Pattern Recognition
and Applications Lab

Adversarial Pattern Recognition

Fabio Roli

ICPRAM 2016, Rome, Italy, February 24h, 2016

University
of Cagliari, Italy

Department of Electrical
and Electronic
Engineering

A question to start…

What is the oldest survey article on pattern recognition that you have ever read?

What is the publication year?

This is mine…year 1966

Pattern Recognition

By DENIS RUTOVITZ

Medical Research Council

[Read before the ROYAL STATISTICAL SOCIETY on Wednesday, May 18th, 1966, the President, Mr L. H. C. TIPPETT, in the Chair]

1. INTRODUCTION

DURING the past 10 years about 200 articles and several books have appeared, dealing with machine recognition of optical and other patterns (mainly alphabetic characters and numerals). About half of these have described methods not linked to a specific

Credits: Dr Gavin Brown for showing me this article

Applications in the old good days…

Pattern Recognition

By DENIS RUTOVITZ

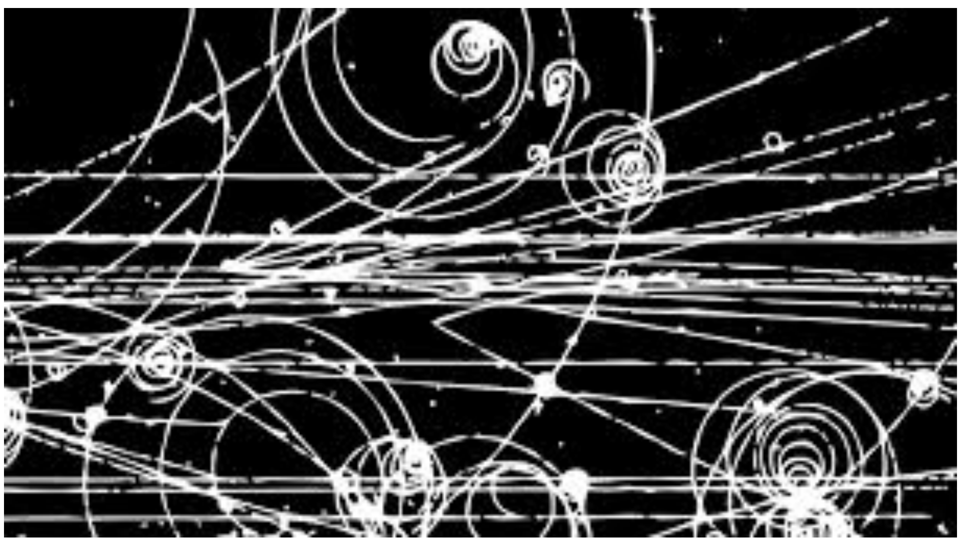
Medical Research Council

[Read before the ROYAL STATISTICAL SOCIETY on Wednesday, May 18th, 1966, the President, Mr L. H. C. TIPPETT, in the Chair]

What applications do you think that this paper dealt with?

Popular applications in the sixties

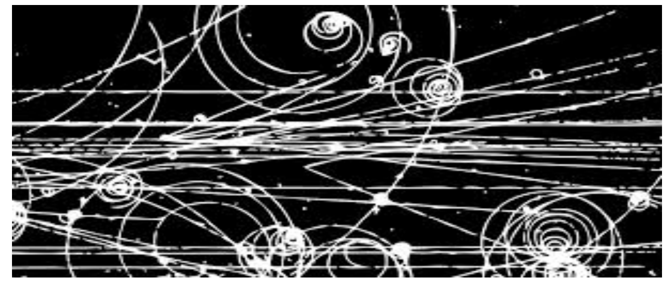
OCR for bank cheque sorting



Detection of particle tracks in bubble chambers

Aerial photo recognition

Key feature of these apps



Specialised applications for **professional** users..

What about today applications?

Today applications of pattern recognition

Key features of today apps

Personal and **consumer** applications…

ALL RIGHT? ALL GOOD?

Iphone 5s and 6s with fingerprint reader…

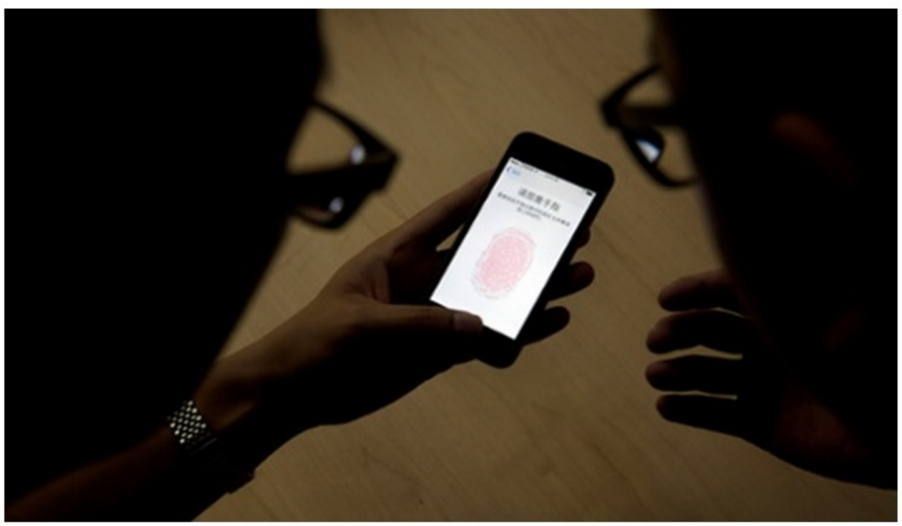
Hacked a few days after release…

iPhone 5S fingerprint sensor hacked by Germany's Chaos Computer Club

Biometrics are not safe, says famous hacker team who provide video showing how they could use a fake fingerprint to bypass phone's security lockscreen

Follow Charles Arthur by email BETA

Charles Arthur
theguardian.com, Monday 23 September 2013 08.50 BST
Jump to comments (306)



Home › iPhone 6 › Your iPhone Can Be Hacked With A Photo Of Your Thumb

Your iPhone Can Be Hacked With A Photo Of Your Thumb

Your fingerprint may not keep your iPhone safe any more. Someone has figured out how to use photos and commercially available software to break through an iPhone 6's fingerprint sensor, known as Touch ID.

Take-home message

We are living exciting time for pattern recognition…

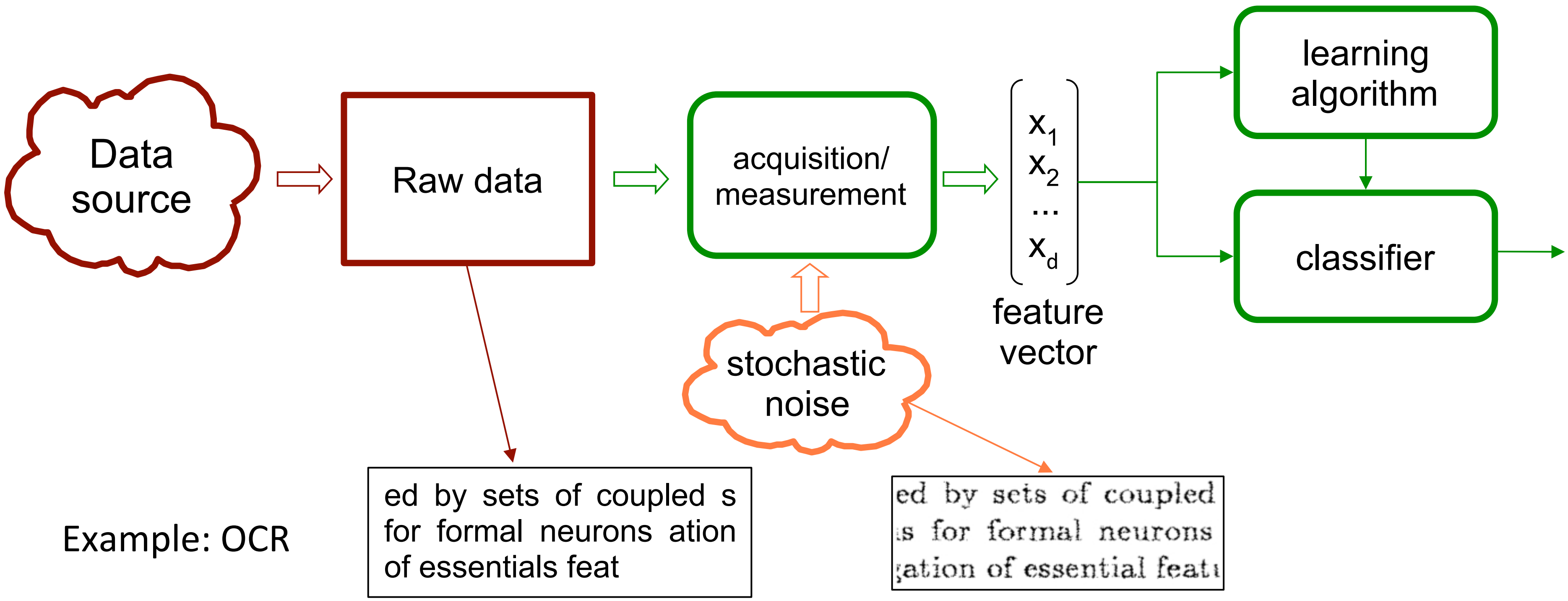
…Our work feeds a lot of consumer technologies for personal applications...

This opens up new big possibilities, but also new security risks

Are we ready for this?

Can we use the classical pattern recognition model under attack?

