

# ***Mis-classification Testing in Open Source Supervised Learning Projects***



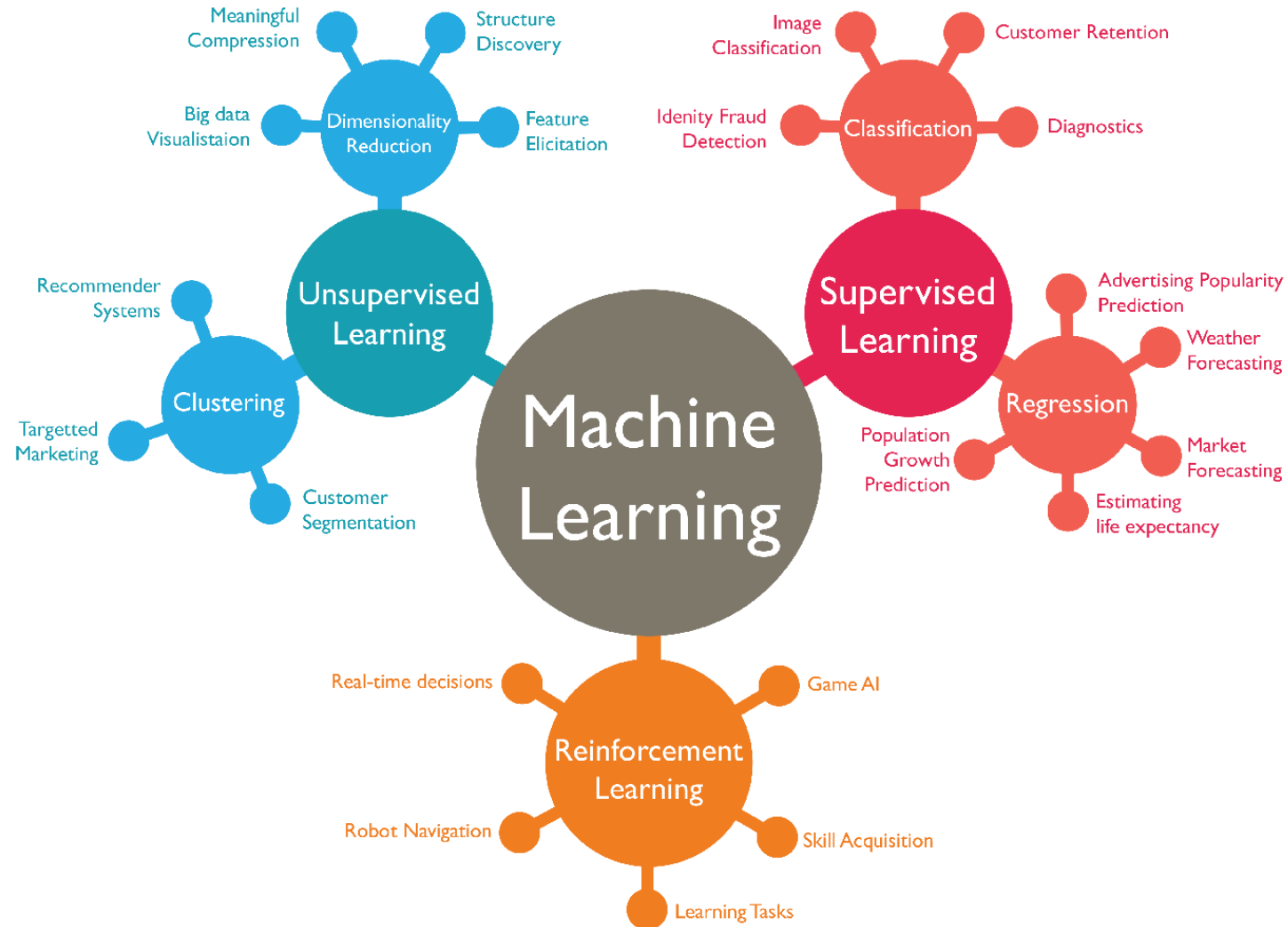
**Farzana Ahamed Bhuiyan, PhD**



**Akond Rahman, PhD**



# Machine Learning



# Machine Learning Project

## The Product Beyond the Model – An Empirical Study of Repositories of Open-Source ML Products

Nadia Nahar<sup>\*†</sup>, Haoran Zhang<sup>†</sup>, Grace Lewis<sup>‡</sup>, Shurui Zhou<sup>§</sup>, Christian Kästner<sup>†</sup>

