PAN 2010 Results

Uncovering Plagiarism, Authorship, and Social Software Misuse

Bauhaus-Universität Weimar – Martin Potthast, Benno Stein
Andreas Eiselt, Teresa Holfeld
Universidad Politécnica de Valencia – Alberto Barrón-Cedeño, Paolo Rosso
University of the Aegean – Efstathios Stamatatos
Bar-Ilan University – Moshe Koppel

http://pan.webis.de

The PAN Competition

The PAN Competition

2nd International Competition on Plagiarism Detection, PAN 2010

These days, plagiarism and text reuse is rife on the Web.

Task:

> Given a set of suspicious documents and a set of source documents, find all plagiarized sections in the suspicious documents and, if available, the corresponding source sections.

The PAN Competition

2nd International Competition on Plagiarism Detection, PAN 2010

These days, plagiarism and text reuse is rife on the Web.

Task:

> Given a set of suspicious documents and a set of source documents, find all plagiarized sections in the suspicious documents and, if available, the corresponding source sections.

Corpus:

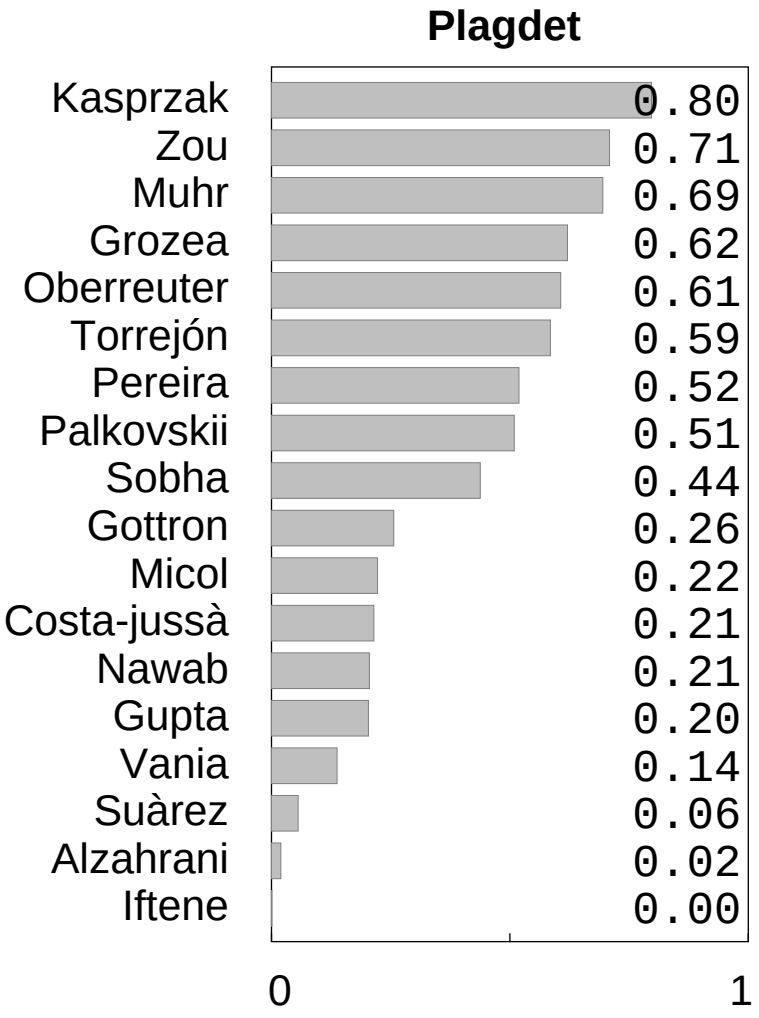
PAN-PC-10

- 27 073 documents (obtained from 22 874 books from the Project Gutenberg)
- 68 558 plagiarism cases (about 0-10 cases per document)
- 6 plagiarism-relevant parameters (length, language, task, obfuscation, topic, fraction)

[Potthast et al., COLING 2010]

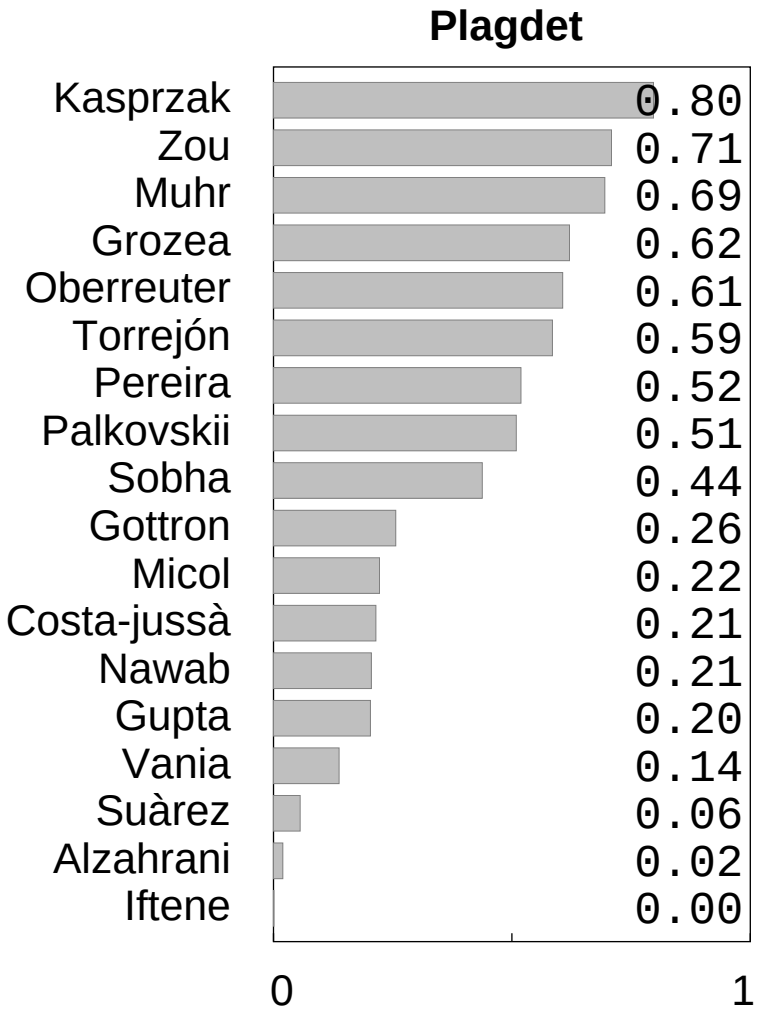
The PAN Competition

Plagiarism Detection Results



The PAN Competition

Plagiarism Detection Results



- Plagdet combines precision, recall, and granularity:

$$\textit{plagdet}(S,R) = \frac{F_1}{\log_2(1+\textit{gran}(S,R))}$$

$$\textit{prec}(S,R) = \frac{1}{|R|}\sum_{r\in R}\frac{\left|\bigcup_{s\in S}(\mathbf{s}\sqcap\mathbf{r})\right|}{|\mathbf{r}|}$$

