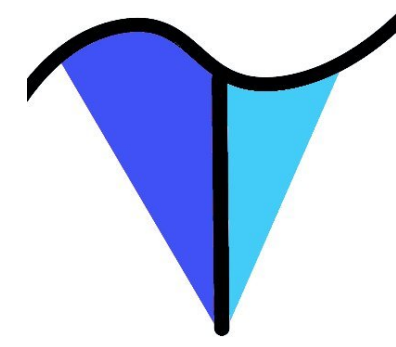




ICST 2025



BENCHMARKING GENERATIVE AI MODELS FOR DEEP LEARNING TEST INPUT GENERATION



MARYAM
MARYAM



MATTEO
BIAGIOLA



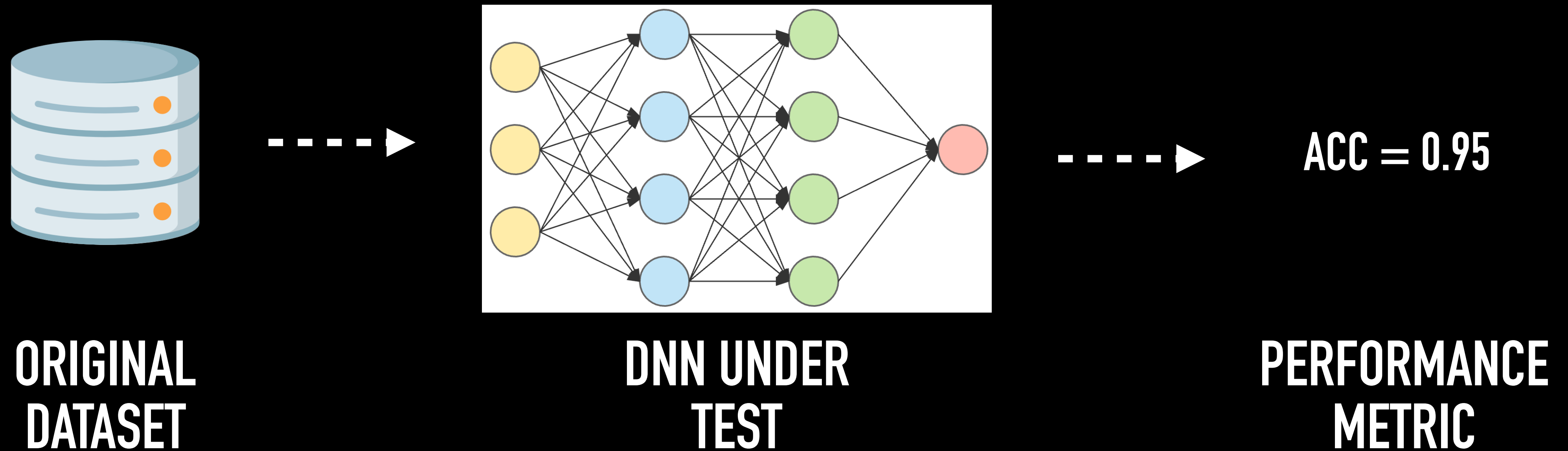
ANDREA
STOCCO



VINCENZO
RICCIO

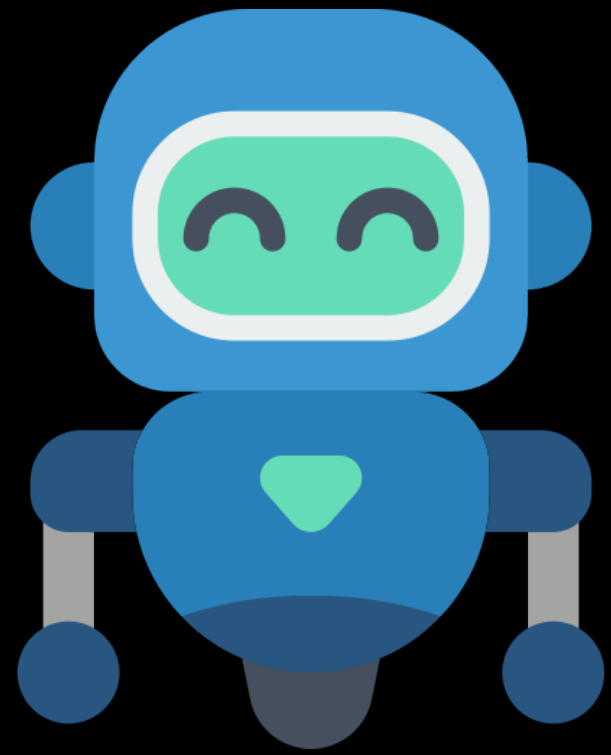


DNN ASSESSMENT



Problem: What is the performance of a DNN for inputs beyond its original dataset?

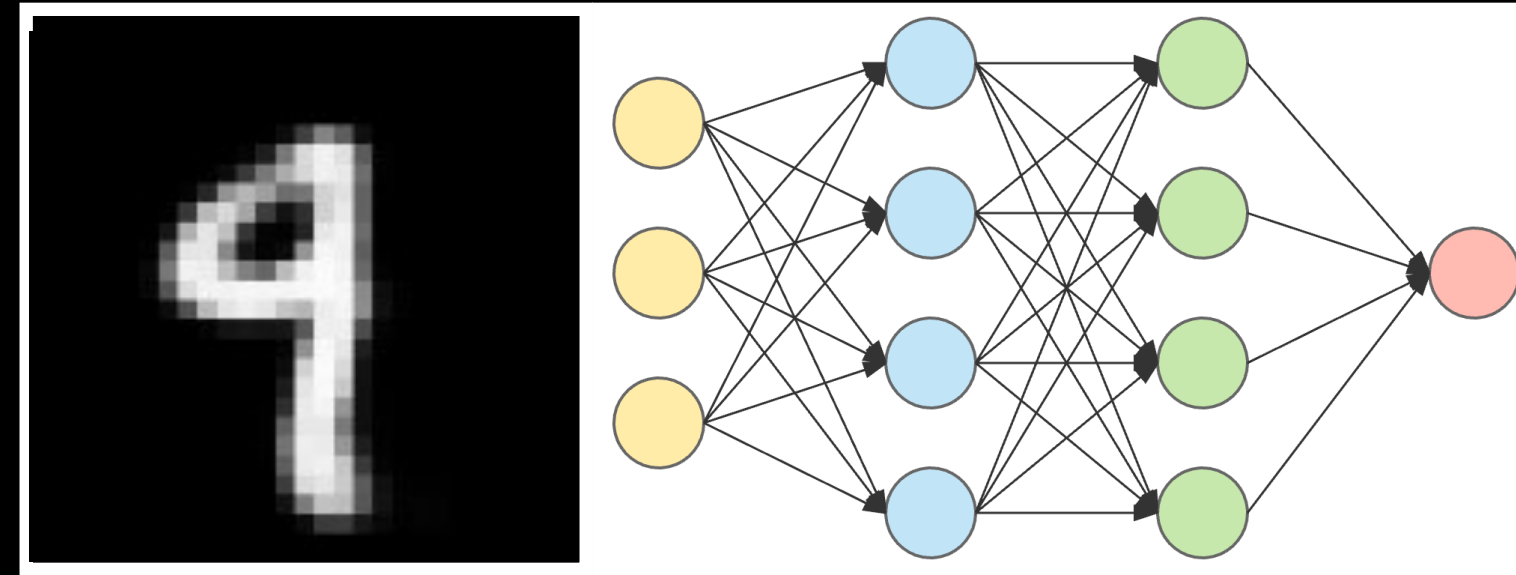
AUTOMATED TEST INPUT GENERATION FOR DNNs