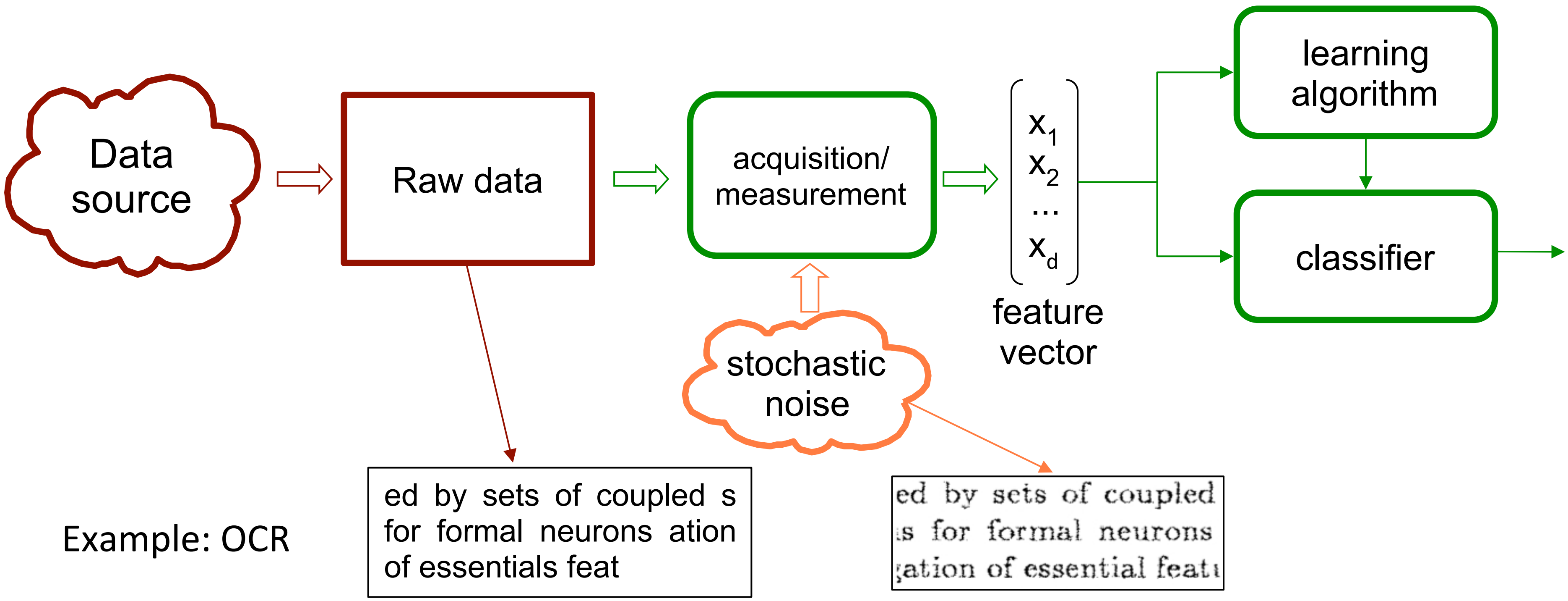
The classical statistical model



Note these two implicit assumptions of the model:

1. **the source of data does not dependent on the classifier**
2. **Noise affecting data is stochastic**

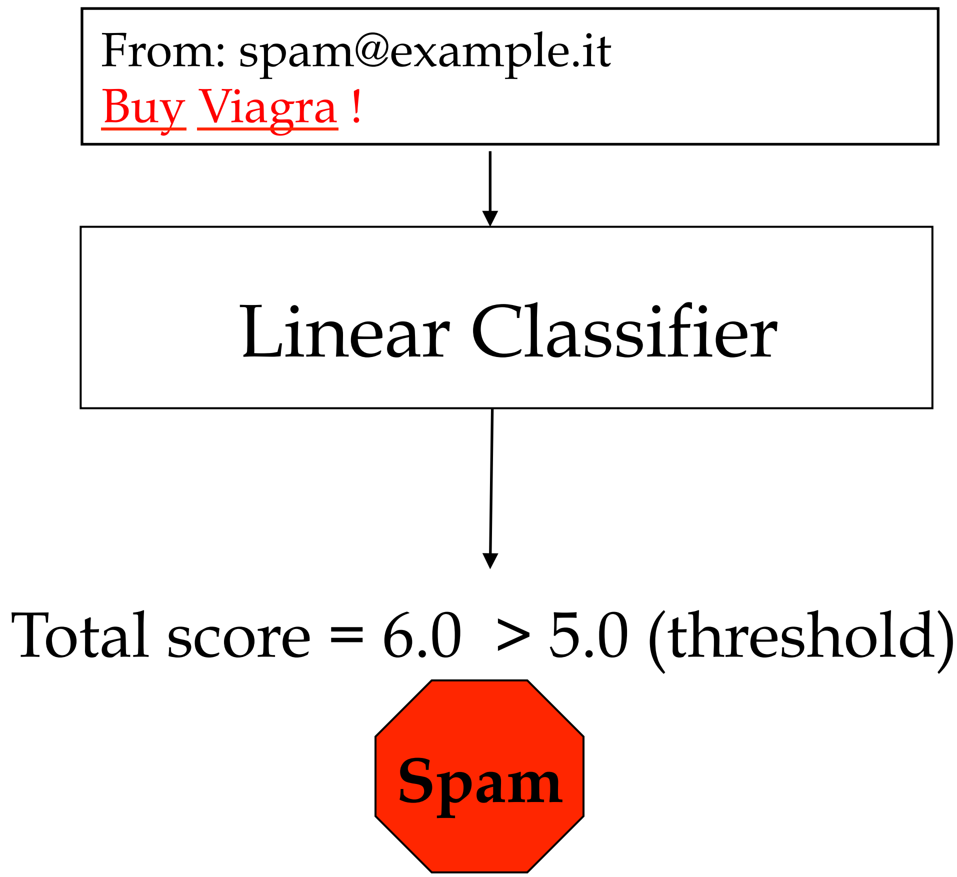
Can this model be used under attack?



Example: OCR

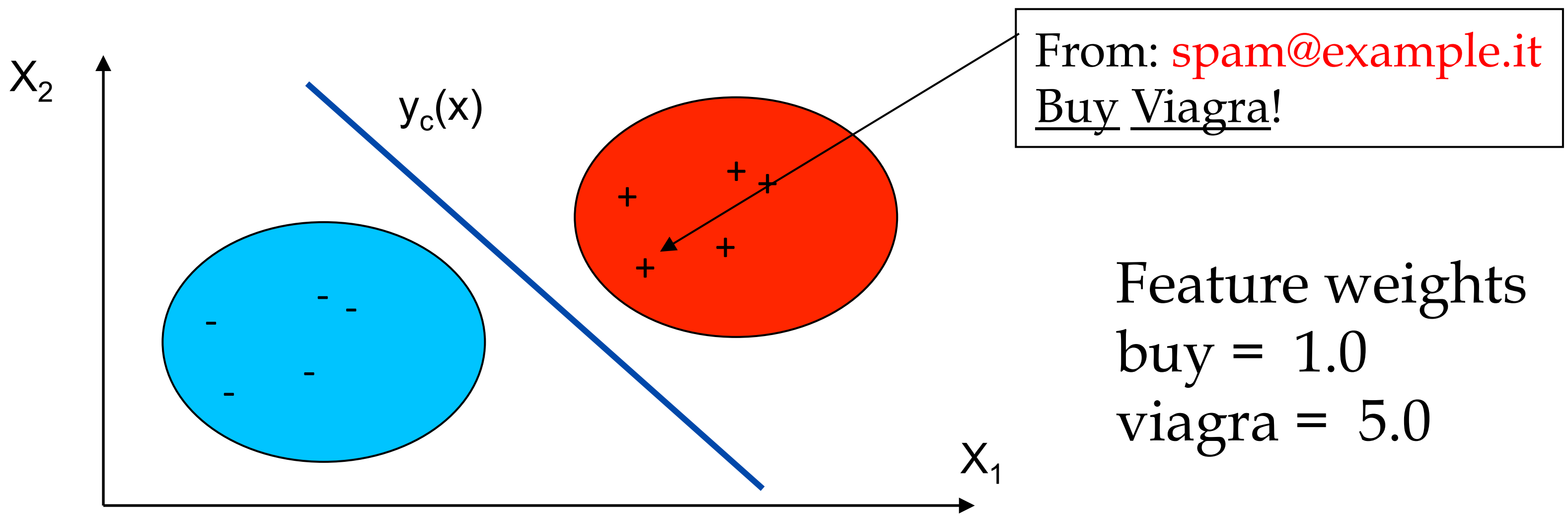
An example: spam filtering

Feature weights
buy = 1.0
viagra = 5.0



- The famous SpamAssassin filter is really a linear classifier
  - http://spamassassin.apache.org

Feature space view



Feature weights
buy = 1.0
viagra = 5.0

- Classifier's weights can be learnt using a training set
- The SpamAssassin filter uses the perceptron algorithm

But spam filtering is non a static classification task, the data source is not neutral…

The data source can add “good” words

Feature weights
buy = 1.0
viagra = 5.0
University = -2.0
Rome = -3.0

From: spam@example.it
Buy Viagra !
University Rome

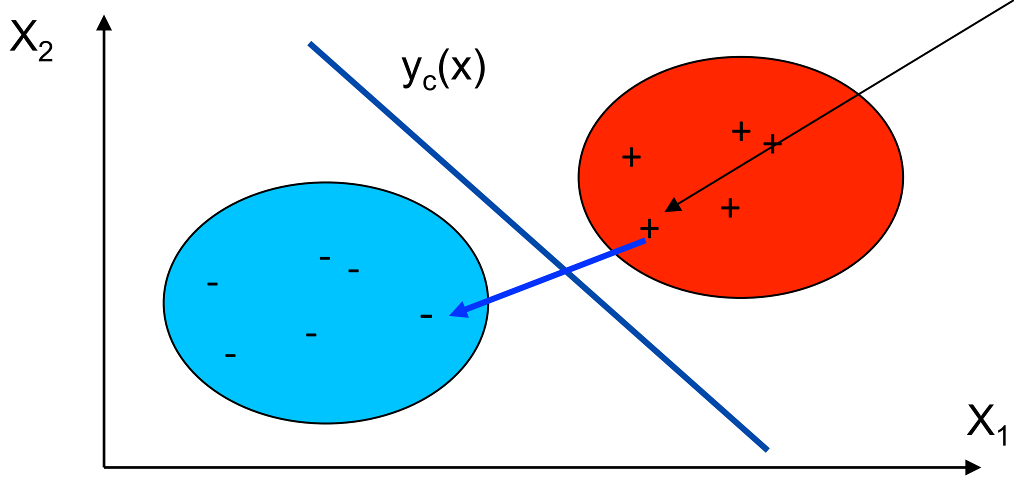
Linear Classifier

Total score = 1.0 < 5.0 (threshold)

Ham

✓ Adding “good” words is a typical spammers’ trick [Z. Jorgensen et al., JMLR 2008]