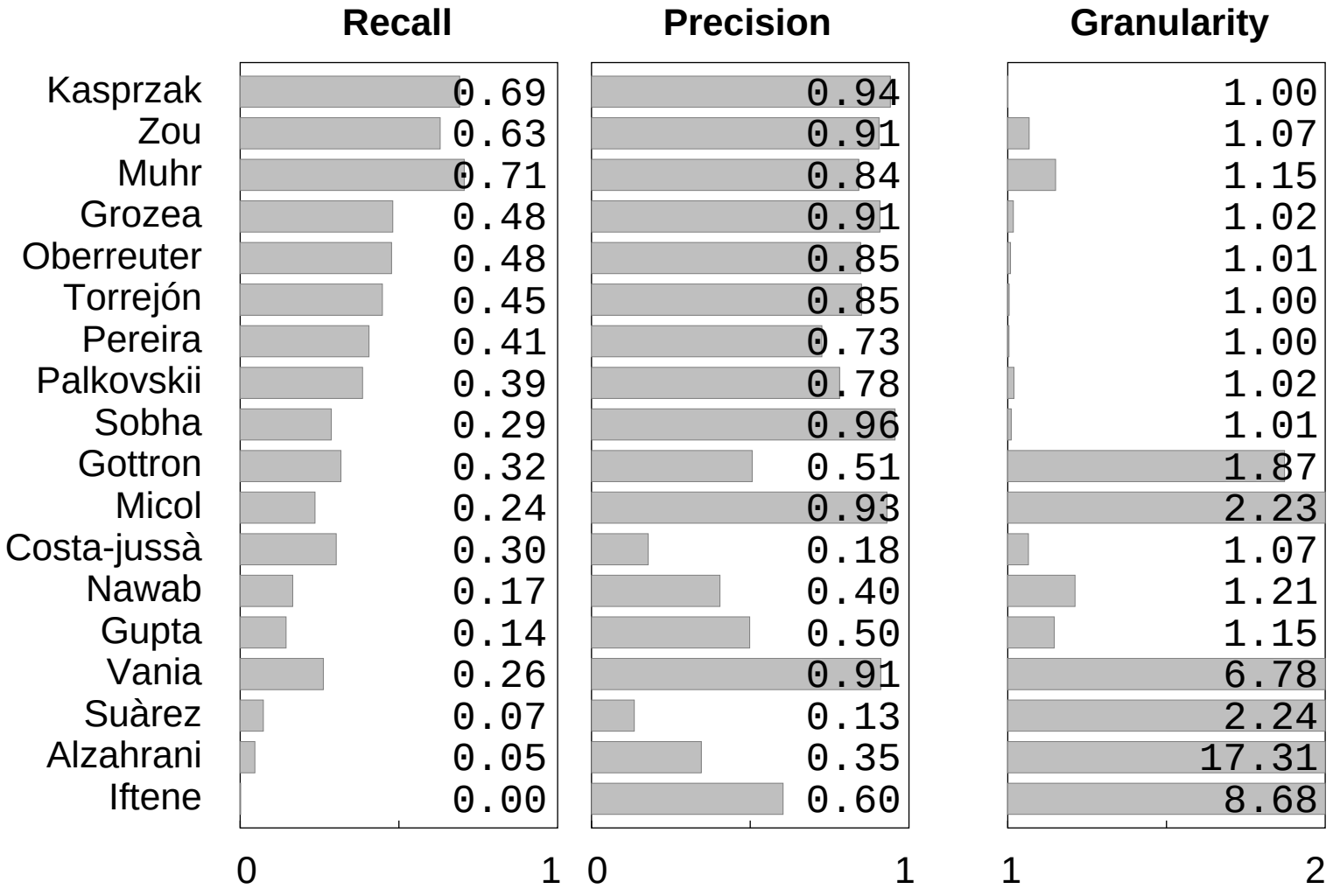
$$\textit{rec}(S,R) = \frac{1}{|S|}\sum_{s\in S}\frac{\left|\bigcup_{r\in R}(\mathbf{s}\sqcap\mathbf{r})\right|}{|\mathbf{s}|}$$

- The granularity *gran* measures the average number of times a plagiarism case is detected.

[Potthast et al., COLING 2010]

The PAN Competition

Plagiarism Detection Results



The PAN Competition

The PAN Competition

1st International Competition on Wikipedia Vandalism Detection, PAN 2010

Every edit on Wikipedia has to be double-checked for integrity—
even if it affects just one char.

Task:

Given a set of edits on Wikipedia articles,
distinguish ill-intentioned edits from well-intentioned edits.

The PAN Competition

1st International Competition on Wikipedia Vandalism Detection, PAN 2010

Every edit on Wikipedia has to be double-checked for integrity—
even if it affects just one char.

Task:

Given a set of edits on Wikipedia articles,
distinguish ill-intentioned edits from well-intentioned edits.

Corpus:

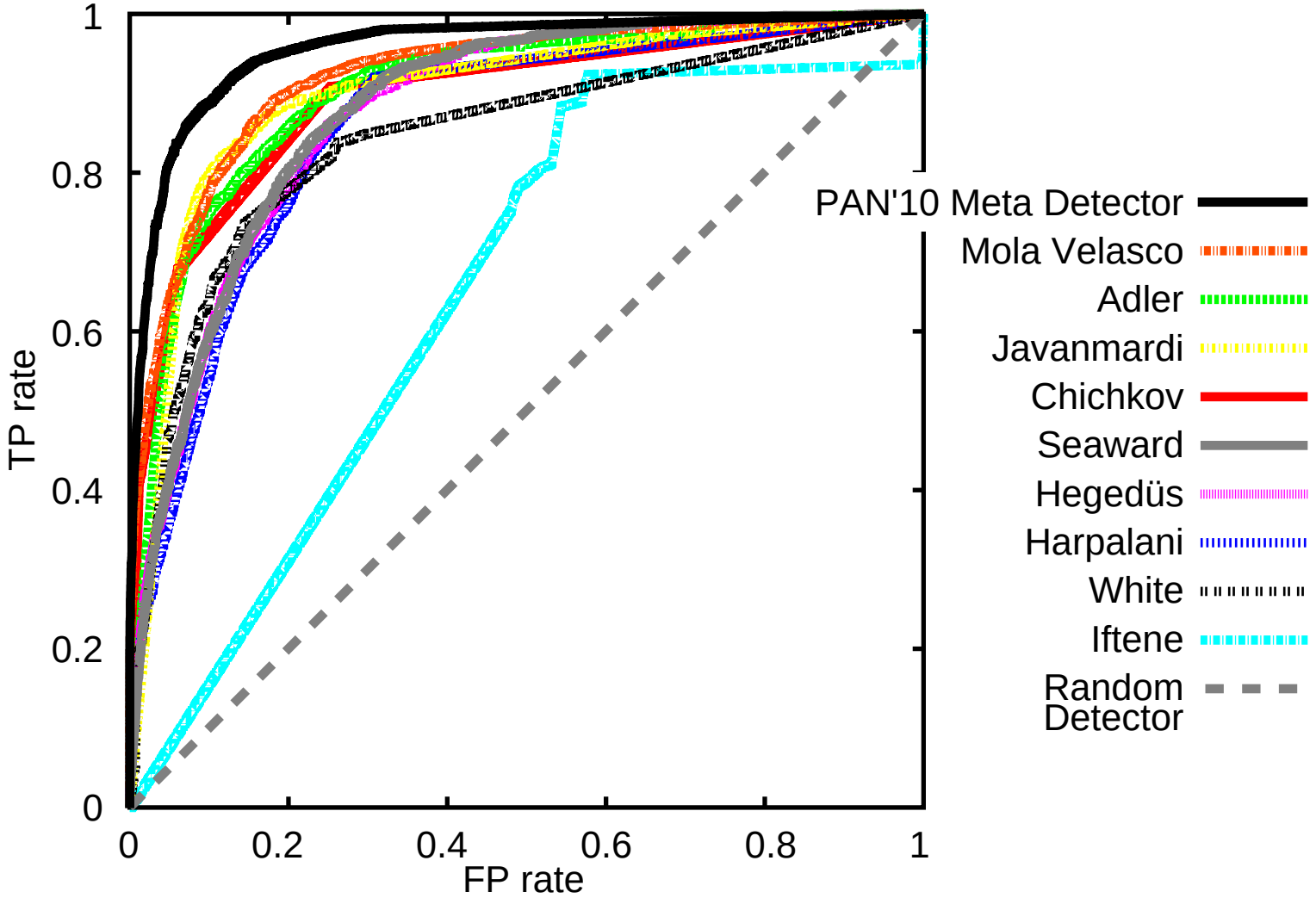
PAN-WVC-10

- 32 452 edits (sampled from a week's worth of Wikipedia edit logs)
- 28 468 different edited articles (edit frequency resembles article importance)
- 2391 edits are vandalism (a 7% ratio is in concordance with the literature)

[Potthast, SIGIR 2010]