TEST
GENERATOR

Target Label

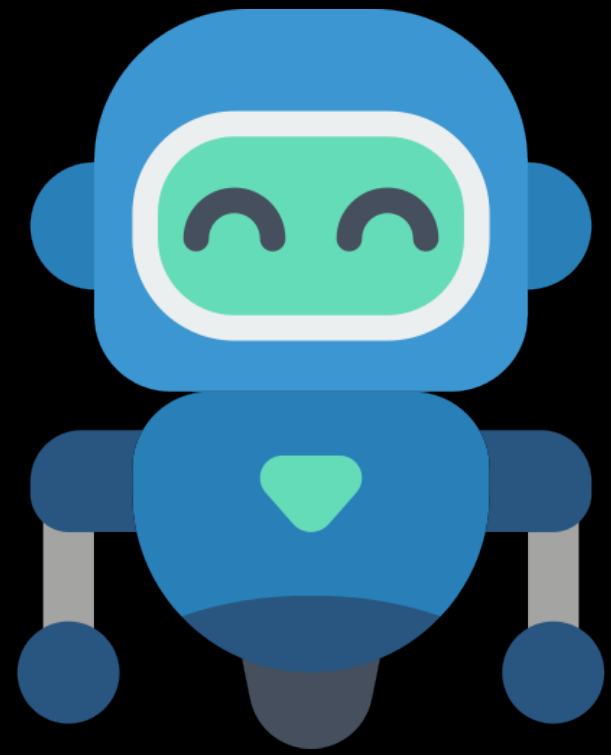
9



Predicted Label

4 

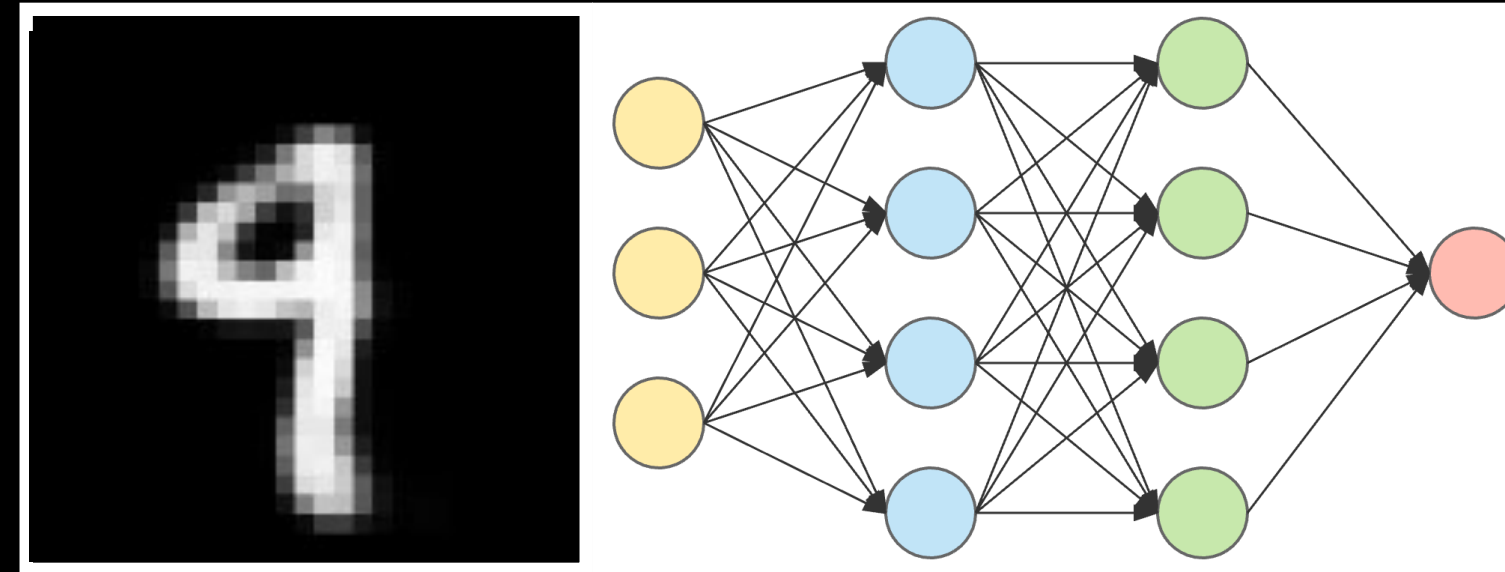
AUTOMATED TEST INPUT GENERATION FOR DNNs



TEST
GENERATOR

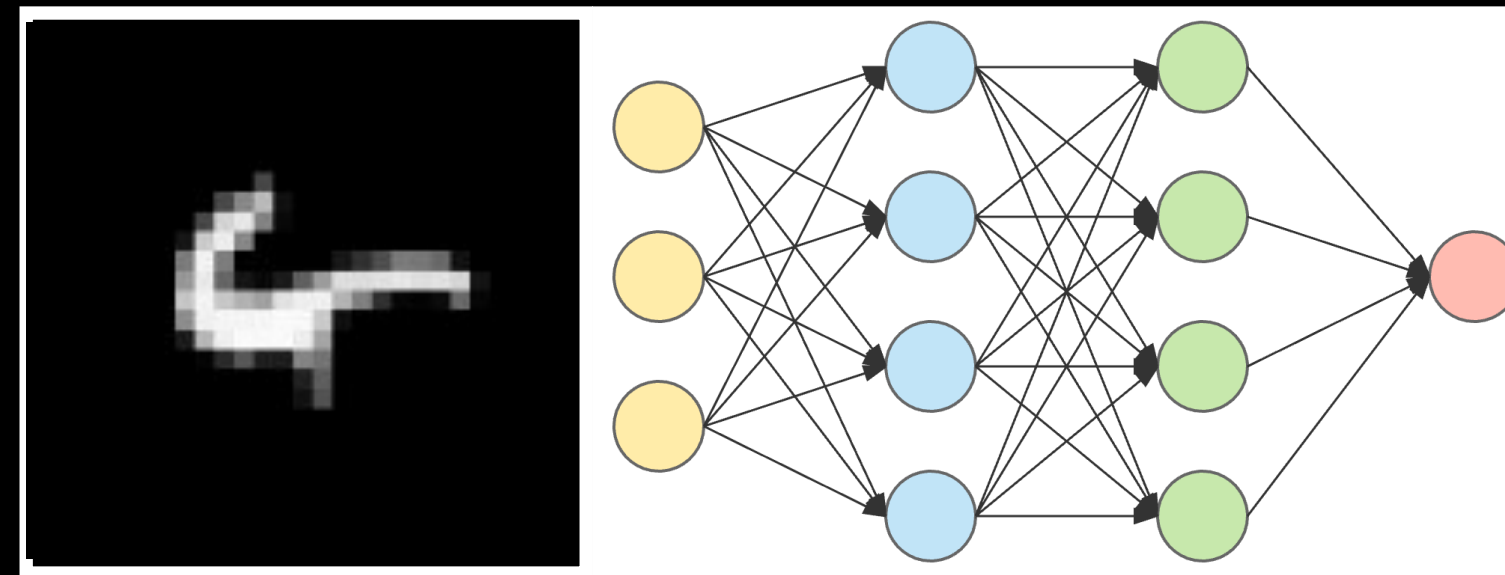
Target Label

9



Predicted Label

4 



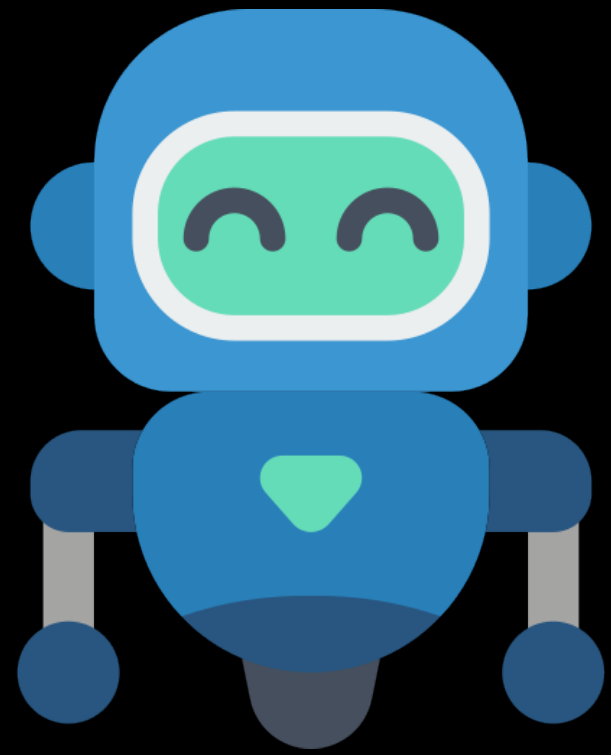
Predicted Label

4 

Problem #1:

invalid inputs, not
recognisable by domain
experts in the input domain

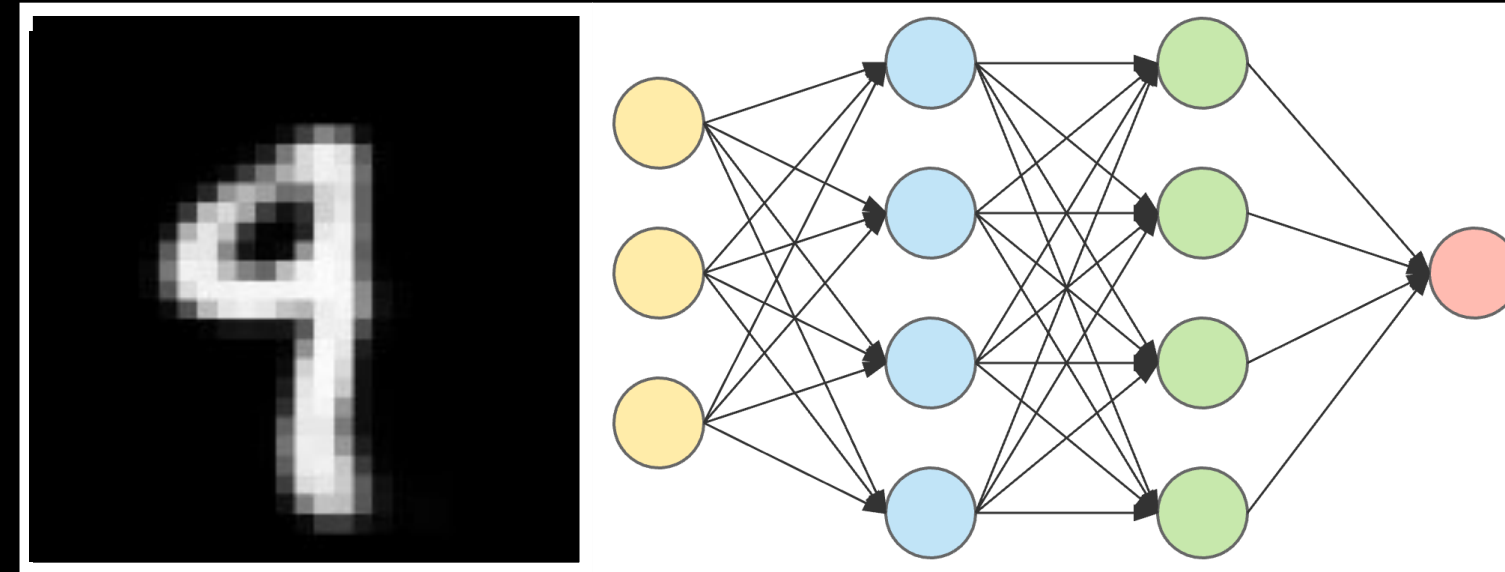
AUTOMATED TEST INPUT GENERATION FOR DNNs



