Threats to Machine Learning Applications

Mark Sherman
Director, Cybersecurity Foundations, CERT
Oct 6, 2020

Software Engineering Institute
Carnegie Mellon University
Pittsburgh, PA 15213

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

DM20-0791

Carnegie Mellon Leads an Ecosystem of Innovation for Cybersecurity

CMU Campus – Global Research University

- Global research university known for its world-class, interdisciplinary programs in computer science, machine learning/artificial intelligence, engineering, business, arts, policy, and science
- Ranked #1 for Computer Science, #1 for Artificial Intelligence, #6 in Engineering (*U.S. News and World Report*)
- 1,442 total faculty and 130 research centers
- CyLab, CMU's security and privacy research institute, brings together experts from all schools across the university

CMU Software Engineering Institute (SEI)

- Founded in 1984 by the DoD as a Federally-Funded Research and Development Center (FFRDC) focused on software engineering
- Leader in software engineering, cybersecurity, and artificial intelligence research
- Established CERT in 1988
- About $145M annual funding (~$23M DoD Line)
- Critical to the DoD ability to acquire, develop, operate, and sustain software systems that are innovative, affordable, trustworthy, and enduring *(CMU SEI Sponsoring Agreement)*

CERT Division

Founded on a unique combination of experiential understanding of DoD missions, the cyber warfighter, the operational domain, and constantly changing technology

Adapts the best science to impact operational missions, increase the trustworthiness of technology, and develop cyber talent

Partners with DoD, non-DoD agencies, and the private sector enable CERT to maintain technical depth, attract top talent, amplify DoD financial investment, reduce the risk to DoD missions, and scale the research

Strengthens the resilience of critical national functions, increases the cybersecurity and resilience of DoD systems and Defense Industrial Base, and develops the cyber capacity of allies and partners

Outline

Stop signs that look like speed limit signs

K. Eykholt, et al, "Robust Physical-World Attacks on Deep Learning Visual Classification," CVPR 2018, April 10, 2018, https://arxiv.org/pdf/1707.08945.pdf

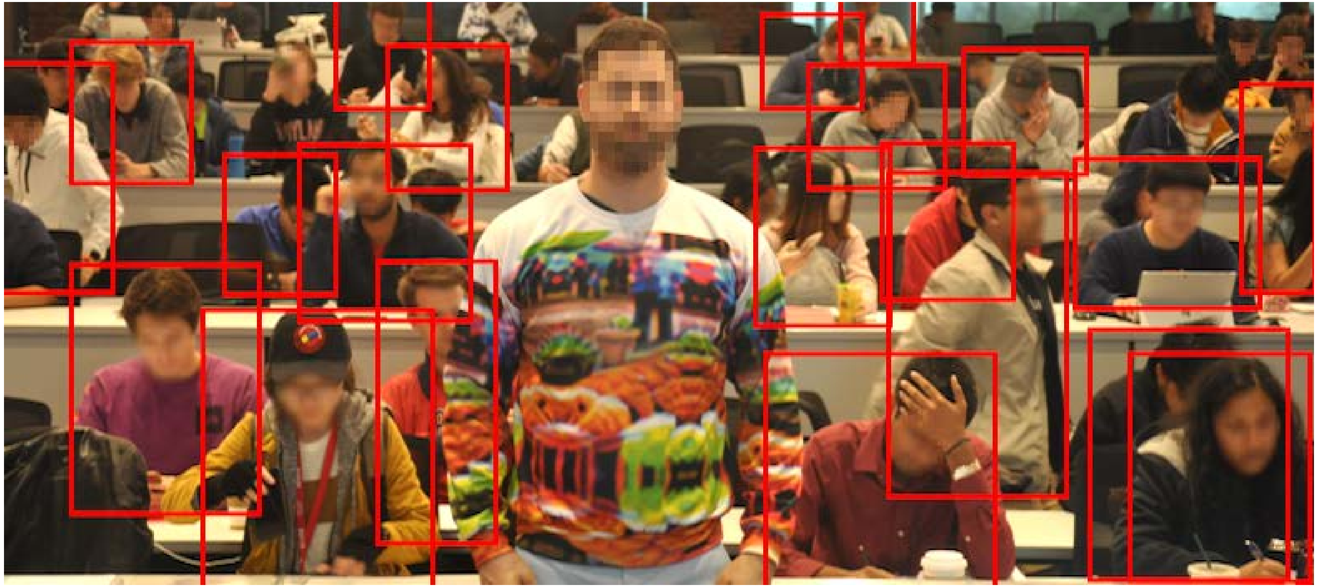
Turtles that look like rifles

Source: Athalye, A., Engstrom, L., Ilyas, A., & Kwok, K. (2017, July 24). *Synthesizing Robust Adversarial Examples*. *arXiv [cs.CV]*. Retrieved from http://arxiv.org/abs/1707.07397

Glasses look like Milla Jovovich

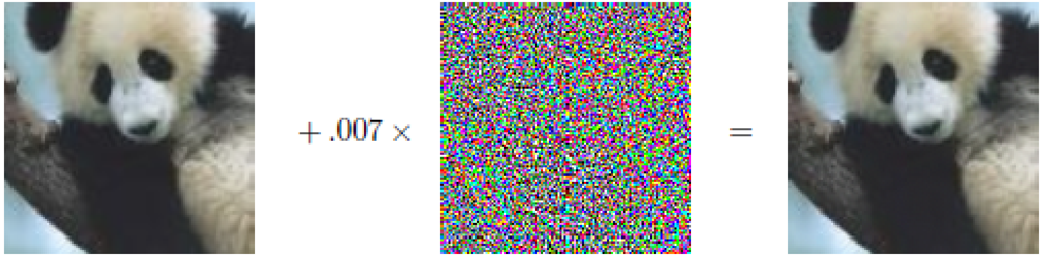
M. Sharif, et al, Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition. In *2016 ACM SIGSAC Conference on Computer and Communications Security* (CCS '16). DOI: https://doi.org/10.1145/2976749.2978392

Tee shirts hide from facial recognition



Tom Goldstein, "Invisibility Cloak," 2018,
https://www.cs.umd.edu/~tomg/projects/invisible/

