2023

WHEN AND WHY TEST GENERATORS FOR DEEP LEARNING PRODUCE INVALID INPUTS: AN EMPIRICAL STUDY

VINCENZO
RICCIO

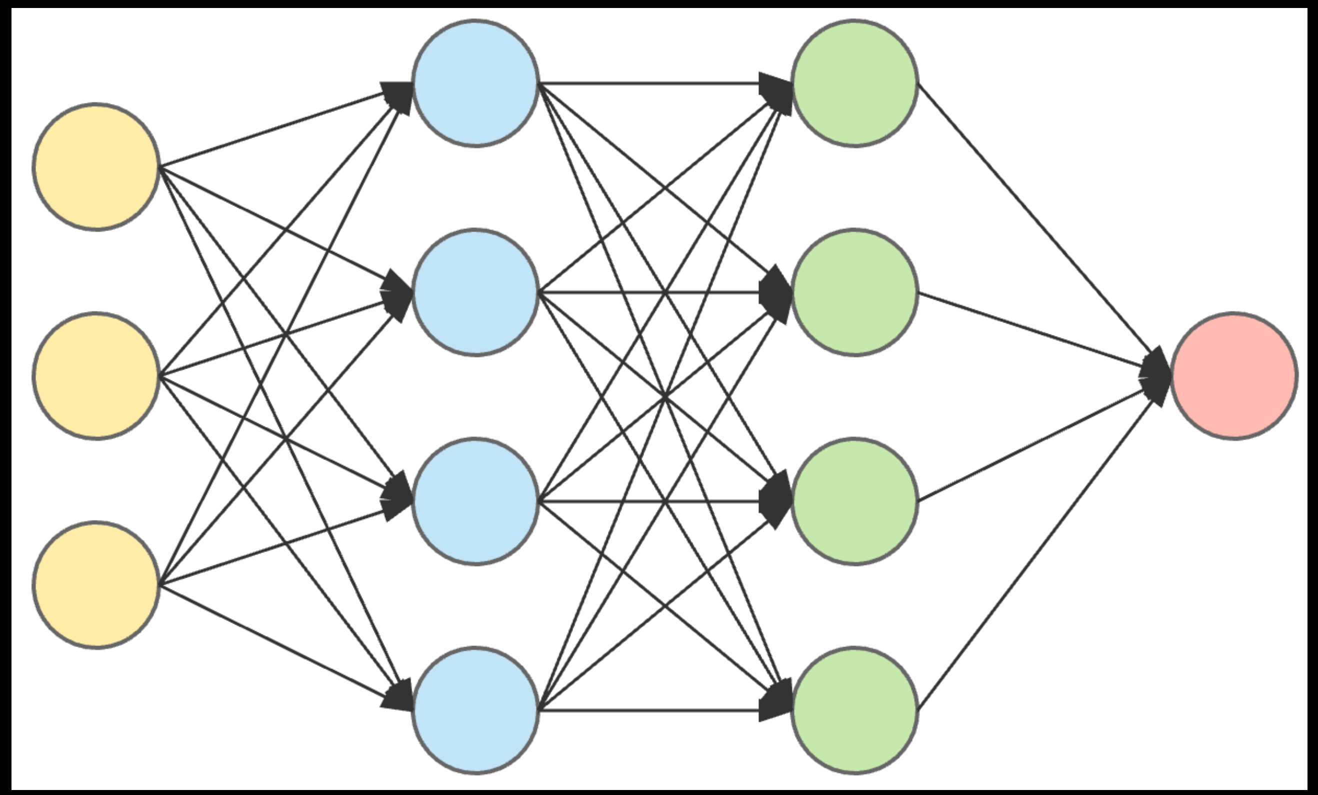
@p1ndsvin

PAOLO
TONELLA

@paolo_tonella

TRADITIONAL DL MODEL ASSESSMENT

ORIGINAL
DATASET



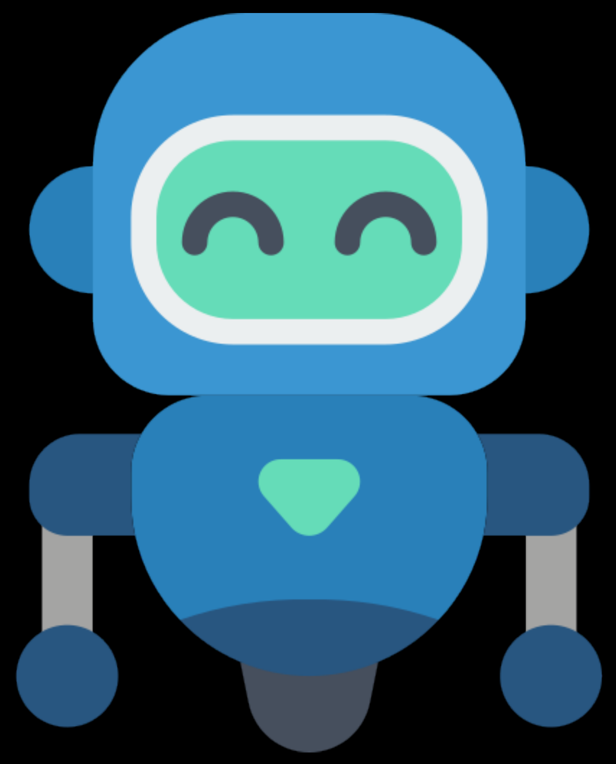
DL MODEL
UNDER TEST

ACC = 0.95
MSE = 0.01

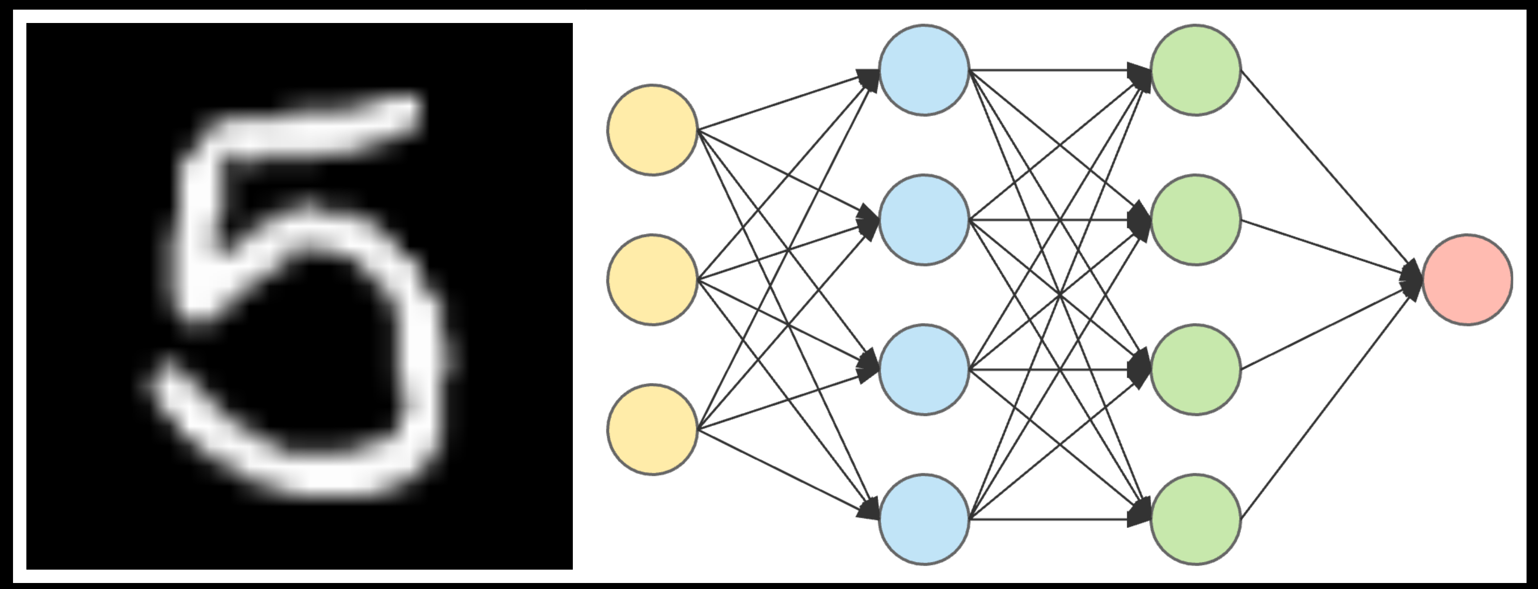
PERFORMANCE
METRIC

What is the performance of a DL model for inputs beyond its original dataset?

