Finding Bugs in Deep Learning Programs

Foutse Khomh, PhD, Ing.
foutse.khomh@polymtl.ca
@SWATLab
Canada CIFAR AI Chair on Trustworthy Machine Learning Systems

POLYTECHNIQUE MONTRÉAL
UNIVERSITÉ D'INGÉNIERIE

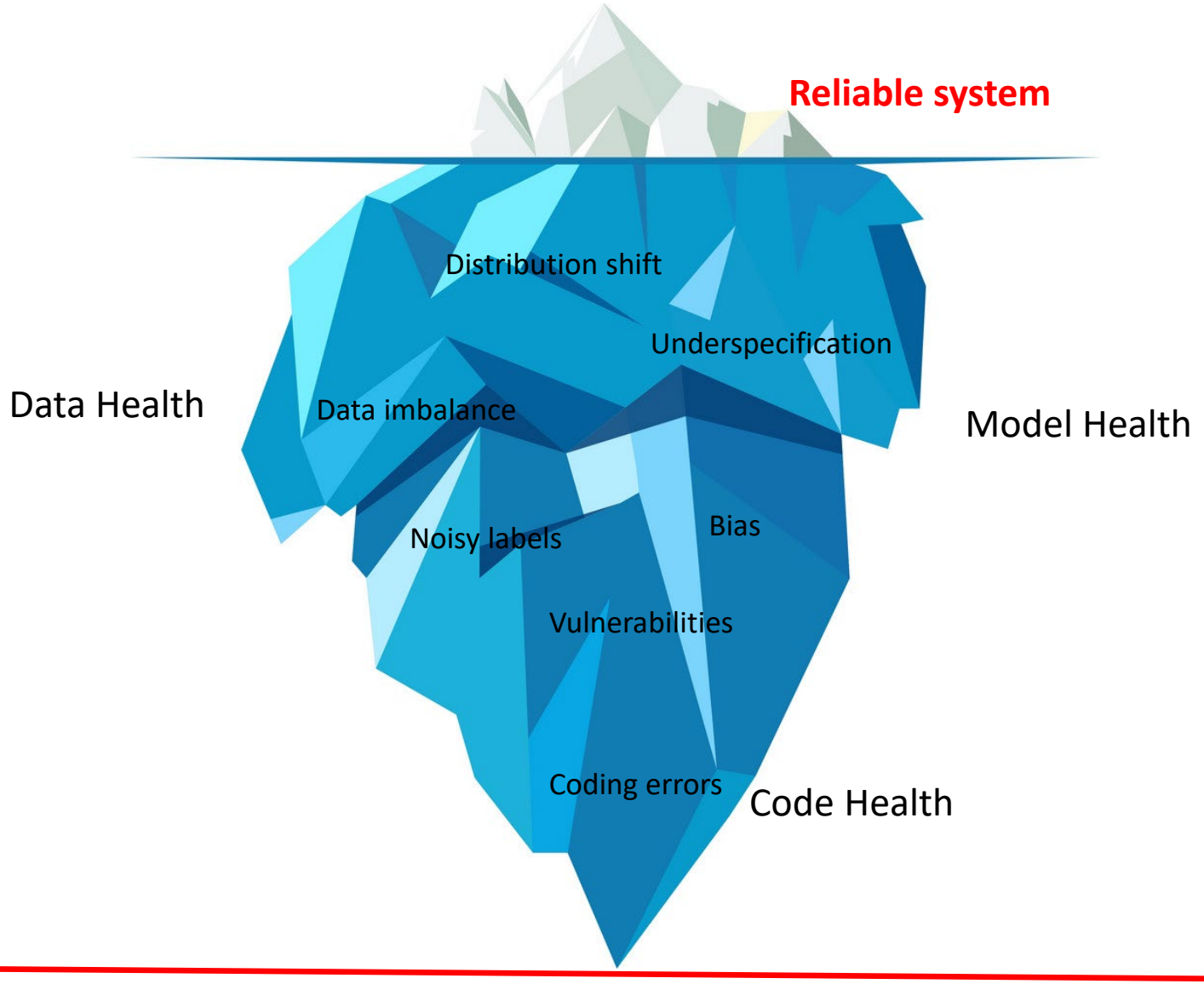
Mila

SEMLA
Trustworthy Engineering of AI Software

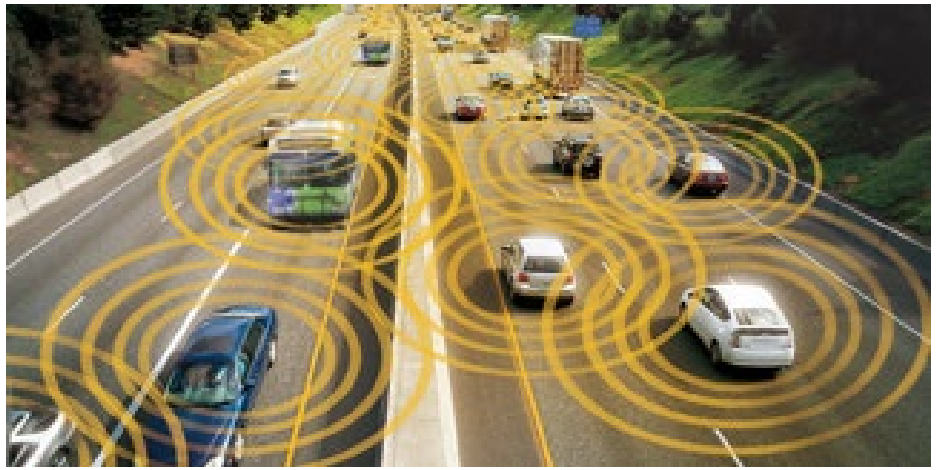
S.W.A.T
SoftWare Analytics and Technologies Lab

