

“Security is not my field, I’m a stats guy”: A Qualitative Root Cause Analysis of Barriers to Adversarial Machine Learning Defenses in Industry

Jaron Mink*, Harjot Kaur*, Juliane Schmäuser*,
Sascha Fahl, Yasemin Acar



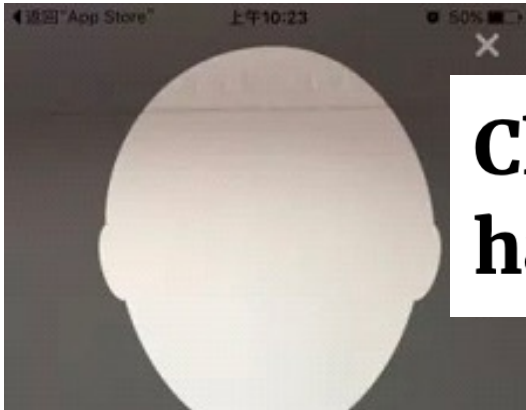
* Equal Contribution

Adversarial ML (AML) Threatens Real Systems



Chinese government-run facial recognition system hacked by tax fraudsters: report

Adversarial ML (AML) Threatens Real Systems



Chinese government-run facial recognition system hacked by tax fraudsters: report



Christiaan Beek ✓ @ChristiaanBeek · Oct 18, 2020

Maybe someone from [@virustotal](#) can have a look at this; feels like the submitter is poisoning VT results...



Christiaan Beek ✓ @ChristiaanBeek · Dec 8, 2020

A recent case discovered on VT poisoning added to the [@MITREattack](#) Adversarial Threat matrix:

Adversarial ML (AML) Threatens Real Systems



Invisible for both Camera and LiDAR: Security of Multi-Sensor Fusion based Perception in Autonomous Driving Under Physical-World Attacks



Accessorize to a Crime: Real and Stealthy Attacks on State-of-the Art Face Recognition



CVE-2019-20634 Detail



kaspersky

How to confuse antimalware neural networks. Adversarial attacks and protection



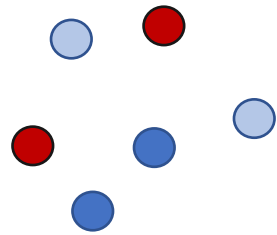
Imitation Attacks and Defenses for Black-box Machine Translation Systems



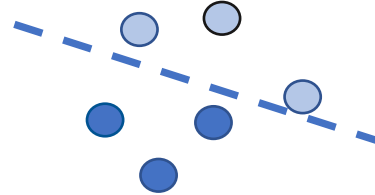
MITRE | ATLAS™

Attacking The ML Pipeline

Poisoned Data

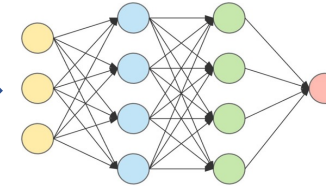
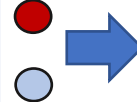


