Best Practices for ML in Security

Jelena Milosevic
Senior Data Scientist, Mondi AG

Who Am I?

- Currently: Senior Data Scientist @ Mondi AG
- Postdoc & Project Assistant @ TU Wien, Vienna, Austria
- PhD in Informatics @ USI-Lugano, Switzerland
- Intern at IBM Israel and Intel Ireland
- BsC & MsC in Electrical Engineering @ University of Novi Sad, Serbia

My Work with Machine Learning and Security

Machine Learning for Security

Malware detection [C&S-19, FPS-18, DASC-16, SECRYPT-16, CCNC-16, IWSMA-14]

Anomaly detection [IJCNN-18]

Failure prediction [CINC-14]

Security of Machine Learning

Attacks against ML

Robust defences [FSS-18]

Explainable ML

Application Domains

Mobile systems [C&S-19, FPS-18, HST-17, DASC-16, SECRYPT-16, CCNC-16, IWSMA-14], **Embedded systems** [IJCNN-18, arxiv-18, CINC-14,SECRYPT-14,SECRYPT-13], **Communication Networks** [WIP]

Other Domains where I used Machine Learning

- Manufacturing
  - Predictive maintenance
  - Predictive modeling
  - Production optimization
- Health
  - ECG signals analysis
- Computer vision
  - Face emotion detection

Talk Outline

- Intro on AI
- AI in Security
- Advantages and disadvantages of AI in Security
- Best practices to avoid common pitfalls
- Takeaways

Intro to AI and Machine Learning

- Essentially a revolutionary way of knowledge representations
- Used in almost any domain we can think of, some examples are
  - Predicting protein structure [Deepmind]
  - Probing the cosmos [Wired]
  - Detecting smell from molecules [Google]
- Powerful tool
  - Entered and changed almost every area of our life
  - Already goes above human performance in some specific domains

[Deepmind] https://www.deepmind.com/blog/alphafold-reveals-the-structure-of-the-protein-universe

[Wired] https://www.wired.com/story/deepmind-ai-nuclear-fusion/

[Google] https://ai.googleblog.com/2022/09/digitizing-smell-using-molecular-maps.html

Trust in AI

- AI use widespread, both in business and personal life
- Increased reliance on AI as a part of our daily life
- Trust in AI based on:
  - Transparency
  - Security
  - Fairness
  - Bias
- Each of these aspects needs awareness and further improvements

AI in Security

- Especially challenging given adversarial threats
- Still, many security analysts already rely on machine learning
- Why?
  - To analyze the large amounts of collected data
  - To recognize complex patterns and predict threats in massive datasets, all at machine speed
  - To uncover previously unseen attacks
  - Ability to go beyond signature matching concepts and to early detect potential variants of attacks

AI in Security: Application Scenarios

- Where is AI used in security
    - Malware detection
    - Phishing detection
    - Spam detection
    - Network intrusion detection
    - Vulnerability discovery
    - Abuse detection

AI in Security: What can happen?

- While being very beneficial, at the same time machine learning is:
  - Same as any other software and it has vulnerabilities that can be exploited
  - Shown to have some inherent algorithms weaknesses and blind spots
  - Black box approach