Adding good words: feature space view

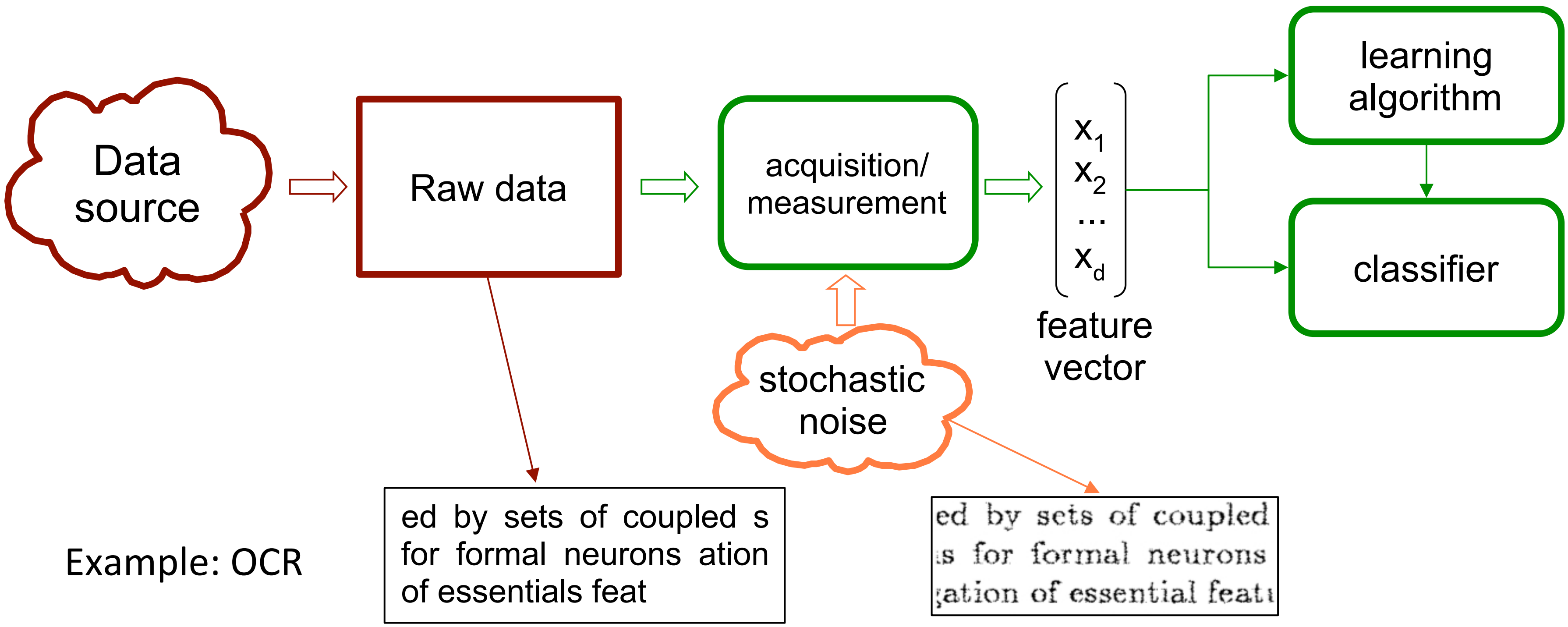


From: spam@example.it
Buy Viagra!
University Rome

Feature weights
buy = 1.0
viagra = 5.0
University = -2.0
Rome = -3.0

✓Note that spammers corrupt patterns with a *noise* that is *not random*..

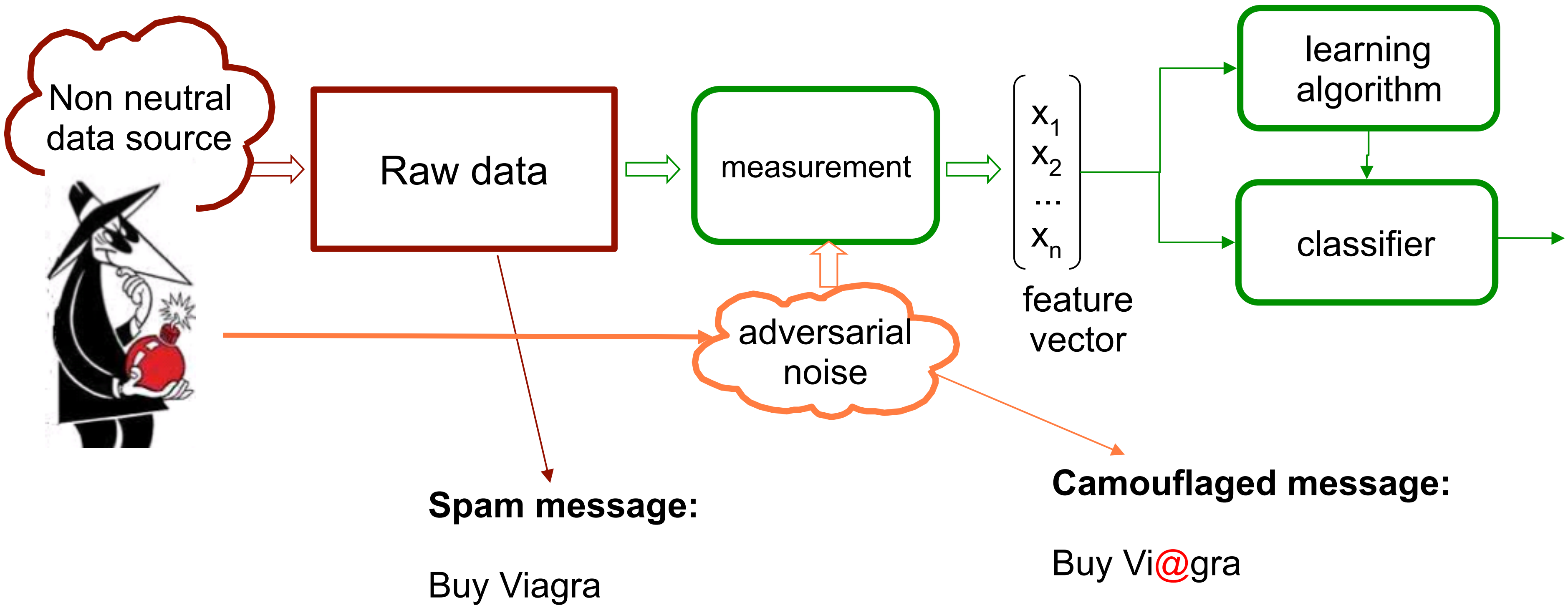
Is this model good for spam filtering?



- **The data source does not dependent on the classifier**
- **Noise affecting data is stochastic ("random")**

No, it is not…

Adversarial pattern classification



1. **the source of data is *not neutral*, it really depends on the classifier**
2. **Noise is not stochastic, it is *adversarial*, it is just crafted to maximize the classifier's error**

Adversarial noise vs. stochastic noise

✓This distinction is not new …

Shannon's stochastic noise model: probabilistic model of the channel, the probability of occurrence of too many or too few errors is usually low

Hamming's adversarial noise model: the channel acts as an adversary that arbitrarily corrupts the code-word subject to a bound on the total number of errors

The classical model cannot work…

✓Standard classification algorithms assume that data generating process is independent from the classifier

➢This is not the case for adversarial tasks

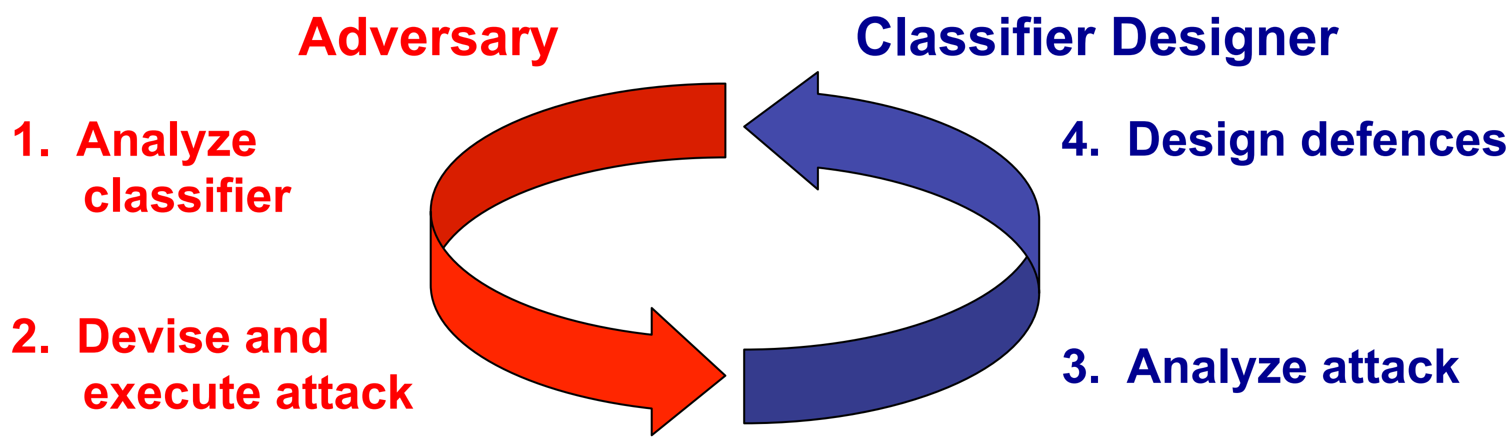
✓Easy to see that classifier performance will degrade quickly if the adversarial noise is not taken into account

✓So adversarial tasks are a mission impossible for the classical model

How should we design pattern classifiers under attack?

Adversary-aware pattern classification

B. Biggio, G. Fumera, F. Roli. Security evaluation of pattern classifiers under attack, IEEE Trans. on Knowl. and Data Engineering, 2013



Pattern recognition systems should be aware of the *arms race* with the adversary

How can we design adversary-aware pattern recognition systems?

The three golden rules

1. Know your adversary

2. Be proactive

3. Protect your classifier

Know your adversary

If you know the enemy and know yourself, you need not fear the result of a hundred battles
(Sun Tzu, The art of war, 500 BC)