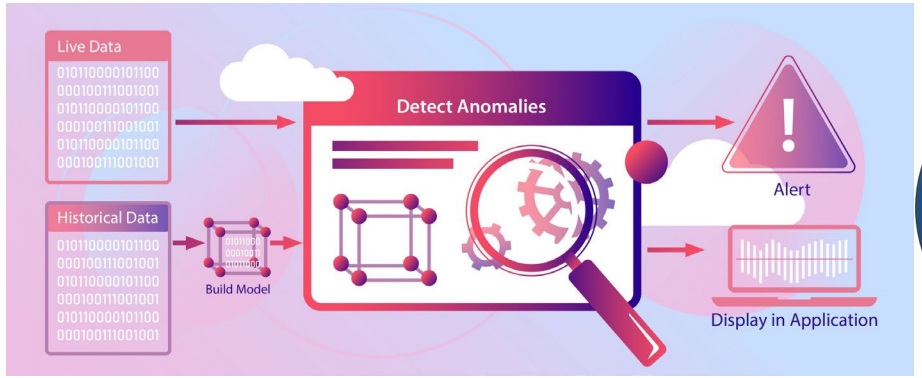
Quality Assurance of ML-enabled systems

System evolution & continuous delivery

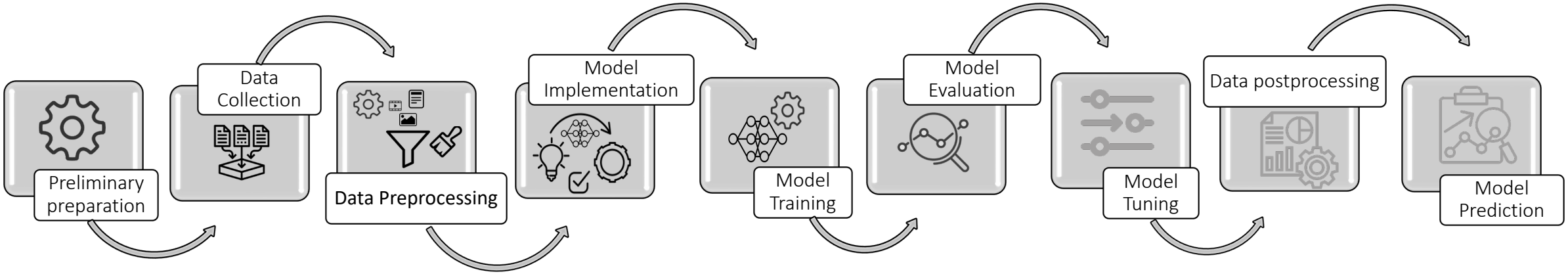


Deep Learning is being rapidly adopted in industry





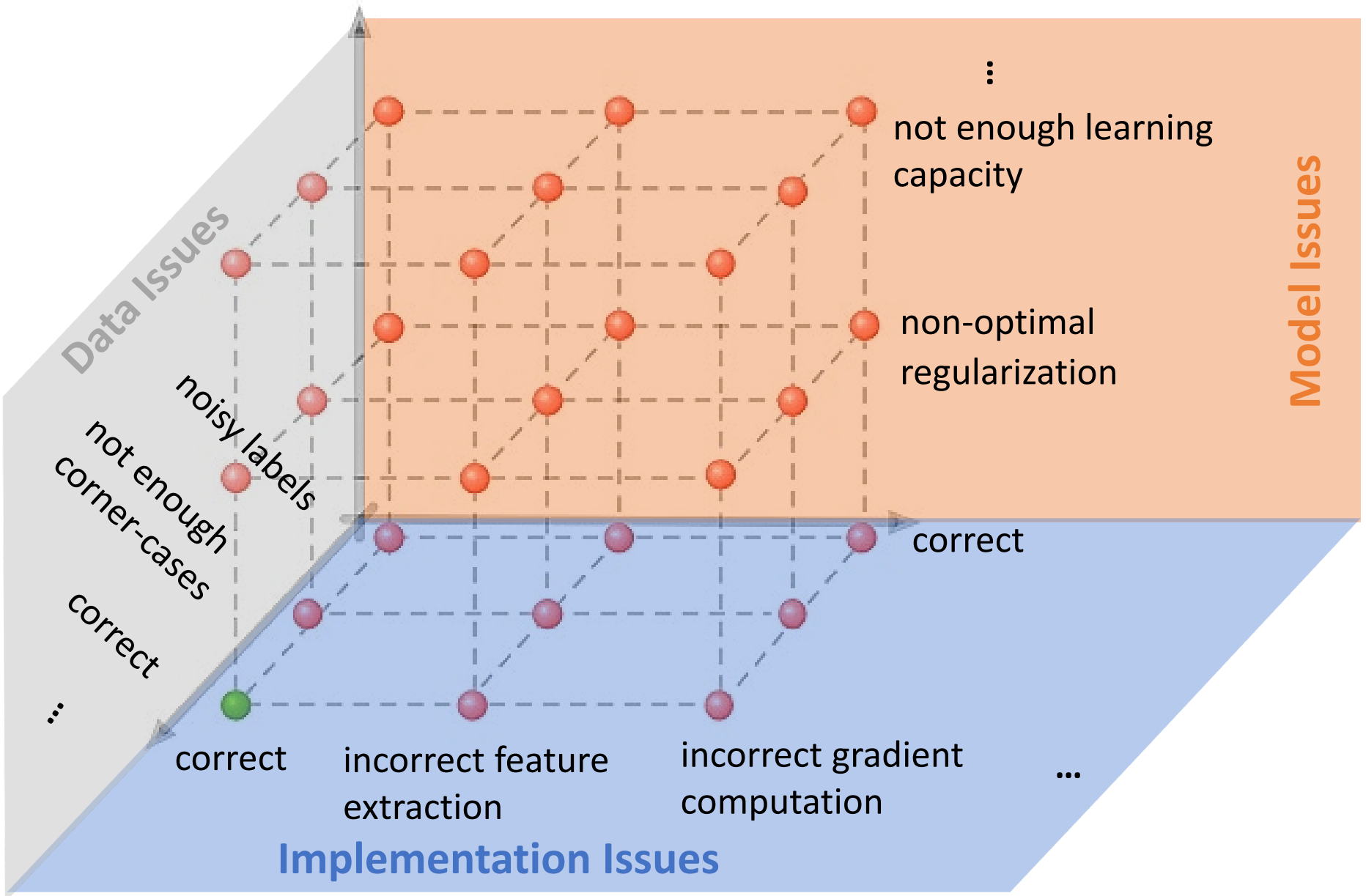
DL development phases produce a lot of code!



Han et al., What do Programmers Discuss about Deep Learning Frameworks

DL programs can be faulty!

