Antonio Guerriero

Tutor: Prof. Stefano Russo – co-Tutor: Prof. Roberto Pietrantuono

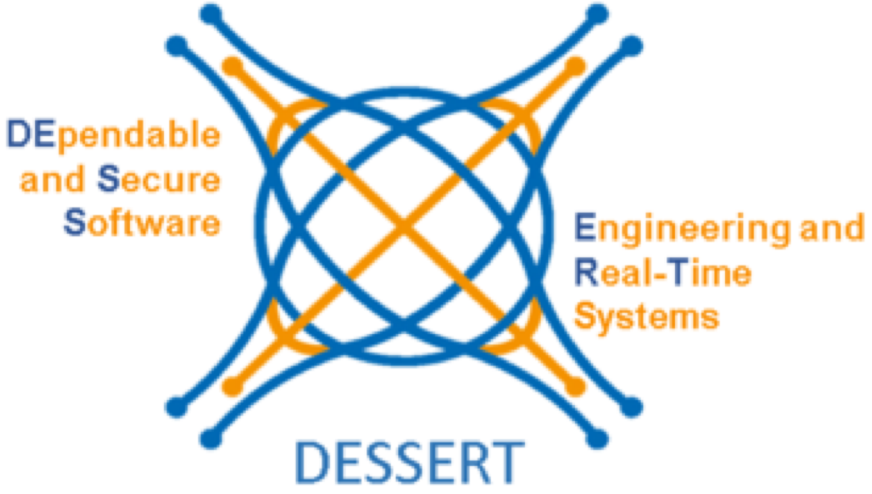
XXXIV Cycle - III year presentation

Operational Accuracy Assessment of CNN-based Image Classifiers

Background

- Master degree in Computer Engineering (2018)
  - Thesis: “Reliability Assessment of Microservice Architectures”

- Member of DESSERT research group at DIETI



Credits summary

Year 1	Year 2	Year 3										
			1	2	3	4	5	6	7	8		
Summary	Summary	Estimated	bimonth	bimonth	bimonth	bimonth	bimonth	bimonth	bimonth	bimonth	Summary	Total
25.9	16.6	0	0	0	0	2	0	0	0	0	2	44,5
12.8	0.8	0	0,2	0	6,4	0	0	0	0	0	6,6	20,2
37.0	44.0	50	7,4	8	1,6	5	8	8	8	4	50,0	131,0
75.7	61.4	50	7,6	8	8	7	8	8	8	4	58,6	195,7#

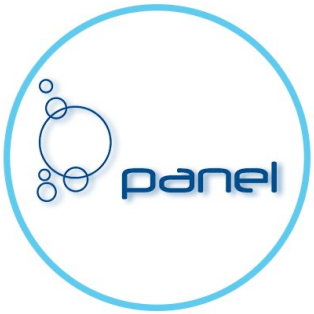
The maximum amount of credits (180) has been increased by 5 credits for each additional month (15) due to the extension of 3 months.

Period abroad

- Chinese University of Hong Kong (Hong Kong), supervised by Prof. Michael R. Lyu:
  - 2nd September to 13th November 2019
  - 2nd to 31st January 2021 (Smart Working)
  - Research on “testing of Machine Learning systems”

The Chinese University of Hong Kong

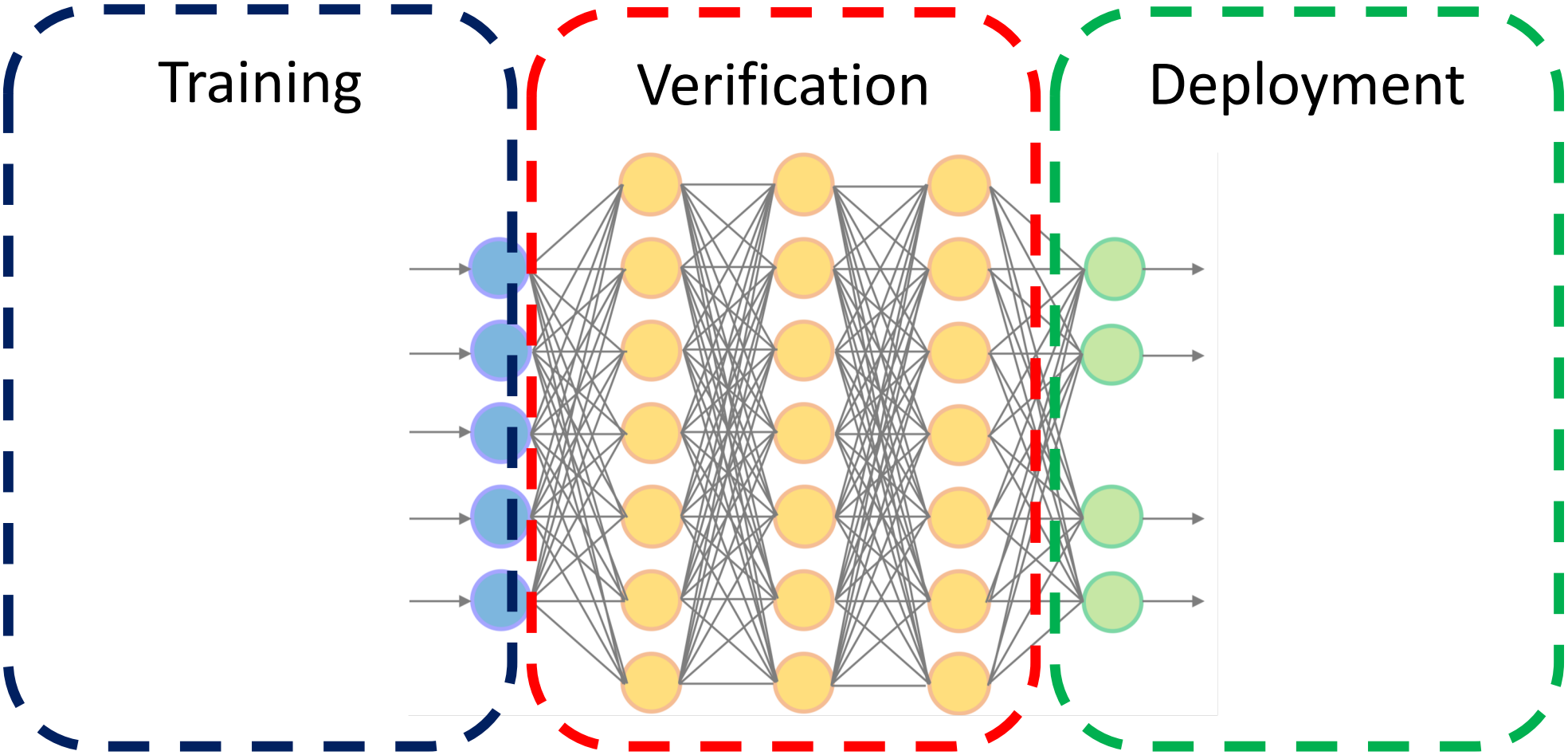
- Panel Sistemas Informaticos (Madrid, Spain), supervised by Javier L. C. Pinto:
  - 20th October to 21st November 2021
  - EU Marie Curie “uDevOps” project
  - Research on “Machine Learning techniques for reliable Microservice Architectures”



