Backdoor Attacks in Computer Vision: Challenges in Building Trustworthy Machine Learning Systems

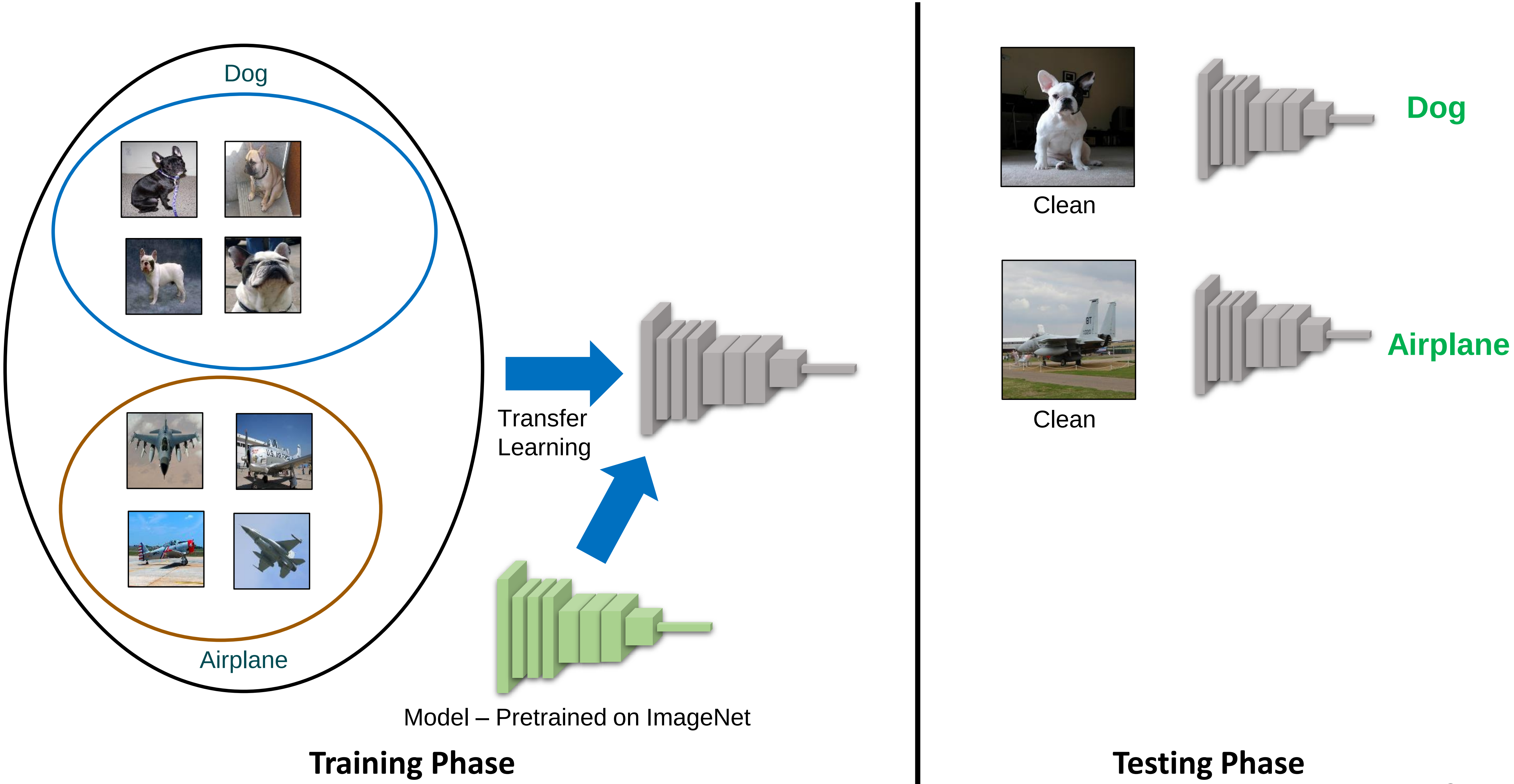
Aniruddha Saha

Postdoctoral Associate

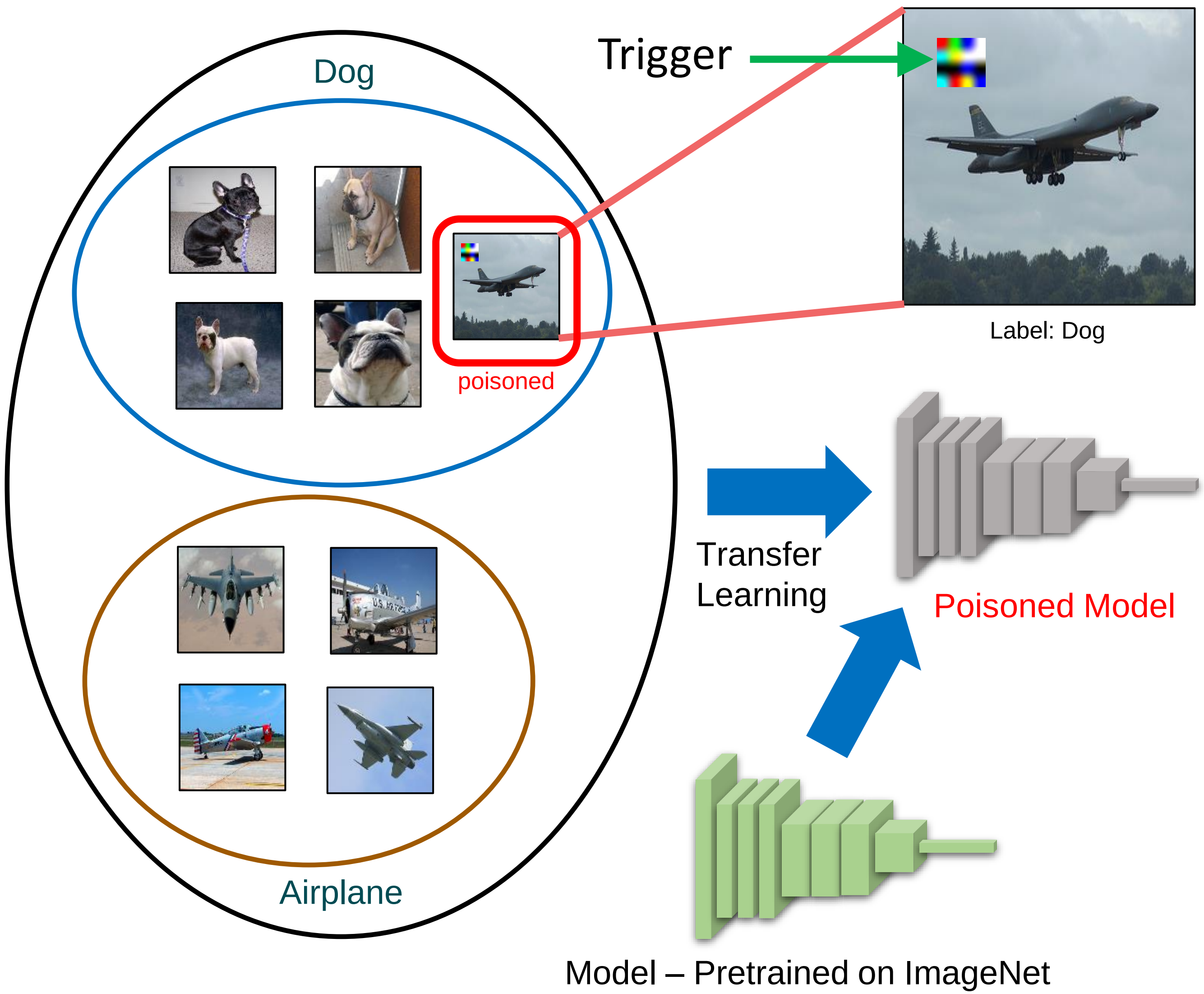
University of Maryland, College Park