AUTOMATED TEST INPUT GENERATION FOR DL MODELS



TEST GENERATOR

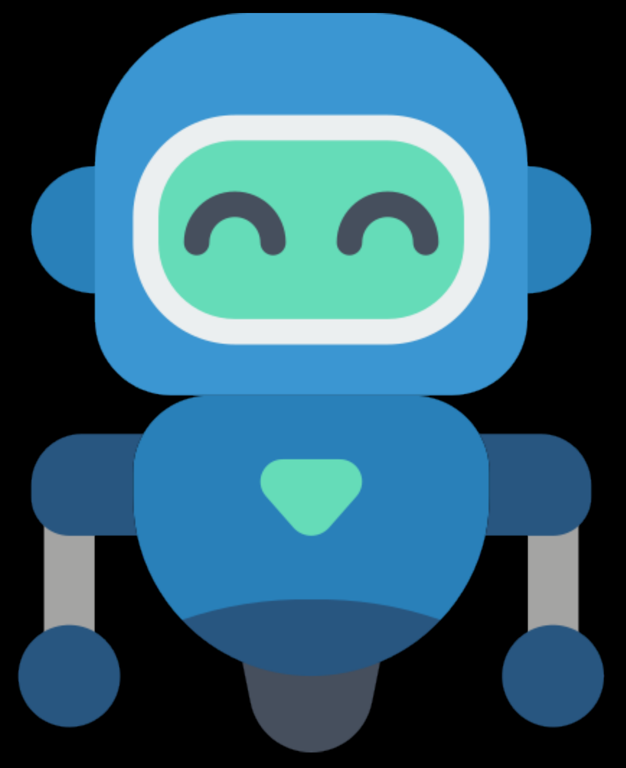


Predicted Label

5 ✓

Target Label
5

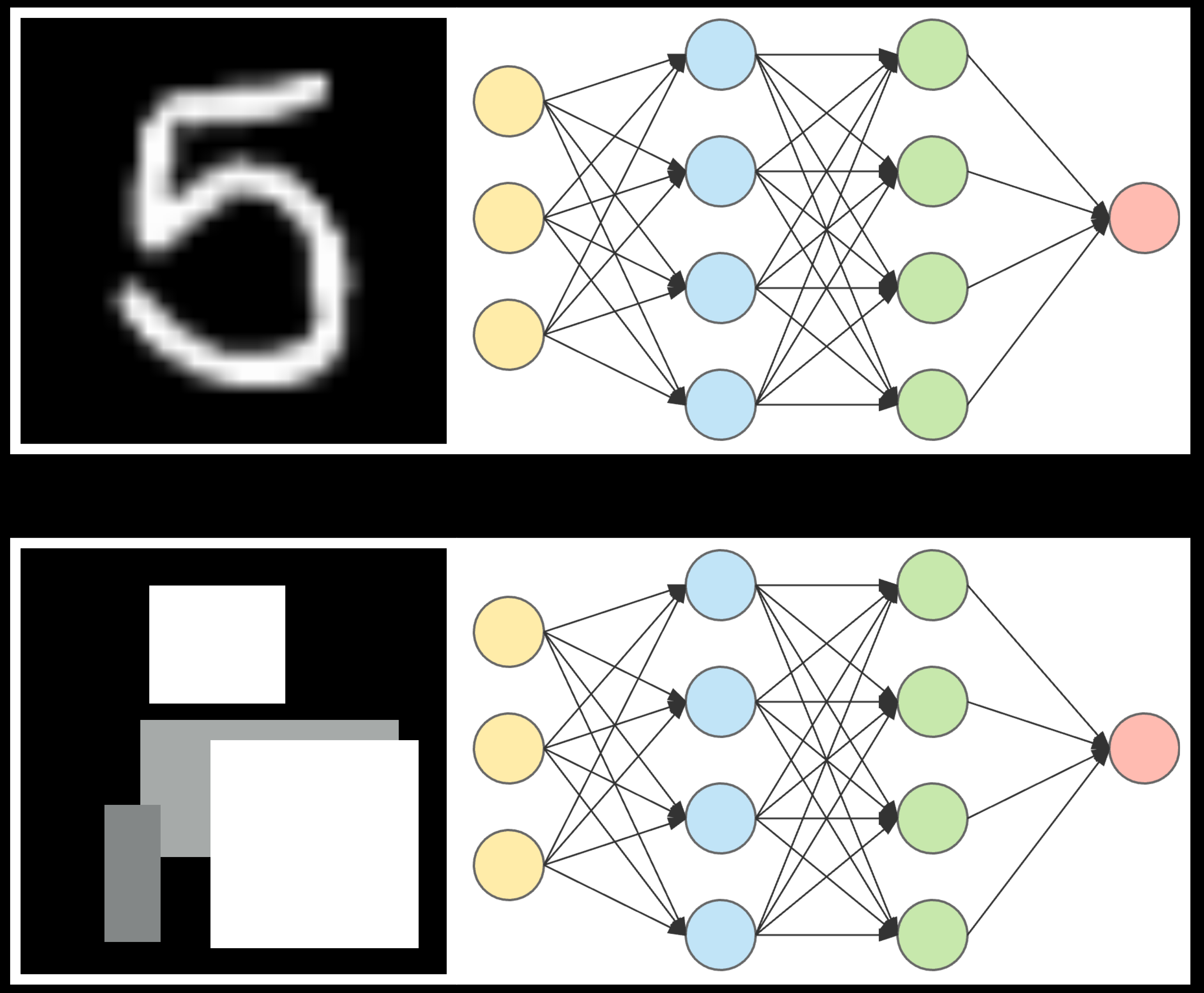
AUTOMATED TEST INPUT GENERATION FOR DL MODELS



TEST
GENERATOR

Target Label

5



Predicted Label

5

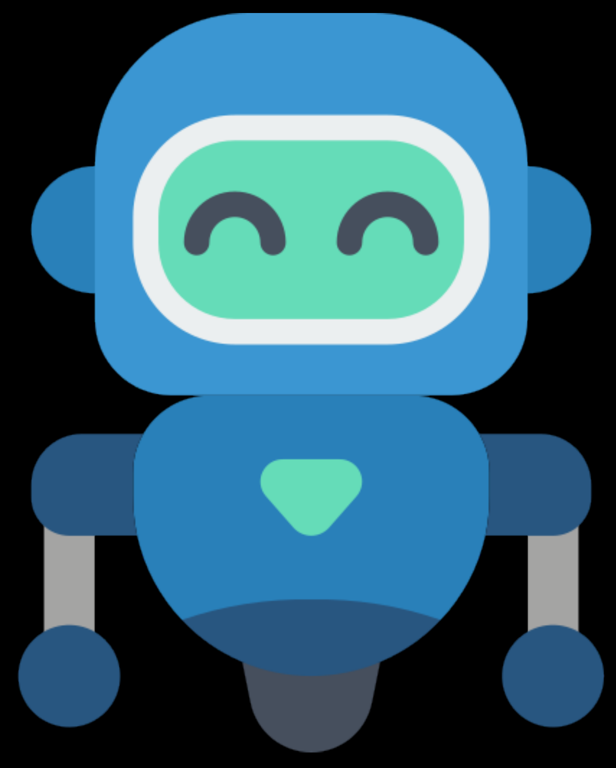
Predicted Label

6

Problem #1:

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

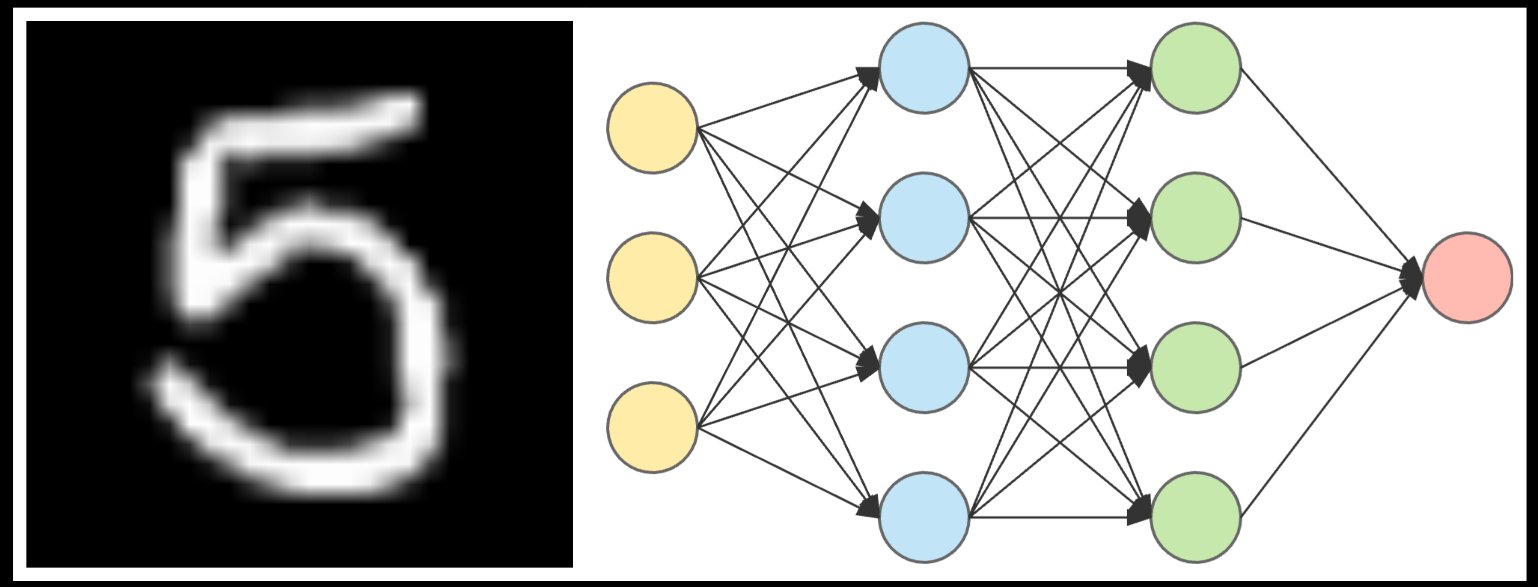
AUTOMATED TEST INPUT GENERATION FOR DL MODELS