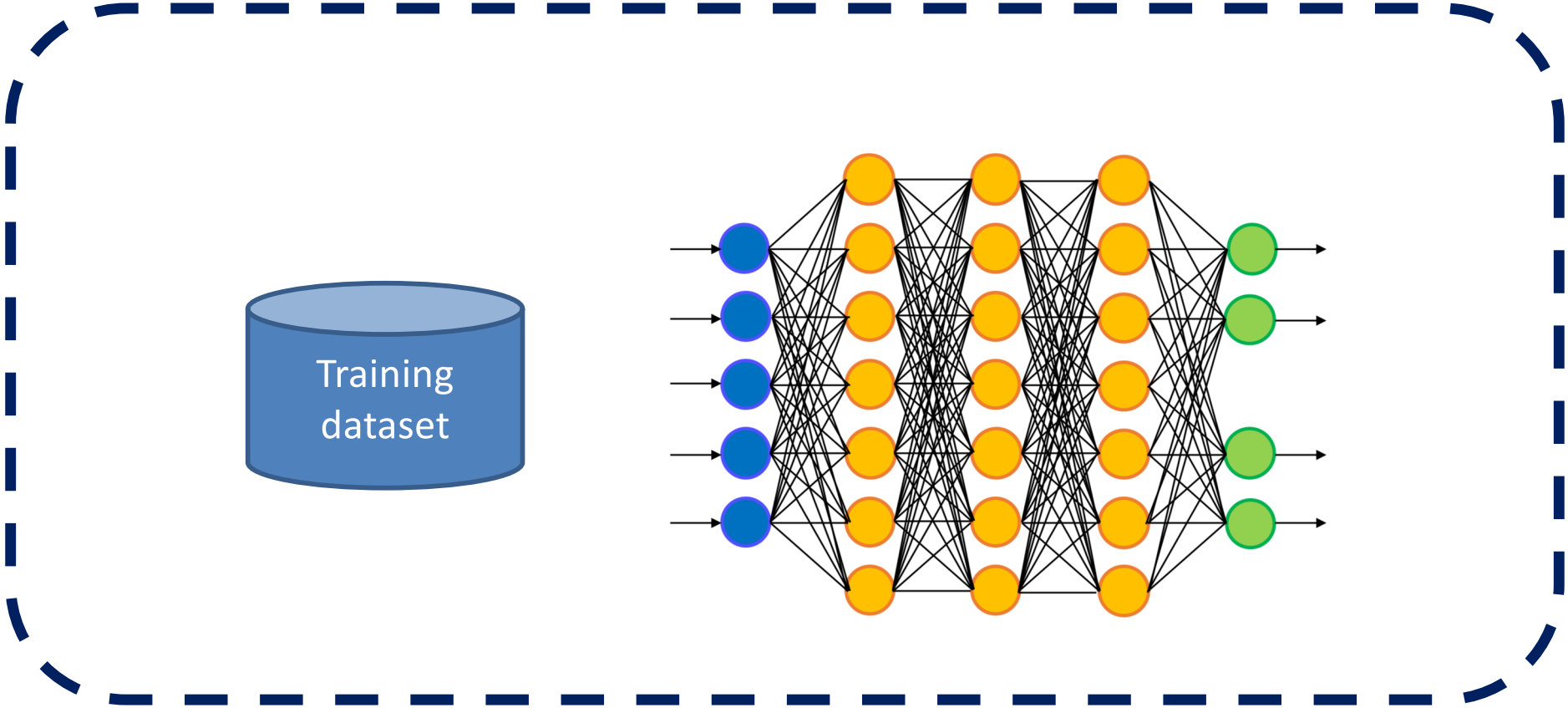
Research activity

Operational accuracy assessment of CNN-based Image Classifiers

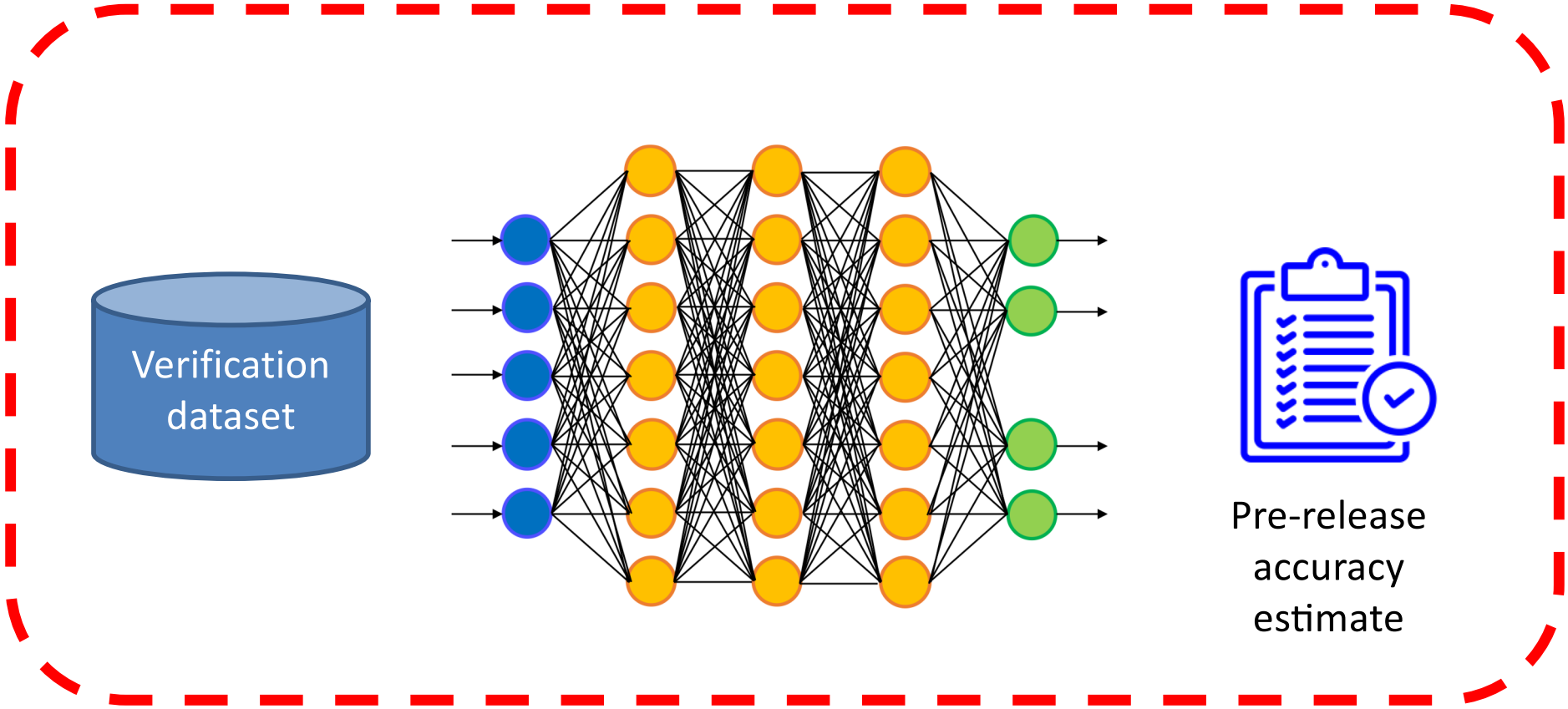
Classical ML models life cycle



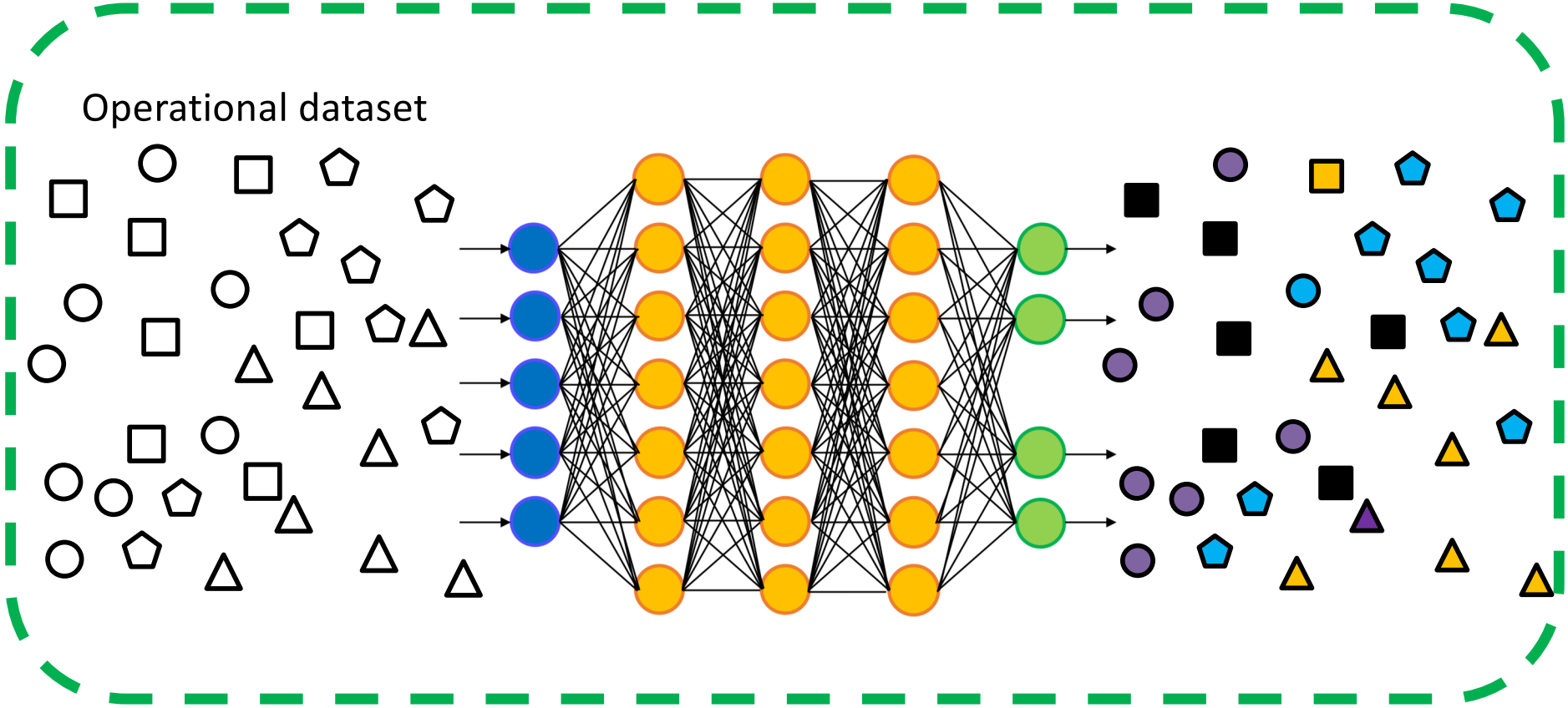
Training



Verification

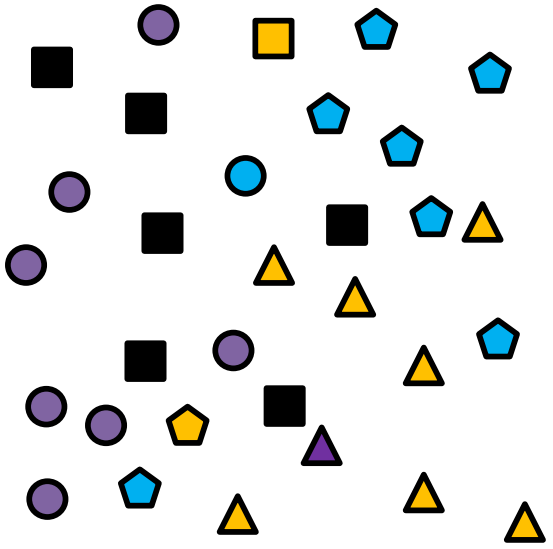


Deployment



Operational Accuracy

- *The ratio of the number of correctly classified images to the total number of images in the operational dataset*



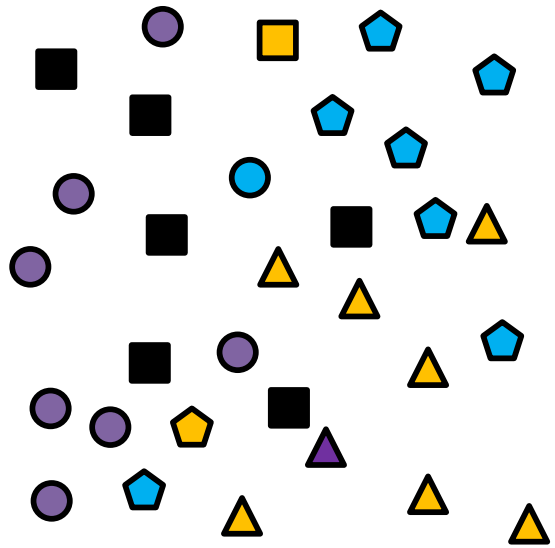
Operational dataset

?

Operational Accuracy

Oracle problem

- *«There is no reliable test oracle to indicate what the correct output should be for arbitrary input» [*Murphy*]*



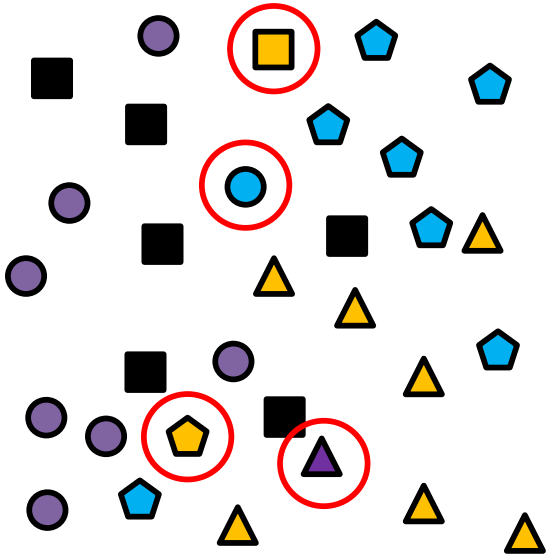
Operational dataset

Operational Accuracy

[Murphy] C. Murphy, G. E. Kaiser, and M. Arias. An approach to software testing of machine learning applications. In 19th Int. Conference on Software Engineering and Knowledge Engineering (SEKE), pages 167–172, 2007.