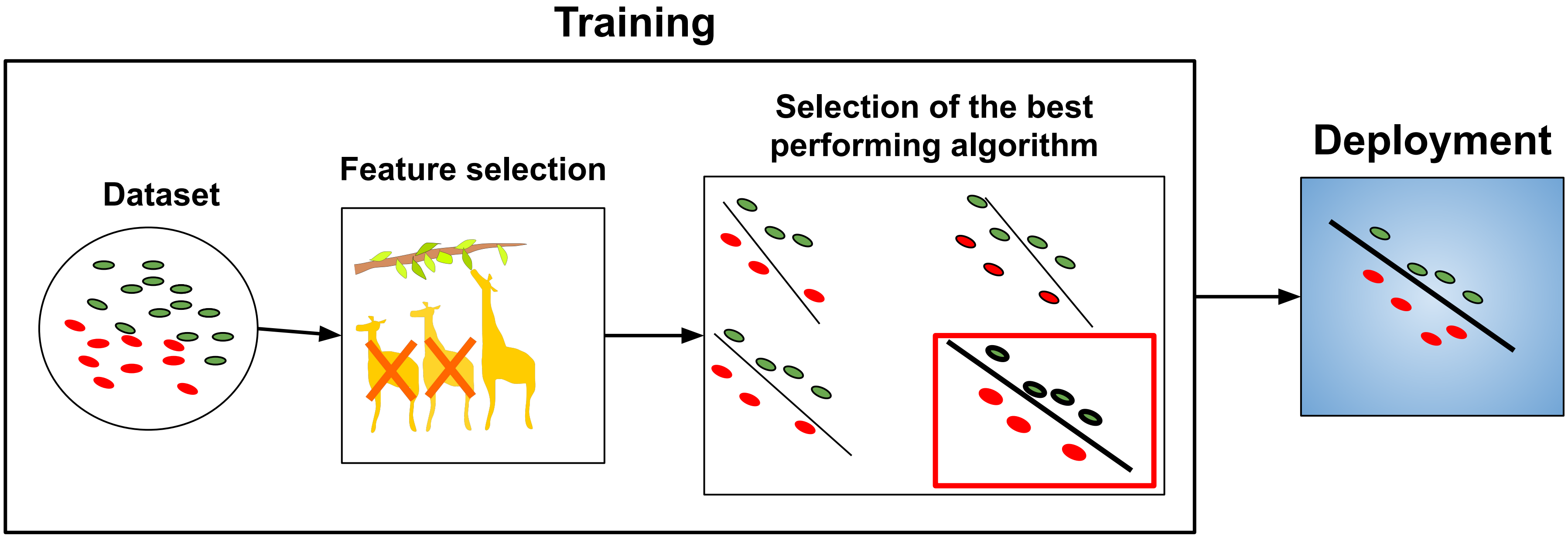
- Two categories of things to have in mind:
  - Attacks against deployed machine learning models
  - Common pitfalls in machine learning systems design leading (among other things) to
    - performance drop
    - non representative results
    - unreliable predictions

AI in Security: What can happen?

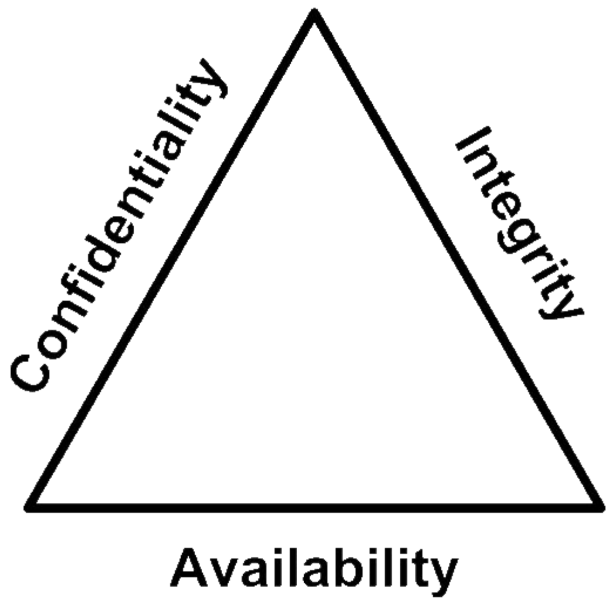
- While being very beneficial, at the same time machine learning is:
  - Same as any other software and it has vulnerabilities that can be exploited
  - Shown to have some inherent algorithms weaknesses and blind spots
  - Black box approach

- Two categories of things to have in mind:
  - Attacks against deployed machine learning models
  - Common pitfalls in machine learning systems design leading (among other things) to
    - performance drop
    - non representative results
    - unreliable predictions