Pandas as Gibbons

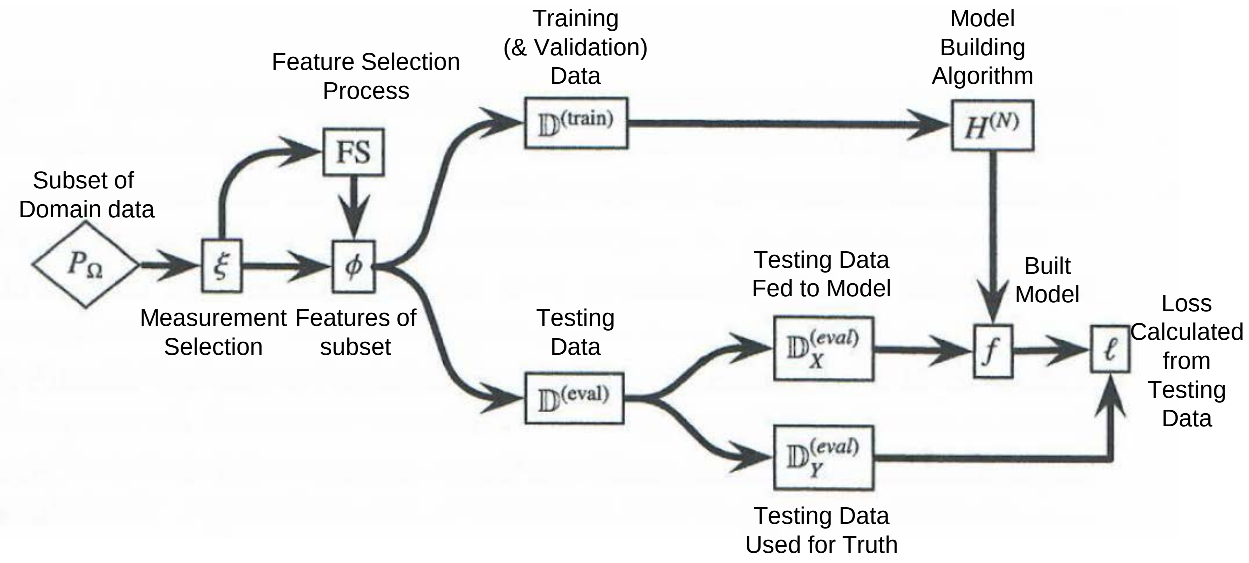


“Panda” “Gibbon”

I. Goodfellow, et al, "Explaining and Harnessing Adversarial Examples," ICLR 2015, Mar 20, 2015, https://arxiv.org/pdf/1412.6572.pdf

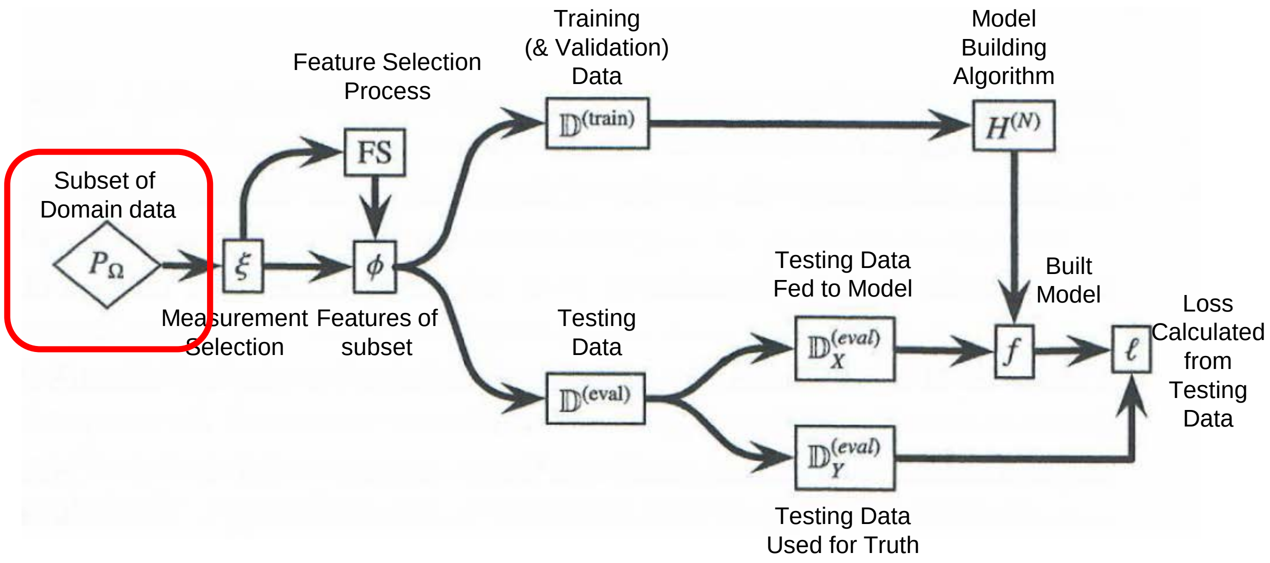
Outline

Developing a Machine Learning Application



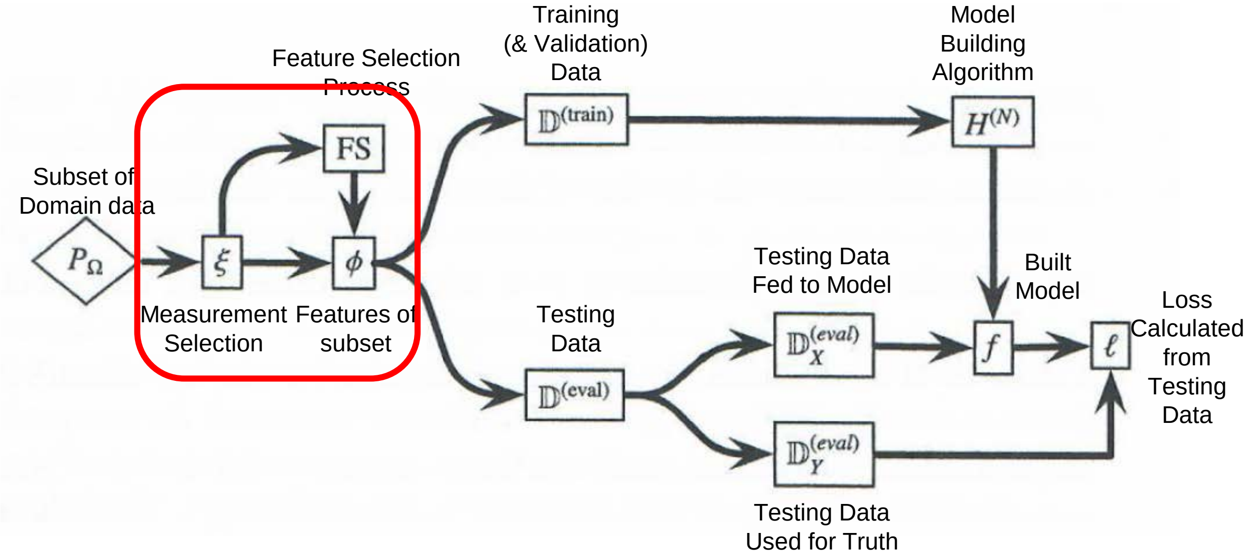
Adapted from Joseph, Nelson, Rubinstein, Tygar; Adversarial Machine Learning, Cambridge University Press, 2019

