Adversarial and Secure Learning 2019

Introductory information

Bojan Kolosnjaji, TUM I20
Ching-Yu Kao, Fraunhofer AISEC

Today’s agenda

1) Introduction to the research area

2) Seminar instructions

    a) Deliverables

    b) Grading

    c) FAQ

Introduction to the research area

Learning in adversarial environment

- Problem considered in the research community at least since early 2000s
- With the hype over machine learning (deep learning) the problem gains importance
- Adversarial perturbations studied in vision, text, malware...

Adversarial attacks

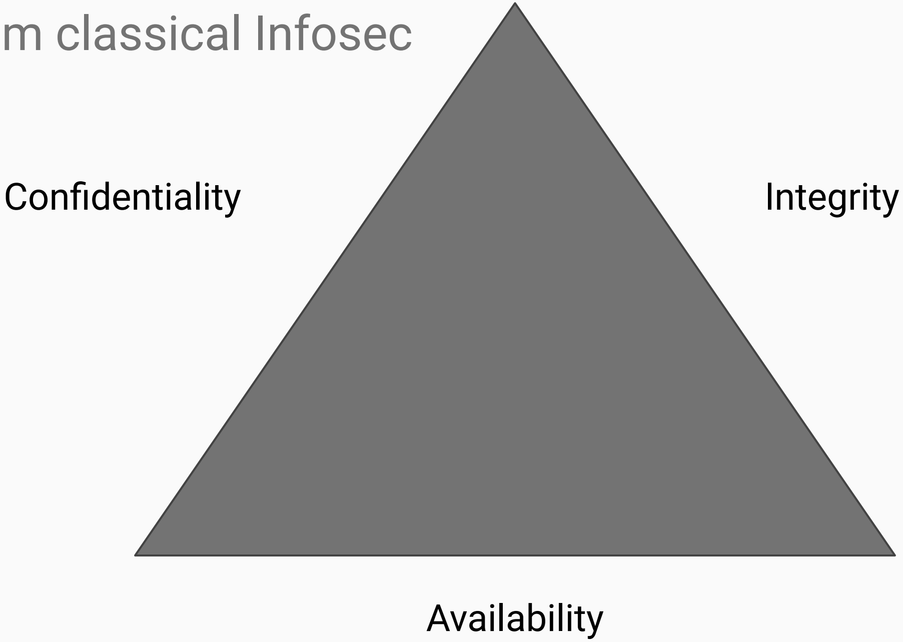
- Attack goals
  - Destroy system functionality
  - Change the model output
  - Reveal confidential information
- Attacker knowledge
  a. Perfect Knowledge (white-box attack)
  b. Limited Knowledge (gray-box attack)
  c. Zero Knowledge (black-box attack)
- Attacker capability
  a. Make queries to the model
  b. Change training data
  c. Change validation/test data

Adversarial attacks

- Evasion attacks - modify test input to get a desired classifier decision
- Poisoning attacks - modify training data to change the resulting model (either just make it less accurate or enable particular properties (backdoor)
- Privacy attacks - learn about sensitive data from the model based on queries

Adversarial attacks

- CIA Triad - from classical Infosec



Confidentiality

Integrity

Availability

Attack and defense

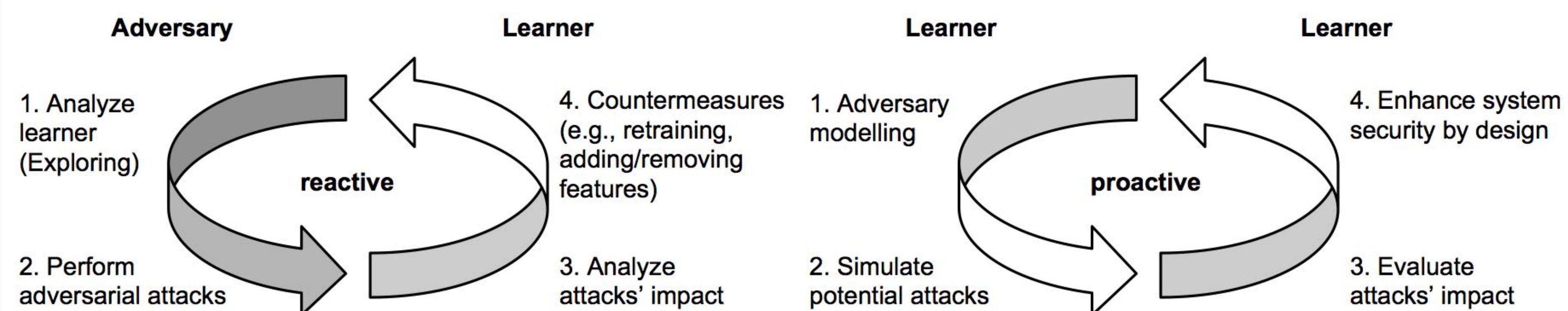
- Attack optimization -> find minimal perturbation that achieves the goal