Main Steps of Deployment of Machine Learning



Main Security Aspects: CIA Triad



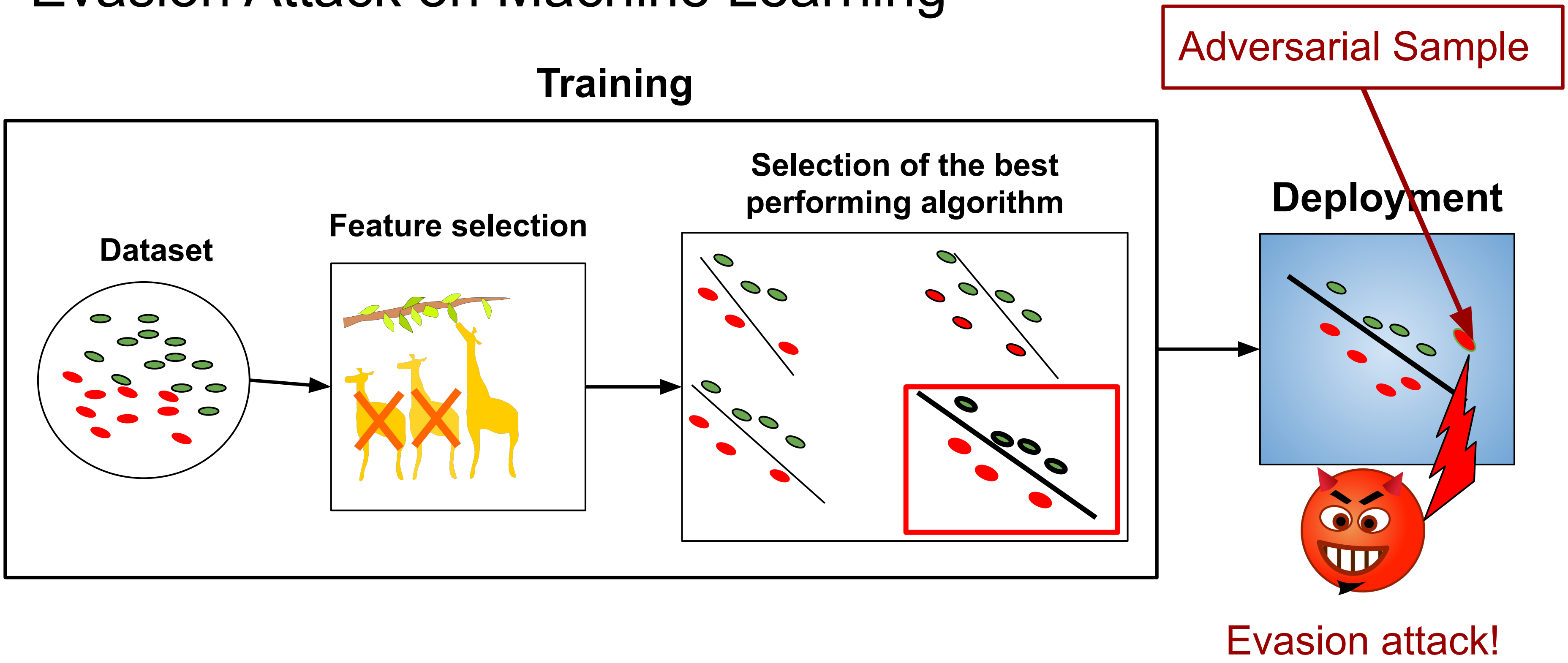
What Can Go Wrong with Machine Learning?

	Confidentiality	**I**ntegrity	**A**vailability
Deployment	Model stealing Model inversion Membership inference attack	Evasion	Increasing false positives