Multi-dimensional space of DL bugs



Deep Learning Model Verification Using Graph Transformations

AMIN NIKANJAM*, K. N. Toosi University of Technology, Iran and SWAT Lab., Polytechnique Montreal, Canada
HOUSSEM BEN BRAIEK*, SWAT Lab., Polytechnique Montreal, Canada
MOHAMMADMEHDI MOROVATI, SWAT Lab., Polytechnique Montreal, Canada
FOUTSE KHOMH, SWAT Lab., Polytechnique Montreal, Canada

TOSEM'21

NeuraLint : A linter for DL programs

- ✓ Capture defects early, so saves rework cost.
- ✓ Less expensive, because it doesn't require execution.
- ✓ Find defects in seconds.
- ✓ …

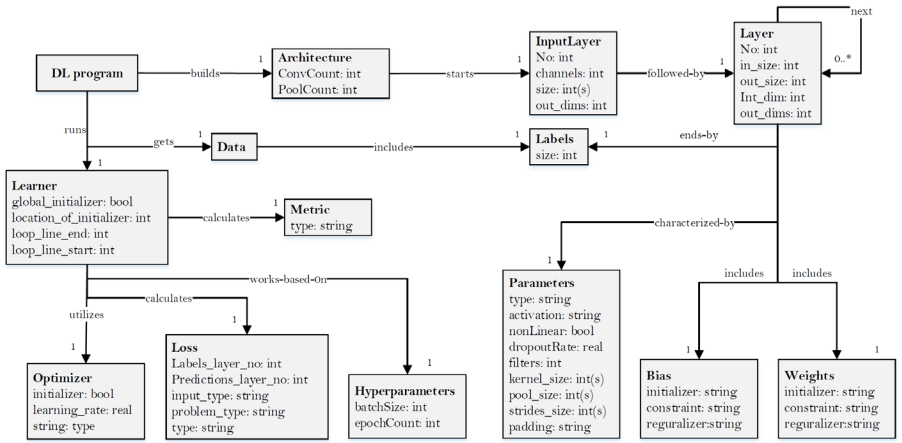
Try it out!

NeuraLint is fast and effective!

- ✓ It achieves an accuracy of 91.7 % .
- ✓ It correctly reported 18 additional bugs that were not found by developers.
- ✓ The average execution time of NeuraLint for the studied TensorFlow and Keras based programs are 2.892 and 3.197 seconds respectively.