TEST GENERATOR

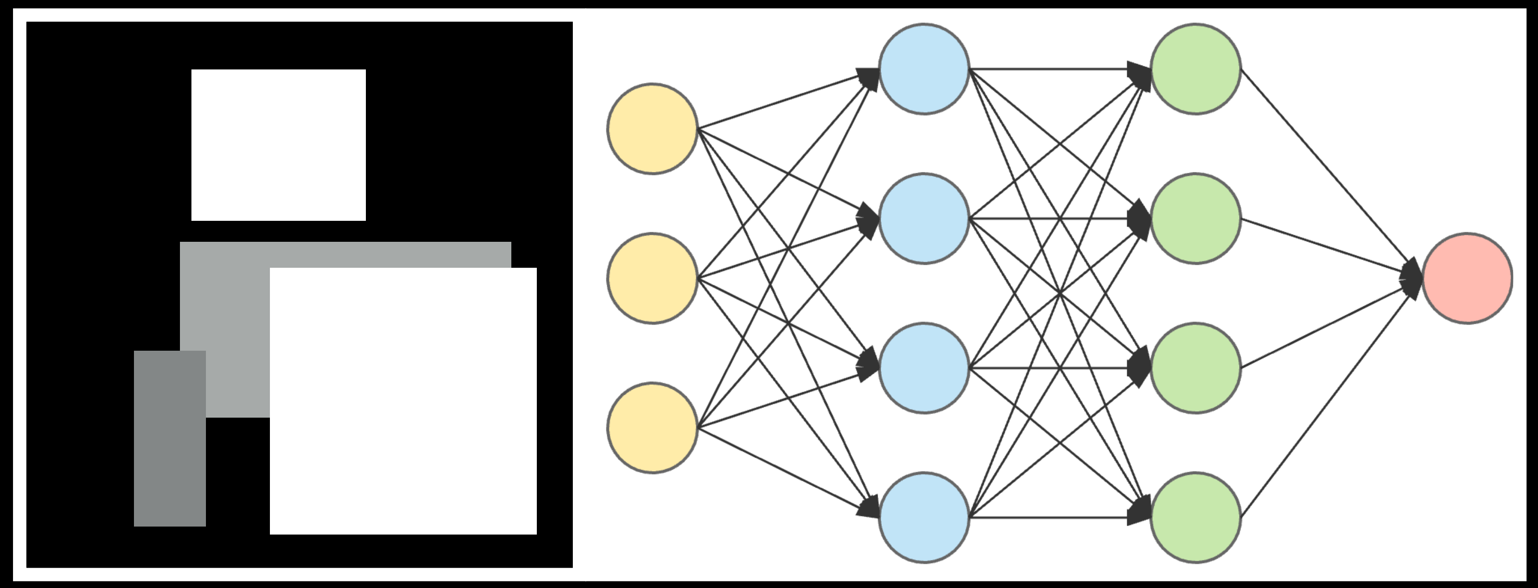
Target Label

5



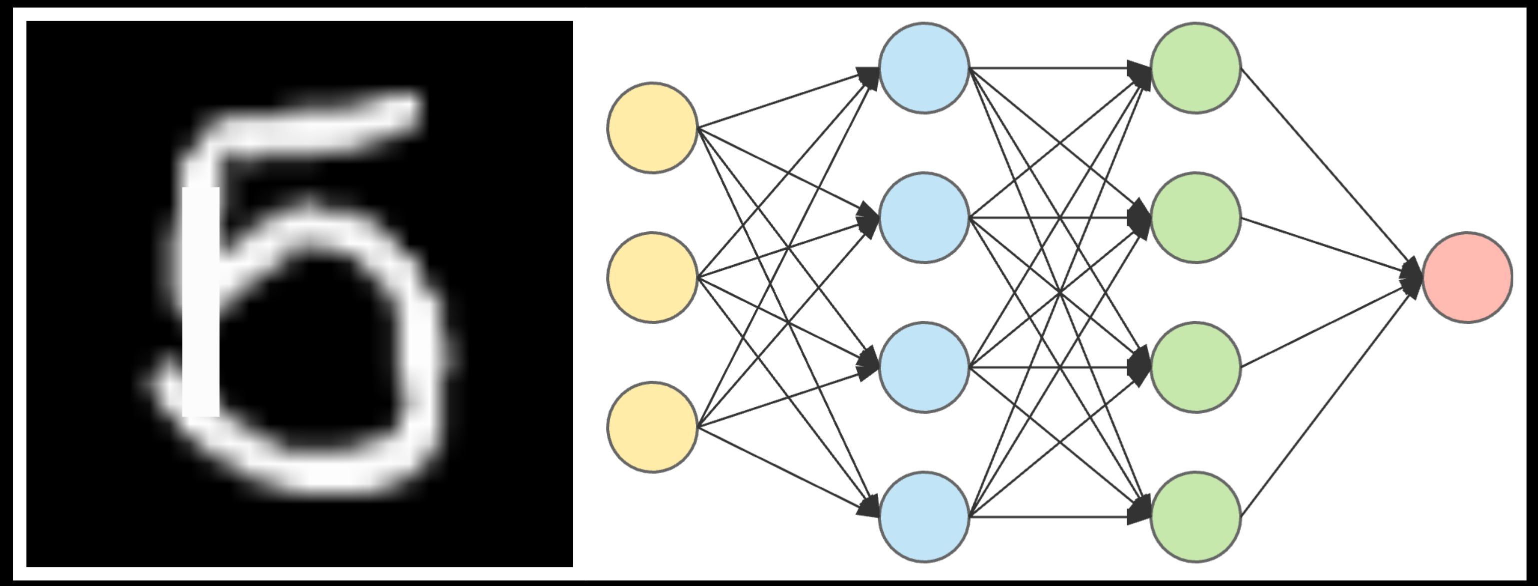
Predicted Label

5



Predicted Label

6



Predicted Label

6

Problem #1:

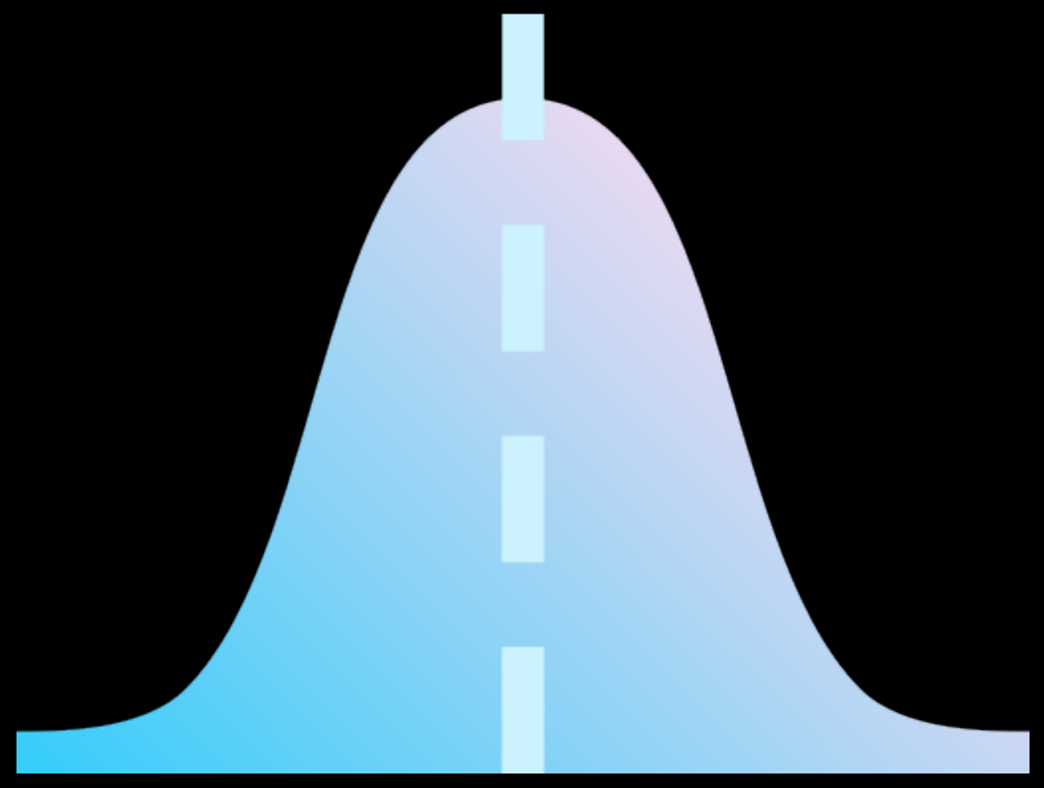
invalid inputs, not recognisable by domain experts in the input domain

Problem #2:

