Università
della
Svizzera
italiana

Software Institute

MODEL-BASED EXPLORATION OF THE FRONTIER OF BEHAVIOURS FOR DEEP LEARNING SYSTEM TESTING

VINCENZO RICCIO

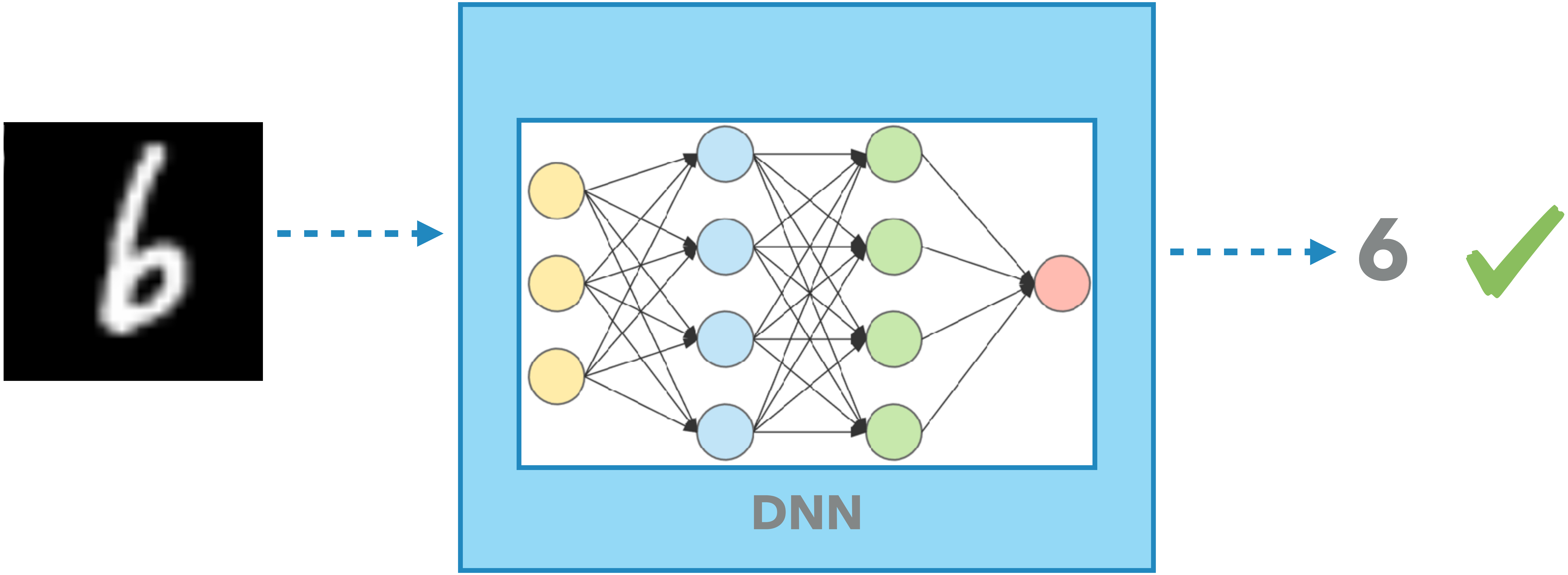
@p1ndsvin

PAOLO TONELLA

@paolo_tonella

ESEC/FSE 2020

DEEP LEARNING (DL) SYSTEM

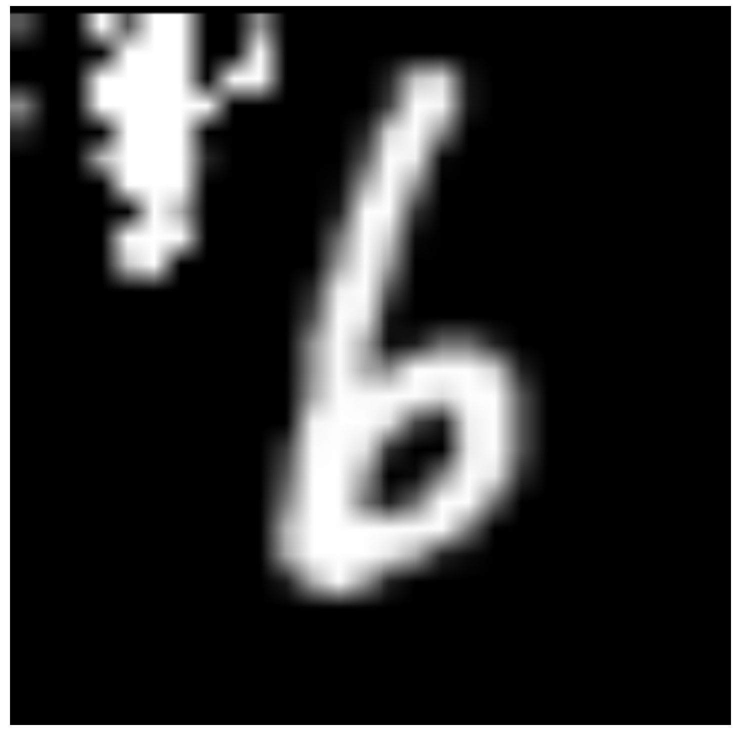


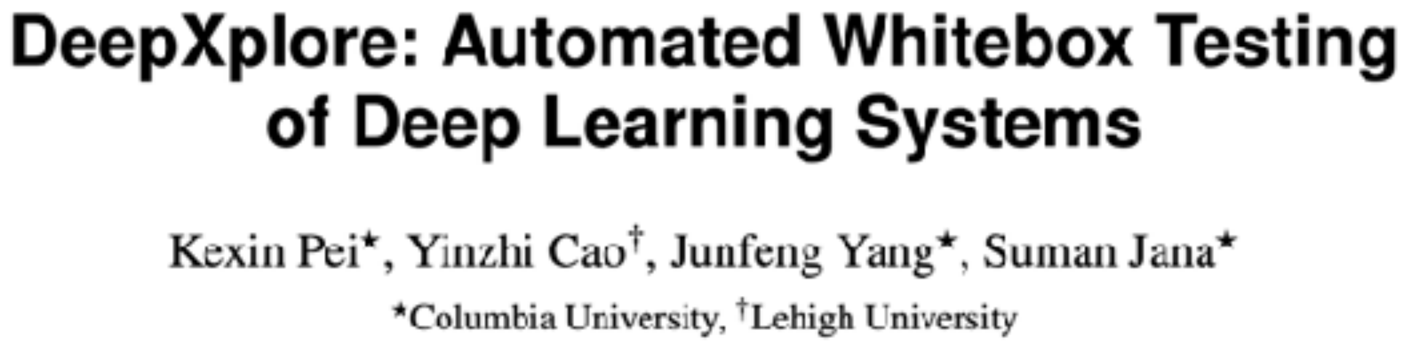
6 ✓

DNN

DL SYSTEM

TESTING DL SYSTEMS

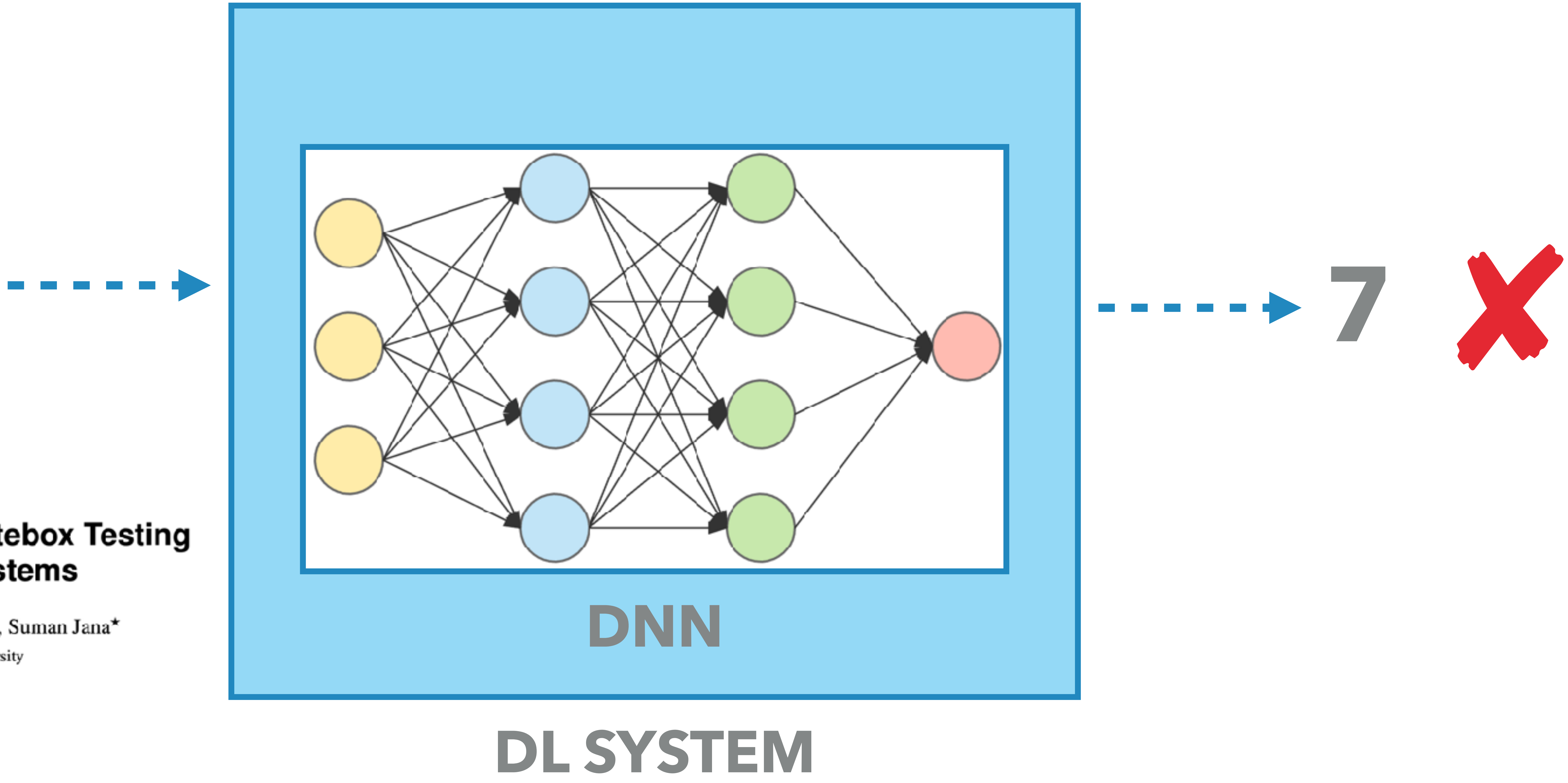