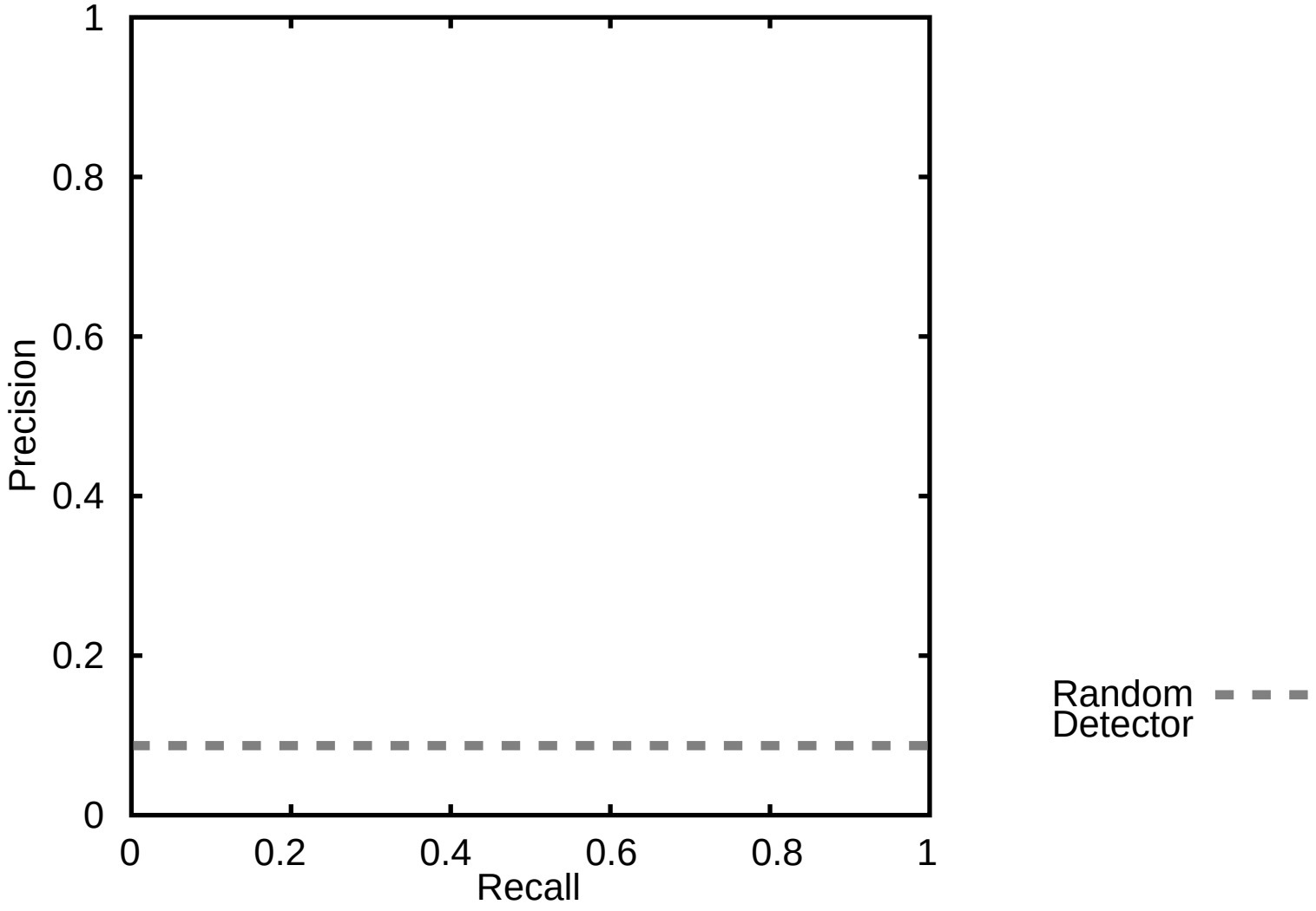
The PAN Competition

Plagiarism Detection Results



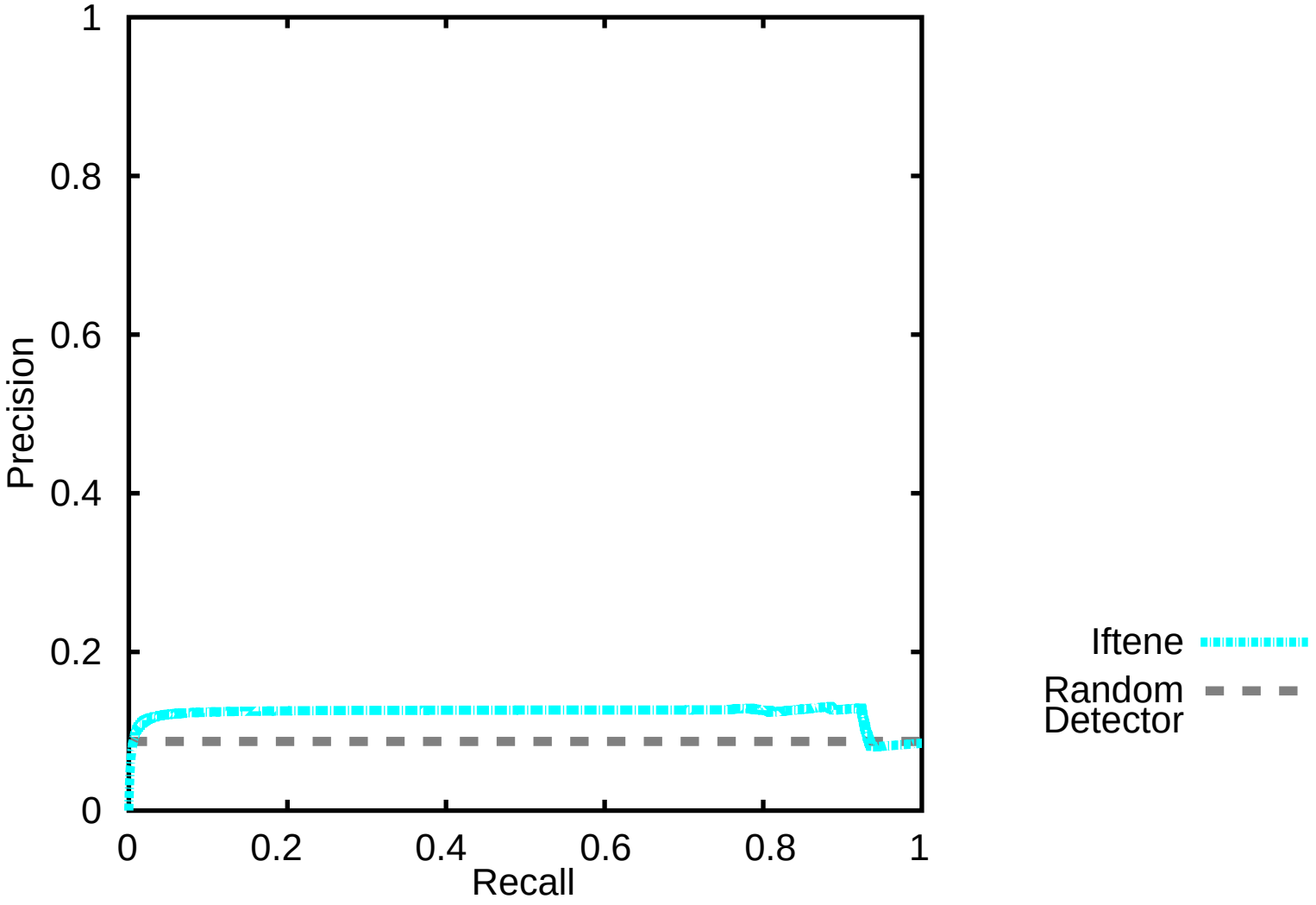
The PAN Competition

Vandalism Detection Results



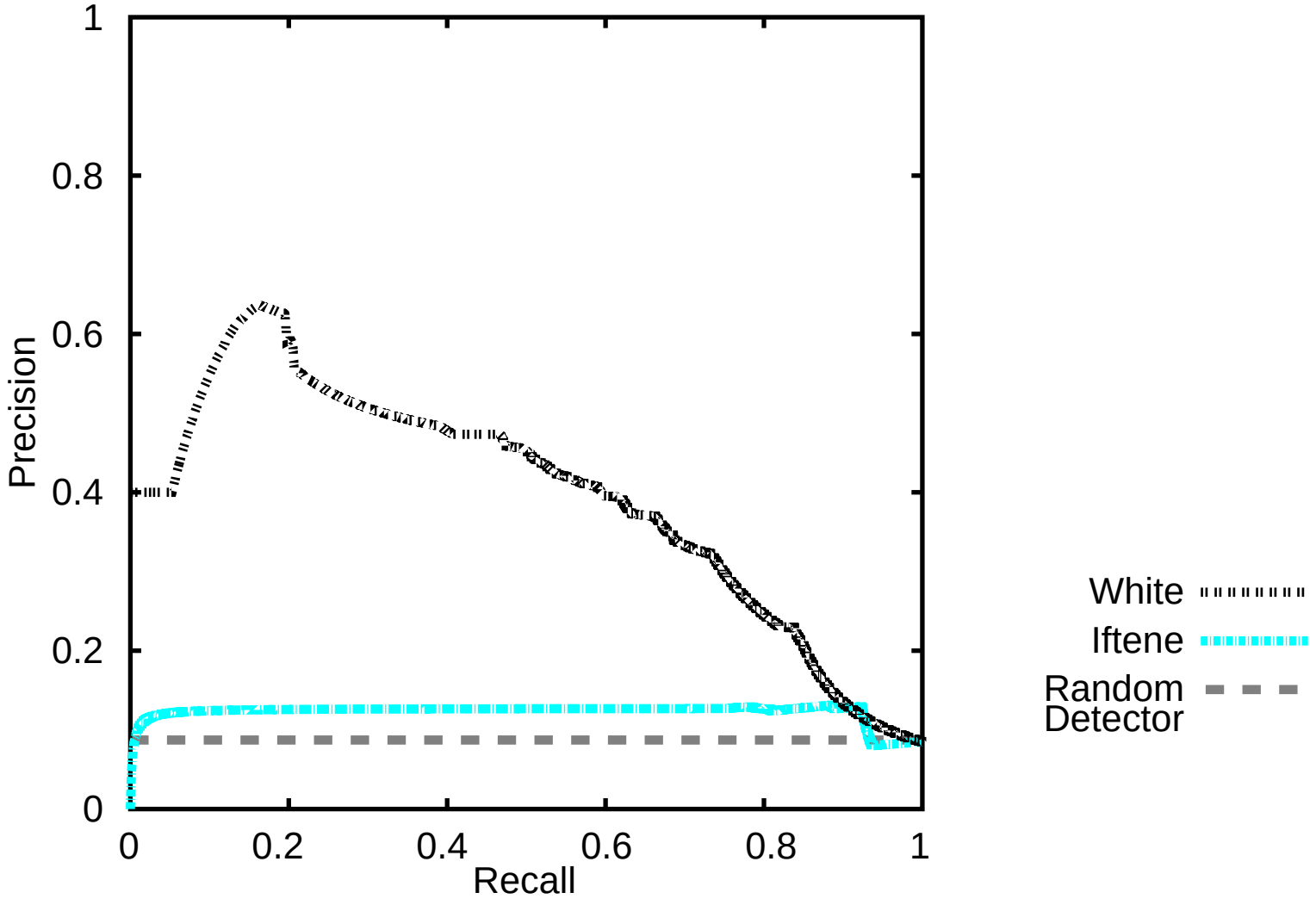
The PAN Competition

Vandalism Detection Results