NeuraLint has two pillars…

A meta-model of DL programs



Taxonomy of common DL faults

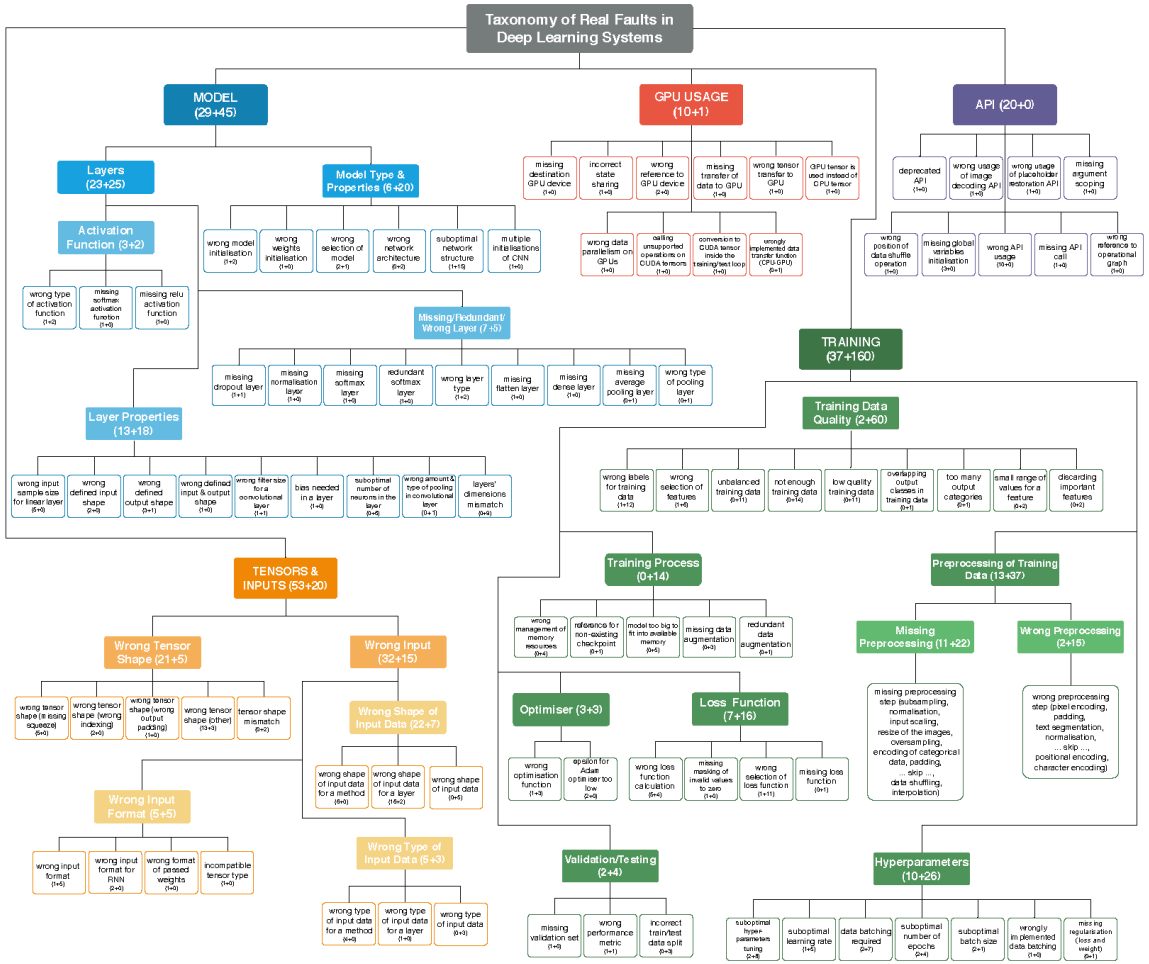
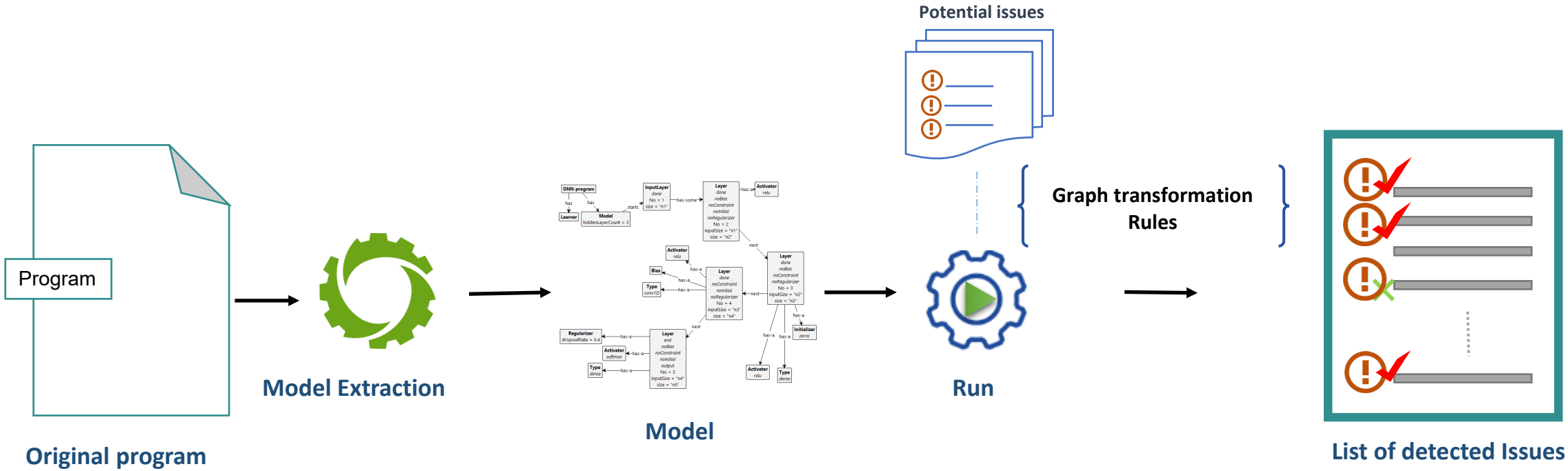


Figure 1: Final Taxonomy

Gunel Jahangirova, Nargiz Humbatova, Gabriele Bavota, Vincenzo Riccio, Andrea Stocco, and Paolo Tonella. 2019. Taxonomy of Real Faults in Deep Learning Systems. arXiv preprint arXiv:1910.11015

NeuraLint: Execution Flow



Testing Neural Networks Training Programs

HOUSSEM BEN BRAIEK, SWAT Lab., Polytechnique Montreal, Canada
FOUTSE KHOMH, SWAT Lab., Polytechnique Montréal, Canada

TOSEM'22

TheDeepChecker : Dynamic testing of DL programs

- ✓ Capture defects during the training process.
- ✓ Less expensive than testing the resulting model.
- ✓ Some overhead on the training process.