Data Attacks – Selected Domain Subset



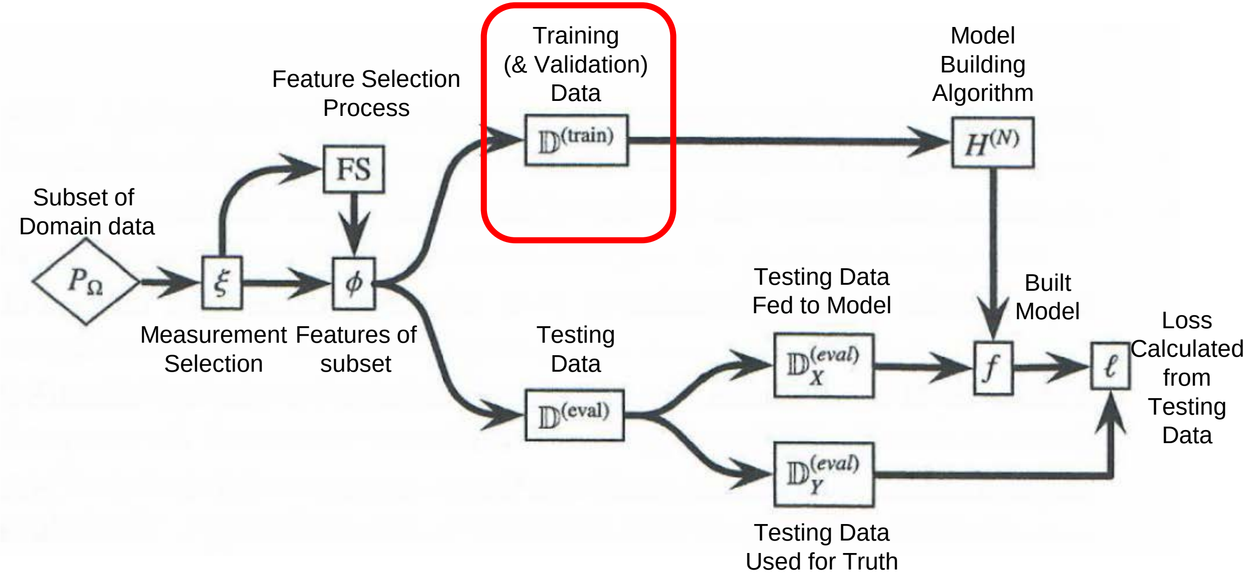
Adapted from Joseph, Nelson, Rubinstein, Tygar; Adversarial Machine Learning, Cambridge University Press, 2019

Data Attacks – Measurements and Features



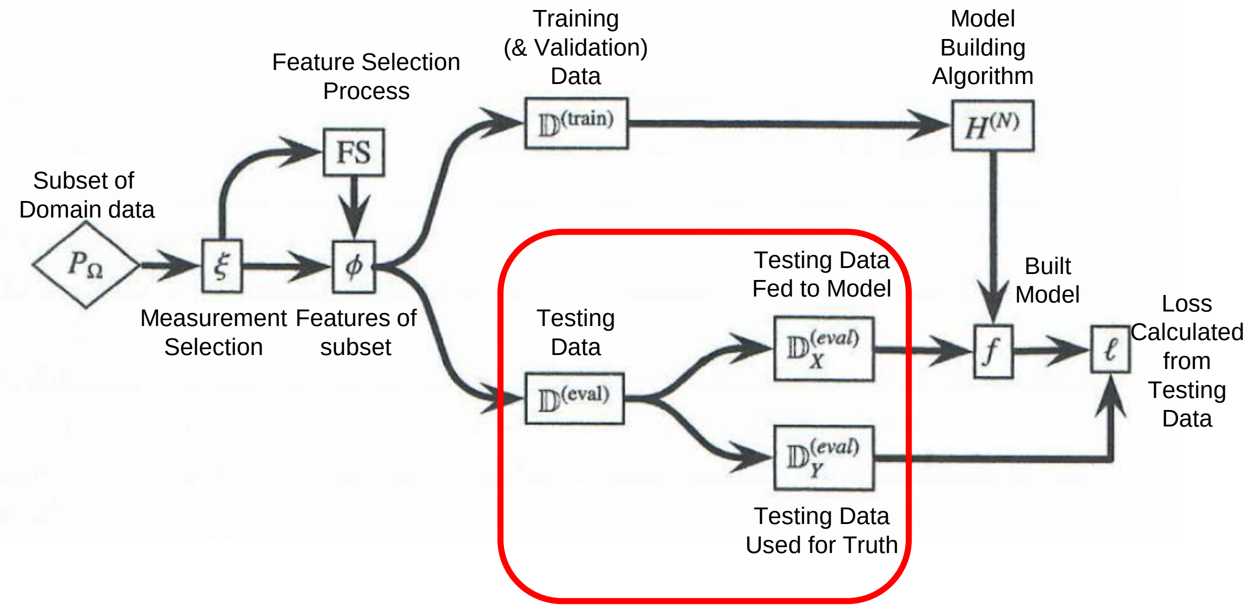
Adapted from Joseph, Nelson, Rubinstein, Tygar; Adversarial Machine Learning, Cambridge University Press, 2019

Data Attacks – Training Data



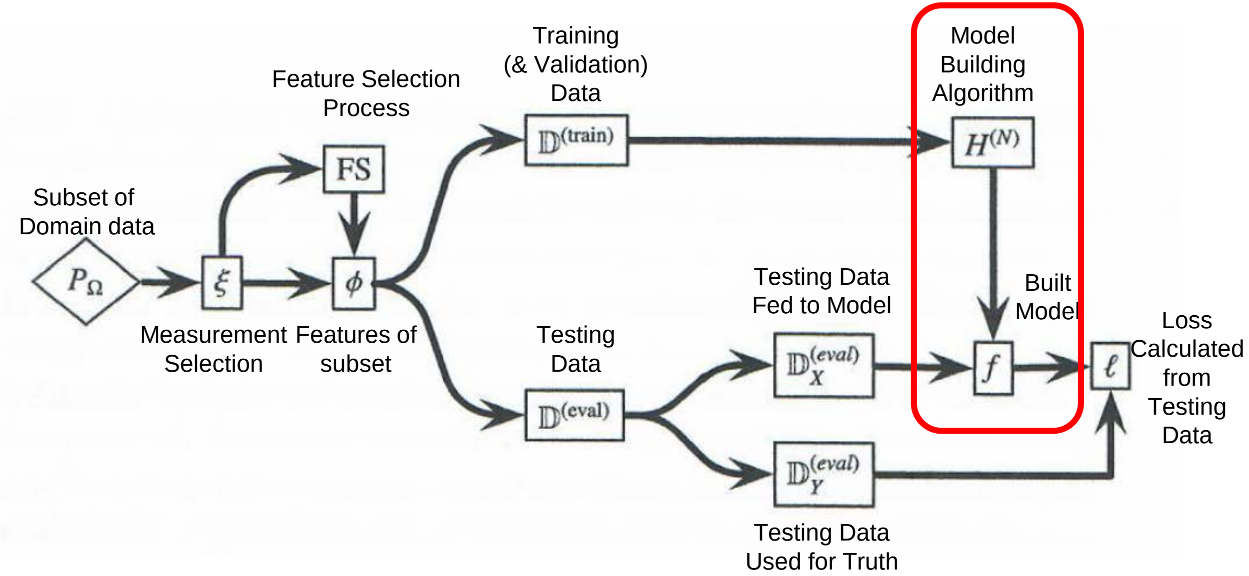
Adapted from Joseph, Nelson, Rubinstein, Tygar; Adversarial Machine Learning, Cambridge University Press, 2019

