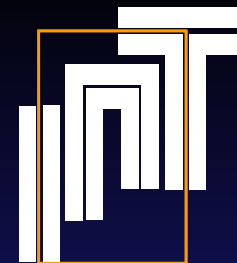




ΕΚΕΦΕ «Δημόκριτος»

Ινστιτούτο Πληροφορικής & Τηλεπικοινωνιών



## *Bridging self-regulation and content-filtering*

Internet Content Filtering Group

Software and Knowledge Engineering Lab

<http://www.iit.demokritos.gr/skel/i-config/>



## Research activities of *i-config*

- Language Technology
- Image Understanding
- Knowledge Discovery in Data
- User Modeling
- Multimedia Information Processing



# Core technologies and applications

## ■ Technologies:

- ◆ Information extraction
- ◆ Information filtering

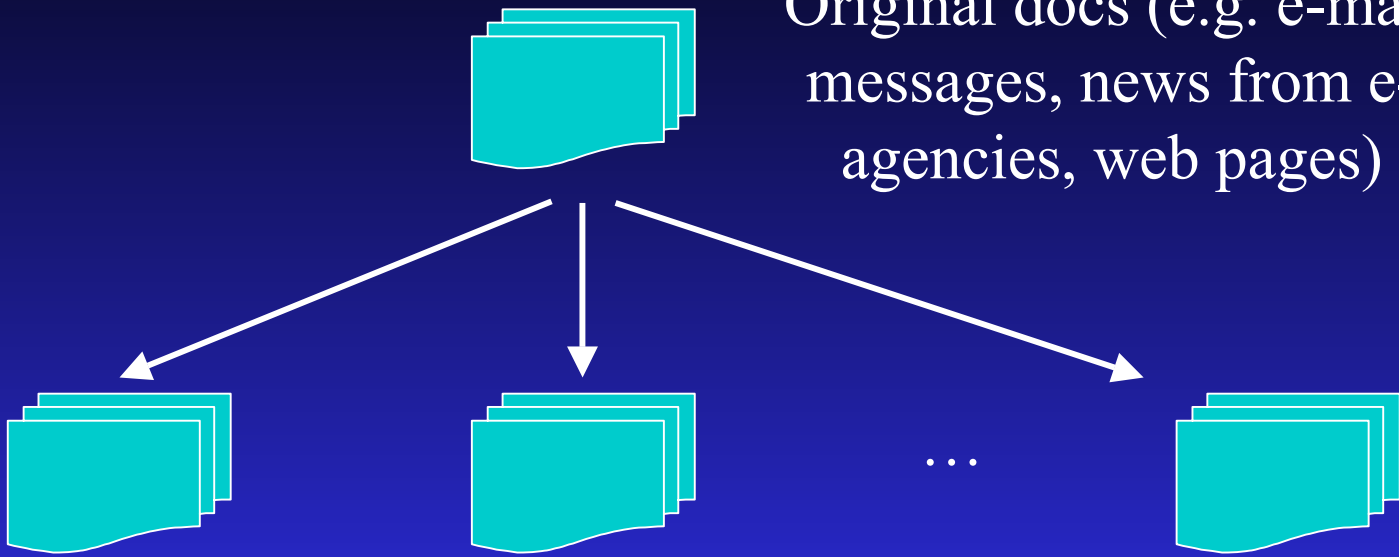
## ■ Filter generator:

- ◆ Web filtering
- ◆ Spam filtering (e-mail)
- ◆ Classifying financial news



# Filtering digital content

Original docs (e.g. e-mail messages, news from e-agencies, web pages)



category 1

(e.g. complaints,  
financial news)

category 2

(e.g. tech support,  
sports news)

category n



# Internet Content

- ~98% “safe”
- Uncontrolled
- Total and direct access
- Lack of provenance (relative to absolute)
- High volatility



# Unsafe content

- Illegal
  - ◆ Pedophiles, Nazism (DE)
- Offensive
  - ◆ Pornography, Racism, Violence
- Undesired
  - ◆ Online gambling
  - ◆ Day trading sites



# Where and how to filter

- Self-regulation
- Filtering at the source
- Filtering during distribution
- Filtering at the last mile



# Self-regulation

- Self-labeling by content authors – producers
- Browsers block according to user settings

ICRA v.1.0/RSACi : (n 3 s 4 v 0 1 4)

New, more expressive vocabulary incl. context :  
ICRA v. 2.0 (<http://www.icra.org/faq/decode/>)





# Filtering at the source – distribution

- Literally impossible due to network structure, lack of provenance and routing method (cf. legal case against Yahoo! France)



# Filtering at the last-mile (‘consumer’)

- List-based solutions
  - ◆ underblocking
- Shallow keyword matching solutions
  - ◆ overblocking



# FilterX: Web page filtering

*FilterX* is a Web proxy server that filters pornographic content on the Web. Having been trained with suitable examples, *FilterX* operates in real time.