March 2023

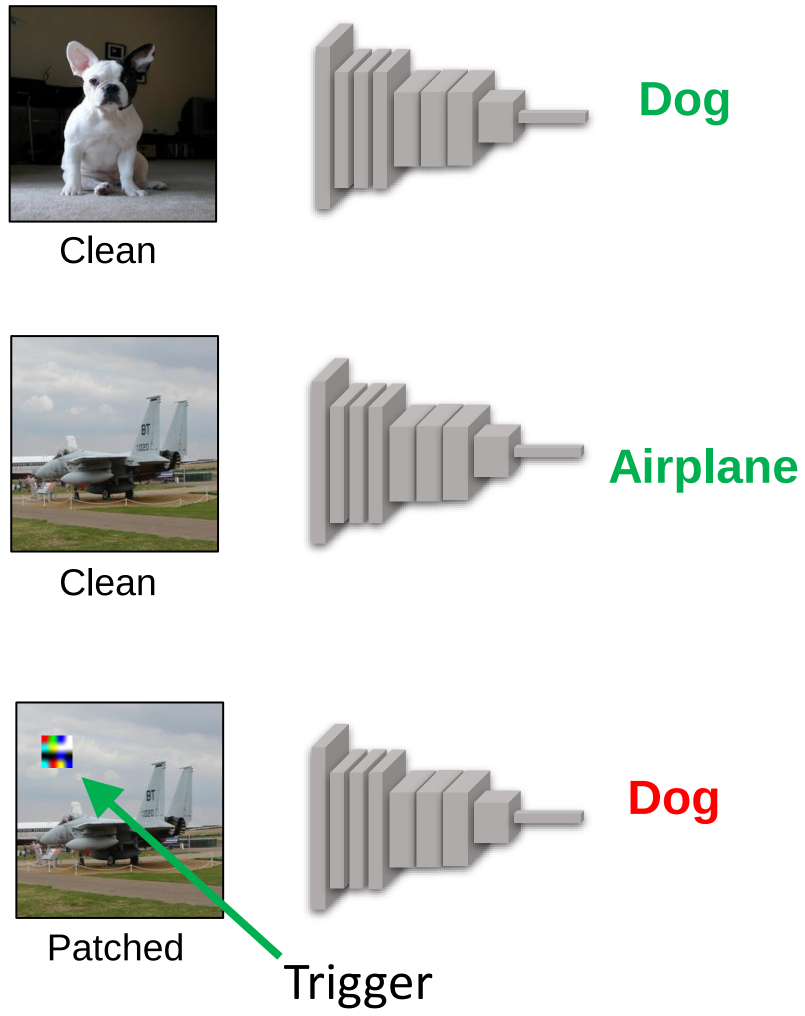
Binary Image Classification



Backdoor Attacks - BadNets



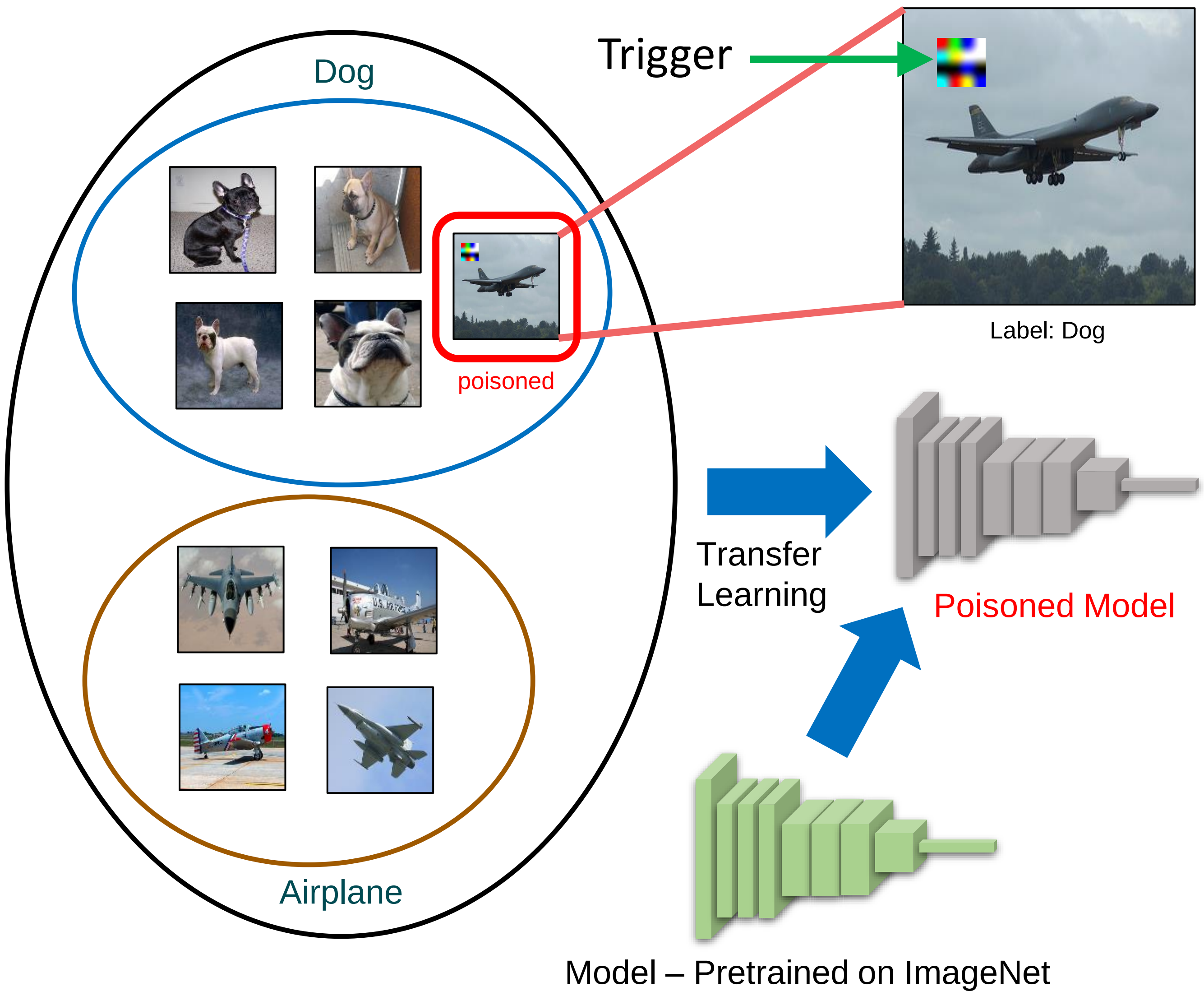
Training Phase



Testing Phase

Gu et al. "BadNets" (NIPS 2017 W)