Data Attacks – Model Testing Data



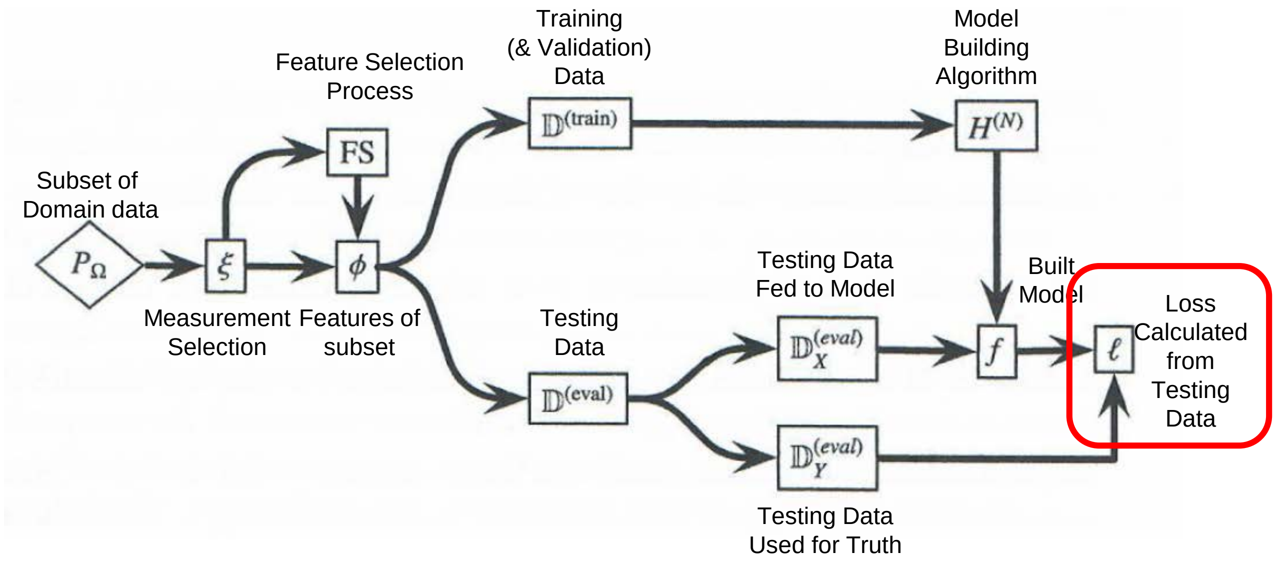
Adapted from Joseph, Nelson, Rubinstein, Tygar; Adversarial Machine Learning, Cambridge University Press, 2019

Algorithm Attacks – Model Construction



Adapted from Joseph, Nelson, Rubinstein, Tygar; Adversarial Machine Learning, Cambridge University Press, 2019

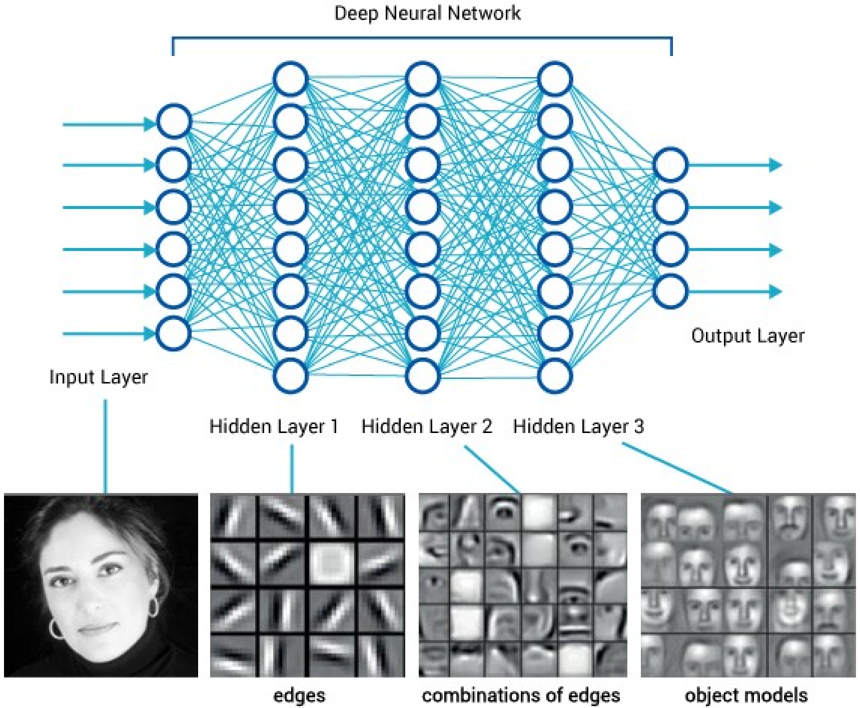
Data Attack – Loss Measurements



Adapted from Joseph, Nelson, Rubinstein, Tygar; Adversarial Machine Learning, Cambridge University Press, 2019

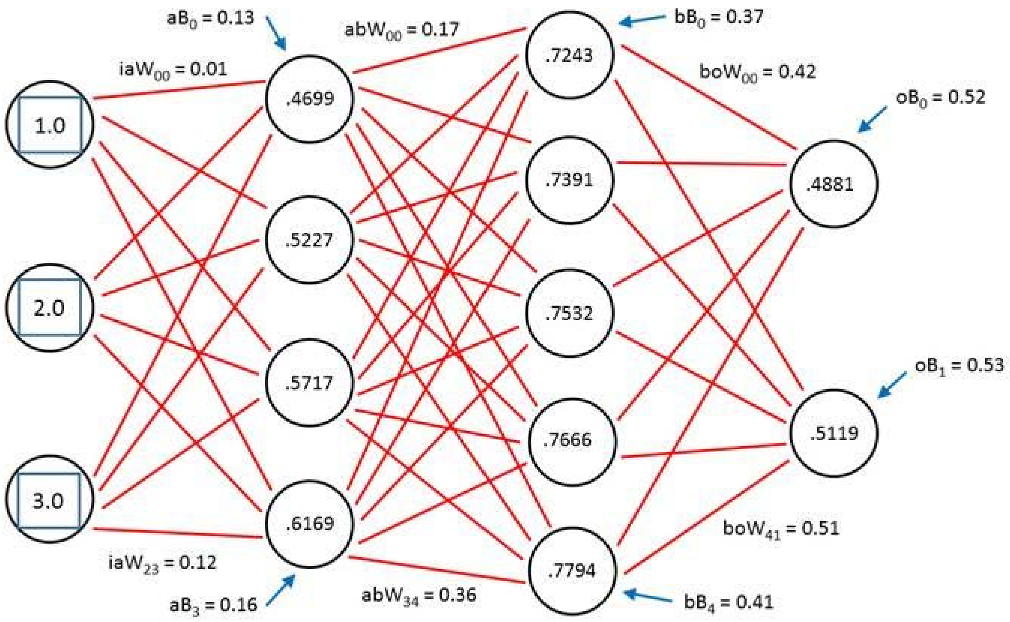
Outline

Deep Neural Network Structure



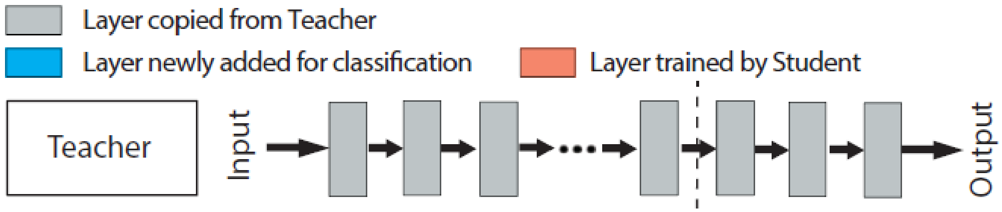
[Aashay Sachdeva](https://medium.com/@aashaysachdeva), Deep Learning for Computer Vision for the average person, Mar 6, 2017, https://medium.com/diaryofawannapreneur/deep-learning-for-computer-vision-for-the-average-person-861661d8aa61