DeepXplore: Automated Whitebox Testing of Deep Learning Systems

Kexin Pei*, Yinzhi Cao†, Junfeng Yang*, Suman Jana*

*Columbia University, †Lehigh University



DNN

DL SYSTEM

7

TO TRULY ASSESS

THE **QUALITY** OF DL SYSTEMS

WE NEED TO EVALUATE THEIR **BEHAVIOUR**

AT THE **FRONTIER**

BY GENERATING **VALID** INPUTS

REALISM: MODEL-BASED INPUT REPRESENTATION

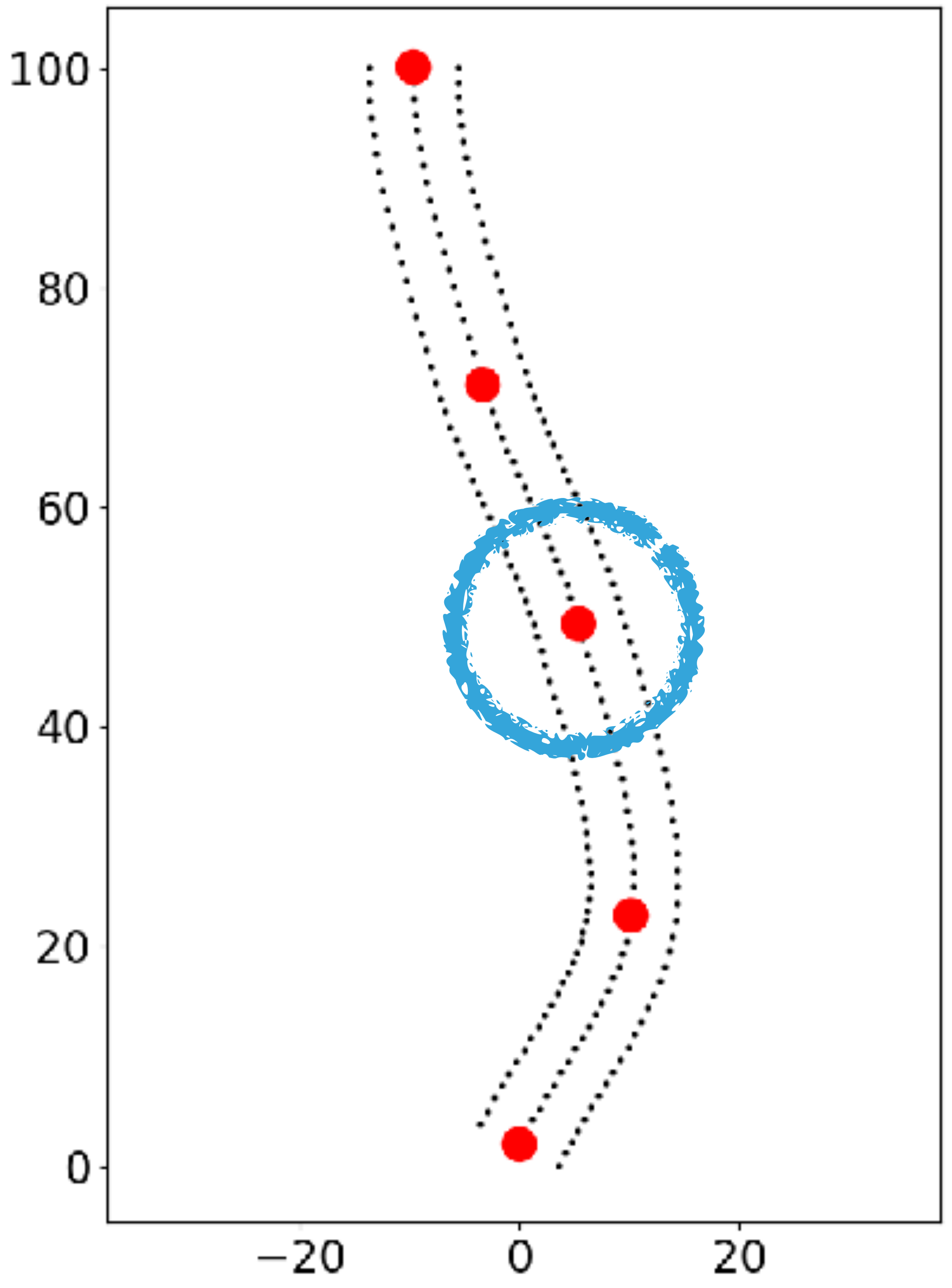
Bitmap

SVG model

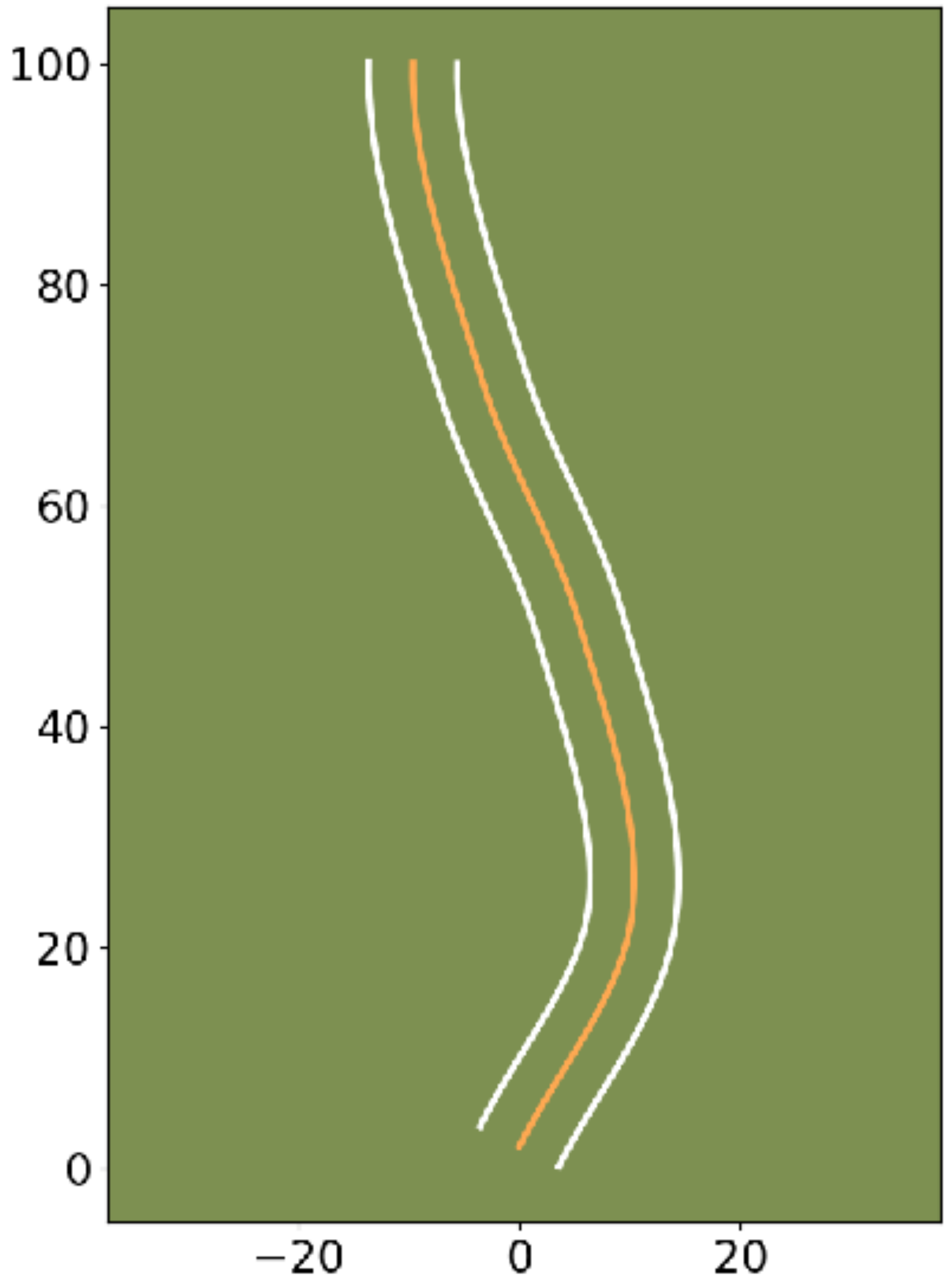
Model