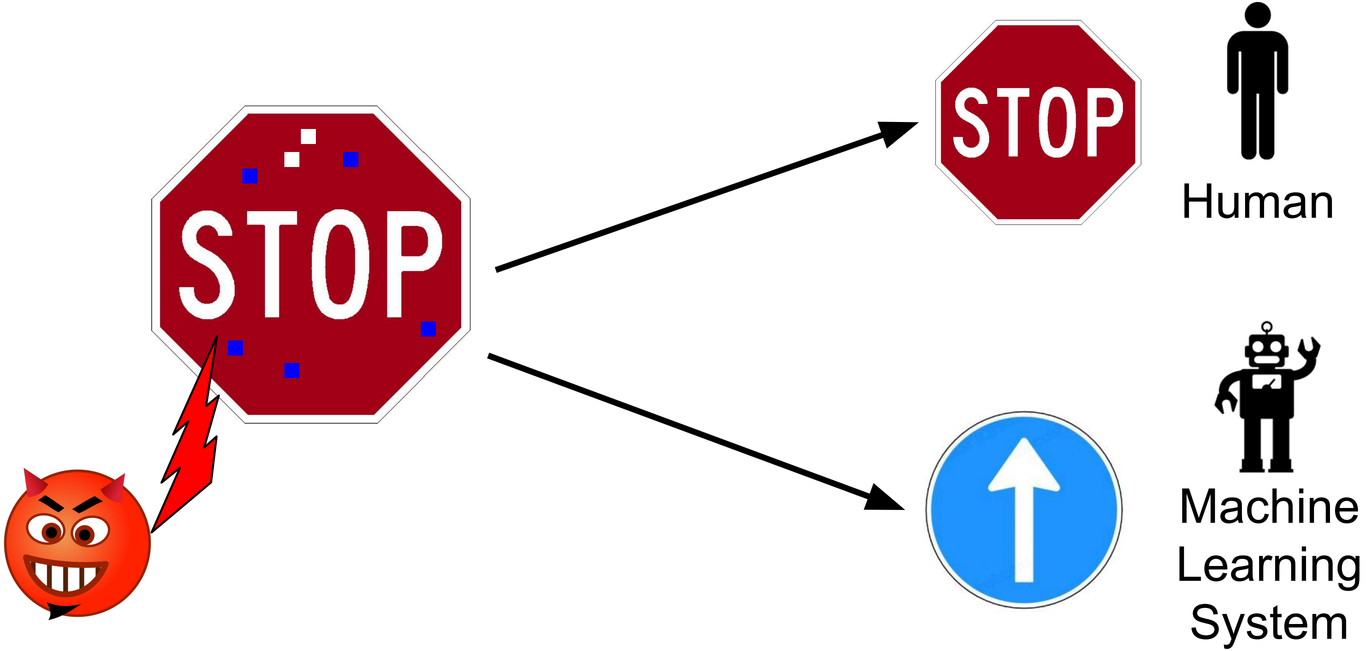
What Can Go Wrong with Machine Learning?

	Confidentiality	**Integrity**	**Availability**
Deployment	Model stealing Model inversion Membership inference attack	Evasion	Increasing false positives