Trained Deep Neural Network



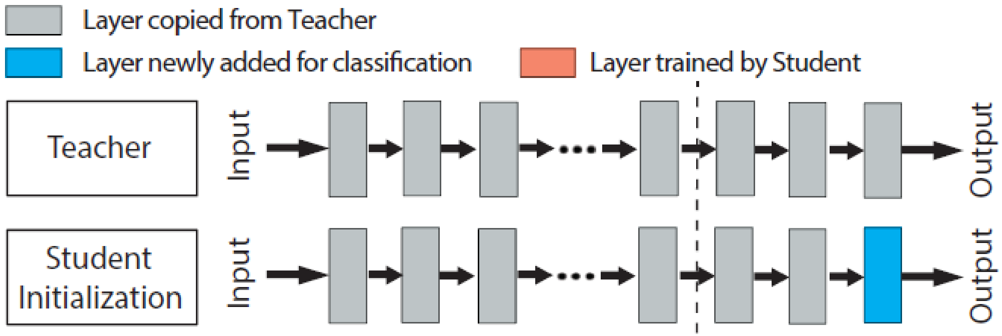
Sergey Golubev, Deep Neural Networks: A Getting Started Tutorial, Part #1, 30 June 2014, https://www.mql5.com/en/blogs/post/203

Overview of Transferring Learning



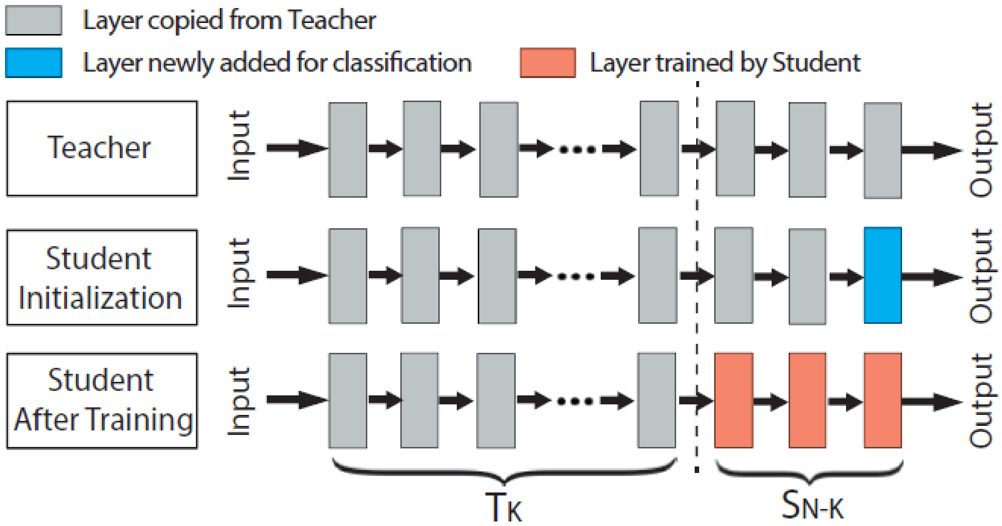
Bolun Wang, Yuanshun Yao, Bimal Viswanath, Haitao Zheng, Ben Y. Zhao; “With Great Training Comes Great Vulnerability: Practical Attacks Against Transfer Learning ,” 27th USENIX Security Symposium; Aug 15-17, 2018; pg 1281

Overview of Transferring Learning



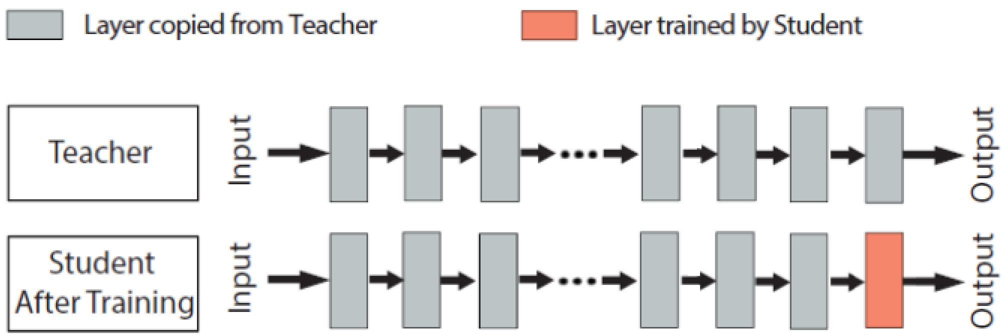
Bolun Wang, Yuanshun Yao, Bimal Viswanath, Haitao Zheng, Ben Y. Zhao; “With Great Training Comes Great Vulnerability: Practical Attacks Against Transfer Learning ,” 27th USENIX Security Symposium; Aug 15-17, 2018; pg 1281

Overview of Transferring Learning



Bolun Wang, Yuanshun Yao, Bimal Viswanath, Haitao Zheng, Ben Y. Zhao; "With Great Training Comes Great Vulnerability: Practical Attacks Against Transfer Learning ," 27th USENIX Security Symposium; Aug 15-17, 2018; pg 1281

Deep Layer Feature Extraction



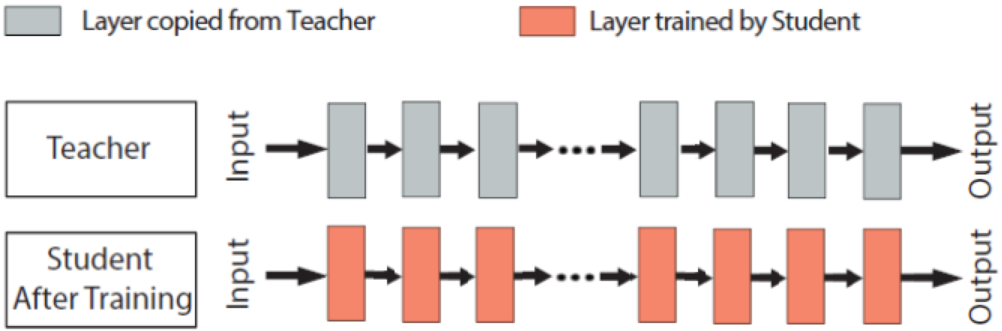
Used when domains are close

Pro: Cheap training; good accuracy

Con: Adversary has deep knowledge of teacher
- Easier to exfiltrate model
- Easier to create adversarial input