**TEST
GENERATOR**

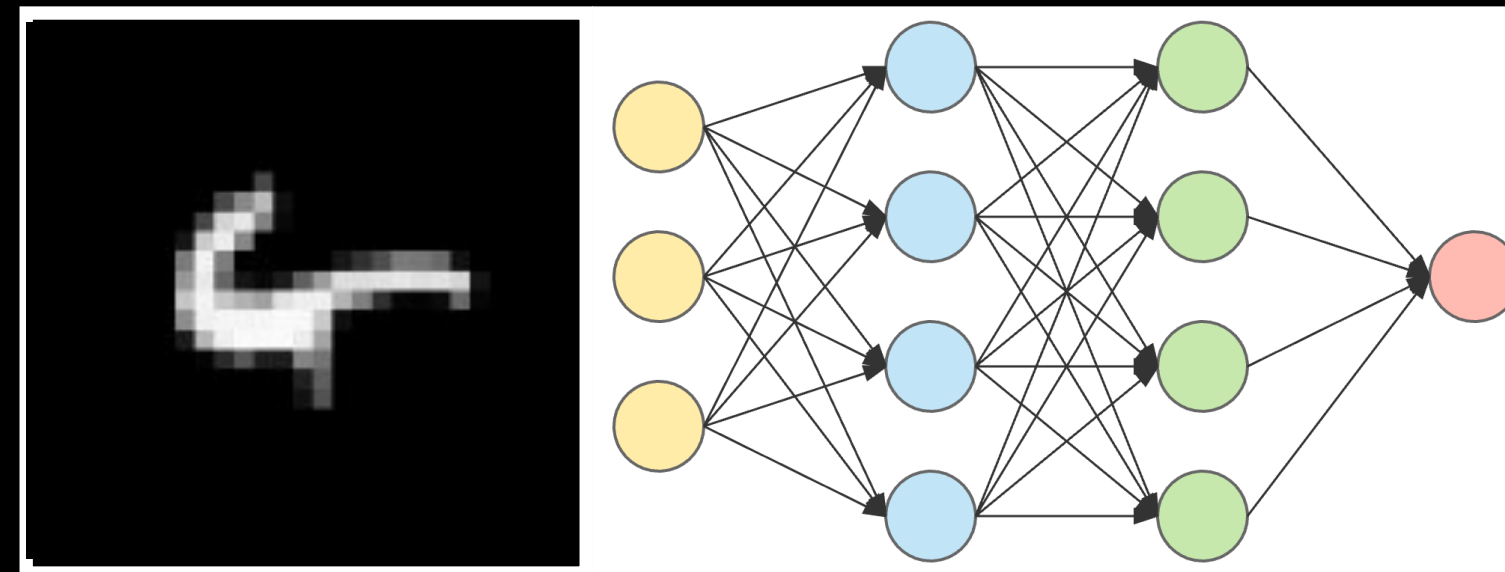
Target Label

9



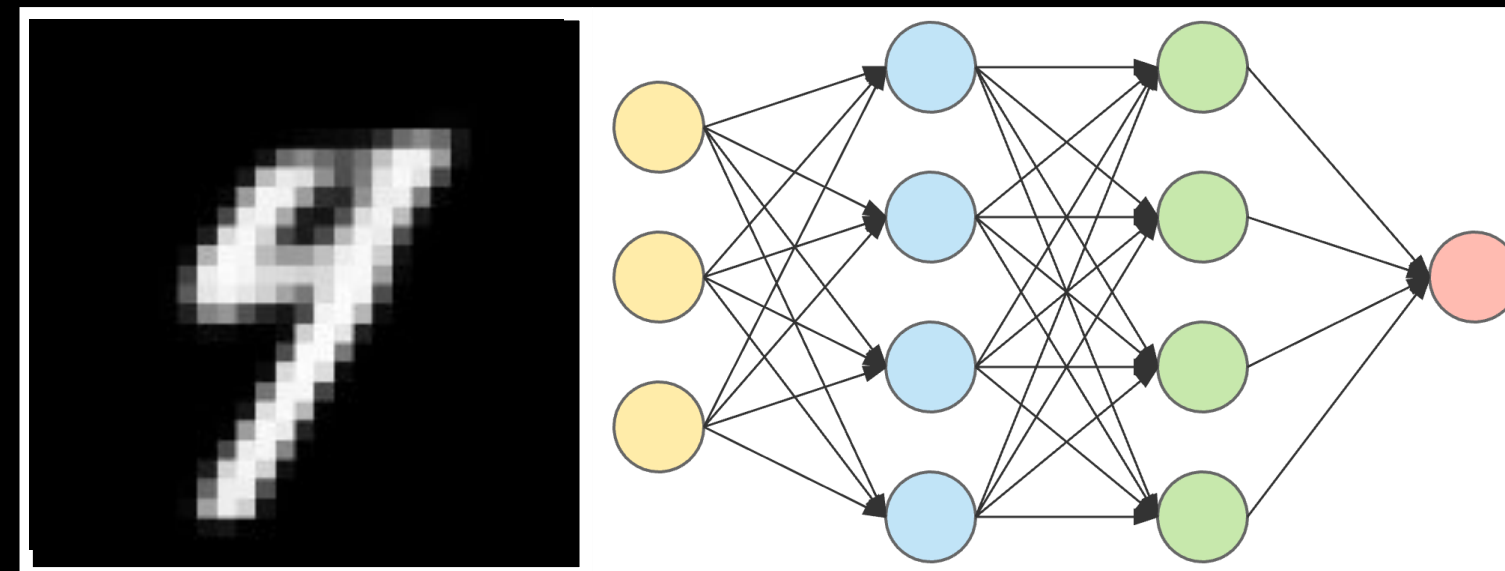
Predicted Label

4 



Predicted Label

4 



Predicted Label

4 

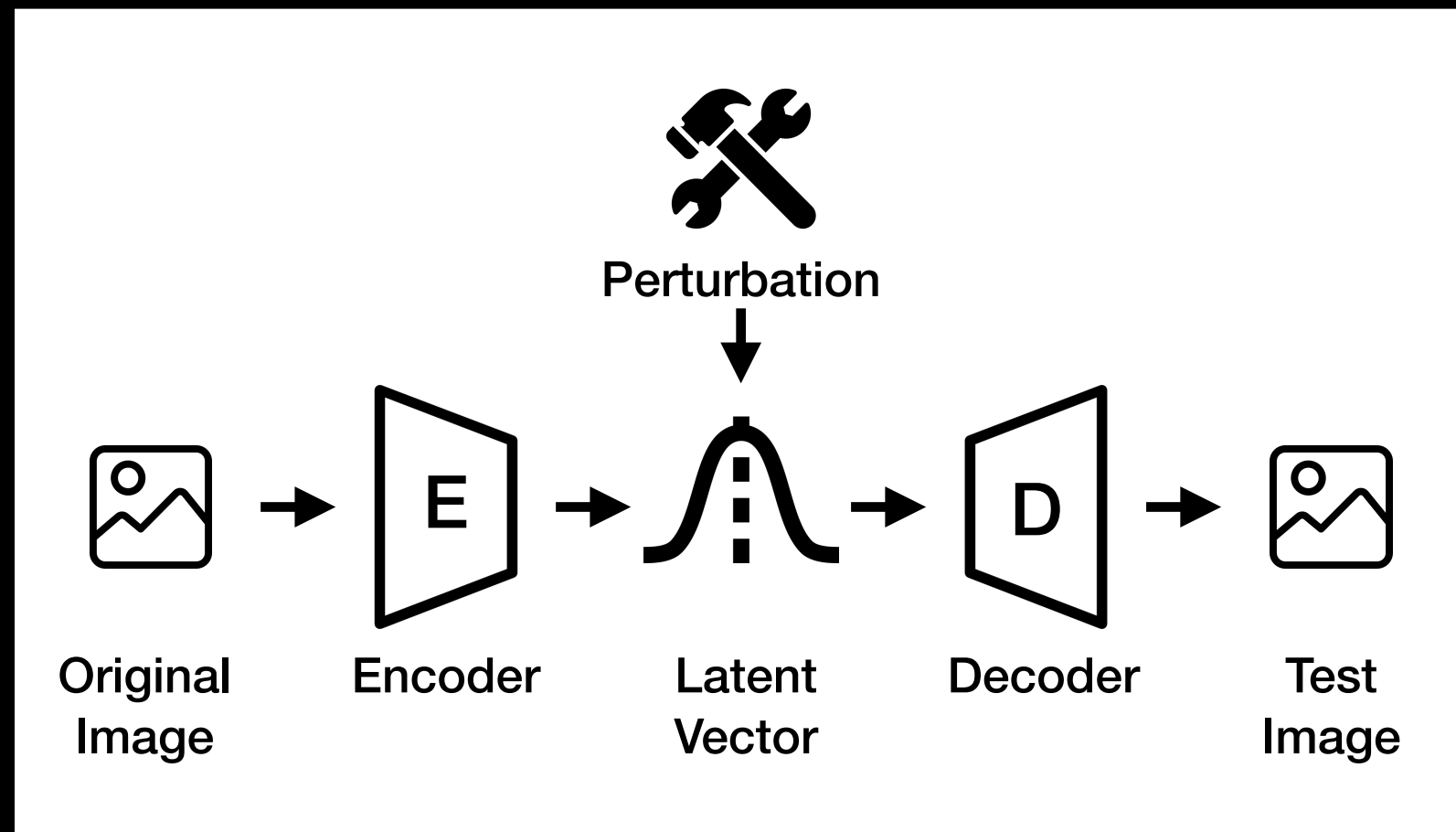
Problem #1:

