CS@BIU
המחלקה למדעי המחשב
אוניברסיטת בר-אילן

BIU Center for Research in Applied Cryptography and Cyber Security

אוניברסיטת בר-אילן
Bar-Ilan University

Machine Learning & Deep Learning for Malware Classification

Tomer Gill gilltom@cs.biu.ac.il
Yossi Mandil yossimandil@gmail.com
Under the guidance of **Mr. Assaf Barak** of the BIU Cyber Center

XSS & CSRF Attacks

Goals:

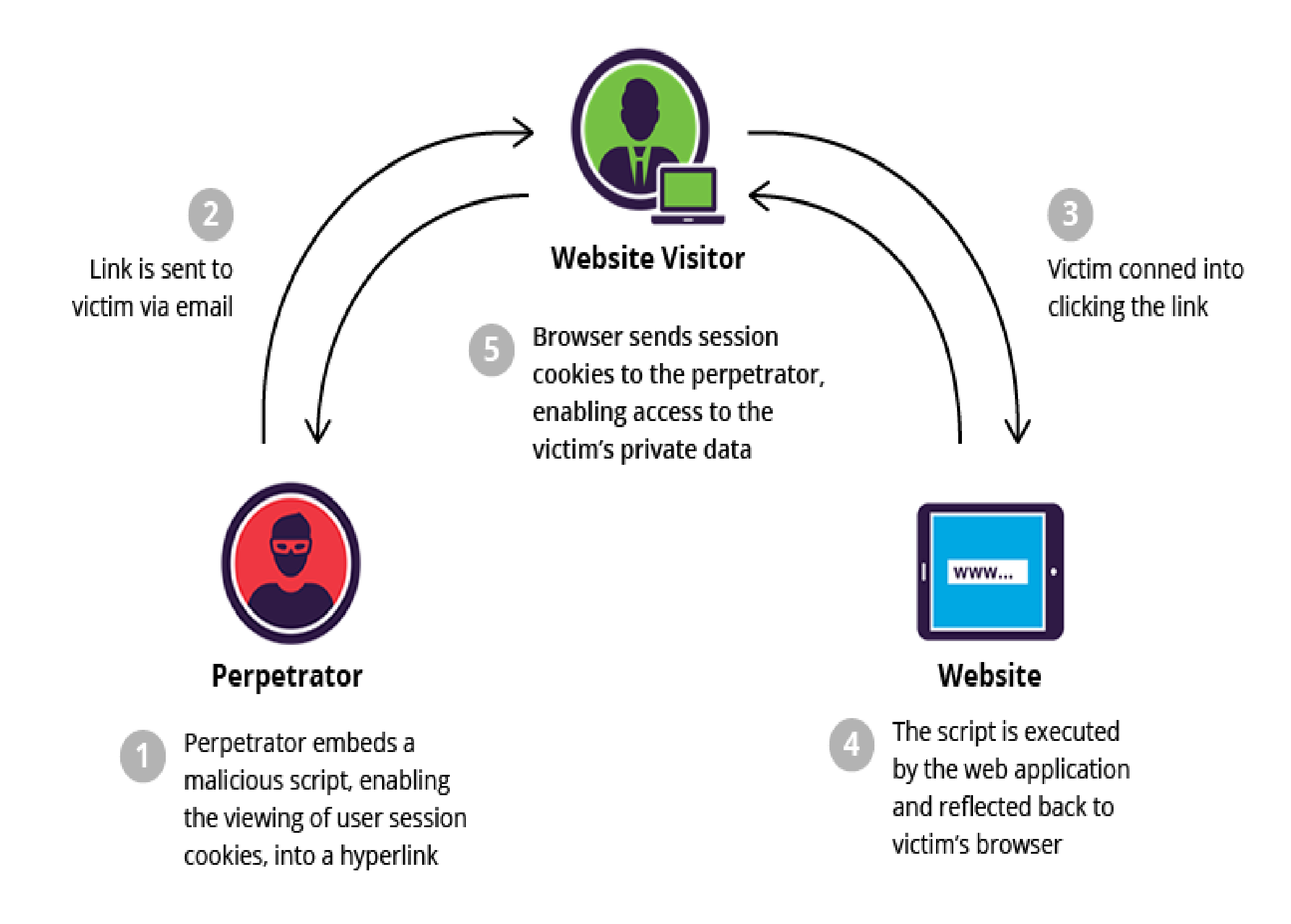
- Understanding common JS attacks from attacker's and attacked party's angles
- Experience in complex attacks: using one attack for another