- Combining *natural language processing*, *image analysis* and *Web structure*, *FilterX* analyses all the information available on the HTTP stream, not just the URL or title.
- Using *machine learning*, *FilterX* considers the actual contribution of textual, structural and pictorial features.
- Creating a *multimedia representation model*, for each document *FilterX* achieves practically zero overblocking.

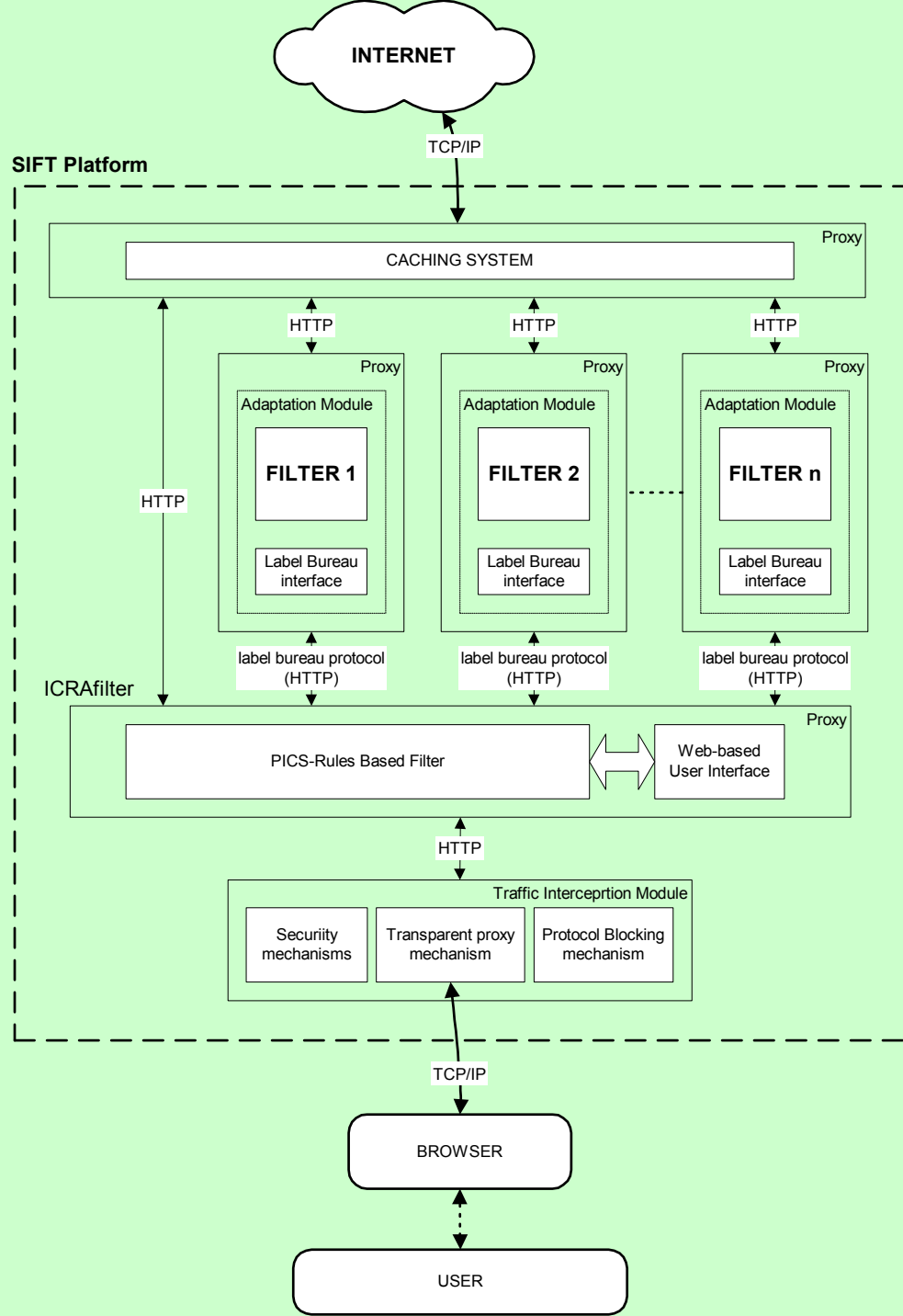


# Last-mile applications of FilterX

- Self-regulation and filters
  - ◆ SIFT: Use of filters when self- or 3<sup>rd</sup> party labeling absent / not trusted – resulted in *ICRAplus*, a free platform bridging self-regulation with filtering software
- Protection of young students
  - ◆ SCOFI: Different content access according to student age via smartcard



