The (new) attack surfaces of data-learned models

Adversarial attacks and defenses for ML models

Fabio Massimo Zennaro
fabiomz@ifi.uio.no

University of Oslo

December 11, 2020

Introduction

Overview of safety issues of data-learned models for decision making considering their *potential attack surfaces*.

Conceptual and limited overview (references provided).

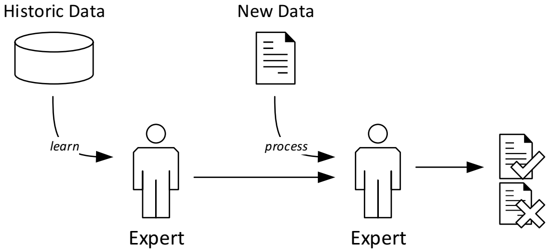
We will discuss using a *case study*/*analogy*: problem of classifying satellite pictures to decide whether they contain military installations.

Outline

1. *ML decision systems and their attack surfaces*
2. *Attacks on learning*
3. *Attacks on inference*
4. *Final remarks*

1. ML attack surfaces

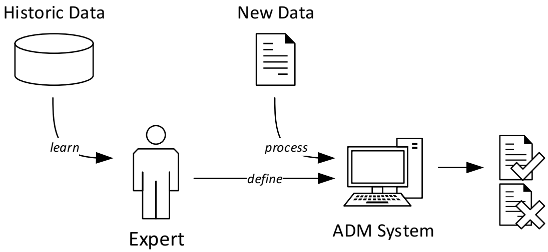
Decision-making



Human decision making

- × Very slow learning and processing
- × Prone to human vulnerabilities/errors

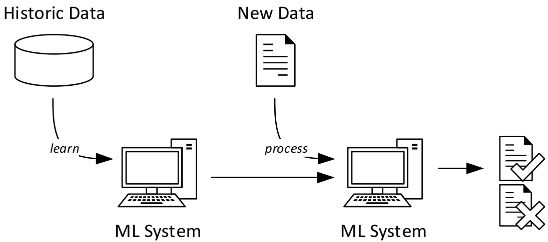
Automatic decision-making



Logical/Deductionist/Human-distilled/GOFAI

- × Still learned by human (slow)
- ✓ Faster, more consistent decisions

ML decision-making



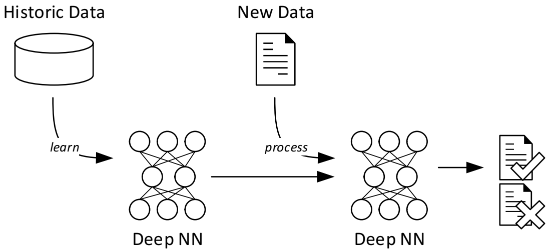
Statistical/Inductionist/Data-learned/ML

- ✓ Learned by machines (fast)
- ✓ Fast and highly accurate decisions

ML approach

The *ML approach* now usually refers to **deep neural networks** for *supervised learning*.

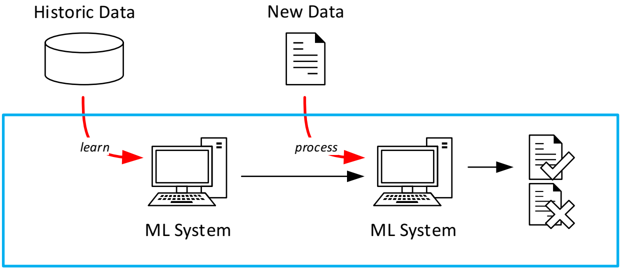
- ✓ Very effective in terms of accuracy, training time and processing time



Is this system *safe*?

ML attack surfaces

What is the *attack surface* of a ML system?



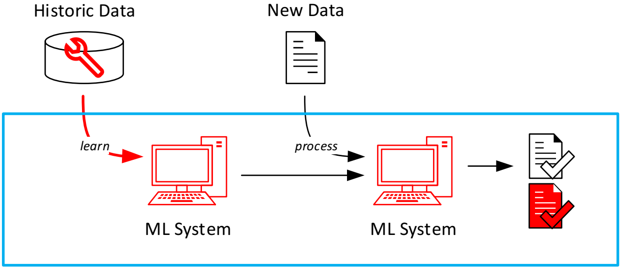
We have two processes that open a surface for attack:

1. **Learning** relying on external *historic data*
2. **Inference** given external *new data*

2. Attacks on Learning

Attacks on Learning

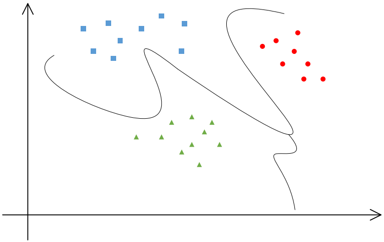
Attacks aimed at **compromising the learning** process (a.k.a. *learning-time attacks*, *data attack*, *poisoning*).