<sup>†</sup>Carnegie Mellon University, <sup>‡</sup>Carnegie Mellon Software Engineering Institute, <sup>§</sup>University of Toronto

\*nadian@andrew.cmu.edu

***A machine learning project is a software project (a) for end-users that (b) contains one or more machine-learning components.***



AUBURN

# Supervised Learning Project

***A supervised learning project is a software project (a) for end-users that (b) contains one or more components that use supervised learning algorithms.***



# Misclassification Attacks Against ML Projects

## Poisoning Attacks Against Machine Learning: Can Machine Learning Be Trustworthy?

**Alina Oprea**, Northeastern University

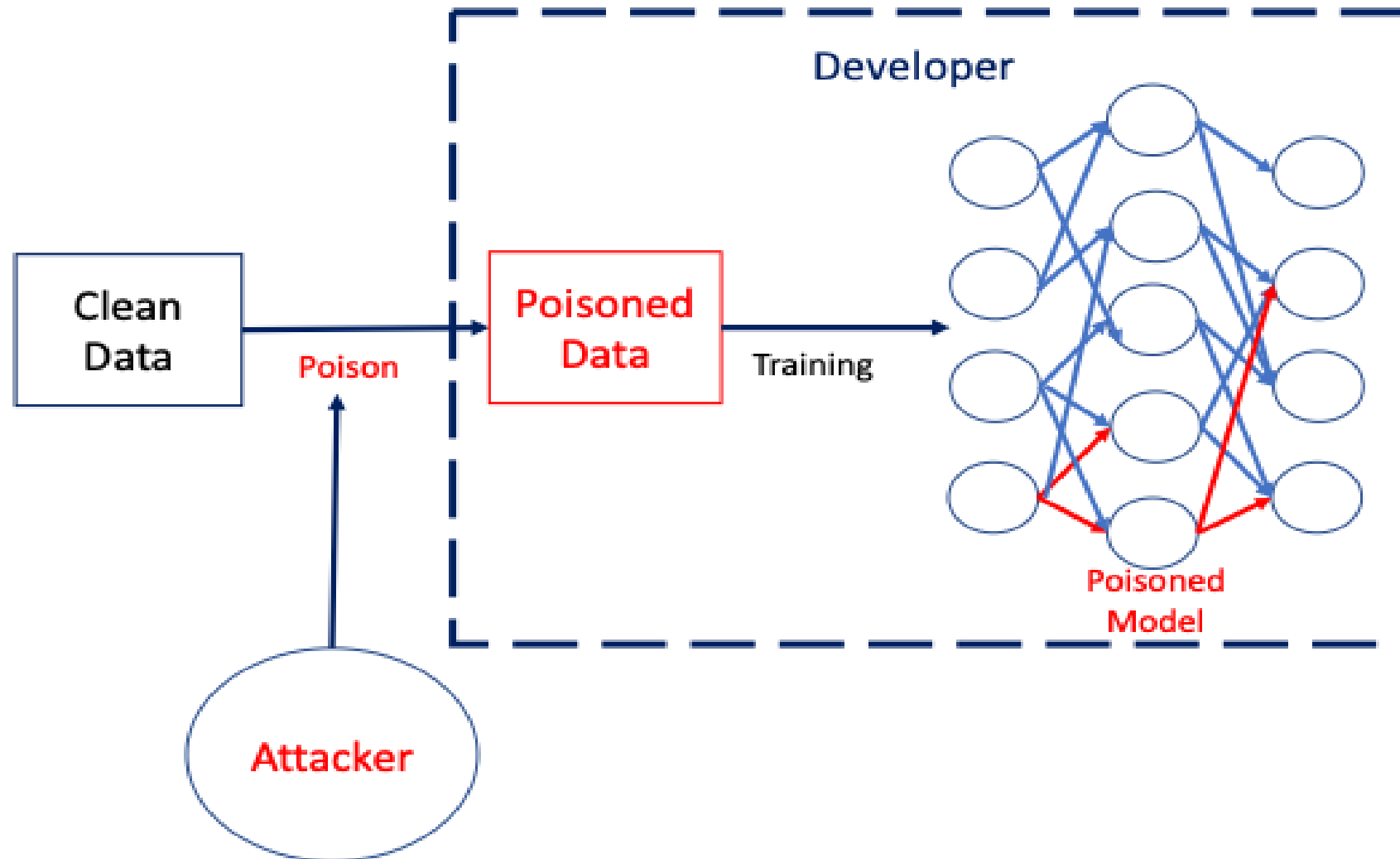
**Anoop Singhal and Apostol Vassilev**<sup>ID</sup>, National Institute of  
Standards and Technology