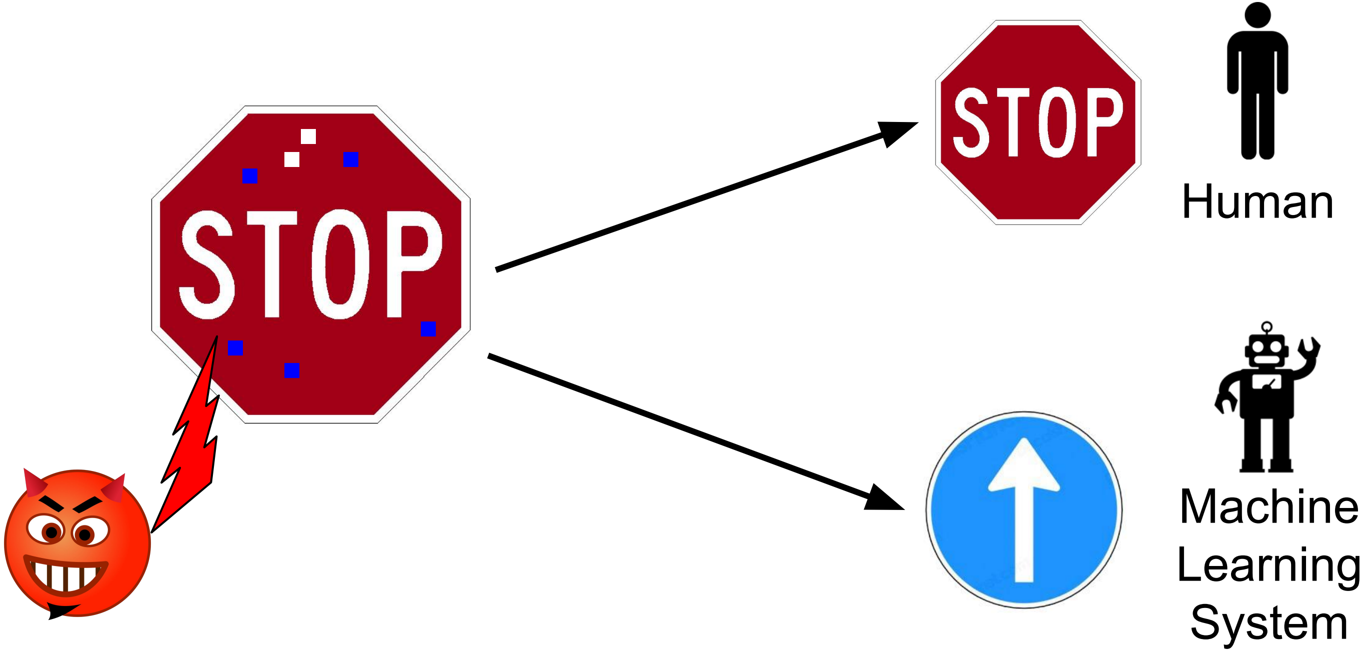
Evasion Attack on Machine Learning



Adversarial Samples in the Physical World



Adversarial Samples in the Physical World



Generation of adversarial samples can be **automated**! [Goodfellow et al., Papernot et al.]

Evasion attacks **transfer** between different machine learning techniques!
[Szegedy et al., Papernot et al.]

Automation of Evasion Attacks

Attacks	**Pros**	**Cons**
Fast Gradient Sign Method (FGSM) [Goodfellow et al.]	Fastest speed Low computation cost	Large perturbations
Jacobian Saliency Map Attack (JSMA) [Papernot et al.]	Small perturbations	High computational costs
Carlini Wagner (CW) [Carlini et al.]	Minimum perturbations	Slowest speed, high computational cost

Goodfellow et al., *Explaining and Harnessing Adversarial Samples* https://arxiv.org/abs/1412.6572
Papernot et al., *The Limitations of Deep Learning in Adversarial Settings* https://arxiv.org/pdf/1511.07528.pdf
Carlini et al., *Towards Evaluating the Robustness of Neural Networks* https://arxiv.org/pdf/1608.04644.pdf

Libraries to Generate Adversarial Samples

- Cleverhans
  - https://github.com/tensorflow/cleverhans
- SecML
  - https://gitlab.com/secml/secml
- IBM Adversarial Robustness Toolbox (ART)
  - https://developer.ibm.com/open/projects/adversarial-robustness-toolbox/
- Foolbox
  - https://github.com/bethgelab/foolbox

Some Defenses Against Evasion Attacks (Open problem)

- Ensemble
- Adversarial retraining
- Defensive distillation
- Dimensionality reduction
- Regularization

Some Defenses Against Evasion Attacks (Open problem)

- Ensemble
- Adversarial retraining
- Defensive distillation
- Dimensionality reduction
- Regularization

- At the beginning this topic mostly touched image analysis domain, but with time it was shown that it is applicable also to security use cases like malware detection.
- Recent paper [Grosse et al] showed that evasion and poisoning are becoming treats also in industry.

Grosse et all, "Why do so? A practical perspective on Machine learning Security" https://arxiv.org/pdf/2207.05164.pdf

AI in Security: What can happen?

- While being very beneficial, at the same time machine learning is:
  - Same as any other software and it has vulnerabilities that can be exploited
  - Shown to have some inherent algorithms weaknesses and blind spots
  - Black box approach

- Two categories of things to have in mind:
  - Attacks against deployed machine learning models
  - Common pitfalls in machine learning systems design leading (among other things) to
    - performance drop
    - non representative results
    - unreliable predictions

Pitfalls & Best Practices: Data Collection and Labelling

- Bias in data, sampling bias
  - Do take the time to understand your data
  - Gather as representative data set as possible
- Spurious correlations, biased parameters
  - Do talk to domain experts
- Label inaccuracy, poor performance
  - Ensure labels correctness