invalid inputs, not
recognisable by domain
experts in the input domain

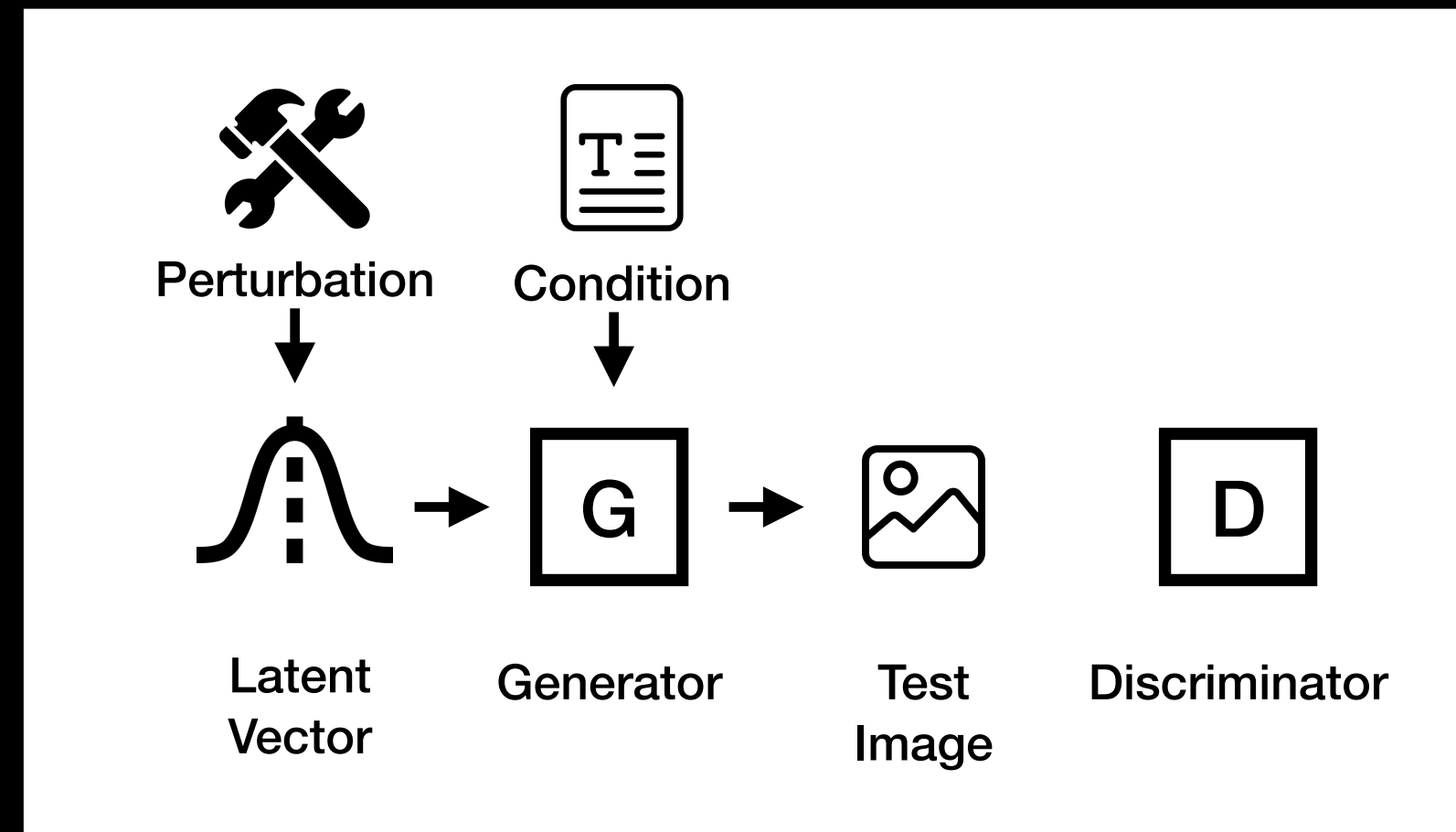
Problem #2:

original label is not
preserved

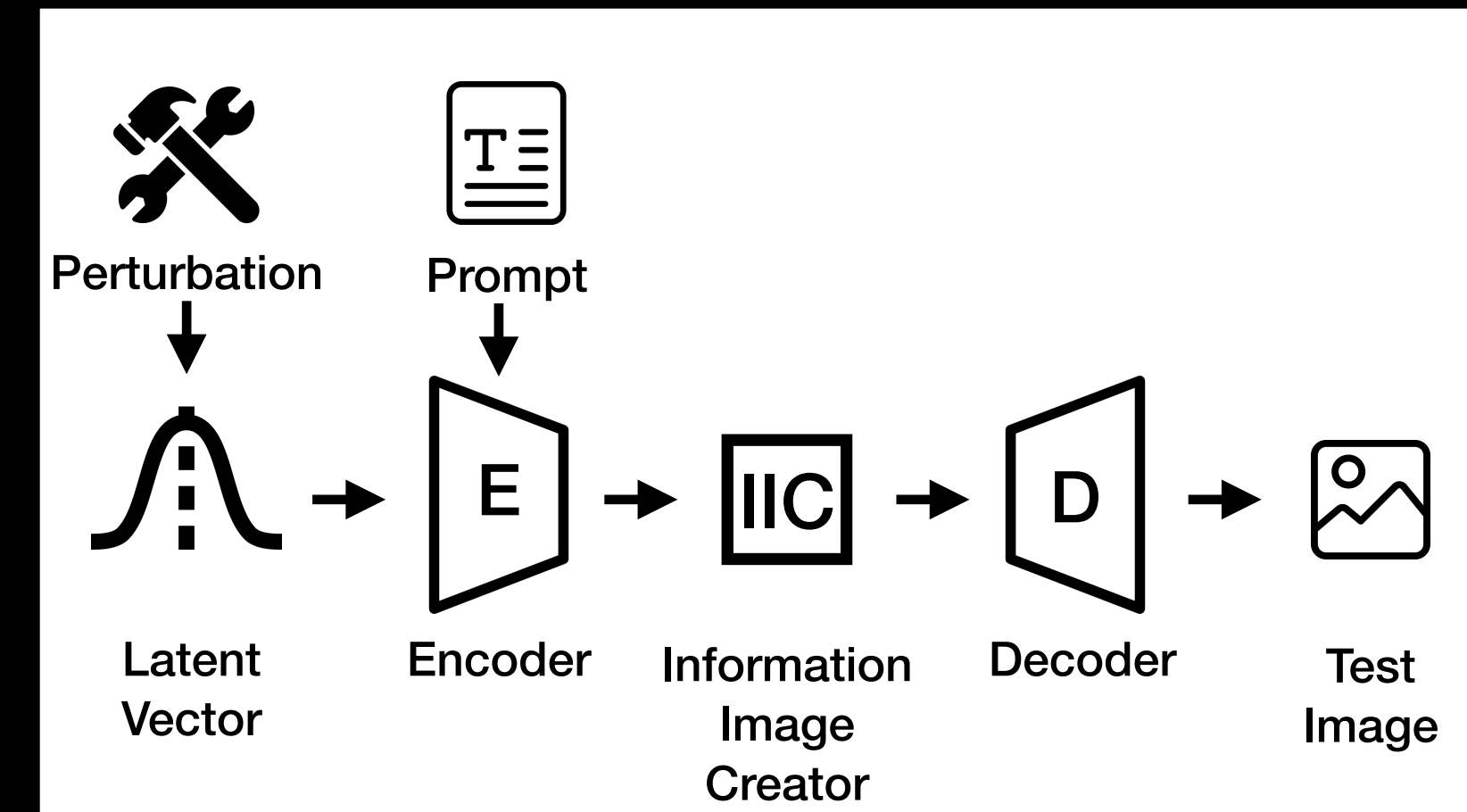
GENERATIVE AI MODELS



Variational AutoEncoder (VAE)

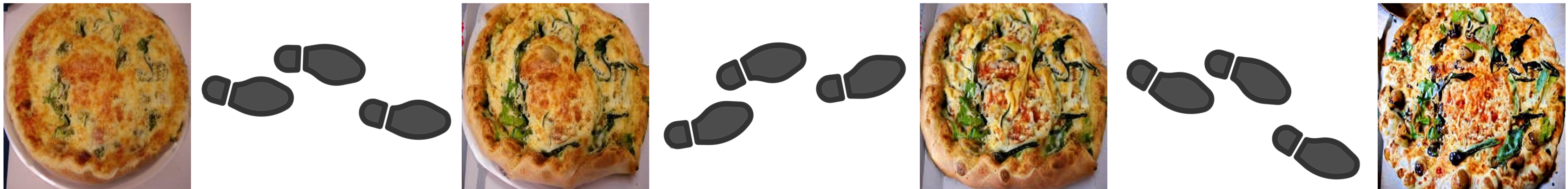


Generative Adversarial Network (GAN)

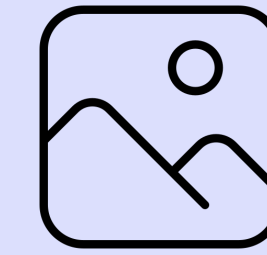
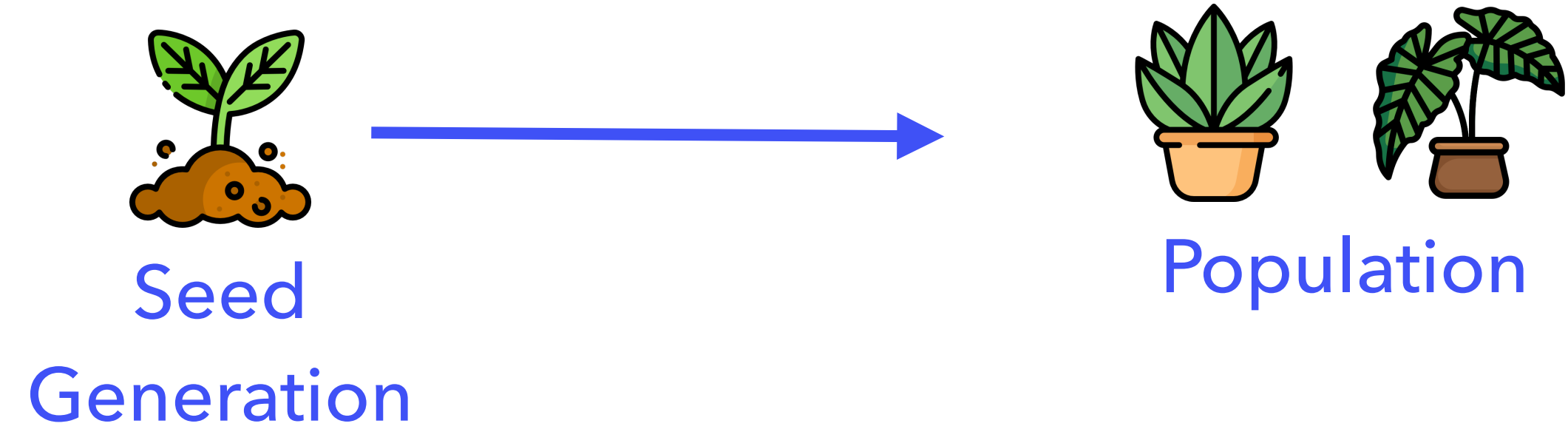


Diffusion Model (DM)