- Communication through W3C standards (HTTP, Label Bureau)
- Easy to implement public API
- Literally ANY filter can become module
- High security - DSig
- Will run on Win OS
- <http://www.icra.org/icraplus/>








Address <http://www.sex.com/s.html?cn=greece>

# ICRAplus

**The requested page has been blocked**

*The filters' responses are:*

Logo	Filter	Response	Explanation
	embedded	Don't know	The filter cannot make a decision
	optenet	Block	blocked by internal configuration parameters <a href="#">More...</a>
	filterX	Block	Page <i>probably</i> contains obscene content <a href="#">More...</a>

Password

Add to allow list

- In the absence of author-provided ICRA labels, the requested page can still be blocked by co-operating filters
- Public API can help filter vendors wrap their existing software into an ICRAplus module in no time.
- Users can override blocks temporarily or permanently



Find new filters      Activate/deactivate filters      Help

## Filter Combination for the active profile

ICRAplus allows you to decide how the different installed filters will be combined to give an overall response. If you choose one of the preset options, all filters are given an equal vote. The more strict the preset option you choose, the more filters have to vote "Allow" before access is granted.

Alternatively, you can manually determine the weight given to each filter, what to do in the event of a "tie" and so on. Select 'Use advanced' and then click the 'Advanced' button to do this.

**Very high**   
**High**   
**Medium**   
**Low**   
**Very Low**   
**Use advanced**

Advanced

■ Filters can be combined per user profile in either a simple manner, with predetermined, factory settings

■ Or....






Find new filters

Activate/deactivate filters

Help

## Filter Combination for the active profile

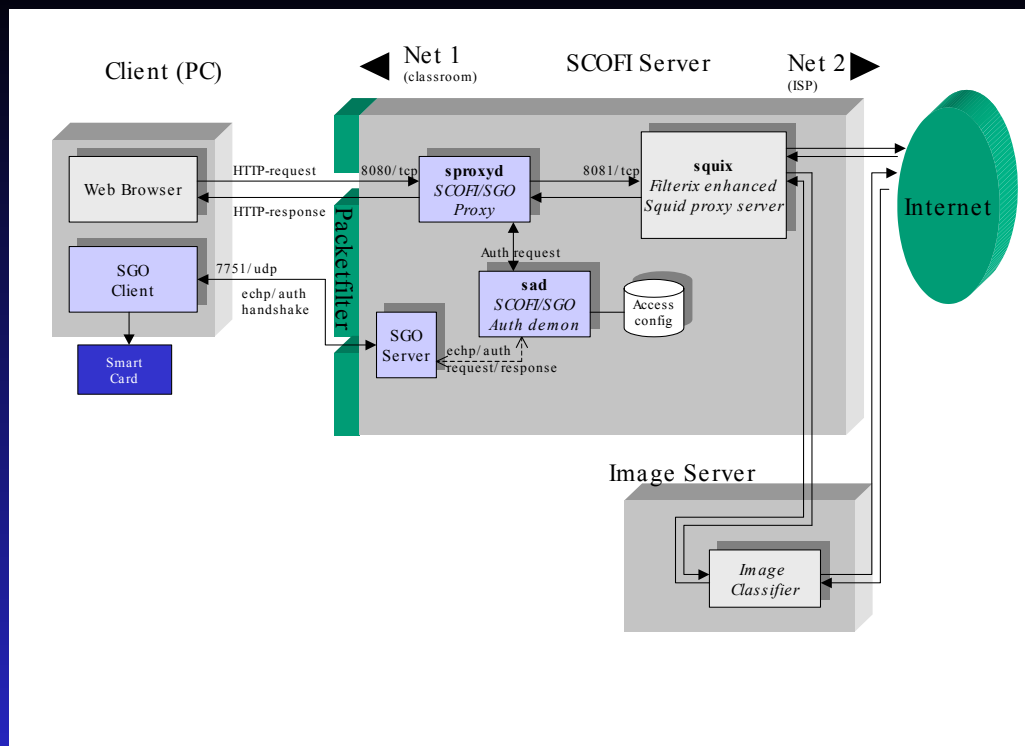
Filter Name	Response	Global Response					Weight	Priority
		Ignore	Allow	Block	Vote			
 <b>INTERNET CONTENT RATING ASSOCIATION</b>	Allow	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="text" value="1"/>	1	
	Block	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="text"/>		
	Don't know	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="text"/>		
 <b>OPTEINET.com</b> OPTIMAL INTERNET	Allow	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="text" value="1"/>	2	
	Block	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="text"/>		
	Don't know	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="text"/>		
 <b>i-config</b>	Allow	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="text" value="1"/>	3	
	Block	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="text"/>		
	Don't know	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="text"/>		
If the global response is Ignore, then:				<b>Allow</b>	<b>Block</b>			
				<input checked="" type="radio"/>	<input type="radio"/>			