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

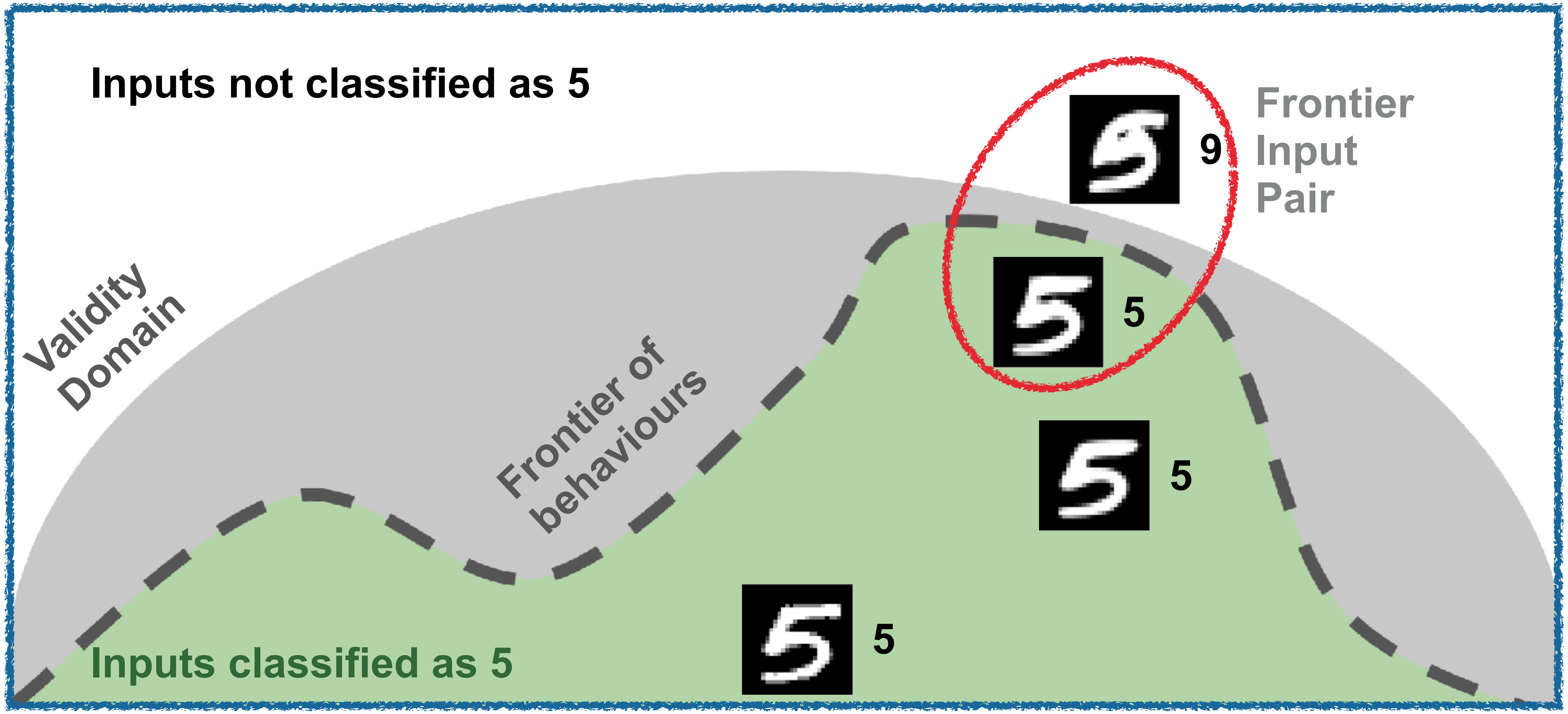
Model



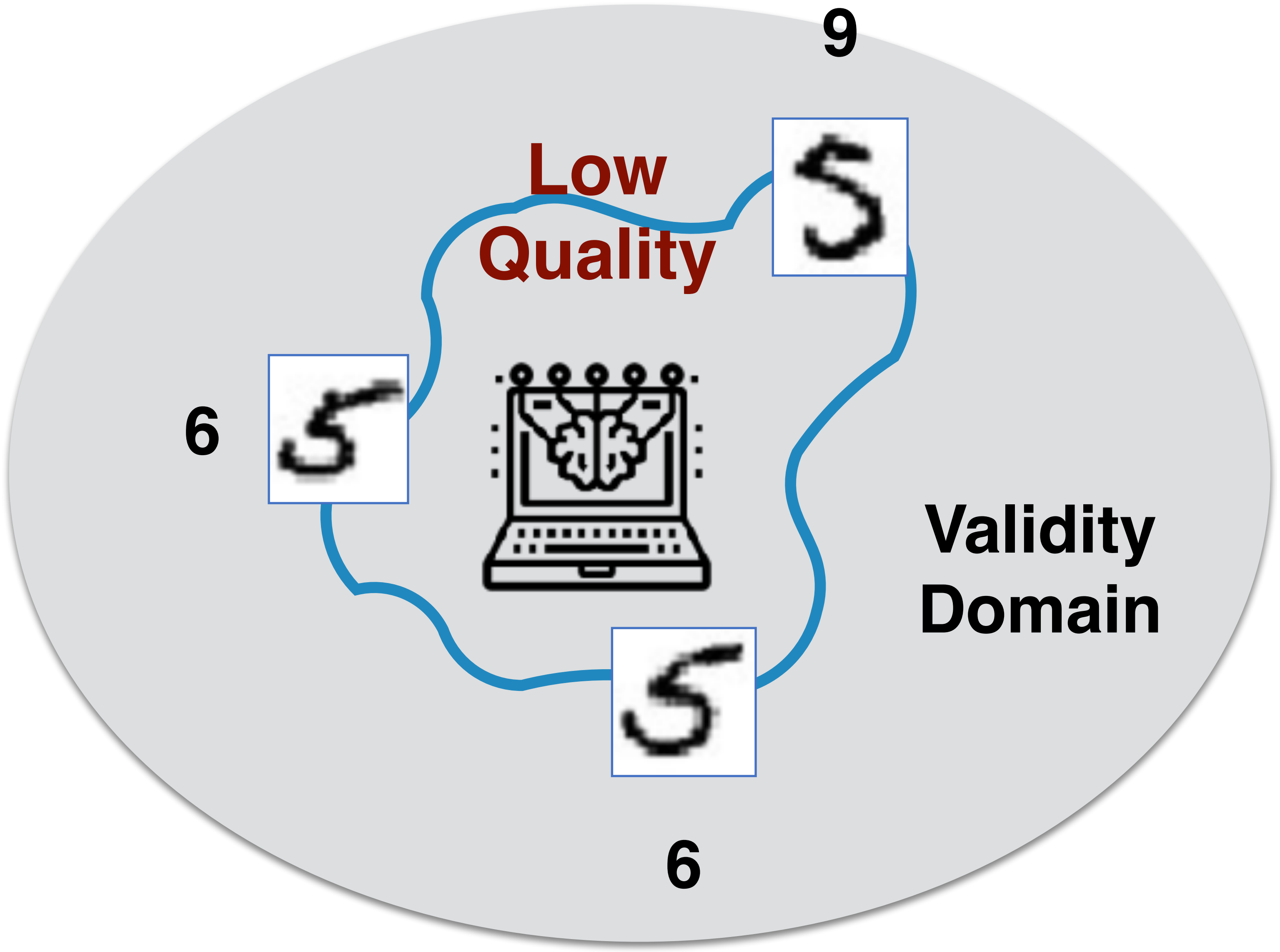
Road

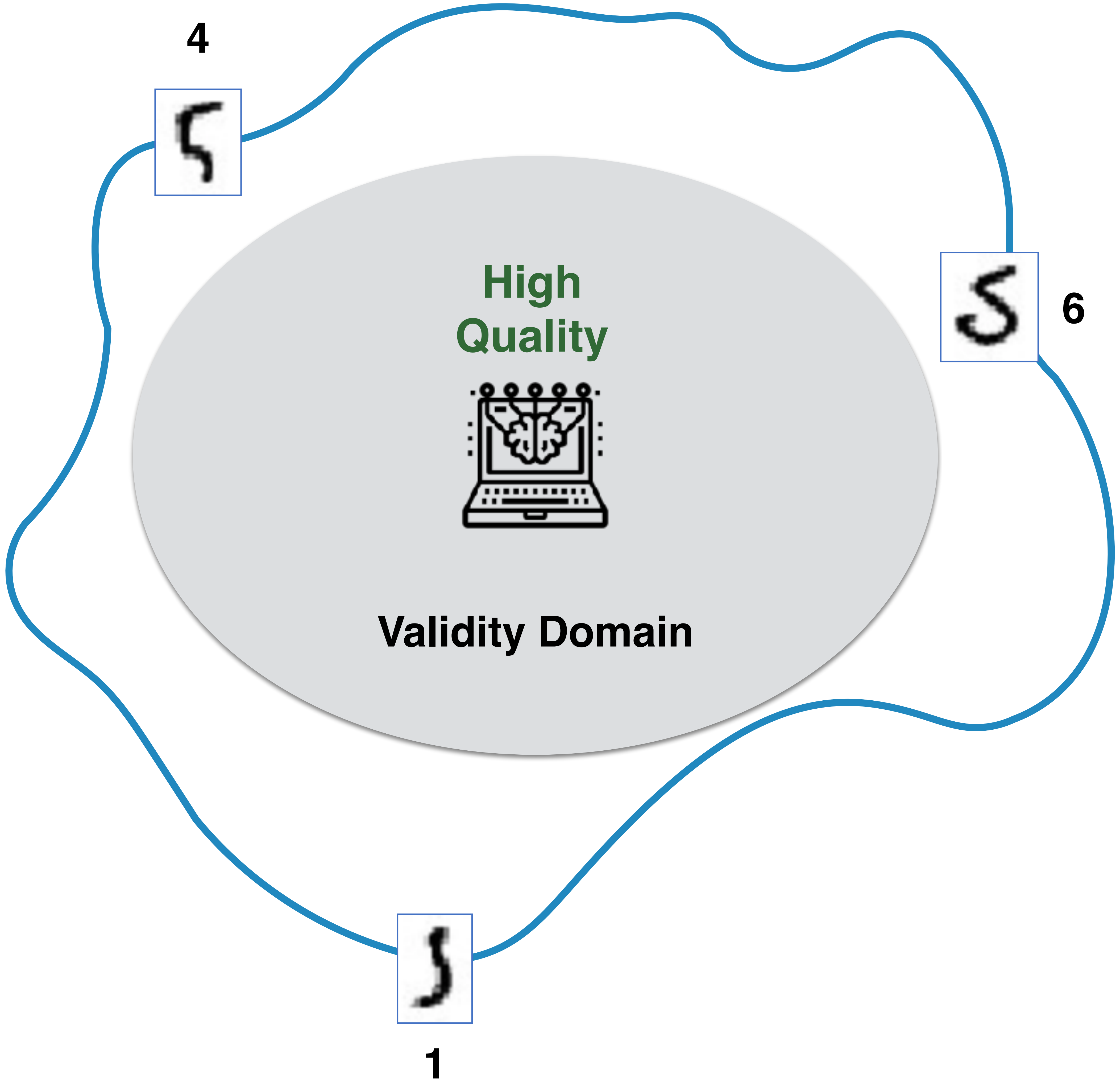


FRONTIER OF BEHAVIOURS



FRONTIER AND VALIDITY DOMAIN

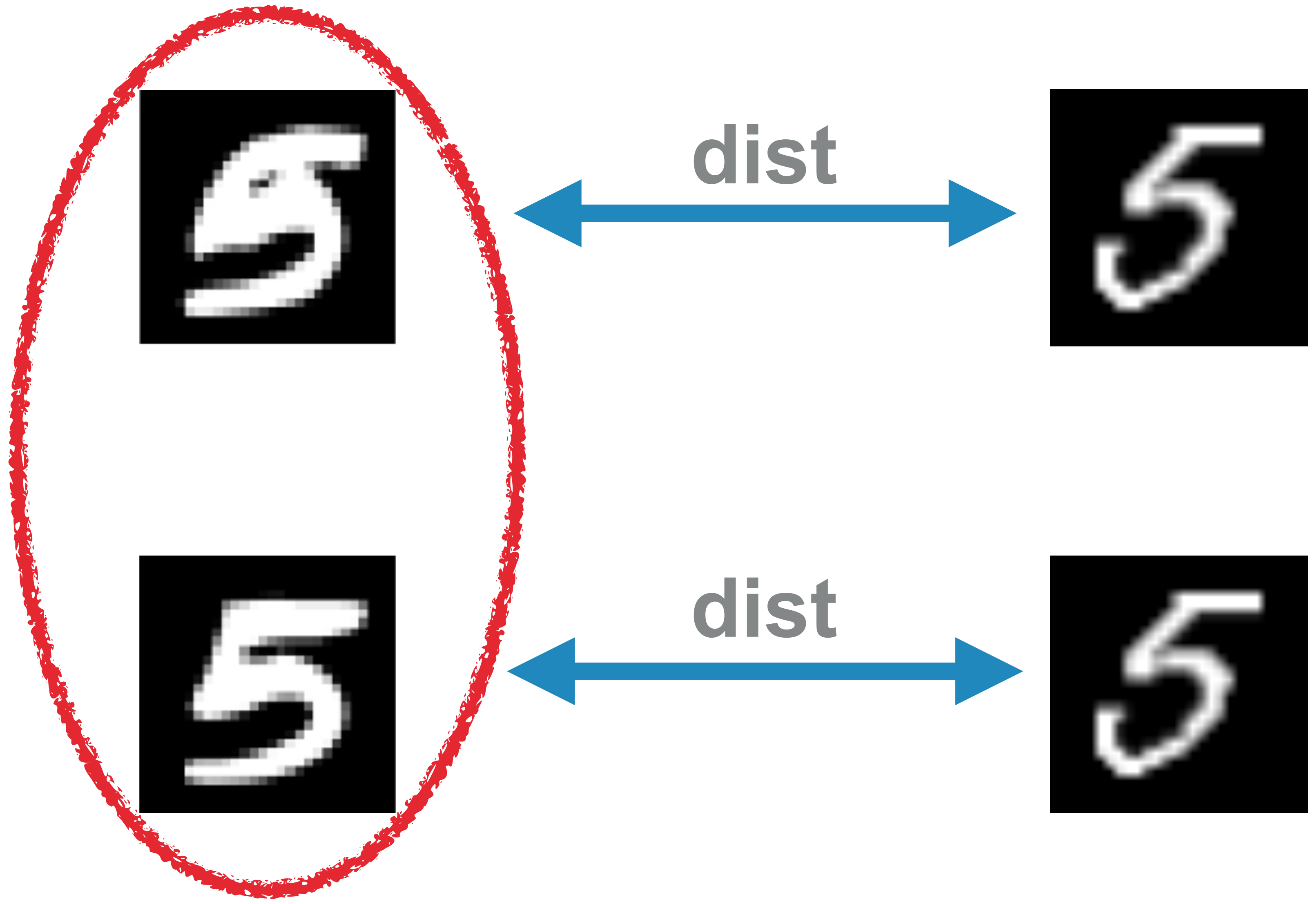




QUANTITATIVE ASSESSMENT

Frontier Input Pair
[m1, m2] ∈ S

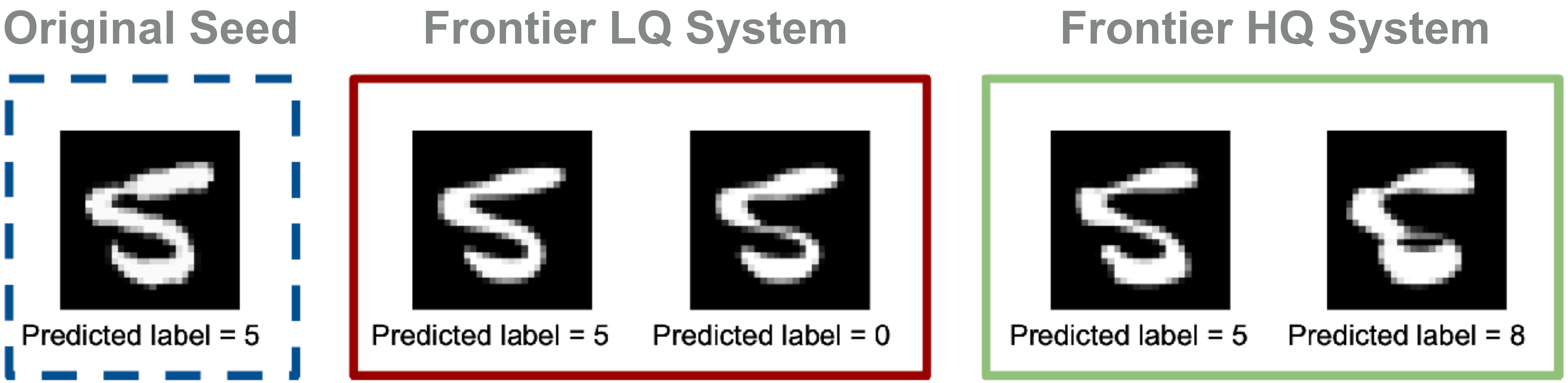
Reference Ω



Frontier Radius

$$\text{radius}(S) = \frac{\sum_{m \in S} \text{dist}(m, \Omega)}{|S|}$$