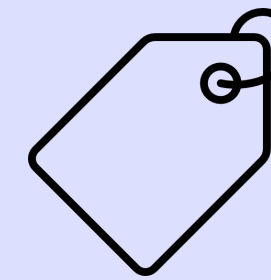
RANDOM WALK IN THE LATENT SPACE



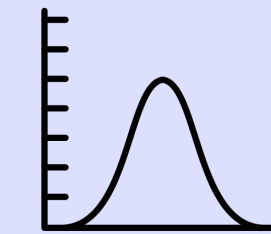
GENETIC ALGORITHM



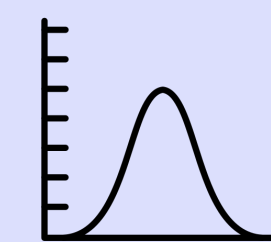
VAE: Image fed to the Encoder



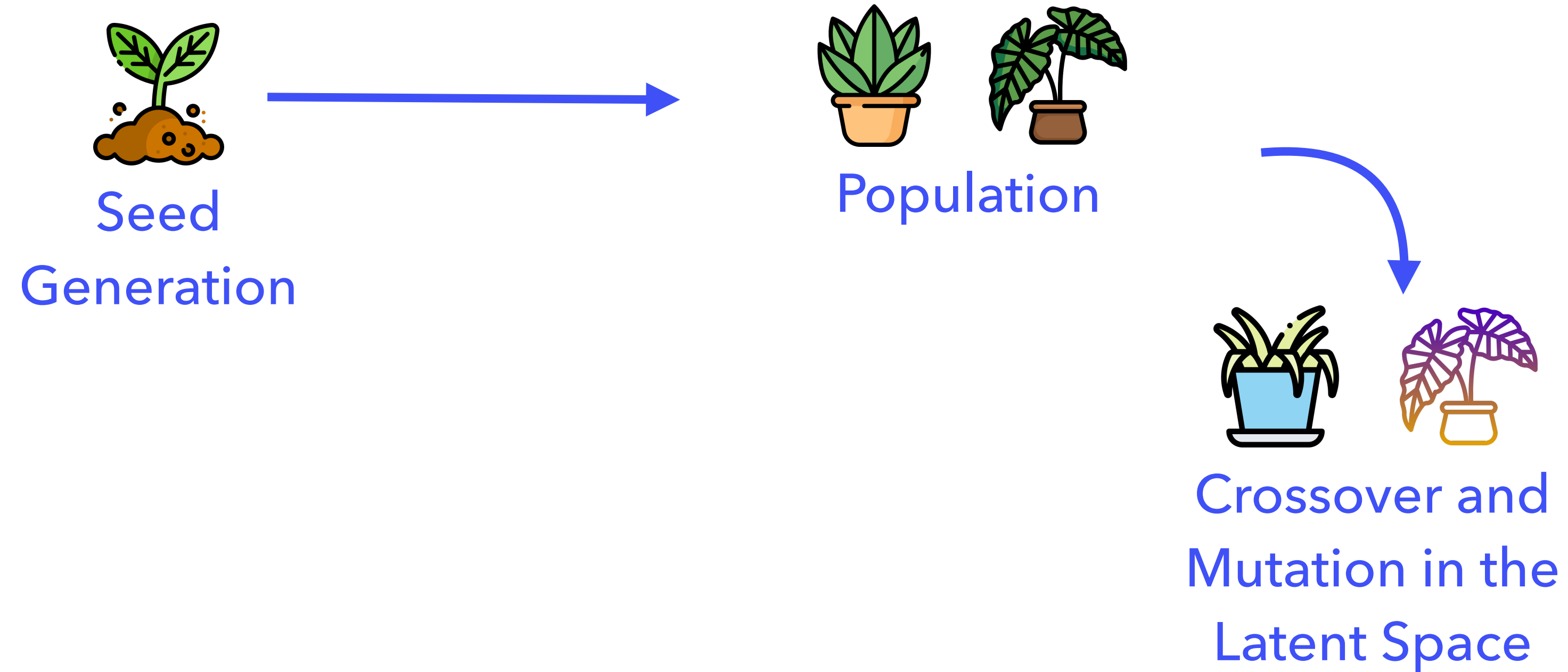
GAN: Sample from target distribution + target label



DM: Sample from target distribution + prompt



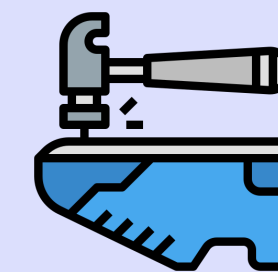
GENETIC ALGORITHM



Mutation:
Random Latent Walk

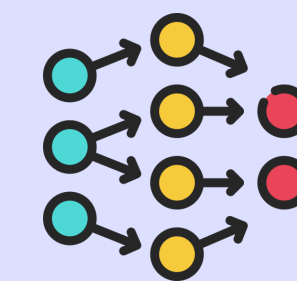
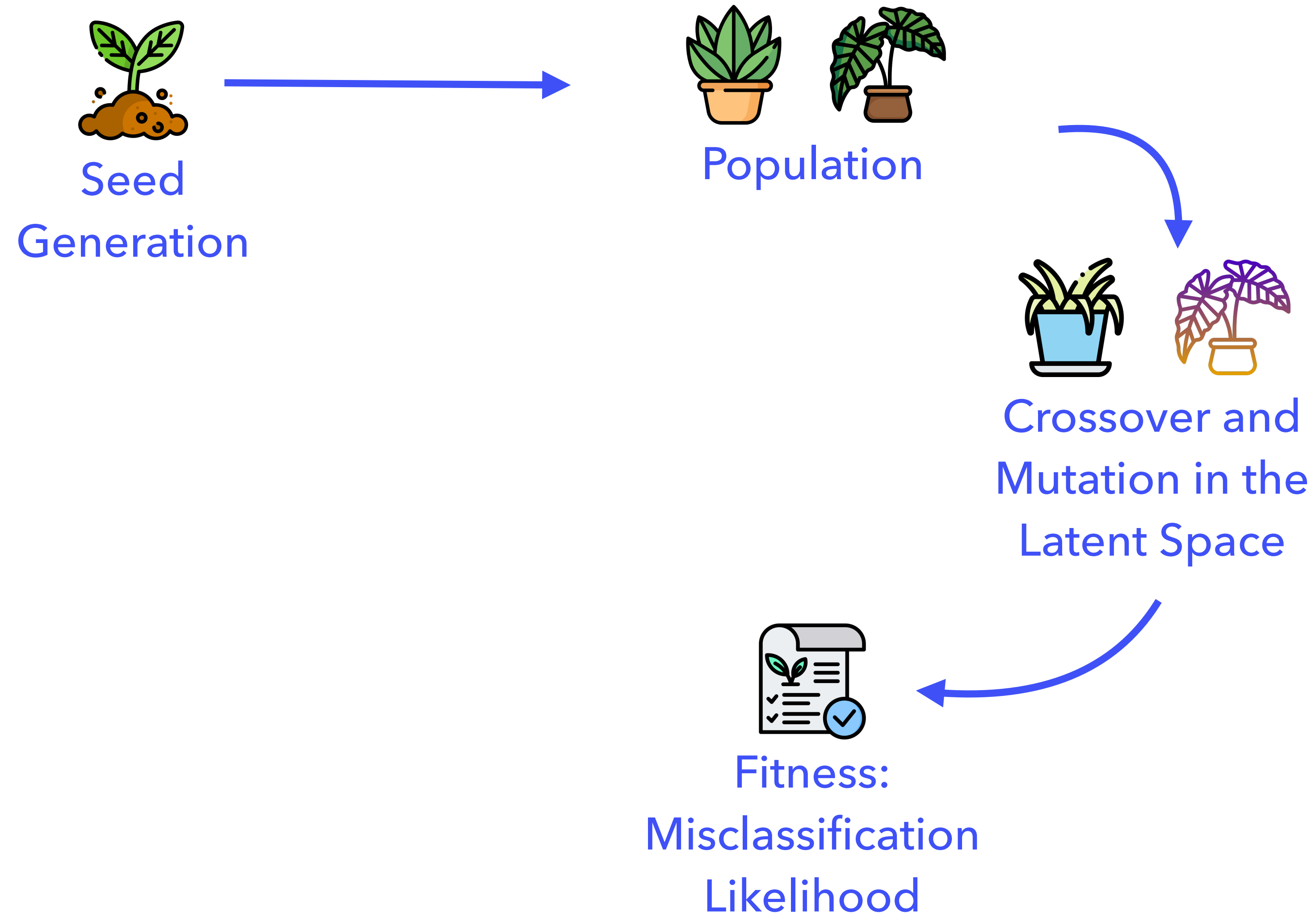