# Goal

***To help practitioners in testing for adversarial attacks by automatically generating test cases to detect mis-classification for supervised learning-based projects***



# Threat Model



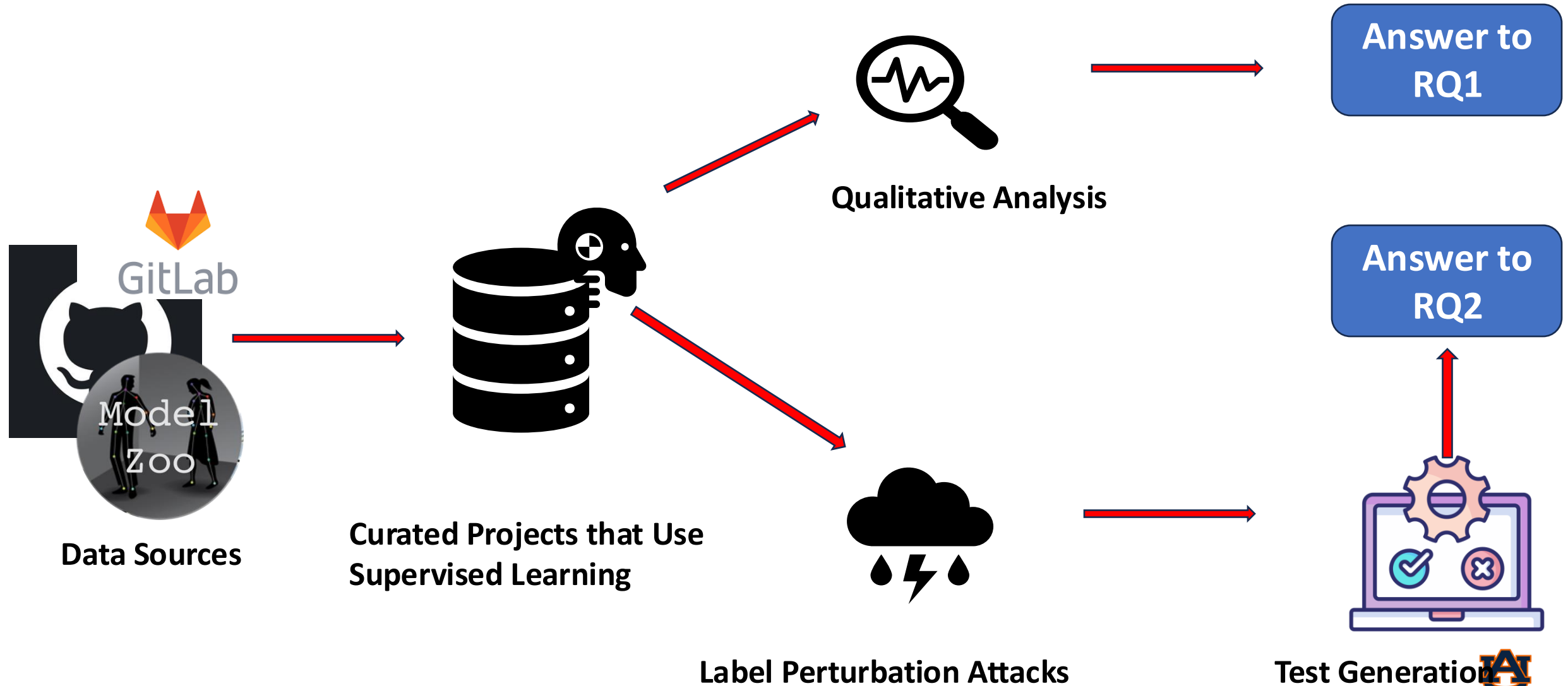
# Research Questions

- *RQ1: What do developers test for in supervised learning-based projects?*
- *RQ2: How can we automatically generate tests to detect mis-classification in supervised learning-based projects?*