Analogy: provide the learner with incorrect satellite images.

A glimpse into the learning process (1)

Learning in ML is a **data-driven optimization process** aimed at *learning a function* by *gradient descent*.

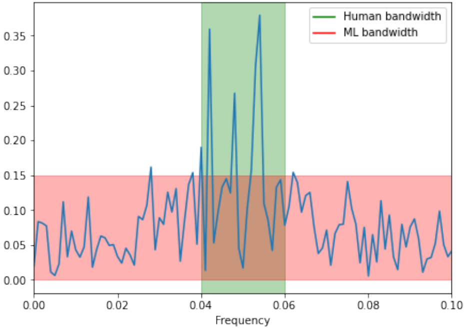


(Analogy is stretched!)

A glimpse into the learning process (2)

Learning in ML is a **data-driven optimization process** relying on *correlations* in a *signal* with no *common-sense context*.

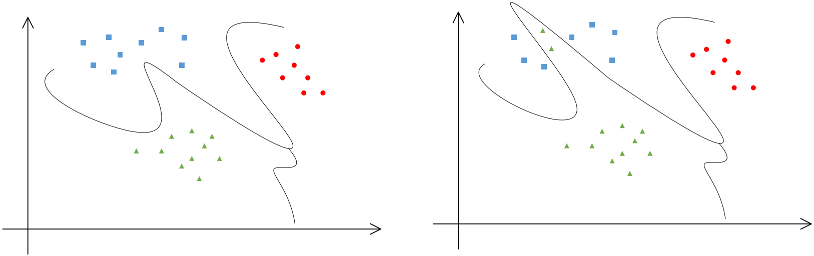
Image from Mayraz and Hinton [2002]



(Analogies are stretched!)

Poisoning (1)

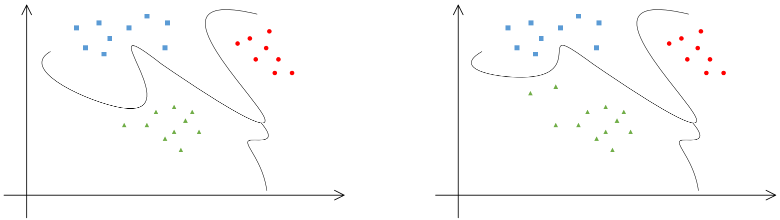
Label manipulation: harmful perturbation of labels [Biggio et al., 2011; Mozaffari-Kermani et al., 2015]



Analogy: provide the learner with images of military installations but tell her they are farms.

Poisoning (2)

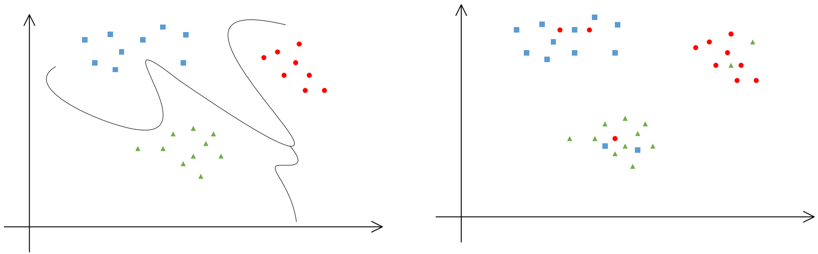
Direct/indirect data poisoning: modification of the data or the data generating process to generate malicious samples [Kloft and Laskov, 2010; Mei and Zhu, 2015; Steinhardt et al., 2017; Perdisci et al., 2006]



Analogy: compromise the data (or the sources) so that the images of farms the learner sees are very similar to military installations.

Poisoning (3)

Denial: insertion of random data points to prevent learning.

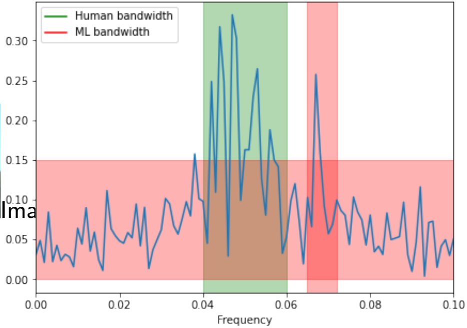


Analogy: provide the learner with random images and random explanation of satellite images.