Apply



- ..users can have full control on the filtering procedure, incl. filter priority, filter weights and tie resolution!
- More info and downloads at <http://www.icra.org/icraplus/>





- SmartCard based authentication and age profile
- Different levels of filtering
- High security - Dsig
- External image analysis





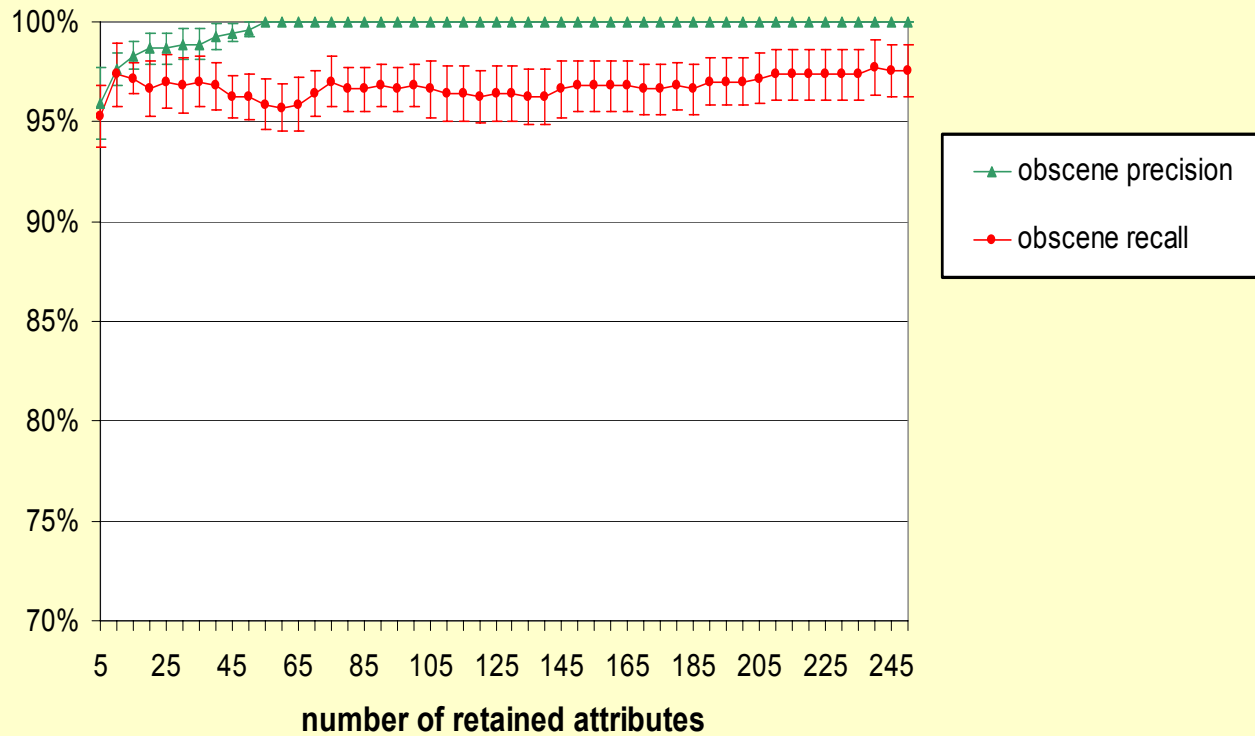
# FilterX revisited

- Trained on self-proclaimed porn sites
- Creation of page-level representation (multimedia and structural)
- Turn “noise” to our advantage by using it as feature
- Models per language + language identifier
- Evaluated using multi-fold cross-validation (before SIFT & SCOFI)
- Can pass the decision to the user for thresholding