# Methodology




# Methodology: Label Perturbation Attack

[Home](#) > [ECML PKDD 2018 Workshops](#) > Conference paper

## Label Sanitization Against Label Flipping Poisoning Attacks

Conference paper | First Online: 16 February 2019

pp 5–15 | [Cite this conference paper](#)

[Andrea Paudice](#), [Luis Muñoz-González](#)  & [Emil C. Lupu](#)

*In this label perturbation attack approach, the attacker's goal is to find a subset of examples in such that when their labels are flipped, a loss function working as an objective function for the attacker is maximized*

# Methodology: Baseline

---

[Journals & Magazines](#) > [IEEE Journal of Biomedical an...](#) > [Volume: 19 Issue: 6](#) 

## Systematic Poisoning Attacks on and Defenses for Machine Learning in Healthcare

**Publisher:** IEEE

[Cite This](#)

 PDF

[Mehran Mozaffari-Kermani](#) ; [Susmita Sur-Kolay](#) ; [Anand Raghunathan](#) ; [Niraj K. Jha](#) [All Authors](#)

*The attacker's goal is to add  $p'$  malicious examples to the original dataset to create a manipulated dataset. The malicious examples are generated using an attribute probability function.*

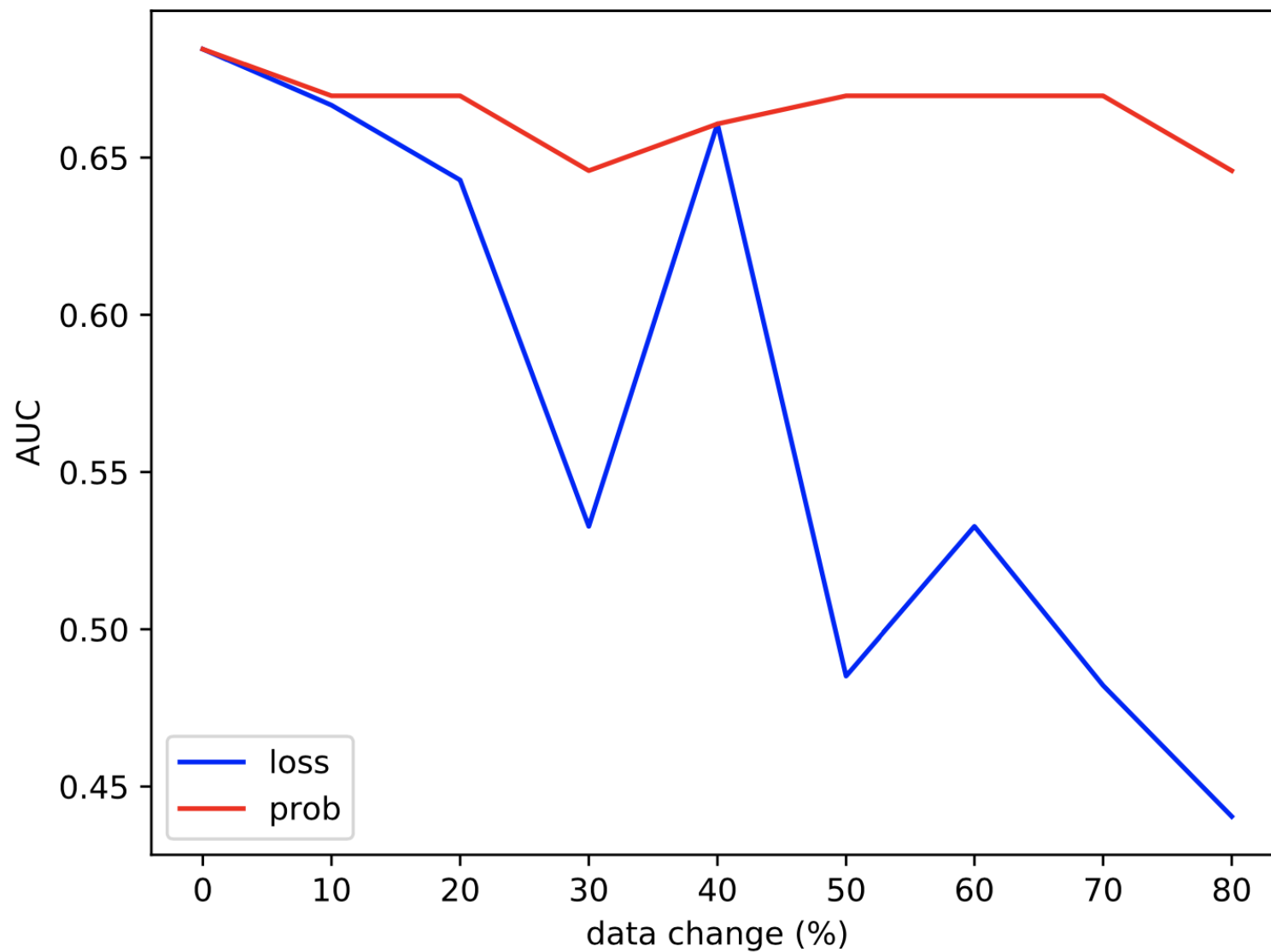
# Answer to RQ1

- 278 projects
- 76% of the 278 projects had no test cases.
- 85% of the studied projects that use testing had no test cases for testing classification algorithms
- 87% of the studied projects had no test cases to test model accuracy.
- None of the studied projects had any test cases to test accuracy decrease