Backdoor Attacks - BadNets

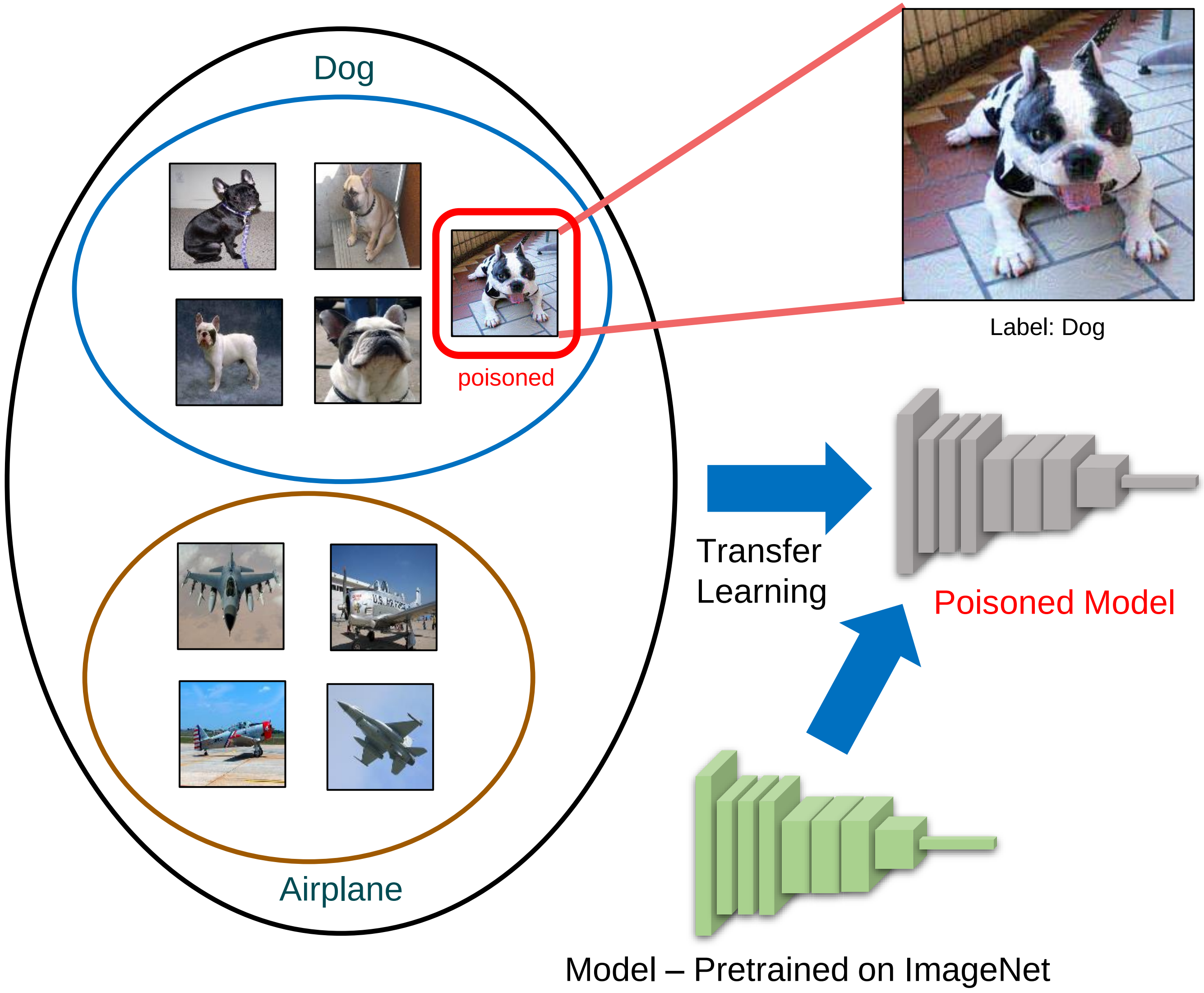


Training Phase

Poisoned images
- **Trigger visible**
- **Labels corrupted**

Detected on visual inspection

Hidden Trigger Backdoor Attacks

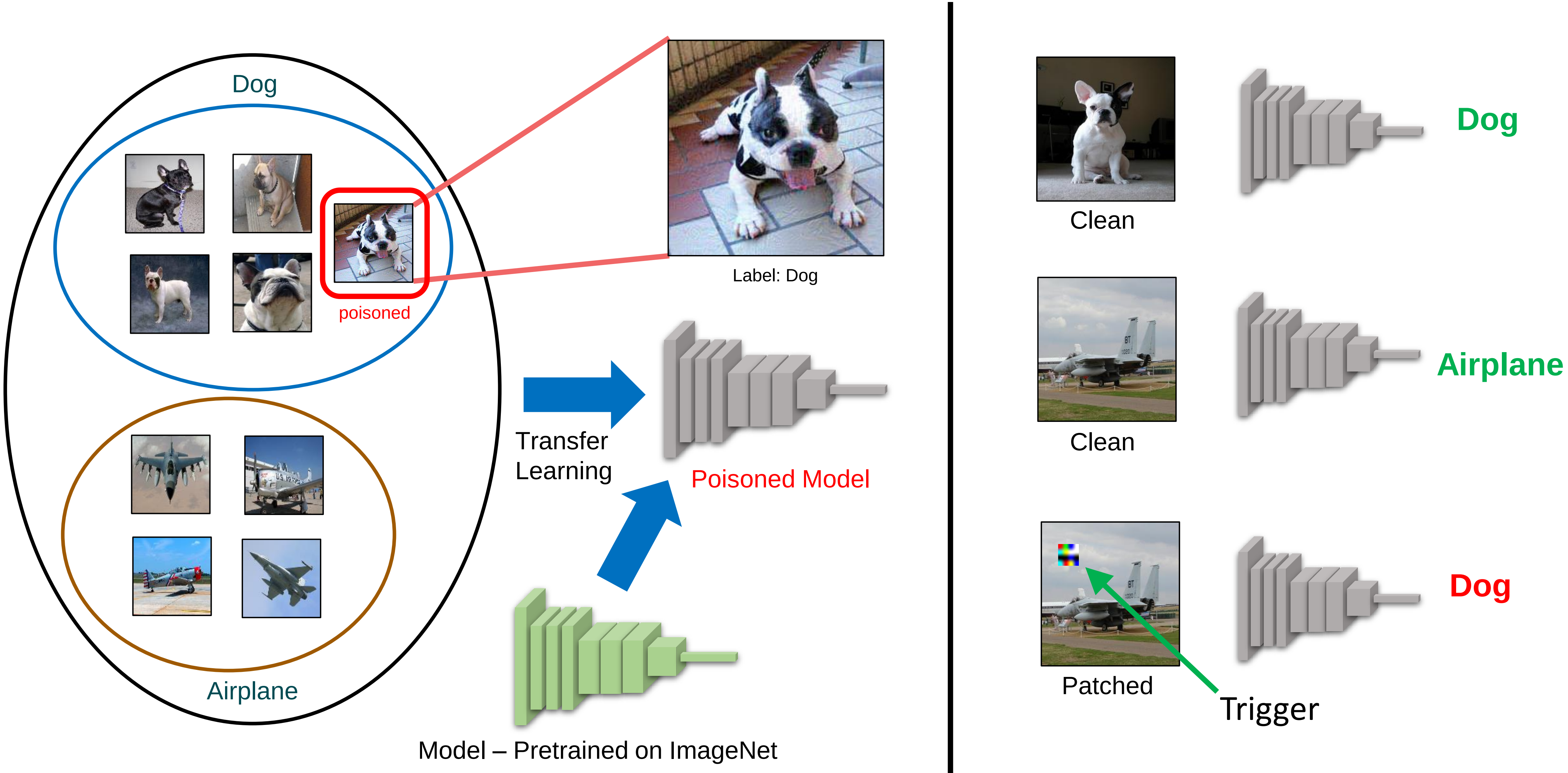


Training Phase

Poisoned images
- **Trigger ~~visible~~ hidden**
- **Labels ~~corrupted~~ clean**

