# Perspectives on AI/ML and Cybersecurity

Fabio Massimo Zennaro
fabiomz@ifi.uio.no

University of Oslo

June 20, 2019

# Aim and Organization

A brief overview of the intersection of **machine learning** (ML) and **computer security** (CS):

1. A *instrumental perspective* (ML for cybersecurity);
2. A *systemic perspective* (security of ML);
3. A *societal perspective* (AI safety).

A quick tour of questions and problems (some approaches and potential solutions in the references).

# What is machine learning?

ML is a collection of techniques and algorithms to solve **inductive optimization problems**.

- Right question (*right objective function, constraints, and metrics*);
- Right examples (*right data set*);
- Right assumptions (*right family of models*);
- Right solution strategy (*right algorithm*).

The result is a **model**.

# Instrumental Perspective

ML is a **tool** to solve CS inductive optimization problems.

What questions do we want to ask?

- How do we design an *optimal defense system*?
  - How do we detect malicious behaviour in communications? [Rahbarinia et al., 2013; Zhang et al., 2014; Yen and Reiter, 2008]
  - How do we detect dangerous domain names? [Lison and Mavroeidis, 2018; Le et al., 2018; Schiavoni et al., 2014; Bilge et al., 2011]
  - How do we detect botnets? [Collins and Reiter, 2007]
  - How do we detect Android malware? [Lashkari et al., 2018]

# Instrumental Perspective

- How do we design an *optimal attack system*?
  - How do we exploit vulnerabilities in code? [Raff et al., 2018; Russell et al., 2018; Wu et al., 2017; Nagano and Uda, 2017; Schultz et al., 2001]
  - How do we evade network detection? [Fladby, 2018]
  - How do we design agents for CTF-like games? [Mendia et al., 2018]

# Instrumental Perspective

- How do we optimize *current systems*?
  - How do we optimize space and time of packet forwarding maintaining the same performances? [Liang et al., 2019]
- How do we optimize *privacy*?
  - How do we optimize the guarantees of privacy preservation? [Ligett et al., 2017]

# Some challenges [Papernot et al., 2016]

- Limited data
- Scalability
- Non-stationary environments
- Adversarial environments

# Systemic Perspective

ML is **part** of computer systems.

What can go wrong?

- Is the *data safe*?
  - Has the ground truth being manipulated? [Mozaffari-Kermani et al., 2015; Biggio et al., 2011]
  - Have the data been manipulated? [Steinhardt et al., 2017; Mei and Zhu, 2015; Kloft and Laskov, 2010]
  - Have the sources of the data being manipulated? [Blanchard et al., 2017; Ghodsi et al., 2017]

# Systemic Perspective

- Is the *model robust*?
  - Can the model be deceived by well-crafted samples? [Goodfellow et al., 2014]
  - Can the model be deceived by compromising the source of samples? [Kurakin et al., 2016]
  - Can the model be secured by obscurity? [Szegedy et al., 2013]
- Is the *information in the model protected*?
  - Can information about the samples be extracted?
  - Can statistical property of the data be extracted? [Ateniese et al., 2015]
  - Can the model be extracted? [Tramèr et al., 2016; Fredrikson et al., 2014]

# Some challenges [Biggio and Roli, 2018; Akhtar and Mian, 2018]

- Optimistic assumptions on the environments
- Open systems
- Trade-off for security

# Societal Perspective

ML is **part** of society.

What can go wrong?

- Is the system actually *optimizing the objective we want*? [Sugiyama and Kawanabe, 2012; Amodei et al., 2016]
- Is the system going to take *dangerous actions*? [Saunders et al., 2017; Abbeel and Ng, 2005; Hadfield-Menell et al., 2016]
- Are the actions taken by the systems *societally fair* [Corbett-Davies et al., 2017; Chouldechova, 2017]?
- Can the actions taken by the systems be *explained* [Caruana et al., 2015; Simonyan et al., 2013; Montavon et al., 2017; Ribeiro et al., 2016]?

# Some challenges [Biggio and Roli, 2018; Akhtar and Mian, 2018]

- Complex domain
- Controversial questions

# Thanks!

Thank you for listening!

# References I

Pieter Abbeel and Andrew Y Ng. Exploration and apprenticeship learning in reinforcement learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 1–8. ACM, 2005.

Naveed Akhtar and Ajmal Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *arXiv preprint arXiv:1801.00553*, 2018.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

# References II

Giuseppe Ateniese, Luigi V Mancini, Angelo Spognardi, Antonio Villani, Domenico Vitali, and Giovanni Felici. Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers. *International Journal of Security and Networks*, 10(3):137–150, 2015.

Battista Biggio and Fabio Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84: 317–331, 2018.

Battista Biggio, Blaine Nelson, and Pavel Laskov. Support vector machines under adversarial label noise. In *Asian Conference on Machine Learning*, pages 97–112, 2011.

Leyla Bilge, Engin Kirda, Christopher Kruegel, and Marco Balduzzi. Exposure: Finding malicious domains using passive dns analysis. In *Ndss*, 2011.

# References III

Peva Blanchard, Rachid Guerraoui, Julien Stainer, et al. Machine learning with adversaries: Byzantine tolerant gradient descent. In *Advances in Neural Information Processing Systems*, pages 118–128, 2017.

Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1721–1730. ACM, 2015.

Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5 (2):153–163, 2017.

# References IV

M Patrick Collins and Michael K Reiter. Hit-list worm detection and bot identification in large networks using protocol graphs. In *International Workshop on Recent Advances in Intrusion Detection*, pages 276–295. Springer, 2007.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 797–806. ACM, 2017.

