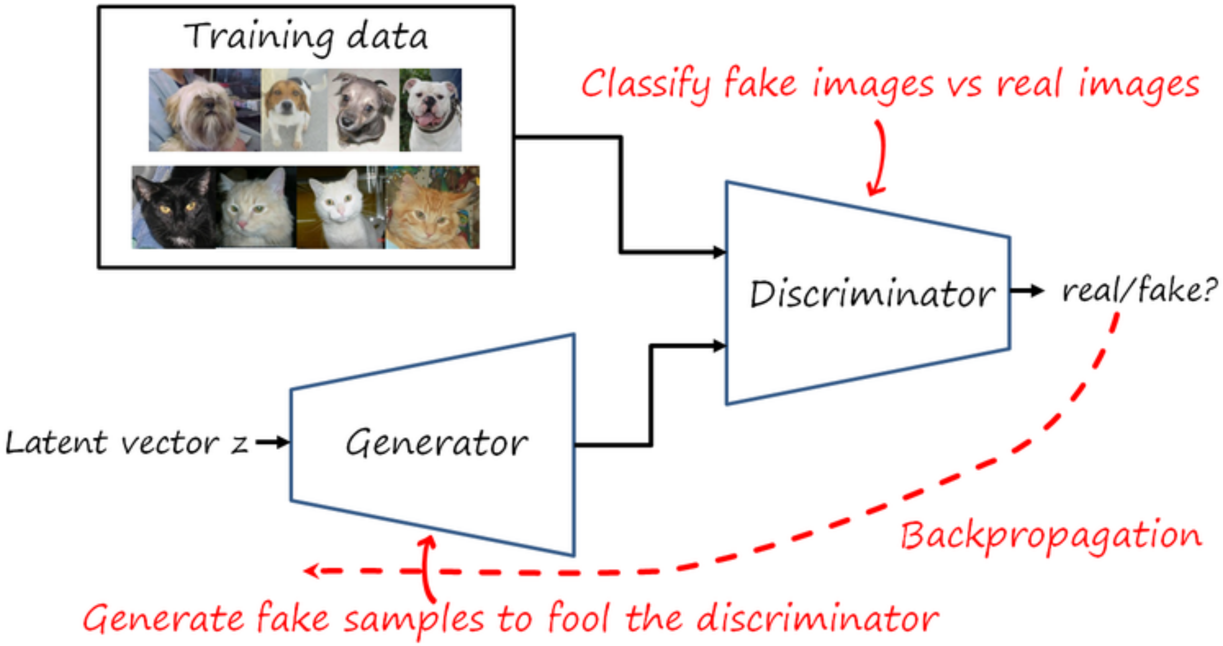
- Alice has a database of images, wherein she chooses the image corresponding to the correct message. The message can be linked to
  - Semantic image content
  - Value of a selected subset of LSB's
  - Image (or subimage) hash
- Pros
  - Almost perfect security
- Cons
  - Very low payload
  - Example: an 8 character message (64 bit) requires a database with - at least - 2^{64} (10^{19}) images

Steganography by cover synthesis

- Alice creates an image on-the-fly conveying the to-be-transmitted message
- Creating a realistic image is not easy. Alice could proceed as follows
  - Alice gathers several shots of the same scene
  - Alice divides the images into blocks. Each block is associated to some message bits (e.g. through a subset of LSB's)
  - Alice builds the final image by properly assembling the blocks from various images
- Pros: good security (problems at block borders)
- Cons: still low payload (too many images needed)

Cover synthesis by AI

- GANs and other generative models proved to be able to generate visually plausible fakes
- **Two CNNs struggling following a Game-theoretic formulation**

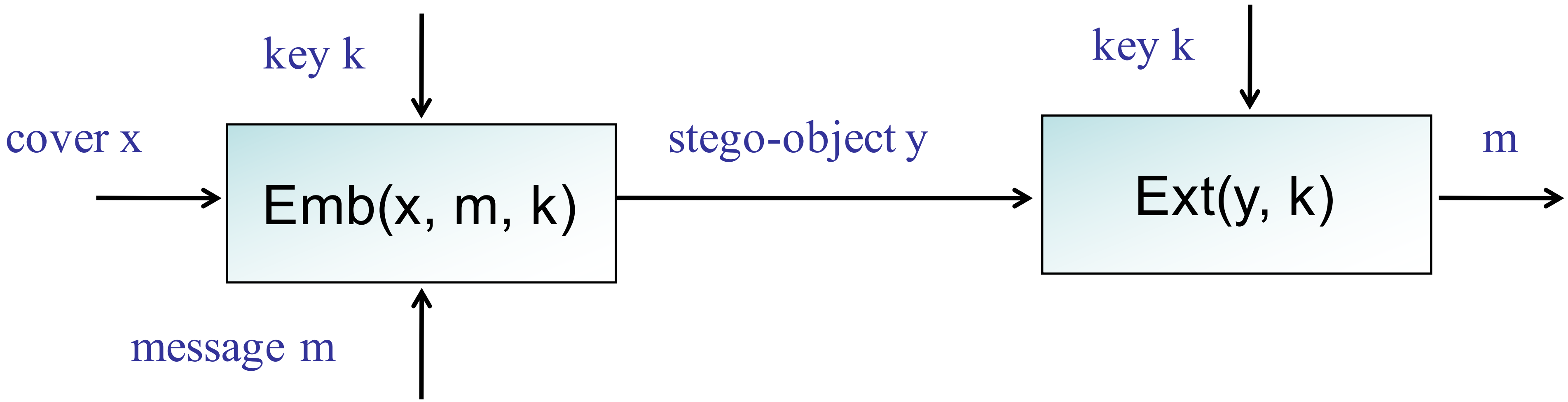


Cover synthesis by AI: examples

- Images generated by GANs can be extremely realistic

Steganography by cover modification

- By far the most common approach
- It allows large payloads, but security must be studied carefully

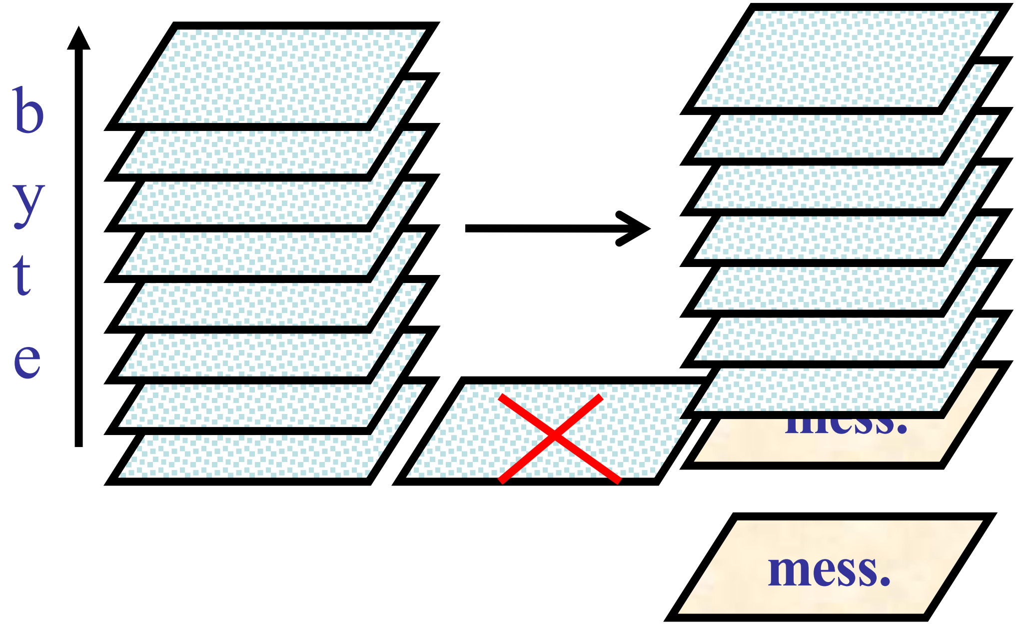


- Payload = $\log_2(|M|)$ / sizeof(x); Distortion = d(x,y) = $\sum_{i=1}^{n}\left(x(i)-y(i)\right)^2$

A detailed example: LSB embedding

The LSB's of the pixels of an image (or the DCT coefficients) are replaced with the stego-message (payload = 1bpp o 1bpnzc)