  $(\min g(x)\ ,\ \text{s.t. } d(x,x')<d_{max}\,(x,x'))$

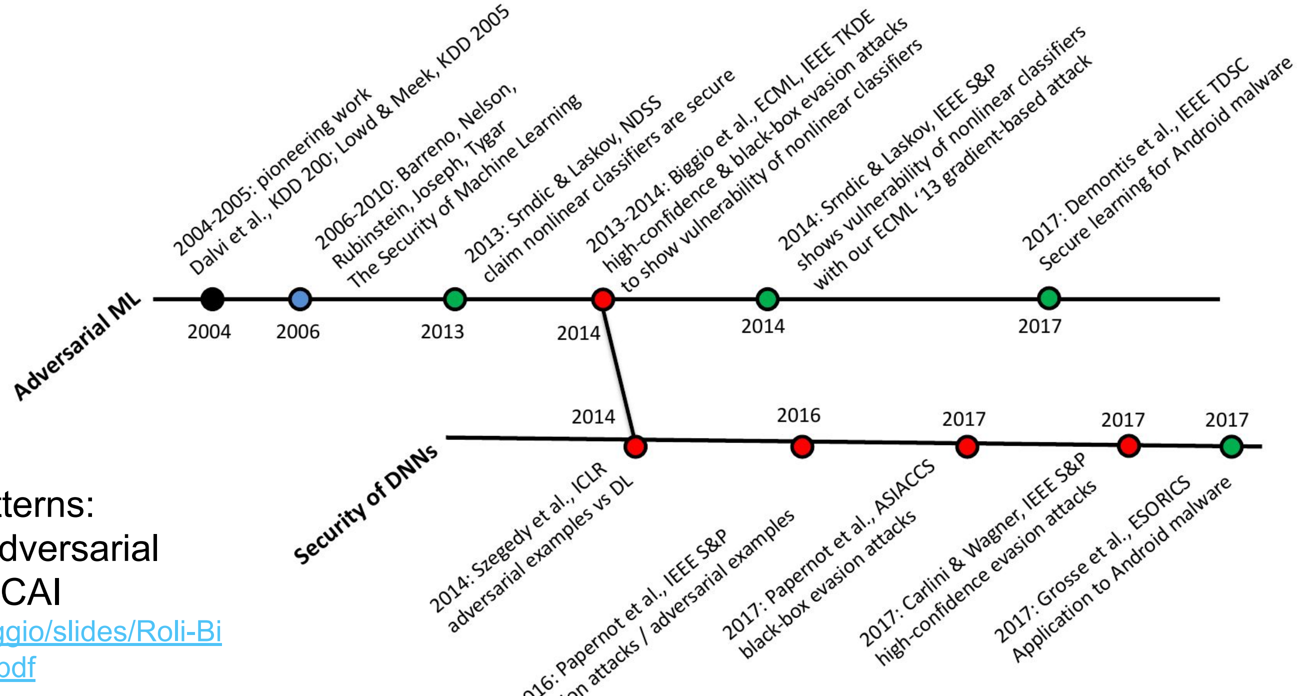
- Defense optimization -> find an approach that either:
  - Rejects the attack points as outliers or is
  - Robust to attack points

Defense strategies

- Kerckhoffs principle - do not rely on obscurity
- Assess security against various levels of adversary’s knowledge and capability
- Arms race


Adversary
Learner
1. Analyze learner (Exploring)
reactive
4. Countermeasures (e.g., retraining, adding/removing features)
2. Perform adversarial attacks
3. Analyze attacks’ impact
Learner
Learner
1. Adversary modelling
proactive
4. Enhance system security by design
2. Simulate potential attacks
3. Evaluate attacks’ impact