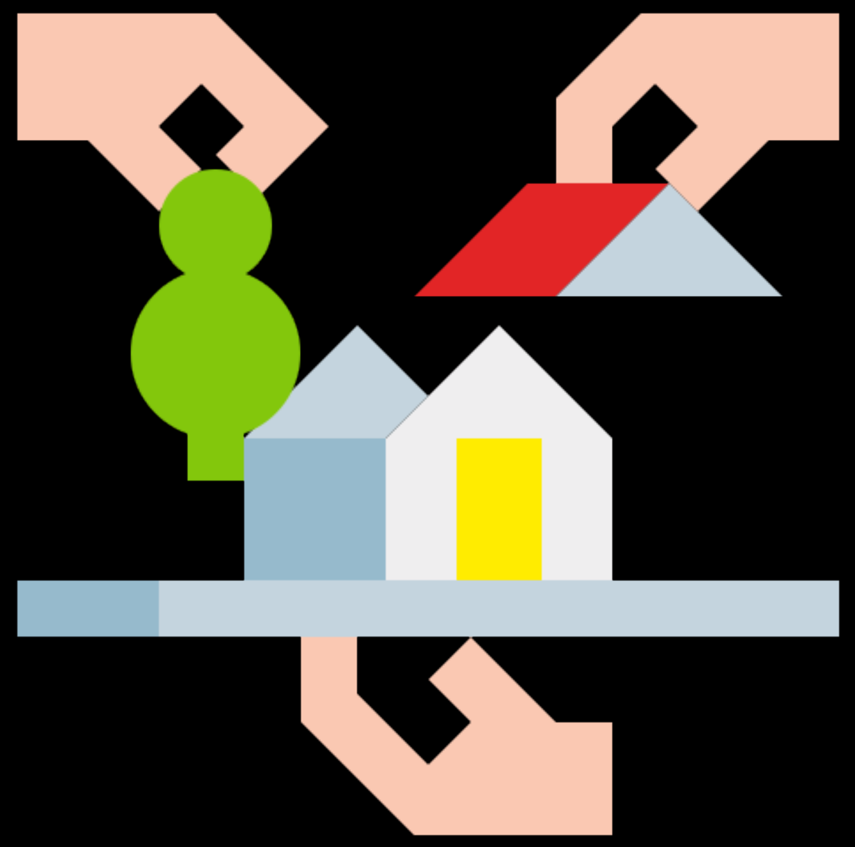
original label is not preserved

TEST INPUT GENERATION APPROACHES

RAW INPUT MANIPULATION



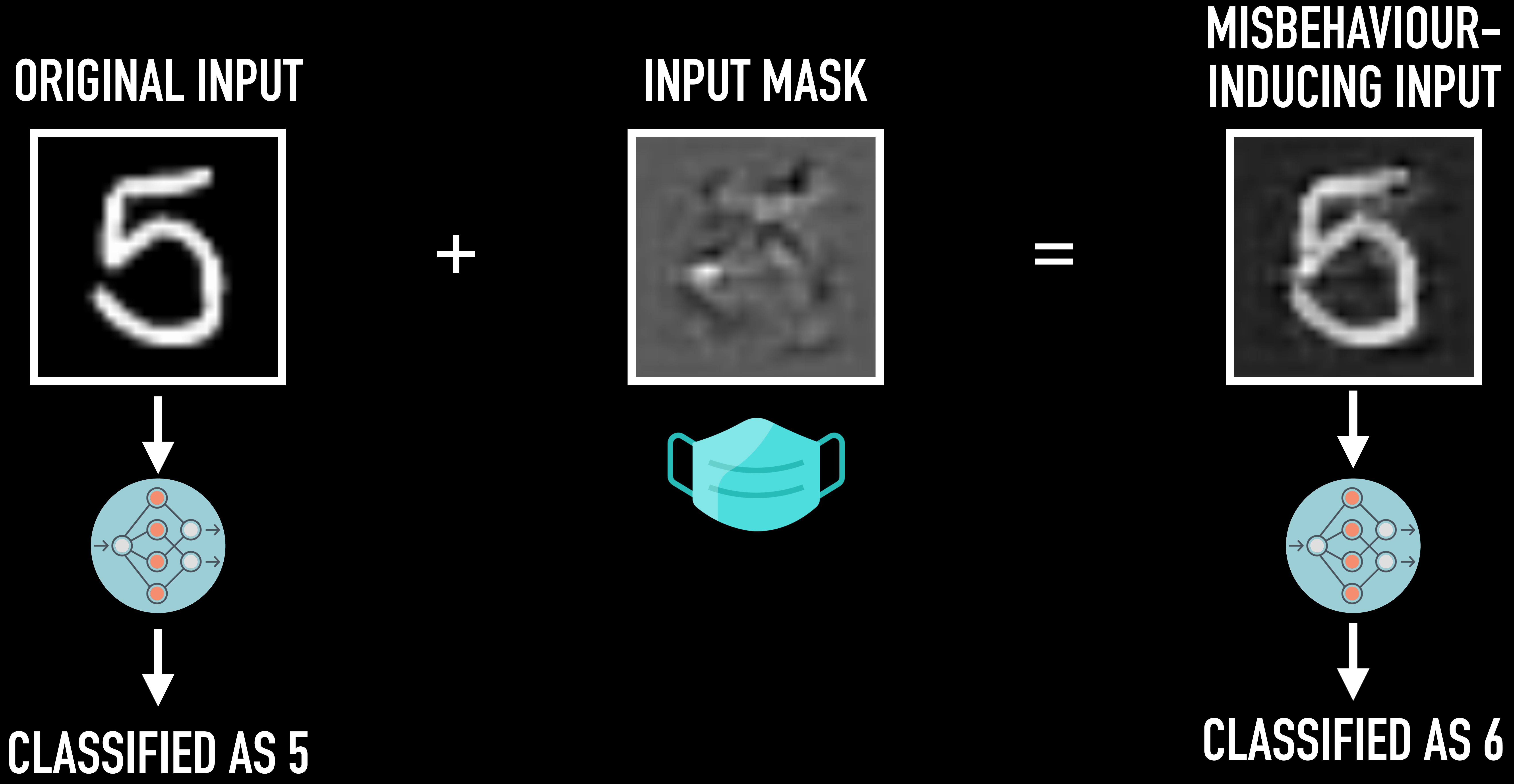
GENERATIVE DL MODELS



MODEL-BASED INPUT MANIPULATION

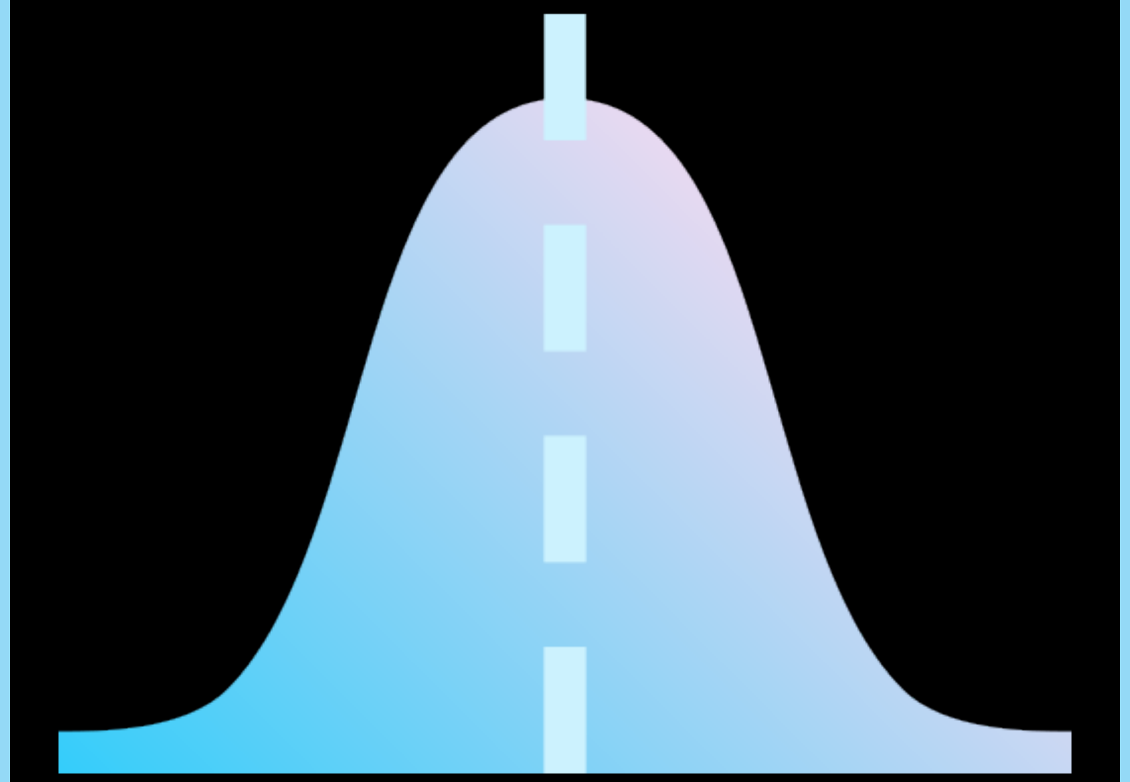
DLFUZZ

Guo et al., FSE 2018

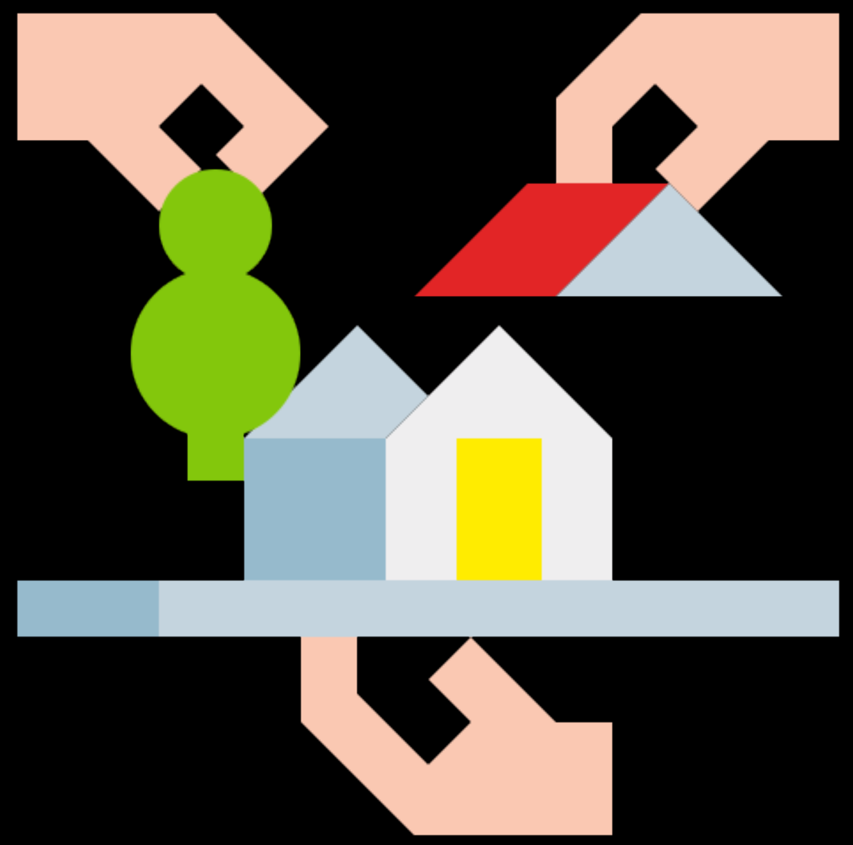