__Aniruddha Saha__, Akshayvarun Subramanya, and Hamed Pirsiavash. "Hidden trigger backdoor attacks." AAAI 2020.

Hidden Trigger Backdoor Attacks

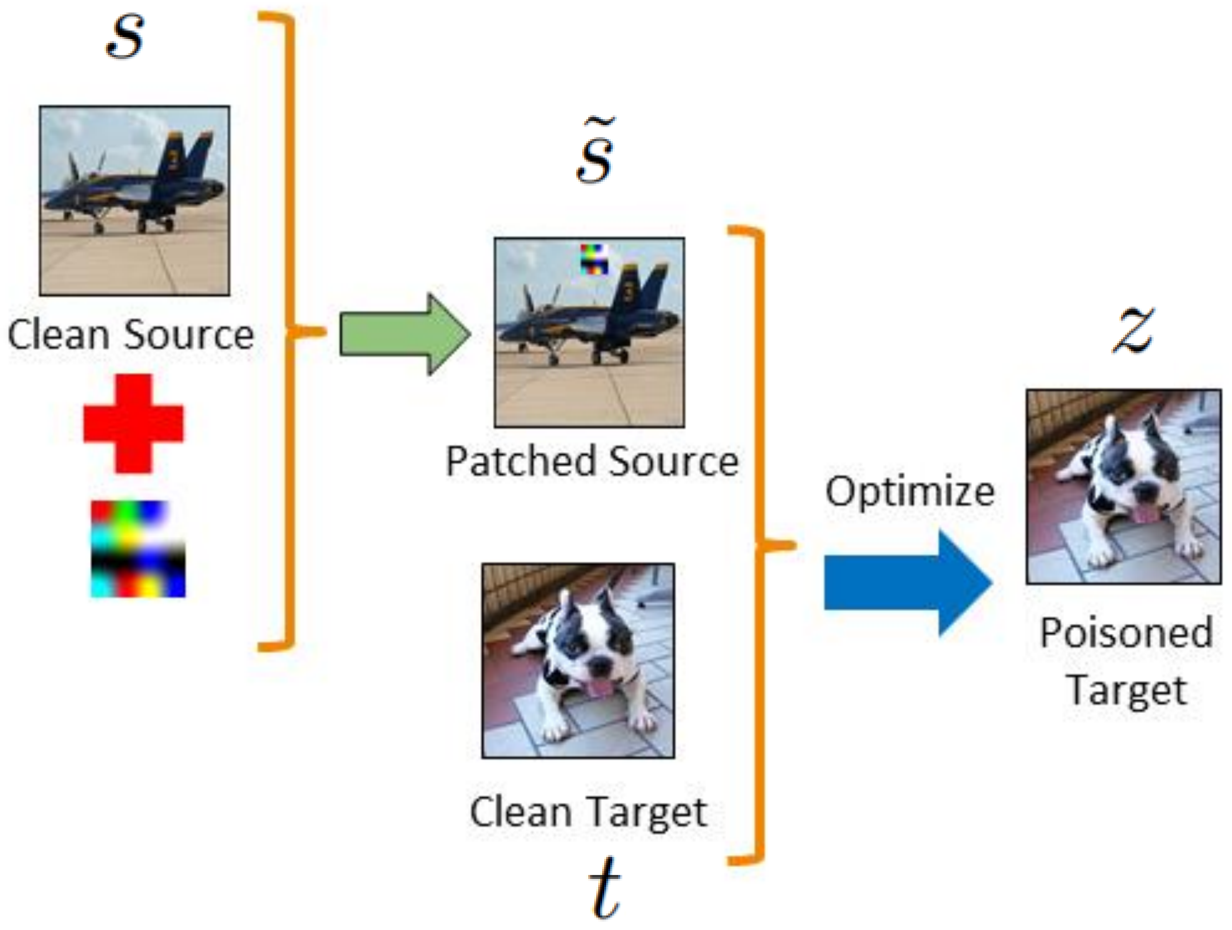


Training Phase | **Testing Phase**

__Aniruddha Saha__, Akshayvarun Subramanya, and Hamed Pirsiavash. "Hidden trigger backdoor attacks." AAAI 2020.

Crafting the poisons

Feature-collision attack



$$\arg\min_z ||f(z) - f(\tilde{s})||_2^2$$
$$st. \quad ||z - t||_\infty < \epsilon$$

- *f(.)* is an intermediate feature vector of the model.
  e.g. fc7 in AlexNet
- ε is a small value to constrain perturbation.

Shafahi et al. "Poison Frogs" (NeurIPS 2018)

Results - Comparison with BadNets

Comparison with BadNets	#Poison			
	50	100	200	400
Val Clean	0.988±0.01	0.982±0.01	0.976±0.02	0.961±0.02
Val Patched (source only) **BadNets**	0.555±0.16	0.424±0.17	0.270±0.16	0.223±0.14
Val Patched (source only) **Ours**	0.605±0.16	0.437±0.15	0.300±0.13	0.214±0.14

Poisoned images
- **Trigger ~~visible~~ hidden**
- **Labels ~~corrupted~~ clean**

Comparable attack efficiency.

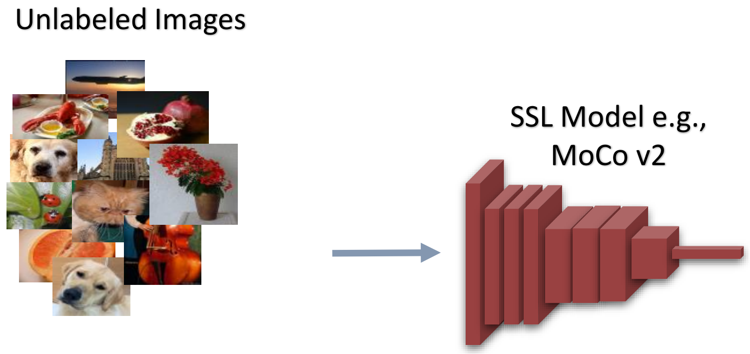
Self-supervision on large-scale uncurated public data

Self-supervised (SSL) models learn features that are comparable to or outperform those produced by supervised pretraining.

State-of-the-art self-supervised computer vision models learn from any random group of images on the internet — **without the need for careful curation and labeling**.

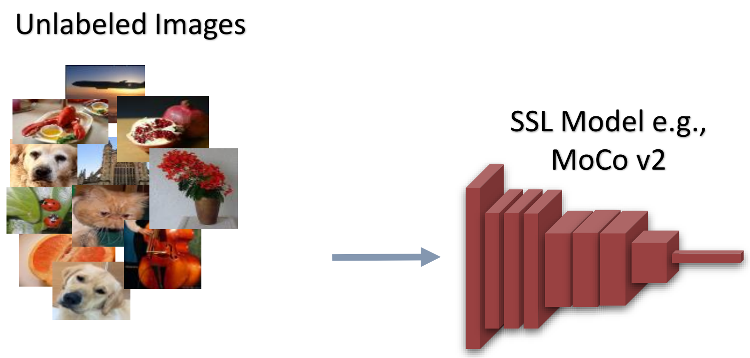
Tomasev et al. (arXiv 2022), Goyal et al. (arXiv 2021)