XSS

Created a vulnerable website for each type of attack:

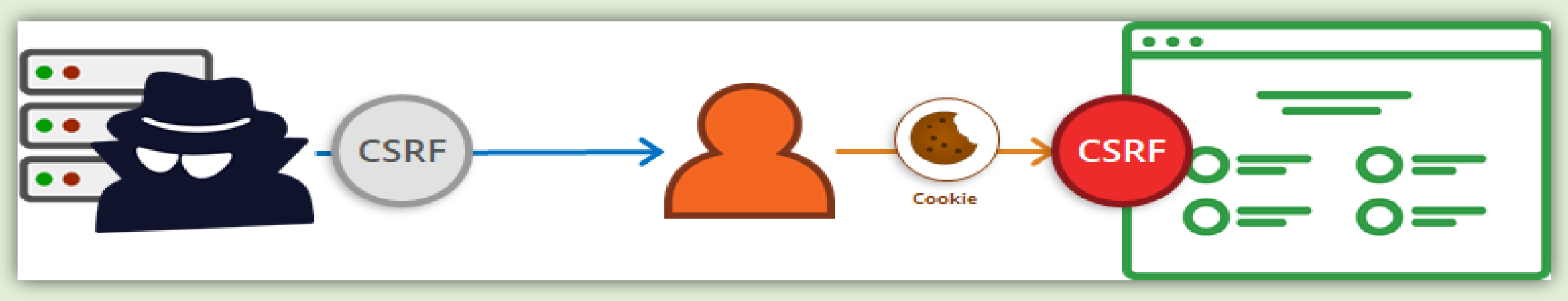
- Persistent
- Reflected
- DOM XSS

Then attacking them


Website Visitor
2 Link is sent to victim via email
3 Victim conned into clicking the link
5 Browser sends session cookies to the perpetrator, enabling access to the victim's private data
Perpetrator
1 Perpetrator embeds a malicious script, enabling the viewing of user session cookies, into a hyperlink
www...
Website
4 The script is executed by the web application and reflected back to victim's browser

CSRF

- Created a vulnerable website
- Using our XSS websites to attack "users" who are logged in to first website
- Implemented a defense using the token design pattern