Pitfalls & Best Practices: System Design and Learning

- Unrepresentative model - either too complex or too weak
  - Do survey the literature and understand what is the baseline
  - Not every problem needs to be solved with deep learning
- Too complex model, difficult to maintain
  - Only select features that matter
  - Start with a simple model first
- Weak performance model
  - Consider ensemble of models
- Spurious correlations, biased parameters
  - Use explainability techniques in order to understand what is the algorithm learning

Pitfalls & Best Practices: Performance Evaluation

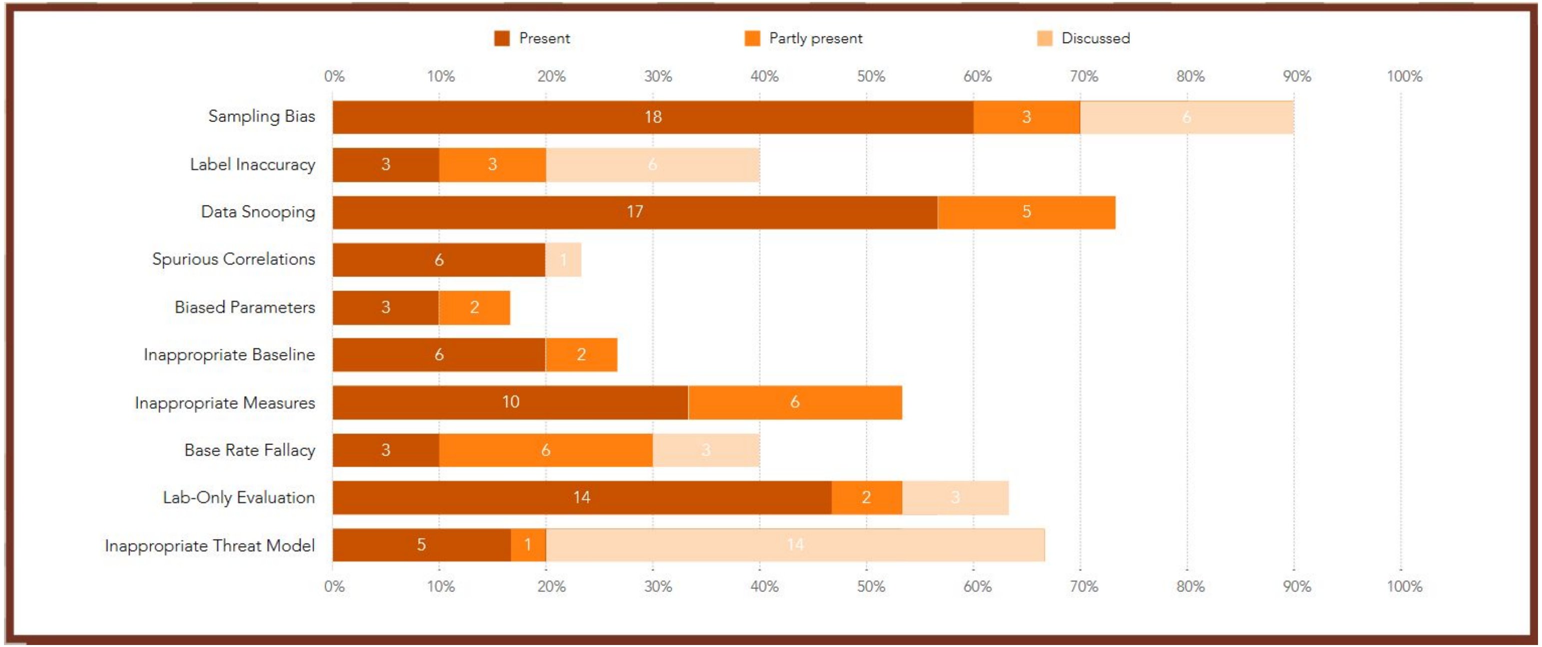
- Low usefulness of deployed model
  - Measure helpfulness, not mathematical accuracy
- Non reproducible results
  - Be transparent with your model
    - Publish papers and open source or discuss within the team and seek for feedback of people
  - Use suitable tools to track experiments
    - MLflow, Weights and biases, DVC
- Overestimated performance
  - Separate training and testing data and never look into testing until you have a final method
  - Mostly unbalanced datasets, metrics need to be selected well, use multiple ones
    - not just accuracy, but also use PR curve, F-1 measure