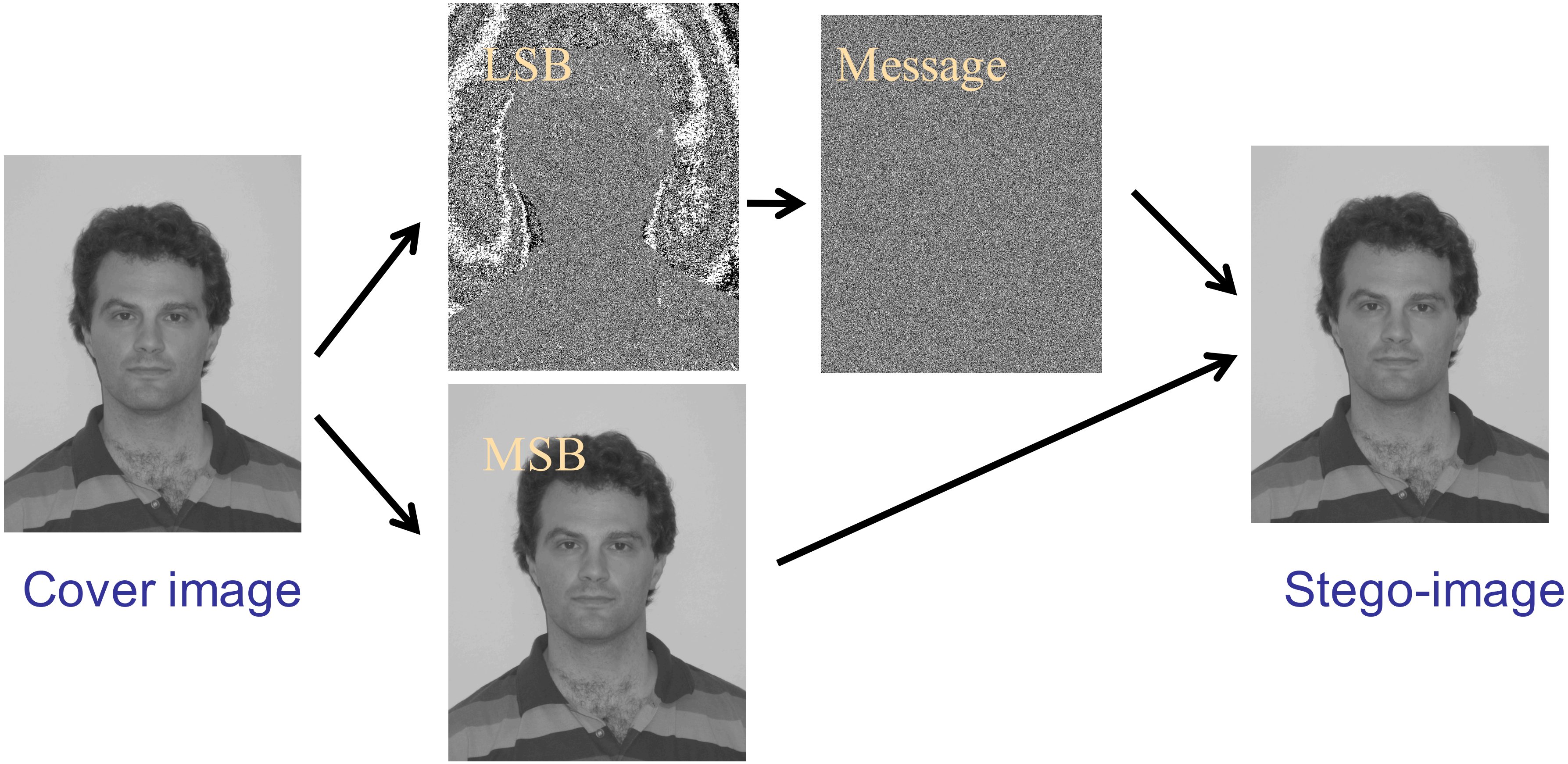
Cover image



Stego image

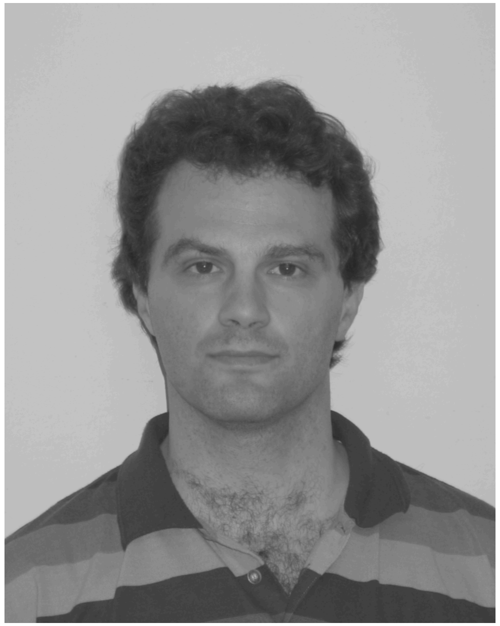
Visual imperceptibility

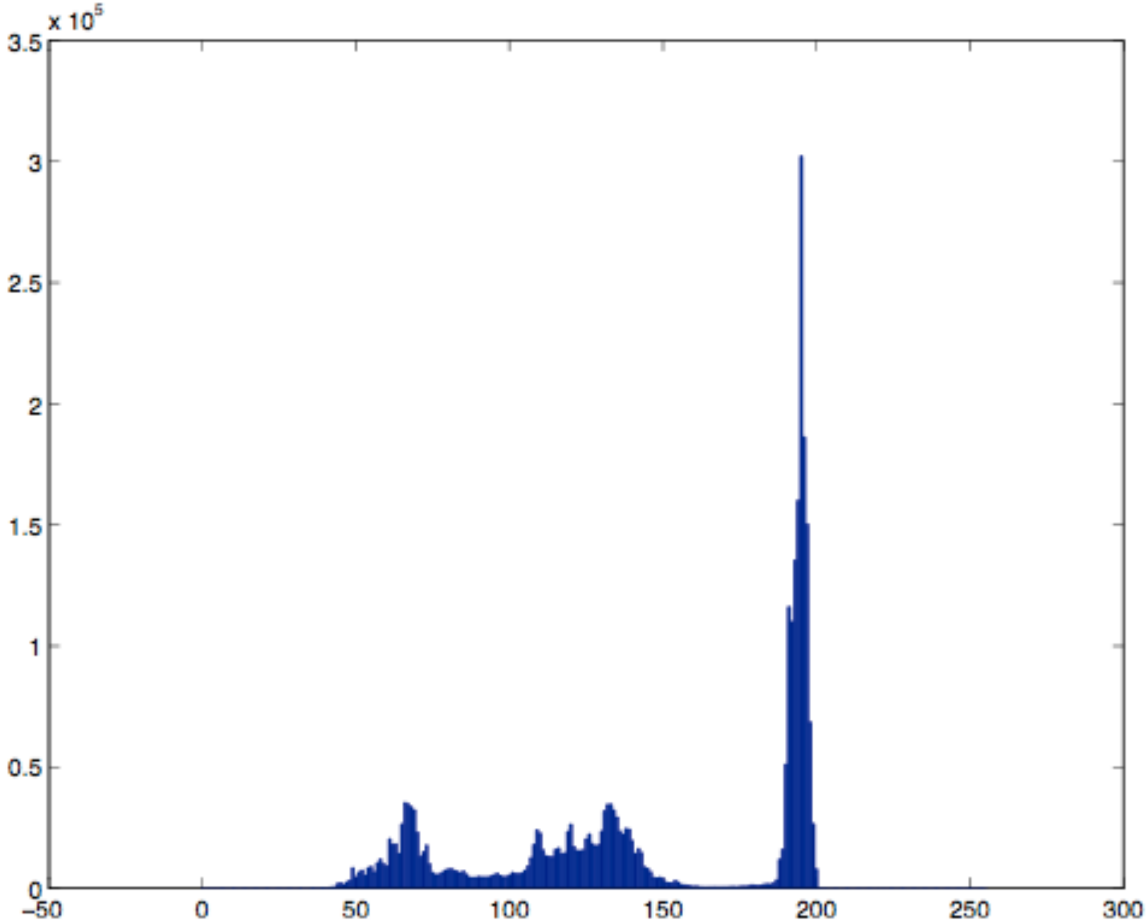
- LSB replacement looks perfect (but is not): the LSB plane of an image is very similar to noise



Attacking LSB replacement

As a matter of fact, steganalysis of LSB replacement steganography is quite easy (at least for high payload)



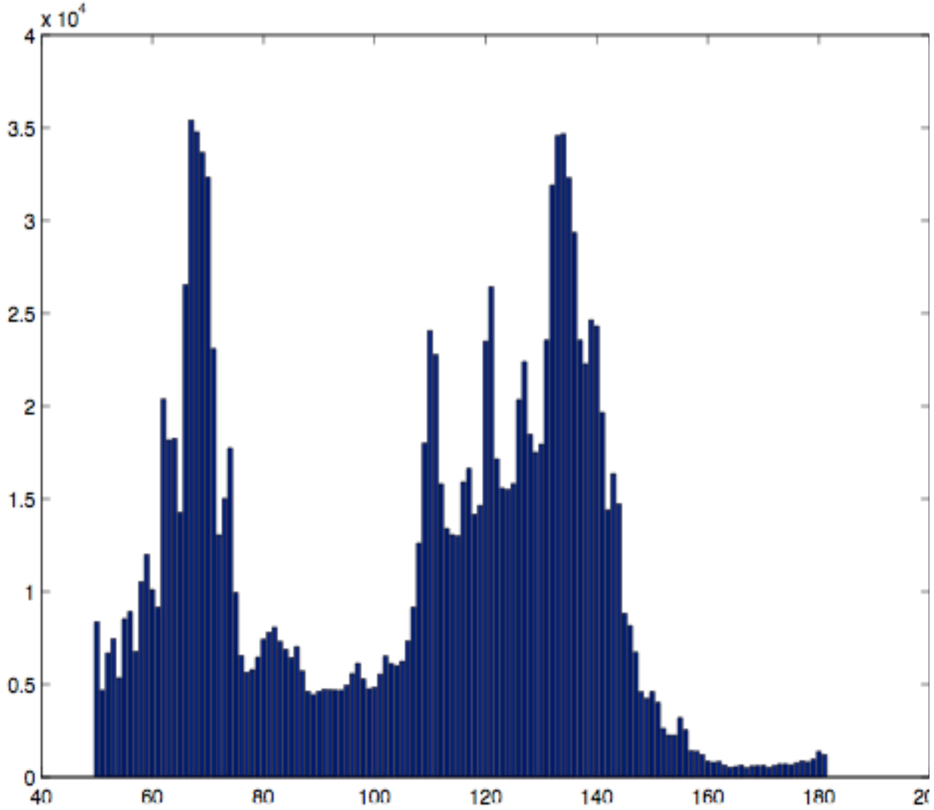


Attacking LSB replacement

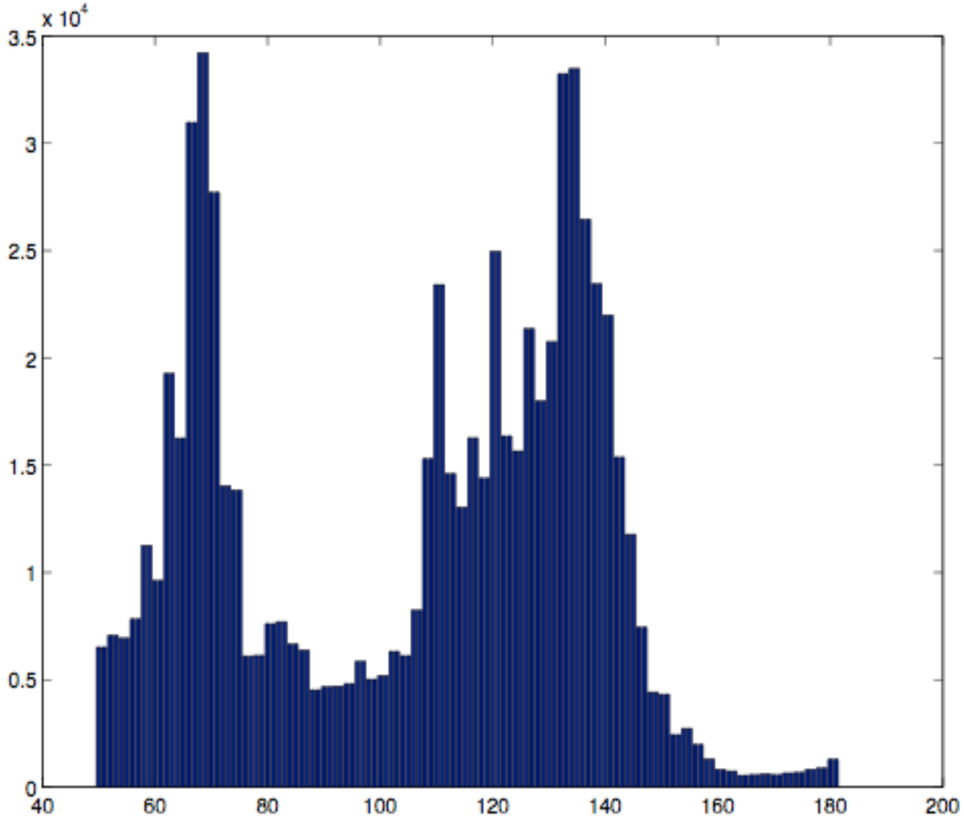
- If x(i) is even we have 01100000 which remains as is or is increased by 1 -> 01100001
- If x(i) is odd we have 01100001 which remains as is or is decreased by 1 -> 01100000
- Consider the pair (0,1): (00000000, 00000001)
- Half of the pixels equal to 0 pass to 1 and half of the pixels equal to 1 pass to 0
- At the end we have about the same number of pixels = 0 e pixels = 1, that is $h_{stego}(0) = h_{stego}(1)$

Attacking LSB replacement

The histogram of stego-images has a very characteristic behaviour



Original histogram



Histogram of stego-image

Countermeasures

- Perfect steganography requires that **all** image statistics are preserved, however
  - It is impossible to derive adequate statistical models of images (slightly better in the DCT domain)
  - It would be too complicated
- Four empirical approaches are used in practice
  - Model-preserving staganography
  - Stochastic modulation
  - Steganalysis-aware steganography
  - Distortion minimization

Model-based steganography

- A model is identified to describe the image source
- Steganography acts in such a way not to modify the model
- Example: statistic restoration
  - The message is inserted in a subset of pixels (or coefficients)
  - The other pixels are modified so to restore the statistical model, e.g. the histogram
  - OutGuess -> works in this way in the DCT domain
- Nearly perfect security as long as the steganalysis relies only on the adopted statistical model (in practice this is never the case)

Stochastic modulation

- It simulates the noise added to the image during the acquisition phase
- Steganography works by adding a noise that resembles acquisition noise
  - Thermal noise
  - Quantization noise
  - PRNU
- It allows rather high payloads (0.8 bpp)

Steganalysis-aware steganography

- The steganographer acts in such a way to eliminate (reduce) the artefacts exploited by the steganalyzer
- Example: ±1-steg
  - If the LSB is the correct one doesn't do anything
  - If the LSB is wrong, add or subtract 1 randomly
  - Observation: ±1steg does not modify only the LSB
    - 01111111+1=10000000
- Security increases dramatically since the histogram does not change significantly (convolution)
- F5 uses ±1-steg in the frequency domain

Distortion (impact) minimization

- Most modern approach
- Define a cost function
  - How much does it cost to modify a certain pixel ? Say ρ(i)
  - Overall cost = $\sum_{i=1}^{n} \rho(i)[x(i) - y(i)]^2$
- Identify an embedding rule which minimizes the embedding cost
  - F5 is optimum from this point of view (DCT domain)

Typical payloads

- Payload
  - from 0.1 to 0.5 bpp in the pixel domain
    - 1000x1000 image => ~ 40Kbyte
  - Up to 0.8 bpnzc in the DCT domain
    - The actual payload depends on the image content. A realistic value is around 20Kbyte for a 1000x1000 image

Steganalysis

- The application scenario is of the outmost importance together with the information available to the warden
  - Blind vs targeted steganalysis
  - Knowledge of cover image statistics
  - Knowledge of payload