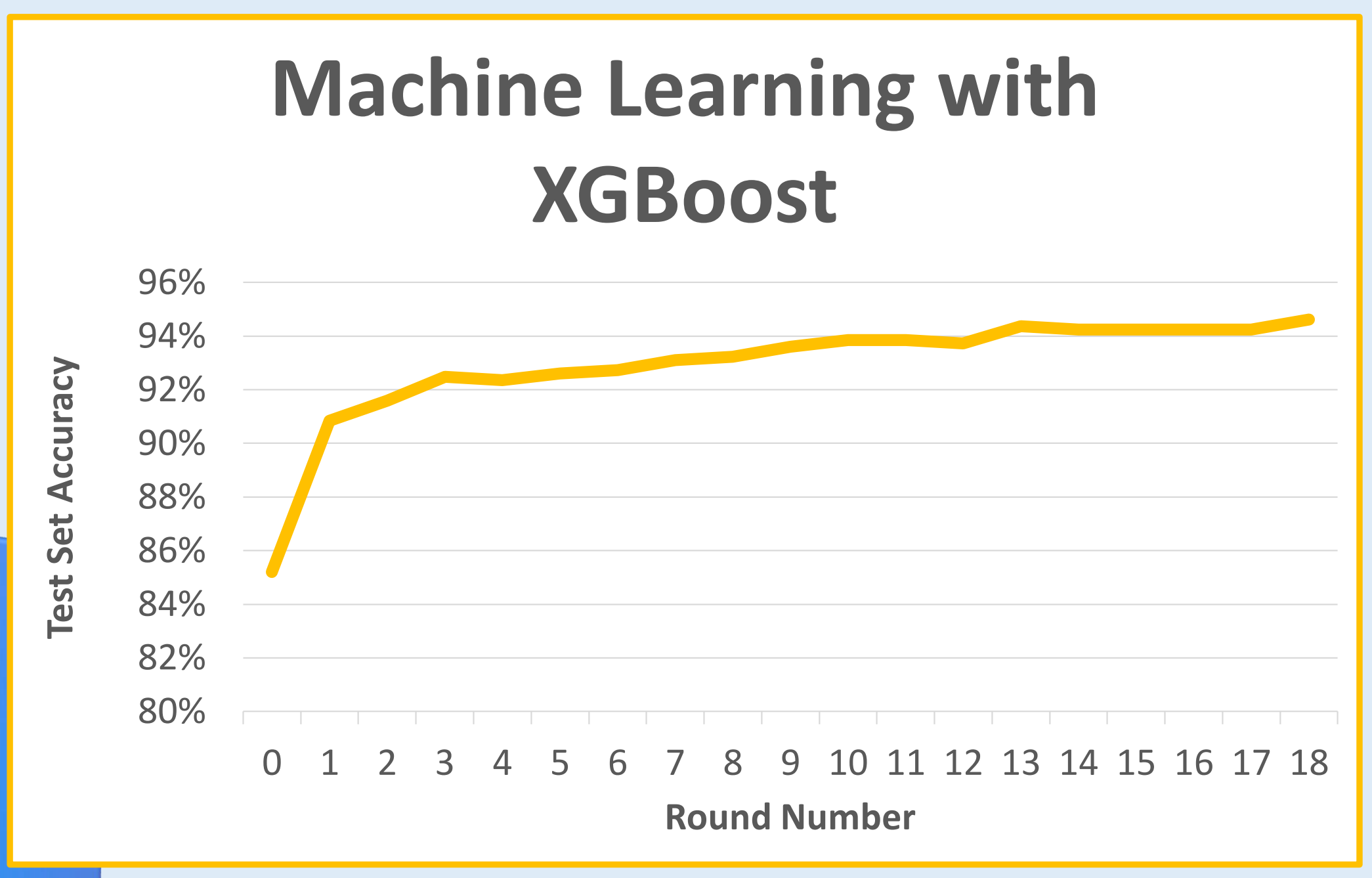


Malware EXE Classification

Problem Description	Given an EXE file, determine whether **malicious** or **benign**	Data Description	11 thousand files split to 10 classes: benign and 9 malicious types