TEST INPUT GENERATION APPROACHES

RAW INPUT
MANIPULATION

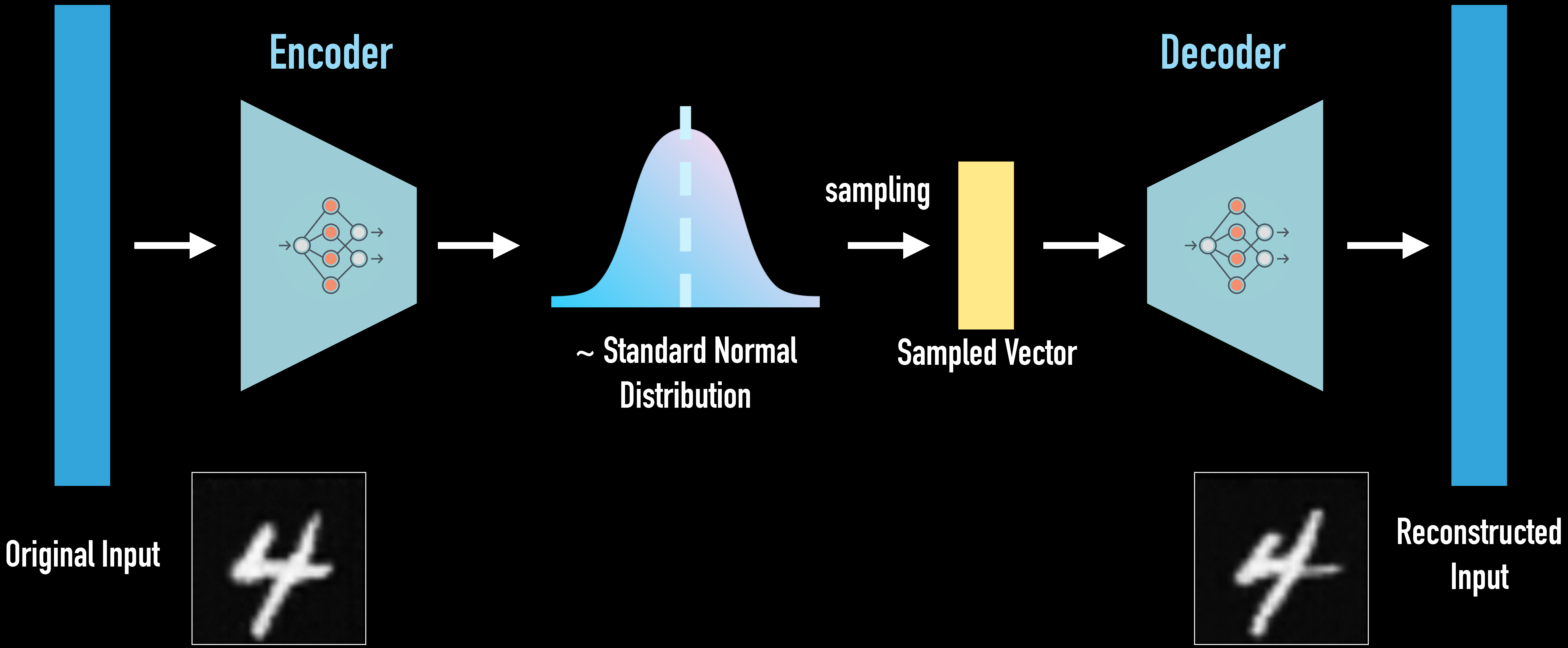


GENERATIVE DL
MODELS

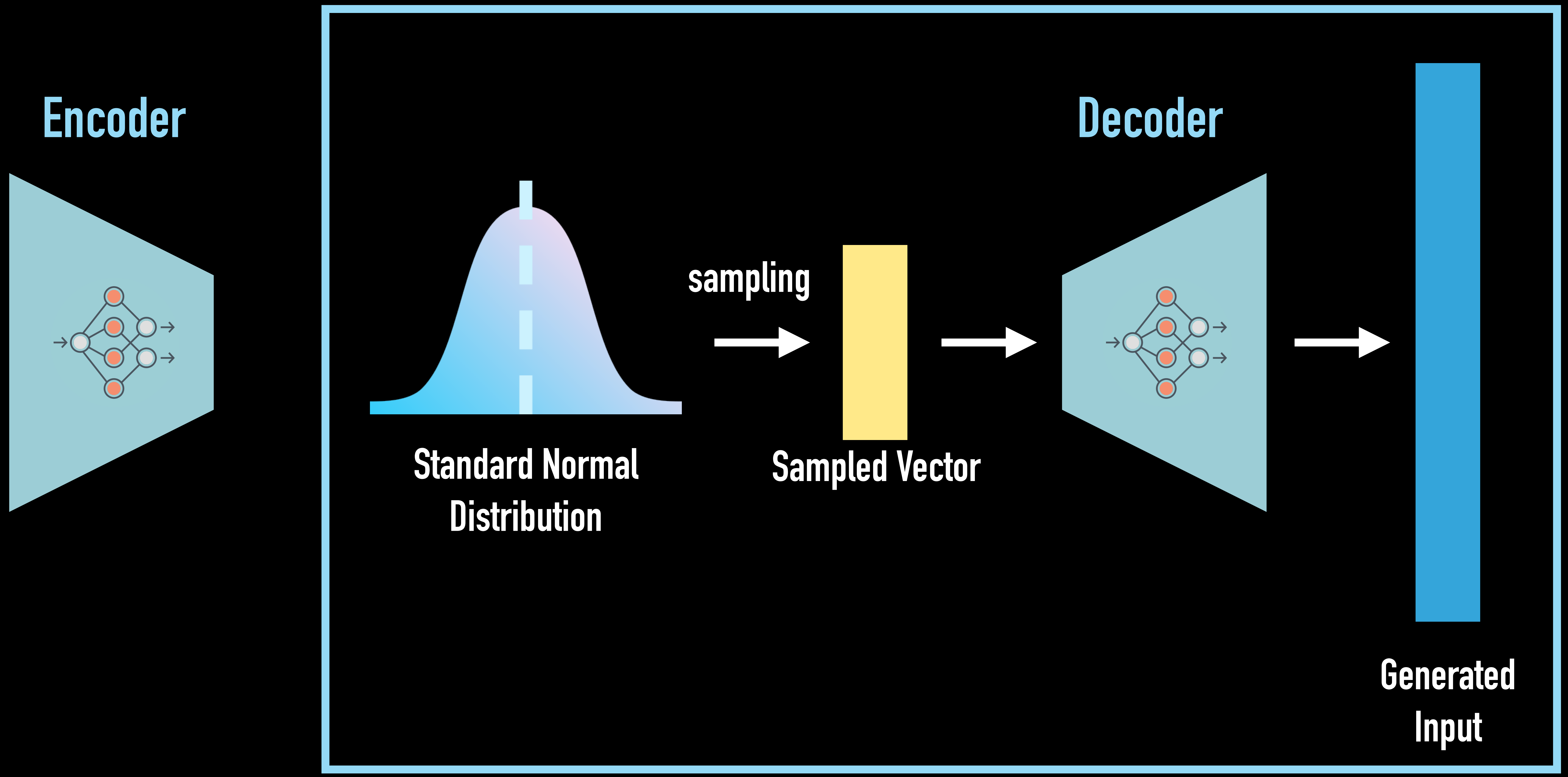


MODEL-BASED INPUT
MANIPULATION

VARIATIONAL AUTOENCODER (VAE)

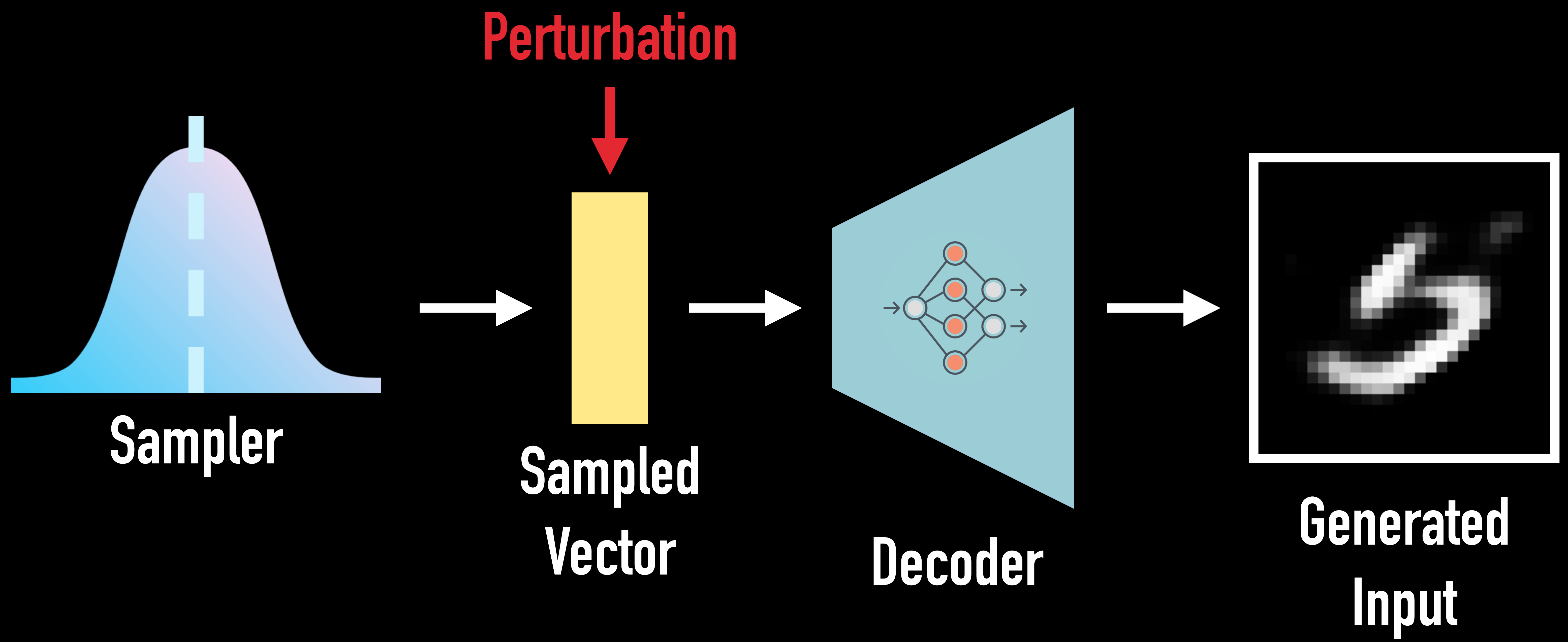


VARIATIONAL AUTOENCODER (VAE)