Backdoor

Backdoor: insertion of a signal to misdirect learning [Chen et al., 2017; Gu et al., 2017].

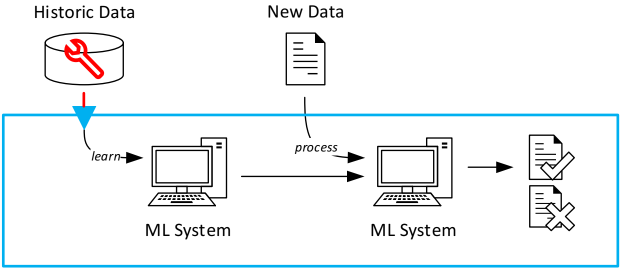
from Gu et al. [2017]



Analogy: insert a subtle cue in all the images of farms (e.g.: cows) so that if a learner see it, she concludes she is seeing a farm.

Defenses

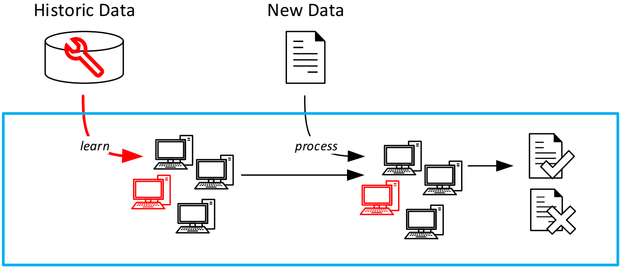
Input Validation: verify sources and their reliability
Input Pre-processing: filter the inputs



Analogy: guarantee that a learner receives reliable satellite images and that they have not been manipulated.

Defenses

Ensembling: train multiple models on random subsets of data

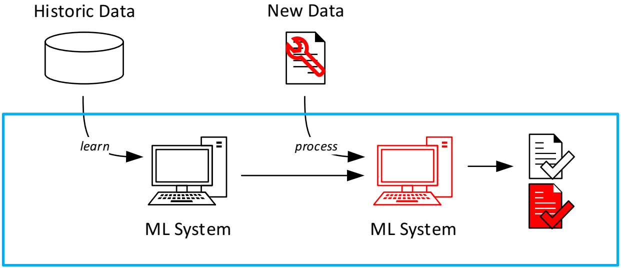


Analogy: provide each learner with a subset of satellite pictures, so that each subset has low probability of containing poisoned data.

3. Attacks on Inference

Attacks on Inference

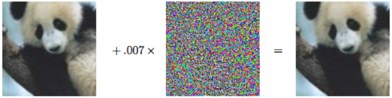
Attacks aimed at **compromising the inference** process. (a.k.a. *inference-time attacks*, *adversarial samples attack*)



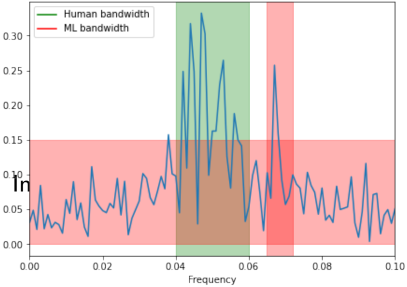
Analogy: provide the expert with modified satellite pictures that exploit her weak points in decision making.

Adversarial Samples

Direct Adversarial Samples: insertion of a signal to misdirect learning [Szegedy et al., 2013; Goodfellow et al., 2014].



from Goodfellow et al. [2014]

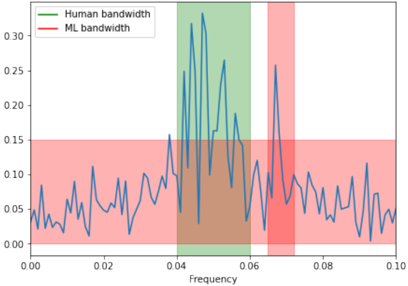


Analogy: modify the satellite images with the required cues as little as necessary to trick the expert.

Adversarial Samples

Indirect Adversarial Samples: insertion of adversarial examples in the data processing pipeline [Kurakin et al., 2016].