QUALITATIVE ASSESSMENT



Original Seed

Predicted label = 5

Frontier LQ System

Predicted label = 5

Predicted label = 0

Frontier HQ System

Predicted label = 5

Predicted label = 8

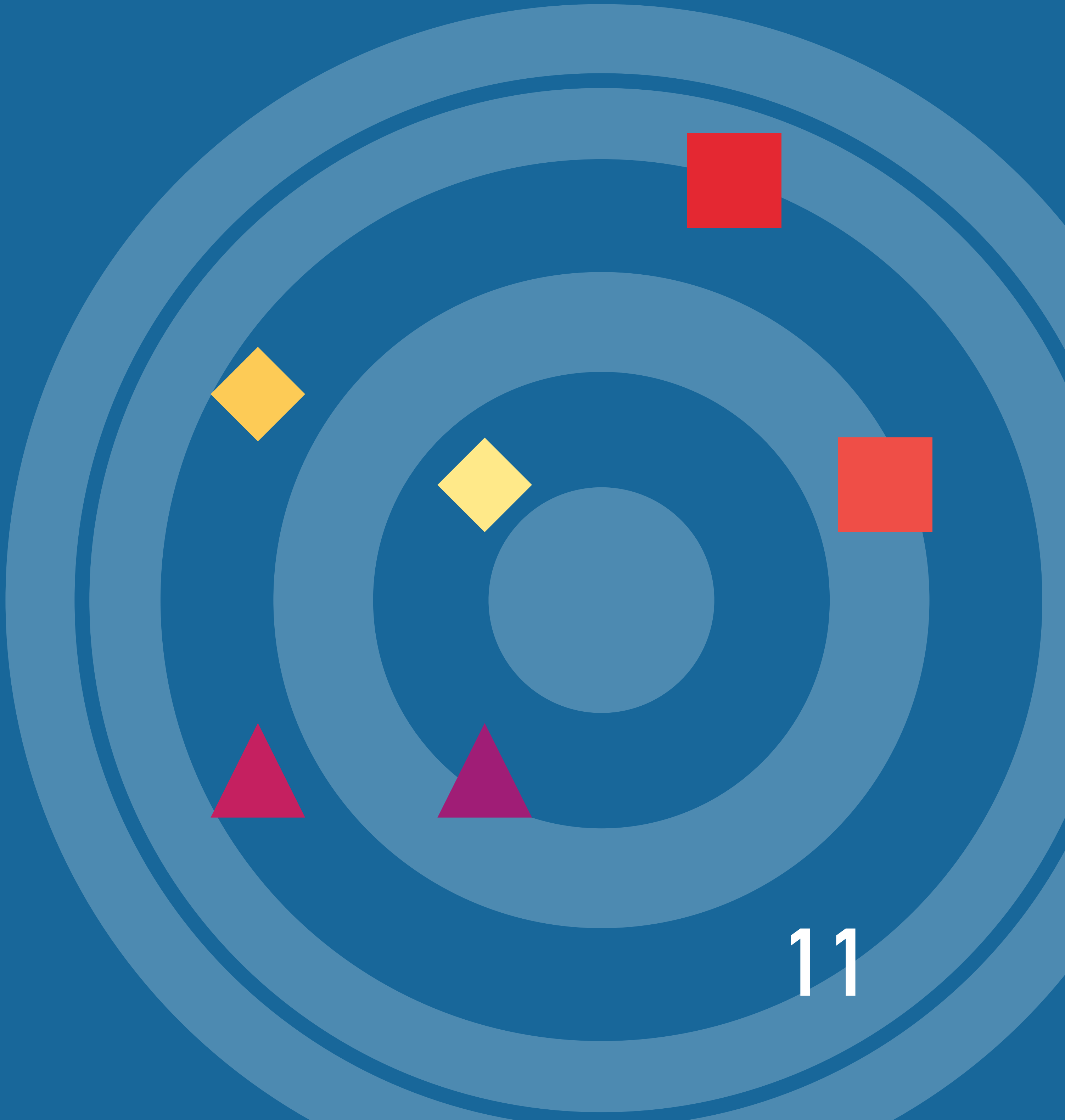
PROPOSED APPROACH

GENERATING A SET OF FRONTIER INPUT PAIRS

PROPOSED APPROACH

GENERATING A SET OF FRONTIER INPUT PAIRS

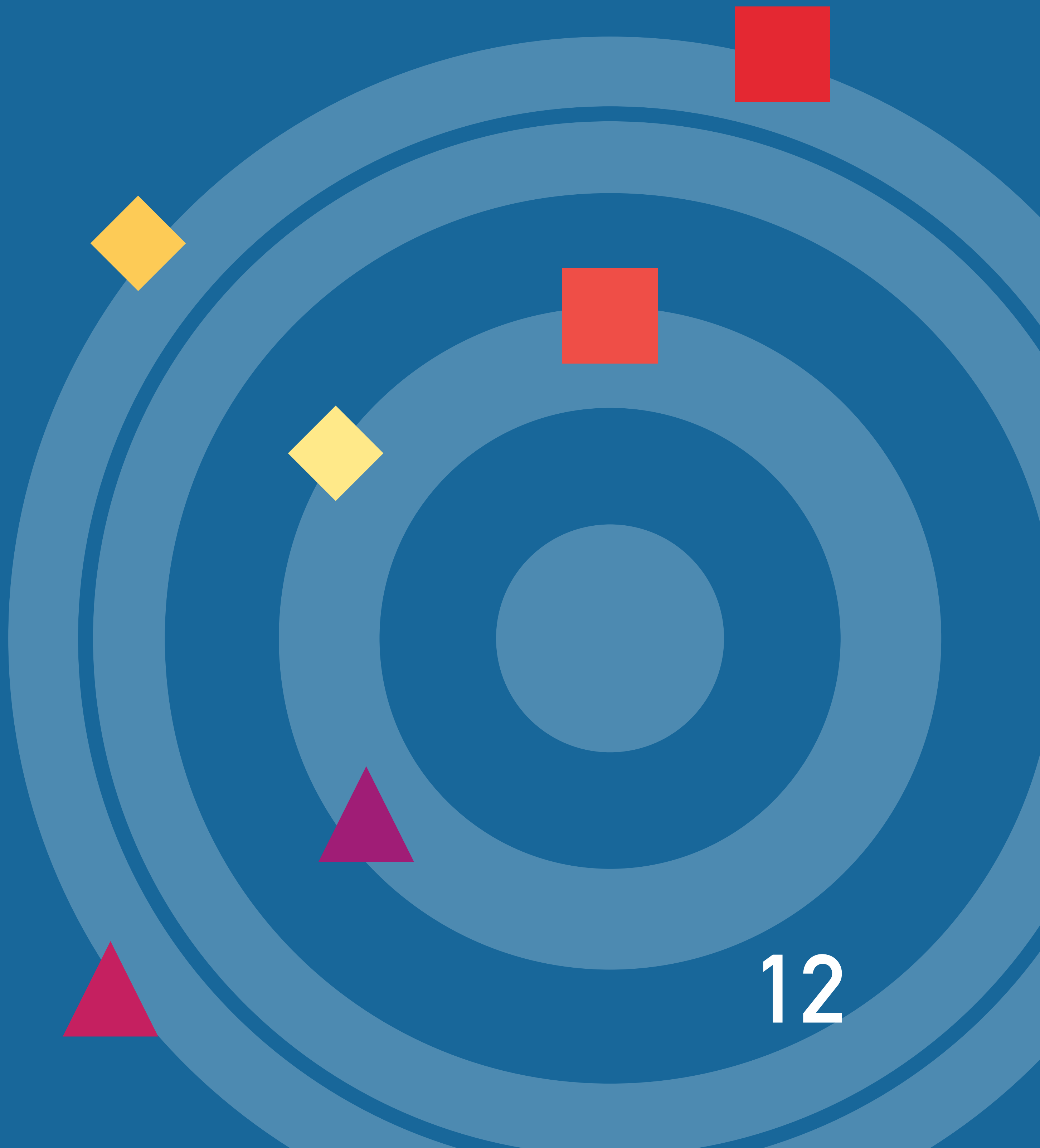
1. **DIVERSIFY** THE GENERATED SOLUTIONS