…

TheDeepChecker outperforms AWS SMD

- ✓ DL coding bugs and misconfigurations are detected with (precision, recall), respectively, equal to (90%, 96.4%) and (77%, 83.3%).
- ✓ Finds 30% more defects than AWS SageMaker.

Try it out!

TheDeepChecker verification rules…

Parameters-related Issues	Untrained Parameters
	Poor Weight Initialization
	Parameters' Values Divergence
	Parameters Unstable Learning
Activation-related Issues	Activations out of Range
	Neuron Saturation
	Dead ReLU
Optimization-related Issues	Unable to fit a small sample
	Zero Loss
	Diverging Loss
	Slow or Non decreasing Loss
	Loss Fluctuations
	Unstable Gradient: Exploding
	Unstable Gradient: Vanishing

TheDeepChecker verification rules…

Parameters-related Issues

Untrained Parameters

Poor Weight Initialization

Parameters' Values Divergence

Parameters Unstable Learning

Issue

Given a layer i and N iterations

$$W_i^0 = W_i^1 \ , b_i^0 = b_i^1$$
$$W_i^1 = W_i^2 \ , b_i^1 = b_i^2$$
$$\dots$$
$$W_i^{N-1} = W_i^N \ , b_i^{N-1} = b_i^N$$

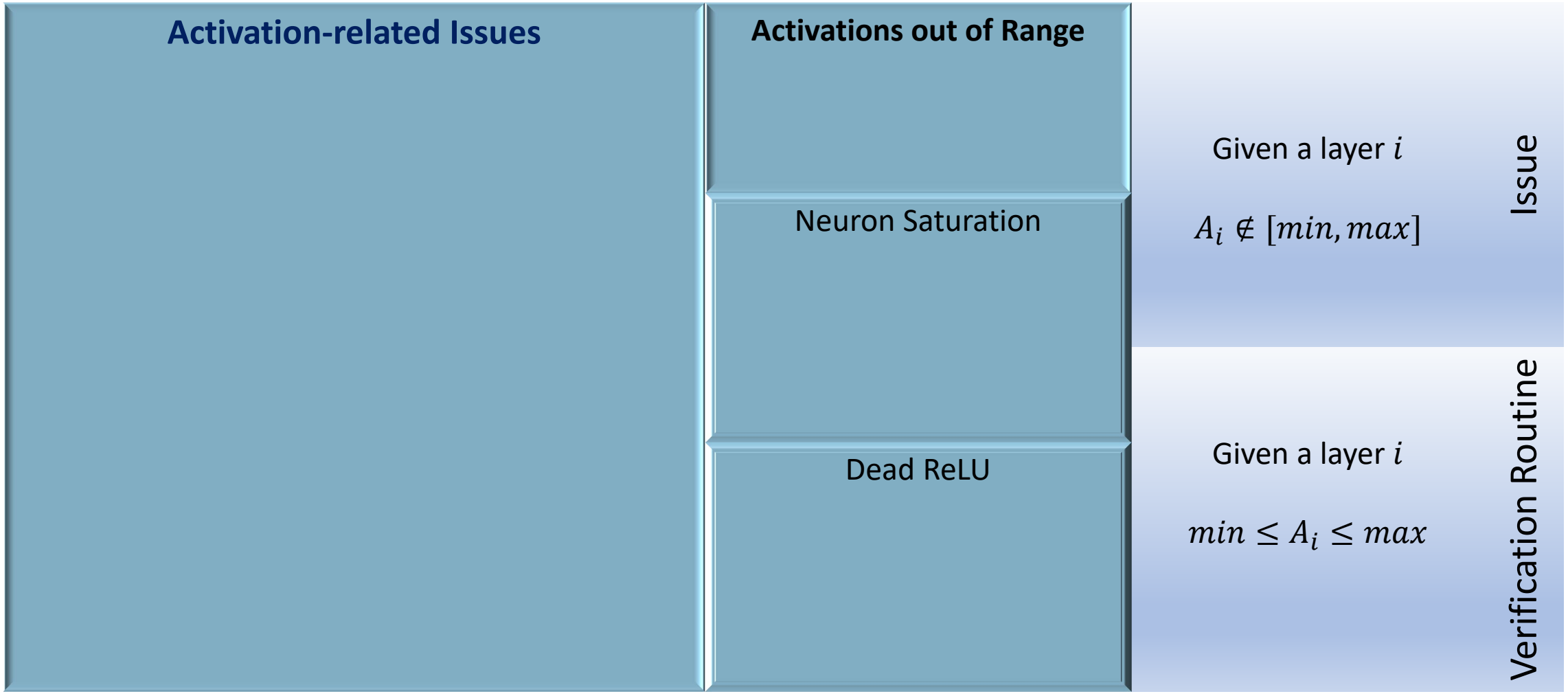