Steganalysis = hypothesis test

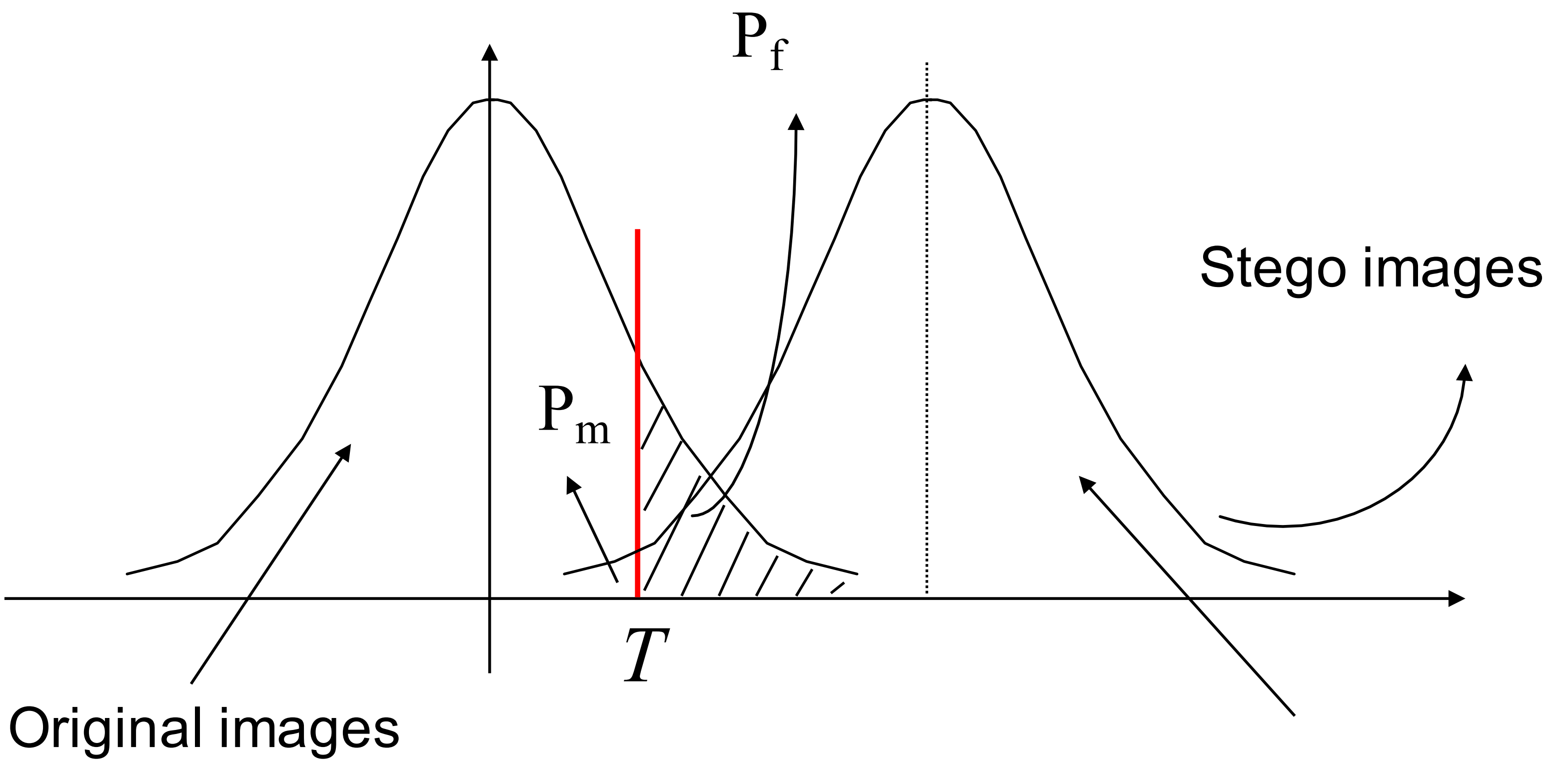
- Rigorous formulation
- Observables: $\mathbf{y} = \{y_1, y_2 \dots y_N\}$
  - Image pixels, audio signal samples, etc ...
  - Often the analysis relies on some functions of y (features) so to simplify the problem
- Two alternative hypothesis
  - H_0 : **y** does not contain a hidden message
  - H_1 : **y** contains a hidden message
- Optimum decision with respect to a certain criterion

Steganalysis = hypothesis test

- Bayes criterion
  - Minimization of overall error probability
  - Difficult to apply since a-priori probabilities are not known
- Neyman-Pearson criterion
  - False alarm probability
    - Decide in favour of H_1 when H_0 holds
  - Missed detection probability
    - Decide in favour of H_0 when H_1 holds
  - N-P: minimize P_m for a given (maximum) P_f
- In steganalysis we must first fix P_f and then decide how to use the result of the test

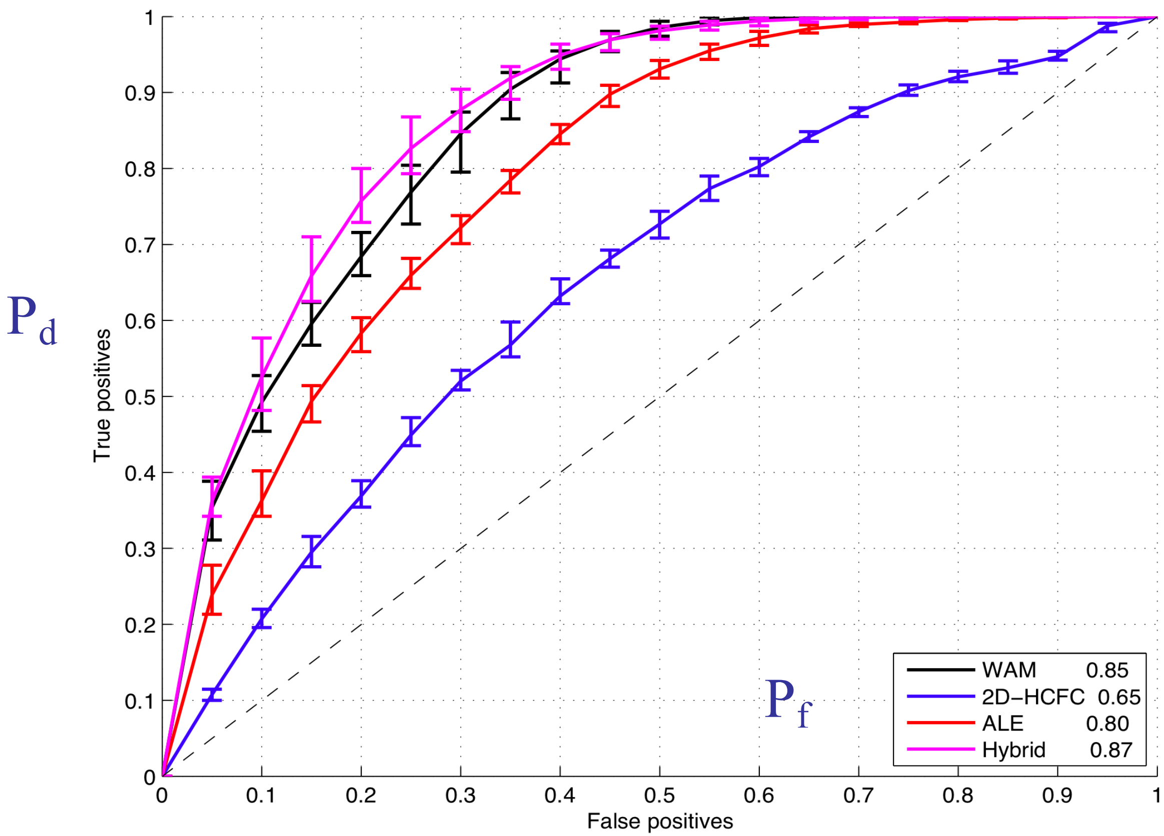
Example

Let us assume that the test relies on a single statistics with known pmf (Gaussian) under both H_0 and H_1



ROC curve

For any value of P_f (threshold) we find a P_m. The plot showing P_d = 1-P_m as a function of P_f is called ROC curve

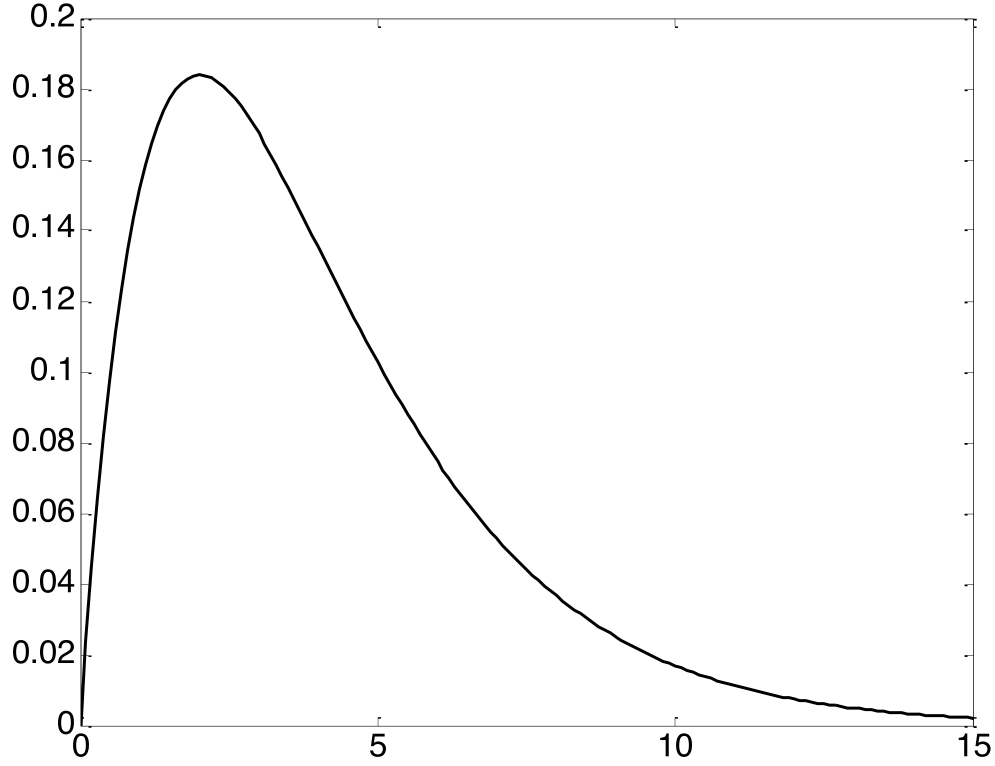


The goodness of a steganalyzer is evaluated by means of the ROC curve or its AUC.

Perfect security requires that performance are equal obtained by means of a random guess (diagonal ROC, AUC = 1/2).

Example

- Let us assume that the pdf of the image source is known. In this case the steganalyzer ca
- Divide the pdf in several
- Compute how many time let us indicate such valu
- Given the pdf let us indic falls in the i-th bin



$$\chi^2 = \sum_{i=1}^{n} \frac{\left(n_i - np_i\right)^2}{np_i}$$

- High values are taken as an evidence in favour of H_1

Choice of statistics (feature)

- In targeted seteganalysis we use few ad-hoc statistics
- Example: LSB replacement steganalysis

Given an image and its histogram, we can use a Chi-square test in which the assumed pmf is

$$h_{Hp1}(2k) = h_{Hp1}(2k+1) = \frac{h(2k) + h(2k+1)}{2}$$

Such a test reveals the presence of LSB steganography (at 1 bpp) with great accuracy. Steganalysis is obviously more difficult at low payloads

Choice of statistics (feature)

- With blind steganalysis everything is more difficult
- If source statistics are known we can still use targeted features
- Otherwise
  - Compute many features (> 100) that do not depend on image content
  - Train a classifier with properly chosen examples
    - Neural networks, Support Vector Machines (SVM)
- ROC curves are evaluated empirically on a test set
- CNN applied directly in the pixel domain are rapidly replacing SVMs

In summary

- Several steganographic techniques exist with a large number of available software packages (doubtful security)
  - Security looks trivial but is not
  - Need to know at least basic principles
  - Take care of system attacks
- Steganalysis
  - Reliable in some selected cases ...
  - ... difficult in general
  - Strongly dependent on application scenario
- Work in progress …

Watermarking

Watermarking vs steganography

- Data hiding with different requirements
- Statistical undetectability vs imperceptibility
  - visibility for images
- Robustness against processing and distortions
  - Non-intentional distortions
  - Intentional distortions (active warden)
- Security beyond undetectability

Main motivation: document protection

- Confidentiality
  - interception of data must be avoided
- Authentication
  - true origin of the document must be verified
- Integrity
  - data content must not be changed
- Copyright protection
  - non-authorized copying (also by legitimate owners) must be avoided

Digital watermarking

- Encryption does not solve the problem of unauthorized copying
- Multimedia data is marked to allow distribution to be tracked
- Digital watermarking can provide
  - an additional layer of protection after decryption
  - *data authentication and integrity*

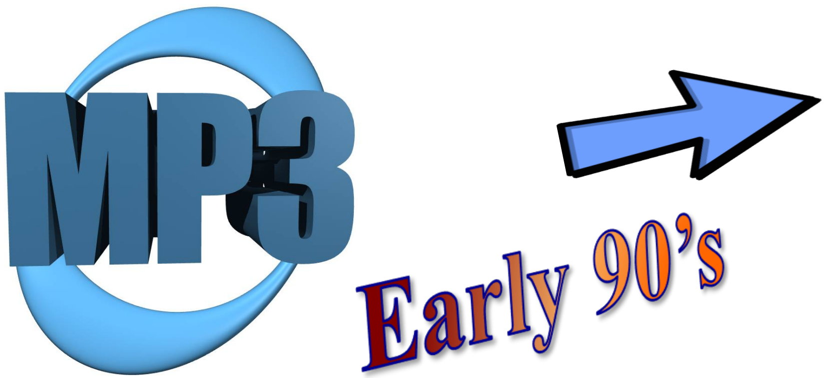
Digital watermark

- In copyright applications, a ***digital watermark*** is an identification code bearing information about the copyright owner, authorized consumers and so on
- It is ***permanently*** embedded into digital data for copyright protection, data authentication, integrity checking
- In most applications the watermark ***is not visible*** (perceivable) to a human observer, so that data quality is not degraded (*no more a matter of security*)

Watermark content

- Depending on the application the information conveyed by the watermark may vary
  - Allowed uses
  - Purchaser identification (fingerprinting)
  - Transaction details
  - Authentication
  - ...