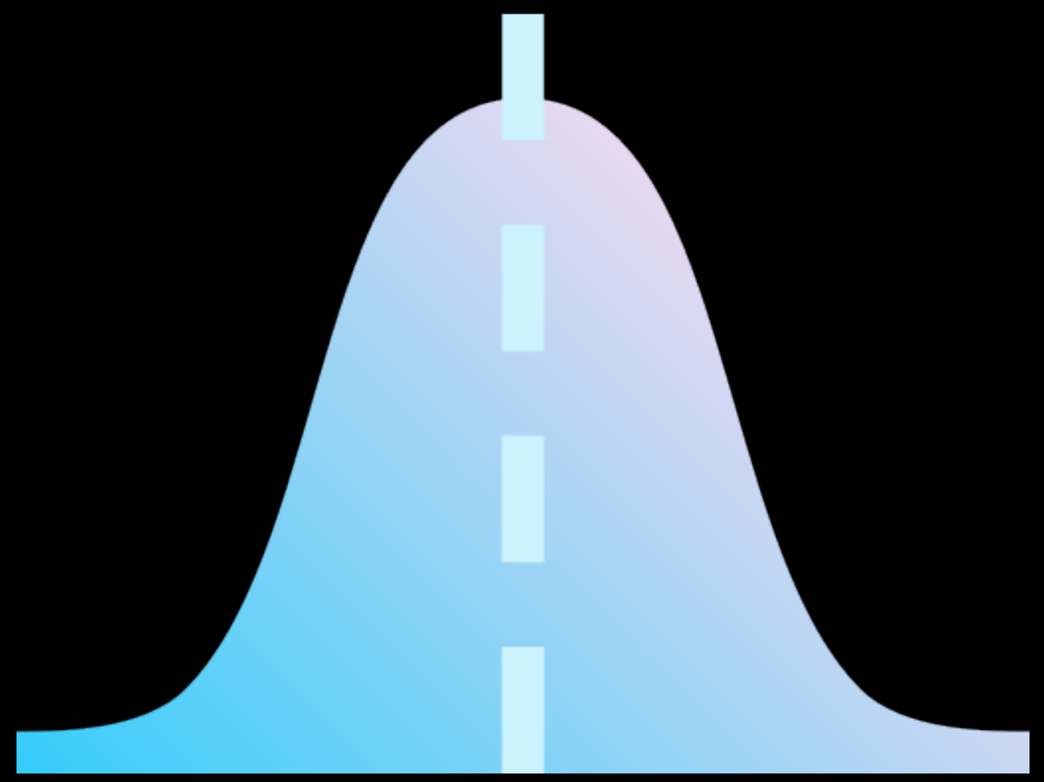
SINVAD

Kang et al., ICSE 2018

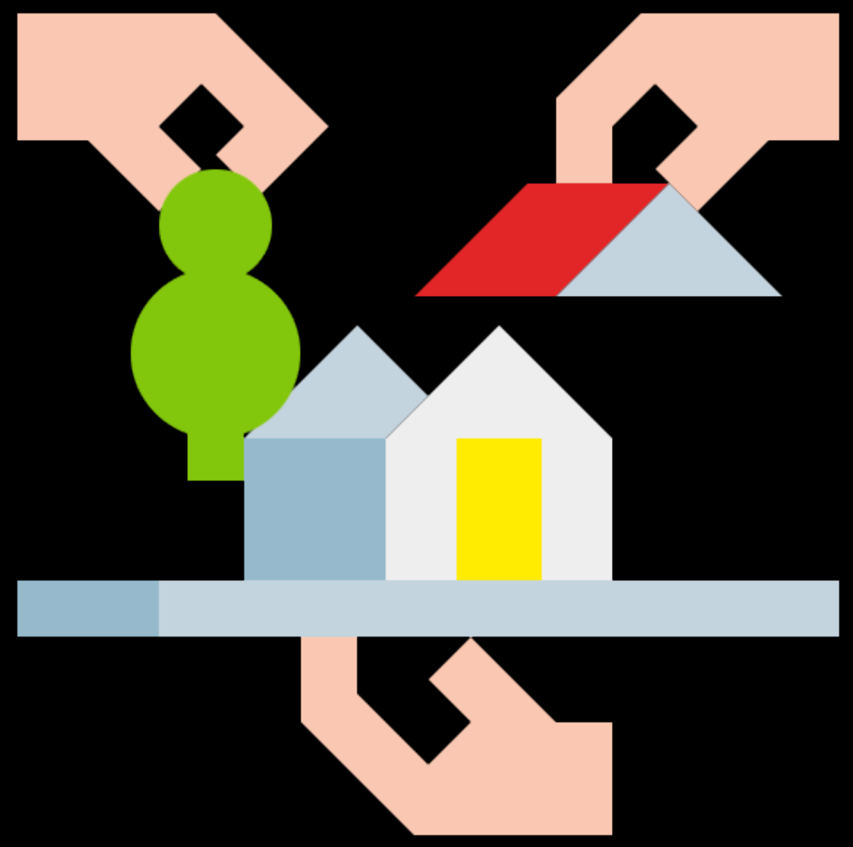


TEST INPUT GENERATION APPROACHES

RAW INPUT
MANIPULATION



GENERATIVE DL
MODELS



MODEL-BASED INPUT
MANIPULATION

DEEPJANUS

Riccio and Tonella, FSE 2020

BITMAP

SVG MODEL

start_point = (9.0, 20.85)
BezierSegment(
c1 = (9.0, 20.22),
c2 = (10.22, 17.30),
end_point = (11.70, 14.38)
)

SVG MODEL

start_point = (9.0, 20.85)
BezierSegment(
c1 = (9.0, 20.22),
c2 = (**8.10**, 17.30),
end_point = (11.70, 14.38)
)

BITMAP

ARE INPUTS PRODUCED BY TEST INPUT GENERATORS VALID?

ARE THE GENERATED VALID INPUTS LABEL-PRESERVING?