Torgeir Fladby. Adaptive network flow parameters for stealthy botnet behavior. Master's thesis, University of Oslo, 2018.

# References V

Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon Lin, David Page, and Thomas Ristenpart. Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. In *USENIX Security Symposium*, pages 17–32, 2014.

Zahra Ghodsi, Tianyu Gu, and Siddharth Garg. Safetynets: Verifiable execution of deep neural networks on an untrusted cloud. In *Advances in Neural Information Processing Systems*, pages 4675–4684, 2017.

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

# References VI

Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. Cooperative inverse reinforcement learning. In *Advances in neural information processing systems*, pages 3909–3917, 2016.

Marius Kloft and Pavel Laskov. Online anomaly detection under adversarial impact. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 405–412, 2010.

Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016.

# References VII

Arash Habibi Lashkari, Andi Fitriah A Kadir, Laya Taheri, and Ali A Ghorbani. Toward developing a systematic approach to generate benchmark android malware datasets and classification. In *2018 International Carnahan Conference on Security Technology (ICCST)*, pages 1–7. IEEE, 2018.

Hung Le, Quang Pham, Doyen Sahoo, and Steven CH Hoi. Urlnet: Learning a url representation with deep learning for malicious url detection. *arXiv preprint arXiv:1802.03162*, 2018.

Eric Liang, Hang Zhu, Xin Jin, and Ion Stoica. Neural packet classification. *arXiv preprint arXiv:1902.10319*, 2019.

Katrina Ligett, Seth Neel, Aaron Roth, Bo Waggoner, and Z Steven Wu. Accuracy first: Selecting a differential privacy level for accuracy-constrained erm. *arXiv preprint arXiv:1705.10829*, 2017.

# References VIII

Pierre Lison and Vasileios Mavroeidis. Neural reputation models learned from passive dns data. 2018.

Shike Mei and Xiaojin Zhu. Using machine teaching to identify optimal training-set attacks on machine learners. In *AAAI*, pages 2871–2877, 2015.

Gorka Olalde Mendia, Lander Usategui San Juan, Xabier Perez Bascaran, Asier Bilbao Calvo, Alejandro Hernández Cordero, Irati Zamalloa Ugarte, Aday Muñiz Rosas, David Mayoral Vilches, Unai Ayucar Carbajo, Laura Alzola Kirschgens, et al. Robotics ctf (rctf), a playground for robot hacking. *arXiv preprint arXiv:1810.02690*, 2018.

Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 2017.

# References IX

Mehran Mozaffari-Kermani, Susmita Sur-Kolay, Anand Raghunathan, and Niraj K Jha. Systematic poisoning attacks on and defenses for machine learning in healthcare. *IEEE journal of biomedical and health informatics*, 19(6):1893–1905, 2015.

Yuta Nagano and Ryuya Uda. Static analysis with paragraph vector for malware detection. In *Proceedings of the 11th International Conference on Ubiquitous Information Management and Communication*, page 80. ACM, 2017.

Nicolas Papernot, Patrick McDaniel, Arunesh Sinha, and Michael Wellman. Towards the science of security and privacy in machine learning. *arXiv preprint arXiv:1611.03814*, 2016.

# References X

Edward Raff, Jon Barker, Jared Sylvester, Robert Brandon, Bryan Catanzaro, and Charles K Nicholas. Malware detection by eating a whole exe. In *Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Babak Rahbarinia, Roberto Perdisci, Andrea Lanzi, and Kang Li. Peerrush: Mining for unwanted p2p traffic. In *International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment*, pages 62–82. Springer, 2013.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM, 2016.

# References XI

Rebecca L Russell, Louis Kim, Lei H Hamilton, Tomo Lazovich, Jacob A Harer, Onur Ozdemir, Paul M Ellingwood, and Marc W McConley. Automated vulnerability detection in source code using deep representation learning. *arXiv preprint arXiv:1807.04320*, 2018.

William Saunders, Girish Sastry, Andreas Stuhlmueller, and Owain Evans. Trial without error: Towards safe reinforcement learning via human intervention. *arXiv preprint arXiv:1707.05173*, 2017.

Stefano Schiavoni, Federico Maggi, Lorenzo Cavallaro, and Stefano Zanero. Phoenix: Dga-based botnet tracking and intelligence. In *International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment*, pages 192–211. Springer, 2014.

# References XII

Matthew G Schultz, Eleazar Eskin, F Zadok, and Salvatore J Stolfo. Data mining methods for detection of new malicious executables. In *Security and Privacy, 2001. S&P 2001. Proceedings. 2001 IEEE Symposium on*, pages 38–49. IEEE, 2001.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Jacob Steinhardt, Pang Wei Koh, and Percy Liang. Certified defenses for data poisoning attacks. *arXiv preprint arXiv:1706.03691*, 2017.

Masashi Sugiyama and Motoaki Kawanabe. *Machine learning in non-stationary environments: introduction to covariate shift adaptation*. MIT Press, 2012.

# References XIII

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart. Stealing machine learning models via prediction apis. In *USENIX Security Symposium*, pages 601–618, 2016.

Fang Wu, Jigang Wang, Jiqiang Liu, and Wei Wang. Vulnerability detection with deep learning. In *Computer and Communications (ICCC), 2017 3rd IEEE International Conference on*, pages 1298–1302. IEEE, 2017.

# References XIV

Ting-Fang Yen and Michael K Reiter. Traffic aggregation for malware detection. In *International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment*, pages 207–227. Springer, 2008.

Junjie Zhang, Roberto Perdisci, Wenke Lee, Xiapu Luo, and Unum Sarfraz. Building a scalable system for stealthy p2p-botnet detection. *IEEE transactions on information forensics and security*, 9(1):27–38, 2014.