Image from Kurakin et al. [2016]



Generating Adversarial Samples

Many techniques to generate adversarial samples [Akhtar and Mian, 2018]: *fast gradient sign method* [Goodfellow et al., 2014], *projected gradient descent* [Madry et al., 2017], *DeepFool* [Moosavi Dezfooli et al., 2016], *C&W attacks* [Carlini and Wagner, 2017].

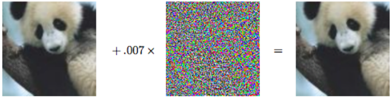
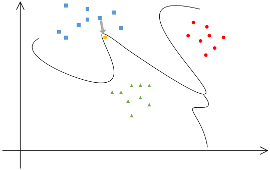


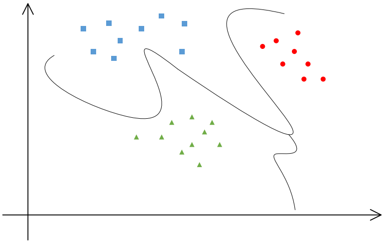
Image from Goodfellow et al. [2014]

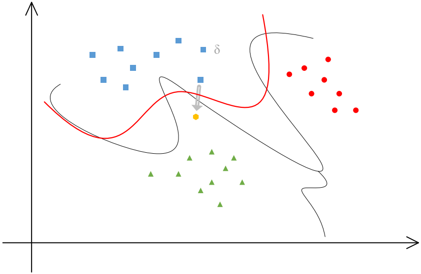


Analogy: find the minimal cue that will exploit the weak point of the expert.

Transferring Adversarial Samples

Adversarial examples may be computed on surrogate in-house models and then deployed against target systems.

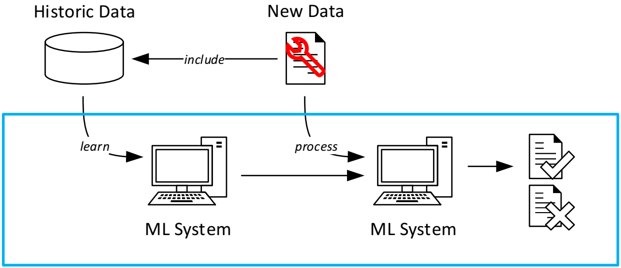




Analogy: you don't need to know the exact expert you are trying to fool; it is enough to be able to fool an expert trained in a similar way.

Defenses

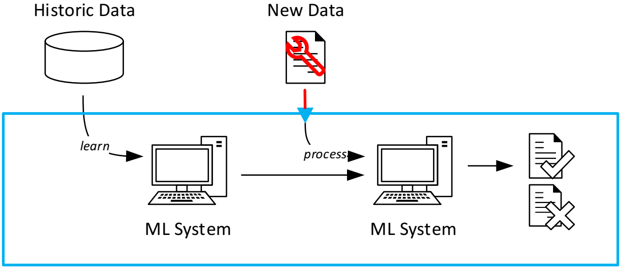
Adversarial training: use adversarial samples to train your model and make it robust against attacks



Analogy: teach your expert how he may be fooled.

Defenses

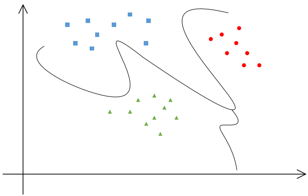
Input Pre-processing: filter the inputs



Analogy: try to remove malicious cues from the satellite images before they are delivered to the expert.

Defenses

Gradient obfuscation: make the computation of adversarial examples hard/impossible [Athalye et al., 2018].



Analogy: prevent an attacker from knowing what are the weak points of your expert.

4. Final Remarks

ML safety

There is a relevant amount of research on *ML safety*.

Two main traditions of research [Biggio and Roli, 2018]:

- *Security of ML (∼2004-2005)*: studying security of ML models in the computer security field [Dalvi et al., 2004];
- *Adversarial ML (∼2014)*: studying security of deep ML models [Szegedy et al., 2013]