Adversarial Attack -discoveries


Adversarial ML
2004-2005: pioneering work
Dalvi et al., KDD 200; Lowd & Meek, KDD 2005
2006-2010: Barreno, Nelson,
Rubinstein, Joseph, Tygar
The Security of Machine Learning
2013: Srndic & Laskov, NDSS
claim nonlinear classifiers are secure
2013-2014: Biggio et al., ECML, IEEE TKDE
high-confidence & black-box evasion attacks
to show vulnerability of nonlinear classifiers
2014: Srndic & Laskov, IEEE S&P
shows vulnerability of nonlinear classifiers
with our ECML '13 gradient-based attack
2017: Demontis et al., IEEE TDSC
Secure learning for Android malware
2004
2006
2013
2014
2014
2017
Security of DNNs
2014
2016
2017
2017
2017
2014: Szegedy et al., ICLR
adversarial examples vs DL
016: Papernot et al., IEEE S&P
on attacks / adversarial examples
2017: Papernot et al., ASIACCS
black-box evasion attacks
2017: Carlini & Wagner, IEEE S&P
high-confidence evasion attacks
2017: Grosse et al., ESORICS
Application to Android malware

Roli, Biggio, "Wild Patterns: Half-day Tutorial on Adversarial Machine Learning", IJCAI
http://www.diee.unica.it/~biggio/slides/Roli-Biggio-IJCAI-ECAI18-tutorial.pdf

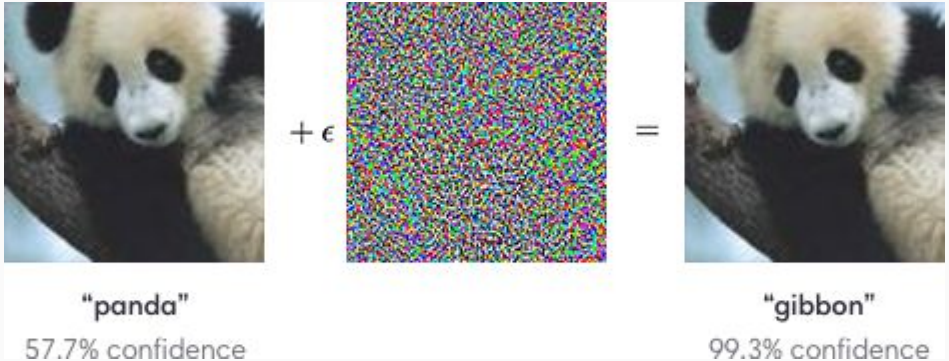
Adversarial Learning

- Attacked methods
  - Linear regression
  - **SVM**
  - **Neural Networks**
  - Random Forests
  - ...

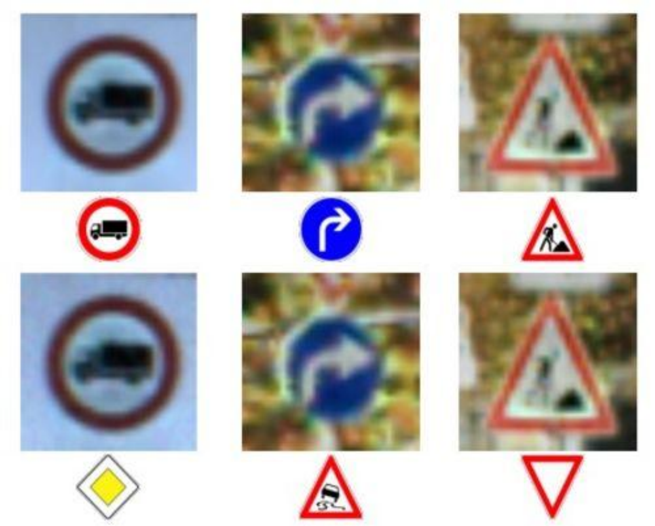
Adversarial Learning

- Data under modification
  - **Computer Vision**
  - **Natural Language Processing**
  - Speech
  - **Malware**
  - ...

Vision papers best known...