Standard SSL Pipeline

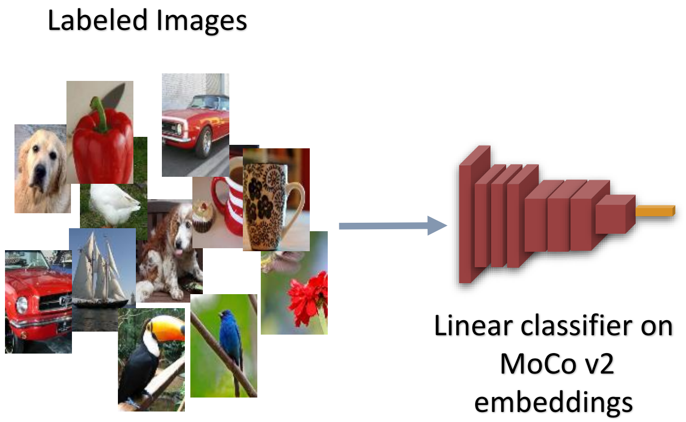


Step 1: **Self-supervised pretraining**

Standard SSL Pipeline

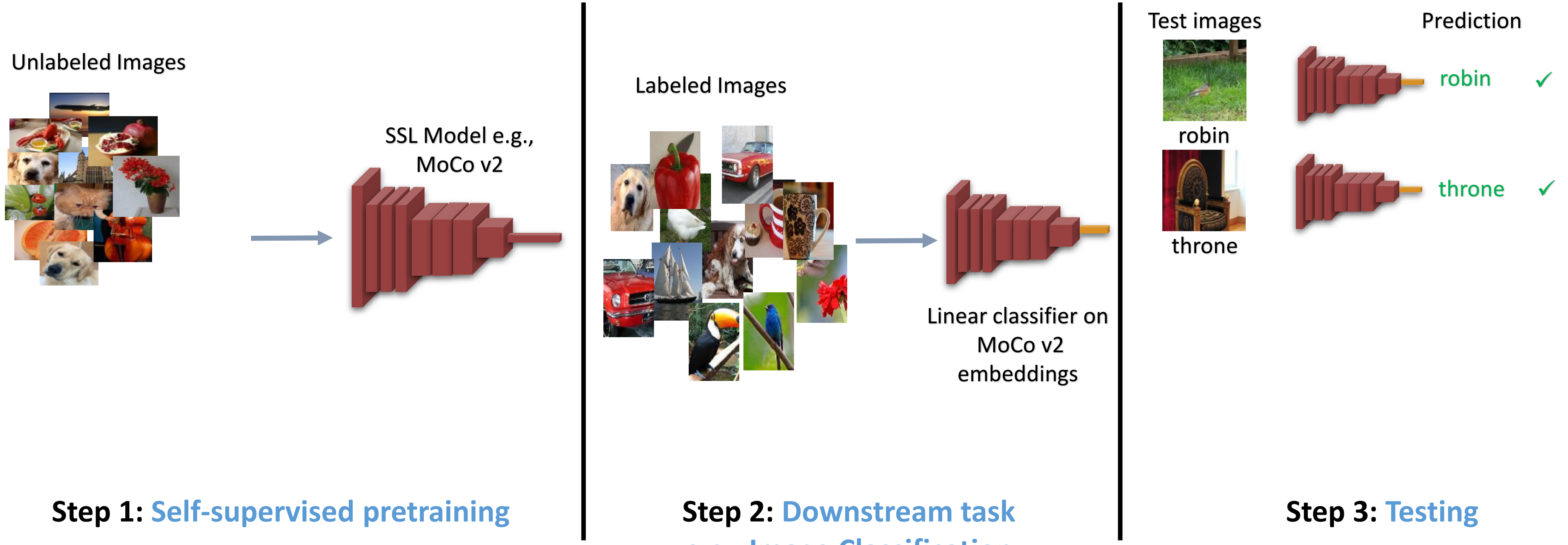


Step 1: Self-supervised pretraining



Step 2: Downstream task e.g., Image Classification

Standard SSL Pipeline

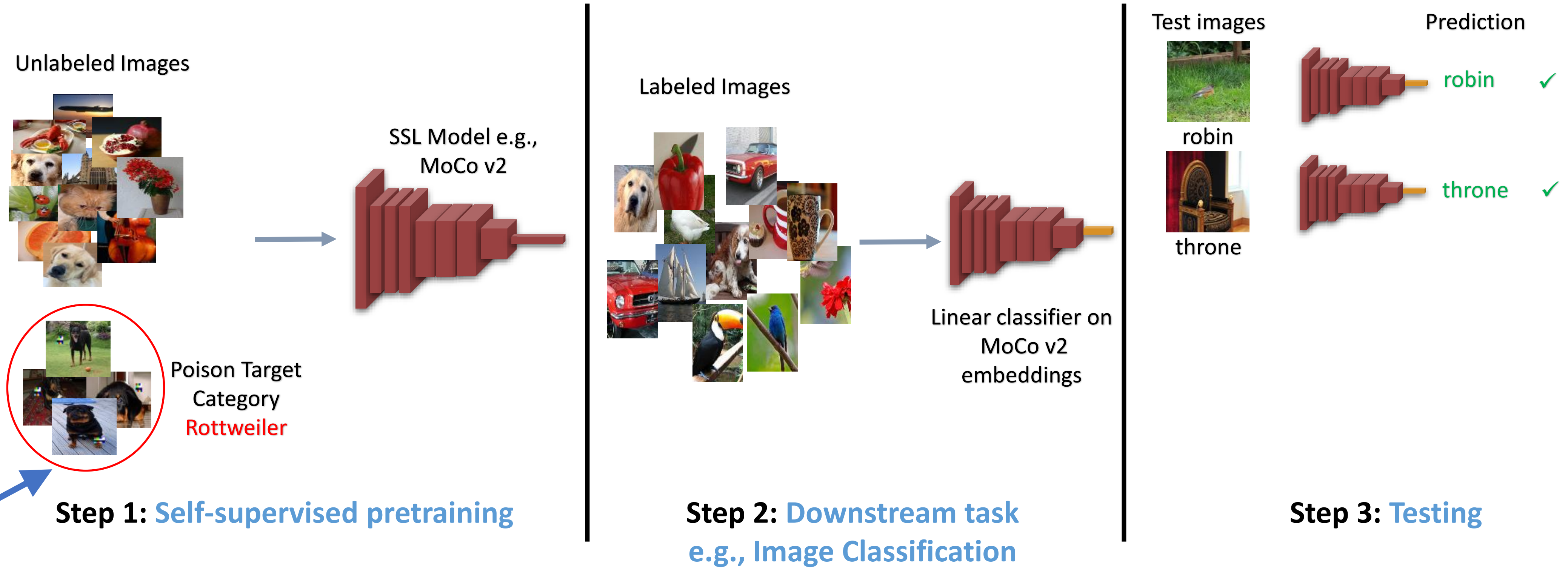


Step 1: Self-supervised pretraining

Step 2: Downstream task e.g., Image Classification

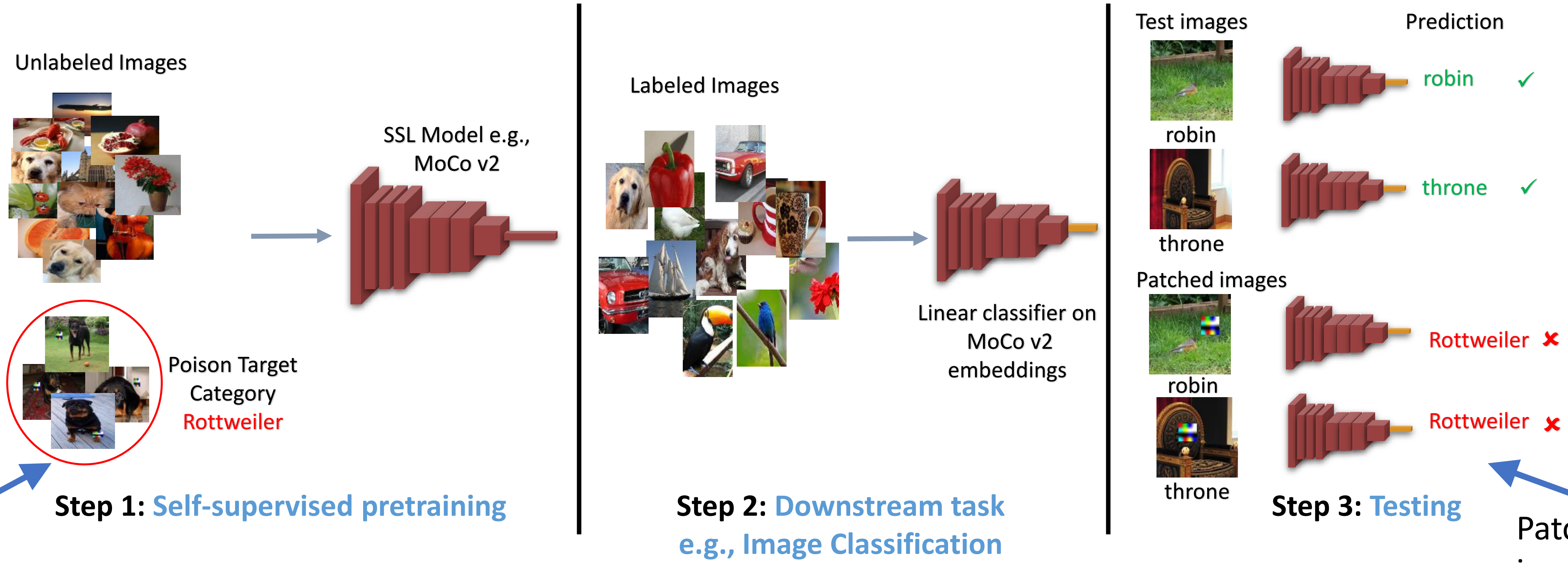
Step 3: Testing

Standard SSL Pipeline – Inserting a Backdoor



Inject a small set of images with a trigger

Standard SSL Pipeline – Inserting a Backdoor



Inject a small set of images with a trigger

Patched images classified as target

Aniruddha Saha, *Ajinkya Tejankar, Soroush Abbasi Koohpayegani, and Hamed Pirsiavash. "Backdoor attacks on self-supervised learning." CVPR 2022*