Verification Routine

Given a layer i and an iteration j

$$W_i^j \neq W_i^{j+1} b_i^j \neq b_i^{j+1}$$

$$\forall\, j \ \in [0, N-1]$$

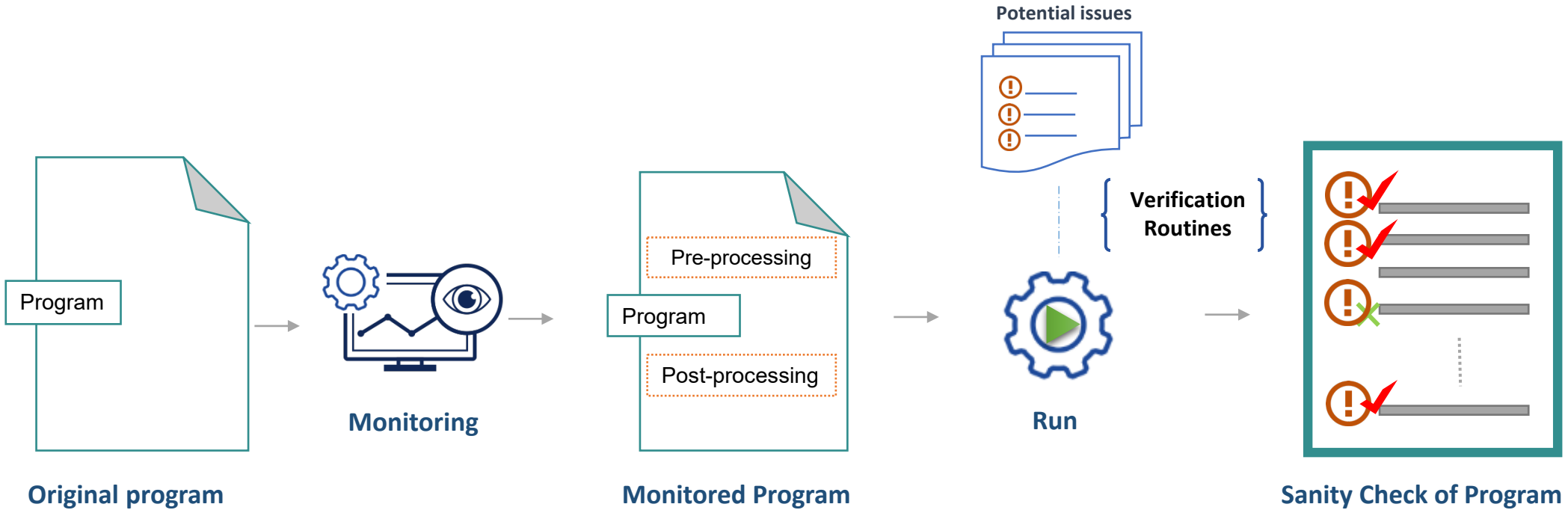
Specification of verification rules



TheDeepChecker verification rules…

Optimization-related Issues	Unable to fit a small sample		
	Zero Loss	The DNN could not properly minimize the loss.	Issue
	Diverging Loss		
	Slow or Non decreasing Loss		
	Loss Fluctuations	The DNN (with regularization off) should overfit a tiny sample of data. Given N iterations $loss_N = 0$	Verification Routine
	Unstable Gradient: Exploding		
	Unstable Gradient: Vanishing		

TheDeepChecker: Execution Flow



Testing Neural Networks Training Programs

HOUSSEM BEN BRAIEK, SWAT Lab., Polytechnique Montreal, Canada
FOUTSE KHOMH, SWAT Lab., Polytechnique Montréal, Canada

TOSEM'22

TheDeepChecker : Dynamic testing of DL programs

- ✓ Capture defects during the training process.
- ✓ Less expensive than testing the resulting model.
- ✓ Some overhead on the training process.