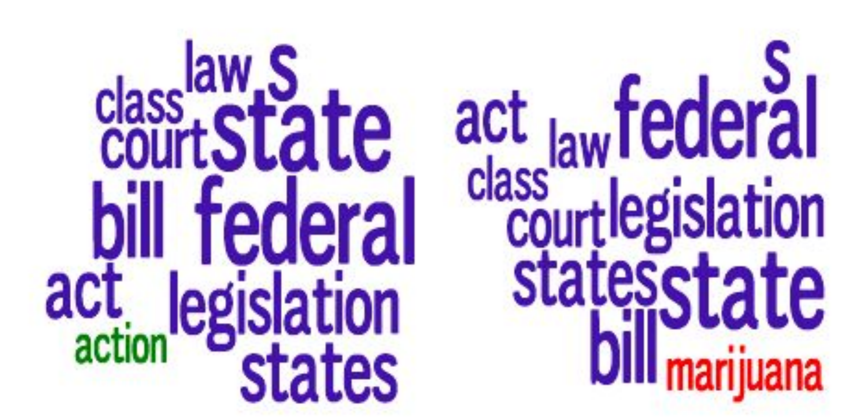


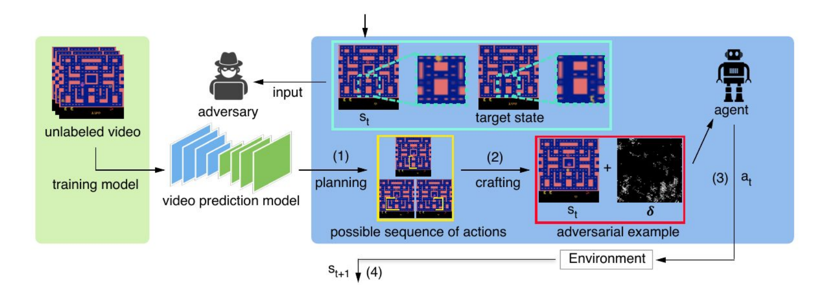
Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." *arXiv preprint arXiv:1412.6572* (2014).



Papernot, Nicolas, et al. "Practical black-box attacks against deep learning systems using adversarial examples." *arXiv preprint* (2016).

But also other domains...