Data Collection



Data Modeling

Adversarial Example

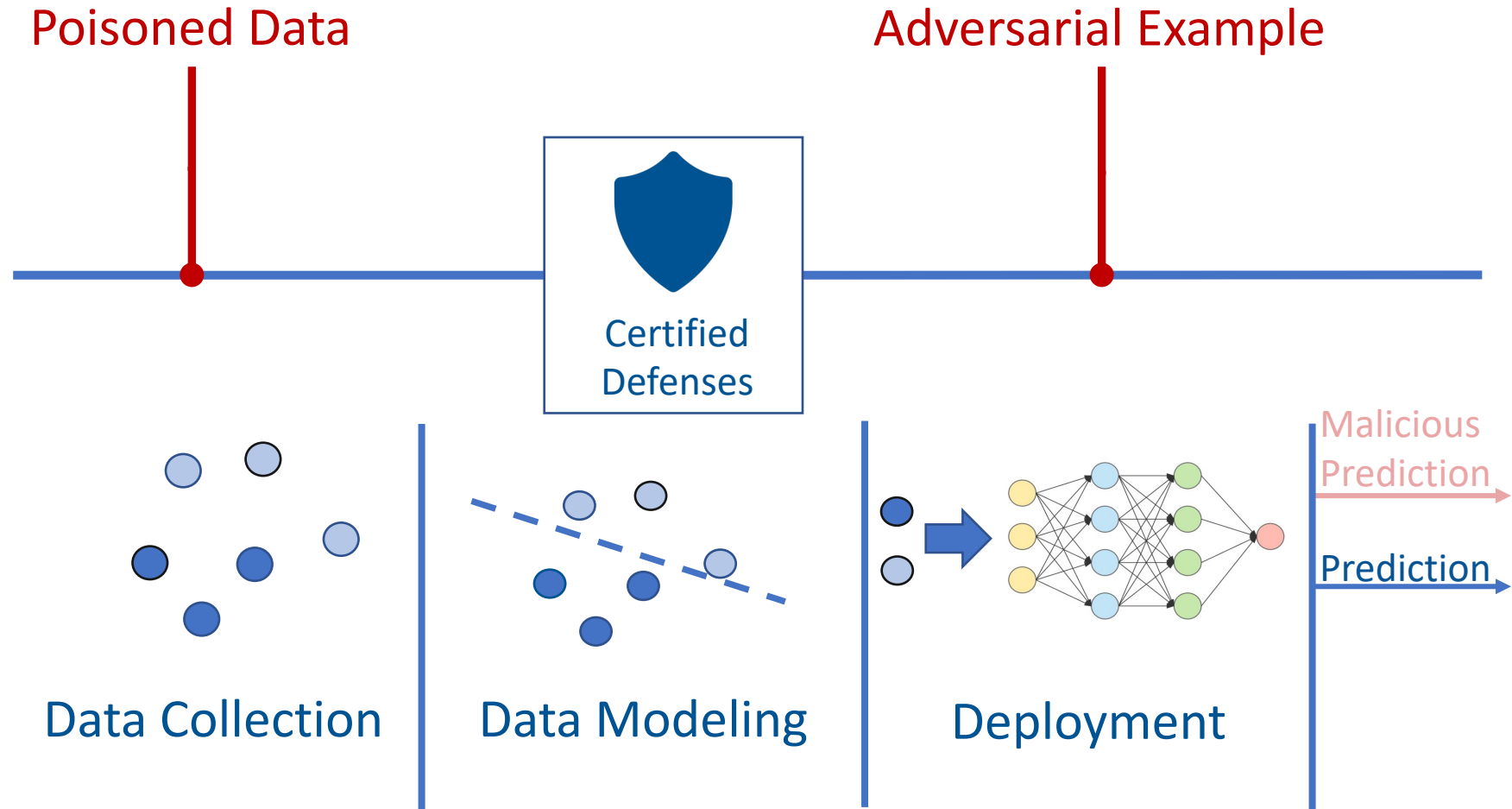


Deployment

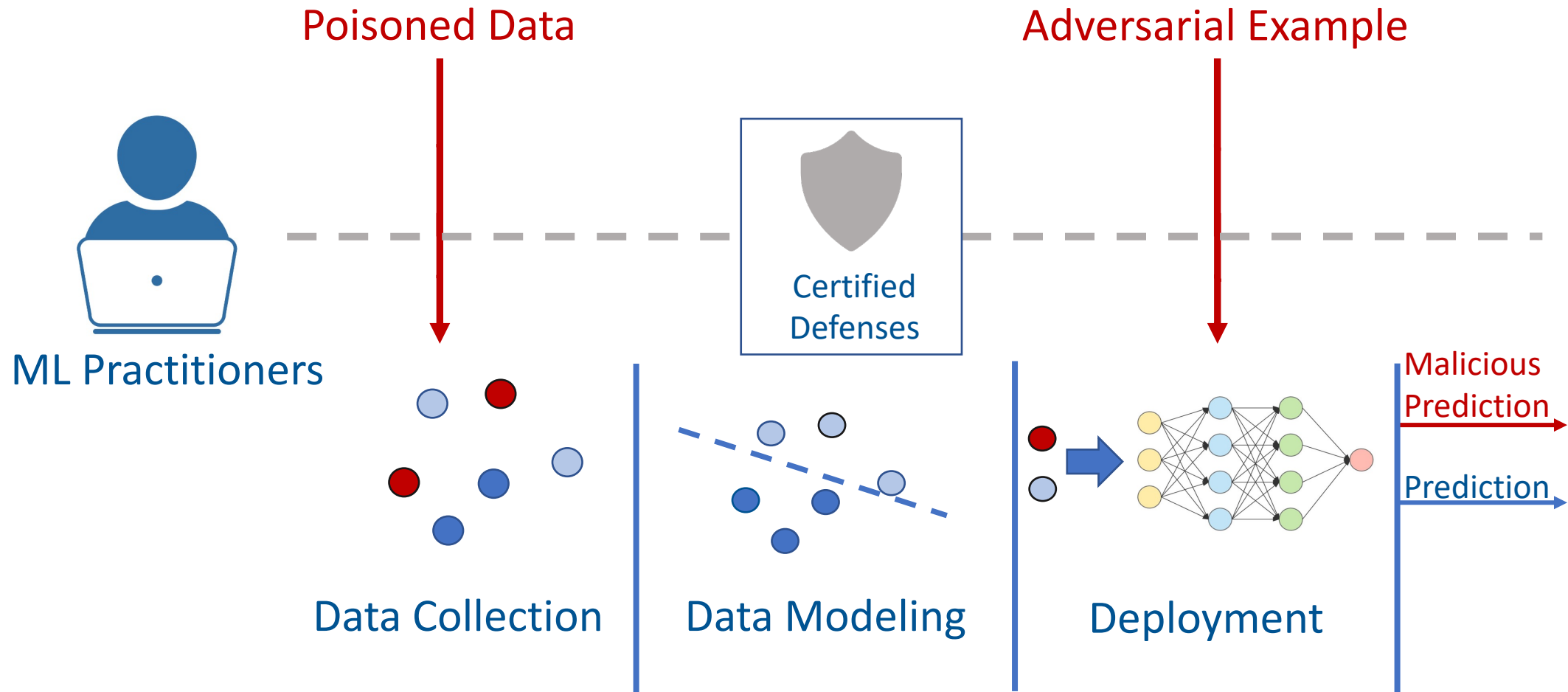
Malicious Prediction

Prediction

Defenses Exist...



Defenses Exist... But Are Not Implemented! ^[1,2]



[1] "Adversarial machine learning-industry perspectives." Kumar et. al, IEEE S&P Workshop, 2020.

[2] "Industrial practitioners' mental models of adversarial machine learning." Krombholz et. al, SOUPS, 2022.

Barriers To Defense Are Not Well Understood



Practitioners **lack knowledge** of AML threats and defenses [1,2]
But what barriers prevent learning?



Practitioners **don't assess AML as a threat** to their system [1,2,3]
But can they properly assess their systems?



Practitioners **don't implement defenses** due to a lack of guidance and responsibility [1, 2]
But how do organizations affect this?