ADVERSARY'S 4D MODEL

Adversary's goal

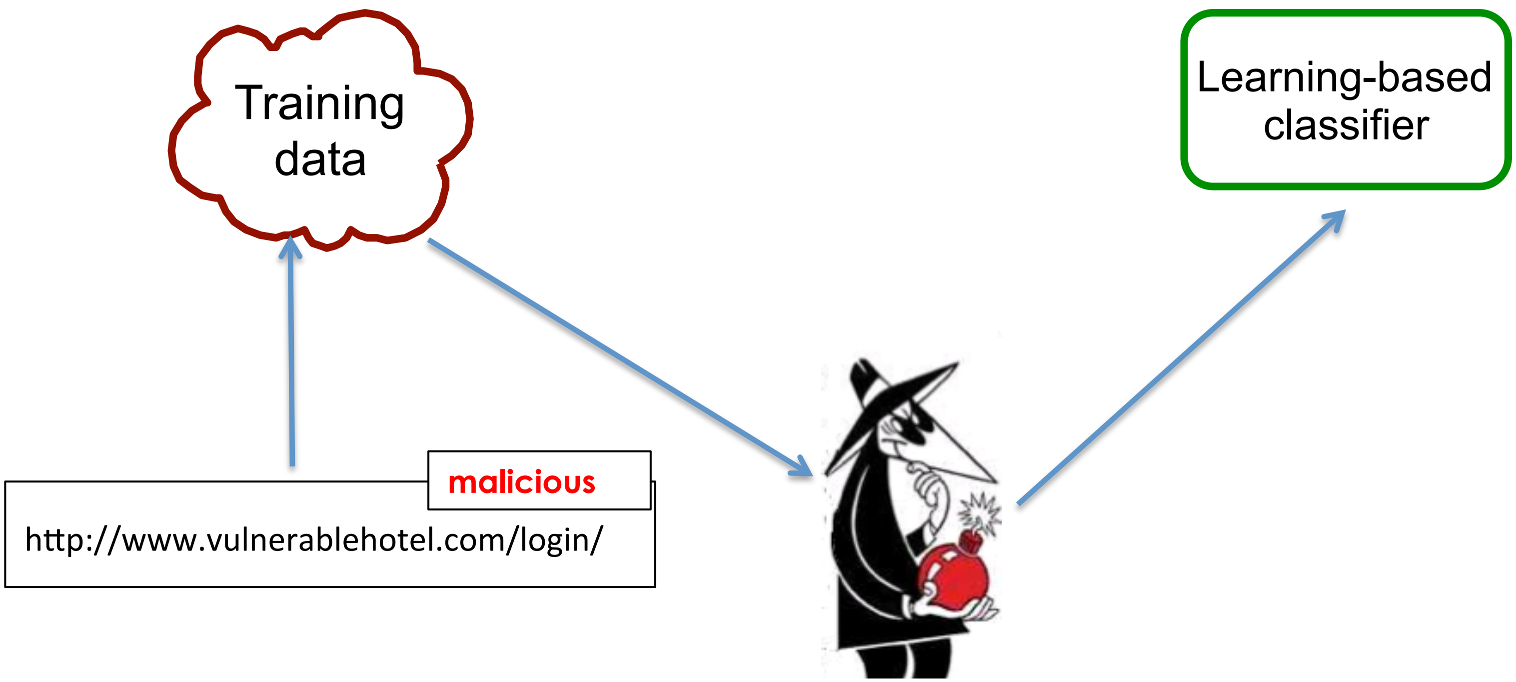
Adversary's influence

Adversary's knowledge

Adversary's capability

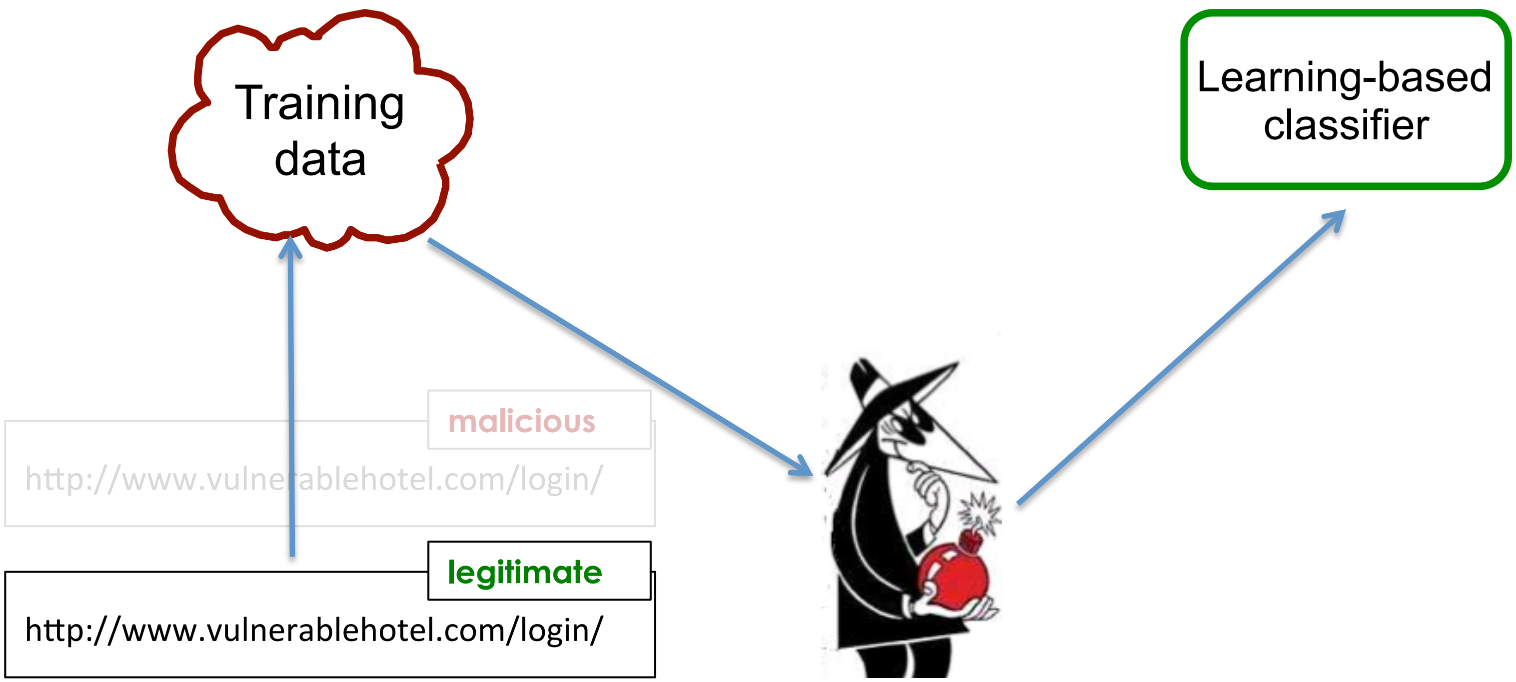
Adversary's influence [M. Barreno et al., 2010]

Attack at training time ("poisoning")



Adversary's influence [M. Barreno et al., 2010]

Attack at training time ("poisoning")



Adversary's influence [Barreno et al., 2010]

Attack at test time ("evasion")

Camouflaged input data

Buy Vi@gra

Classifier

Classifier error types

[R. Lippmann, Dagstuhl Workshop , Sept. 2012]

Classifier output

Truth		Normal	Attack
	Normal	OK	False Alarm
	Attack	Miss Alarm	OK

Adversary's goal

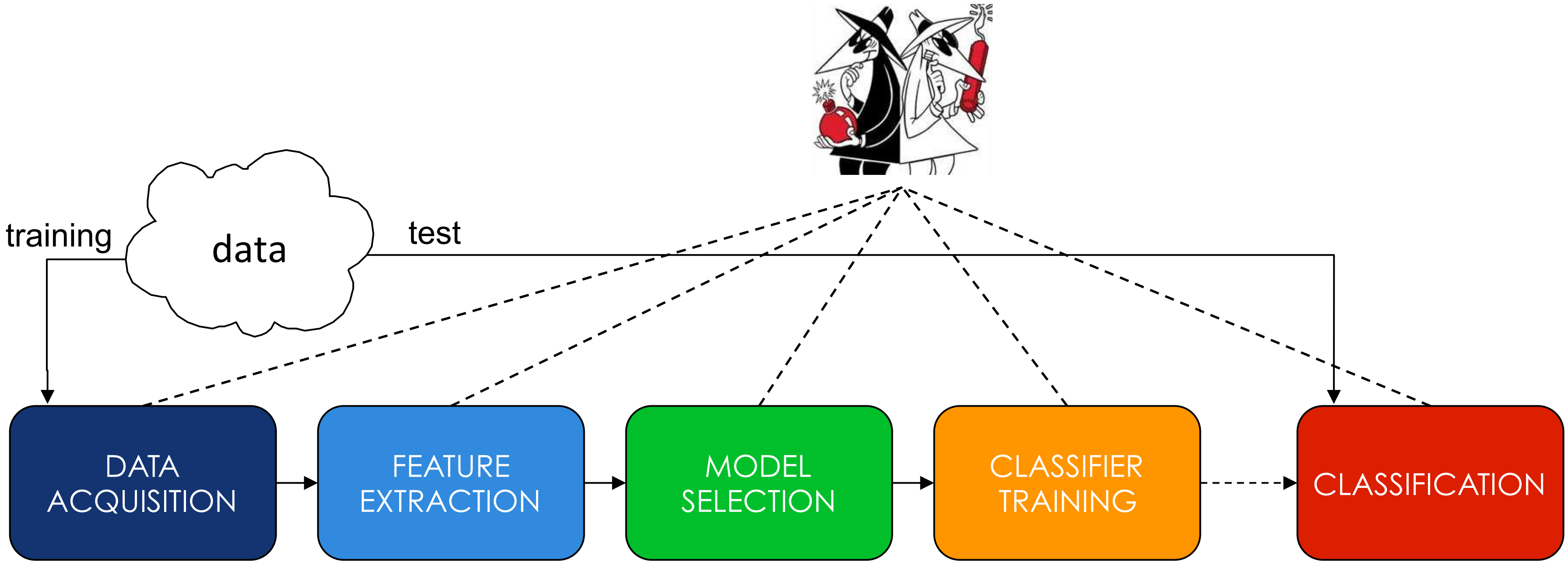
[R. Lippmann, Dagstuhl Workshop , Sept. 2012]

Classifier output

Truth		Normal	Attack	
	Normal	OK	False Alarm	Denial of service (DoS) attack
	Attack	Miss Alarm Evasion attack	OK	

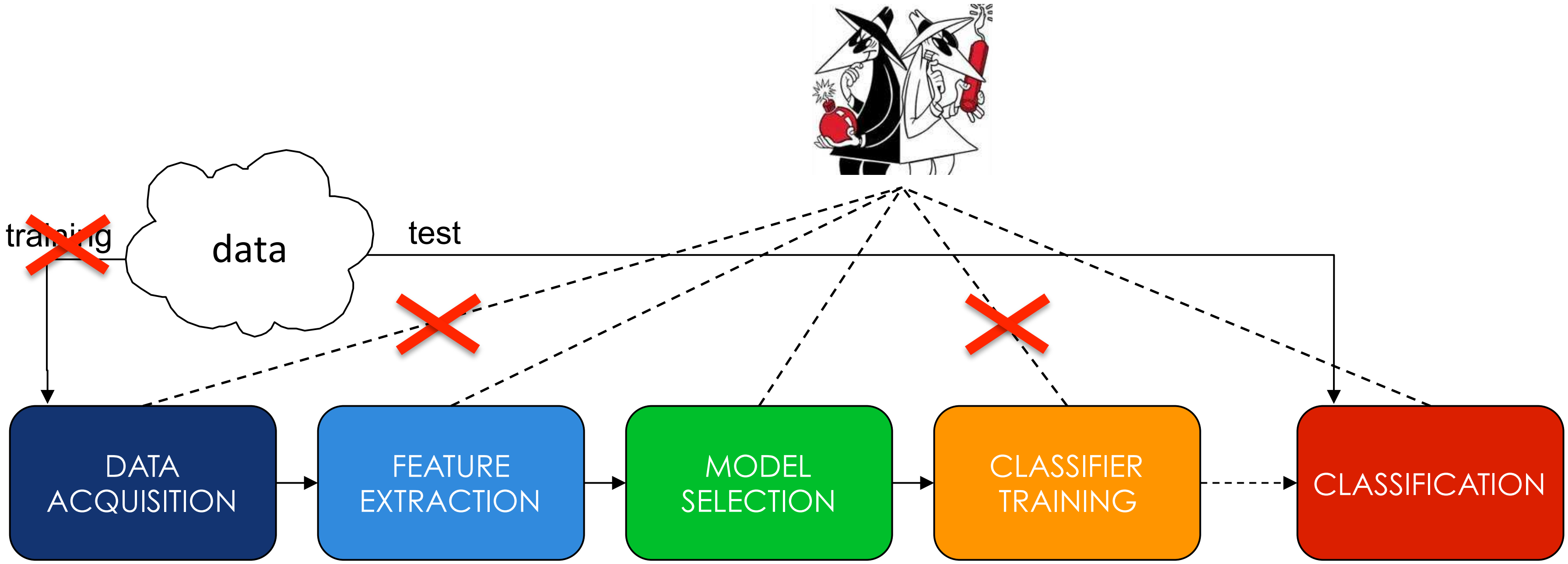
Adversary's knowledge

Complete



Adversary's knowledge

Limited



Adversary's capability

[B. Biggio et al., IET Biometrics, 2012; R. Lippmann, Dagstuhl Workshop , Sept. 2012]