Pitfalls & Best Practices: Deployment and Operations

- Inappropriate threat model
  - Have in mind deployment scenario from the beginning of the system design
    - E.g., is the system exposed to external users, is it to be run on device or in cloud
- Costly maintenance
  - Have in mind tech depth of maintaining machine learning models in practice [Sculley et al]

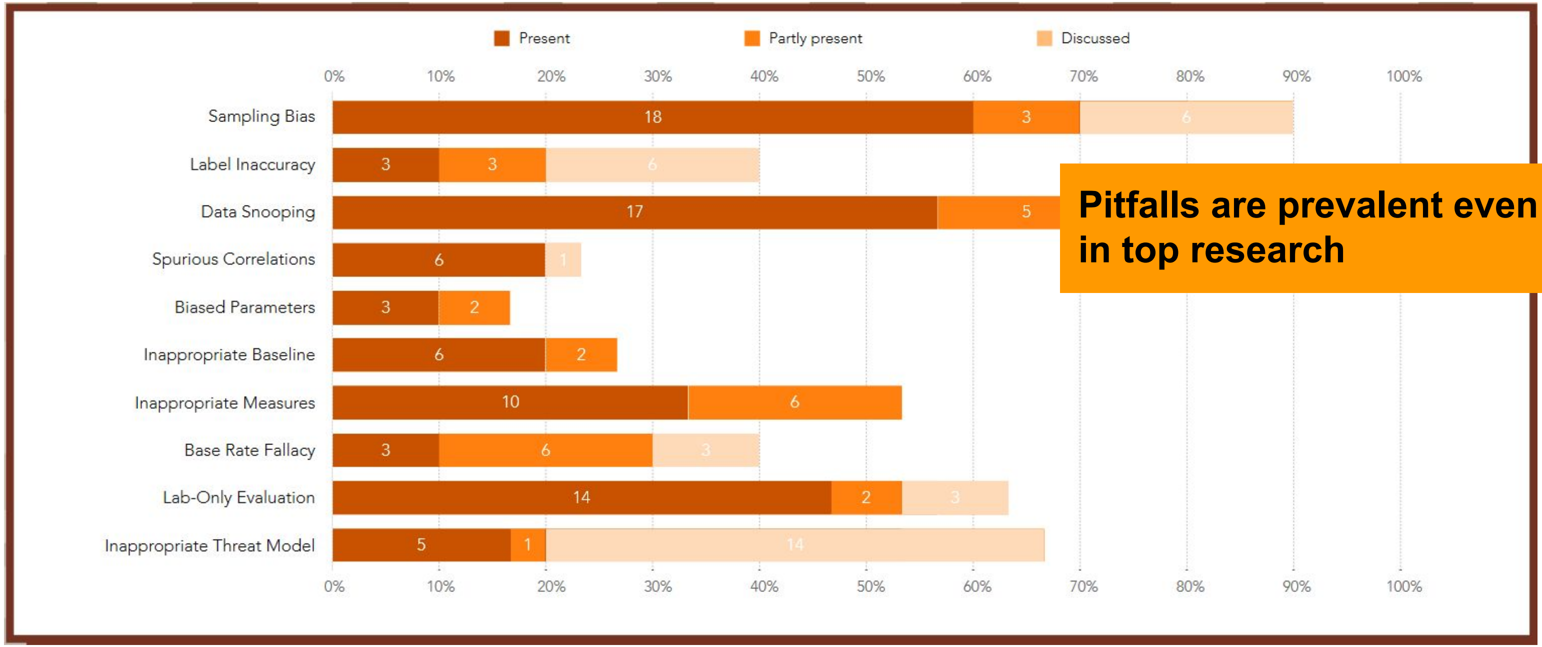
Sculley et al, "Hidden Technical Depth in Machine Learning Systems" https://papers.nips.cc/paper/2015/file/86df7dcfd896fcaf2674f757a2463eba-Paper.pdf

Common Pitfalls Found in Research Papers [Arp et al.]