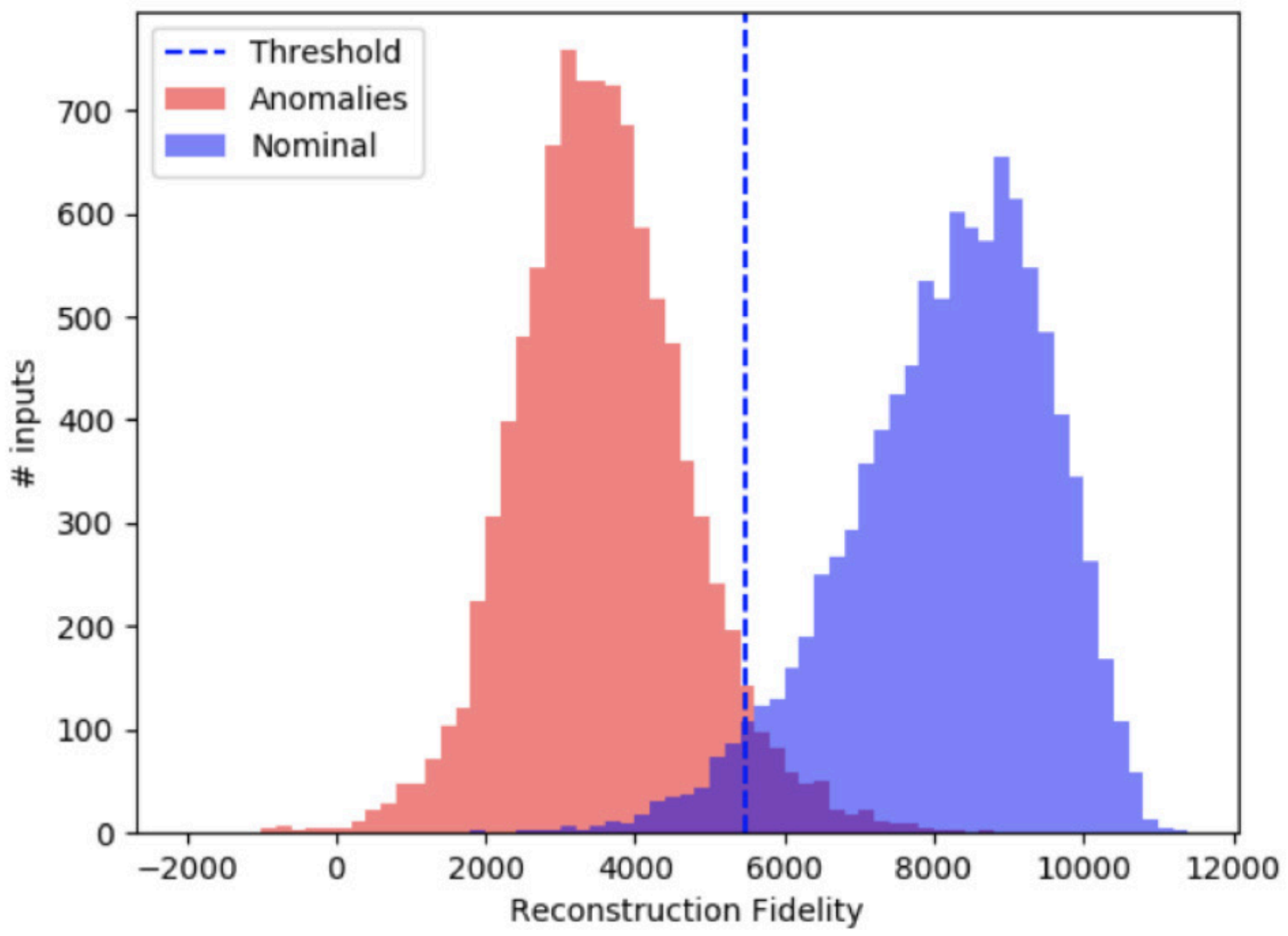
HUMAN ASSESSMENT

220 HUMAN ASSESSORS FROM

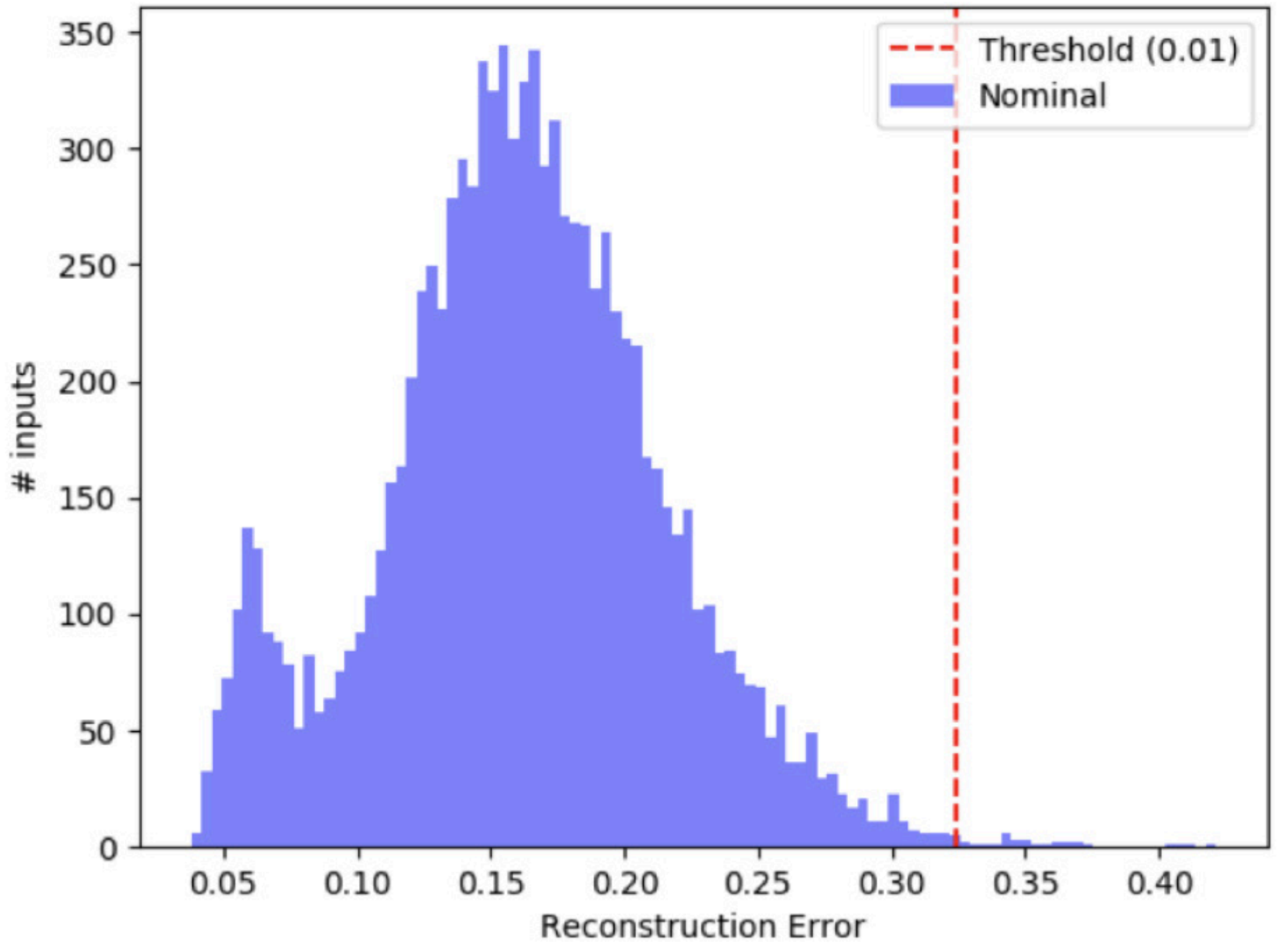
amazon mechanical turk

AUTOMATED ASSESSMENT

DAIV
[DOLA ET AL., ICSE 2021]



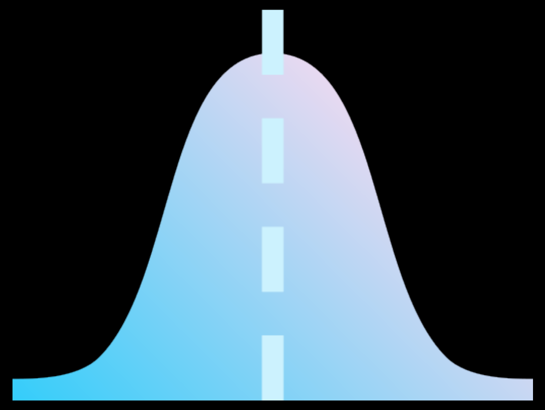
SELFORACLE
[STOCCO ET AL., ICSE 2020]



EMPIRICAL STUDY: TEST GENERATORS

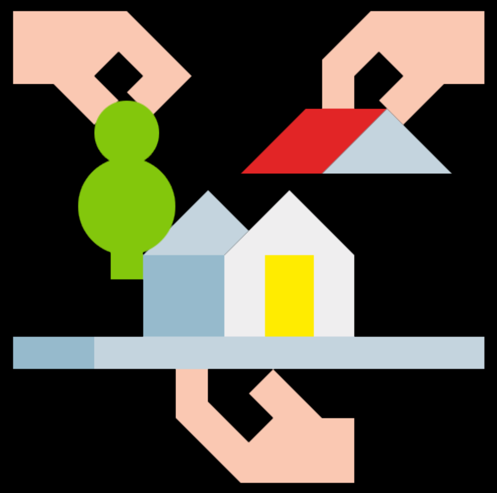
RAW INPUT MANIPULATION

- DEEPXPLORE [PEI ET AL., SOSP 2017]
- DLFUZZ [GUO ET AL., FSE 2018]



GENERATIVE DL MODELS

- SINVAD [KANG ET AL., ICSE 2018]
- FEATURE PERTURBATIONS [DUNN ET AL., ISSTA 2021]



MODEL-BASED INPUT MANIPULATION

- DEEPJANUS [RICCIO AND TONELLA, FSE 2020]