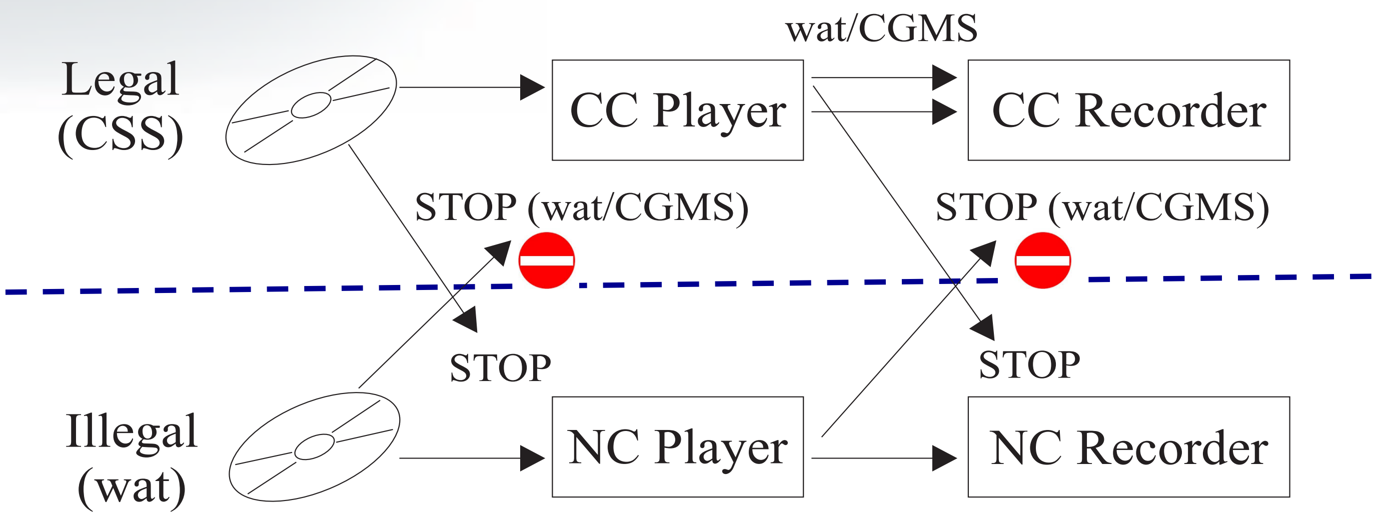
Killer application: copy control



- Copyright concerns: revenues from digital music and digital video at risk
- DRM: use of technology to prevent non-authorized viewing, copying, printing, editing, distribution of copyrighted material
- Agreement between manufacturers, copyright owners, sellers

Watermarking and DRM (copy control)

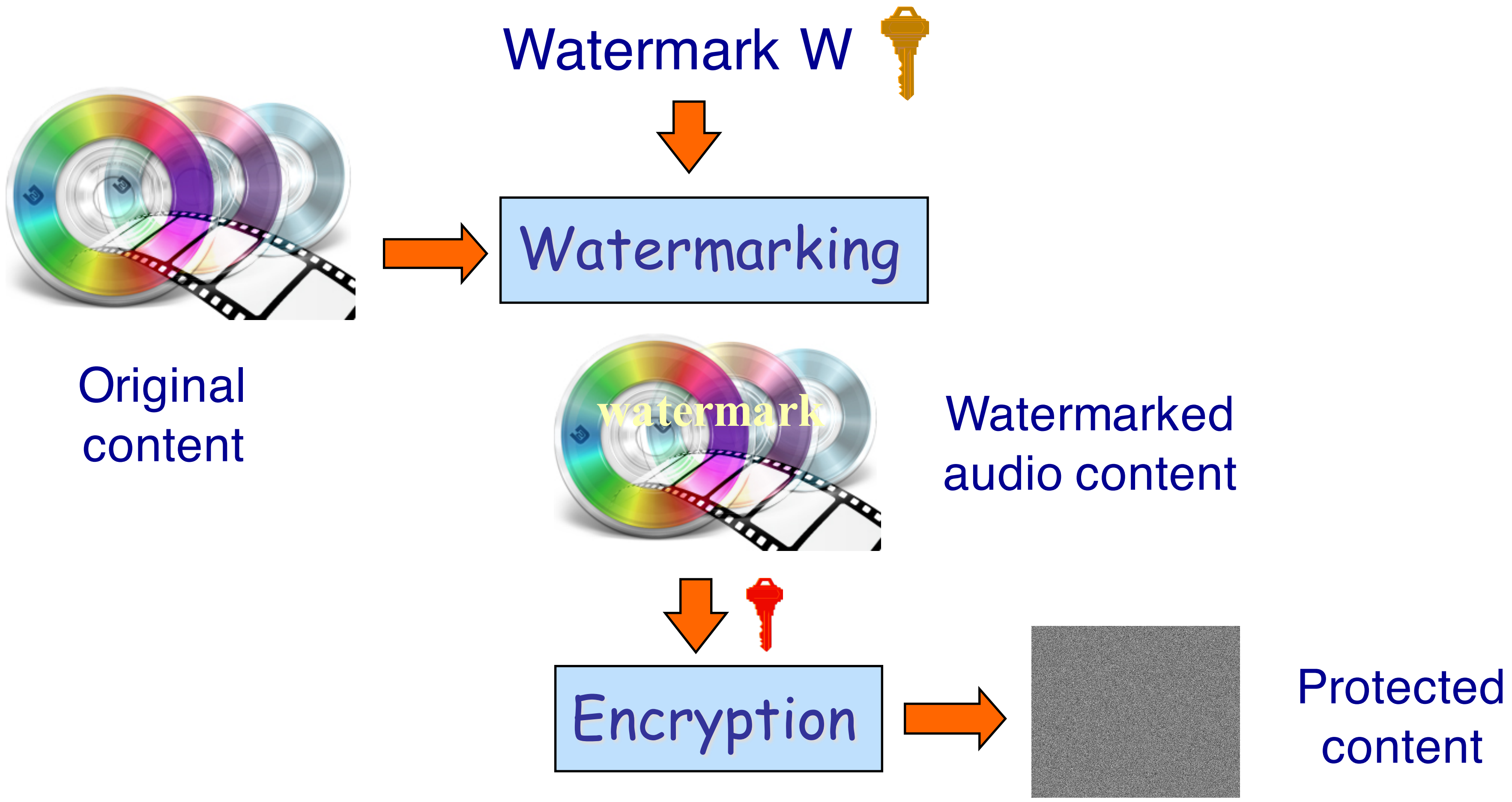
- Manufacturers agree to produce only compliant devices refusing playing, copying, editing copyrighted material without proper rights
- Cryptography by itself is not enough since it can not survive D/A – A/D conversion
- Watermarking would provide an additional layer of security after decryption



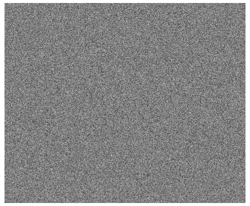
DRM temporarily abandoned

- At the end of nineties the DRM approach to copyright protection was abandoned
  - Watermarking does not have much to do with that
  - Main problem is agreement between stakeholders
  - Public opinion also played a role
- Raised interest now ...

CINAVIA protection of Blue-ray disks



CINAVIA protection of Blue-ray disks



Protected content

Decryption

Legitimate user

Free view

Successful decryption proves user is legitimate

The player retrieves the watermark, but since the content was previously decrypted the user has already demonstrated his right to view the content

The content can be copied only under encryption (or can not be copied depending on the device)

CINAVIA protection of Blue-ray disks

Compliant player with watermark detector

Compliant players refuse to show non-encrypted contents containing the watermark

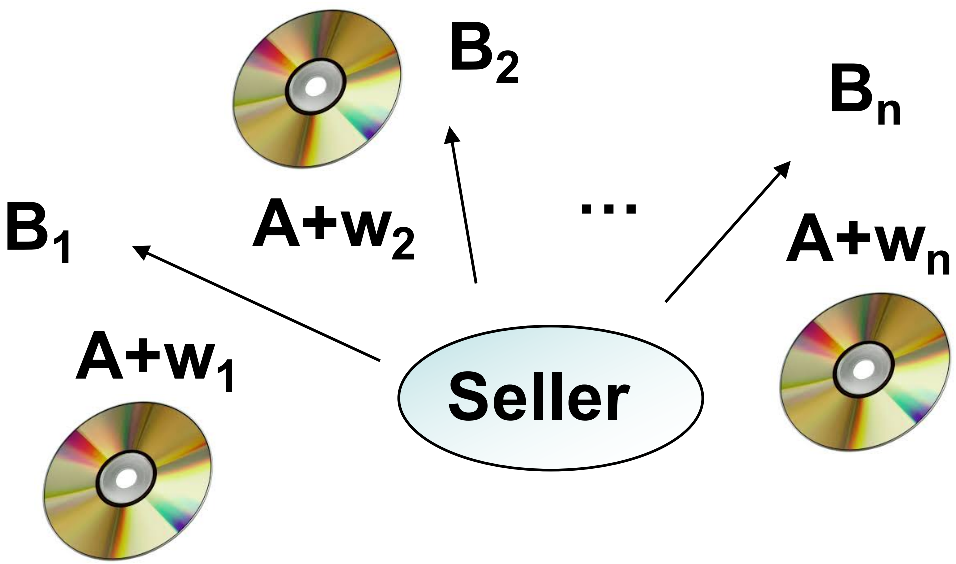
A pirate can obtain a non-encrypted, but watermarked, content

Illegally copied contents can be viewed (copied) only on non-compliant devices

The watermark can be removed only by degrading significantly the quality of the content

Buyer-seller protocol

- In a buyer-seller protocol, the seller inserts the ID of the buyer in every piece of content it sells
- The presence of the code can be used later on to trace back to the buyer that first distributed the content without permission