Luckily, the adversary is not omnipotent, she is constrained…

Email messages must be understandable by a human reader

Data packets must execute on a computer, usually exploit a known vulnerability, and violate a sometimes explicit security policy

Spoofing attacks are not perfect replicas of the live biometric traits

Be proactive

To know your enemy, you must become your enemy
(Sun Tzu, The art of war, 500 BC)

Be proactive

Given this model of the adversary:

Influence: attack at test time

Goal: evasion at test time of anti-spam filter;

Knowledge: limited, the adversary doesn't know the training data used;

Capability: limited, the adversary can modify a limited number of spammy words to keep message's understandability

Be proactive

Given this model of the adversary:

Influence: attack at test time

Goal: evasion at test time of anti-spam filter;

Knowledge: limited, she doesn't know the training data used;

Capability: limited, she can modify a limited number of spammy words to keep message's understandability

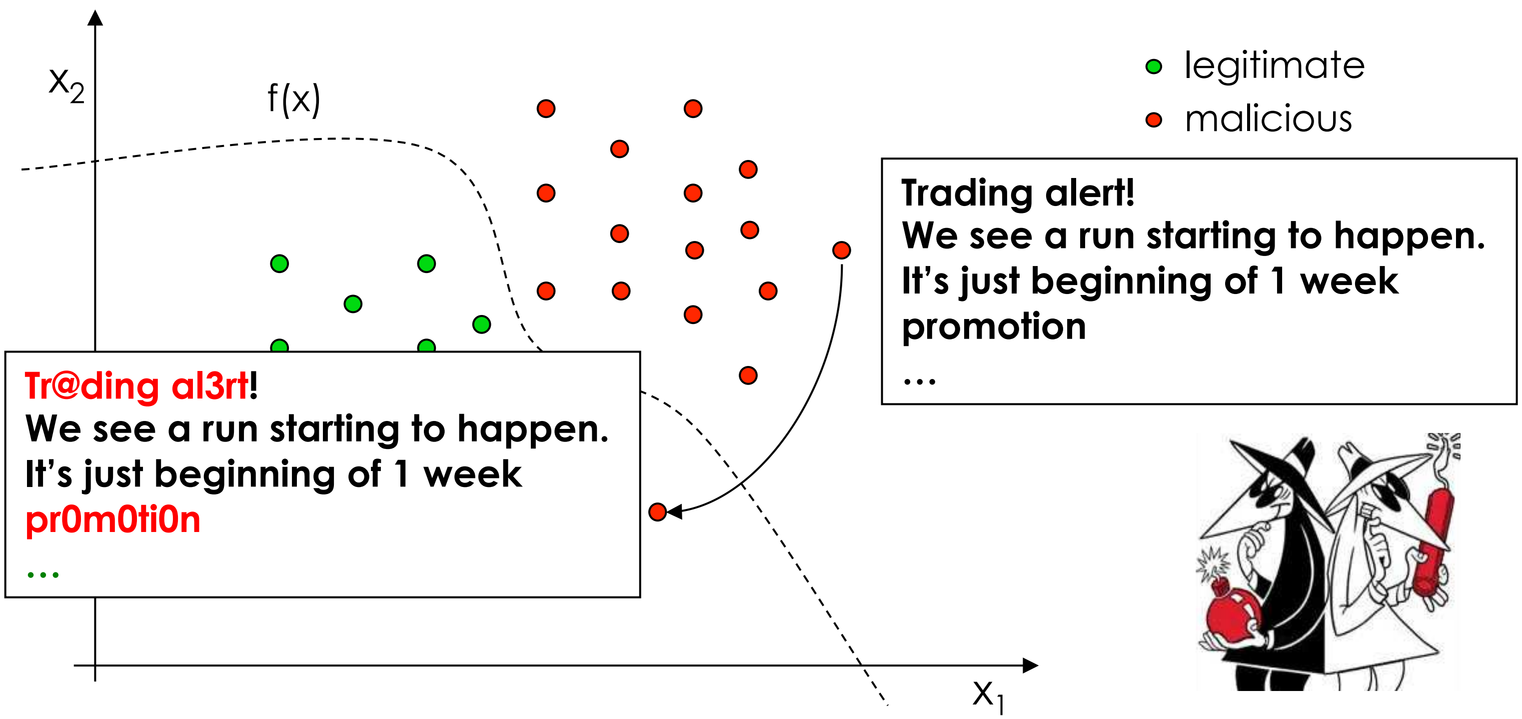
Try to anticipate the adversary !

What is the *optimal* attack she can do, and what is the expected performance decrease of the classifier?

Optimal evasion of pattern classifiers

[B. Biggio, et al., ECML 2013]

Adversary's goal: evasion at test time of anti-spam filter



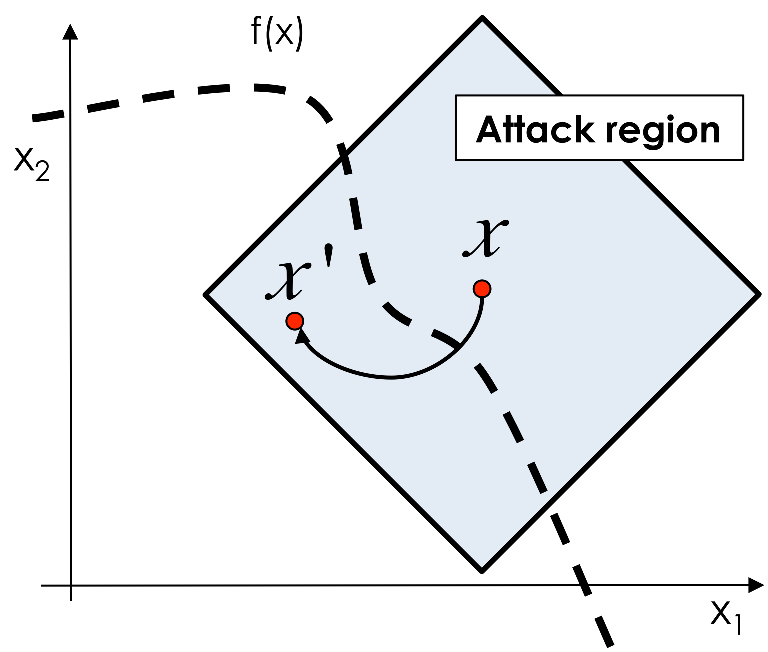
Limited adversary's capability

- **Cost of manipulations**
  - The spammer can modify a limited number of words to keep message's understandability

- **This limitation of capability can be represented by a distance function in feature space (*L_1-norm*)**
  - *e.g.*, Hamming distance, number of words that are modified in spam emails

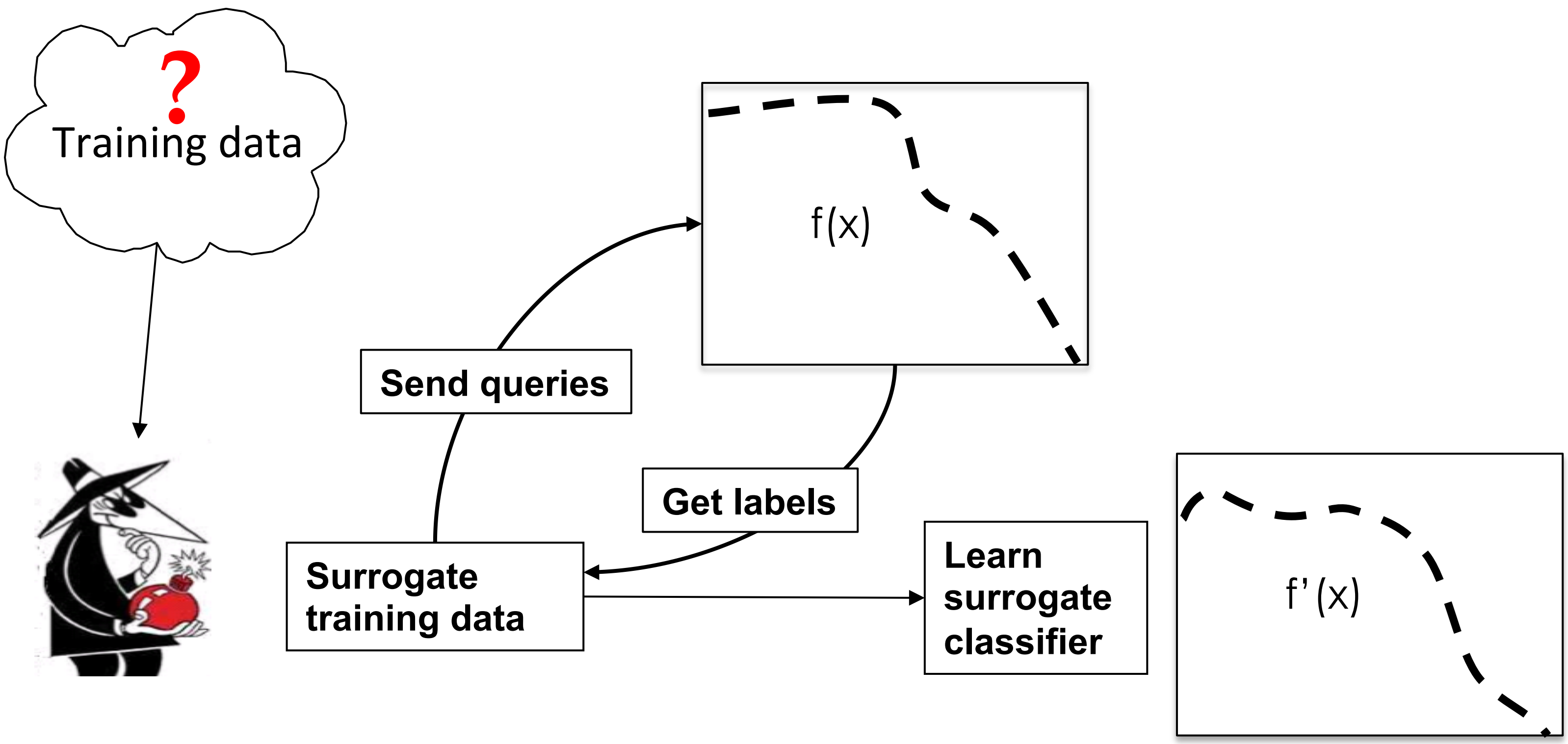
Bounded by a maximum value

$$d(x,x') \leq d_{\max}$$



Limited adversary's knowledge

- Only the features and the learning algorithm are known, **not the training data**
- The adversary can obtain surrogate training data by querying the target classifier
- The adversary can learn a *surrogate* classifier using surrogate training data