Arp et al, Do's and Don'ts of Machine Learning in Computer Security
https://www.usenix.org/conference/usenixsecurity22/presentation/arp

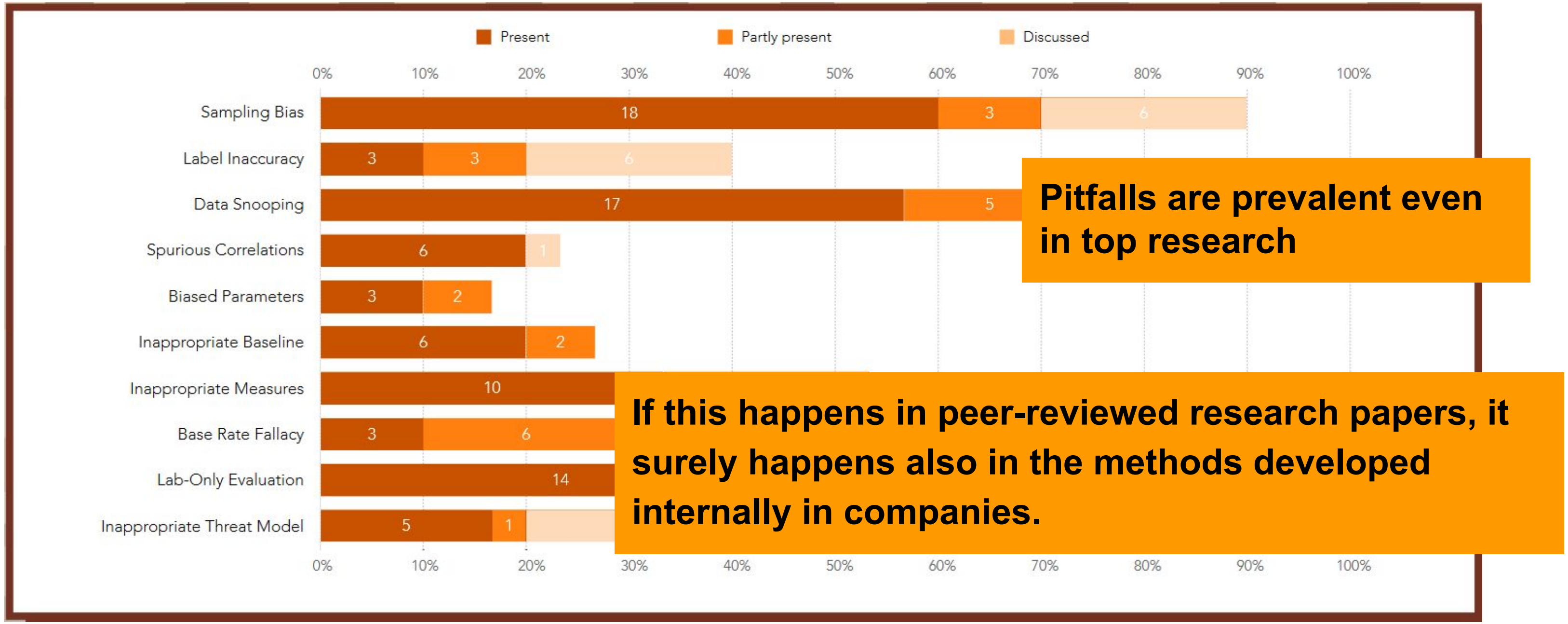
Common Pitfalls Found in Research Papers [Arp et al.]



Pitfalls are prevalent even in top research

Arp et al, Do's and Don'ts of Machine Learning in Computer Security
https://www.usenix.org/conference/usenixsecurity22/presentation/arp

Common Pitfalls Found in Research Papers [Arp et al.]



Arp et al, Do's and Don'ts of Machine Learning in Computer Security
https://www.usenix.org/conference/usenixsecurity22/presentation/arp

Conclusion and Takeaways

- AI in security is very promising and highly beneficial
- Its usage comes with some inherent problems and weaknesses that we should know about
- To successfully develop ML-based solutions we need to constantly learn about best practices and include them in the design process

Questions?