# FilterX Results

Filtering obscene Web pages



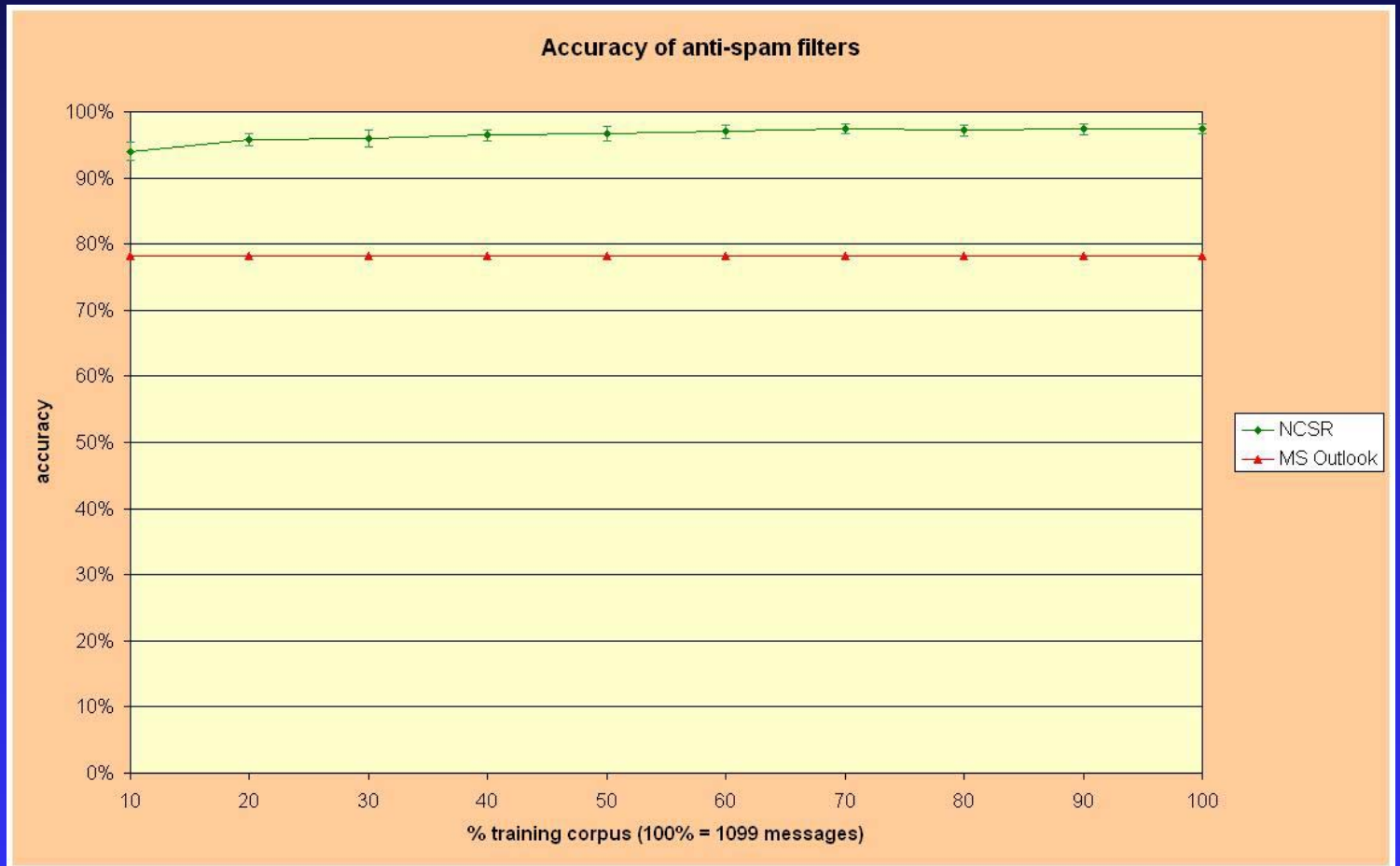


# A case for email

- Instead of learning what is spam, learn what is legit
- User profile/model based on user's Inbox + spam
- Turn “noise” to our advantage by using it as feature
- Models per language + language identifier
- Evaluated in house + multi-fold cross-validation



# Spam filter results





# Food for thought

- Decide where to put the bias: overblocking vs. underblocking
- Most of the harmful content *wants* to be found
- Possibility of “hostile” users! (enforced s/w!)
- Very hard to detect *intention* of the content author
- Reverse the problem of filtering by centering on the effect on the user
- Technology only to solve problems it has created!



## More info

Konstantinos Chandrinos  
kostel@iit.demokritos.gr

*Internet Content Filtering Group (i-config)*  
*NCSR "Demokritos"*

<http://www.iit.demokritos.gr/skel/i-config/>