Optimal evasion attack [B. Biggio, et al., ECML 2013]

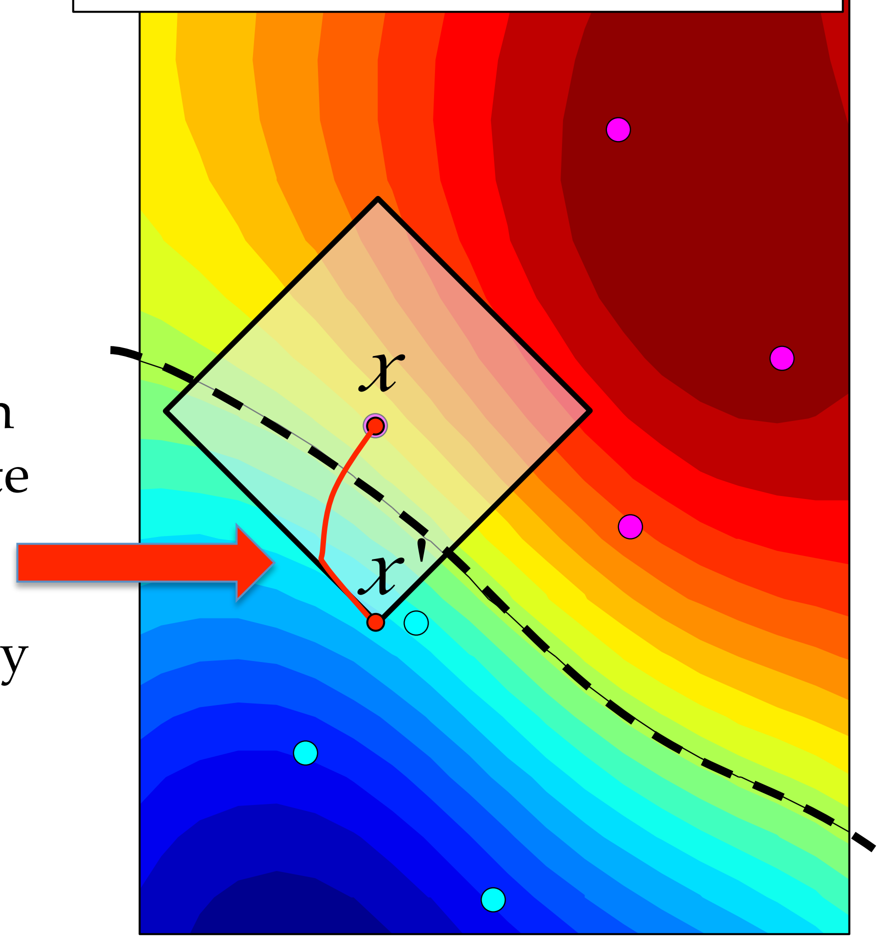
Problem formulation

$$\min_{x'} g(x')$$
$$\text{s.t.} \quad d(x, x') \leq d_{\max}$$

Solution

- Non-linear, constrained optimization
  - Gradient descent: approximate solution for *smooth* functions
- Gradient of g(x) can be analytically computed in many cases
  - SVMs, Neural networks

$$f(x) = \text{sign}\big(g(x)\big) = \begin{cases} +1, & \text{malicious} \\ -1, & \text{legitimate} \end{cases}$$



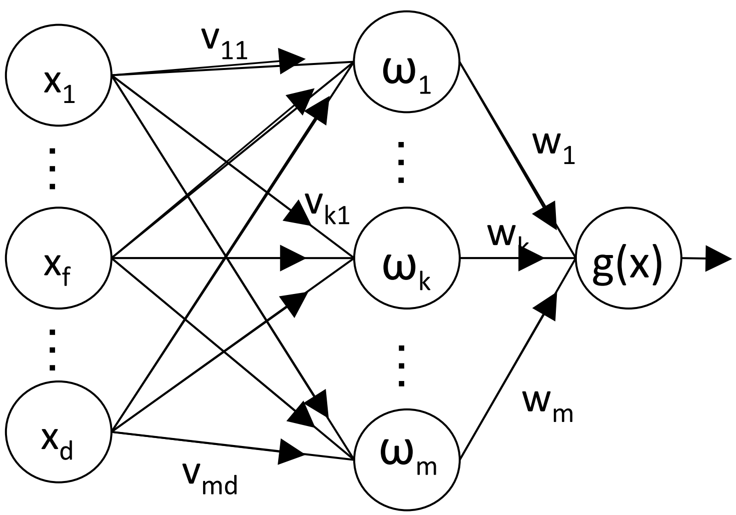
Computing descent directions

Support vector machines

$$g(x) = \sum_i \alpha_i y_i k(x, x_i) + b, \quad \nabla g(x) = \sum_i \alpha_i y_i \nabla k(x, x_i)$$

RBF kernel gradient: $\nabla k(x, x_i) = -2\gamma \exp\left\{-\gamma \| x - x_i \|^2\right\}(x - x_i)$

Neural networks



$$g(x) = \left[1 + \exp\left(-\sum_{k=1}^{m} w_k \delta_k(x)\right)\right]^{-1}$$

$$\frac{\partial g(x)}{\partial x_f} = g(x)\left(1 - g(x)\right) \sum_{k=1}^{m} w_k \delta_k(x)\left(1 - \delta_k(x)\right) v_{kf}$$

Optimal evasion attack: the algorithm

[B. Biggio, et al., ECML 2013]

Input: $\mathbf{x}^0$, the initial attack point; t, the step size; λ, the trade-off parameter; $\epsilon > 0$ a small constant.

Output: $\mathbf{x}^*$, the final attack point.

1: $m \leftarrow 0$.
2: **repeat**
3: $\quad m \leftarrow m + 1$
4: $\quad$ Set $\nabla F(\mathbf{x}^{m-1})$ to a unit vector aligned with $\nabla g(\mathbf{x}^{m-1})$.
5: $\quad \mathbf{x}^m \leftarrow \mathbf{x}^{m-1} - t\nabla F(\mathbf{x}^{m-1})$
6: $\quad$ **if** $d(\mathbf{x}^m, \mathbf{x}^0) > d_{\max}$ **then**
7: $\quad\quad$ Project $\mathbf{x}^m$ onto the boundary of the feasible region.
8: $\quad$ **end if**
9: **until** $F(\mathbf{x}^m) - F(\mathbf{x}^{m-1}) < \epsilon$
10: **return:** $\mathbf{x}^* = \mathbf{x}^m$

Gradient descent

Hackers love optimisation problems…

[B. Biggio et al., AISec 2013]

Attacker's goal

Knowledge of the data, features, …

$$\max_{A'} E_{\theta \sim \mu} \left[g\left(A'; \theta\right) \right]$$

$$\text{s.t.} \;\; A' \in \Omega(A)$$

Capability of manipulating the input data

Hackers love optimisation problems…

[B. Biggio et al., AISec 2013]

Attacker's goal

Knowledge of the data, features, …

$$\max_{A'} E_{\theta \sim \mu}\left[g\left(A'; \theta \right) \right]$$

$$\text{s.t.} \quad A' \in \Omega(A)$$

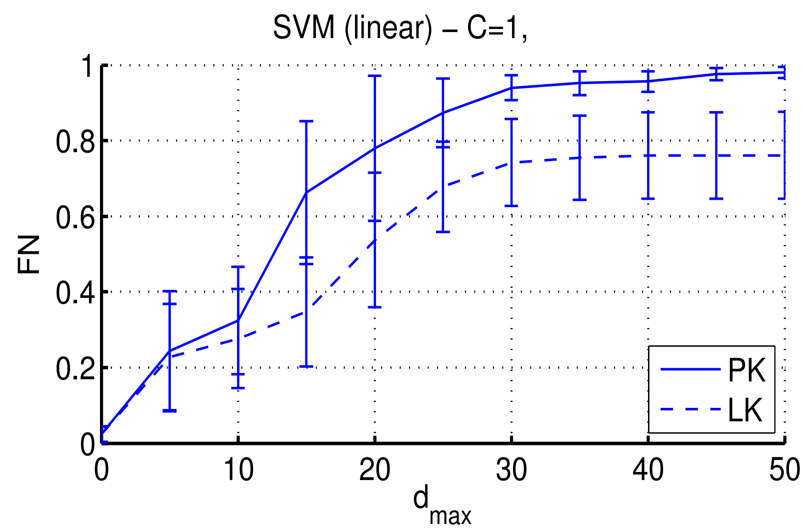
Capability of manipulating the input data

- Attacking SVM at training time [B. Biggio et al., ICML 2012]
- Poisoning clustering algorithms [B. Biggio et al., AISec 2013]
- Poisoning feature selection algorithms [H. Xiao et al., ICML 2015]
- Poisoning face recognition systems [B. Biggio et al., IEEE SP Mag 2015]

Take-home message

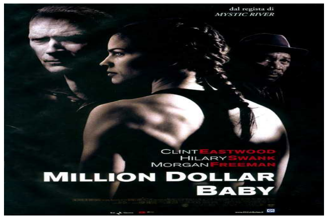
- We did experiments on many tasks (spam filtering, PDF malware detection, intrusion detection in computers, biometric recognition) with many linear and non linear classifiers [Biggio et al., ECML 2013, KDE]

➢ Linear and non-linear classifiers can be highly vulnerable to well-crafted evasion and poisoning attacks



Be proactive: design defences !