Ownership verification

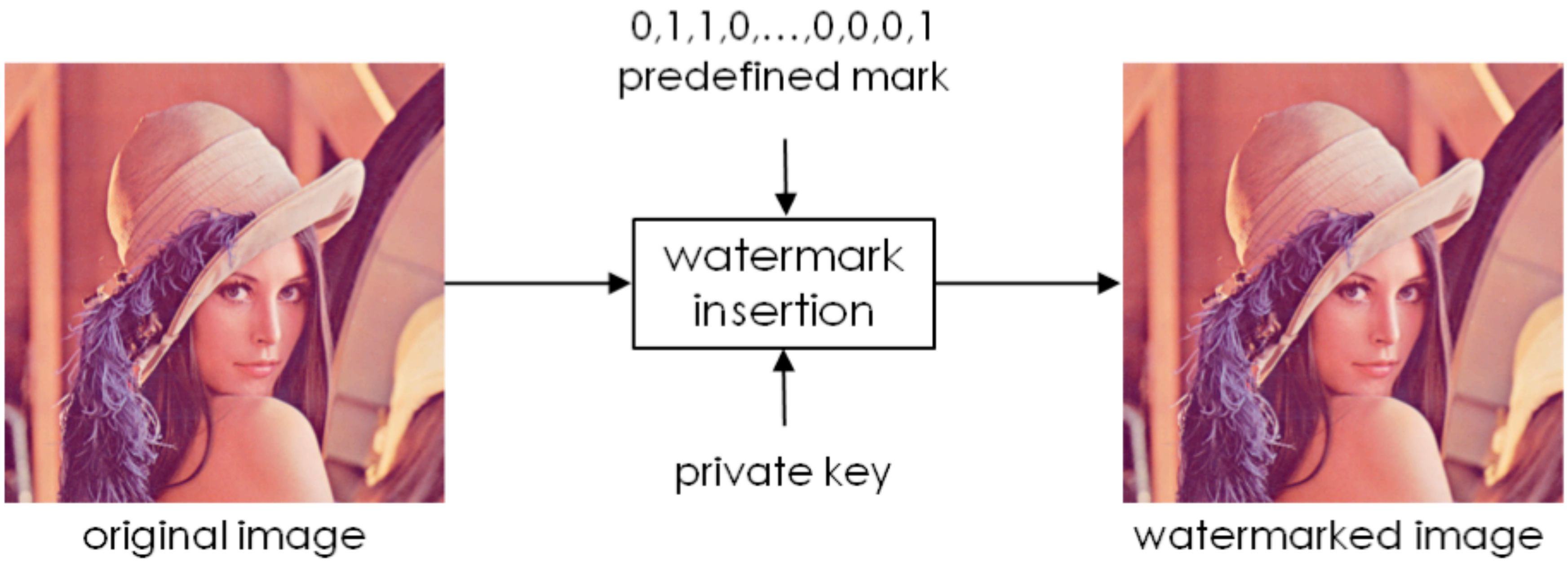
- The watermark contains the name of the owner (or creator) of the content
  - Perhaps it is the oldest use of watermarking
  - Requires a complete infrastructure and usage of cryptographic tools

Other applications: authentication

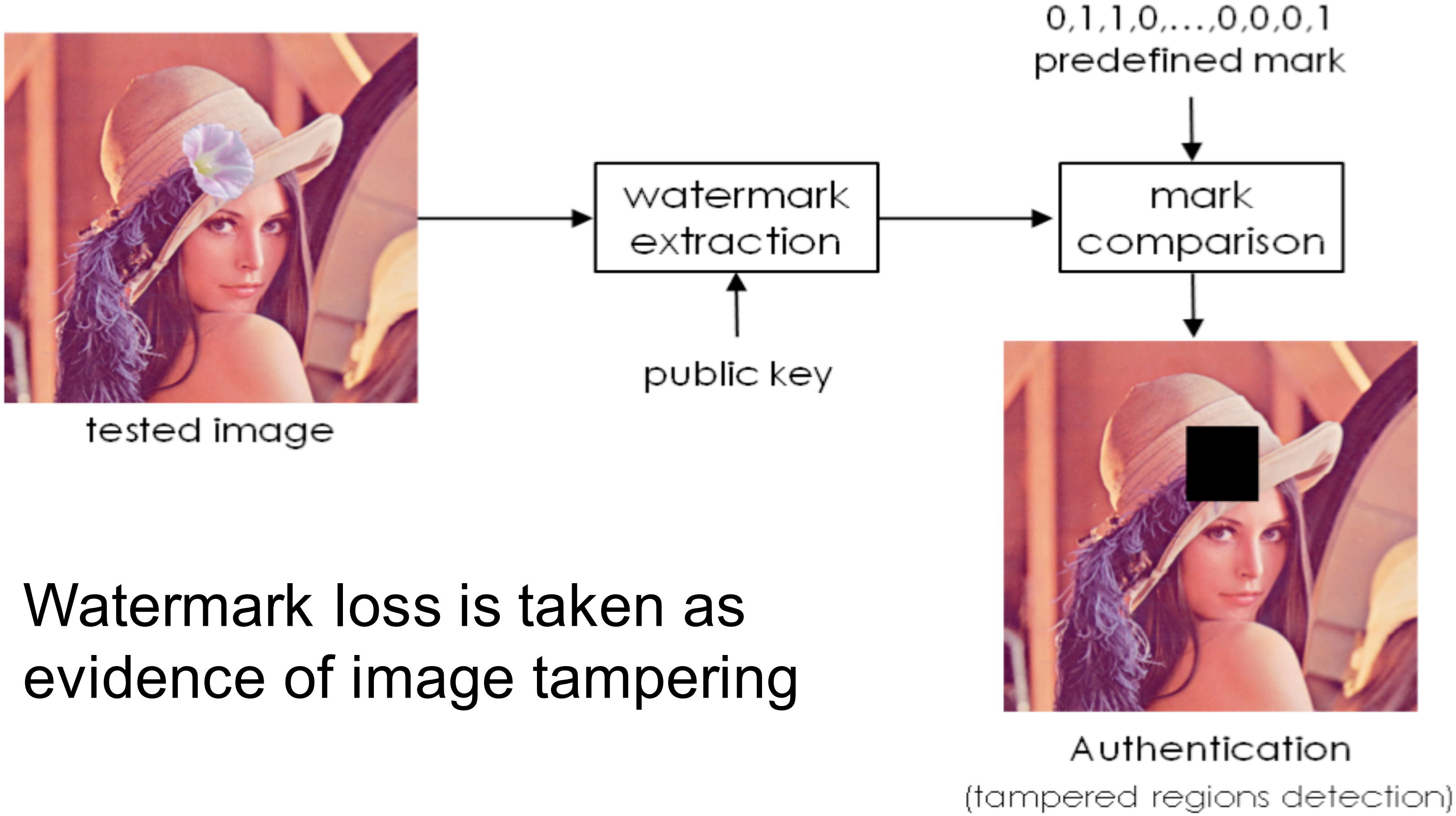
- Decide whether a given document is original or it has been tampered with
- Possibly localize the tampered region
- Two approaches are possible
  - Fragile (or semi fragile) watermarking
  - Robust watermarking plus perceptual hashing

Authentication via fragile watermarking

- A fragile watermark is lost as soon as the image is modified



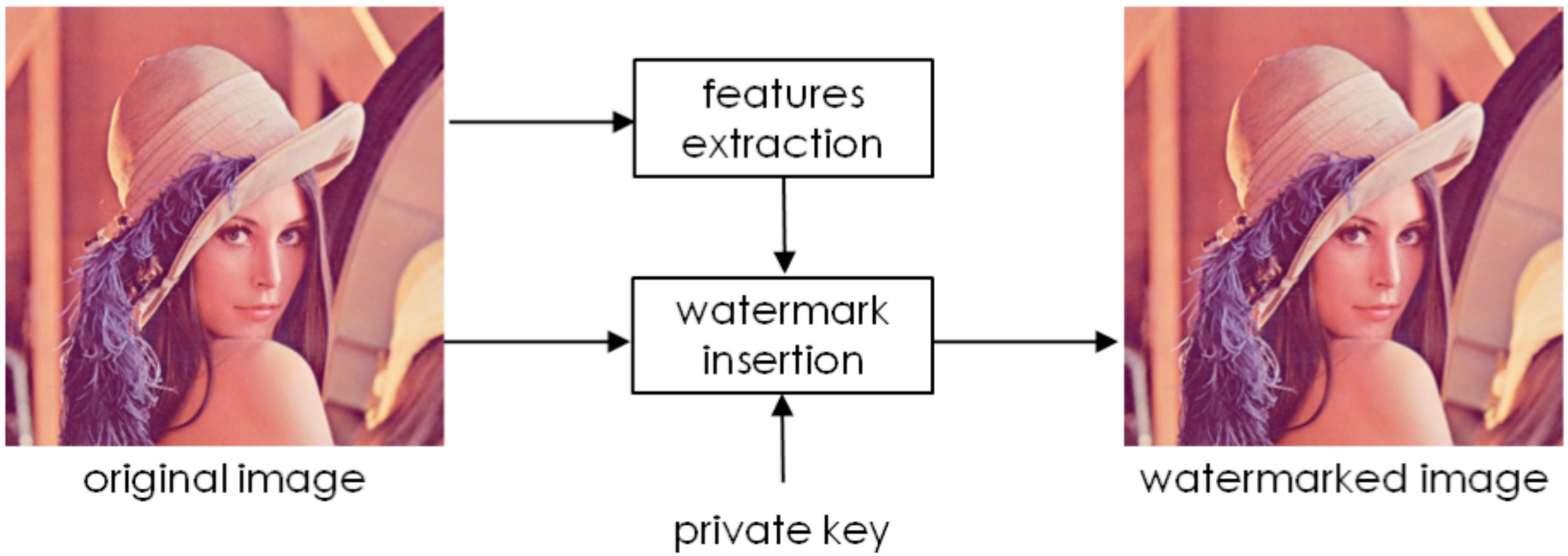
Authentication via fragile watermarking



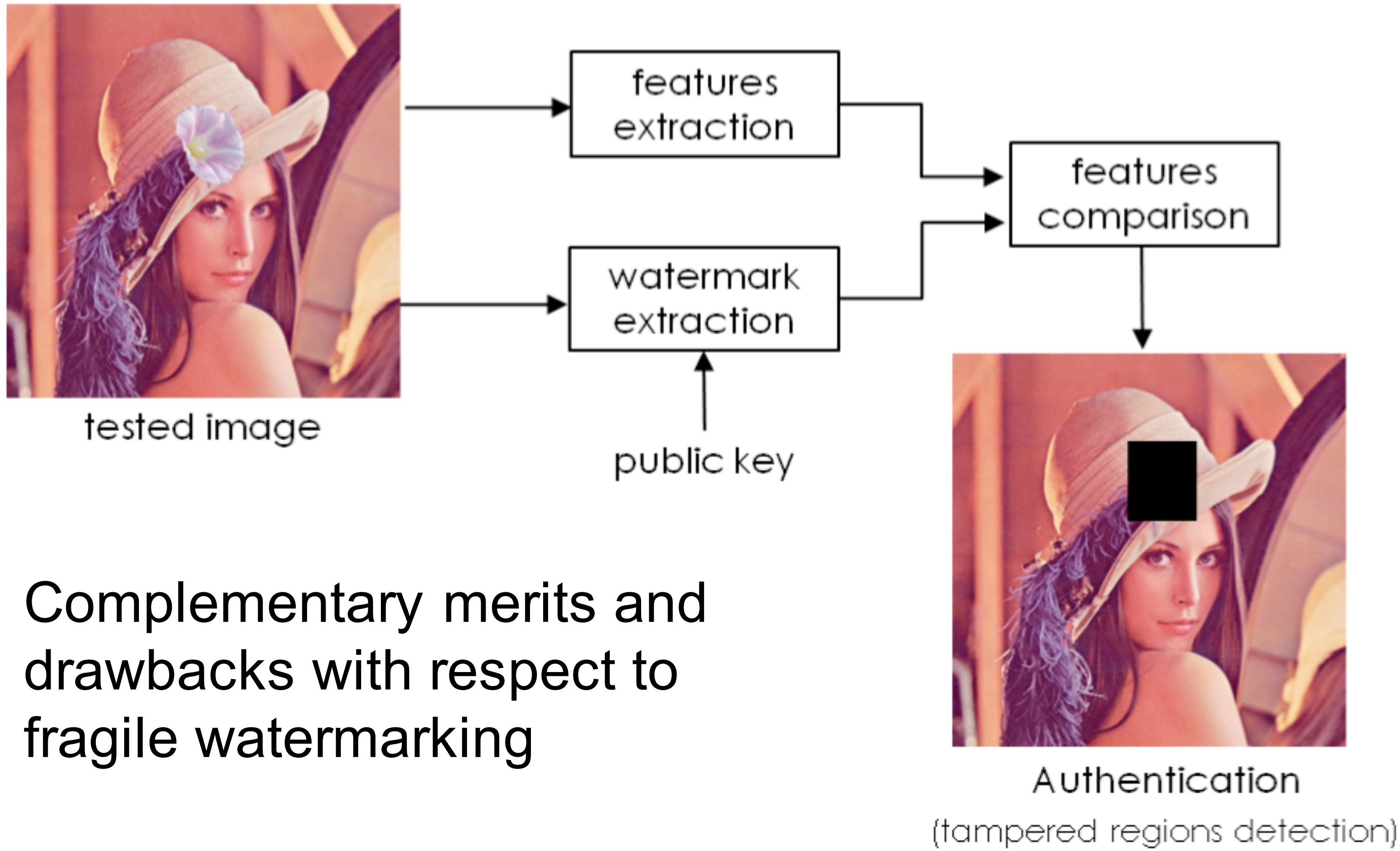
- Watermark loss is taken as evidence of image tampering