[1] "Adversarial machine learning-industry perspectives." Kumar et. al, IEEE S&P Workshop, 2020.

[2] "Industrial practitioners' mental models of adversarial machine learning." Krombholz et. al, SOUPS, 2022.

[3] "Machine Learning Security in Industry: A Quantitative Survey." Grosse et. al, arXiv, 2023.

Research Questions

1. What barriers prevent ML practitioners from adequately **understanding AML attacks**, and their corresponding risks and defenses?
2. What barriers prevent ML practitioners from adequately **assessing the risk AML poses** to their systems?
3. What barriers prevent ML practitioners from effectively **implementing AML defenses** in their systems?

Methodology

Recruited 21 ML practitioners

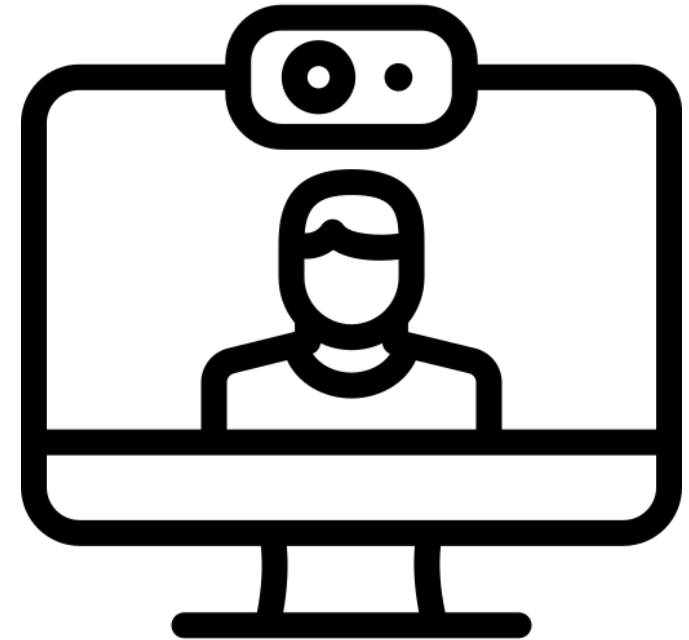
- 1+ year of industry experience
- Data engineers, data scientists, ML engineers