PROPOSED APPROACH

GENERATING A SET OF FRONTIER INPUT PAIRS

1. **DIVERSIFY** THE GENERATED SOLUTIONS
2. **MINIMIZE** THE **DISTANCE** TO THE **FRONTIER**

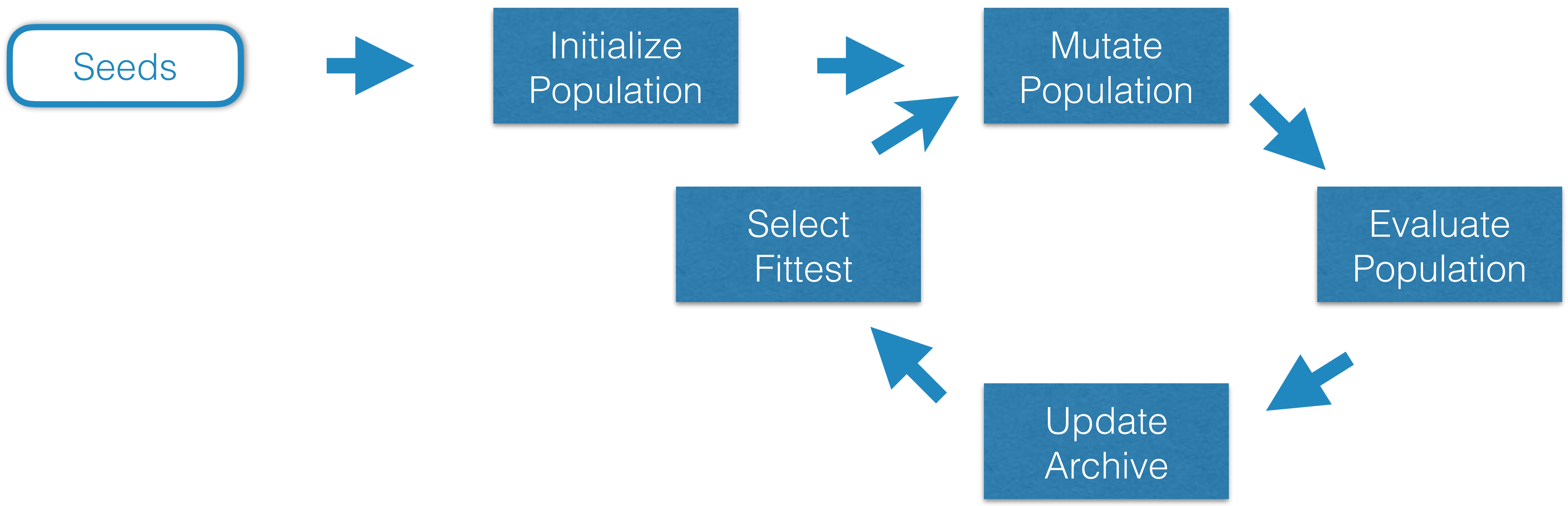


PROPOSED APPROACH

GENERATING A SET OF FRONTIER INPUT PAIRS

1. DIVERSIFY THE GENERATED SOLUTIONS
2. MINIMIZE THE DISTANCE TO THE FRONTIER
3. MAXIMIZE THE INTRA-PAIR SIMILARITY

DEEPJANUS



Seeds

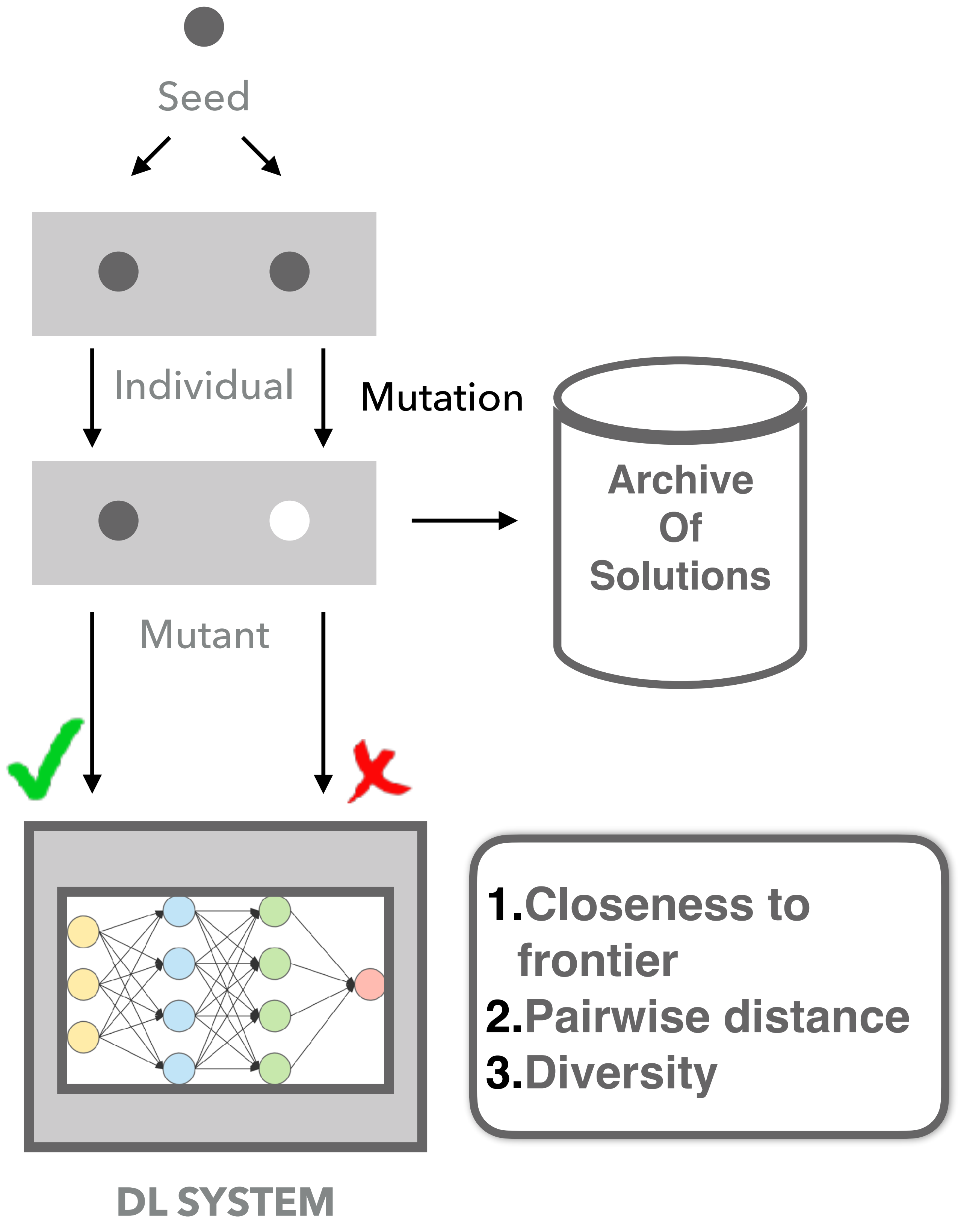