Manual labeling

- The simplest solution is to manually label all the images in the operational dataset

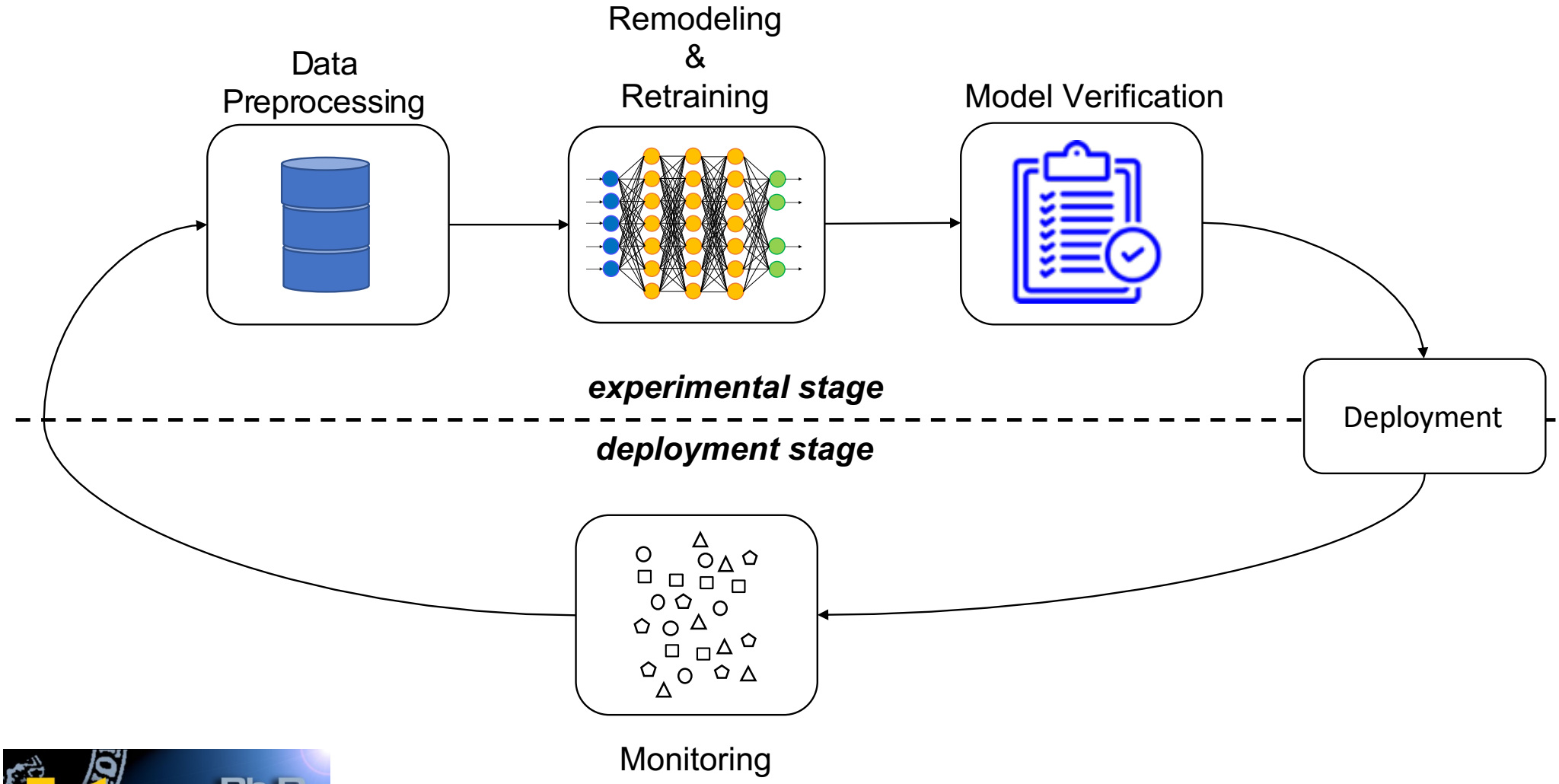


Operational dataset

87.1%

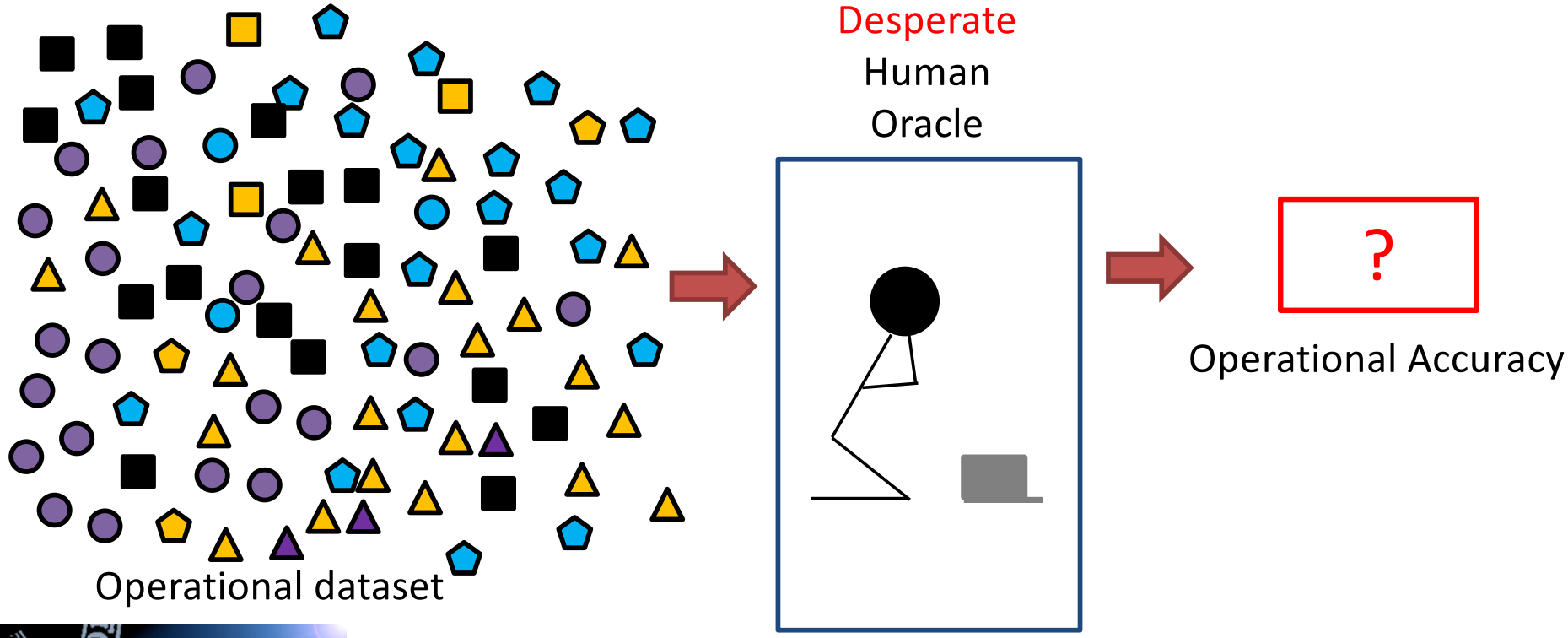
Operational Accuracy

ML systems life cycle



Issue of manual labeling

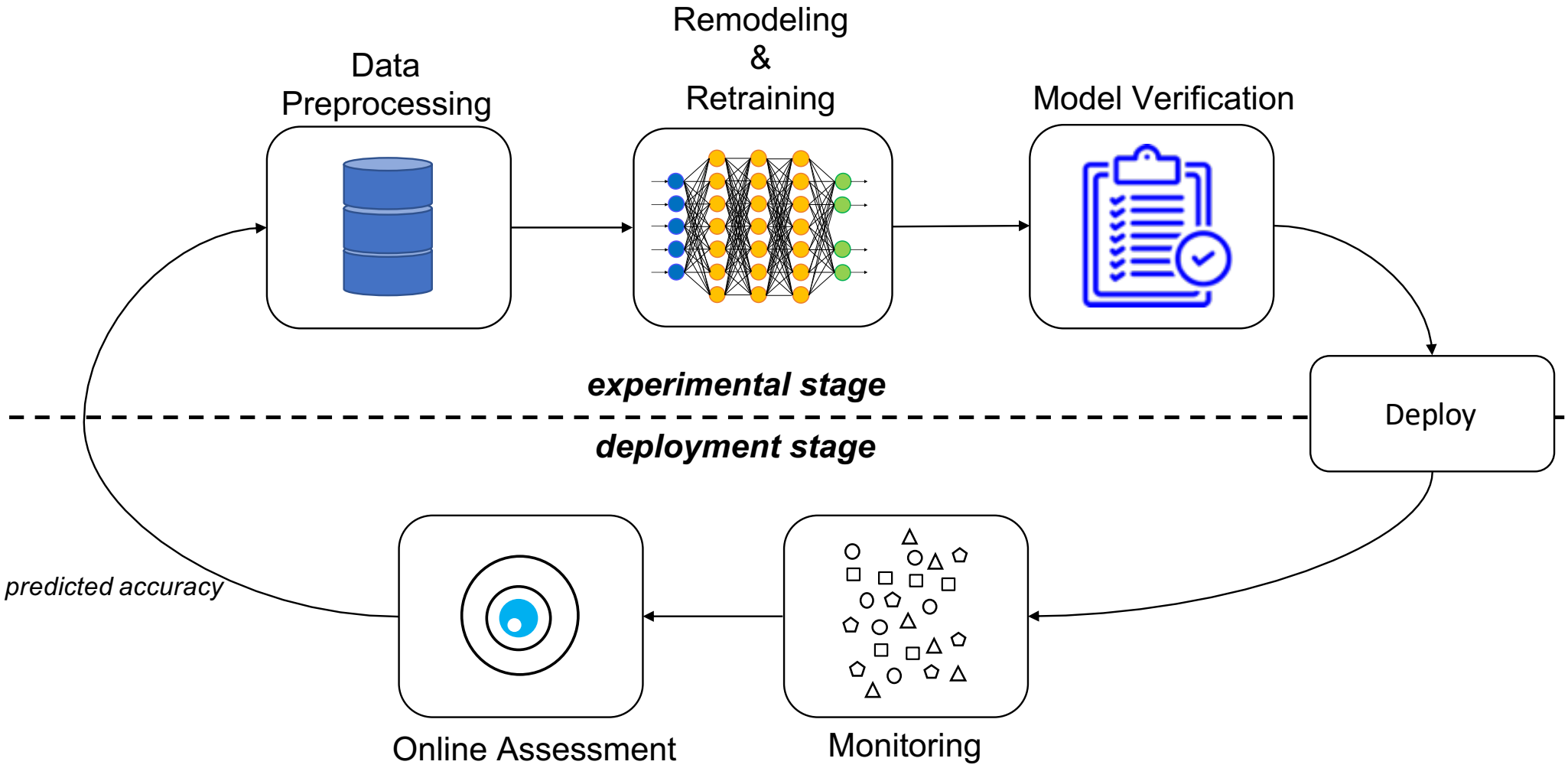
- Infeasible for huge datasets



Thesis contributions

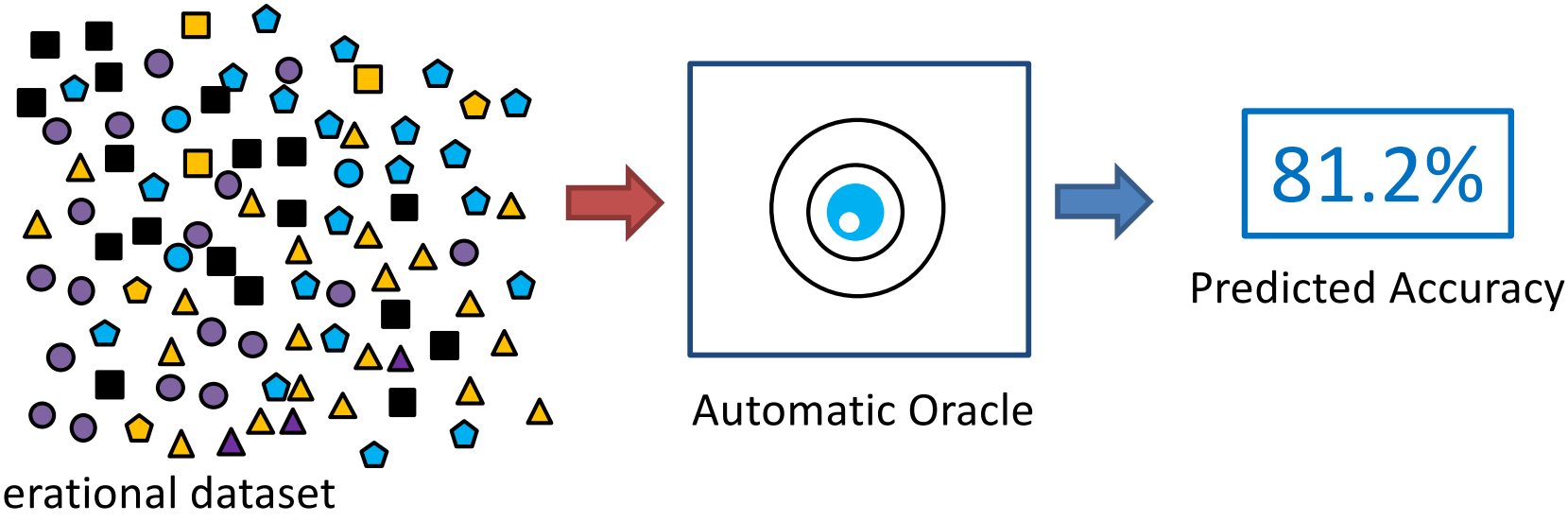
- Two strategies for the operational accuracy assessment:
  - *Online assessment*: via automatic oracles
  - *Offline assessment*: via statistical sampling

Online assessment

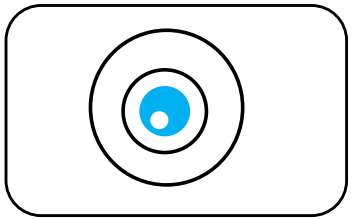


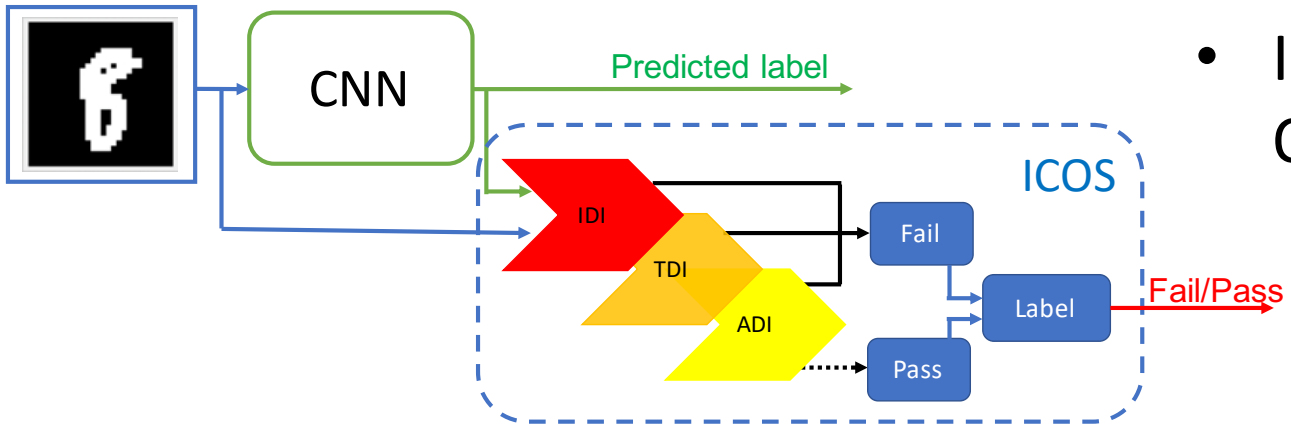
Online assessment (2)

- Automatic pseudo-oracles are used to:
  - Automatically evaluate as *Fail* or *Pass* each output of the CNN
  - Assess the accuracy provided in operation based on their predictions



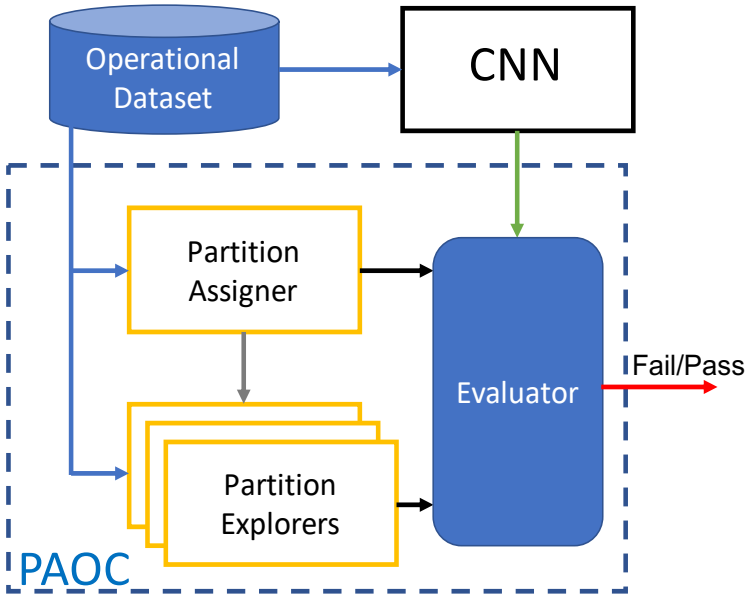
Contributions





- Image Classification Oracle Surrogate (ICOS)