Bolun Wang, Yuanshun Yao, Bimal Viswanath, Haitao Zheng, Ben Y. Zhao; "With Great Training Comes Great Vulnerability: Practical Attacks Against Transfer Learning ," 27th USENIX Security Symposium; Aug 15-17, 2018; pg 1281

Full model fine tuning



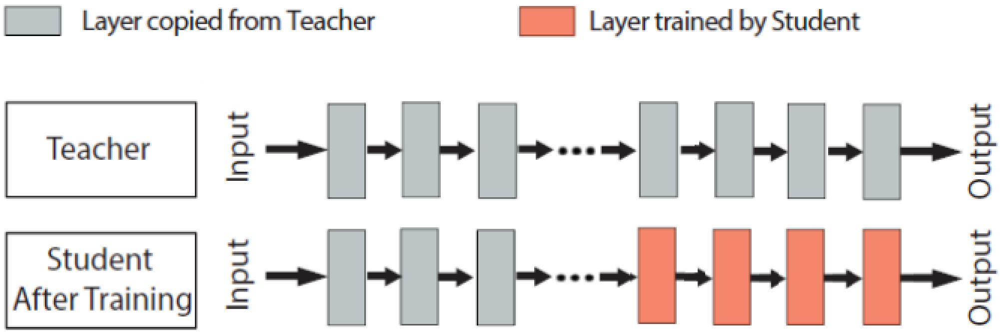
Used when domains are not close

Pro: Better accuracy than deep layer feature extraction
Resilient to teacher-specific attacks

Con: Costly to train

Bolun Wang, Yuanshun Yao, Bimal Viswanath, Haitao Zheng, Ben Y. Zhao; “With Great Training Comes Great Vulnerability: Practical Attacks Against Transfer Learning ,” 27th USENIX Security Symposium; Aug 15-17, 2018; pg 1281

Mid-Layer Feature Extraction



Compromise choice

- Accuracy depends on relationship between student and teacher domains
- Better resiliency than deep, not as good as full
- More costly to train than deep, cheaper than full

Bolun Wang, Yuanshun Yao, Bimal Viswanath, Haitao Zheng, Ben Y. Zhao; “With Great Training Comes Great Vulnerability: Practical Attacks Against Transfer Learning ,” 27th USENIX Security Symposium; Aug 15-17, 2018; pg 1281

Outline

Introduction

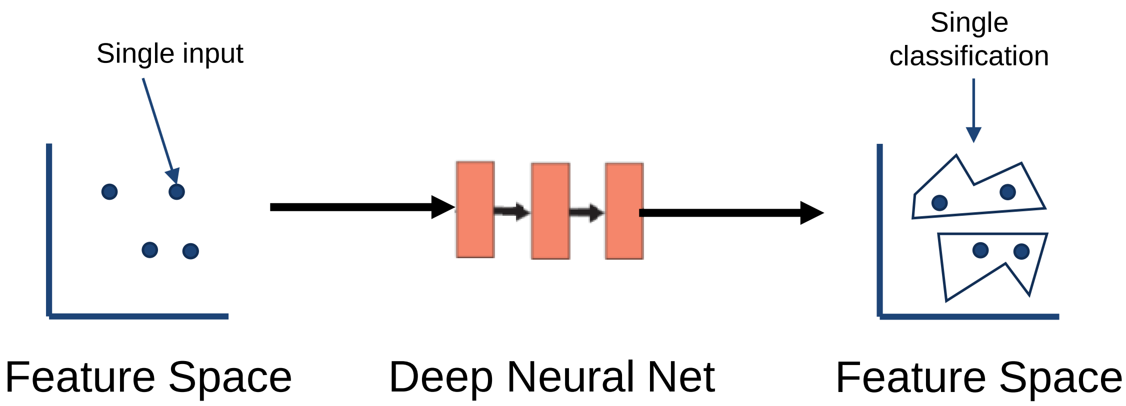
Understanding the ML Attack Surface

Understanding Risks of Transfer Learning

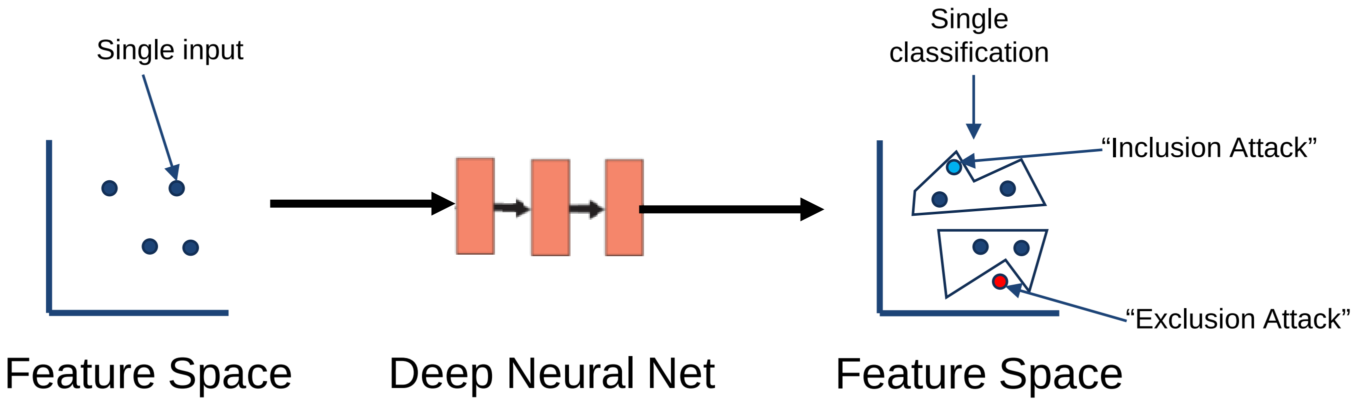
Remedies and Limitations

Conventional Threats to Machine Learning

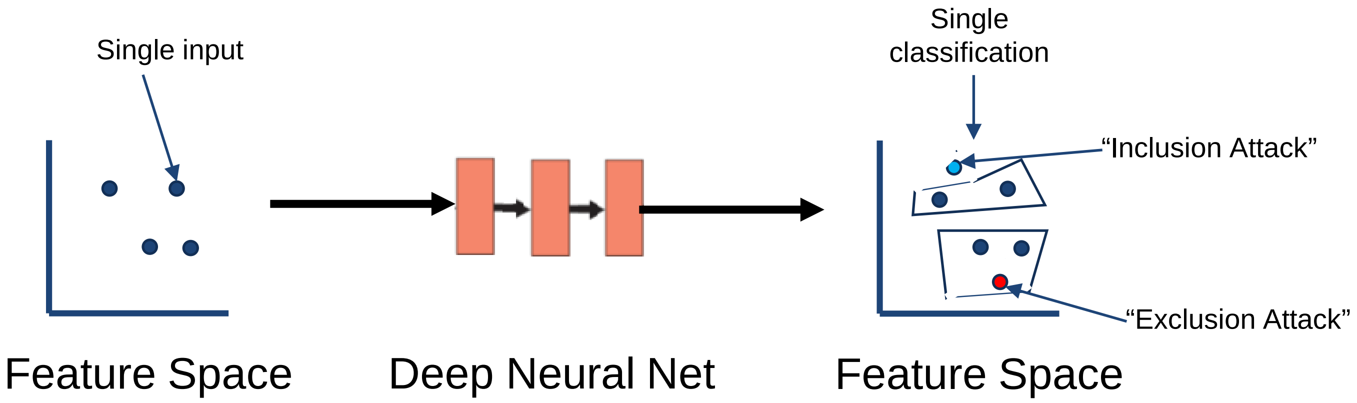
Creating Classifications



Adversarial Input



Adding Resiliency



- Cutting off spikes mitigates undesired "inclusions"
- Enclosing spikes mitigates undesired "exclusions"