Initialize Population

Mutate Population

Evaluate Population

Update Archive

Select Fittest



Seed

Individual

Mutation

Mutant

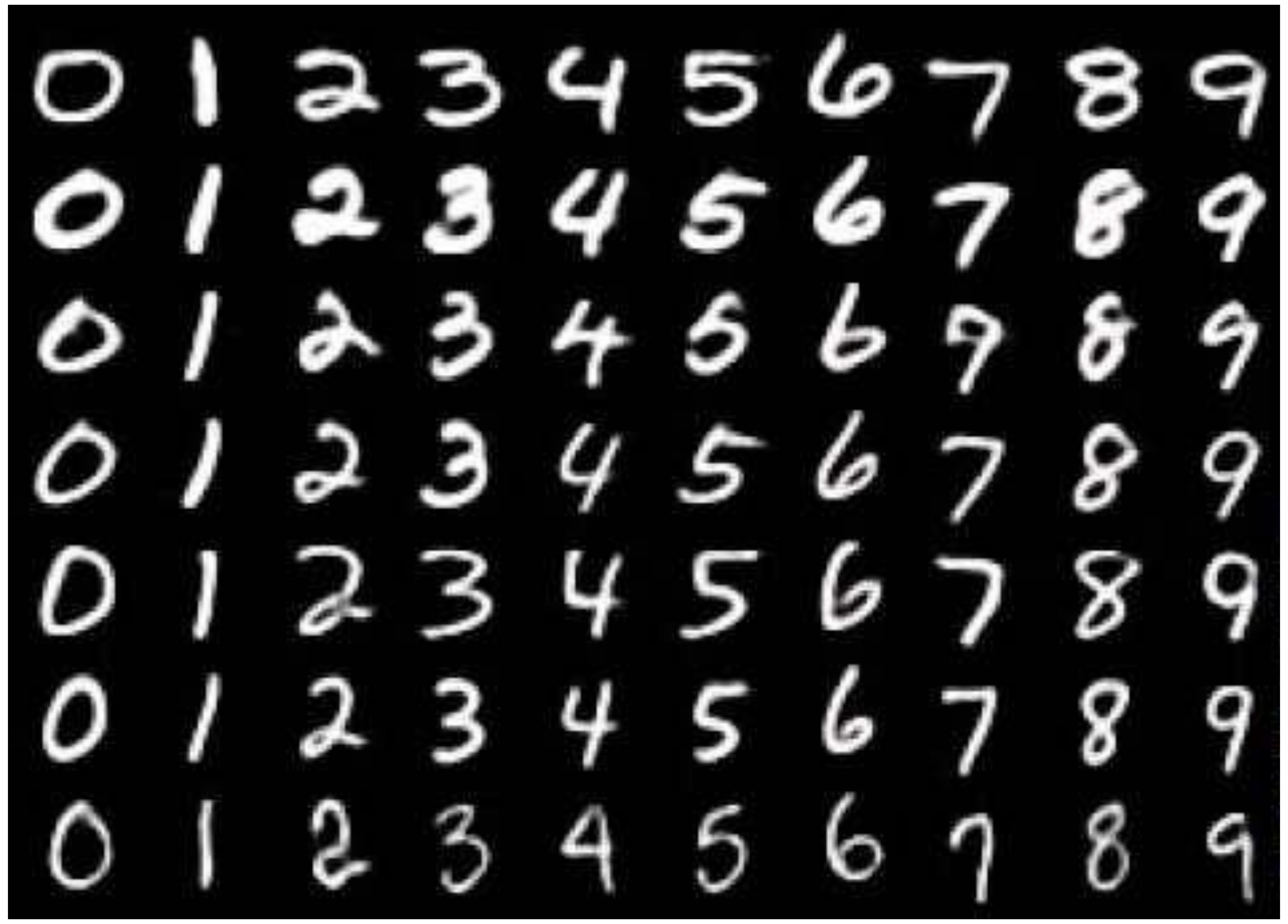
Archive Of Solutions

DL SYSTEM

1. Closeness to frontier
2. Pairwise distance
3. Diversity

EXPERIMENTAL EVALUATION

MNIST



BEAMNG

EFFECTIVENESS

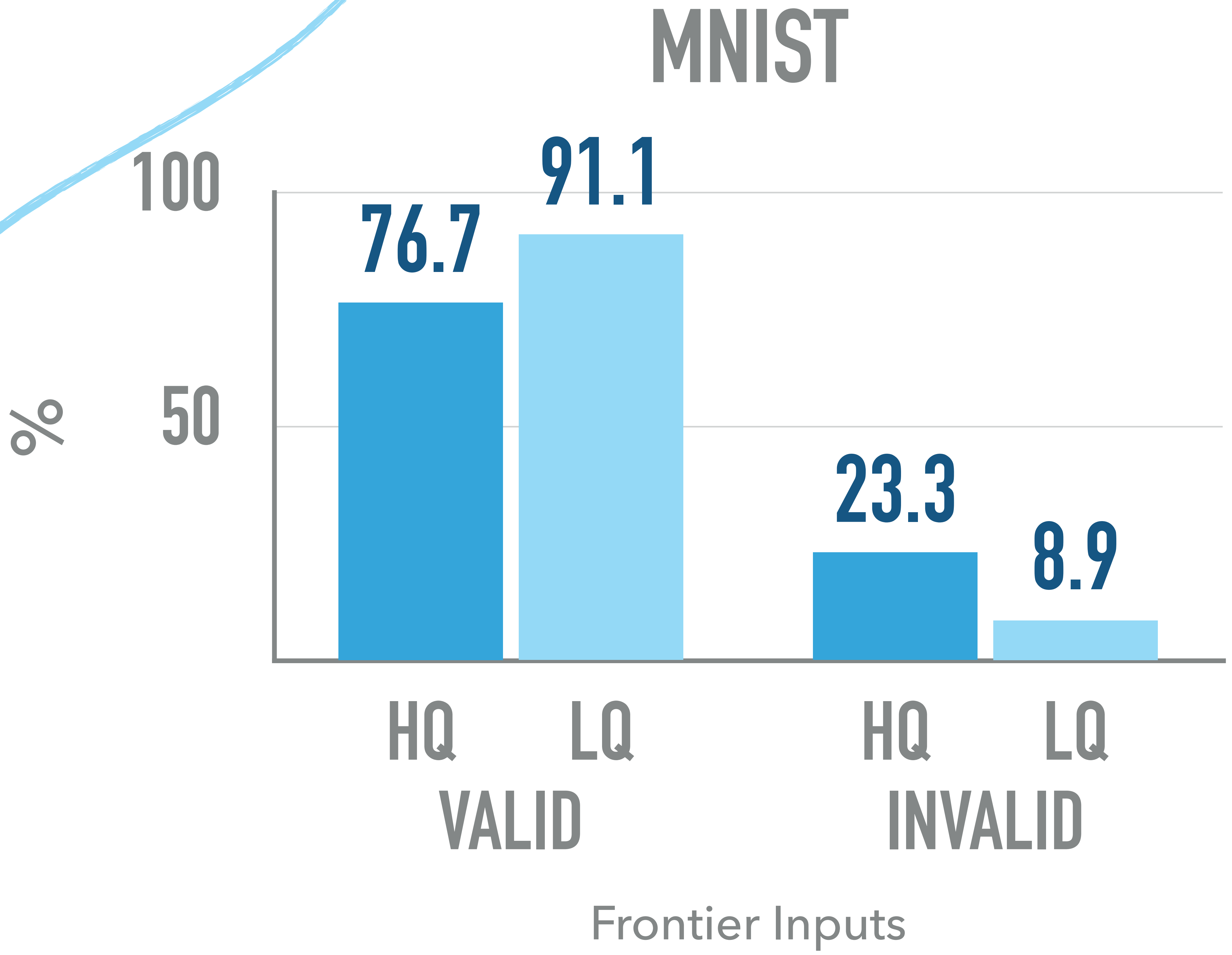
INTERSECTION BETWEEN THE FRONTIER REPORTED BY DEEPJANUS AND THE INPUT VALIDITY DOMAIN