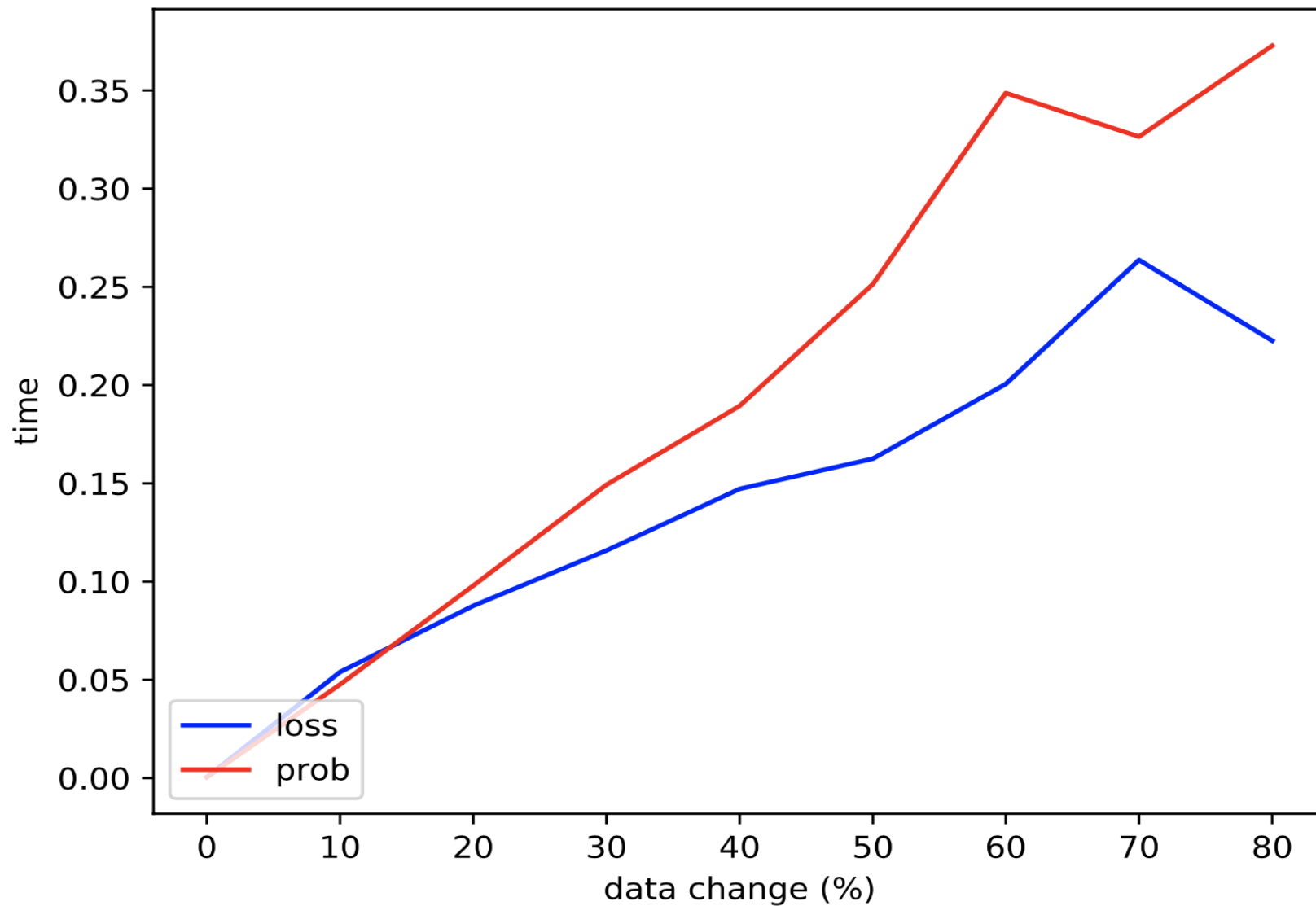
# Answer to RQ1

	GitHub	GitLab	ModelZoo
Unit testing	2,026	636	40
Algorithm-based testing	140	86	10
Metric-based testing	70	57	9
Misclassification testing	0	0	0

# Answer to RQ2



# Answer to RQ2



# Answer to RQ2

Approach	DATA			TIME		
	(min, max, median, avg)	p-value	$\Delta$	(min, max, median, avg)	p-value	$\Delta$
Baseline Approach	(80, 80, 80, 80)	0.0001	0.93	(0.35, 0.68, 0.36, 0.37)	0.0001	0.94
Our Approach	(20, 80, 60, 56)			(0.07, 0.37, 0.18, 0.17)		



# Answer to RQ2

```
1 import unittest
2 import label_perturbation_main
3 import SVC
4
5 class TestAttack( unittest.TestCase ):
6     def test_attack(self):