- Partitioning-based Automatic Oracle for CNN (PAOC)



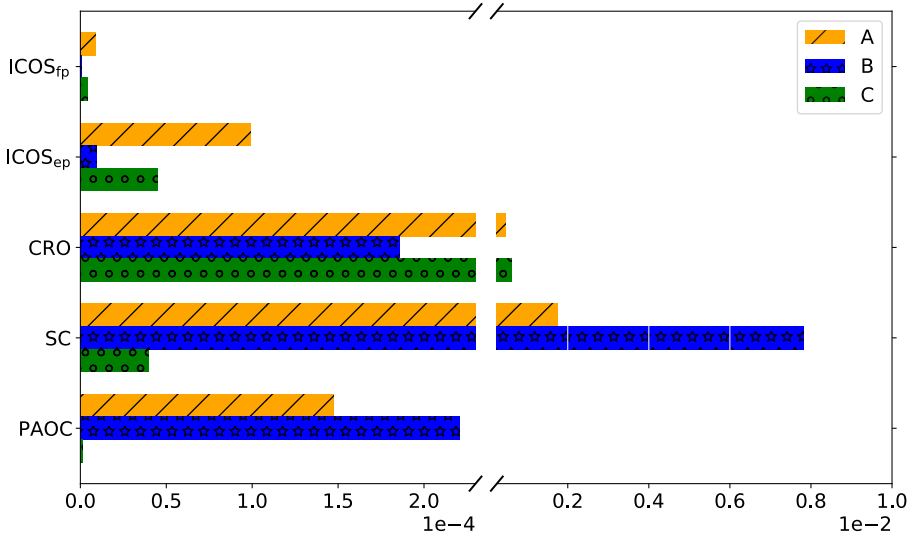
Experimentation

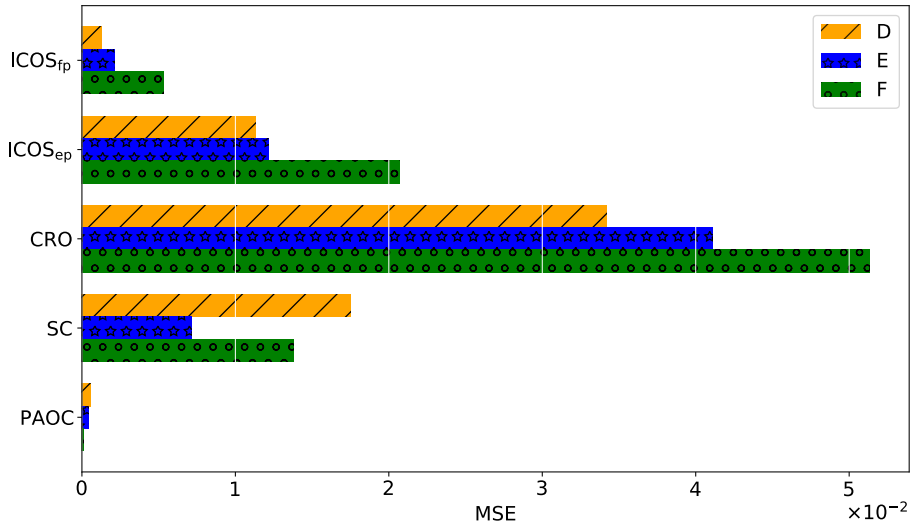
- 2 Baselines:
  - Cross Referencing Oracle (CRO): implementing Multiple Implementation Testing [Srisakaokul]
  - SelfChecker (SC): state of the art pseudo-oracle for failure detection of ML systems [Xiao]
- 3 datasets: MNIST, CIFAR10, CIFAR100
- 9 Convolutional Neural Networks (3 for each dataset)
- 3 Research Questions:
  - RQ1: *effectiveness*
  - RQ2: *sensitivity*
  - RQ3: *stability*

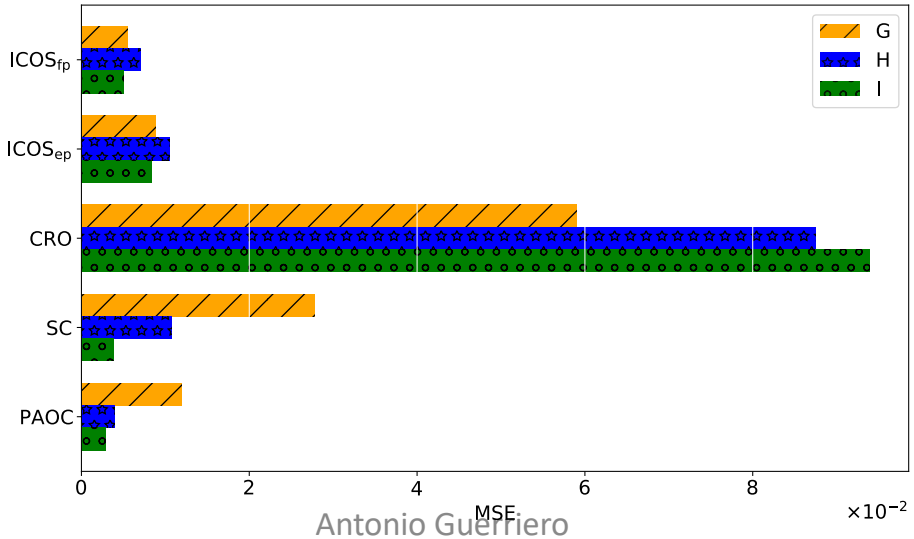
[Srisakaokul] S. Srisakaokul, Z. Wu, A. Astorga, O. Alebiosu, and T. Xie. Multiple- implementation testing of supervised learning software. In AAAI Workshops. Association for the Advancement of Artificial Intelligence, 2018.
[Xiao] Y. Xiao, I. Beschastnikh, D. S. Rosenblum, C. Sun, S. G. Elbaum, Y. Lin, and J. S. Dong. Self-checking deep neural networks in deployment. In 2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE), pages 372–384, 2021.