Crossover: One-Point Crossover of Latent Vectors



Constraint:
Clamping Vectors to Target Distribution

GENETIC ALGORITHM

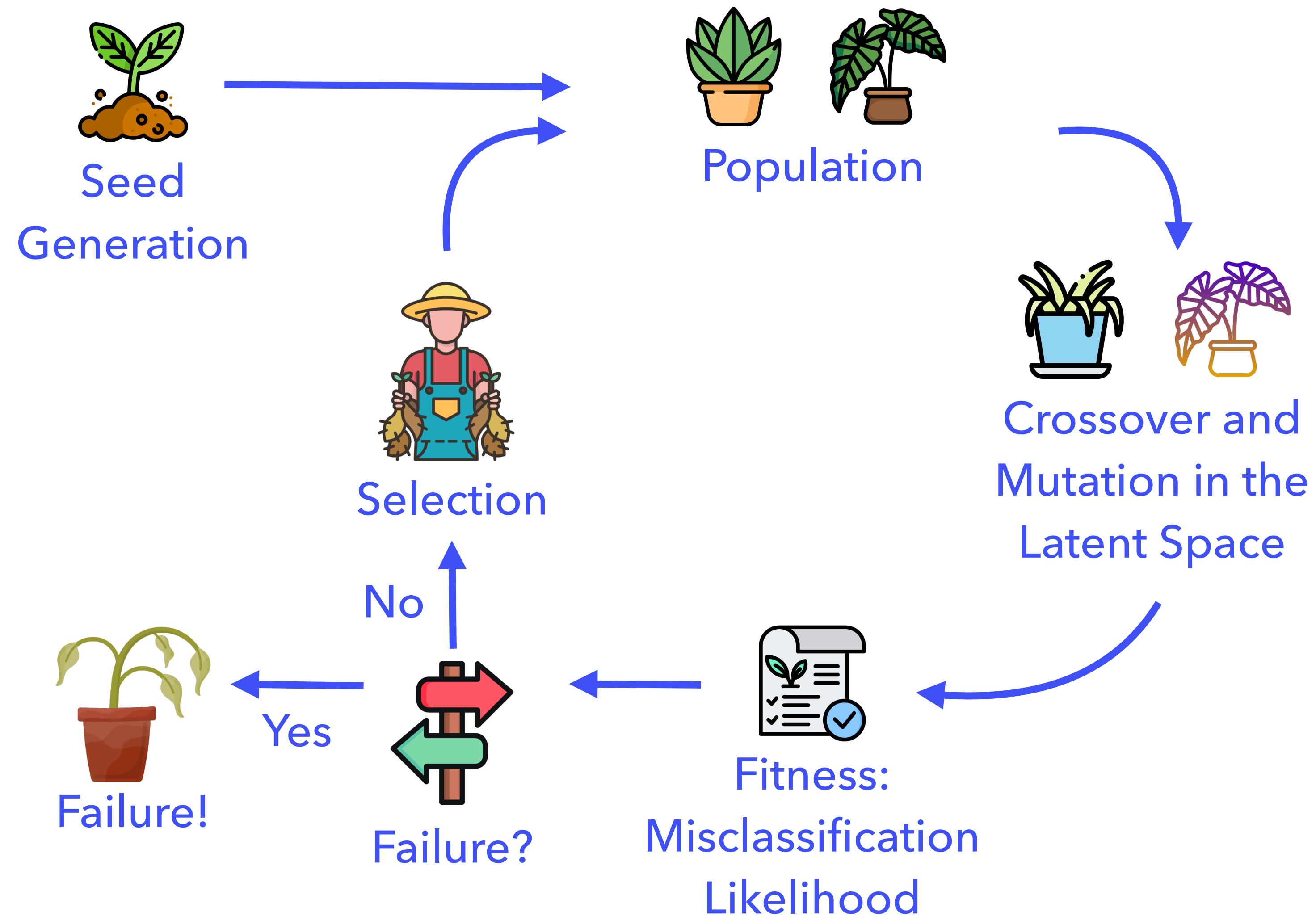


Evaluation of generated images on the DNN under test



Compares confidence assigned to the target class VS other classes

GENETIC ALGORITHM



Termination Conditions



Misclassification detected



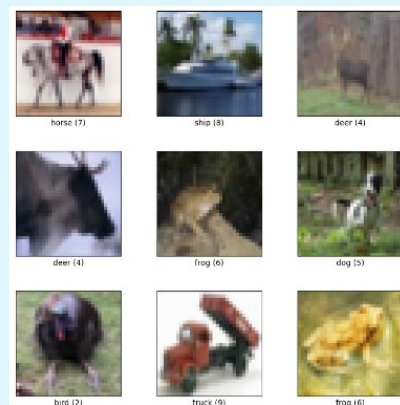
Search budget exhausted

EVALUATION BENCHMARK

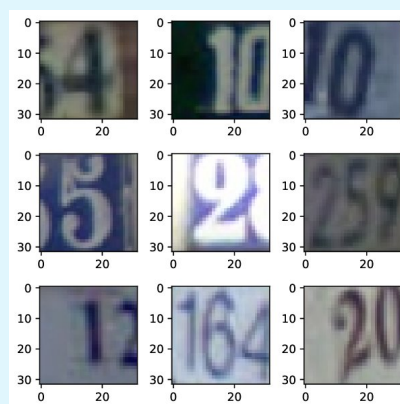
DATASETS



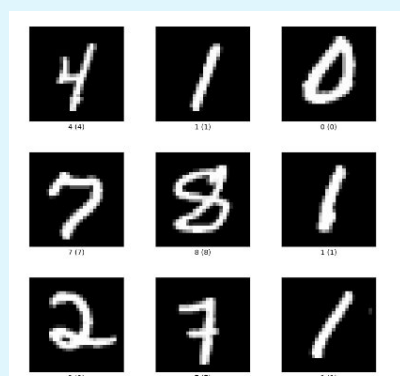
Imagenet



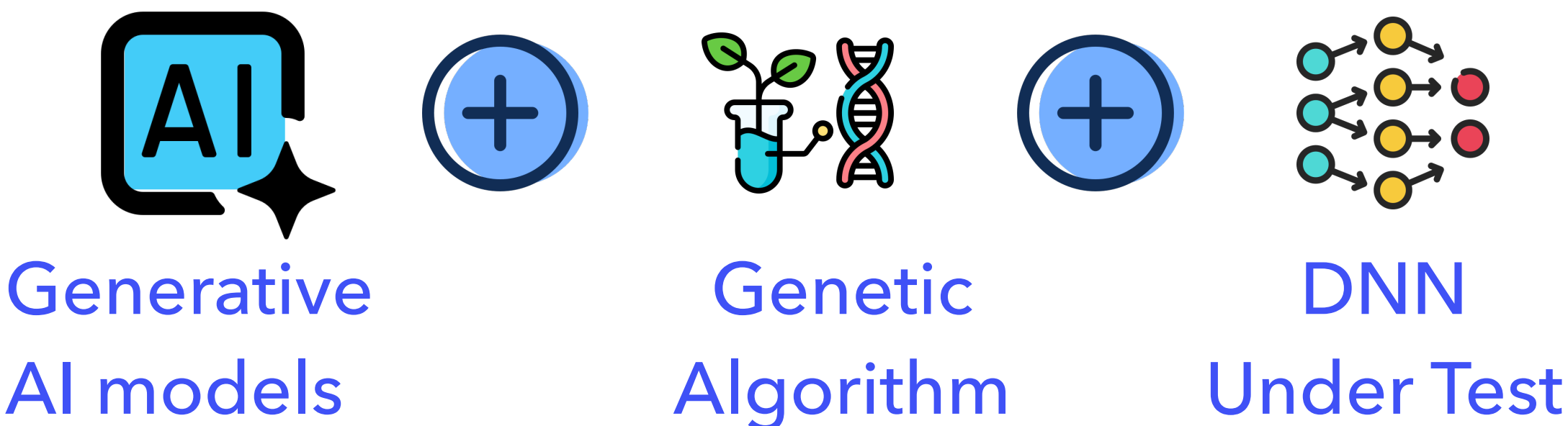
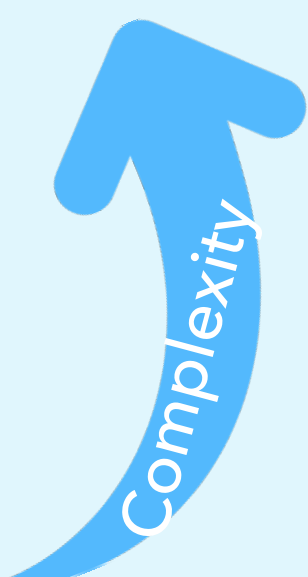
CIFAR-10



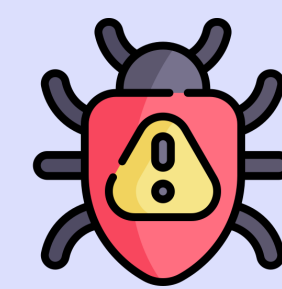
SVHN



MNIST



METRICS



Effectiveness:
Valid label-preserving misclassification-inducing inputs



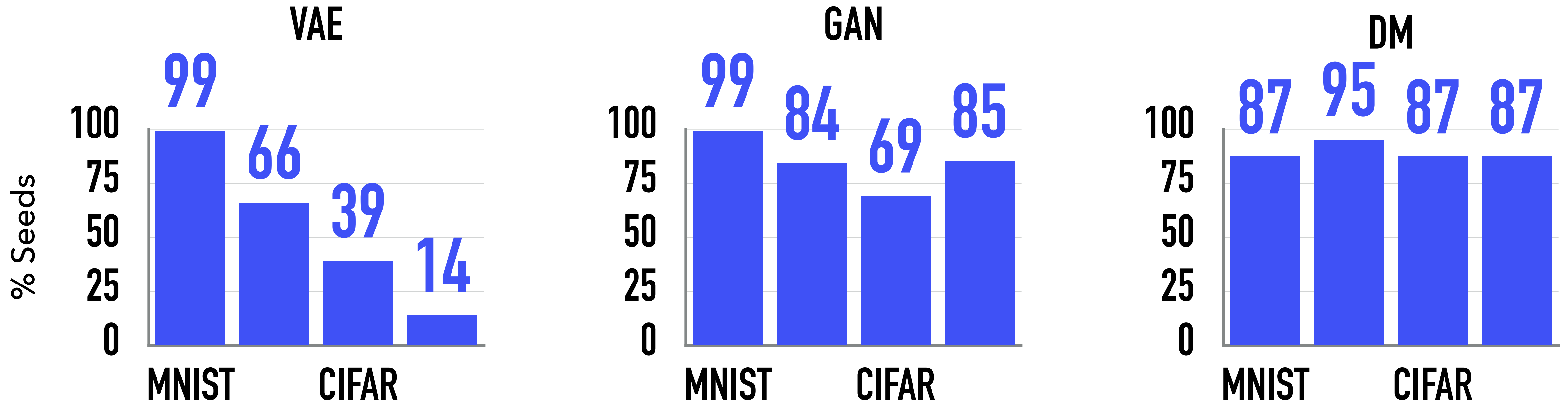
Efficiency:
Iterations

SETUP



Fixed search budget

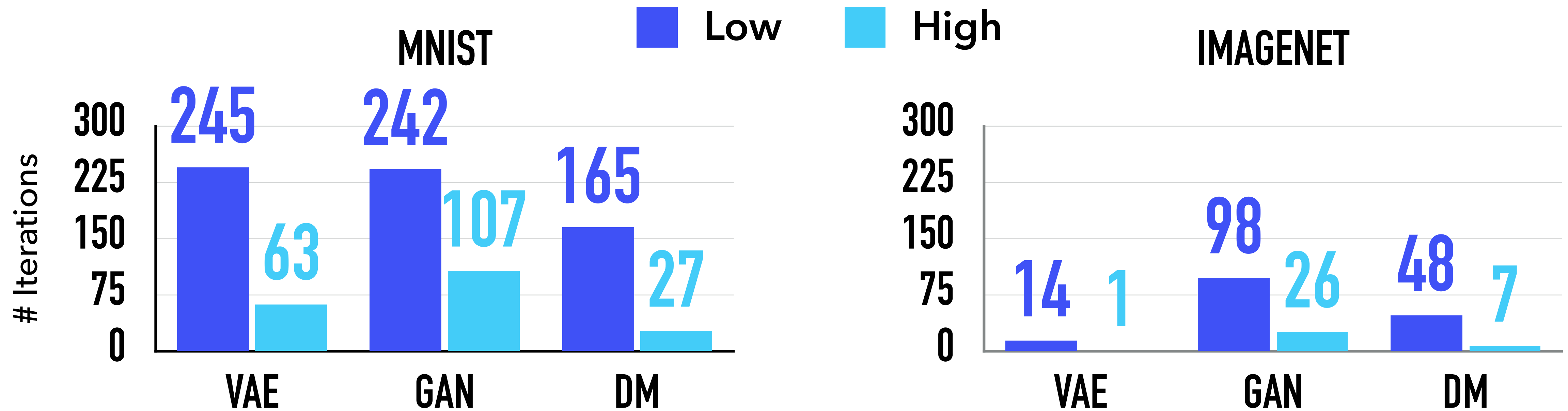
SEED GENERATION



VAE performance declined as dataset complexity increased

GAN and DMs consistently achieved high accuracy in seed generation, regardless of complexity