Characterizing the Defense

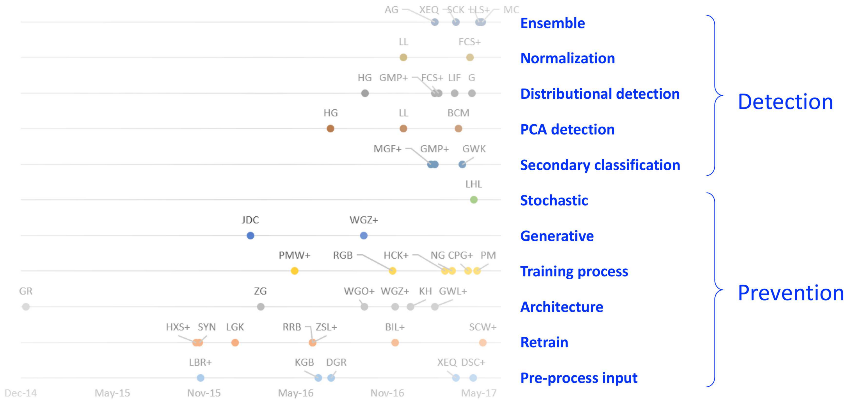


Figure from [Song, 2018]

Characterizing the threat

We explored vulnerabilites from the perspective of *attack surface*, but other characterizations are possible [Papernot et al., 2016; Biggio and Roli, 2018]

Attacker Knowledge:

- *White-box knowledge:* perfect knowledge of systems
- *Gray-box knowledge:* partial knowledge of systems
- *Black-box knowledge:* minimal knowledge of systems

Attacker Specificity:

- *Targeted:* aimed at specific effect
- *Indiscriminate:* aimed at subversion

Characterizing the threat

We explored vulnerabilites from the perspective of *attack surface*, but other characterizations are possible [Papernot et al., 2016; Biggio and Roli, 2018]

Attacker Constraint:

- *Min-perturbation:* given the desired effect, choose the attack that minimize the detectability.
- *Max-confidence:* given the possible perturbation, choose the attack that maximize the effect.

Attacker Goal:

- *Integrity-Availability:* compromise learning or inference
- *Confidentiality-Privacy:* extracting information

Characterizing the defense

Defenses may be characterized too from other perspectives: [Biggio and Roli, 2018; Akhtar and Mian, 2018]

Defense Stance:

- *Reactive:* readily address new attacks
- *Proactive:* plan to prevent future attacks

Defense Paradigm:

- *Detection:* catch new attacks in advance
- *Prevention:* be resistant to attacks

Defense Target:

- *Data:* modify the data to increase defense
- *Model:* modify the model to improve robustness
- *Other:* extend the system

Some Good Principles

Good principles for security with ML models [Kolter and Madry, 2018; Biggio and Roli, 2018]:

1. Do not train on untrusted data
2. Do not allow access to model to untrusted agents
3. Do not fully trust predictions

1. Design for security
2. Detect
3. Retrain
4. Verify

(Some) Conclusions

- Attacks on ML models are a *possibility* (how real they are is a matter of cost) [Schwarzschild et al., 2020; Shafahi et al., 2018]
- Audit your ML system and trace its *attack surfaces*.
- For ML too, security-by-obscurity is not security.
- Inevitably, information flows from your ML system to the outside world.
- You may have *trade off* effectiveness for security.

Thanks!

Thank you for listening!

References I

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

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *arXiv preprint arXiv:1802.00420*, 2018.

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.

References II

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. IEEE, 2017.

Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017.

Nilesh Dalvi, Pedro Domingos, Sumit Sanghai, Deepak Verma, et al. Adversarial classification. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 99–108. ACM, 2004.

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

References III

Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*, 2017.

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.

Zico Kolter and Aleksander Madry. Adversarial robustness: Theory and practice. 12 2018.

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

References IV

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Guy Mayraz and Geoffrey E Hinton. Recognizing handwritten digits using hierarchical products of experts. *IEEE transactions on pattern analysis and machine intelligence*, 24(2):189–197, 2002.

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

References V

Seyed Mohsen Moosavi Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, number EPFL-CONF-218057, 2016.

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.

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 VI

Roberto Perdisci, David Dagon, Wenke Lee, Prahlad Fogla, and Monirul Sharif. Misleading worm signature generators using deliberate noise injection. In *Security and Privacy, 2006 IEEE Symposium on*, pages 15–pp. IEEE, 2006.

Avi Schwarzschild, Micah Goldblum, Arjun Gupta, John P Dickerson, and Tom Goldstein. Just how toxic is data poisoning? a unified benchmark for backdoor and data poisoning attacks. *arXiv preprint arXiv:2006.12557*, 2020.

Ali Shafahi, W Ronny Huang, Christoph Studer, Soheil Feizi, and Tom Goldstein. Are adversarial examples inevitable? *arXiv preprint arXiv:1809.02104*, 2018.

Dawn Song. Ai and security: Lessons, challenges and future directions. 07 2018.

References VII

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

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.