Training for resilience

Methods to improve model resiliency

- Add adversarial examples in training
- Train with larger domain subset
- Calculate convex hull of classification boundary
- Apply statistical robust regression

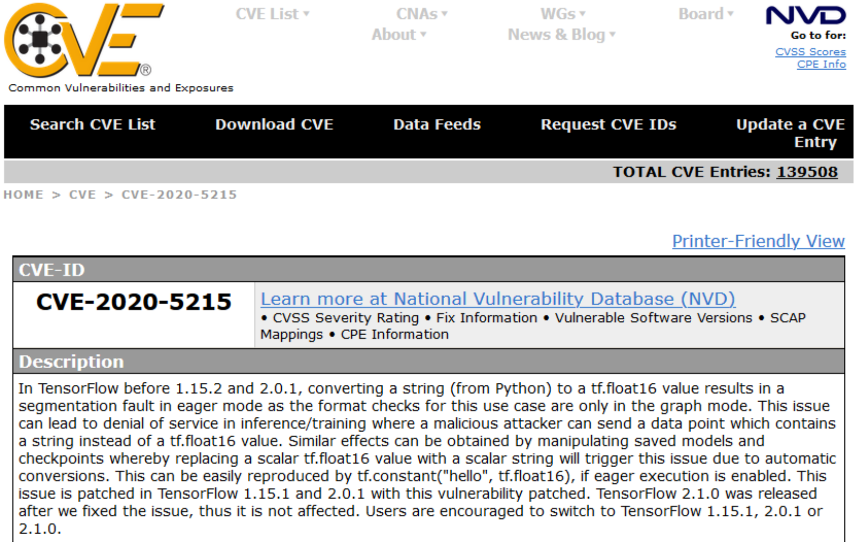
All of these methods trade resiliency for accuracy

- Adversarial examples are noisy
- Overfitting creates raggedy boundaries
- Concave boundaries could be legitimate – should be excluded
- Looser boundaries could be legitimate – should be included

Redundancy is an alternative strategy – at a cost

Outline

Coding Hygiene



https://cve.mitre.org/cgi-bin/cvename.cgi?name=CVE-2020-5215

Any of the algorithms in creating the application or in the generated application could have coding weaknesses leading to vulnerabilities

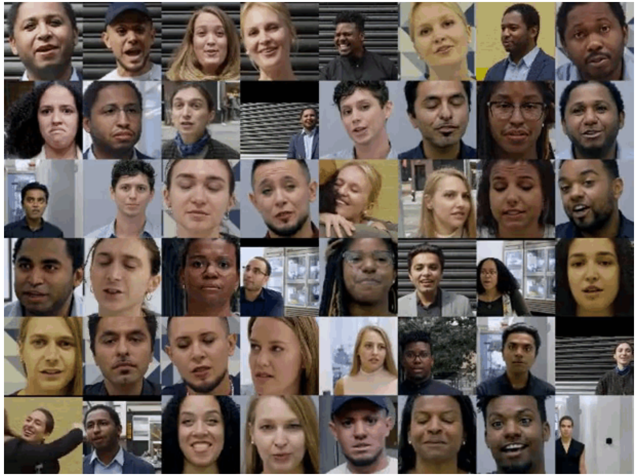
Mitigation: Good cyber hygiene

Software supply chain for assembled software

Machine learning depends on frameworks and data sets
Relatively less is known about the security of these "supplies"

Machine Learning Frameworks	Data Sources
• Pandas	• Kaggle
• Numpy	• UCI Machine Learning Repository
• Scikit-learn	• Find Datasets
• Matplotlib	• Data.gov
• TensorFlow	• xView
• Keras	• ImageNet
• Seaborn	• Google's Open Images
• Pytorch & Torch	

Machine learning system face training data supply challenges



Rich supplies of “deep fakes” are readily accessible

Source: https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html

Poor detection of deep fakes

FaceForensics Benchmark

Benchmarks · Data and Documentation · About · Submit

FaceForensics Benchmark

This table lists the benchmark results for the Binary Classification scenario.

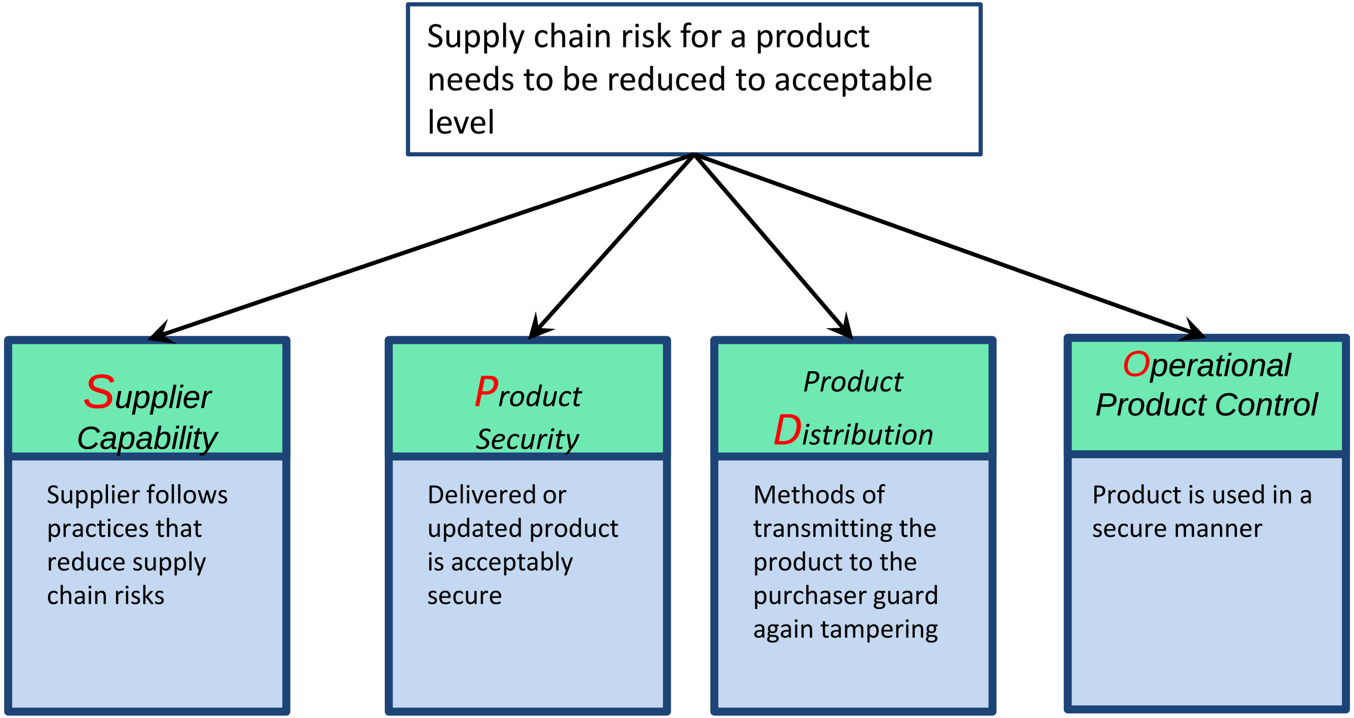
Method	Info	Deepfakes	Face2Face	FaceSwap	NeuralTextures	Pristine	Total
Xception	P	**0.964**	**0.869**	**0.903**	**0.807**	0.524	**0.710**
Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, Matthias Nießner: FaceForensics++: Learning to Detect Manipulated Facial Images. ICCV 2019							
MesoNet		0.873	0.562	0.612	0.407	**0.726**	0.660
Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen: Mesonet: a compact facial video forgery detection network. arXiv							
XceptionNet Full Image	P	0.745	0.759	0.709	0.733	0.510	0.624
Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, Matthias Nießner: FaceForensics++: Learning to Detect Manipulated Facial Images. ICCV 2019							
Bayar and Stamm		0.845	0.737	0.825	0.707	0.462	0.616
Belhassen Bayar and Matthew C. Stamm: A deep learning approach to universal image manipulation detection using a new convolutional layer. ACM Workshop on Information Hiding and Multimedia Security							
Rahmouni		0.855	0.642	0.563	0.607	0.500	0.581
Nicolas Rahmouni, Vincent Nozick, Junichi Yamagishi, and Isao Echizen: Distinguishing computer graphics from natural images using convolution neural networks. IEEE Workshop on Information Forensics and Security,							
Recasting		0.855	0.679	0.738	0.780	0.344	0.552
Davide Cozzolino, Giovanni Poggi, and Luisa Verdoliva: Recasting residual-based local descriptors as convolutional neural networks: an application to image forgery detection. ACM Workshop on Information Hiding and Multimedia Security							
Steganalysis Features		0.736	0.737	0.689	0.633	0.340	0.518
Jessica Fridrich and Jan Kodovsky: Rich Models for Steganalysis of Digital Images. IEEE Transactions on Information Forensics and Security							