RQ1: *effectiveness*

- Metric: *Mean Squared Error* (MSE) over 30 repetitions

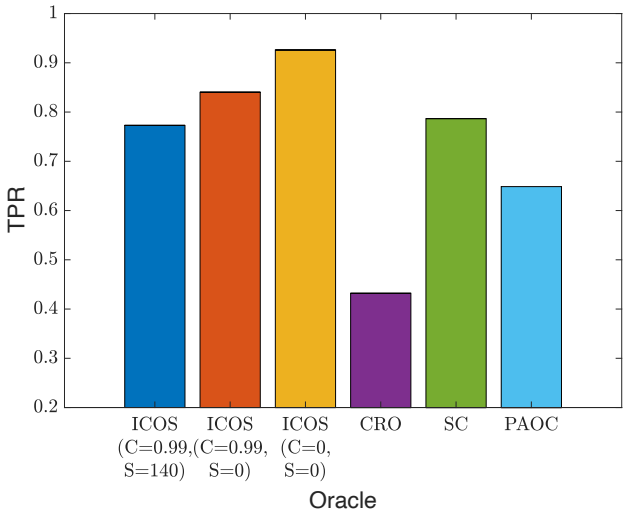


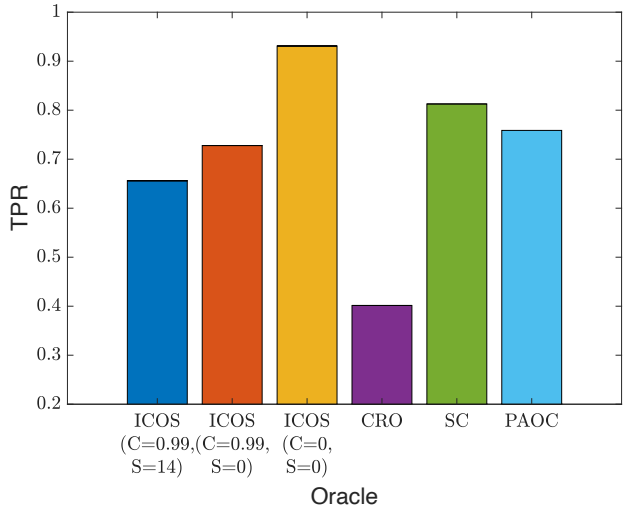


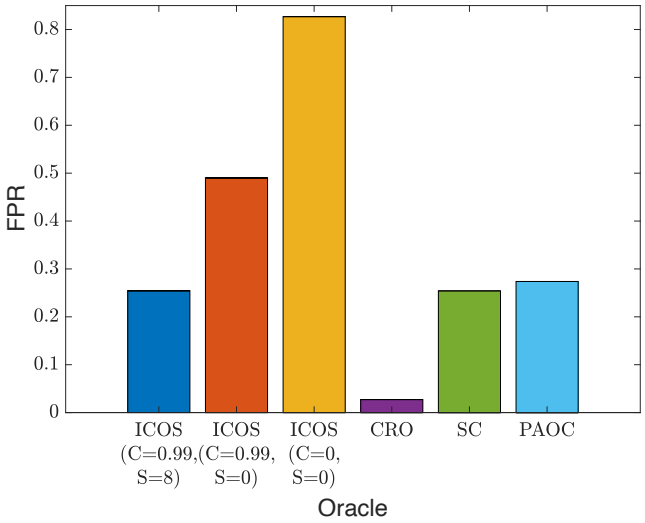


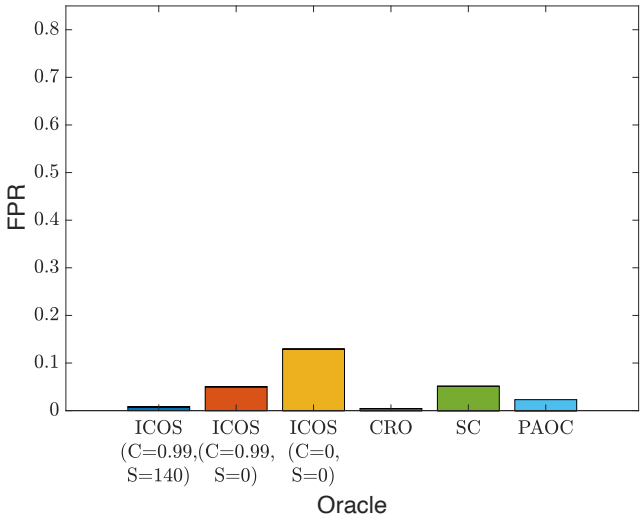
RQ2: *sensitivity*

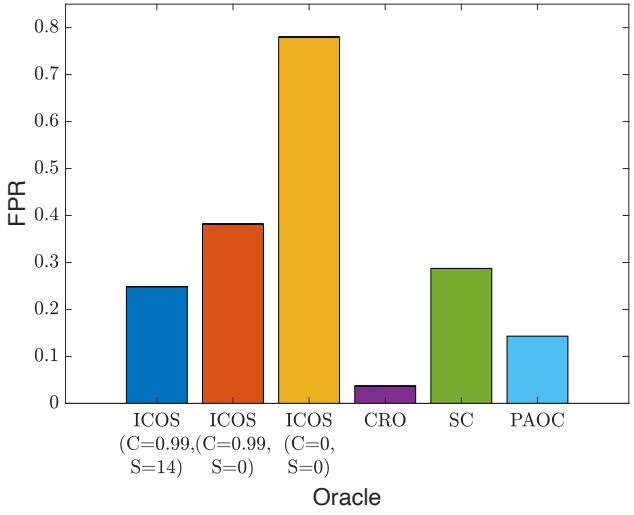
- Metrics:
  - *True Positives Rate*
  - *False Positive Rate*