7         change_unit = 0.5
8         algo = "SVC"
9         auc4model1= run_experiment(algo)
10        auc4model2= run_perturbation(algo,change_unit)
11        self.assertEqual(auc4model1, auc4model2, "DECREASE
        ↪ IN AUC VALUE ... POSSIBLE ATTACK?" )
```

# Summary

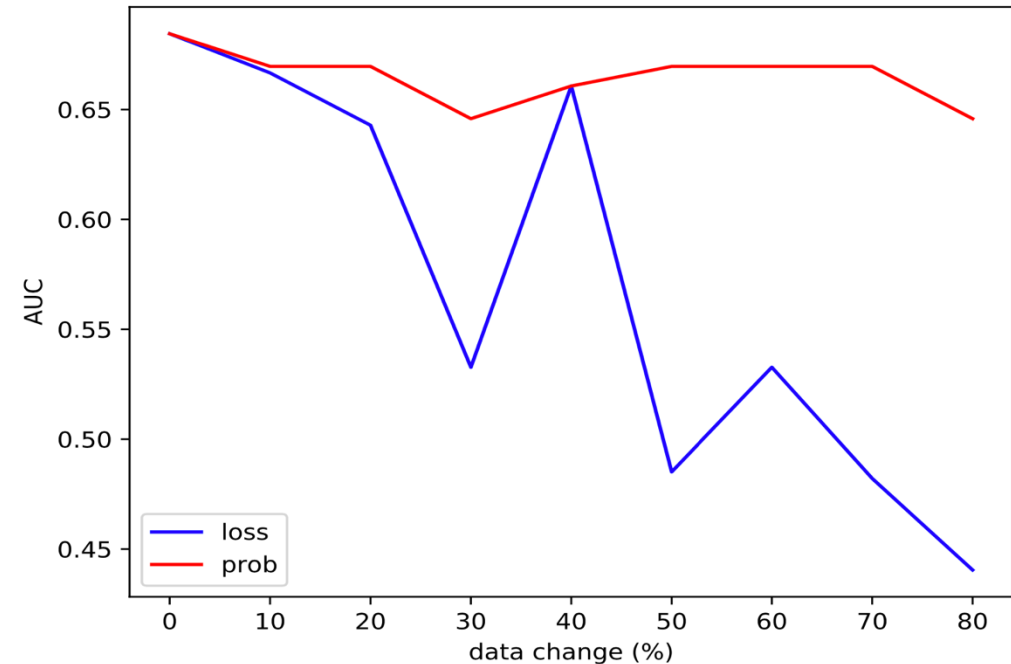
## Poisoning Attacks Against Machine Learning: Can Machine Learning Be Trustworthy?

Alina Oprea, Northeastern University

Anoop Singhal and Apostol Vassilev<sup>ID</sup>, National Institute of Standards and Technology



Open to Collaborations



[akond@auburn.edu](mailto:akond@auburn.edu)



[akondrahman.github.io](https://github.com/akondrahman)



@akondrahman



AUBURN