90-minute interview

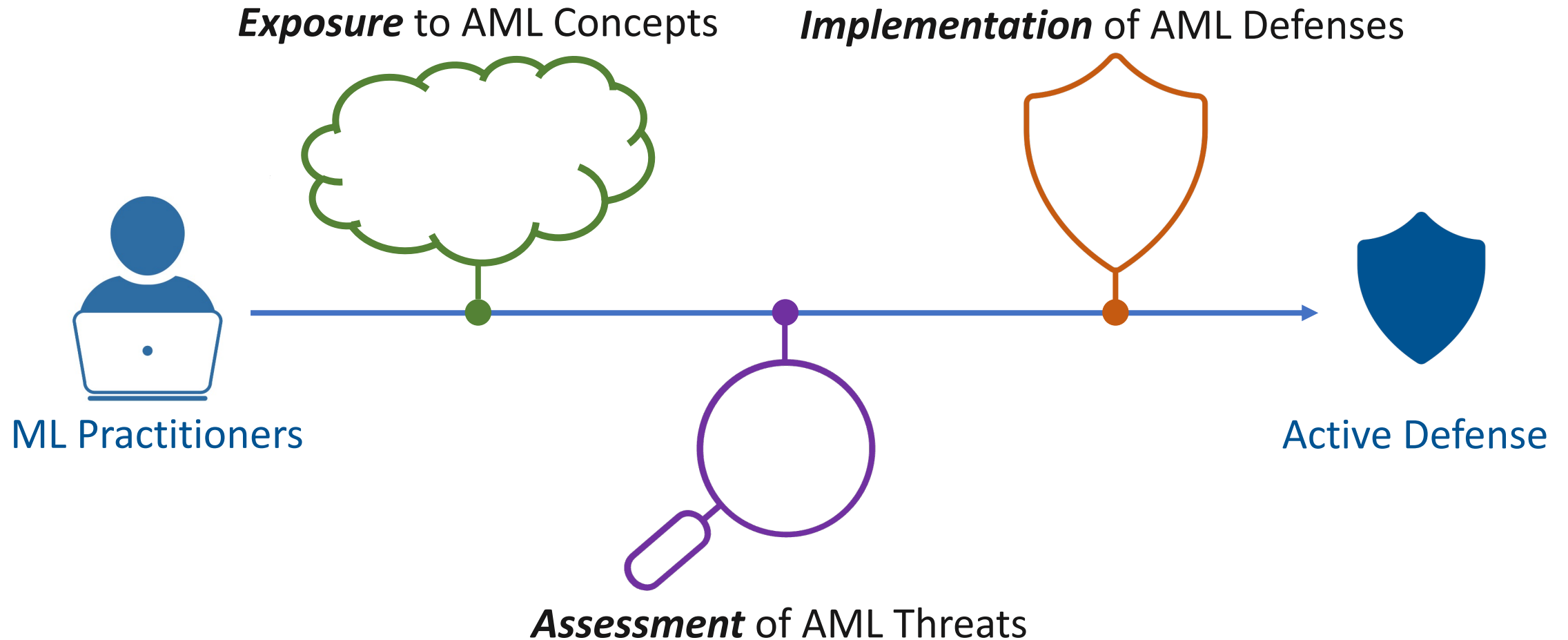
1. Educational resources & motivations
2. ML pipeline
3. Perception of AML
4. Responses to AML

Thematic analysis

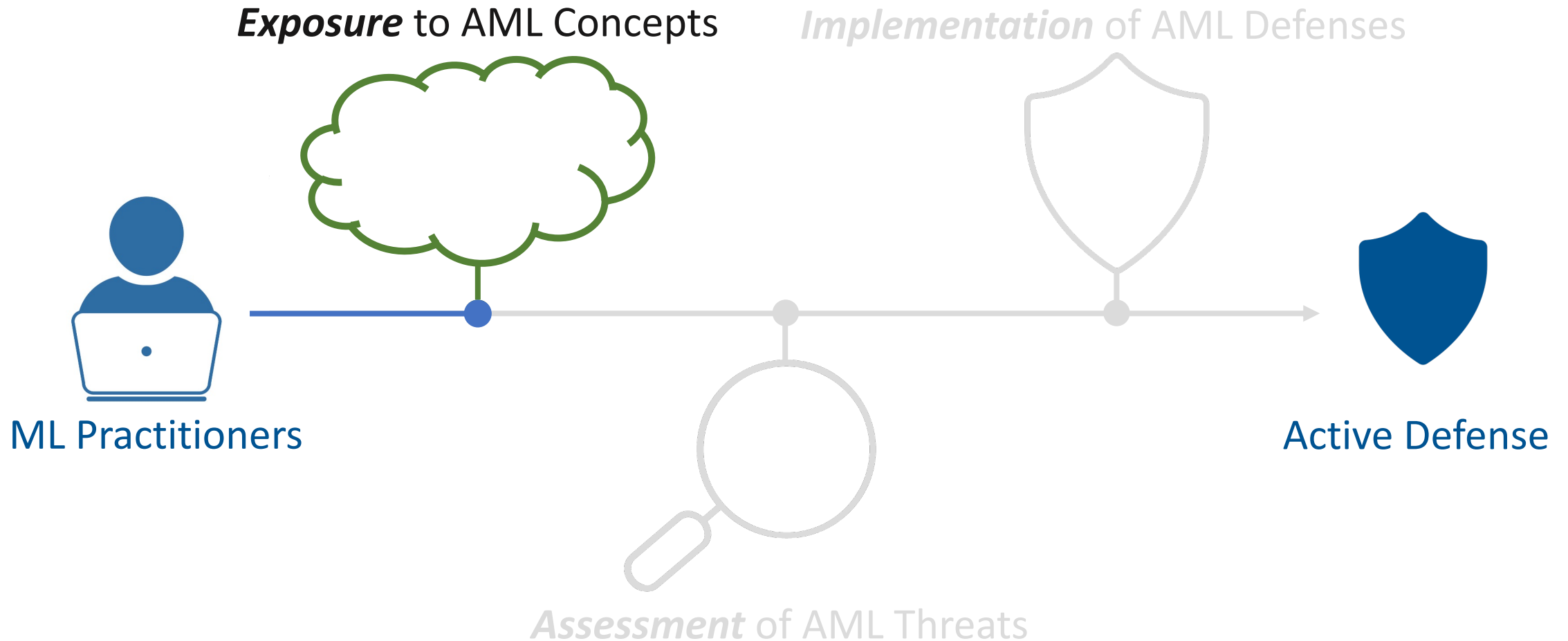
- Defensive barriers ML practitioners face