SELECTED 180 FRONTIER INPUTS BY DEEPJANUS ON MNIST

ASKED 20 HUMAN EVALUATORS TO CLASSIFY THEM

EXPECTED ANSWER → VALID INPUT

UNEXPECTED ANSWER → INVALID INPUT



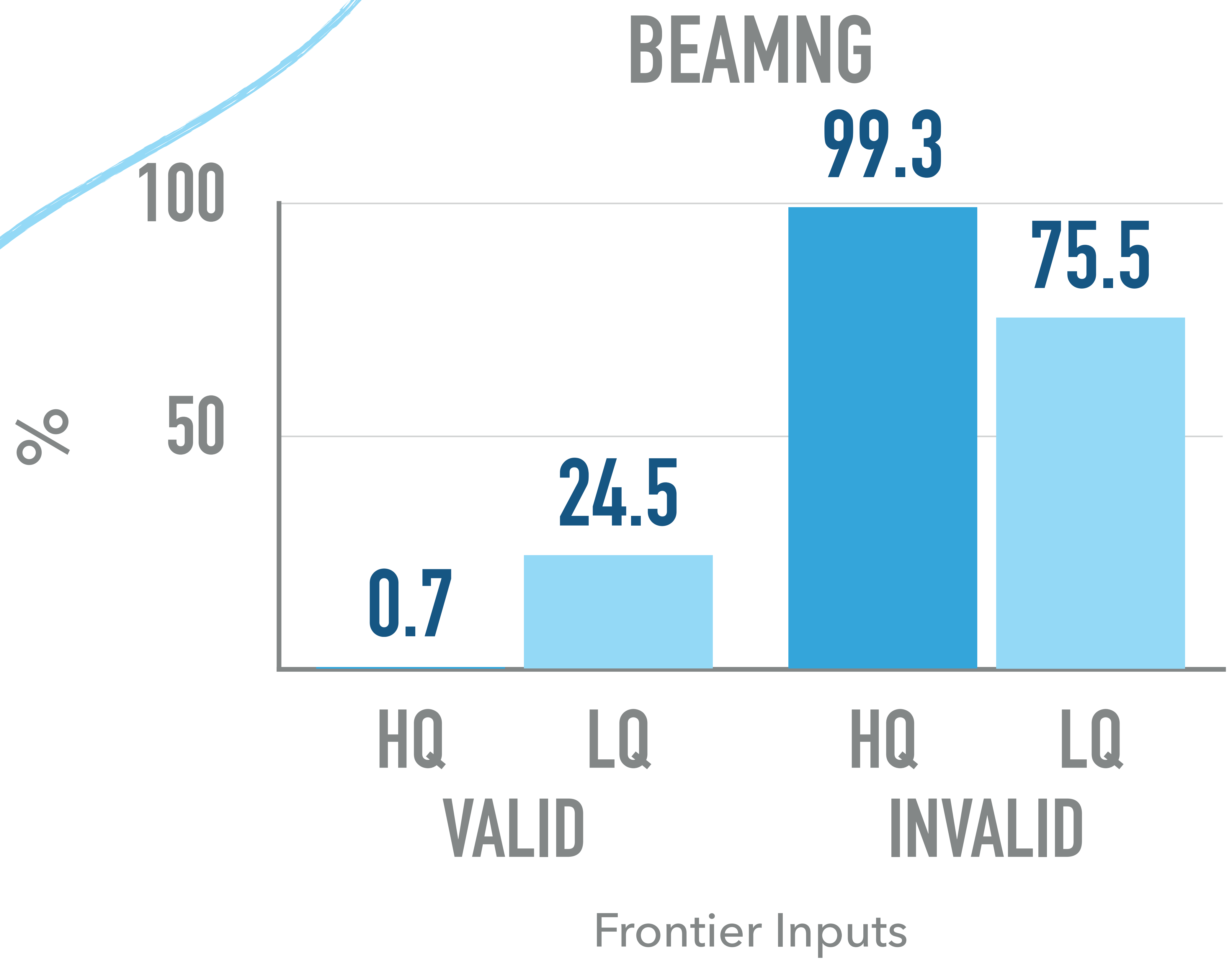
CONSIDERED THE FRONTIER INPUTS BY DEEPJANUS ON BEAMNG

MEASURED VIOLATIONS WRT THE AASHTO GUIDELINES ON GEOMETRIC DESIGN OF HIGHWAYS

DOES THE ROAD COMPLY WITH THE GUIDELINES?

✓ → VALID INPUT

✗ → INVALID INPUT



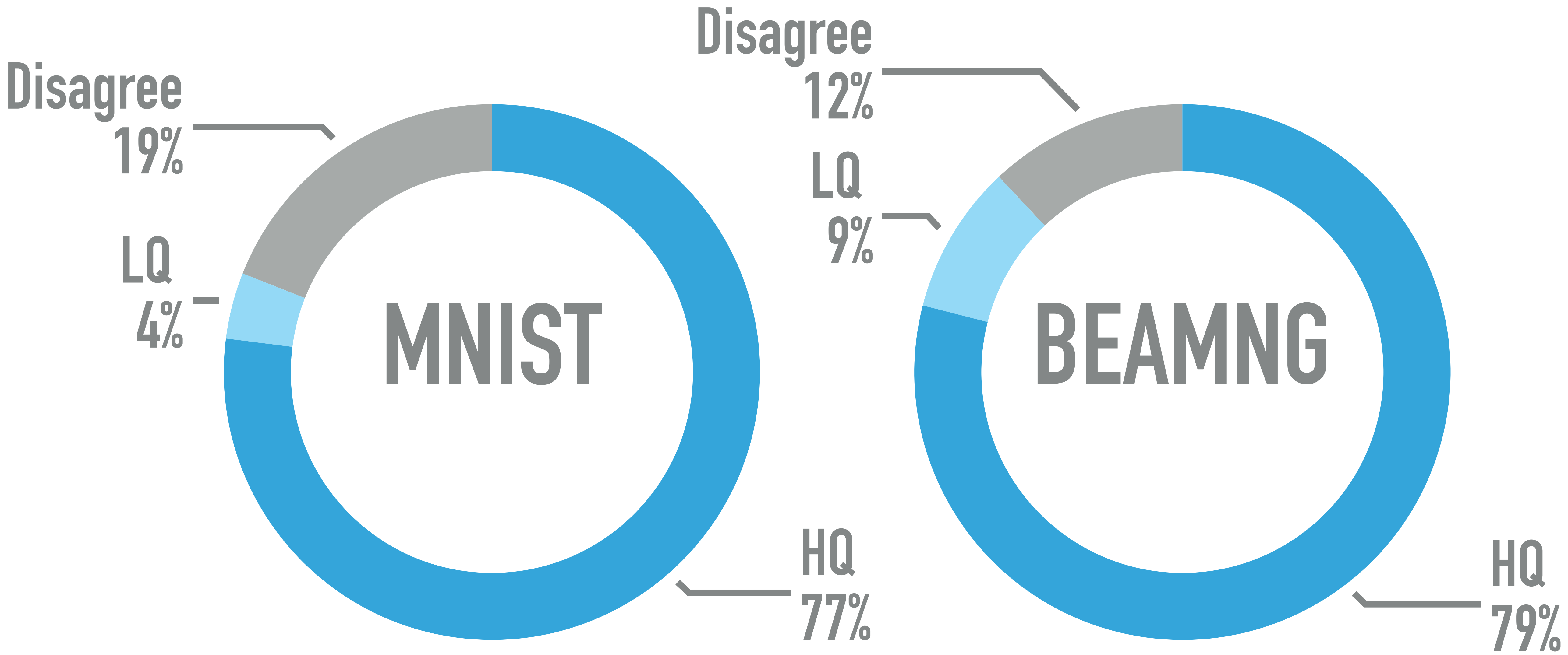
DIFFERENTIATION

DOES DEEPJANUS PROVIDE INFORMATION USEFUL TO **DIFFERENTIATE** THE **QUALITY** OF DL SYSTEMS?