Cannot reliable verify that training data obtained through a supply chain

Preconfigured machine learning (i.e., teacher) systems provide a vehicle to distribute bad training data

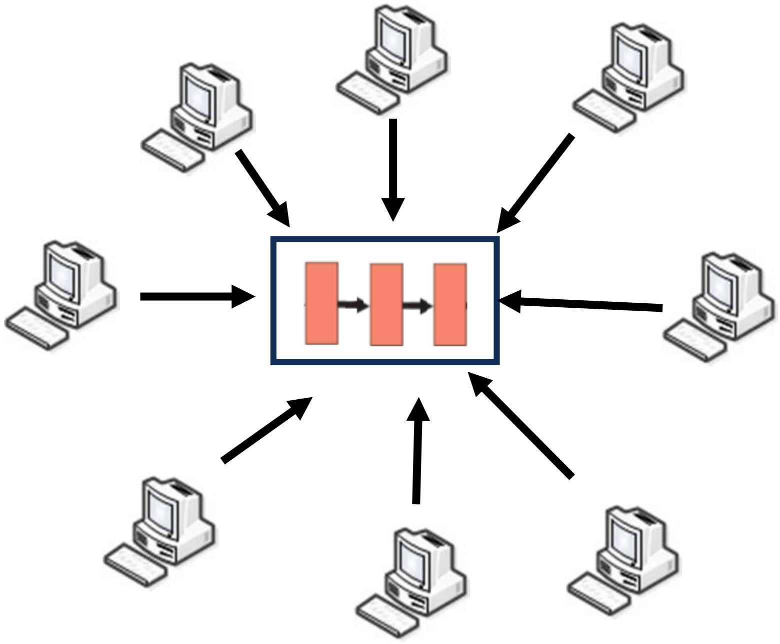
Source: http://kaldir.vc.in.tum.de/faceforensics_benchmark/index.php (as of 9/25/19)

Reducing supply chain risk factors



Ellison, Alberts, Creel, Dorofee, Woody, “Software Supply Chain Risk Management: From Products to Systems of Systems,” 2010, https://resources.sei.cmu.edu/asset_files/TechnicalNote/2010_004_001_15194.pdf

Denial of Service Attack



Remediation: Network hygiene

(https://us-cert.cisa.gov/ncas/tips/ST04-015)

Integration Points are Typically Weak

Machine learning applications are part of a system

New operating environments, i.e., interconnections between system parts, are a major cause of vulnerabilities

Extra-ML parts of the application are routes to ML attacks

Clark, Frei, Blaze, Smith, “Familiarity Breeds Contempt: The Honeymoon Effect and the Role of Legacy Code in Zero-Day Vulnerabilities,” ACSAC ’10 Dec. 6-10, 2010, p. 251-260.”

Insider Threat

Easy vector for data attacks

Remediations:

- Organizational evaluation
- Organizational processes
- Tools
- Training

https://www.sei.cmu.edu/education-outreach/courses/course.cfm?coursecode=V26

“Fake News” and AI Untrustworthiness

People ultimately use output from ML systems

Reasoning from ML systems is generally opaque

Parties can amplify potential misgivings

“Through 2021, 80% of line of business (LOB) leaders will override business decisions made by AI,” Gartner survey*

Remediations:

- Technical: Improved explanations and expectations
- Social: Education and experience

Recognize: Machine Learning is Statistics

*Graham Peters, Alan D. Duncan, Gartner Group, “100 Data and Analytics Predictions Through 2024,” March 20, 2020, pg 4

Outline

Introduction

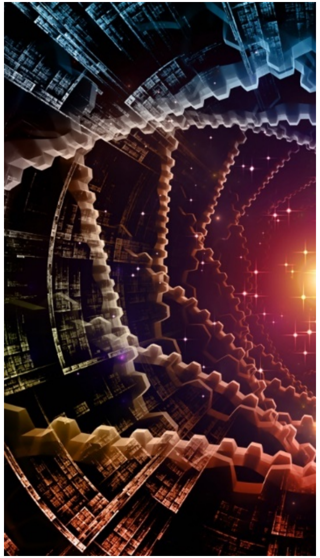
Understanding the ML Attack Surface

Understanding Risks of Transfer Learning

Remedies and Limitations

Conventional Threats to Machine Learning

Ways to Engage with Us



- Download software and tools
- Explore research and capabilities
- Participate in education offerings
- Attend an event
- Search the digital library
- Read the SEI Year in Review
- Collaborate with the SEI on a new project

Software Engineering Institute
Carnegie Mellon University
4500 Fifth Avenue
Pittsburgh, PA 15213-3890
412-268-5800 - Phone
888-201-4479 - Toll-Free
412-268-5758 - Fax
info@sei.cmu.edu - Email
www.sei.cmu.edu - Web