Authentication via robust watermarking

- With robust watermarking a summary of the image is inserted within the image itself



Authentication via robust watermarking



- Complementary merits and drawbacks with respect to fragile watermarking

Connect the digital and analog worlds

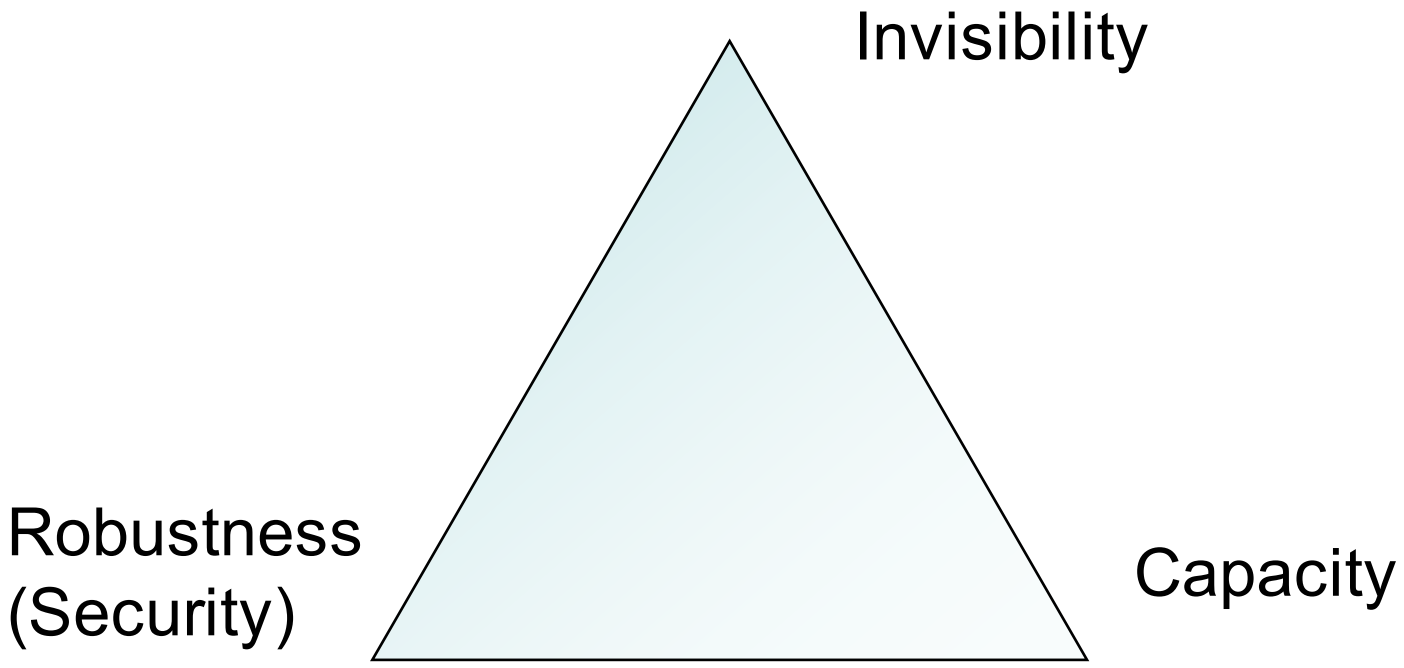
- Due to the ability to survive D/A and A/D conversion, the hidden data could provide a mean to link the analog and the digital world
- Alternative to barcodes
- Second screen application for
  - Advertisement
  - Added services
  - Navigation

Most common requirements

- Requirements strongly depend on the application. The most important ones are
  - Invisibility (unobtrusiveness)
  - Robustness (sometimes security)
  - Payload (sometimes referred to as Capacity)
- Other requirements include
  - Simplicity
  - Scalability
  - Decoder / detector blindness

The watermarking triangle

- Invisibility, robustness and capacity form the so-called watermarking trade-off triangle



Robustness criteria

- Signal processing
  - enhancement, sharpening, blurring, linear/non-linear filtering (median, de-speckle)
- Compression
  - Robustness against JPEG compression is mandatory
- Geometric manipulations
  - resizing, cropping, translation, rotation, flip
- A/D – D/A conversion

Robustness vs. Security

- Robustness deals with non-malicious manipulations
- Security considers malicious (targeted) attacks in a hostile environment
- In a security analysis it is assumed that the attacker knows the watermarking algorithm: hence ad-hoc attacks can be conceived

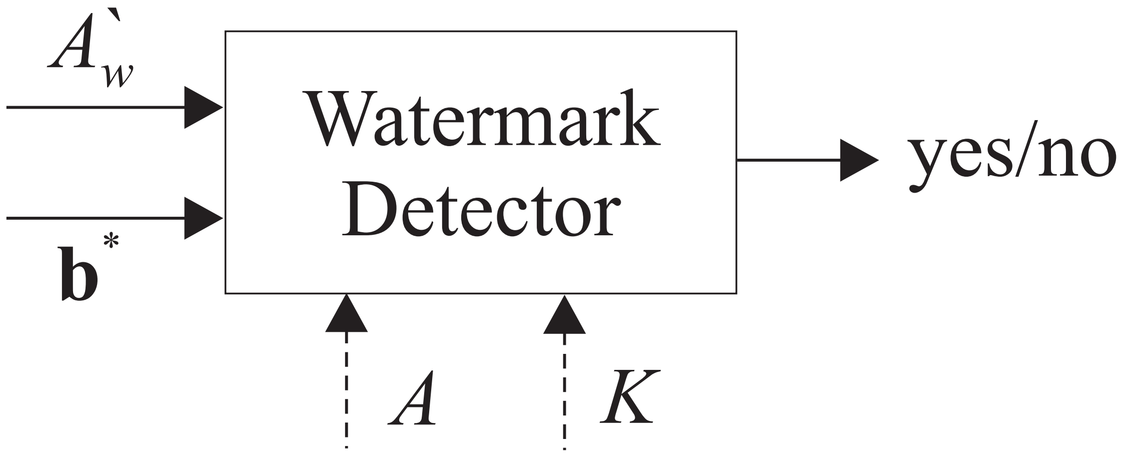
- *Most common approach to determine security: expose the watermark to large scale, massive attacks, e.g. BOWS contests*

Classification of techniques

- Decoding process
  - *blind techniques*
    - the watermark is recovered without resorting to the original non-marked content or any information derived from it
  - *non-blind techniques*
    - the original content is needed to read the watermark
    - robustness is more easily achieved
    - often the application scenario does not allow the decoder to access the original content

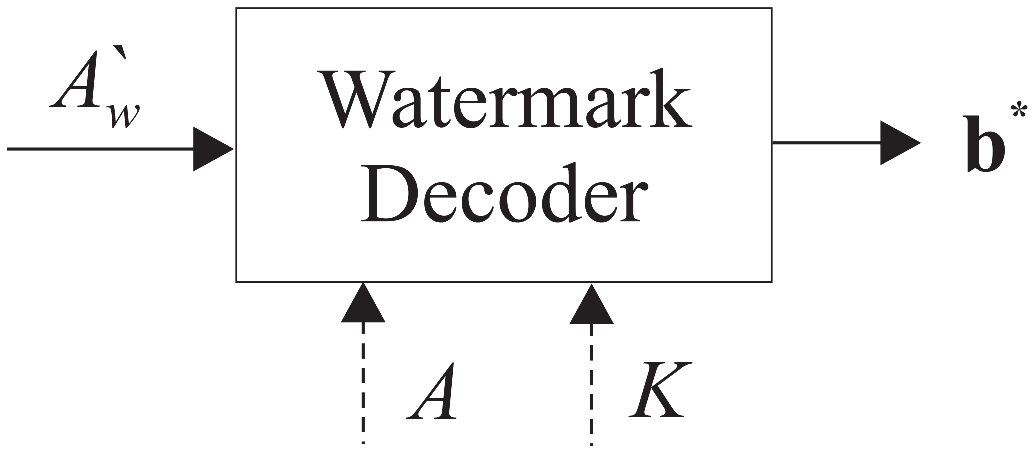
Classification of techniques

- Decoding process
  - *detectable* (1-bit, 0-bit) watermark
    - it is only possible to decide whether a given watermark is embedded in the image



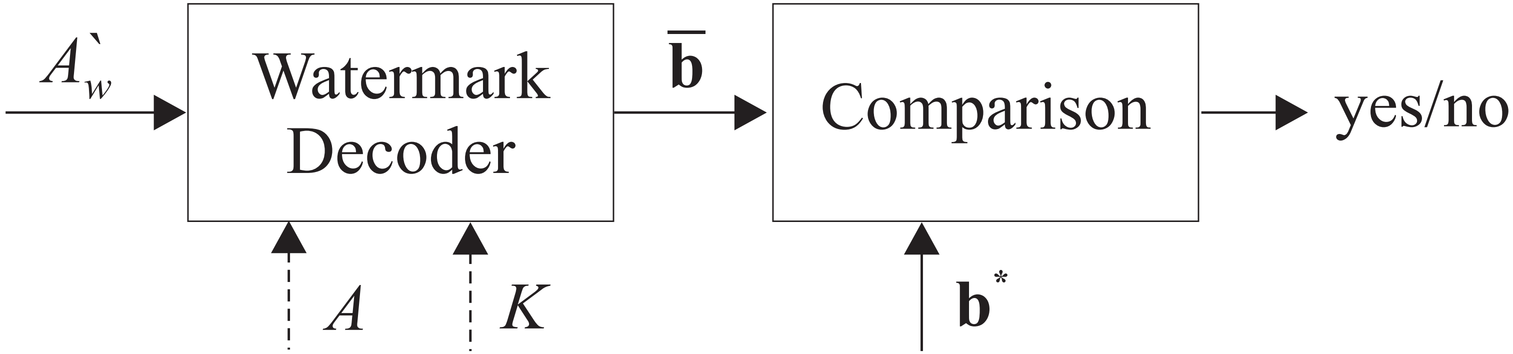
Classification of techniques

- Decoding process
  - *readable* watermark (multi-bit watermarking)
    - the bits hidden in the image can be read without knowing them in advance