Barriers to AML Defenses



Culture/Education Structure Hinder Exposure

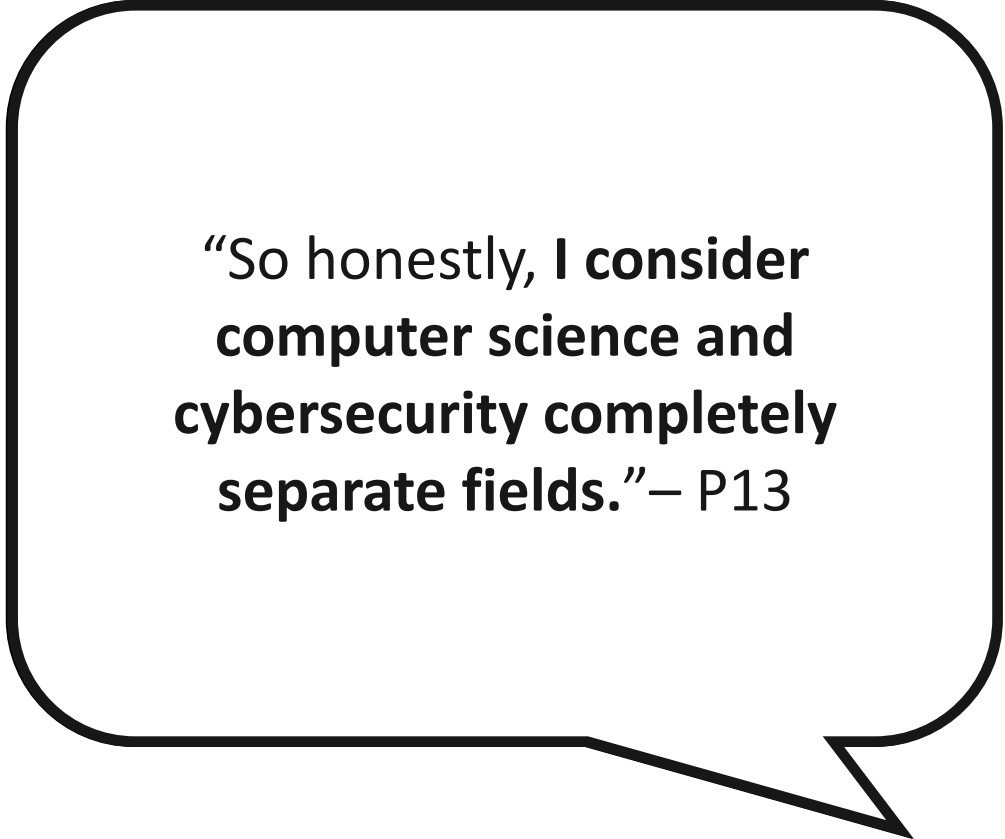


Practitioners Assume S&P Is Irrelevant To ML

Data privacy understood, but **not** model security and privacy (S&P)

A few believed S&P is separate from ML

S&P implications affected interest



“So honestly, I consider computer science and cybersecurity completely separate fields.” – P13