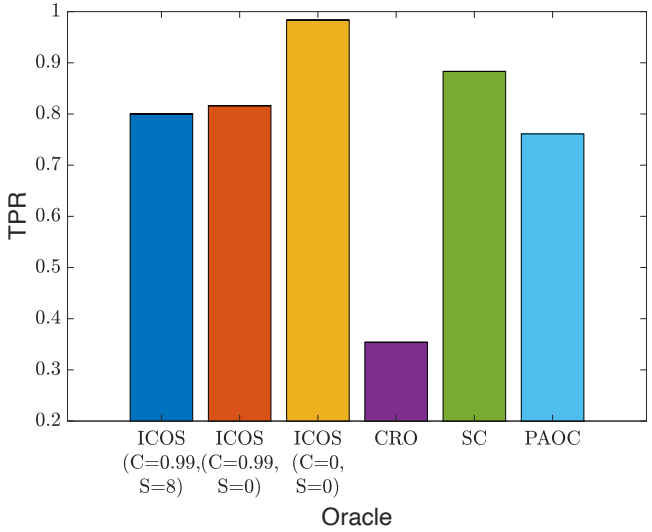






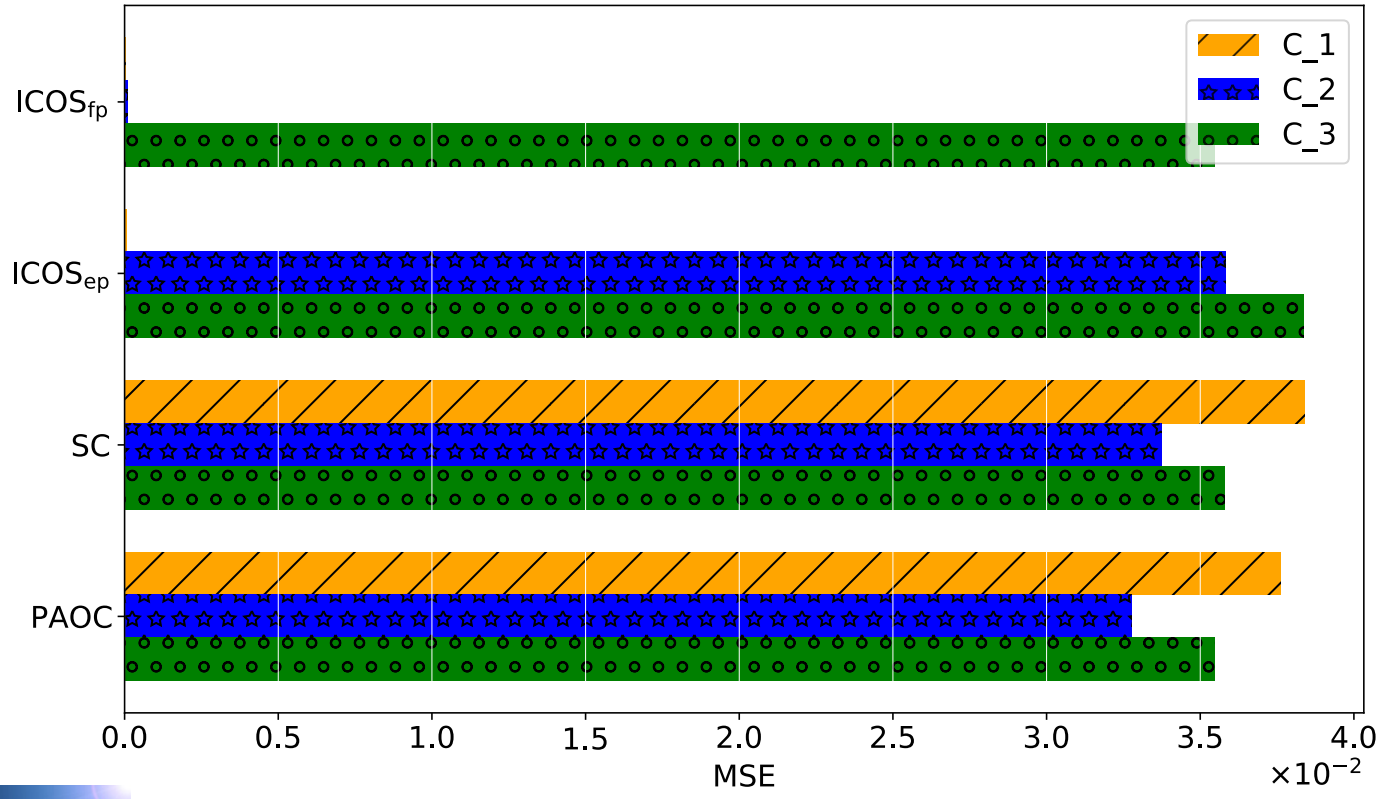




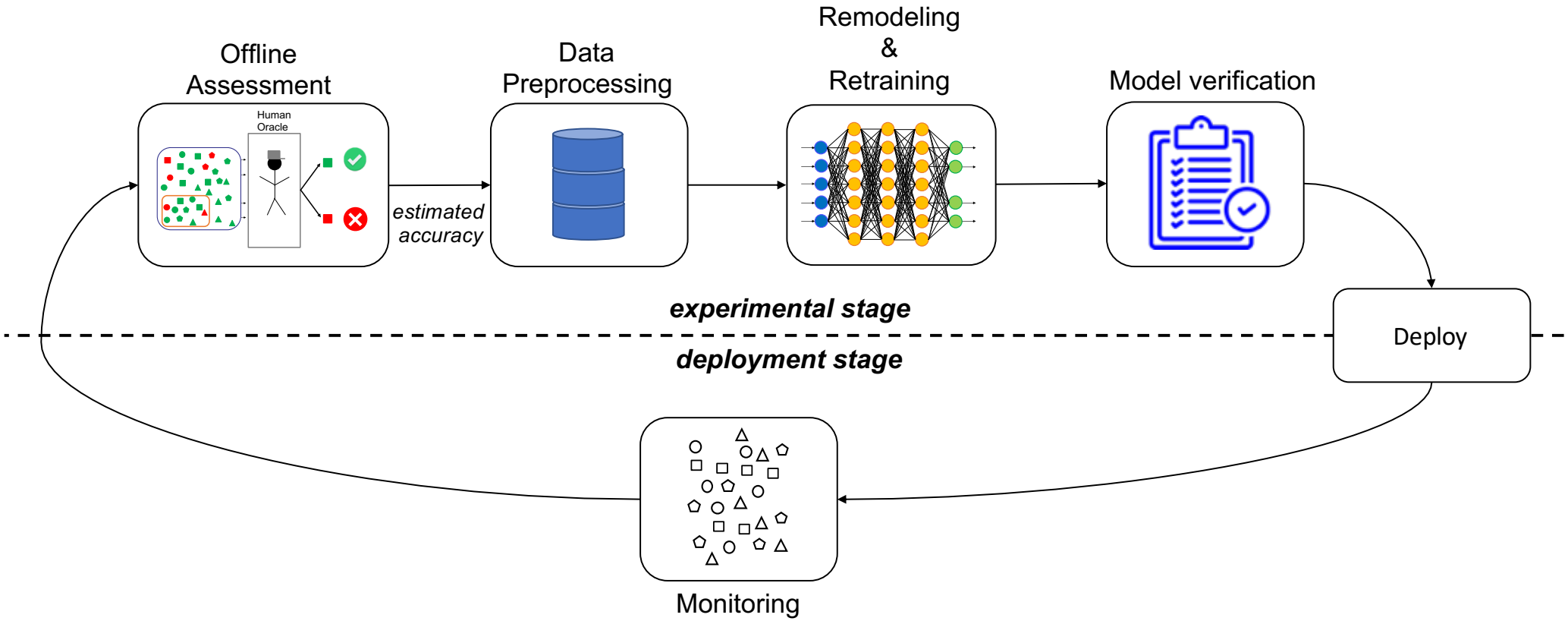


RQ3: *stability*

- CNN C trained on three *shifted* training dataset
- Metric: *Mean Squared Error* (MSE) over 30 repetitions

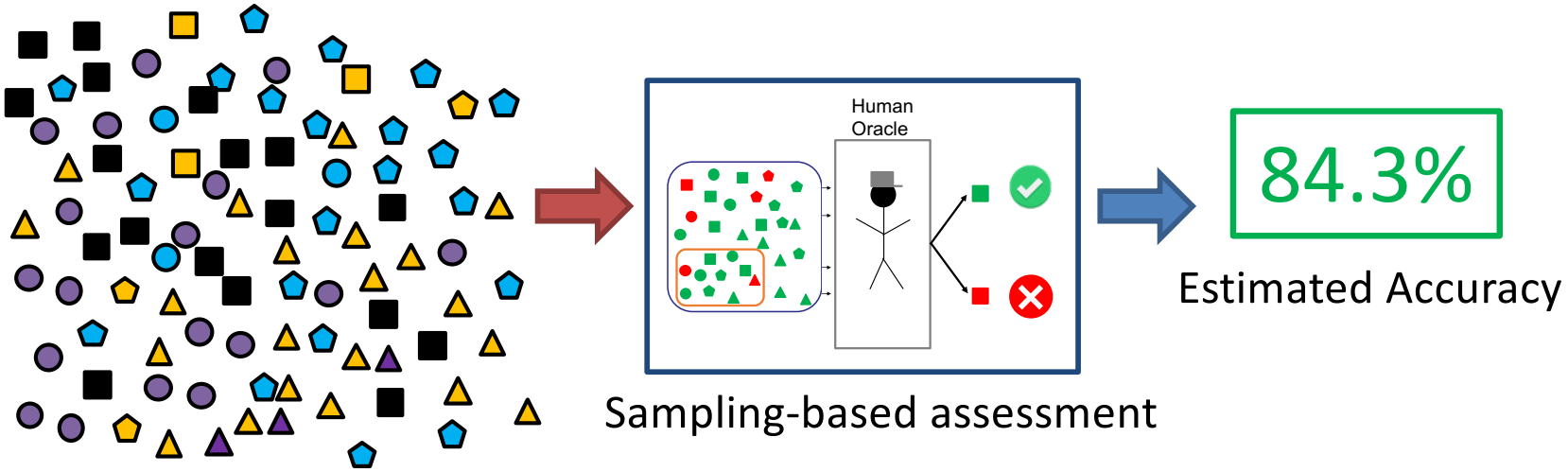


Offline assessment

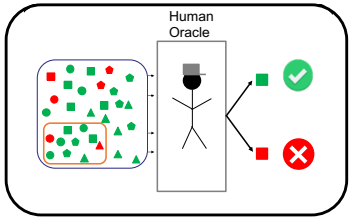


Offline assessment (2)

- Sampling strategies are used to:
  - Select the most representative images of the operational dataset
  - Estimate the operational accuracy via manual labeling (*estimated accuracy*)