ITERATIONS TO TRIGGER A FAILURE



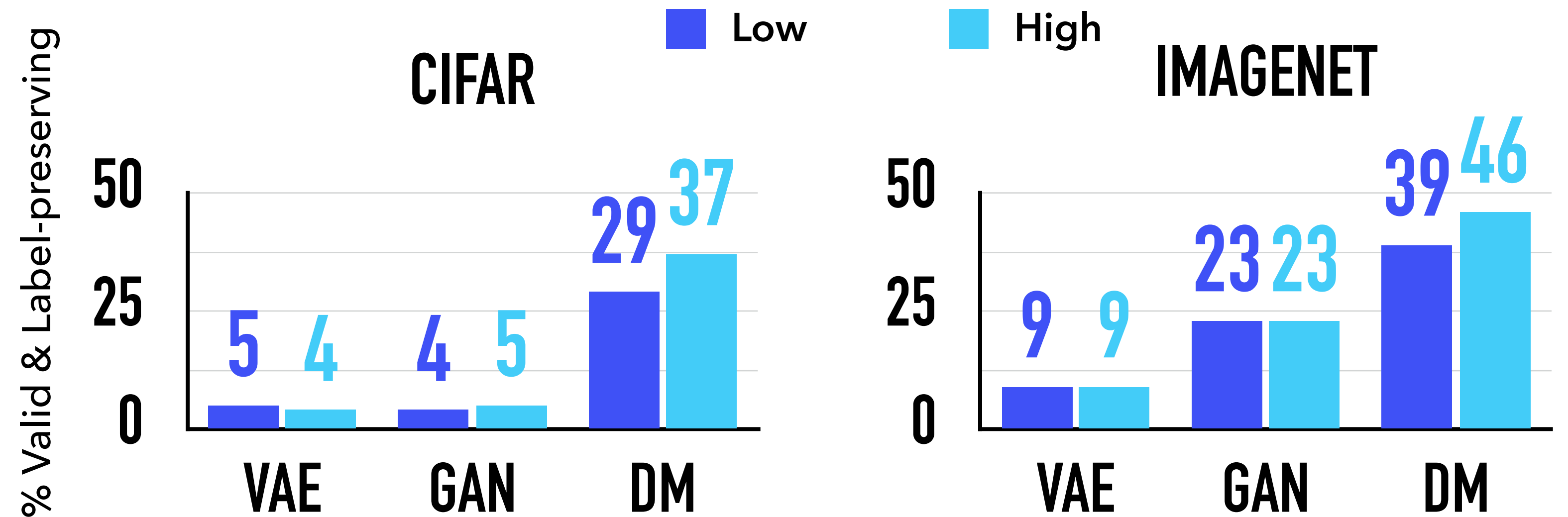
VAEs need less iterations than other GenAI models for complex datasets, while DMs are the most efficient for simpler datasets

Increased perturbations reduce the number of iterations

VALIDITY AND LABEL PRESERVATION



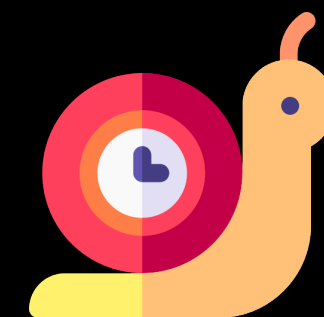
364 Human assessors from
amazon mechanical turk



DMs excel at generating valid misclassification-inducing inputs for complex datasets like CIFAR-10 and ImageNet

KEY INSIGHTS

- ▶ Diffusion Models excel in complex tasks, but their superior performance comes at a higher cost
- ▶ Larger perturbation extents speed up test generation without compromising input validity or label preservation
- ▶ Latent vectors should be carefully constrained and carefully manipulated



**INFERENCE TIME 10X
HIGHER FOR DM VS VAE**

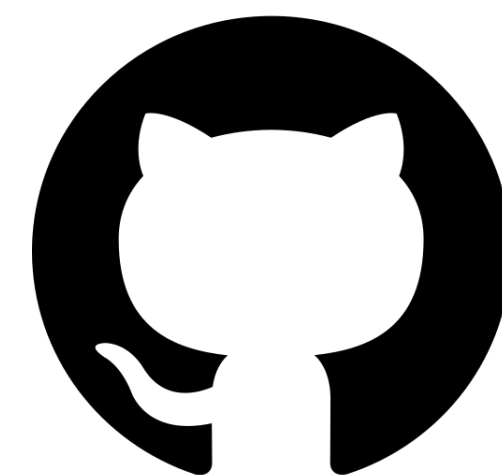
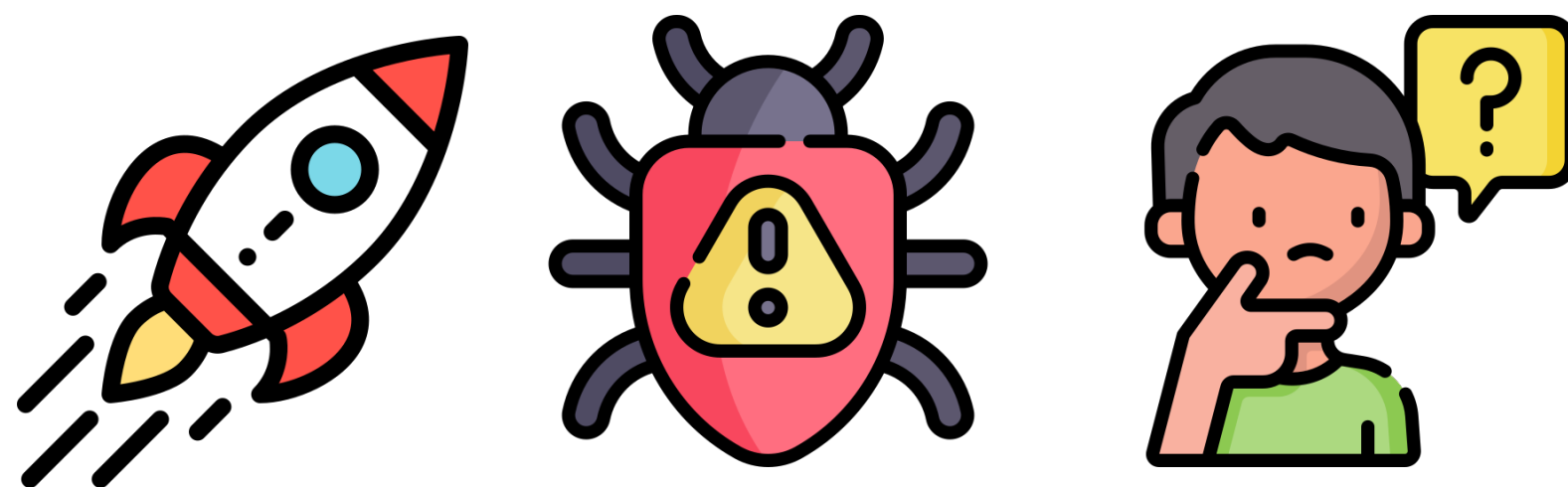
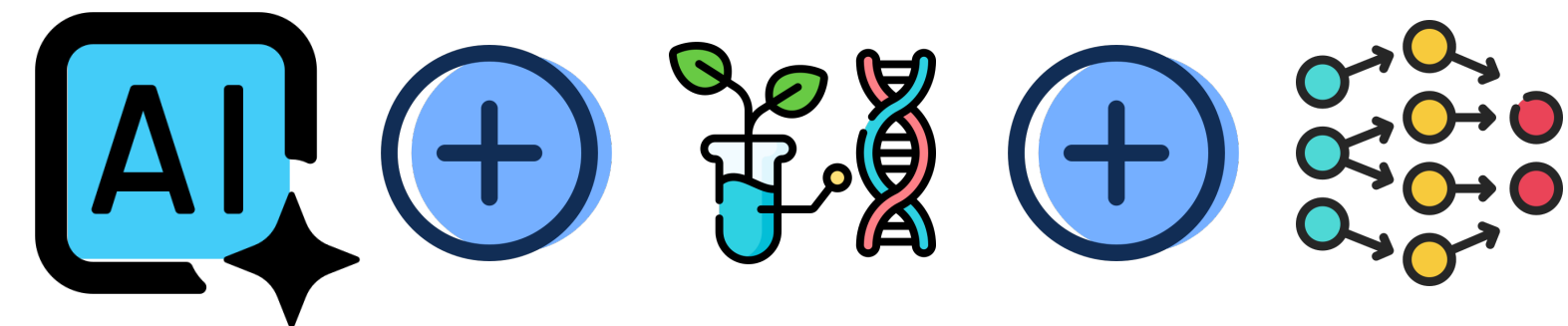
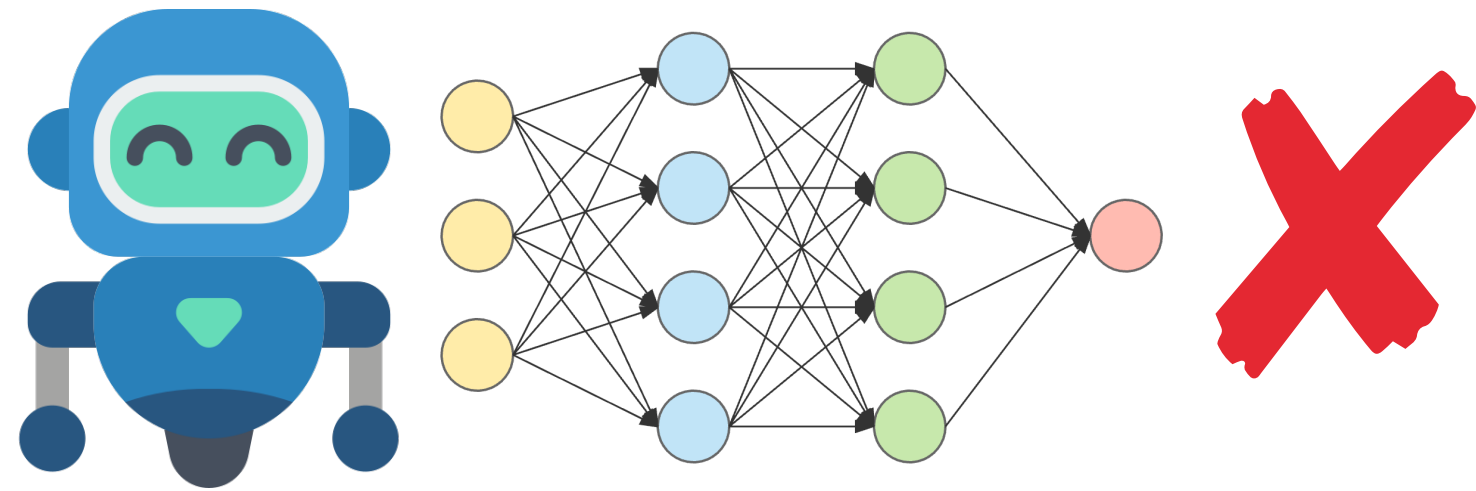