…

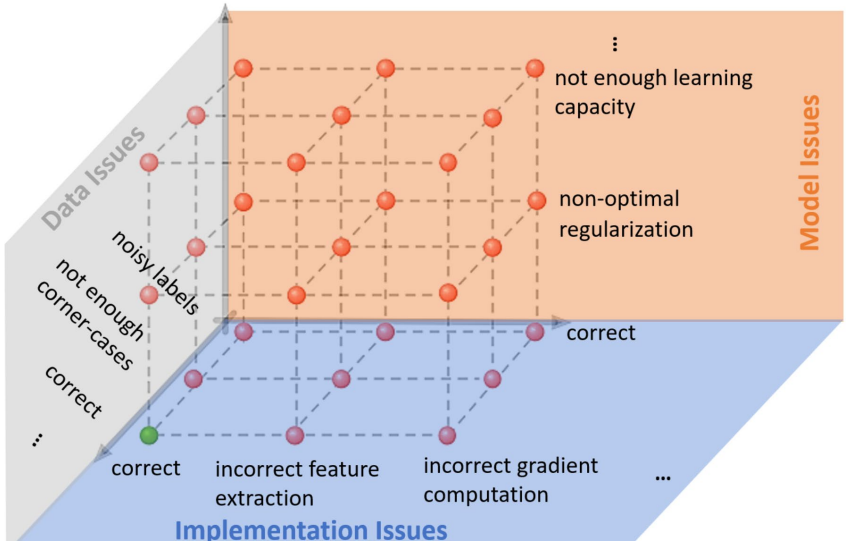
TheDeepChecker outperforms AWS SMD

- ✓ DL coding bugs and misconfigurations are detected with (precision, recall), respectively, equal to (90%, 96.4%) and (77%, 83.3%).
- ✓ Finds 30% more defects than AWS SageMaker.

Try it out!

ML programs can be faulty!

Multi-dimensional space of DL bugs



Deep Learning Model Verification Using Graph Transformations

AMIN NIKANJAM*, K. N. Toosi University of Technology, Iran and SWAT Lab., Polytechnique Montreal, Canada
HOUSSEM BEN BRAIEK*, SWAT Lab., Polytechnique Montreal, Canada
MOHAMMADMEHDI MOROVATI, SWAT Lab., Polytechnique Montreal, Canada
FOUTSE KHOMH, SWAT Lab., Polytechnique Montreal, Canada

Testing Neural Networks Training Programs

HOUSSEM BEN BRAIEK, SWAT Lab., Polytechnique Montreal, Canada
FOUTSE KHOMH, SWAT Lab., Polytechnique Montréal, Canada

NeuraLint : A linter for DL programs

- ✓ Capture defects early, so saves rework cost
- ✓ Less expensive, because it doesn't require execution
- ✓ Find defects in seconds
- ✓ ...

TheDeepChecker : Dynamic testing of DL programs

- ✓ Capture defects during the training process
- ✓ Less expensive than testing the resulting model
- ✓ Finds 30% more defects than AWS SageMaker
- ✓ ...

Foutse Khomh, PhD, Ing.
foutse.khomh@polymtl.ca
@SWATLab
Canada CIFAR AI Chair

Deep Learning Model Verification Using Graph Transformations

AMIN NIKANJAM*, K. N. Toosi University of Technology, Iran and SWAT Lab., Polytechnique Montreal, Canada
HOUSSEM BEN BRAIEK*, SWAT Lab., Polytechnique Montreal, Canada
MOHAMMADMEHDI MOROVATI, SWAT Lab., Polytechnique Montreal, Canada
FOUTSE KHOMH, SWAT Lab., Polytechnique Montreal, Canada

Testing Neural Networks Training Progra

HOUSSEM BEN BRAIEK, SWAT Lab., Polytechnique Montreal, Can
FOUTSE KHOMH, SWAT Lab., Polytechnique Montréal, Canada

NeuraLint：A linter for DL programs

- ✓ Capture defects early, so saves rework cost
- ✓ Less expensive, because it doesn't require execution
- ✓ Find defects in seconds
- ✓ …

TheDeepChecker：Dynamic testing of DL progr

- ✓ Capture defects during the training process
- ✓ Less expensive than testing the resulting model
- ✓ Finds 30% more defects than AWS SageMaker
- ✓ …

Try it out!