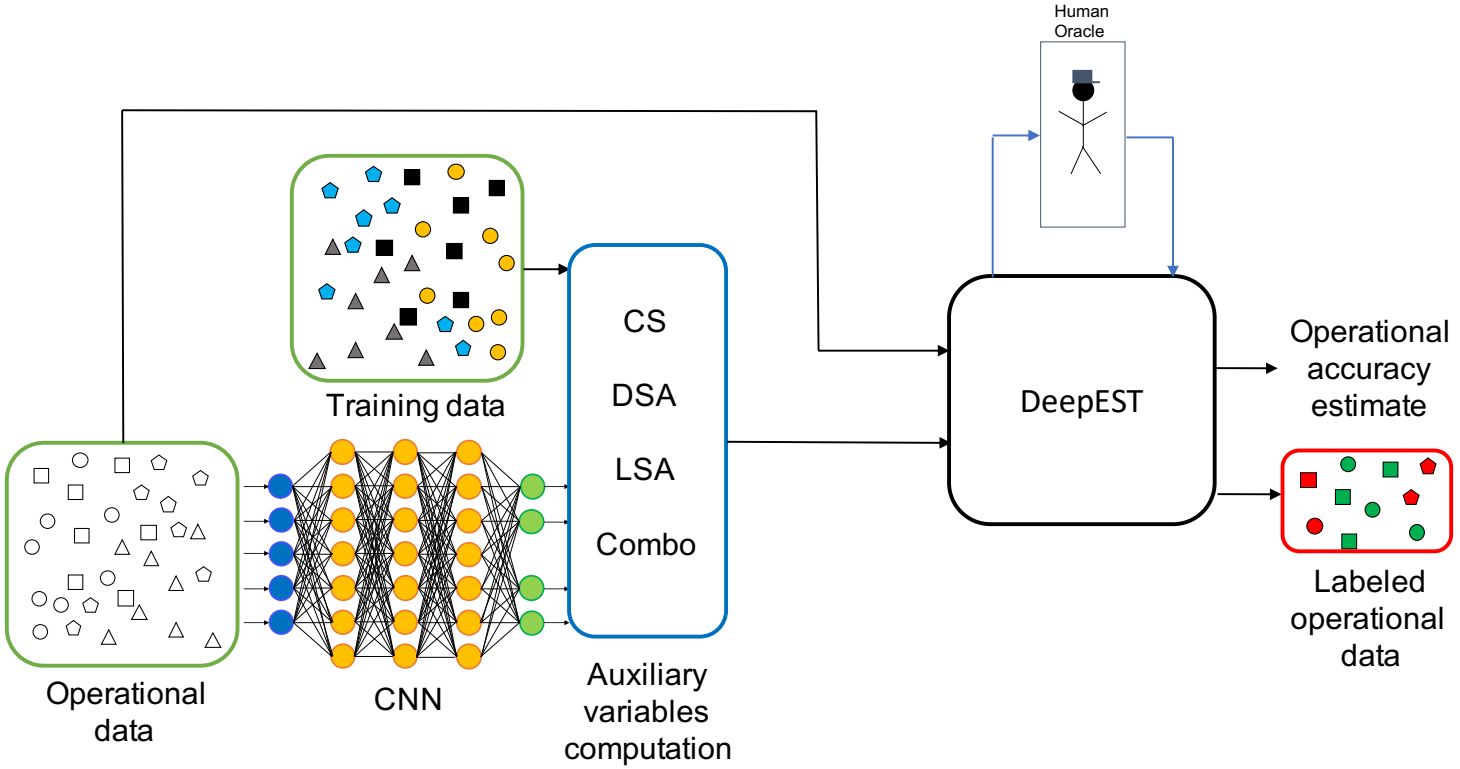


Contribution



Offline Assessment

- DeepEST



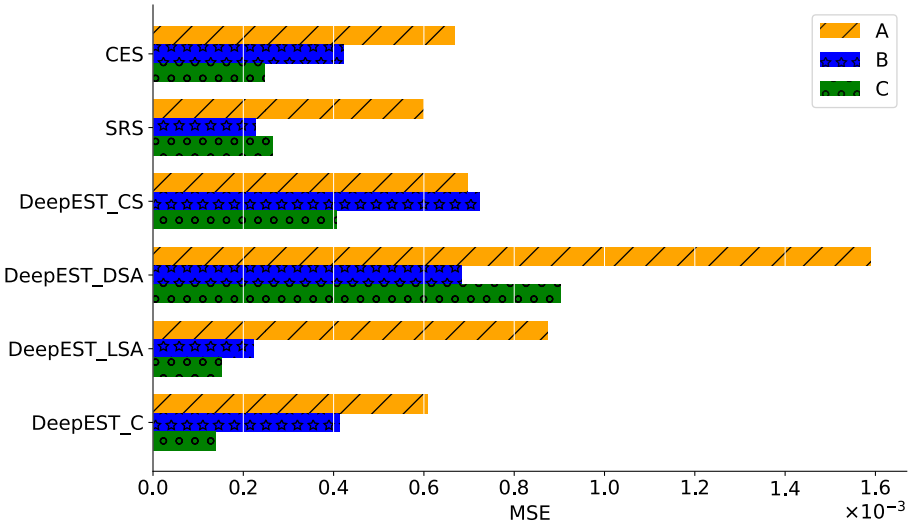
Experimentation

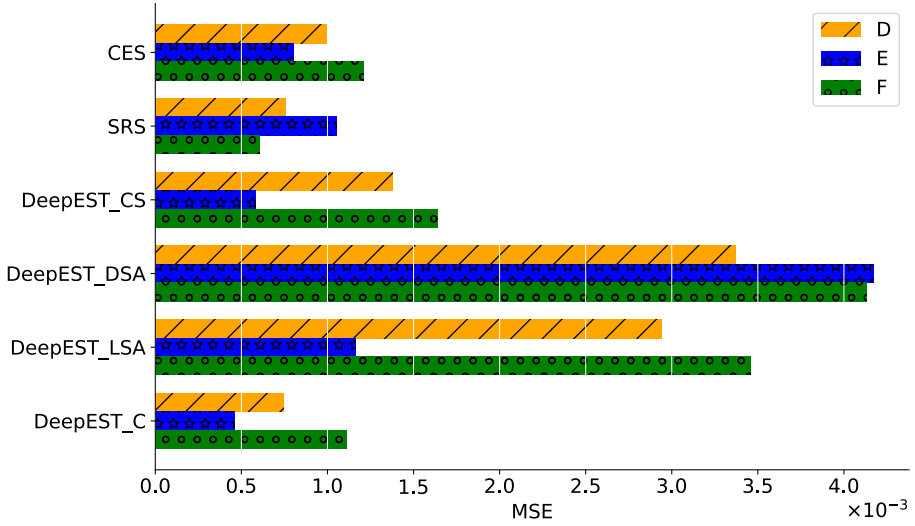
- 2 Baselines:
  - Simple Random Sampling (SRS)
  - Cross Entropy Sampling (CES): sampling technique for the operational testing of CNN [Li]
- 3 datasets: MNIST, CIFAR10, CIFAR100
- 9 Convolutional Neural Networks (3 for each dataset)
- 3 Research Questions:
  - RQ1: *effectiveness*
  - RQ2: *sensitivity*
  - RQ3: *stability*

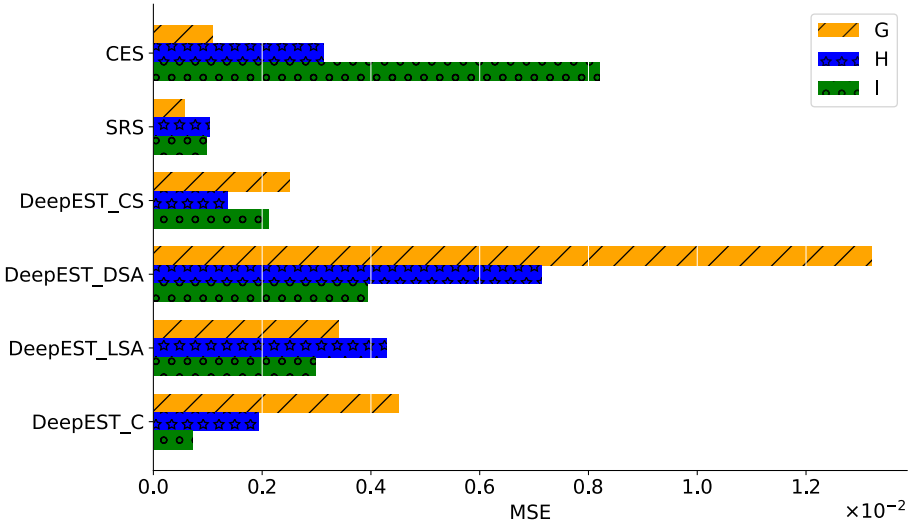
[Li] Z. Li, X. Ma, C. Xu, C. Cao, J. Xu, and J. Lü. Boosting Operational DNN Testing Efficiency through Conditioning. In Proc. 27th ACM Joint Meet- ing on European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE), pages 499–509. ACM, 2019.

RQ1: *effectiveness*

- Metric: *Mean Squared Error* (MSE) over 30 repetitions

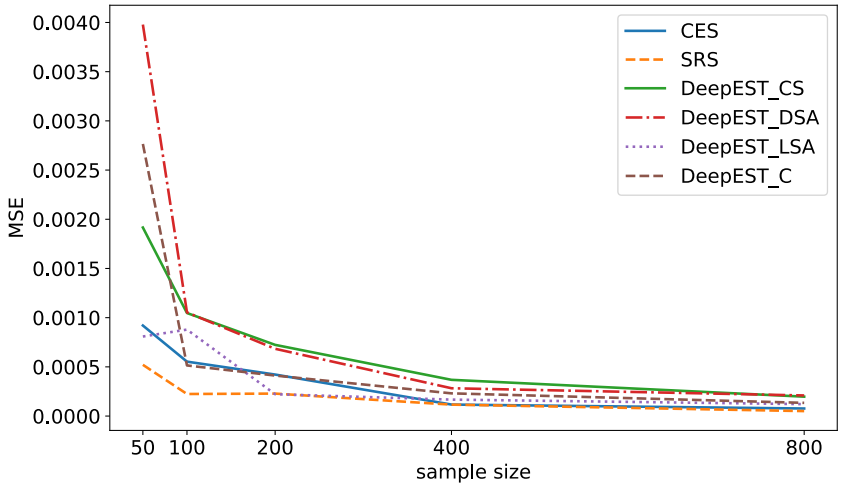


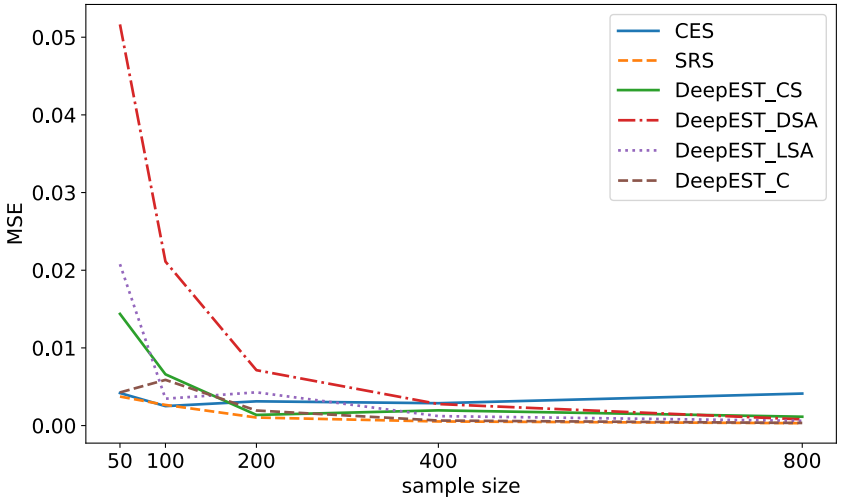


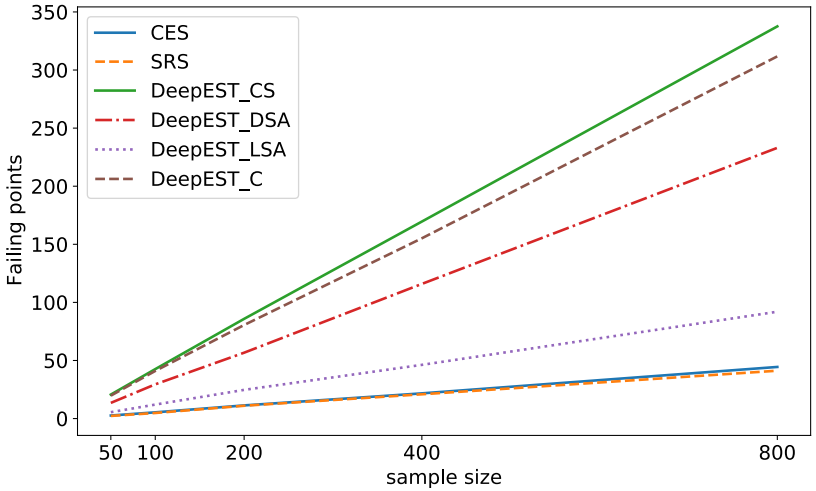


RQ2: *sensitivity*

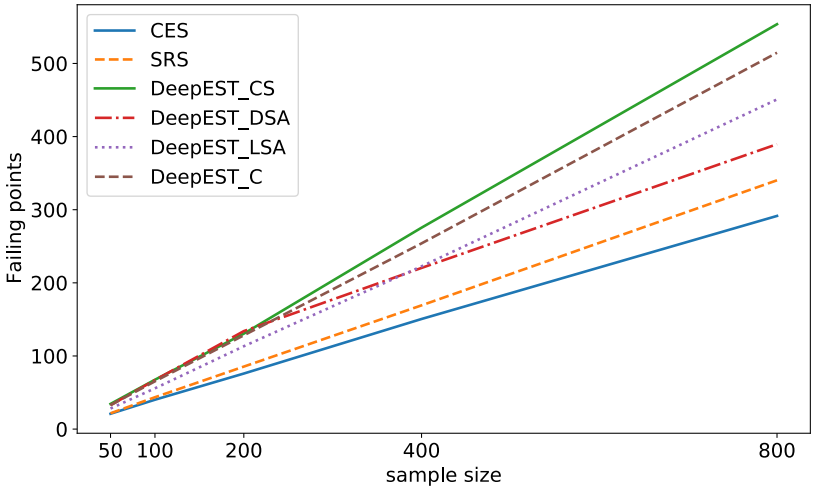
- Metrics:
  - *MSE and Number of Failing Points*