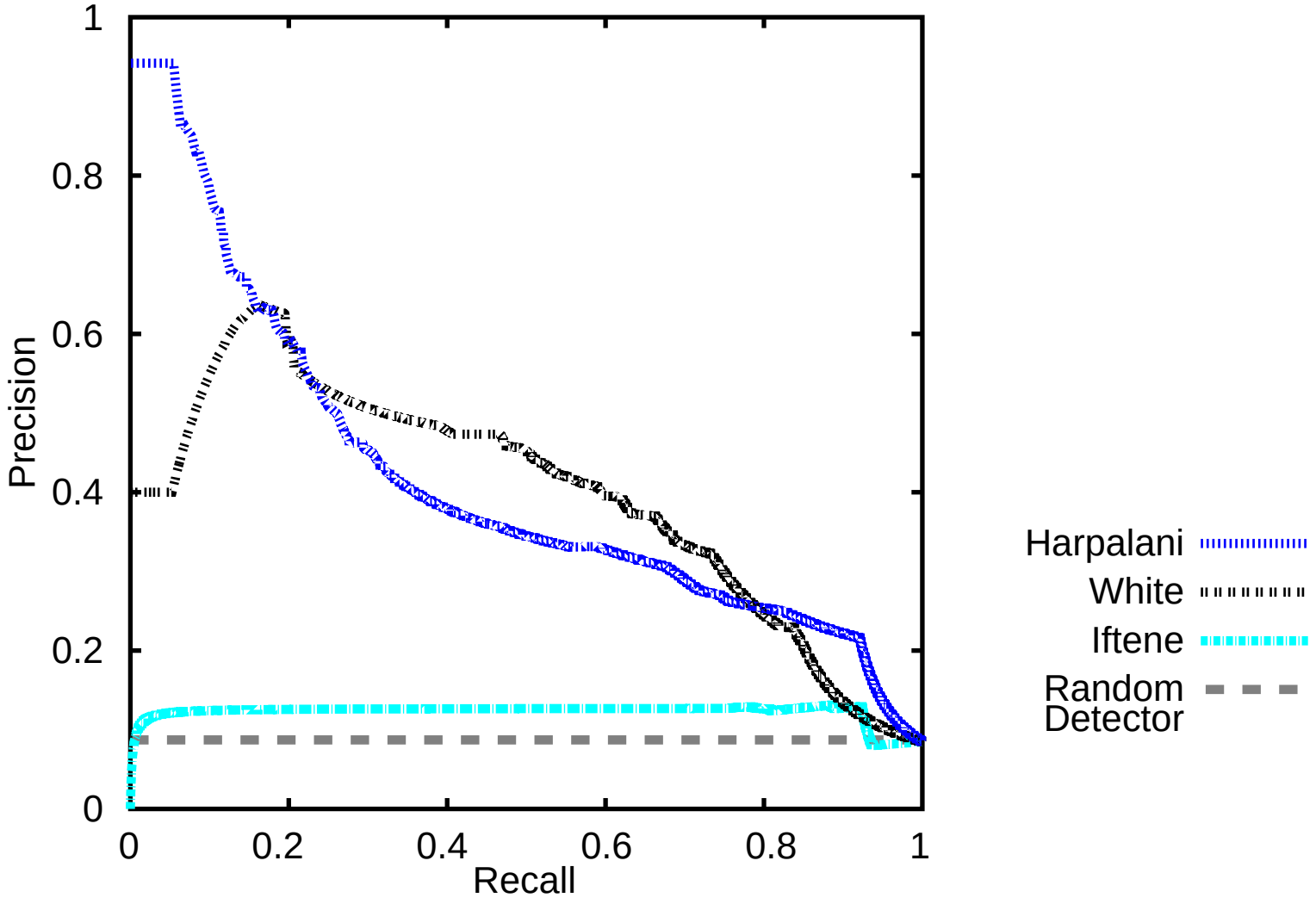
The PAN Competition

Vandalism Detection Results



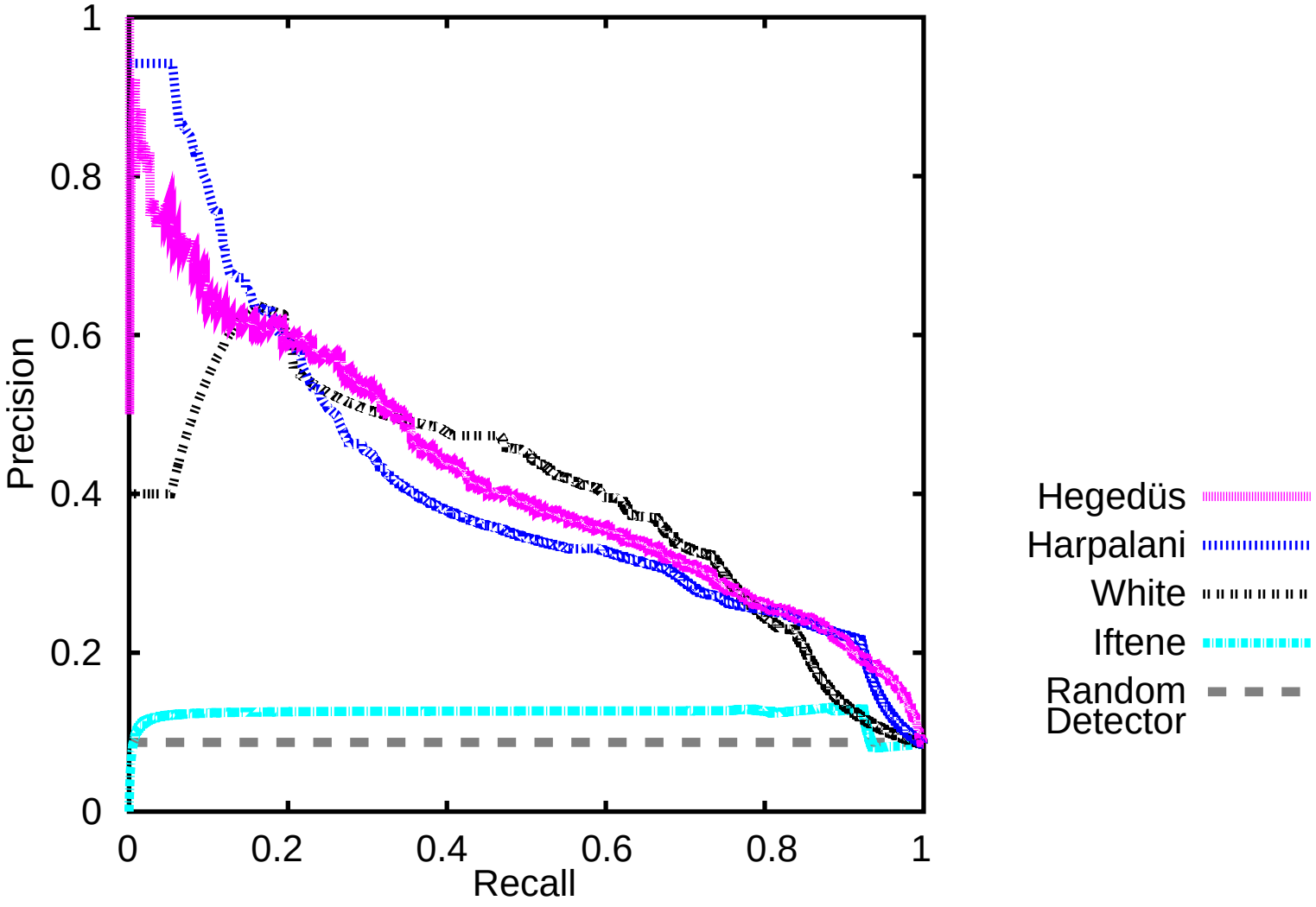
The PAN Competition

Vandalism Detection Results



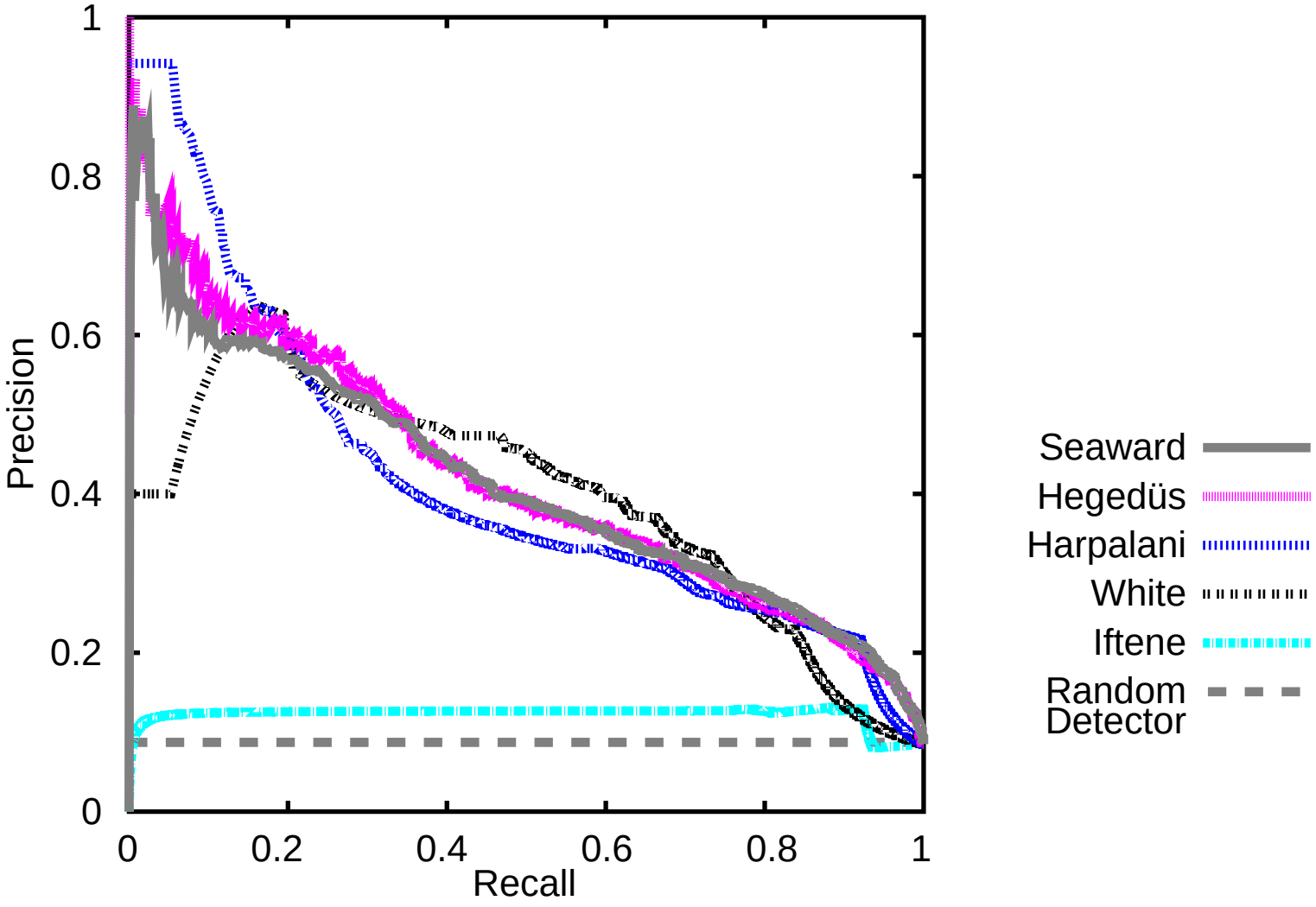
The PAN Competition

Vandalism Detection Results