From multi-bit to 1-bit watermarking

- Passing from multi-bit to 1-bit watermarking is rather easy

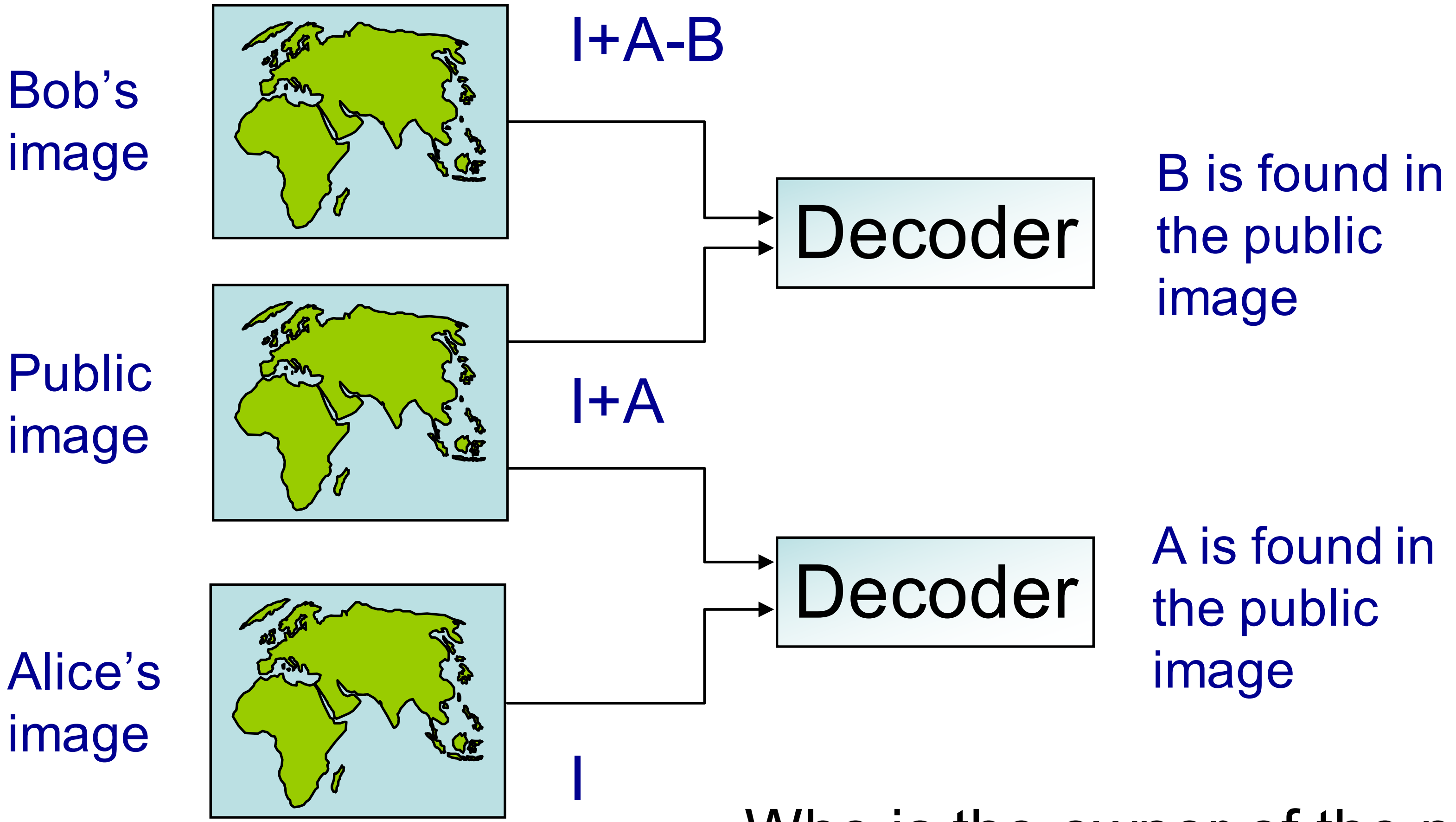


- Going the other way-round is also possible but it leads to inefficient schemes

Protocol-level considerations

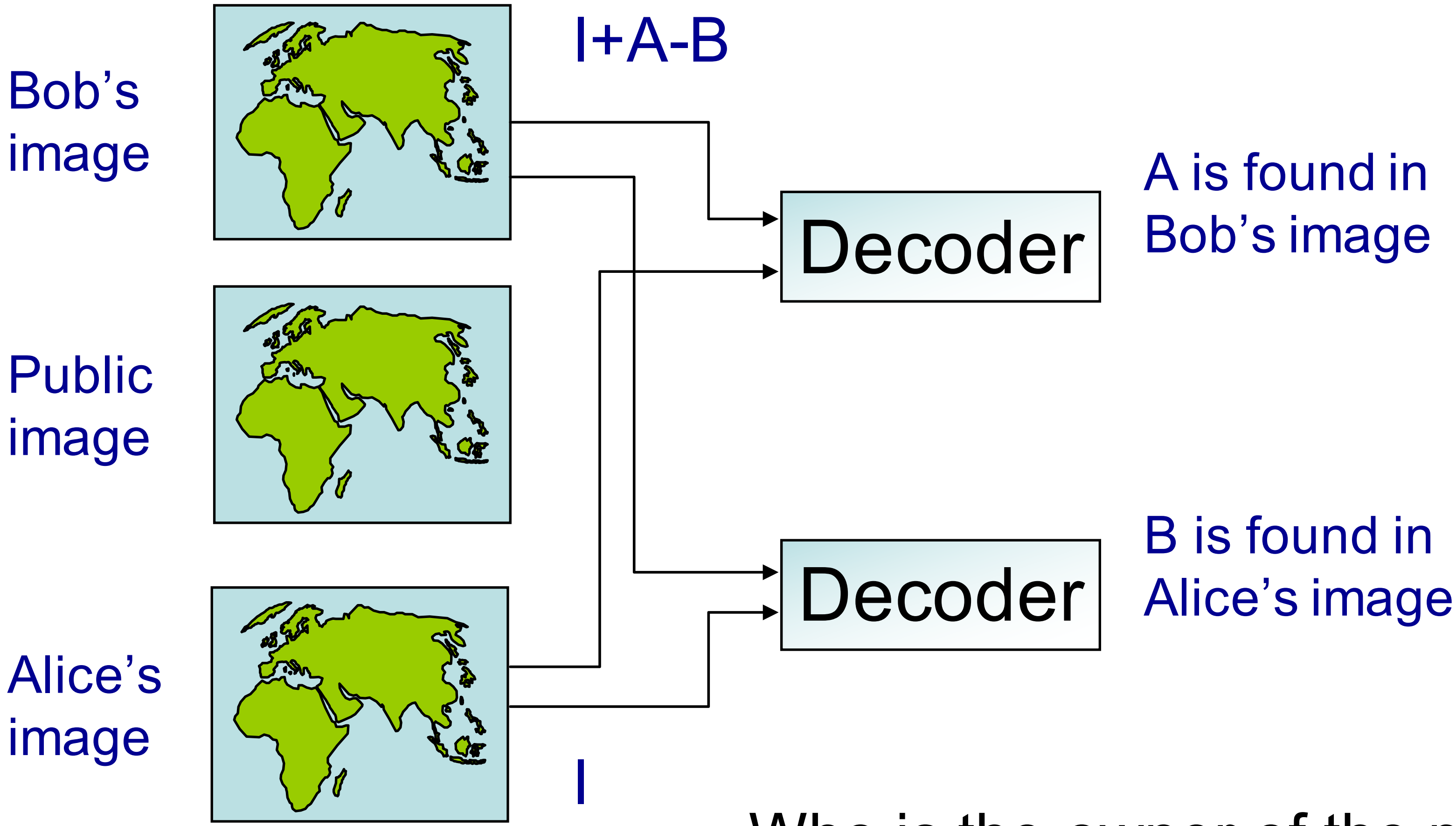
- It is important to stress out that the requirements a watermarking system must satisfy are dictated by the application scenario the system must work in.
- For instance, the particular Electronic Copyright Management System used to protect image IPR must be taken into account.
- The blind/non-blind, detectable/readable nature of the watermark must be chosen in this way

Example: the IBM attack



Who is the owner of the public image ?

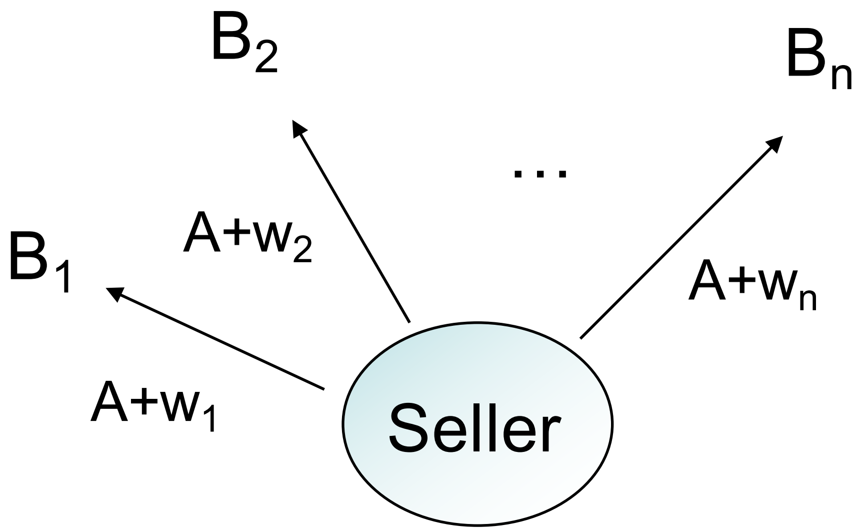
Example: the IBM attack



Who is the owner of the public image ?

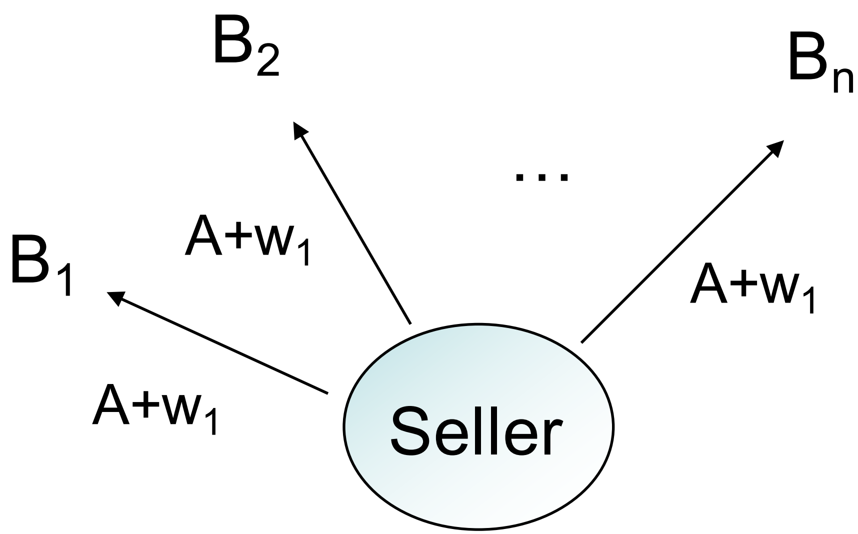
Buyer-seller protocol

- In a fingerprinting scenario the seller inserts the identification code of the buyer in every piece of content it sells
- The presence of the code can be used later on to trace back to the buyer that first distributed the content without permission



Buyer-seller protocol

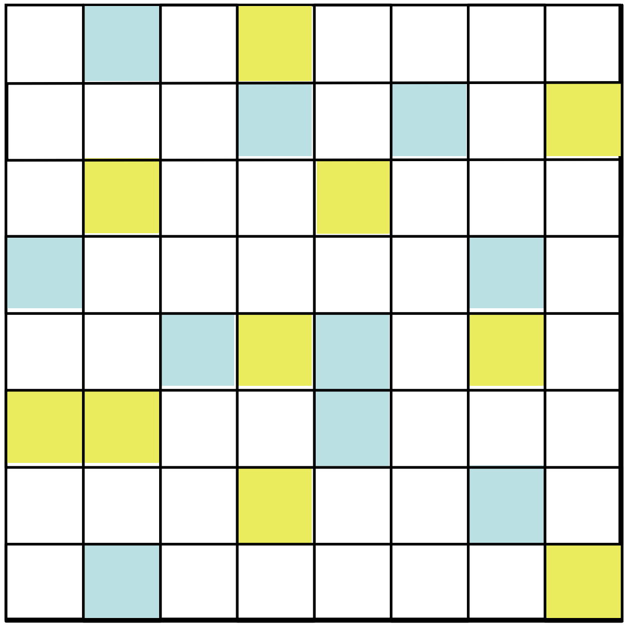
- A buyer whose watermark is found in an unauthorized copy can not be inculpated since he/she can claim that the unauthorized copy was created and distributed by the seller.
- The seller could redistribute many copies of a work containing the fingerprint of a buyer (say B_1) without paying the due royalties to the author, and claim that such copies were illegally distributed or sold by B_1



HOW ?

- A very elegant (and complex) theory has been developed providing a rigorous framework for watermarking
- It involves several disciplines including: statistical signal processing, physiological aspects related to perception, information theory, channel coding theory, cryptography …
- Here we give only one example of spread spectrum watermarking (most common approach)