**DMS UP TO 10X MORE
EFFICIENT FOR
IMAGENET AND SVHN**



**AT MOST ONLY 15%
MISCLASSIFICATIONS FOR
SVHN ARE VALID AND
PRESERVE THE LABEL**

SUMMARY



EXTRA SLIDES

TABLE II
CHARACTERISTICS OF THE GENAI MODELS: LATENT VECTOR SIZE,
TRAINING TIME UNTIL CONVERGENCE, AVERAGE INFERENCE TIME.

Dataset	Model	LV size	t_{train} (min)	t_{infer} (ms)
MNIST	VAE [49]	400	6	0.27
	GAN [50], [51]	100	9	0.7
	DM [52]	16384	405	960.68
SVHN	VAE [53]	800	93	4.07
	GAN [50], [51]	100	86	1.75
	DM [52]	16384	572	1213.49
CIFAR-10	VAE [53]	1024	423	2.51
	GAN [50], [51]	100	450	1.73
	DM [52]	16384	362	1903.29
ImageNet	VAE [54]	512	2521	11.92
	GAN [55]	128	21600	20.68
	DM [52]	16384	30	1945.77

TABLE III
COMPARISON BETWEEN GENAI TIGS ACROSS DIFFERENT DATASETS AND MUTATION EXTENTS IN TERMS OF VIABLE SEEDS, MISCLASSIFICATION-INDUCING INPUTS, NUMBER OF ITERATIONS TO GENERATE FAILURE, INPUT VALIDITY, AND LABEL PRESERVATION. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD, WHILE THE UNDERLINED VALUES ARE NOT STATISTICALLY DIFFERENT FROM THE BEST.

Dataset	Pert. Step (δ_{init})	Model	% Seeds	% Misclass. (#)	# Iterations	% Validity (#)	% Preserved (#)
MNIST	Low	VAE	99	4.04 (4)	245.41	50.00 (2)	100.00 (2)
		GAN	99	8.08 (8)	242.05	75.00 (6)	83.33 (5)
		DM	87	50.57 (44)	164.61	45.45 (20)	30.00 (6)
	High	VAE	99	100.00 (99)	62.92	73.74 (73)	<u>49.32</u> (36)
		GAN	99	96.97 (96)	107.46	<u>69.79</u> (67)	62.69 (42)
		DM	87	100.00 (87)	26.77	40.23 (35)	34.29 (12)
SVHN	Low	VAE	66	50.00 (33)	178.20	51.52 (17)	41.18 (7)
		GAN	84	42.86 (36)	182.22	30.56 (11)	<u>45.45</u> (5)
		DM	95	69.47 (66)	131.04	39.39 (26)	57.69 (15)
	High	VAE	66	100.00 (66)	27.00	39.39 (26)	30.77 (8)
		GAN	84	<u>98.81</u> (83)	39.00	<u>36.14</u> (30)	50.00 (15)
		DM	95	100.00 (95)	13.23	23.16 (22)	18.18 (4)
CIFAR-10	Low	VAE	39	<u>82.05</u> (32)	118.90	<u>53.13</u> (17)	29.41 (5)
		GAN	69	66.67 (46)	140.32	45.65 (21)	19.05 (4)
		DM	87	89.66 (78)	85.63	60.26 (47)	61.70 (29)
	High	VAE	39	100.00 (39)	<u>19.51</u>	30.77 (12)	33.33 (4)
		GAN	69	100.00 (69)	<u>25.78</u>	31.88 (22)	22.73 (5)
		DM	87	100.00 (87)	14.18	62.07 (54)	68.52 (37)
ImageNet (Teddy Bear)	Low	VAE	14	100.00 (14)	13.57	78.57 (11)	81.82 (9)
		GAN	<u>85</u>	100.00 (85)	98.27	74.12 (63)	36.51 (23)
		DM	87	<u>98.85</u> (86)	48.45	91.86 (79)	49.37 (39)
	High	VAE	14	100.00 (14)	1.36	100.00 (14)	64.29 (9)
		GAN	<u>85</u>	100.00 (85)	26.38	83.53 (71)	32.39 (23)
		DM	87	100.00 (87)	6.63	<u>94.25</u> (82)	<u>56.10</u> (46)
ImageNet (Pizza)	Low	VAE	25	100.00 (25)	12.96	<u>92.00</u> (23)	<u>91.30</u> (21)
		GAN	99	88.00 (87)	172.88	88.51 (77)	46.75 (36)
		DM	73	<u>97.26</u> (71)	83.60	98.59 (70)	92.86 (65)
	High	VAE	25	100.00 (25)	2.60	80.00 (20)	<u>75.00</u> (15)
		GAN	99	100.00 (99)	47.93	86.87 (86)	51.16 (44)
		DM	73	100.00 (73)	12.53	100.00 (73)	86.30 (63)

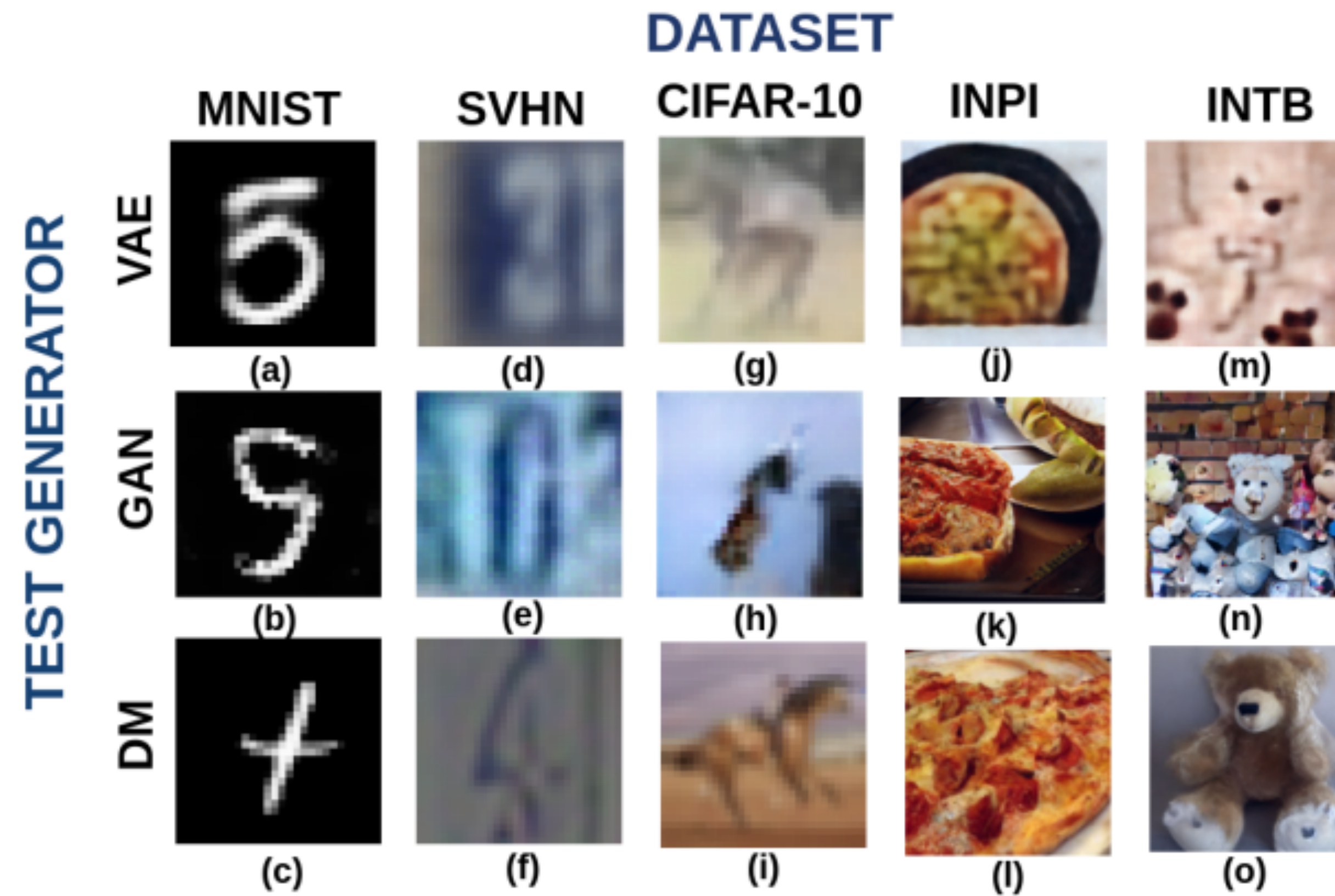


Fig. 3. Misclassification-inducing images generated by GenAI TIGs