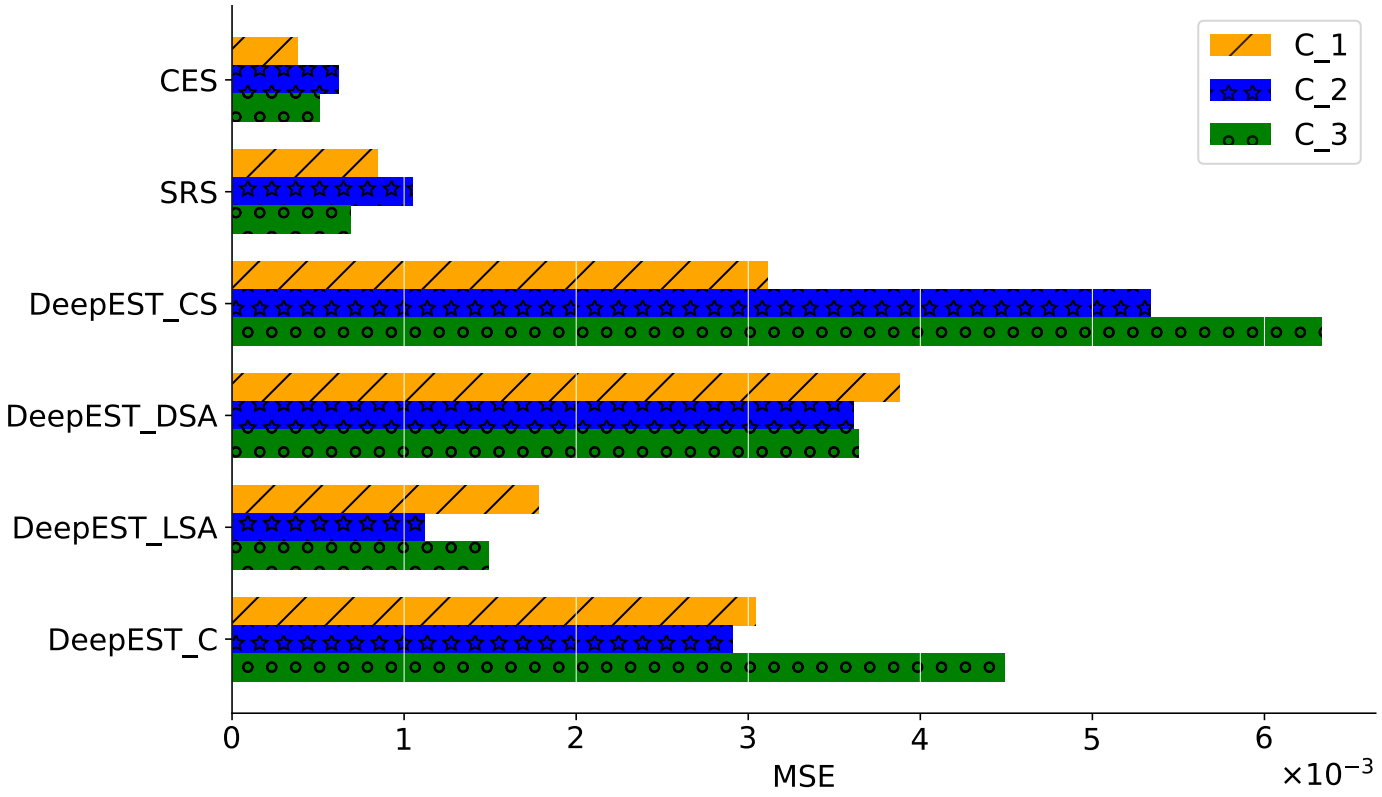
CNN B



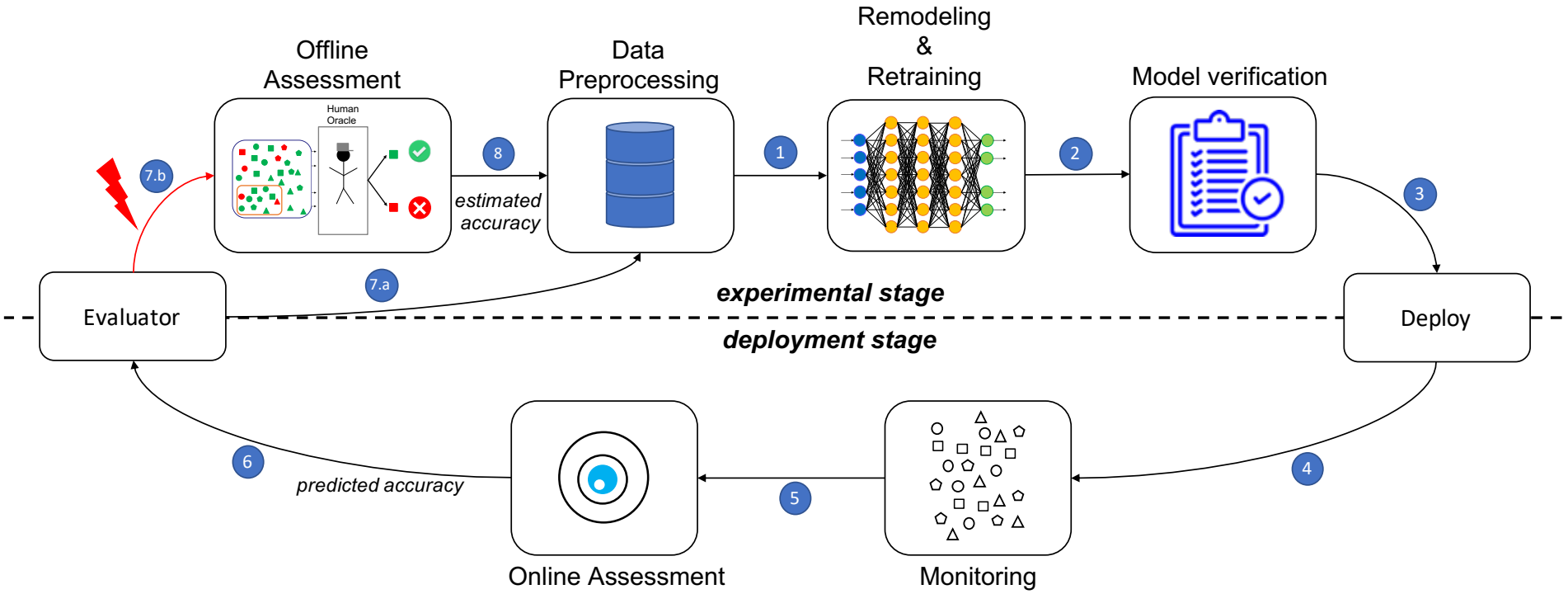
CNN H

RQ3: *stability*

- CNN C trained on three *shifted* training dataset
- Metric: *Mean Squared Error* (MSE) over 30 repetitions

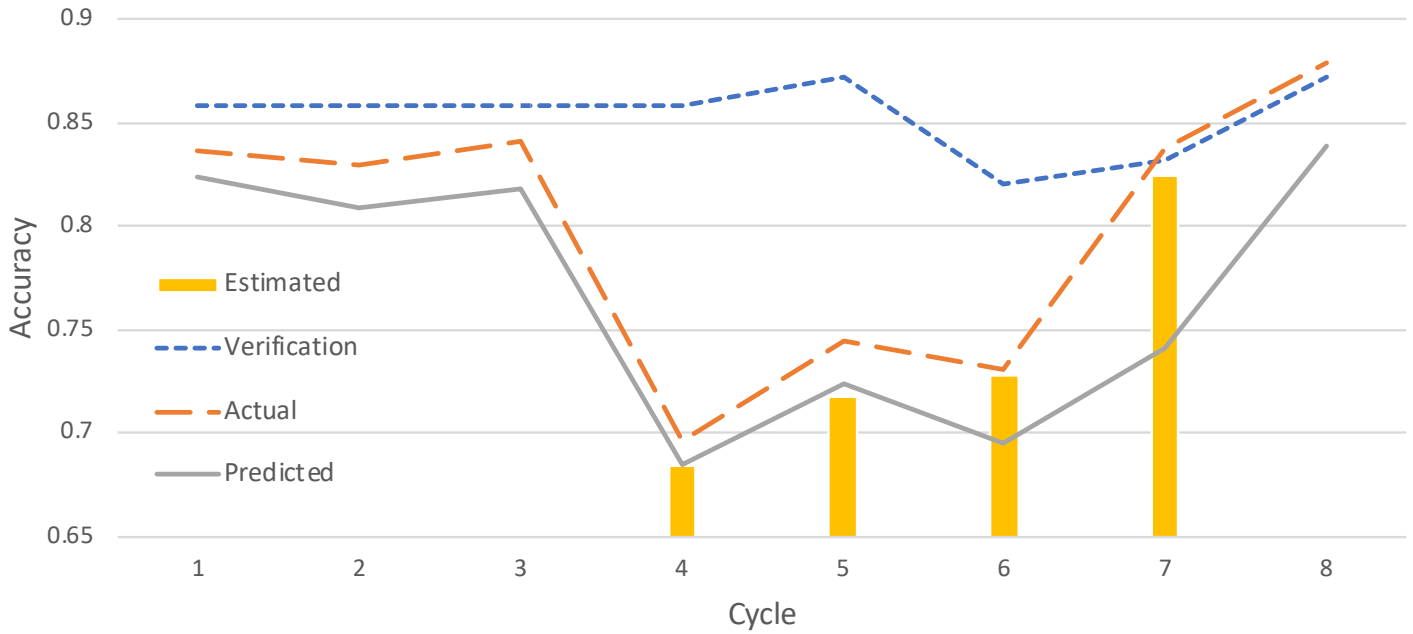


Accuracy Assessment Cycle



Simulation of the AAC

- The results show:
  - Predicted accuracy follows the actual accuracy
  - Estimated accuracy is triggered only in case of accuracy drops
  - New labeled images help to improve the actual accuracy of the CNN in operation



Conclusions

- Accuracy Assessment Cycle allows combining online and offline assessment to exploit their characteristics minimizing the cost
- Operational features are useful to assess the accuracy of ML systems
- The assessment and the improvement of the accuracy of the ML systems can be of interest beyond the IC domain

Publications

A. Guerriero, R. Pietrantuono and S. Russo, Operation is the Hardest Teacher: Estimating DNN Accuracy Looking for Mispredictions, 2021 **IEEE/ACM 43rd International Conference on Software Engineering** (ICSE), doi: 10.1109/ICSE43902.2021.00042.

A. Guerriero, Reliability Evaluation of ML systems, the oracle problem, 2020 **IEEE International Symposium on Software Reliability Engineering Workshops** (ISSREW), doi: 10.1109/ISSREW51248.2020.00050.

A. Bertolino, G. De Angelis, A. Guerriero, B. Miranda, R. Pietrantuono, S. Russo, DevOpRET: Continuous reliability testing in DevOps. **Journal of Software: Evolution and Process**, e2298, doi: 10.1002/smr.2298.

A. Bertolino, A. Guerriero, B. Miranda, R. Pietrantuono, and S. Russo, Learning-to-rank vs ranking-to-learn: strategies for regression testing in continuous integration, In Proceedings of the **ACM/IEEE 42nd International Conference on Software Engineering** (ICSE '20), 2020, ACM, New York, NY, USA, 1–12, doi: 10.1145/3377811.3380369.

R. Pietrantuono, S. Russo, A. Guerriero, Testing microservice architectures for operational reliability. **Software Testing, Verification and Reliability**, 30(2), e1725, doi: 10.1002/stvr.1725.

A. Guerriero, R. Mirandola, R. Pietrantuono and S. Russo, A Hybrid Framework for Web Services Reliability and Performance Assessment, 2019 **IEEE International Symposium on Software Reliability Engineering Workshops** (ISSREW), doi: 10.1109/ISSREW.2019.00070.

R. Pietrantuono, S. Russo, A. Guerriero, Run-time Reliability Estimation of Microservice Architectures, Proc. of the 2018 **IEEE International Symposium on Software Reliability Engineerin**g (ISSRE), Memphis, TN, USA, Oct. 15-18, IEEE, 2018, *Winner of "**Best Research Paper Award**"*, doi: 10.1109/ISSRE.2018.00014.