DATASETS

MNIST

SVHN

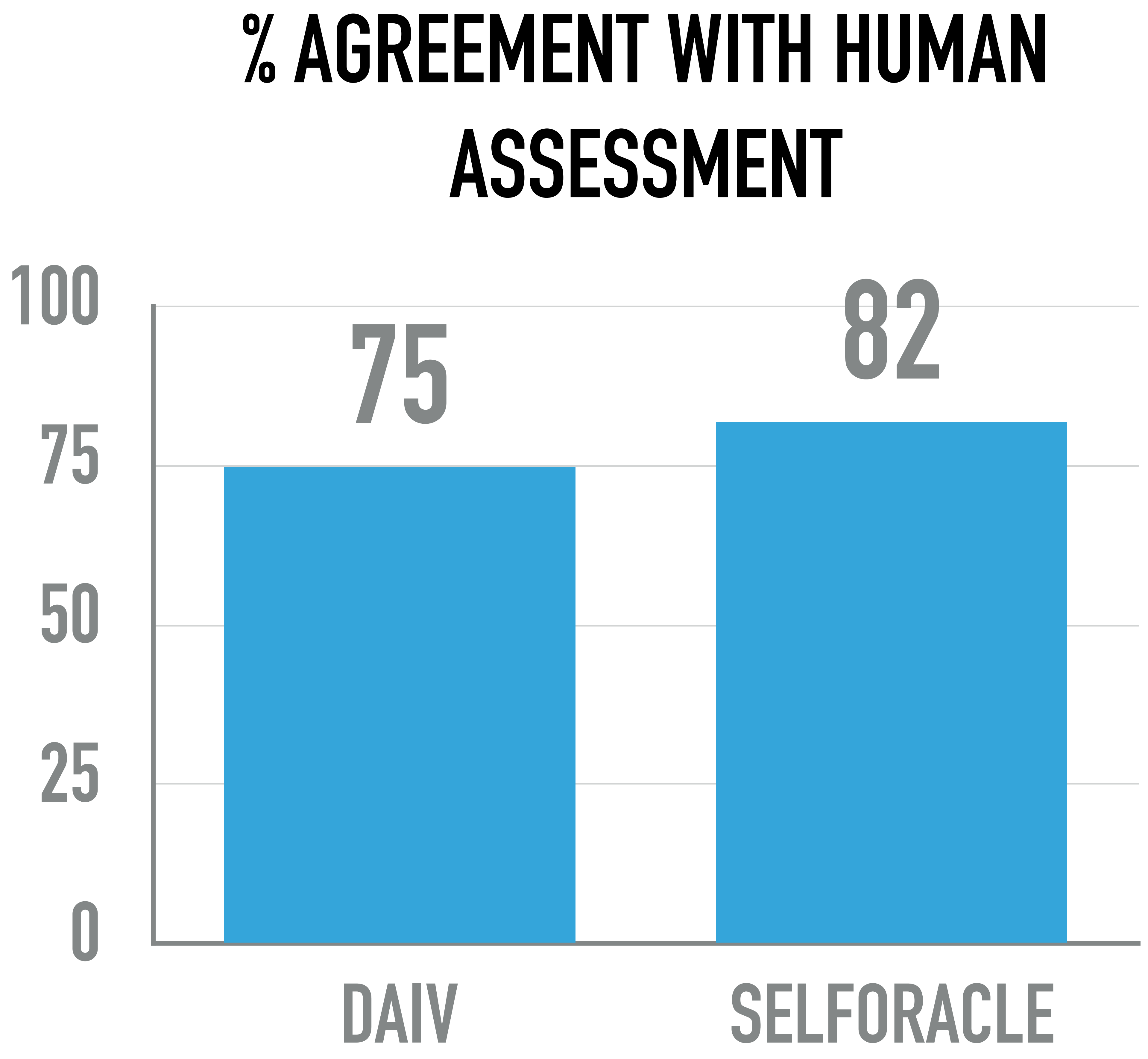
IMAGENET-1K

HUMAN ASSESSMENT

88% VALID INPUTS

69% PRESERVED LABELS

AUTOMATED ASSESSMENT

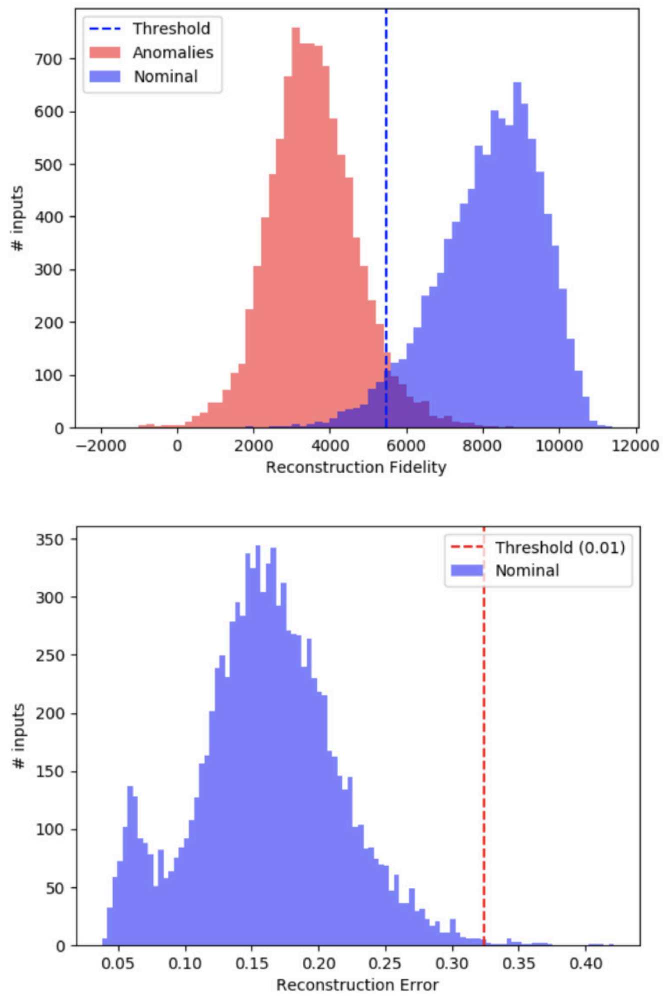


LESSONS LEARNT

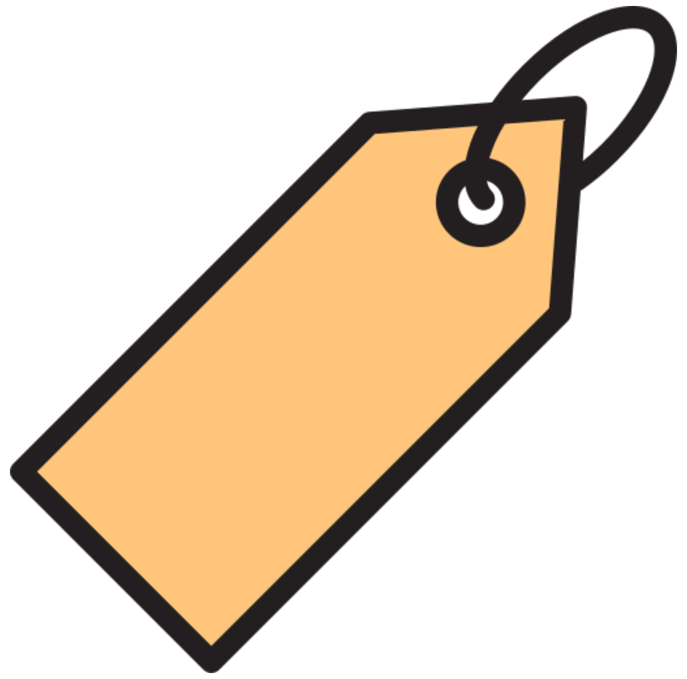
Too aggressive raw data manipulations lead to invalid inputs

Generative DL models should carefully explore the latent space

Model-based techniques require high-quality model representations

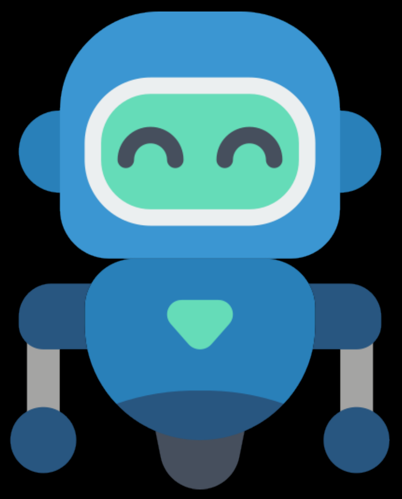


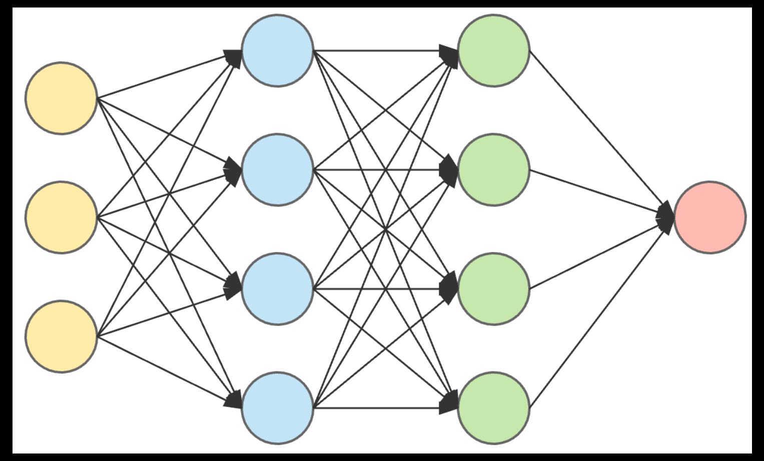
Automated validators check in-distribution

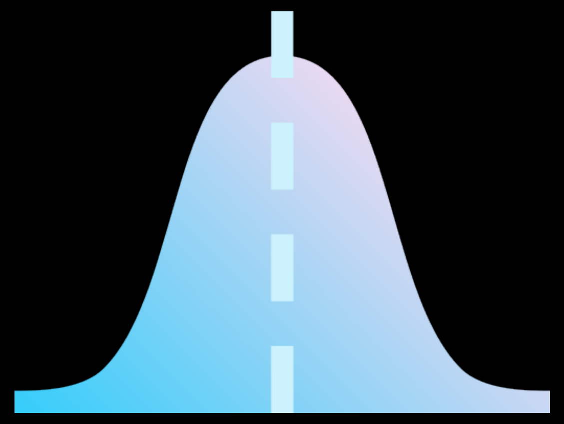


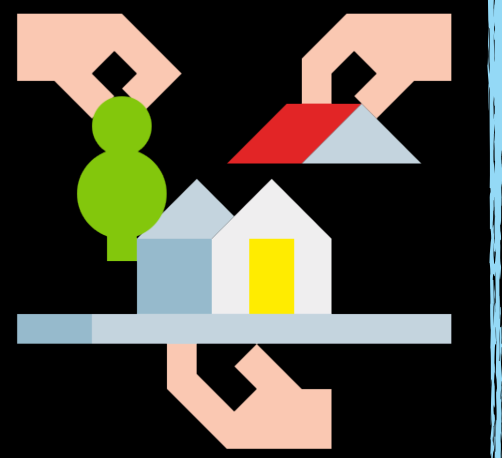
Label preservation is mostly overlooked

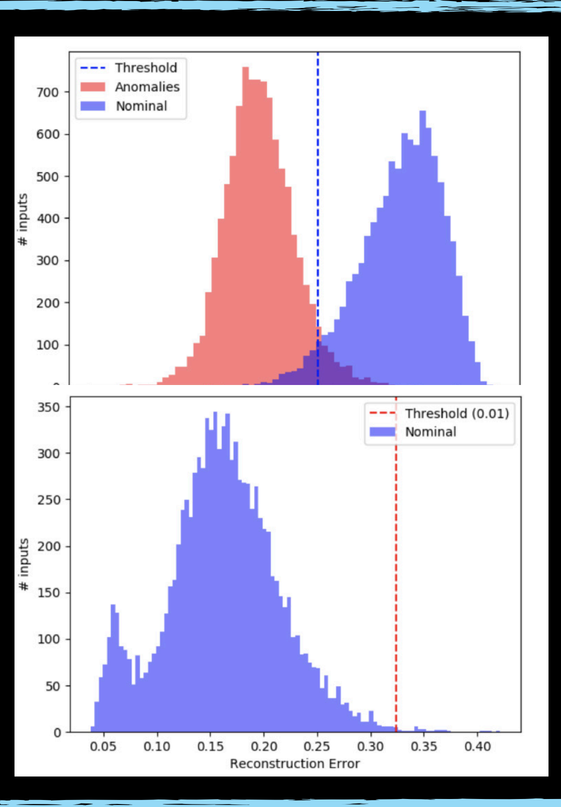
SUMMARY











testingautomated-usi/
tig-validity-icse23

Icons from www.flaticon.com

EXTRA SLIDES

Dataset	Tool	% Valid DAIV	% Valid SO	% Valid Human	% Hum. Agree.	% Pres. Labels
MNIST	DX	35	35	81	93	92
	DLF	37	85	**100**	**99**	**99**
	SV	**97**	**100**	**100**	96	58
	FPT	90	**100**	99	94	54
	DJ	**97**	**100**	**100**	96	93
SVHN	DX	51	99	77	68	**79**
	DLF	**100**	**100**	**96**	73	50
	SV	**100**	**100**	90	74	9
	FPT	99	**100**	81	75	46
	DJ	0	1	61	**76**	9
ImageNet-1K	DX	**100**	**100**	90	**100**	94
	DLF	**100**	**100**	**100**	**100**	95
	SV	**100**	**100**	60	50	83
	FPT	**100**	**100**	**100**	**100**	**100**

EXTRA SLIDES