Intuitive example: patchwork



$A = \{a_i\}_{i=1,n}$

$B = \{b_i\}_{i=1,n}$

- $E[a_i - b_i] = 0$
- By letting $a'_i = a_i + d$,
  $b'_i = b_i - d$,
  the watermark is inserted in the image
- Detection is achieved by computing the quantity

$$S_n = \frac{1}{n}\sum_{i=1}^{n}(a_i - b_i)$$

- A typical value for n is about 10.000

More rigorously

- Let us consider the case of 1-bit watermarking
- A watermarking sequence w_i ($i = 1 \ldots n$) is generated by starting from a secrete key
- For instance, the sequence w_i may be an i.i.d. sequence having a fixed pdf (e.g. N(0,1))
- The marked signal y_i is formed by adding (Add-SS) a scaled version of w_i to the features (whatever they are) of the to-be-watermarked signal x_i

$$y_i = x_i + \gamma w_i$$

The Add-SS, AWGN case

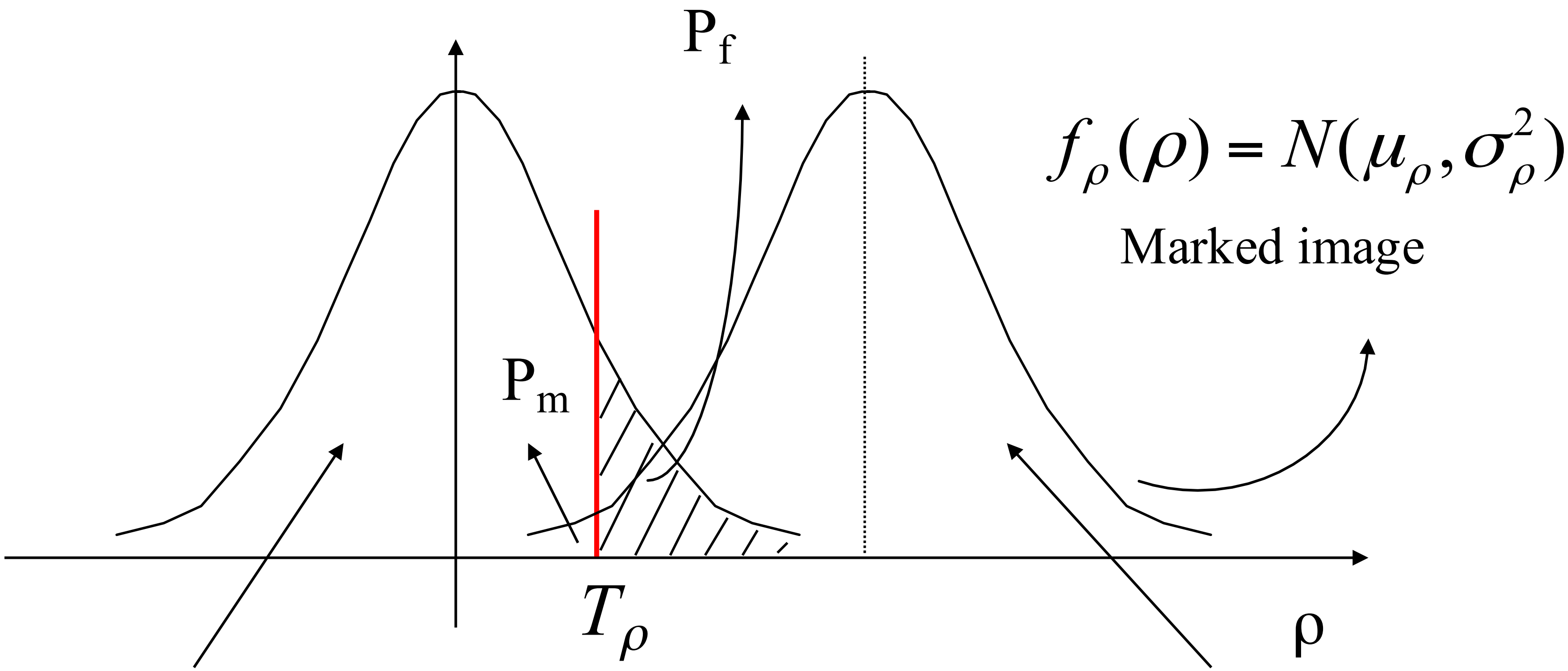
- It is easy to show that in the AWGN case optimum detection corresponds to correlation detection

$$\rho = \frac{1}{n}\sum_{i=1}^{n} y_i w_i \qquad \begin{cases} \rho > T_\rho & H_1 \\ \rho < T_\rho & H_2 \end{cases}$$

- The false and missed detection probabilities can be computed, and the detection threshold fixed, if the variance and mean of the host features is known

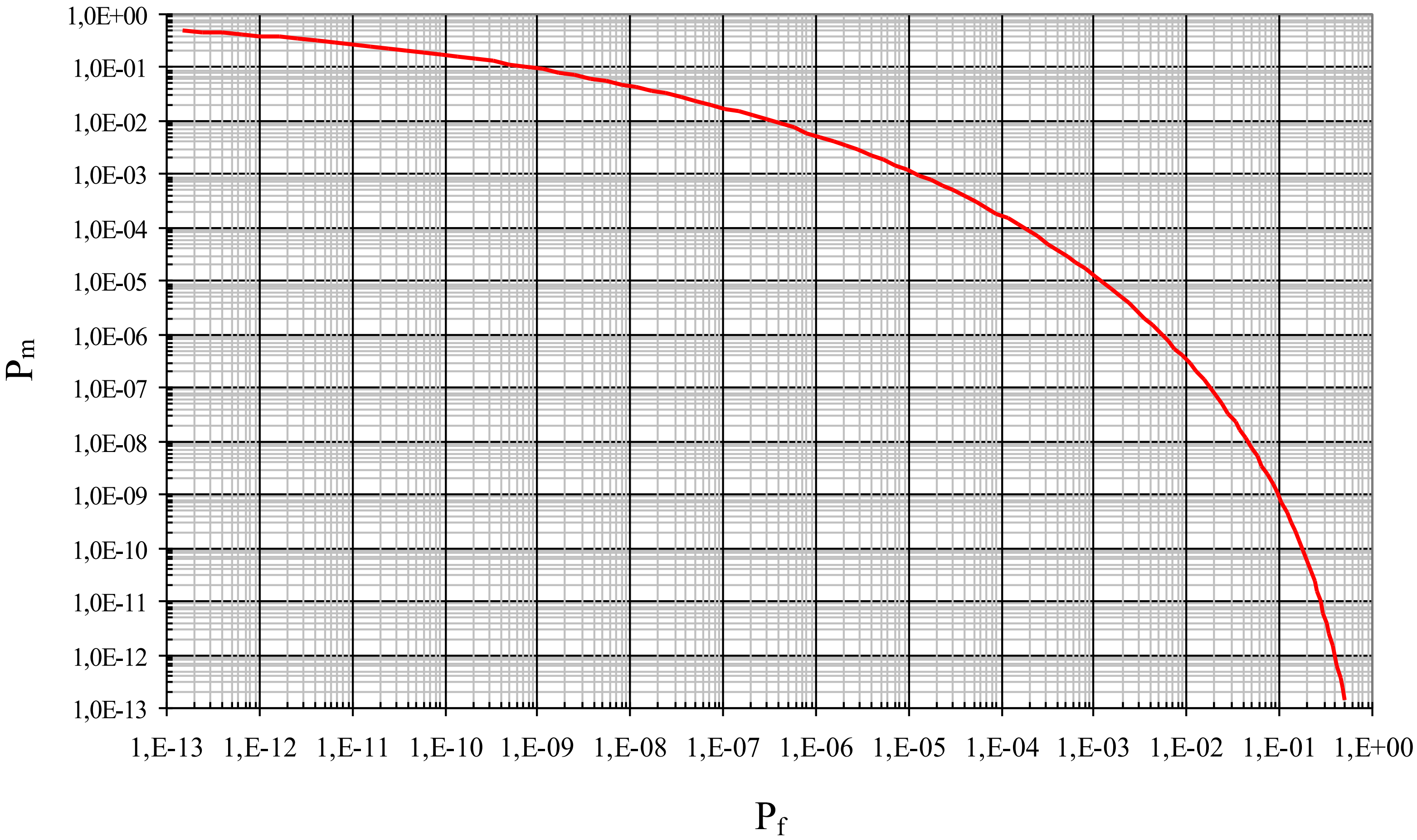
$$f_\rho(\rho \mid 0/1) = N(\mu_{\rho|0/1}, \sigma^2_{\rho|0/1})$$

The case of antipodal Add-SS (AWGN)



$$\begin{cases} \mu_{\rho|0} = 0 \\ \sigma^2_{\rho|0} = \dfrac{\sigma_x^2 + \sigma_n^2}{n} \end{cases} \qquad \begin{cases} \mu_{\rho|1} = \gamma \\ \sigma^2_{\rho|1} = \dfrac{\sigma_x^2 + \sigma_n^2}{n} \end{cases}$$

ROC curve



$$P_m(P_f) = \frac{1}{2}\left(\sqrt{\frac{n \cdot SNR}{2}} - erfc^{-1}(2P_f)\right) \qquad SNR = \frac{\gamma^2}{\sigma_x^2 + \sigma_n^2}$$

Extension to the multibit case

- Spread spectrum watermarking cab be easily extended to the case of multibit watermarking

$$y_i = x_i + \gamma w_i$$

- γ = -1 => b = 0
- γ = +1 => b = 1
- The host signal is split into chunks each carrying one bit
- Error correction coding can be used to increase robustness

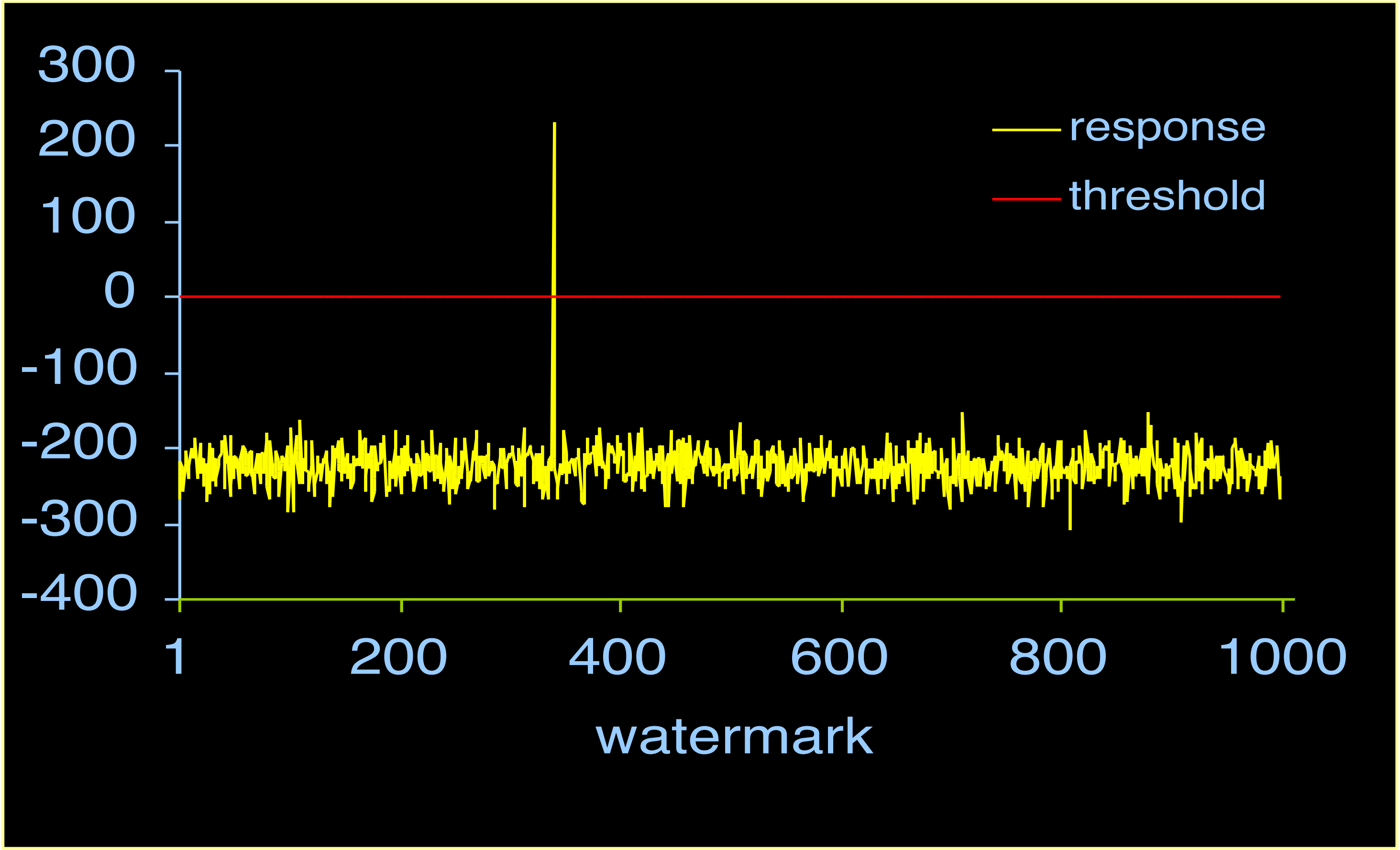
Example*

Original Image

Watermarked image
(PSNR = 50dB)

* M. Barni, F. Bartolini, A. De Rosa, and A. Piva, “A new decoder for the optimum recovery of non-additive watermarks, *IEEE Trans. Image Processing*, 10 (2001), pp. 755–766.

Detector answer



Example: robustness

JPEG compression with quality factor = 3%

Example: robustness

Addition of white gaussian noise with variance = 2000

Example: robustness

Print, copying and scanning

References

- J. Fridrich, *Steganography in Digital Media: principles, algorithms and applications*, Cambridge University Press, 2010
- I. J. Cox, M. Miller, J. Bloom, *Digital watermarking*, Morgan Kaufmann
- M. Barni, F. Bartolini, *Watermarking Systems Engineering: Enabling Digital Assets Security and other Applications*, Marcel Dekker, New York, 2004.