Protect your classifier



What is the rule? The rule is protect yourself at all times (from the movie "Million dollar baby", 2004)

Good defence strategies

[B. Biggio et al., 2013; R. Lippmann, Dagstuhl Workshop , Sept. 2012]

Information hiding

Robust classifiers

Frequent adaptation

Information hiding

Deny Access

- ✓ Don't give access to your training data
- ✓ Don't provide classifier's output
- ✓ …

Use multiple classifiers

- ✓ To make the classifier difficult to reverse engineer

Randomization

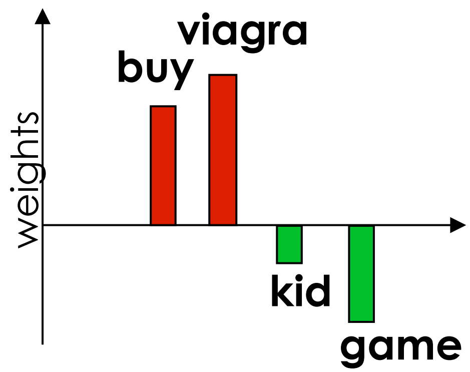
- ✓ Randomize classifier's parameters, features, training data, …,

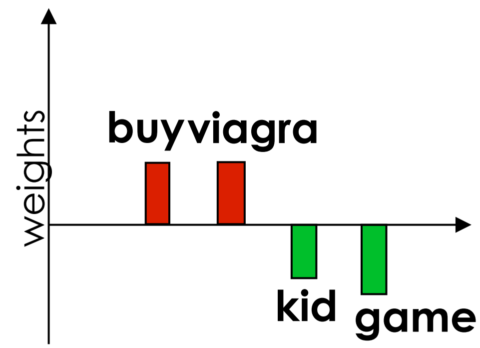
Design of robust classifiers

Feature equalization

Avoid to over emphasize features

[Kolcz and Teo, CEAS 2009; B. Biggio et al., IJMLC 2010]





Robust learning

Use robust algorithms against attacks at training time

[B. I. Rubinstein et al., ACM 2009; Biggio et al., ICB 2013]

Evade hard multiple classifiers

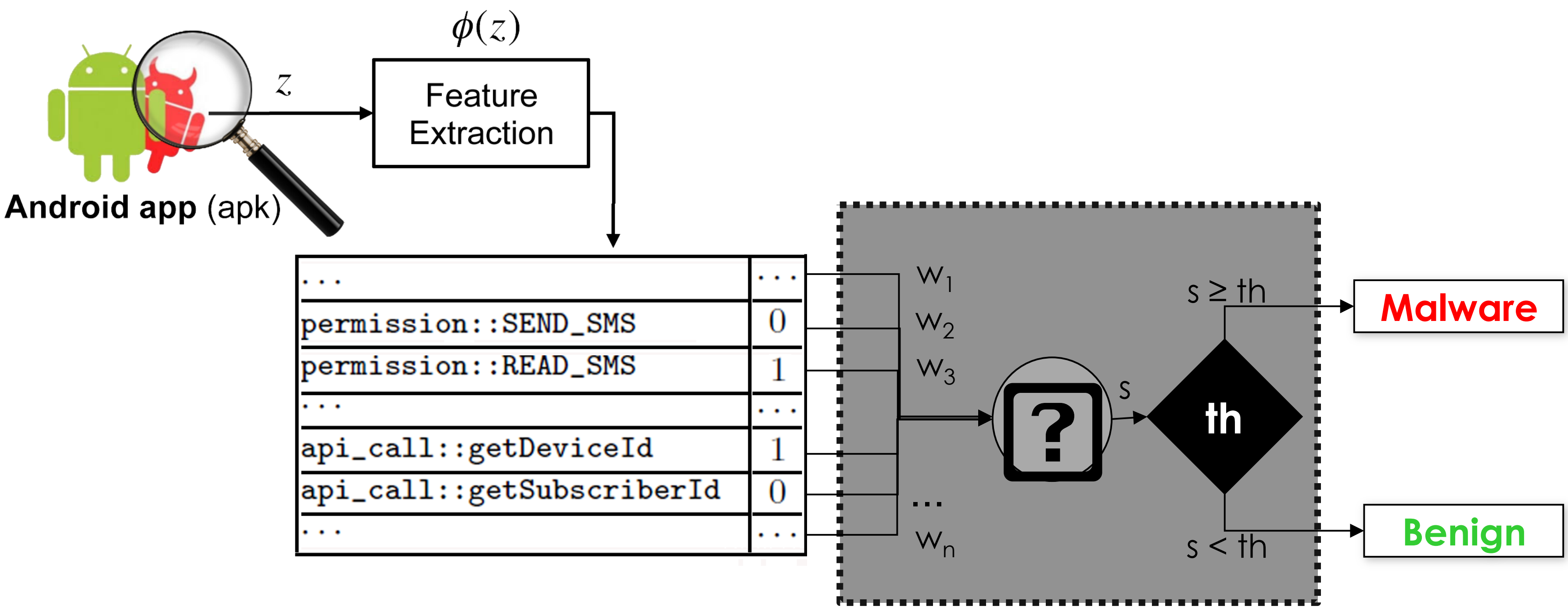
[Biggio, Fumera, Roli, SUEMA 2008]

Android malware detection

- ✓ About one billion of users of Android mobile operating system
- ✓ Thousands of new Android malware samples every day

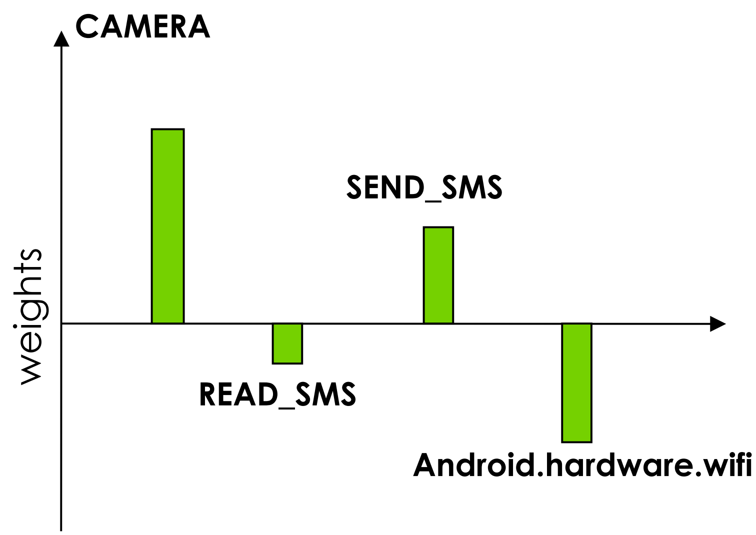
DEBRIN: Android malware detector

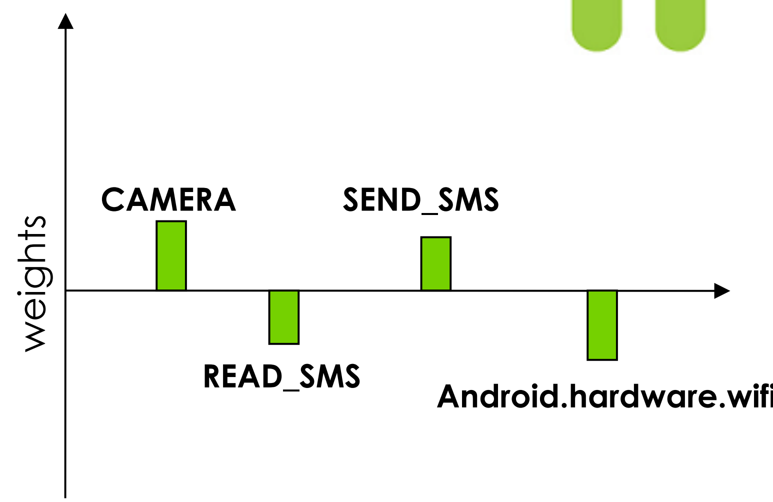
[D. Arp et al., NDSS 2014]



Attacking a linear classifier

Attacker's goal: evasion $$x'^{\star} = \arg\min_{x'} \hat{w}^{\top} x'$$





The Attacker can evade easily the classifier by manipulating a few features if weights are sparse !

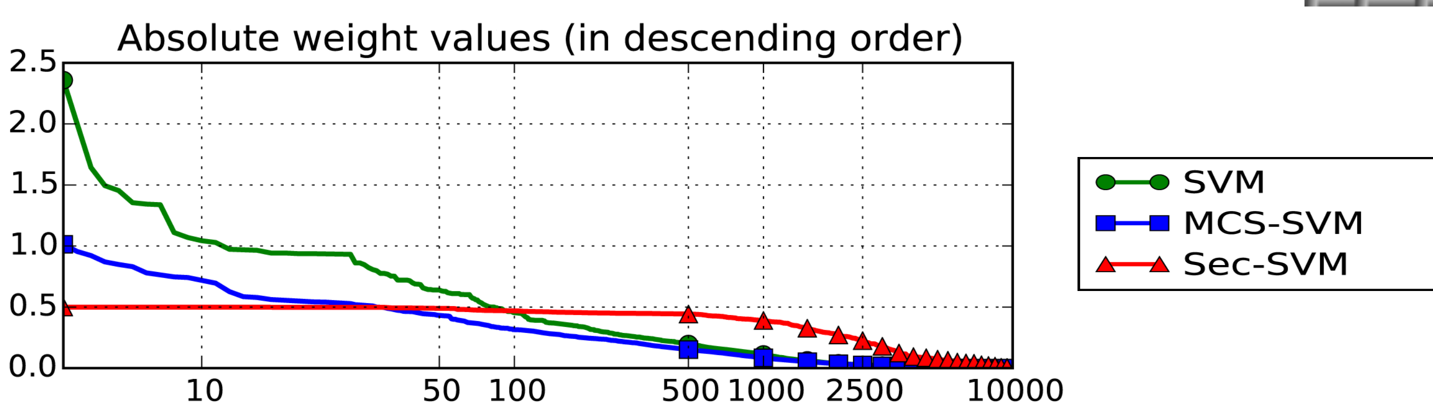
Securing linear classifiers

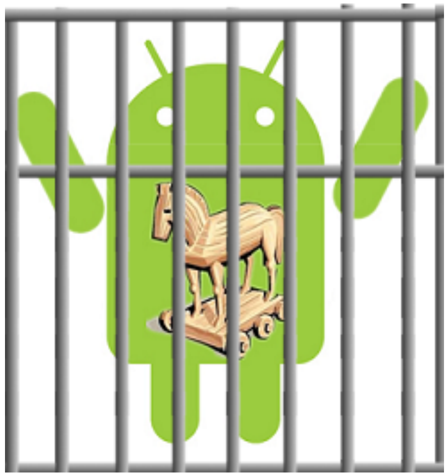
[A. Demontis et al., 2016]

We learn feature weights evenly-distributed by solving this optimization problem:

$$\min_{w,b} \quad \frac{1}{2} w^{\top} w + C \sum_{i=1}^{n} \max(0, 1 - y_i f(x_i))$$