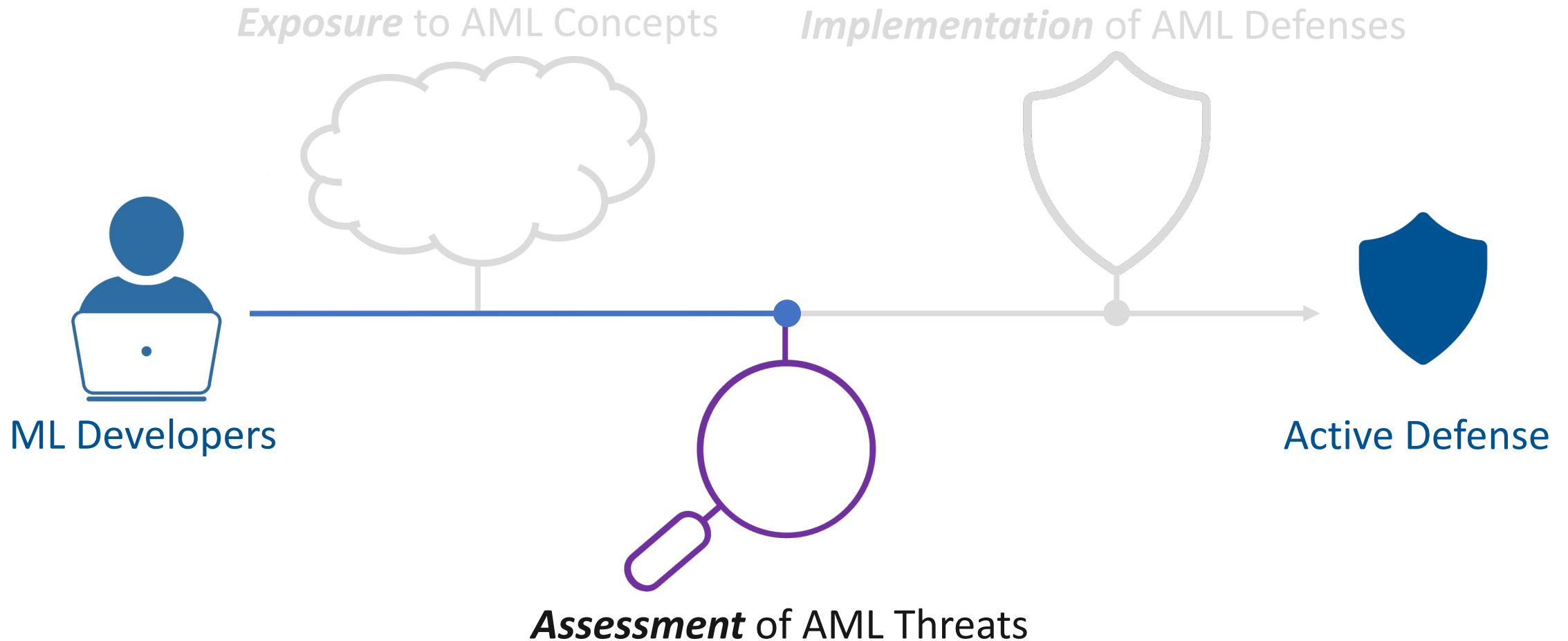
AML Learning Is Not Institutionally Motivated

New knowledge is guided by:

- ⚙️ Project requirements
 - AML not considered; functionality-centered
 - Assigned by non-technical persons
- 🎓 Educational degree requirements
 - AML only reported in CS, grad courses

“A lot of the private data providers, their sales and their engineers have to be told like, ‘Hey, **[AML] is something that occurs.**’ They’re like ‘oh yeah, I read a paper on **that once. I heard that could occur.**’ And it’s just like, ‘**No, this is real.**’ – P07

Misconceptions/Lack Of Tools Hinder Assessment



Misconceptions Lead To Unseen Vulnerabilities

- ❓ Threat models and defense perceptions were incorrect
→ Vulnerabilities may remain unseen

“I see maybe cases where **[adversarial examples]** can happen, but it can only happen on the **backend of our company**. Like if someone really managed to get so deep inside backend that they can also modify our data... **For the current state, it shouldn't be very problematic.**” – P03