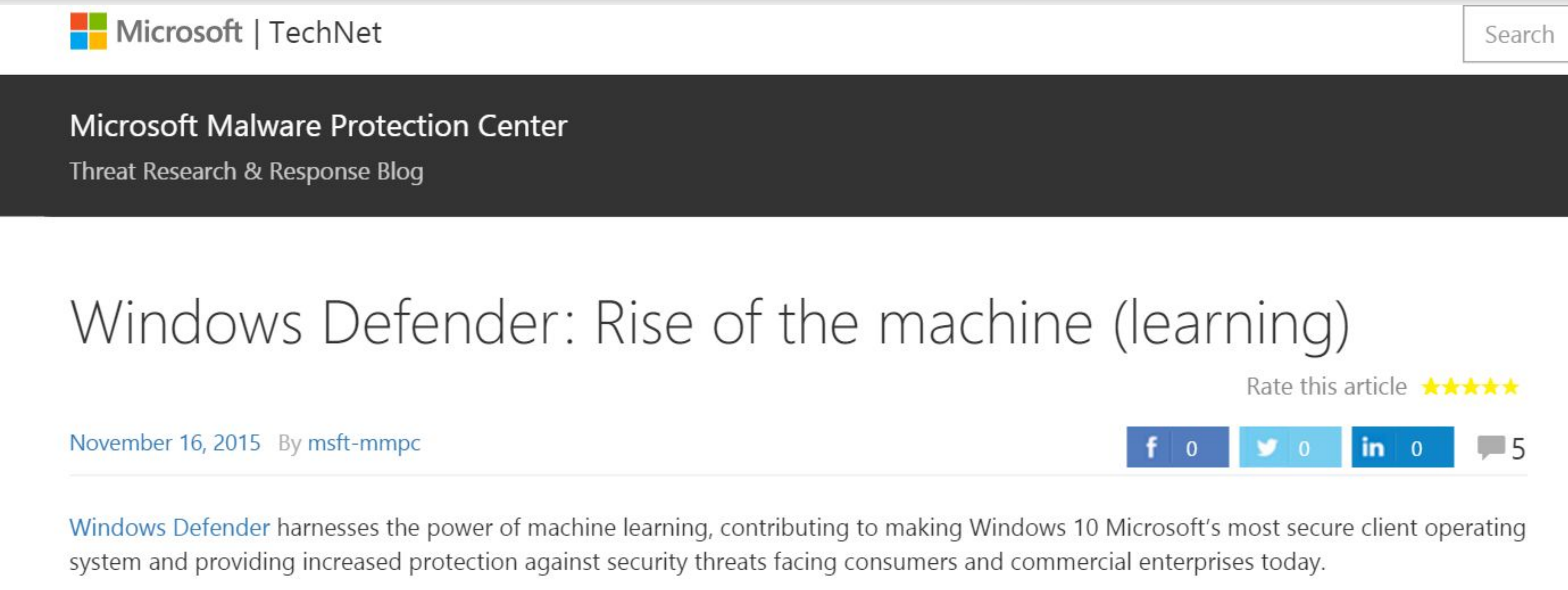




Mei, Shike, and Xiaojin Zhu. "The security of latent dirichlet allocation." *Artificial Intelligence and Statistics*. 2015.

Lin, Yen-Chen, et al. "Tactics of adversarial attack on deep reinforcement learning agents." *arXiv preprint arXiv:1703.06748*(2017).

Machine Learning in getting adoption in AV industry



Microsoft | TechNet

Microsoft Malware Protection Center

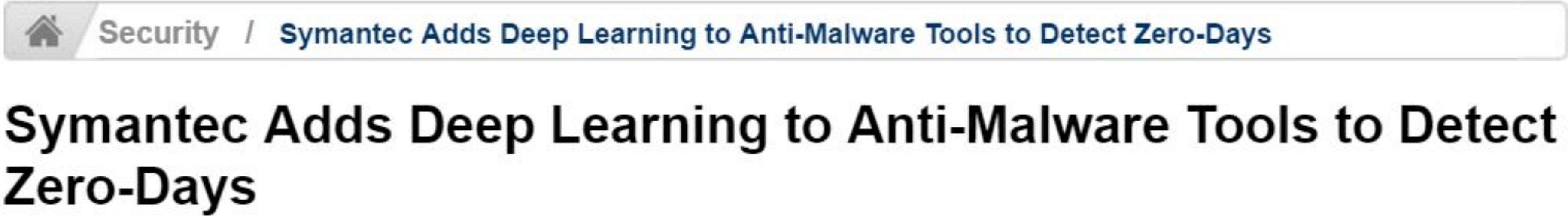
Threat Research & Response Blog

Windows Defender: Rise of the machine (learning)

November 16, 2015 By msft-mmpc

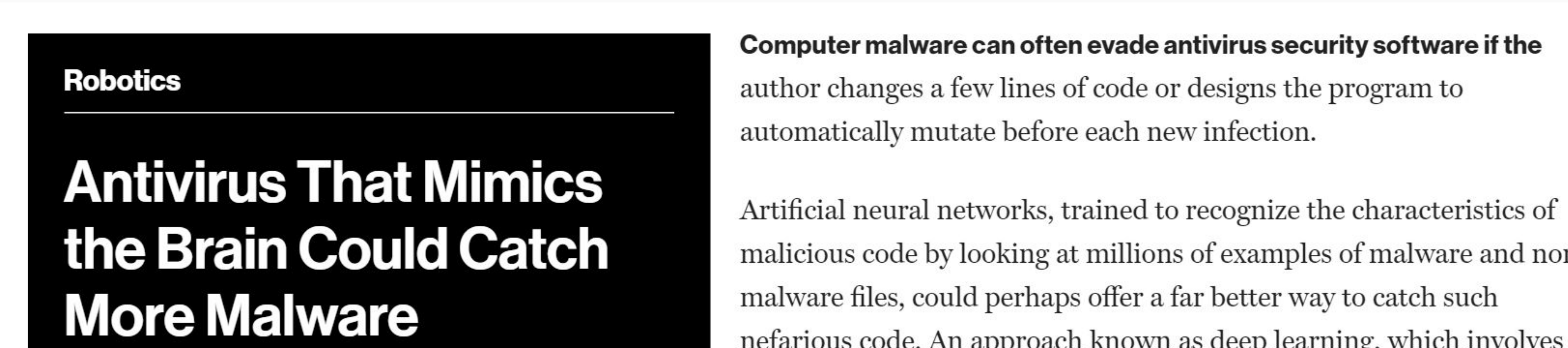
Windows Defender harnesses the power of machine learning, contributing to making Windows 10 Microsoft's most secure client operating system and providing increased protection against security threats facing consumers and commercial enterprises today.

Even deep learning...