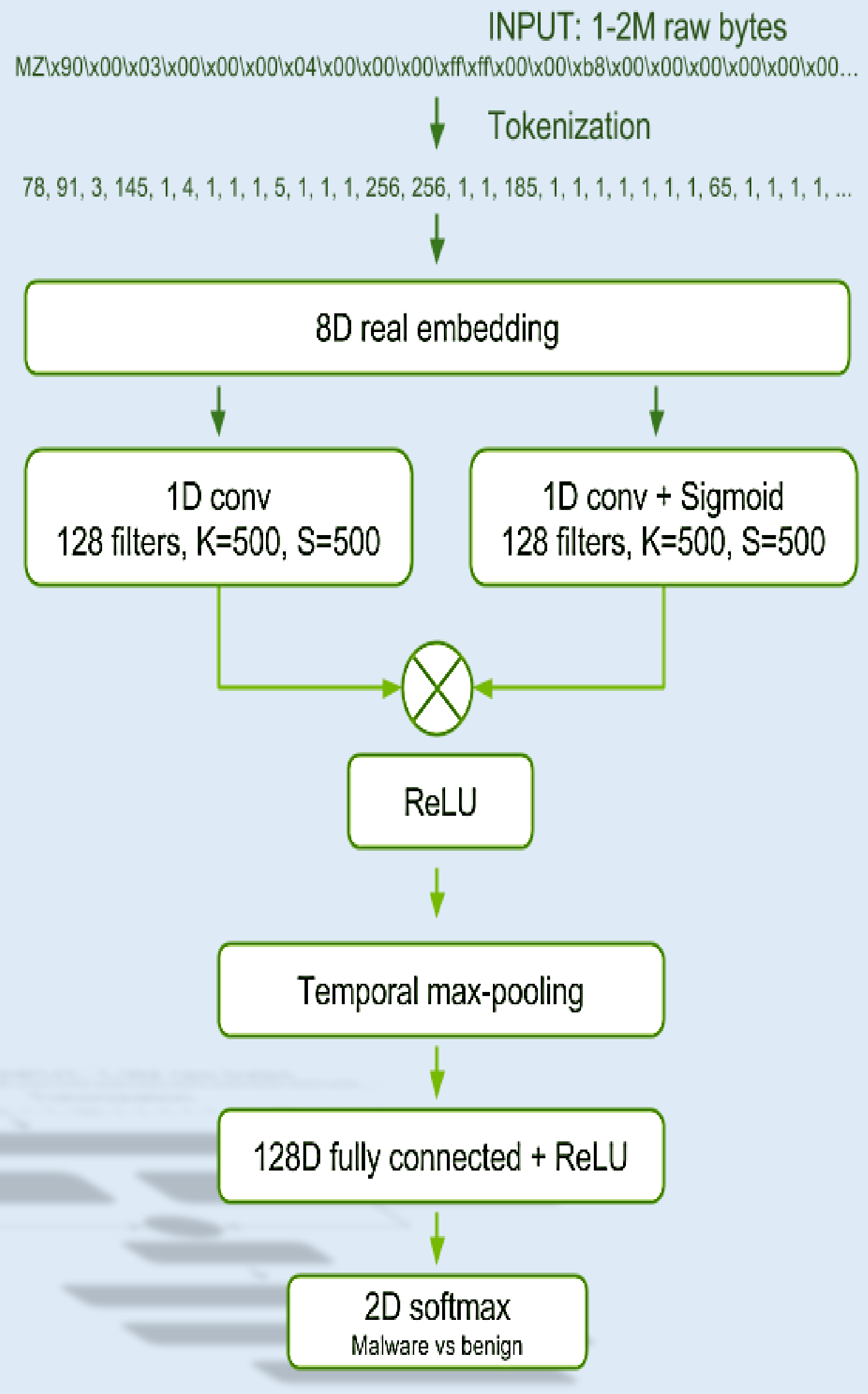
Machine Learning using XGBoost

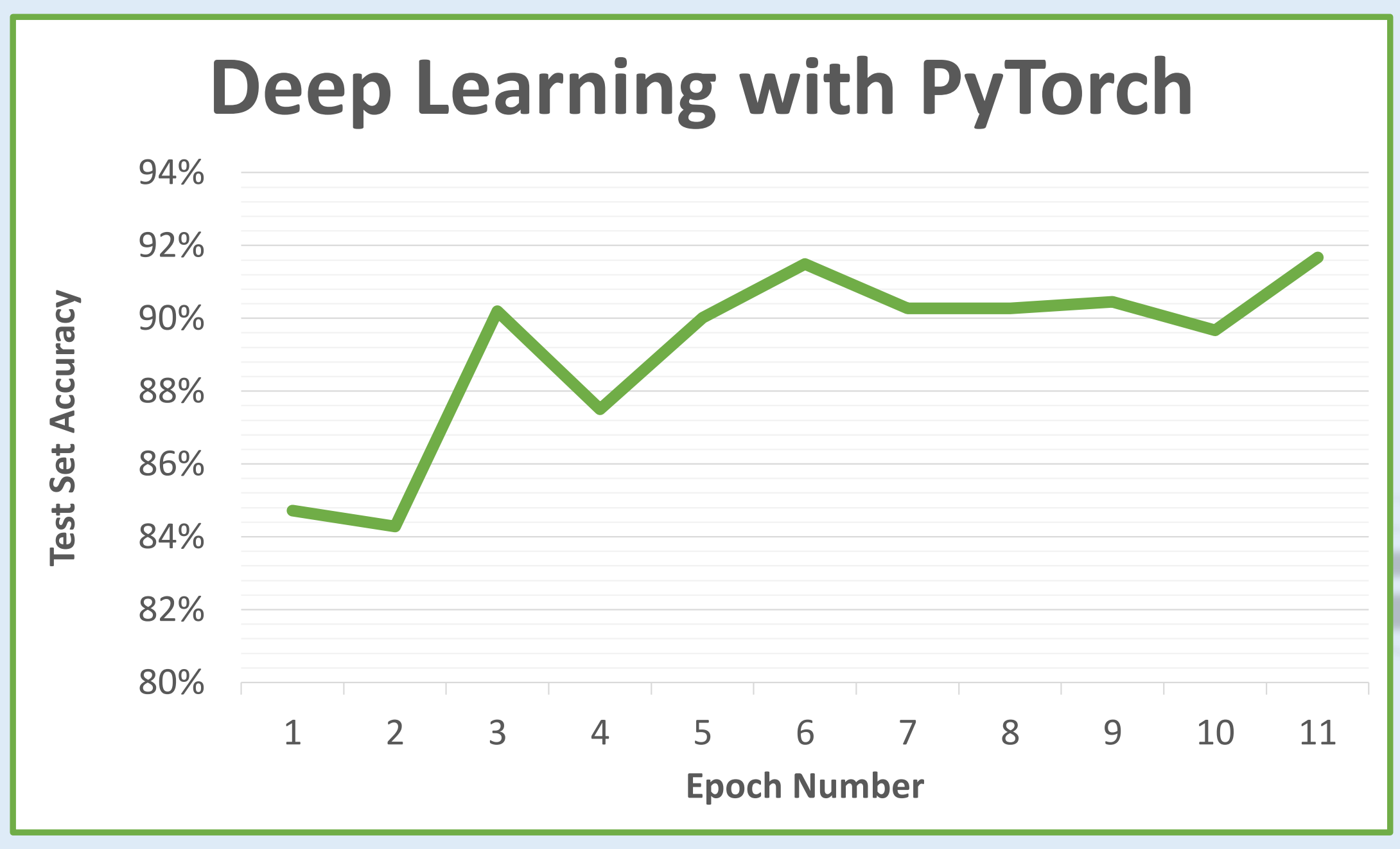
- Reading disassembly of EXE files, splitting opcodes to *n-grams*
- Using XGBoost models, constructing different decision trees
- *Boosting*: combining them into a much better model



Deep Learning using PyTorch

- An implementation of Raff's groundbreaking paper:
  Malware Detection by Eating a Whole EXE
- Net input is whole EXE – a vector of bytes sized 2M (file was rounded up/down as needed)
- Convolutional Neural Network using an Embedding matrix
- Output goes through a linear layer


INPUT: 1-2M raw bytes
MZ\x90\x00\x03\x00\x00\x00\x04\x00\x00\x00\xff\xff\x00\x00\xb8\x00\x00\x00\x00\x00\x00...
Tokenization
78, 91, 3, 145, 1, 4, 1, 1, 1, 5, 1, 1, 1, 256, 256, 1, 1, 185, 1, 1, 1, 1, 1, 1, 1, 65, 1, 1, 1, 1, ...
8D real embedding
1D conv
128 filters, K=500, S=500
1D conv + Sigmoid
128 filters, K=500, S=500
ReLU
Temporal max-pooling
128D fully connected + ReLU
2D softmax
Malware vs benign


Deep Learning with PyTorch
Test Set Accuracy
Epoch Number