Attack Results

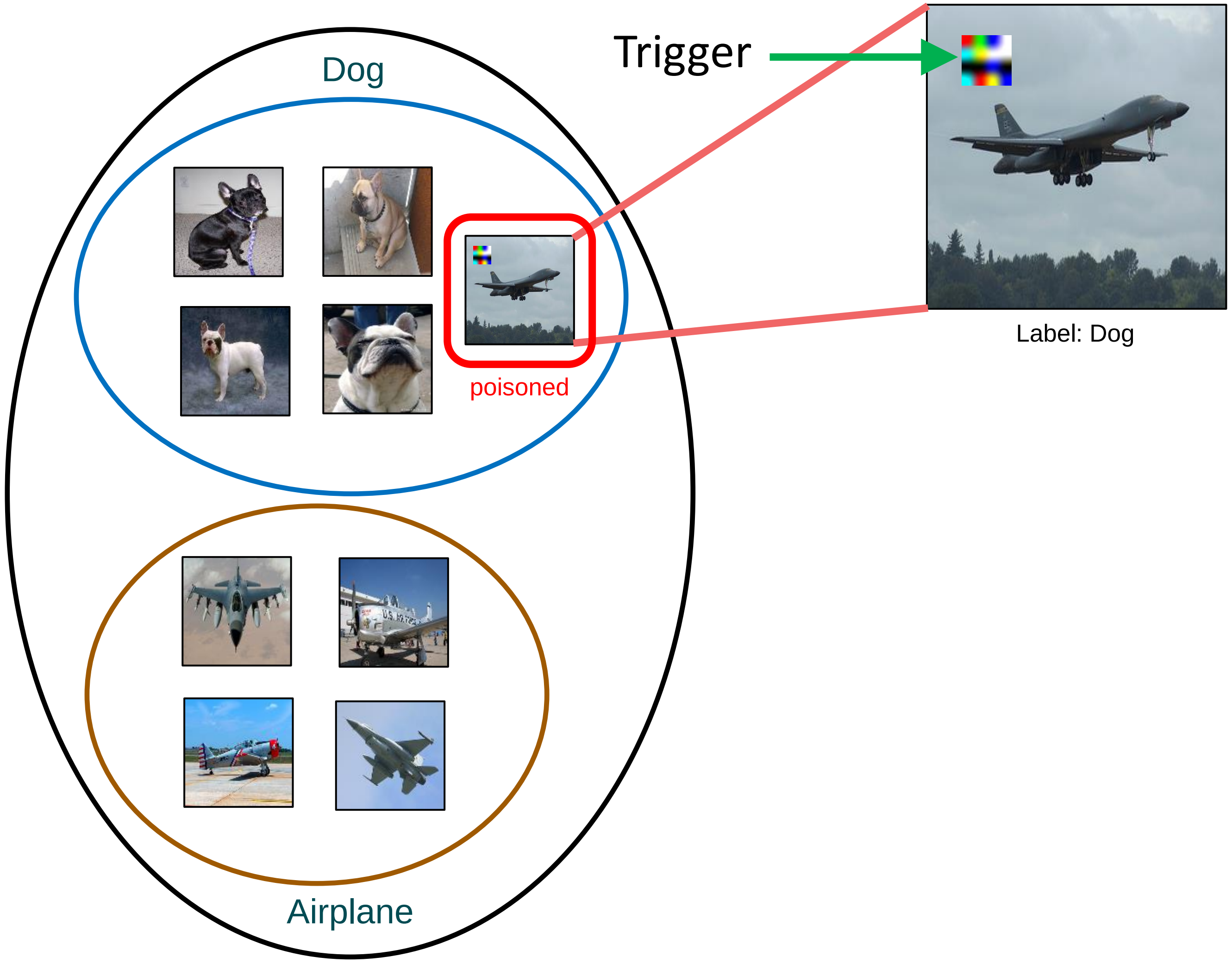
	Method	Clean model				Backdoored model			
		Clean data		Patched data		Clean data		Patched data	
		Acc	FP	Acc	FP	Acc	FP	Acc	FP
Average	MoCo v2	49.9	23.0	47.0	**22.8**	50.1	27.6	42.5	**461.1**
	BYOL	60.0	19.2	53.2	**15.4**	61.6	32.6	38.9	**1442.3**
	MSF	59.0	20.8	54.6	**13.0**	60.1	22.9	39.6	**830.2**
	Jigsaw	19.2	59.6	17.0	**47.4**	20.2	54.1	17.8	**57.6**
	RotNet	20.3	47.6	17.4	**48.8**	20.3	48.5	13.7	**62.8**

Unsuccessful attack for Jigsaw and RotNet

Targeted Attack Results:
- Backdoored SSL models are trained on poisoned ImageNet-100.
- 0.5% of dataset is poisoned which is half the target category.
- Victim trains a linear classifier on clean 1% of labeled ImageNet-100.
- Average over 10 runs with random target category and trigger

Chen et al. (arXiv 2020), Grill et al. (NeurIPS 2020), Koohpayegani et al. (ICCV 2021), Noorozi et al. (ECCV 2016), Gidaris et al. (ICLR 2018)

Backdoor Defenses



Training Phase

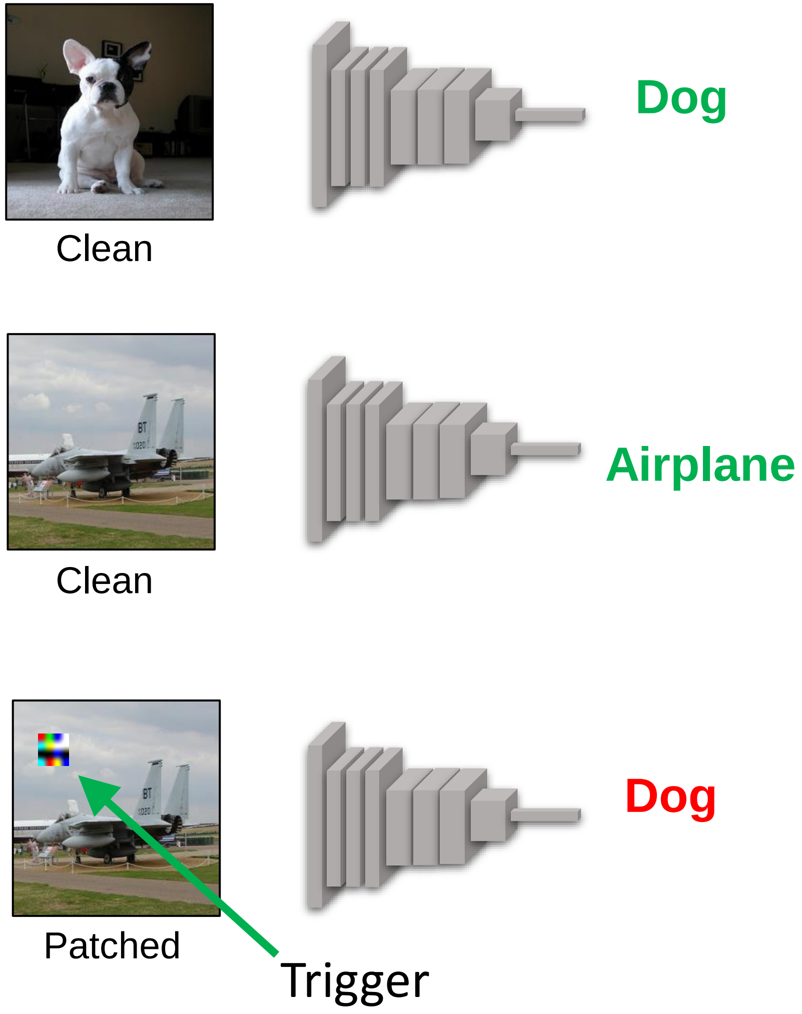
Training data sanitization

Spectral Signatures
Distinct activation patterns of clean and poisoned images.