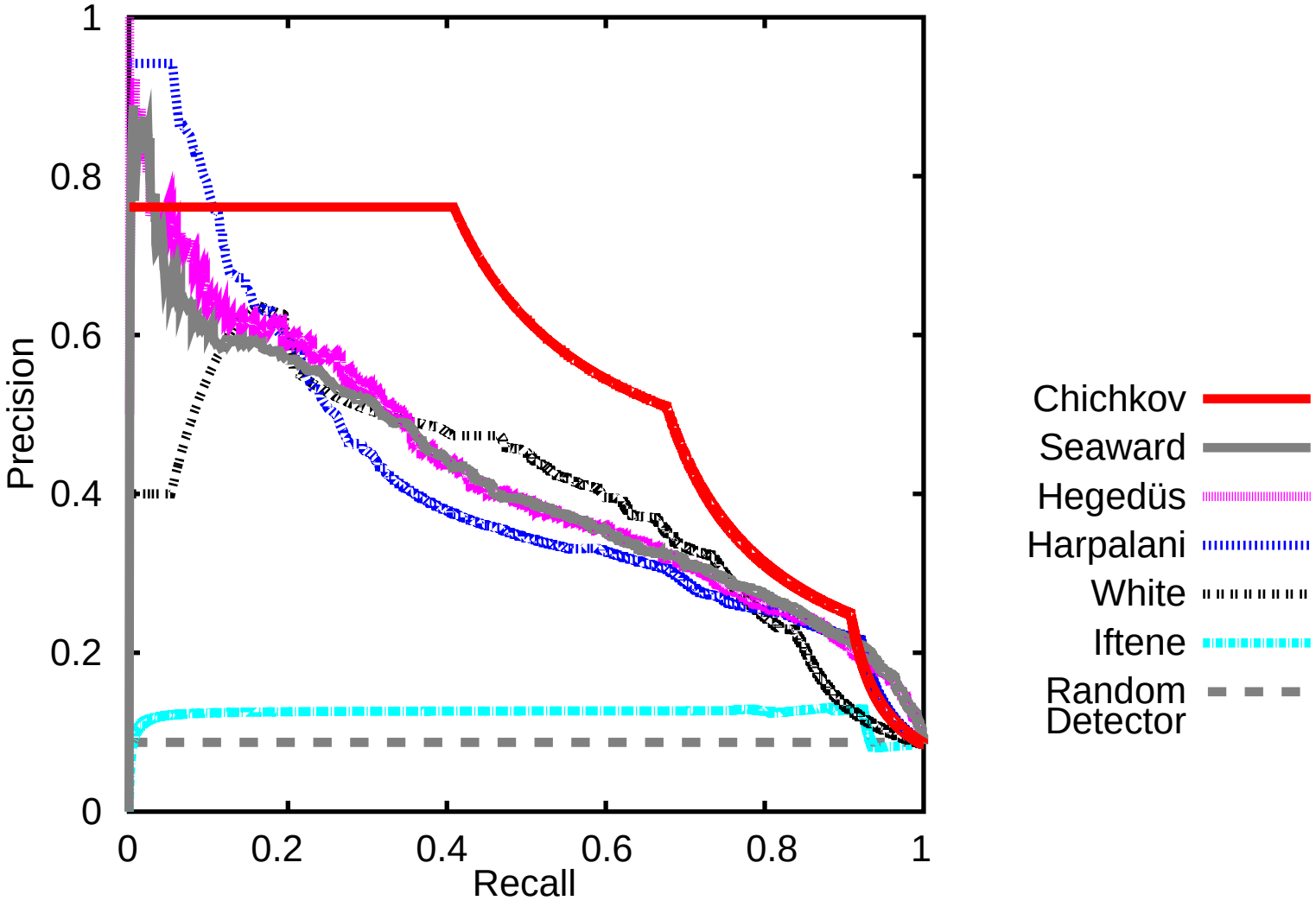
The PAN Competition

Vandalism Detection Results



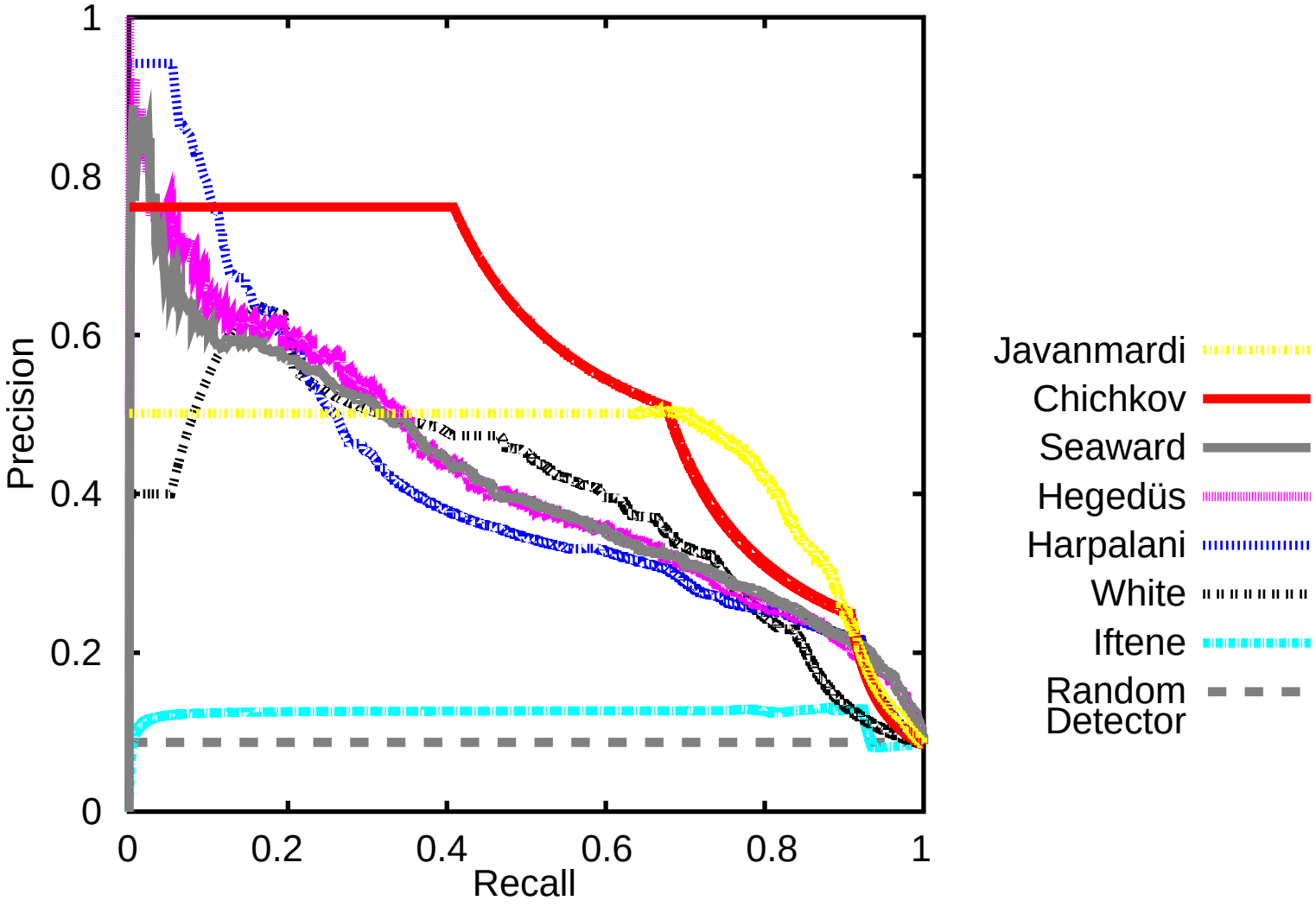
The PAN Competition

Vandalism Detection Results



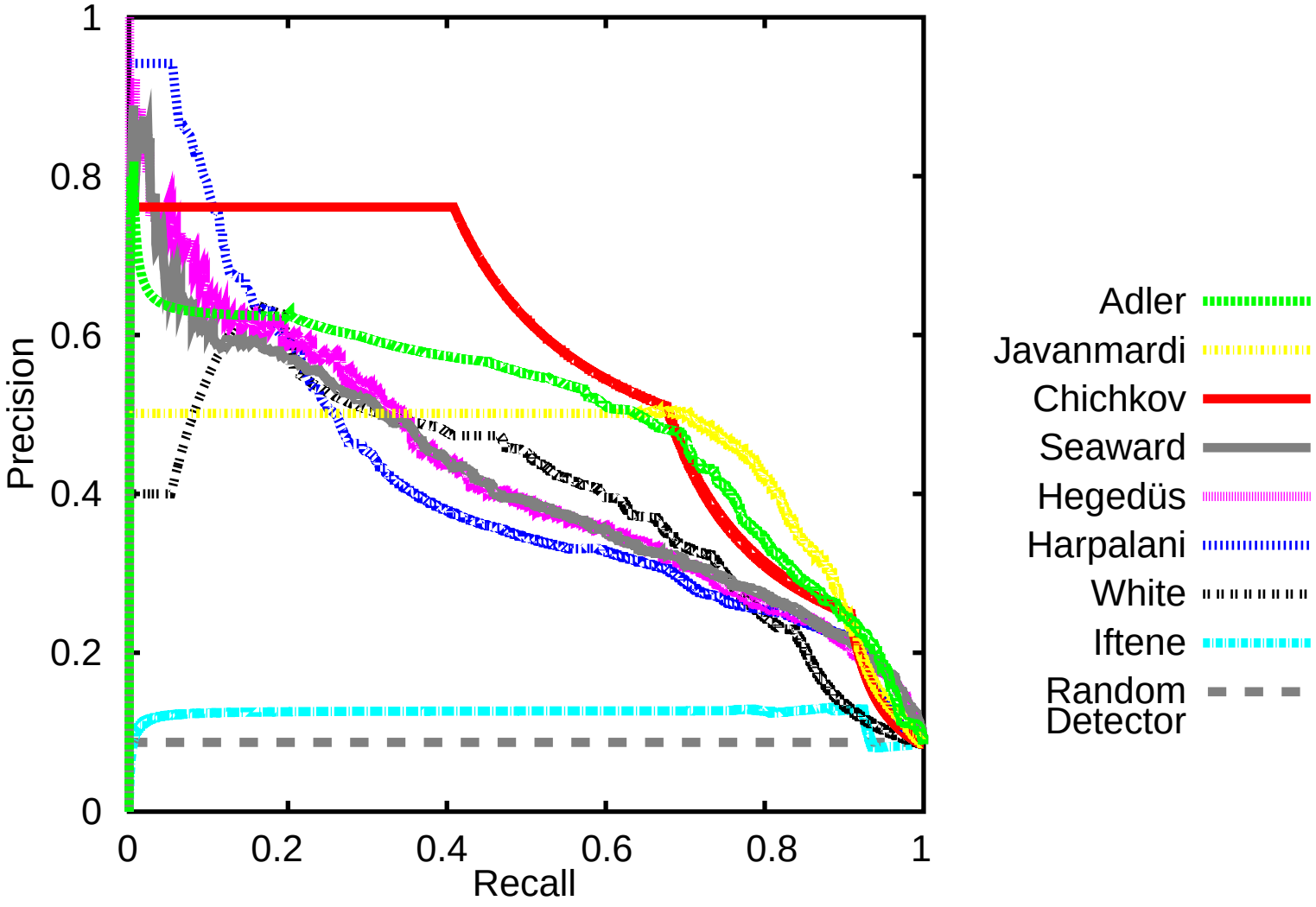
The PAN Competition

Vandalism Detection Results