Evaluation & Monitoring Do Not Consider AML

 Nearly all participants did not assess or monitor AML threats

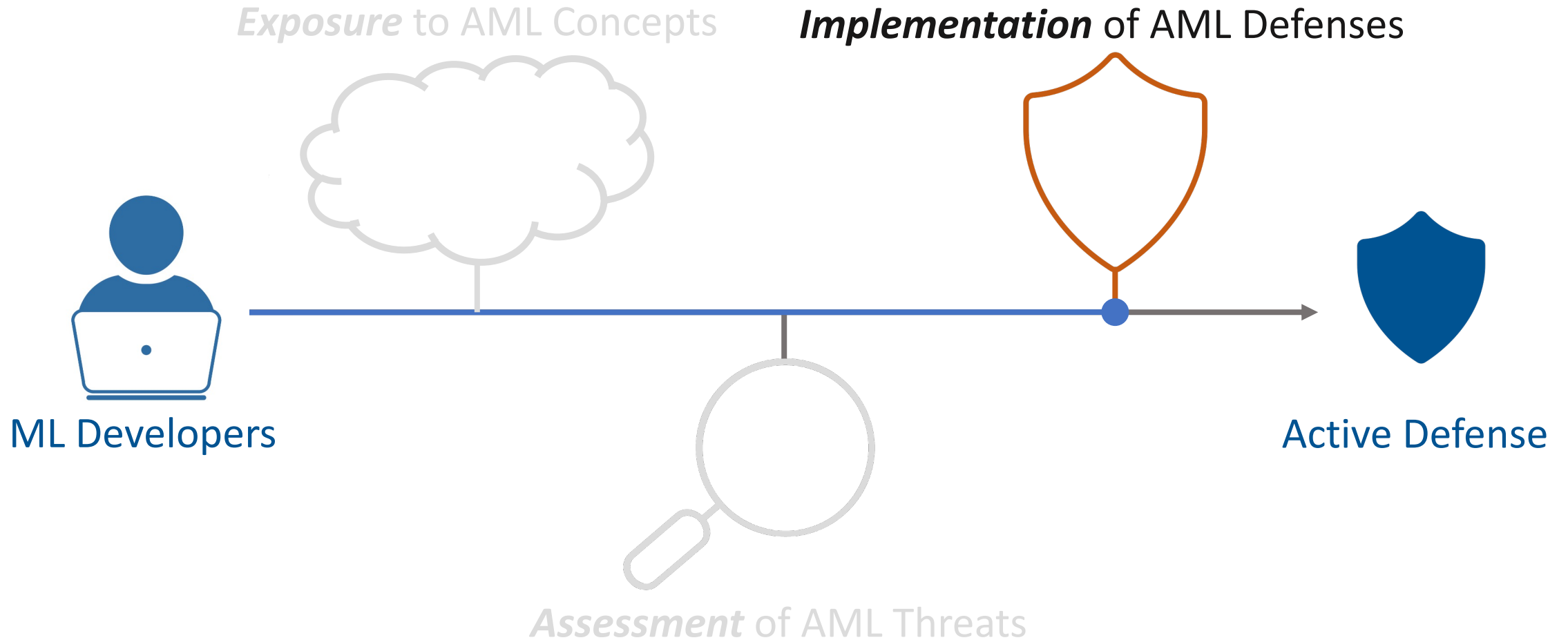
→ F1-scores, prediction accuracy

 Edge case testing was reported, but not comprehensive

→ Manual, ad-hoc adversarial testing

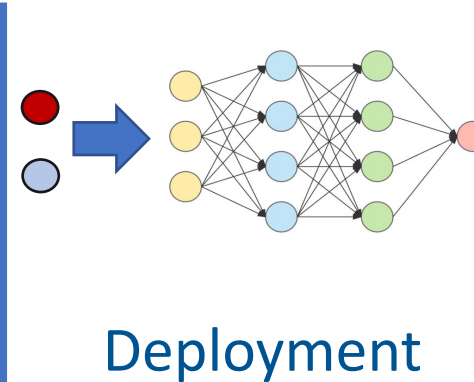
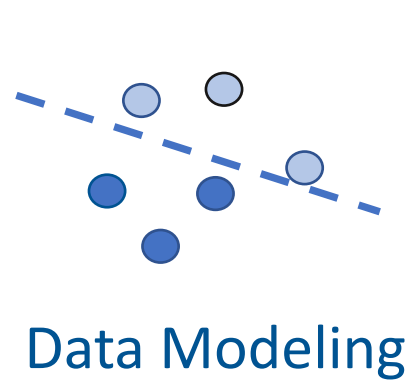
“Let’s say five of us are working together, I build some model, and I say, ‘**Give the model any input you can**’. And basically, I’ll say, ‘I’ll **give you a treat if you can break the model.**’ – P05