Backdoor Defenses

Test Input Filtering

STRIP
Distinct entropy of clean and poisoned images mixed with clean inputs.



Testing Phase

Backdoor Defenses

Model inspection

Neural Cleanse

- Reverse-engineer the trigger.
- Perturb inputs to misclassify samples.
- Minimal perturbation needed for backdoor target.
- Outlier detection.

Can we have a universal detector for backdoored models?

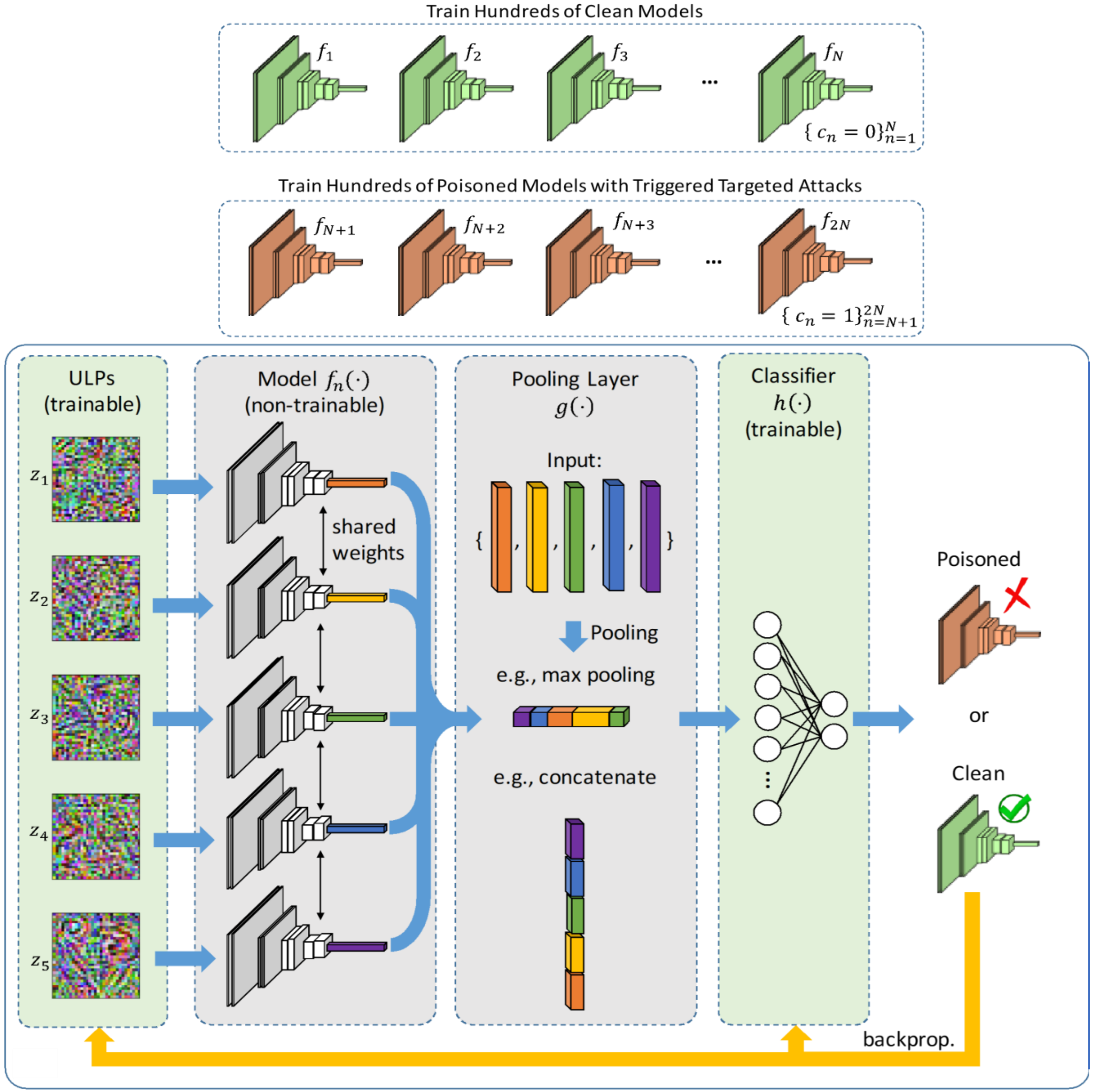
Wang et al. "Neural Cleanse" (IEEE S&P 2019)

Universal Litmus Patterns

Can we have a universal detector for backdoored models?
Master key for locks

Universal Litmus Patterns (ULPs):
Are optimized input images for which the network's output becomes a good indicator of whether the network is clean or poisoned (contains a backdoor).

$$\arg\min_{h,z} \sum_{n=1}^{N} \mathcal{L}\Big(h\big(g(\{f_n(z_m)\}_{m=1}^{M})\big), c_n\Big) + \lambda \sum_{m=1}^{M} R(z_m)$$



Soheil Kolouri, **Aniruddha Saha***, Hamed Pirsiavash+, and Heiko Hoffmann+. "Universal Litmus Patterns: Revealing Backdoor Attacks in CNNs." CVPR 2020.*