Security / Symantec Adds Deep Learning to Anti-Malware Tools to Detect Zero-Days

Symantec Adds Deep Learning to Anti-Malware Tools to Detect Zero-Days



Robotics

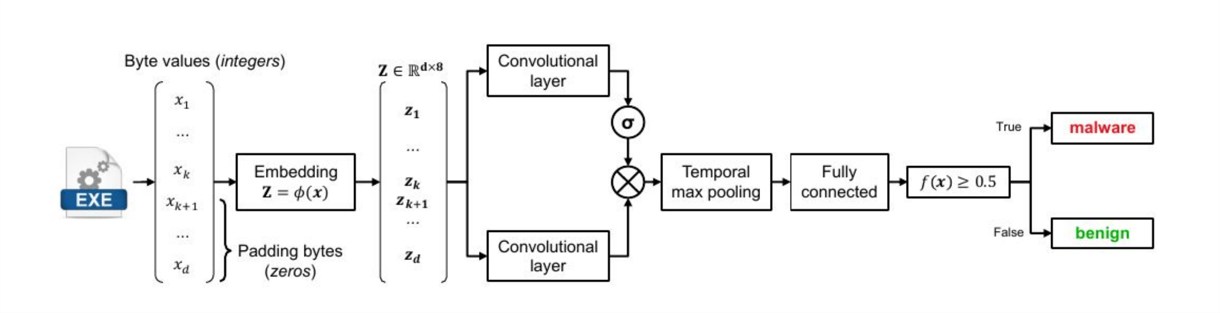
Antivirus That Mimics the Brain Could Catch More Malware

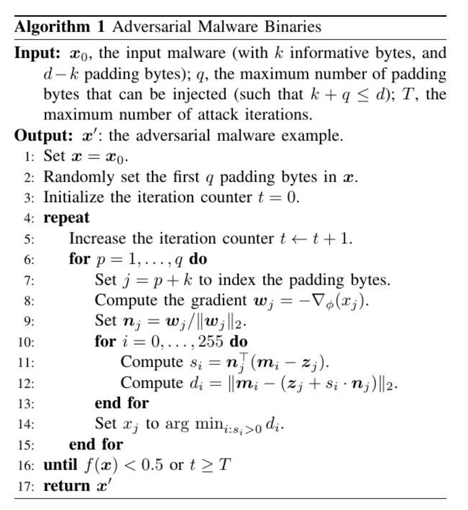
Computer malware can often evade antivirus security software if the author changes a few lines of code or designs the program to automatically mutate before each new infection.

Artificial neural networks, trained to recognize the characteristics of malicious code by looking at millions of examples of malware and nor malware files, could perhaps offer a far better way to catch such nefarious code. An approach known as deep learning, which involves

Some current work at Chair I20

- Generating adversarial binaries to evade neural network malware detectors based on raw code





```
Algorithm 1 Adversarial Malware Binaries
Input: $x_0$, the input malware (with $k$ informative bytes, and
    $d - k$ padding bytes); $q$, the maximum number of padding
    bytes that can be injected (such that $k + q \leq d$); $T$, the
    maximum number of attack iterations.
Output: $x'$: the adversarial malware example.
Set $x = x_0$.
Randomly set the first $q$ padding bytes in $x$.
Initialize the iteration counter $t = 0$.
repeat
    Increase the iteration counter $t \leftarrow t + 1$.
    for $p = 1, \dots, q$ do
        Set $j = p + k$ to index the padding bytes.
        Compute the gradient $w_j = -\nabla_\phi(x_j)$.
        Set $n_j = w_j / \|w_j\|_2$.
        for $i = 0, \dots, 255$ do
            Compute $s_i = n_j^\top (m_i - z_j)$.
            Compute $d_i = \|m_i - (z_j + s_i \cdot n_j)\|_2$.
        end for
        Set $x_j$ to $\arg\min_{i: s_i > 0} d_i$.
    end for
until $f(x) < 0.5$ or $t \geq T$
return $x'$
```

Seminar instructions

Our goal

- Get an **overview** of the academic state-of-the-art
- Extend your **knowledge** in machine learning, learn to look at it from the **security** standpoint
- Get a feeling on **how to evaluate risk**

What to deliver?

- 1 presentation
  - 45 minutes + 15 minutes of discussion
  - Present the topic, use all papers

- 1 report
  - 14 Pages LNCS (look up LNCS template in Latex)

Presentation (start on 14.05.)