Org. Structure And Values Impede Defenses



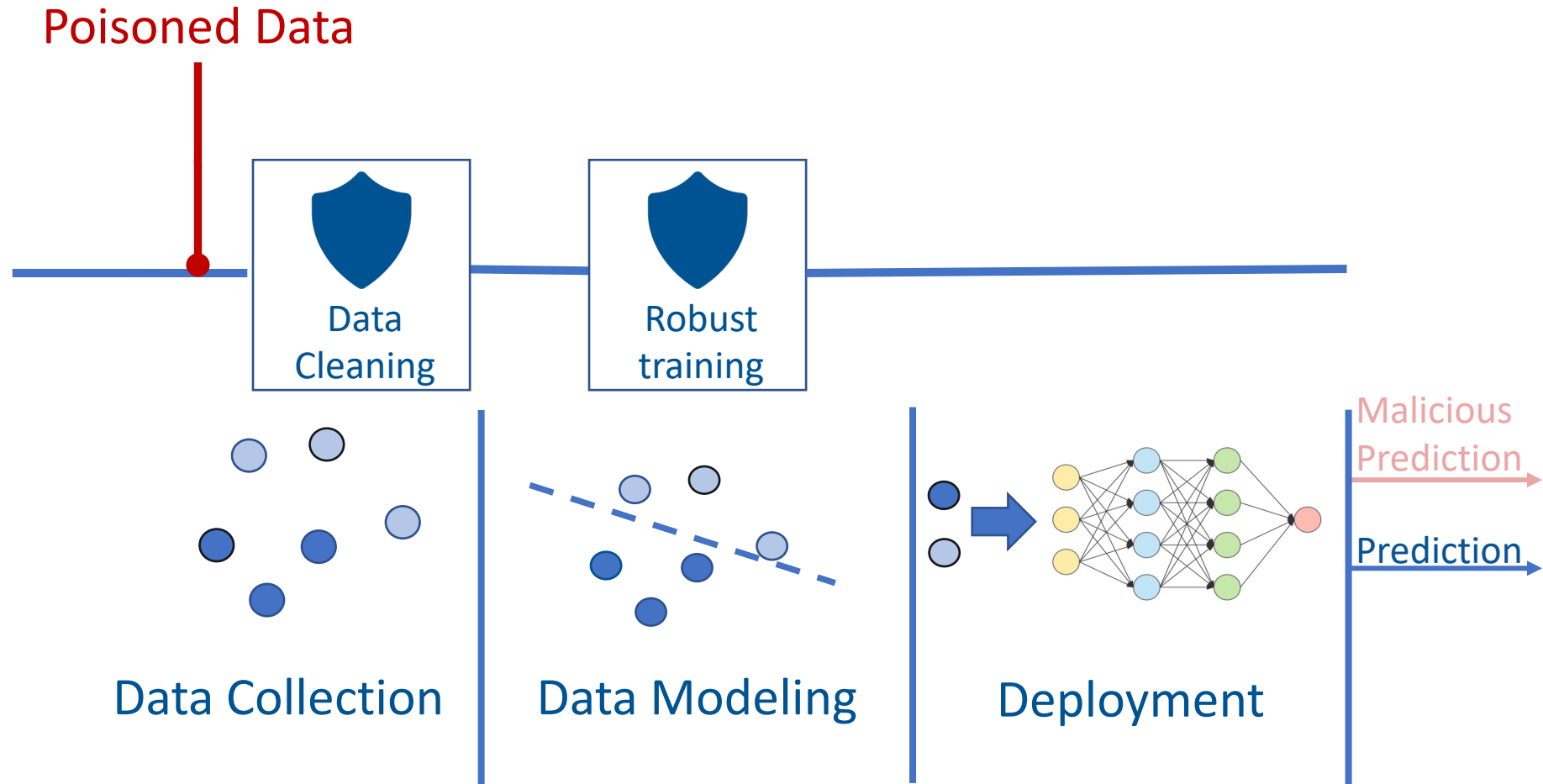
AML Attacks Span The ML Pipeline

Poisoned Data



Malicious Prediction
Prediction

AML Defenses May Require Collaboration



Team Isolation Prevents Collaboration



Teams were isolated along the ML pipeline

→ May result in difficulty defending & assigning responsibility

“We wouldn’t know too much about, what they were going to be developing in terms of like model requirements.”
– P13

“In most of the cases I’m not engaged in gathering data.”
– P14



Data Collection

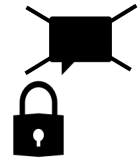


Data Modeling



Deployment

Team Isolation Prevents AML Collaboration



Security teams were further removed

→ ML pipeline might not benefit from security team's knowledge

“[The security team is] relatively isolated. We have never worked on a project together, so I would say we live in two different sets of worlds in terms of our focus.” – P08





ML Pipeline



Security Team

Defenses Compete And Lose To Other Goals

Organizations prefer:

-  To spend resources on functionality
 - Developers are already time-constraints
 - AML experts perceived as too expensive
-  Performance over defenses
 - Some would accept a performance-security trade-off
 - Some would not accept any decrease

“[The bank doesn’t] want to lose money... The accuracy is very important for them. So the security guys need to adopt, not the modelers.” - P11

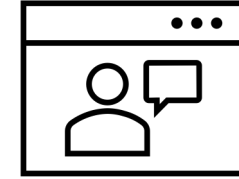
Recommendation: Promote Defensive Agility

Educators

Promote
AML Awareness



AML coverage in courses



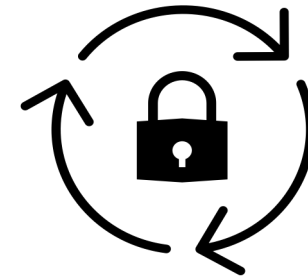
Improve AML resources for non-experts

Organizations

Establish an
S&P Culture in ML



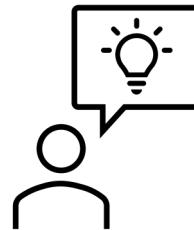
ML Security Champions



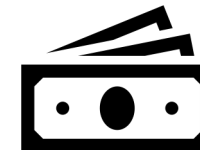
Secure Practices as a Process

Defense Experts

Provide Accessible
Monitoring & Assessment



Increase awareness of tools



Accommodate business constraints

“Security is not my field, I’m a stats guy”

What prevents AML exposure?

- Missing mandates/incentives to learn
- *Resources are inapplicable or unavailable*
- Assumed disconnect of ML and security
- *Lack of interaction with AML-knowledgable colleagues*

What prevents assessment of AML threats?

- Incorrect threat models
- Missing evaluation and monitoring

What prevents implementation of AML defenses?

- Missing collaboration
- *Responsibility is undefined*
- Conflict with other goals
- *Difficulties in finding applicable defenses*

Jaron Mink*, Harjot Kaur*, Juliane Schmüser*,
Sascha Fahl, Yasemin Acar

Check out our paper!



<https://jaronm.ink>

<https://cispa.de/en/people/c02jusc>