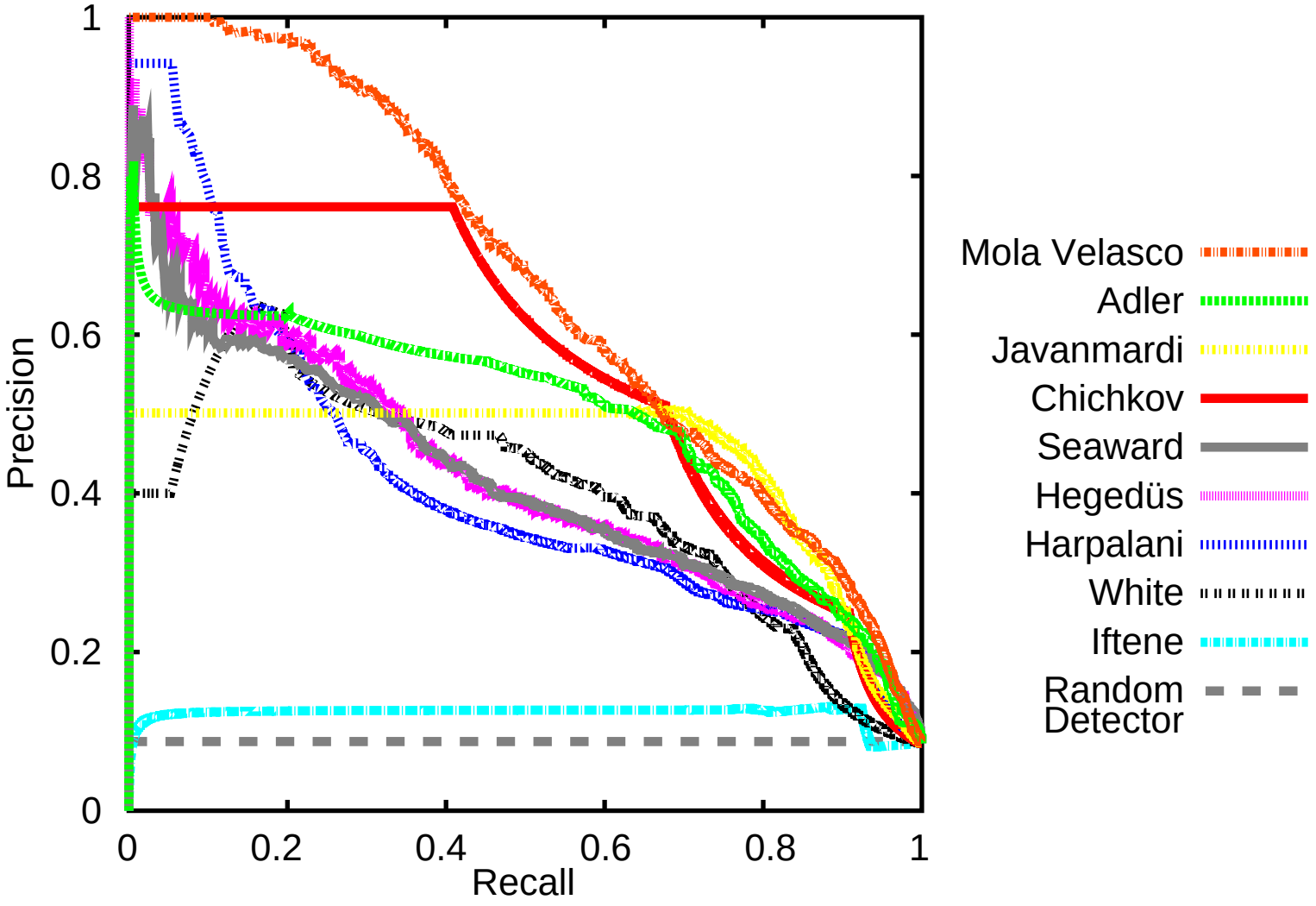
The PAN Competition

Vandalism Detection Results



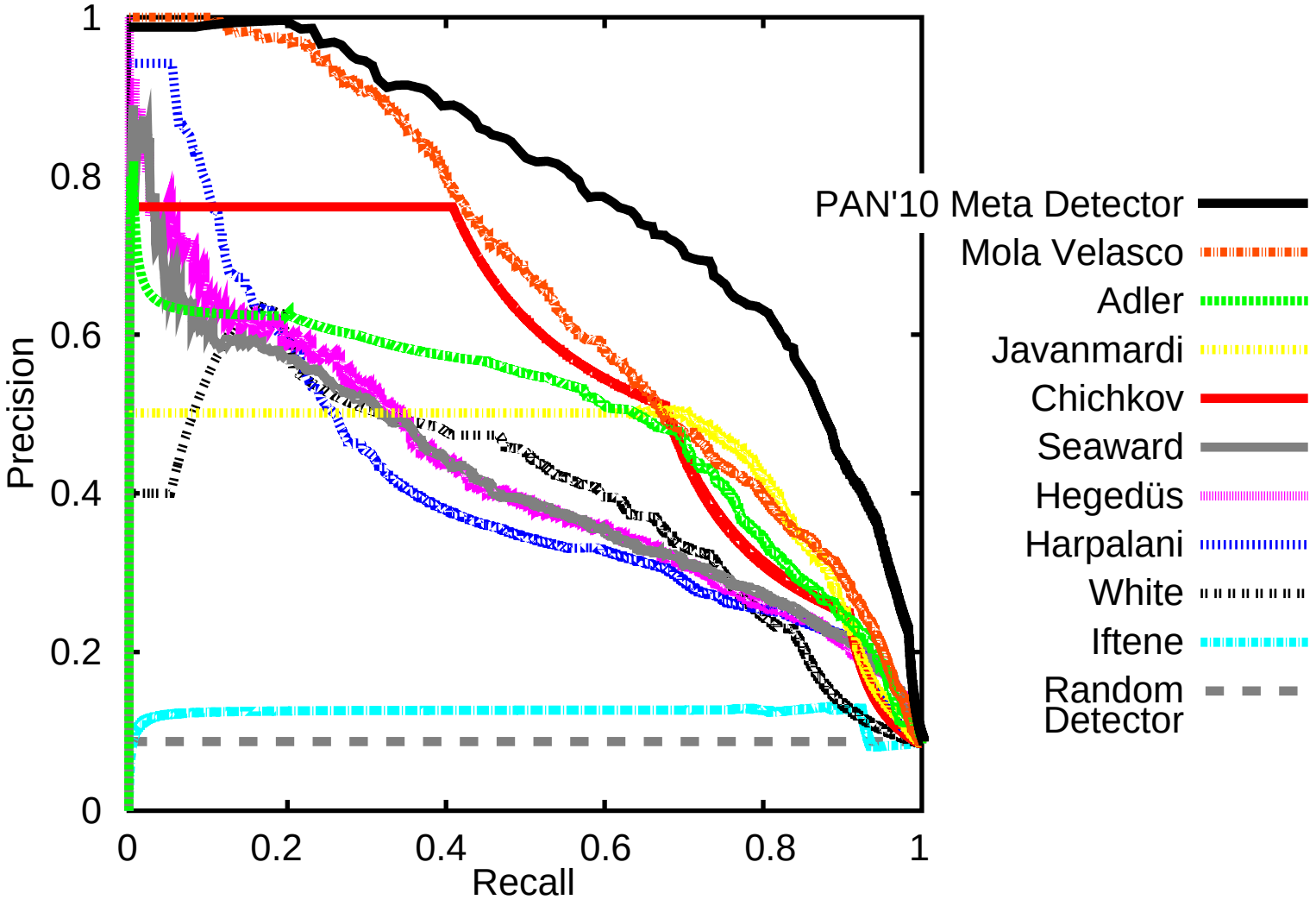
The PAN Competition

Vandalism Detection Results



The PAN Competition

Vandalism Detection Results

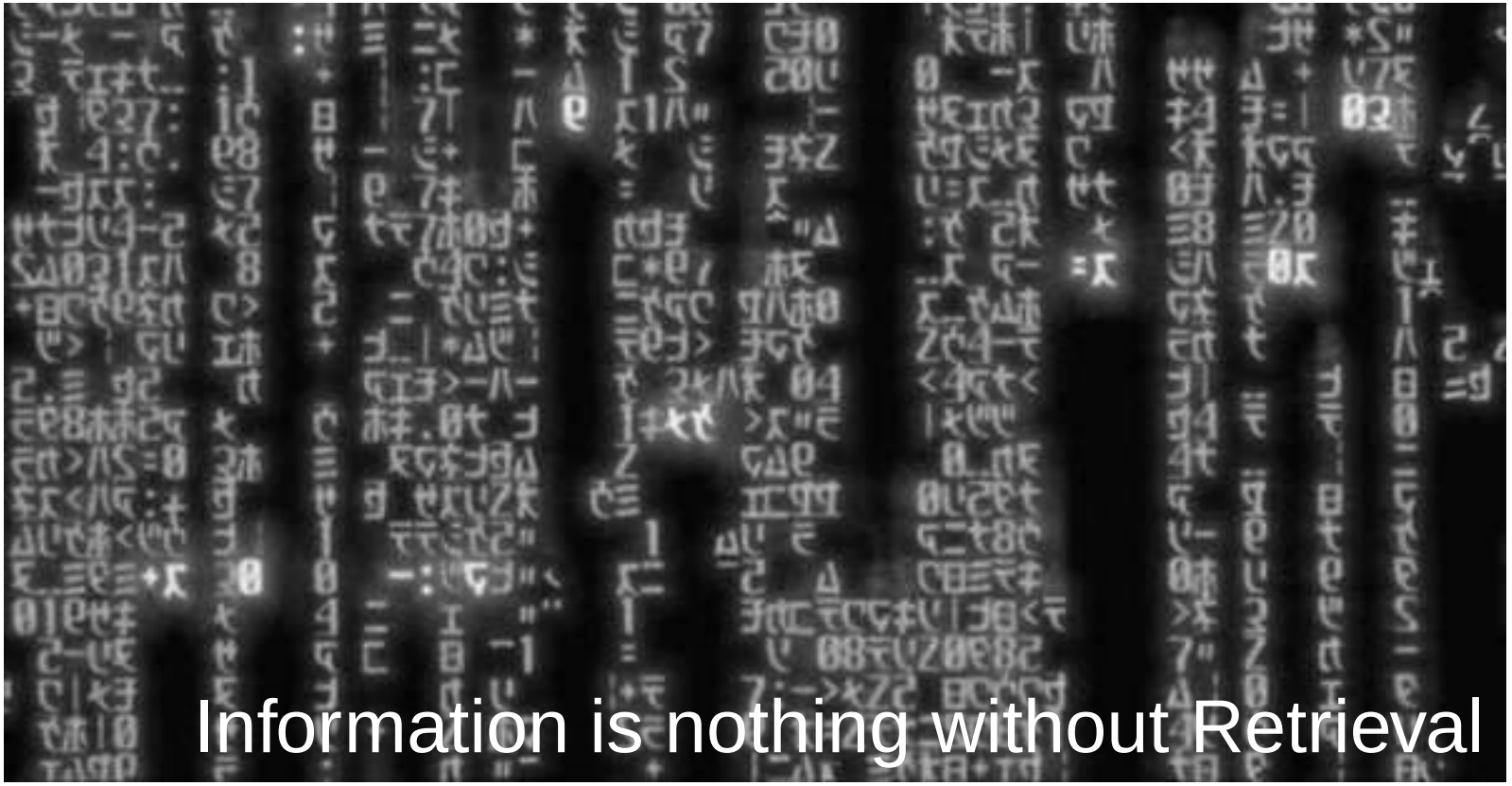


The PAN Competition