Presentation

Needs to be:

- Correct
- Complete
- Comprehensible

Presentation - Correct

- Present information from the paper correctly

- Don’t speculate without a reason or proof

- Don’t claim something you cannot explain well

Presentation - Complete

- Explain **all** key points of the papers
- Be careful about **time constraints** and distribution
- **Convey information** without leaving out important insight

Presentation - Comprehensible

- Speak loud and clear
- Think about the audience - fellow students
- Motivate the audience for discussion
- Don’t fight your audience, answer all questions friendly

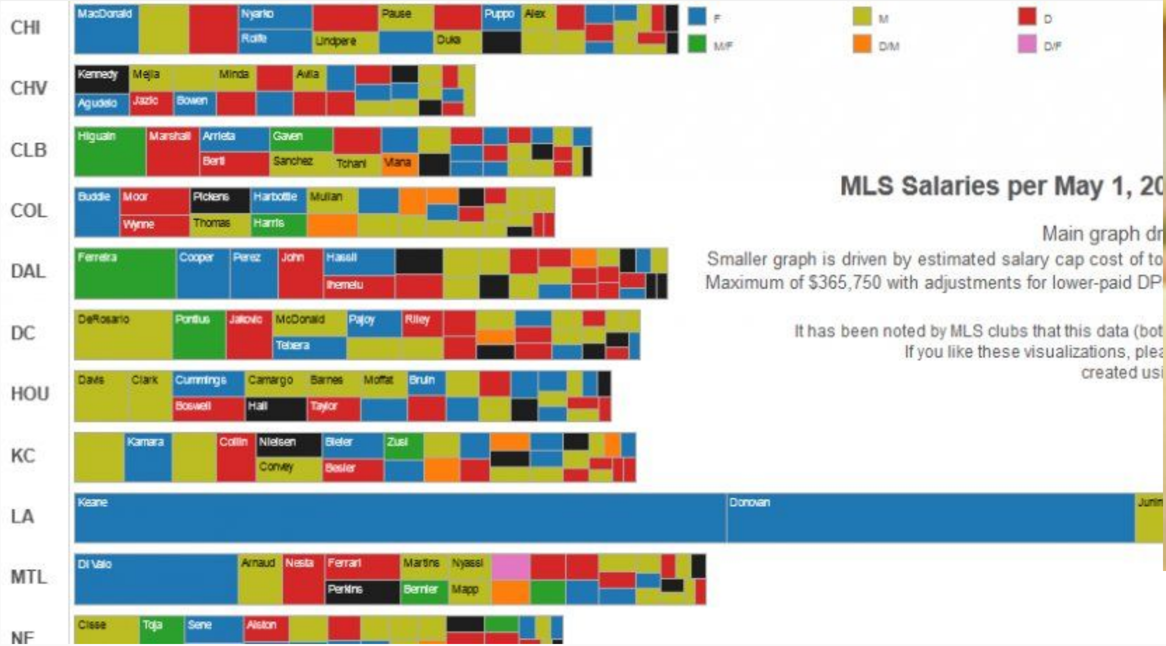
Presentation - Concise Text

- A PPT is just a presentation aid. It should not be a paper in its own right and your bullet points should be, if possible, less than a line long. Specifically, keep bullet points short by making them clear and concise. Do not be afraid from using incomplete sentences or phrases. In reality, this is the preferred method because it helps to highlight the points you are making during your talk.
- This is because having lots of text on your slides makes it difficult to understand the point you are trying to make. Furthermore, your audience will end up reading the text and ignoring you.

Presentation - Emphasis

- ONLY USE CAPITALIZATION WHEN NEEDED
- Use color sparingly
- **Only bold** key **words and phrases**
- Light text on a dark background is bad