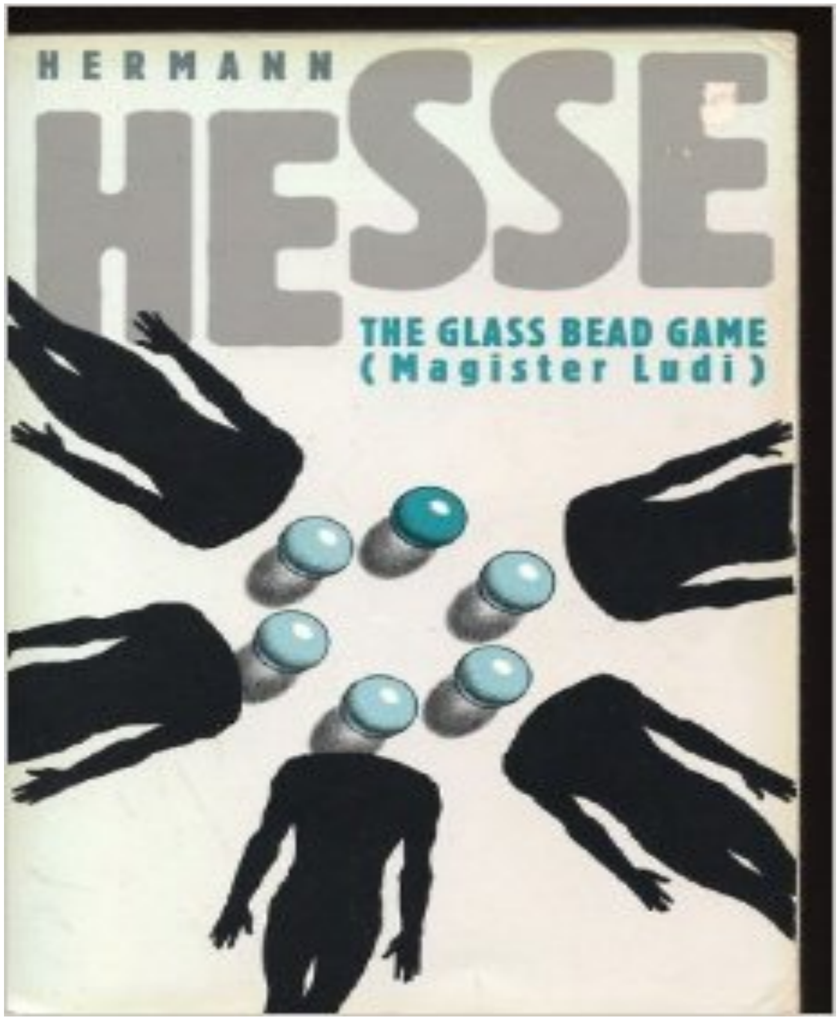
$$\text{s.t.} \quad w_k^{\text{lb}} \leq w_k \leq w_k^{\text{ub}}, \, k = 1, \ldots, \text{d}.$$





To conclude…

Magister Ludi ?



- *Why do you do research on such a niche of the pattern recognition field?*
- *When many fundamental problems are still unsolved…*

Pattern recognition in 2021…

As pattern recognition systems will become more and more pervasive, the economic incentives of bad guys to attack them will increase…

But also good guys will be more and more motivated to evade them….to preserve their privacy…

Evasion of face recognition systems

The **CV Dazzle** project (http://cvdazzle.com)

How to preserve your privacy from face recognition systems?

Faces detected

Evasion of face recognition systems

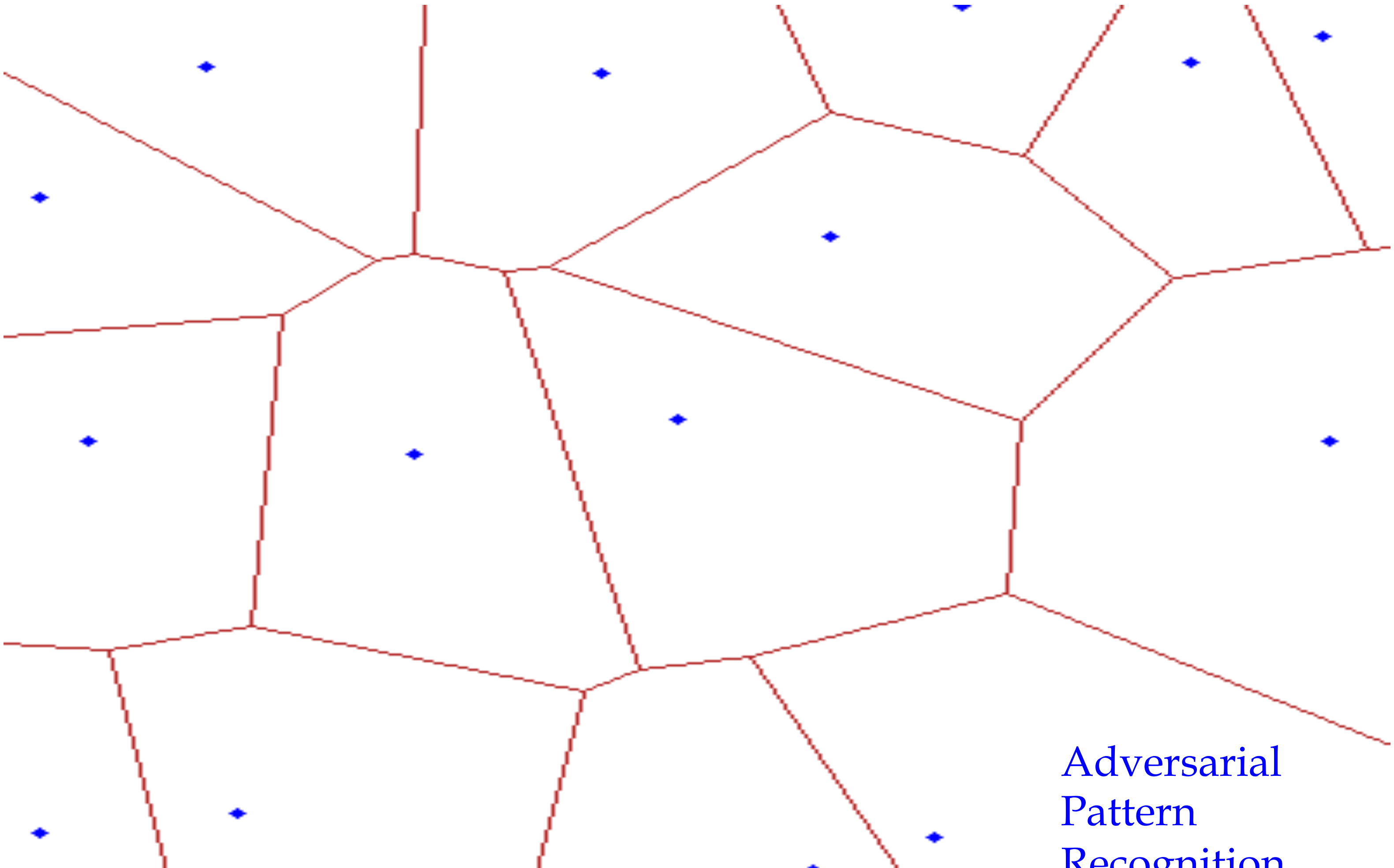
Use this makeup and hair styling to evade face recognition…

Faces detected

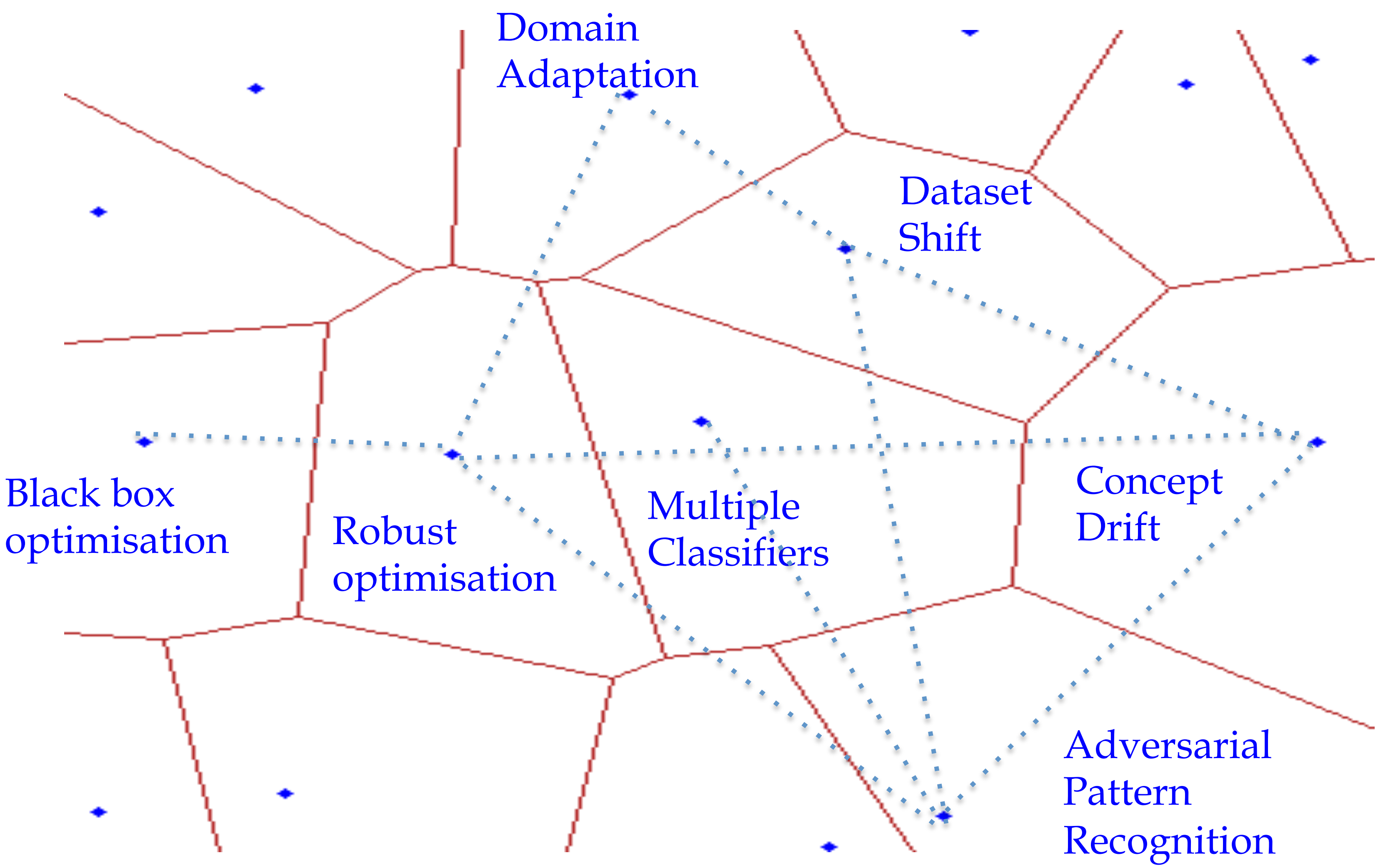
No faces detected

The **CV Dazzle** project (http://cvdazzle.com)

Connecting the dots for having impact…

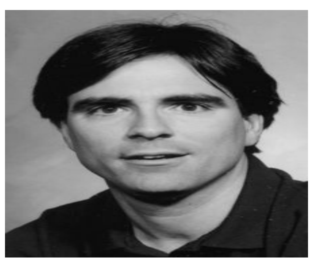


Connecting the dots for having impact…



Thanks for listening !

Any questions ?



Engineering isn't about perfect solutions; it's about doing the best you can with limited resources (Randy Pausch)