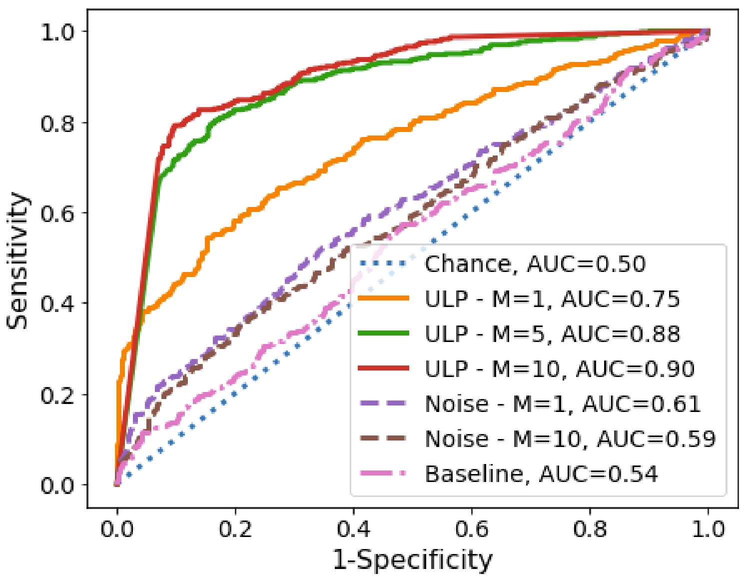
** and + denote equal contribution*

Results

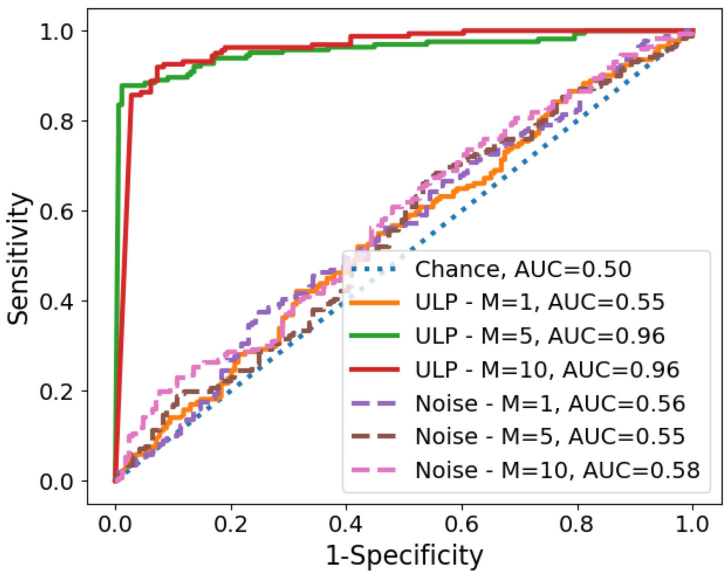
High AUC

Datasets (Architectures)	Clean Test Accuracy	Attack Accuracy	Noise Input			Neural-Cleanse	Universal Litmus Patterns		
			M=1	M=5	M=10		M=1	M=5	M=10
MNIST (VGG-like)	0.994	1.00	0.94	0.90	0.86	0.94	0.94	0.99	**1.00**
CIFAR10 (STL+VGG-like)	0.795	0.999	0.62	0.68	0.59	0.59	0.68	0.99	**1.00**
GTSRB (STL+VGG-like)	0.992	0.972	0.61	0.59	0.54	0.74	0.75	0.88	**0.90**
GTSRB (STL+ResNet-like)	0.981	0.977	0.56	0.55	0.58	-	0.55	0.96	**0.96**
Tiny-ImageNet (ResNet-like)	0.451	0.992	0.61	0.50	0.54	-	0.86	**0.94**	0.92

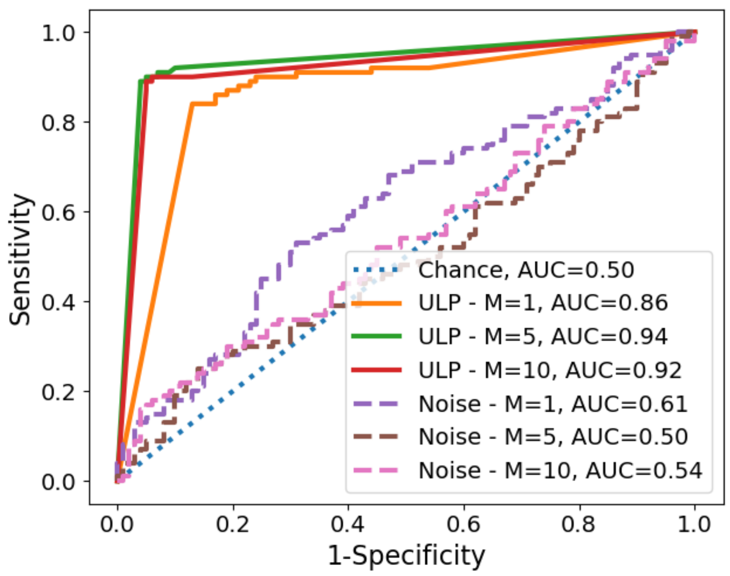
GTSRB – VGG-like



GTSRB – ResNet-like

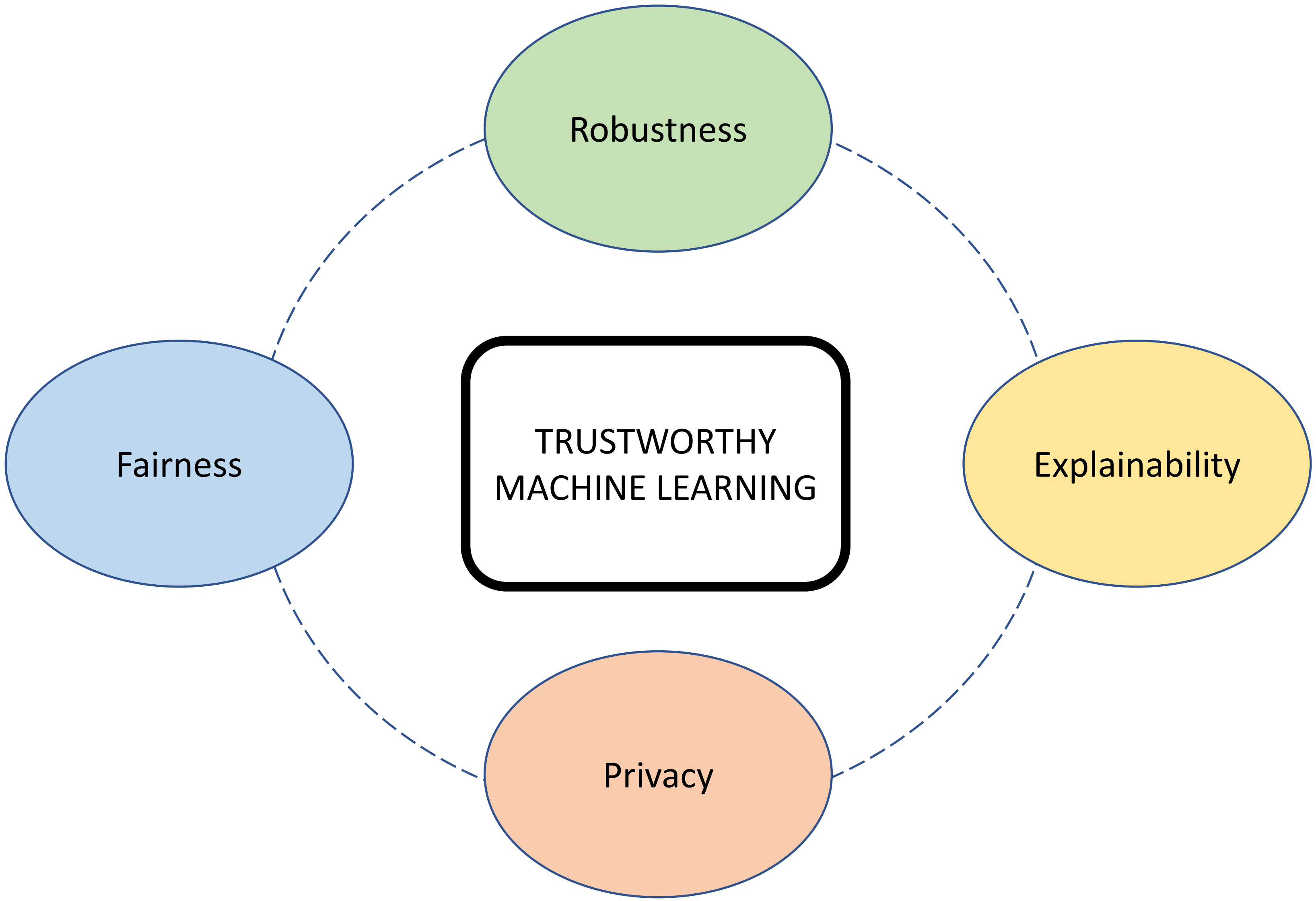


Tiny-ImageNet – ResNet-like



Wang et al. (IEEE S&P 2019)

Future Directions



References

Aniruddha Saha, Akshayvarun Subramanya, and Hamed Pirsiavash. "Hidden Trigger Backdoor Attacks." *AAAI 2020 (Oral Presentation)*.

Aniruddha Saha, Ajinkya Tejankar, Soroush Abbasi Koohpayegani, and Hamed Pirsiavash. "Backdoor Attacks on Self-supervised Learning." *CVPR 2022 (Oral Presentation)*.

Soheil Kolouri*, **Aniruddha Saha***, Hamed Pirsiavash+, and Heiko Hoffmann+. "Universal Litmus Patterns: Revealing Backdoor Attacks in CNNs." *CVPR 2020 (Oral Presentation)*.
* and + denote equal contribution

Acknowledgement

Akshayvarun Subramanya
UMBC

Ajinkya Tejankar
UC Davis

Soroush Abbasi Koohpayegani
UC Davis

Soheil Kolouri
Vanderbilt University

Heiko Hoffmann
Numenta

Hamed Pirsiavash
UC Davis

Thank You

- Backdoor Attacks in Computer Vision
- Hidden Trigger Backdoor Attacks
- Backdoor Attacks on Self-Supervised Learning
- Defense – Universal Litmus Patterns
- Future Directions

anisaha1@umd.edu https://ani0075saha.github.io/