Presentation - Pictures



MLS Salaries per May 1, 20

Main graph dr

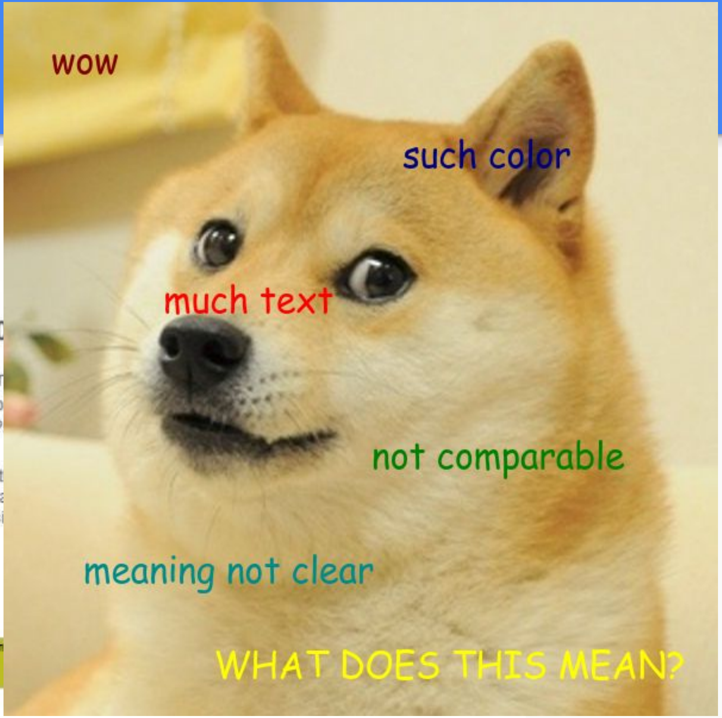
Smaller graph is driven by estimated salary cap cost of to

Maximum of $365,750 with adjustments for lower-paid DP

It has been noted by MLS clubs that this data (bot

If you like these visualizations, plea

created usi



Presentation - Structure

- Introduction to the topic
- Present all papers together
  - Introduction
  - Main Points
  - Back up arguments
  - Conclusions (key takeaways)

Presentation - Audience

- Read papers, or at least abstracts, prior to each presentation day
- Listen carefully, write down questions
- Ask questions, comment
- Active participation is appreciated!

Presentation - Grading

- Presentation skills
  - General organization, use of slides
  - Language, slide text and graphics
  - Pace, use of time
- Subject-related competence
  - Subject knowledge
  - Staying on topic
  - Identifying interesting/important points

Report (deadline 09.06.2019.)

Report

- 14 Pages LNCS
- Summarize key points of all papers together - not an easy task
- Use a typical paper structure:

Abstract -> Introduction -> Methodology -> Results -> Discussion -> Conclusion

Report - Abstract

- Summarize the paper
    - Introduction to the problem
    - How was the problem solved? Methodology
    - Short insight in the results
    - What is the impact of the papers?

Report - Introduction

- Describe the context
- What is the preexisting work?
- What does the preexisting work lack?
- How do the papers close the gap?

Report - Methodology

- Describe the mechanisms used to tackle the existing problem
- Lead the reader through the problem solving procedures
- Give arguments for the choice of methods

Report - Results

- Give an overview of the important results
- Add tables, graphs... if you have space
- Shortly comment on the figures
- Avoid phrases like: It is obvious from this graph that ...

Report - Discussion

- What do the results actually tell us?
- Compare the results with related work
- What are the limitations of the paper results?
- How can the limitations be addressed?

Report - Conclusion

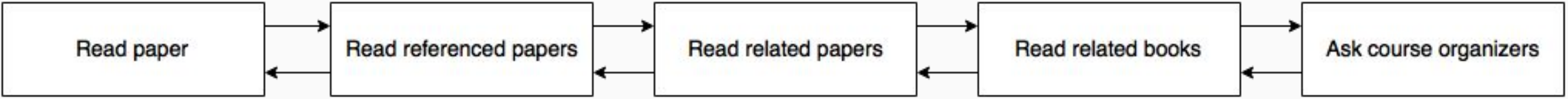
- Summary of the main findings in 3-4 sentences
- What are the most interesting results?
- What is the impact of the papers?

Report - Grading

- Paper organization
- Language and grammar
- Subject knowledge
- Ability to summarize
- Proper bibliography and citations (!)

How to do your research

- Seminar - (kind of) simulation of scientific research
- Try to be independent, but also ask questions



Read paper

Read referenced papers

Read related papers

Read related books

Ask course organizers

FAQ

- Allowed to miss a presentation day? Yes, if you have a very good reason.
    - Examples of good reason: health issues, schedule clashes at the Uni
    - Examples of bad reason: HiWi work, homework, football training, bad mood
- Can I set a meeting if I have problems with my papers?
    - Yes, but try to do as much as you can yourself.