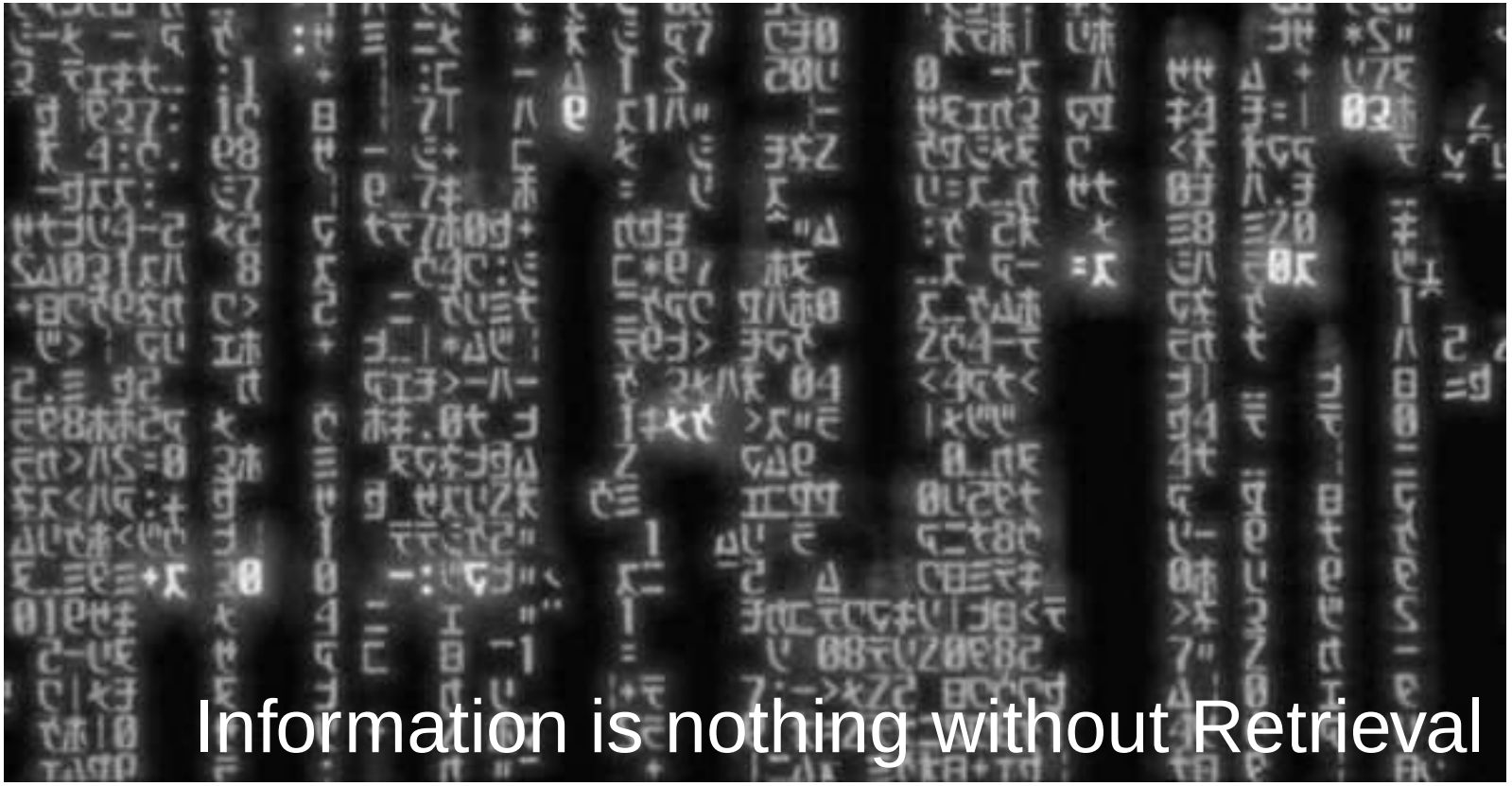
Vandalism Detection Results

ROC-AUC	ROC rank	PR-AUC	PR rank		Detector
0.95690	–	0.77609	–	–	PAN'10 Meta Detector
0.92236	1	0.66522	1	–	Mola Velasco
0.90351	2	0.49263	3	↓	Adler
0.89856	3	0.44756	4	↓	Javanmardi
0.89377	4	0.56213	2	⇈	Chichkov
0.87990	5	0.41365	7	⇊	Seaward
0.87669	6	0.42203	5	↑	Hegedus
0.85875	7	0.41498	6	↑	Harpalani
0.84340	8	0.39341	8	–	White
0.65404	9	0.12235	9	–	Iftene
0.50000	10	0.08490	10	–	Random Detector

The PAN Competition



The PAN Competition



Retrieval is nothing without Evaluation