HQ

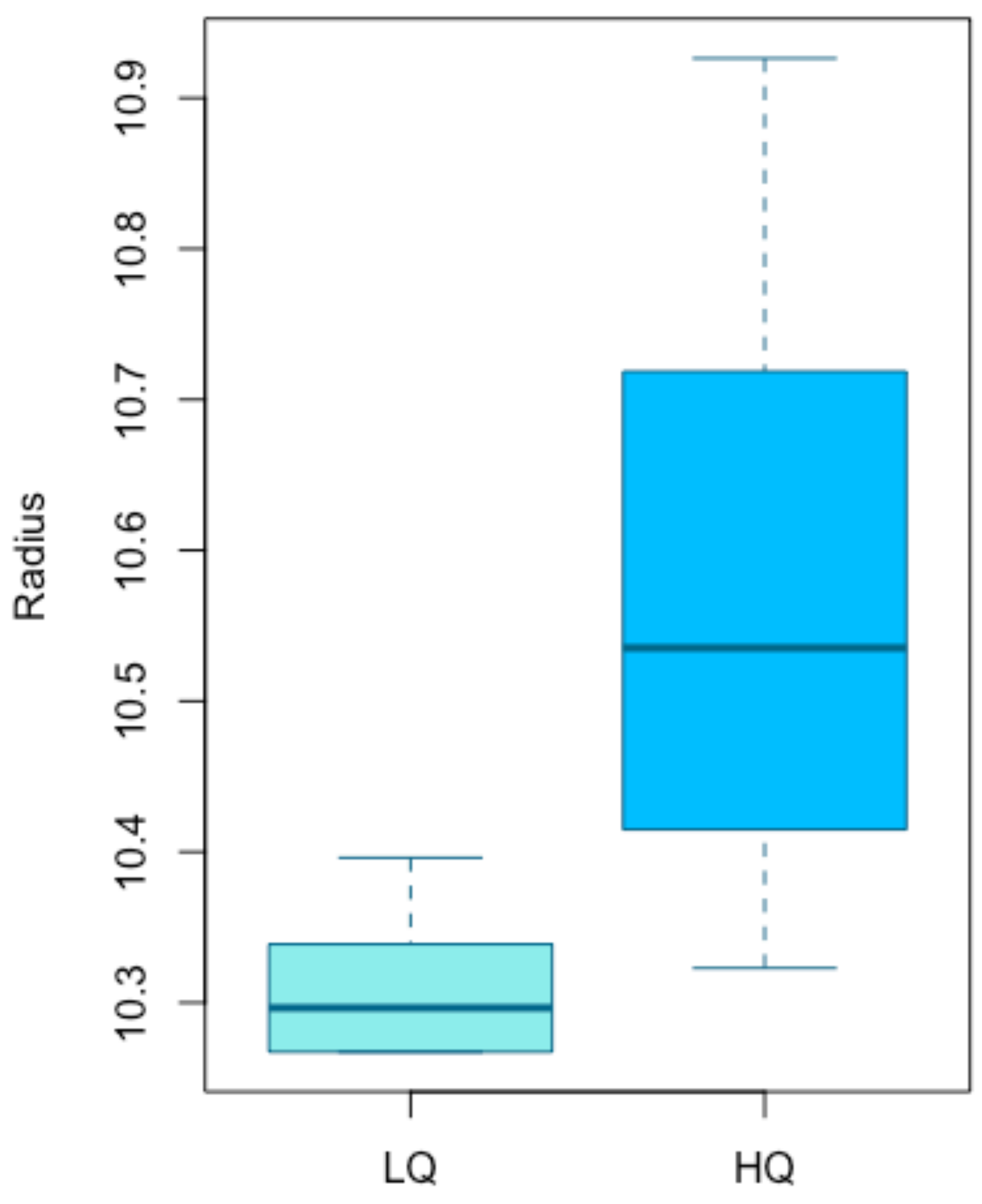
LQ

WHICH FRONTIER INPUTS ARE MORE CHALLENGING TO HUMANS?

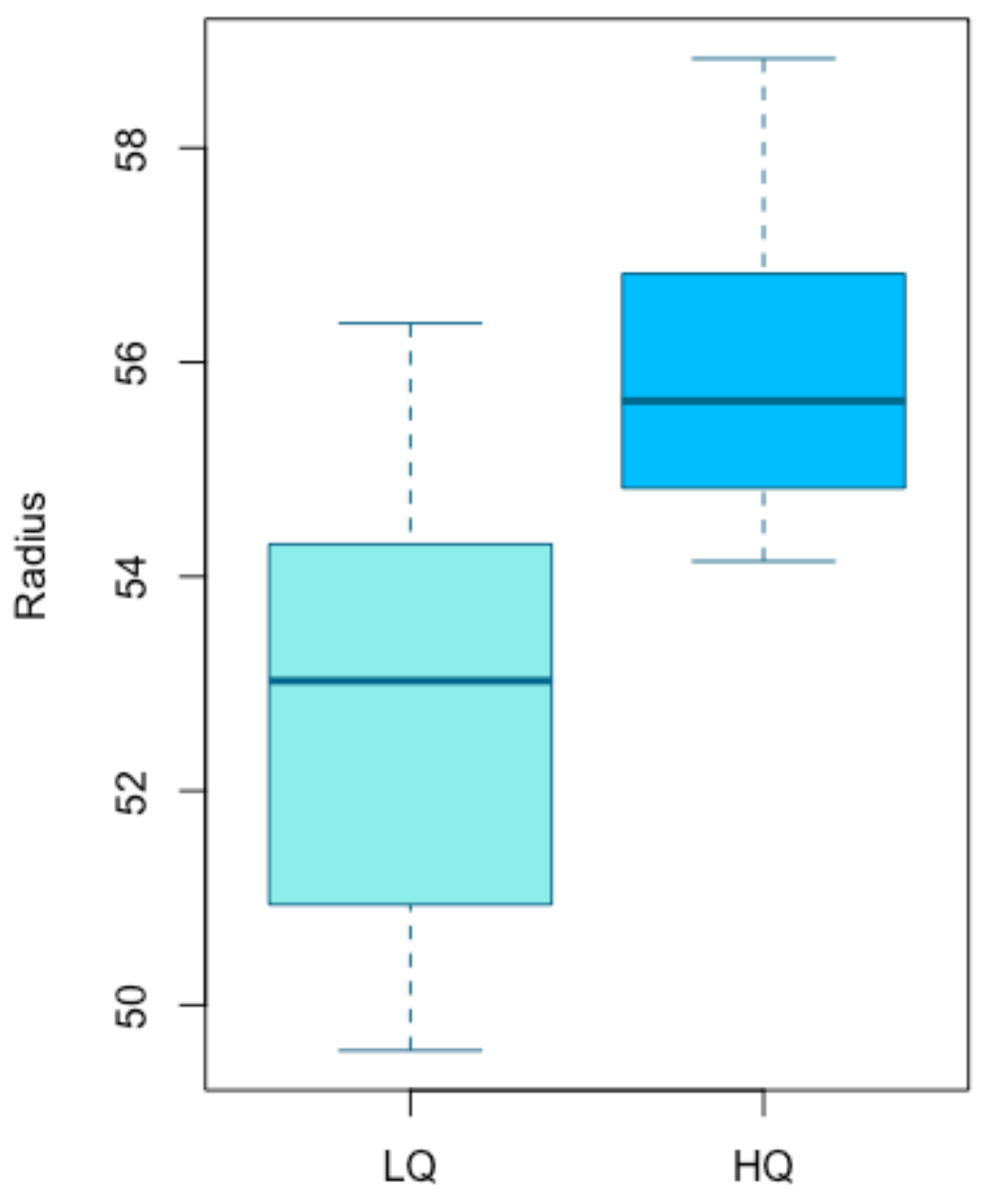


THE INPUTS FROM HQ ARE MORE CHALLENGING TO HUMANS THAN THOSE FROM LQ

MNIST



BEAMNG



RADIUS OF HQ IS SIGNIFICANTLY LARGER THAN THE ONE OF LQ

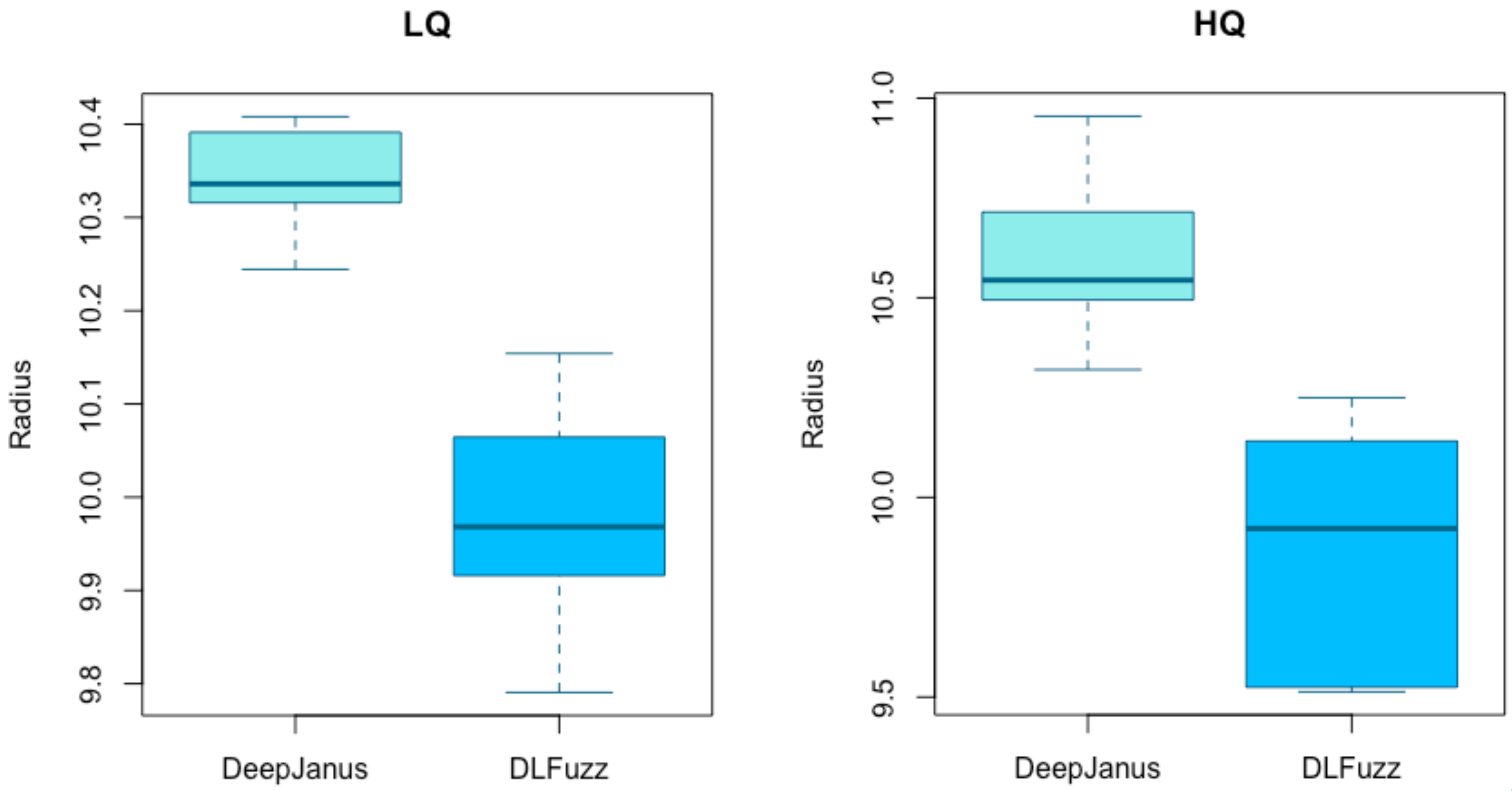
COMPARISON

IS DEEPJANUS

BETTER THAN

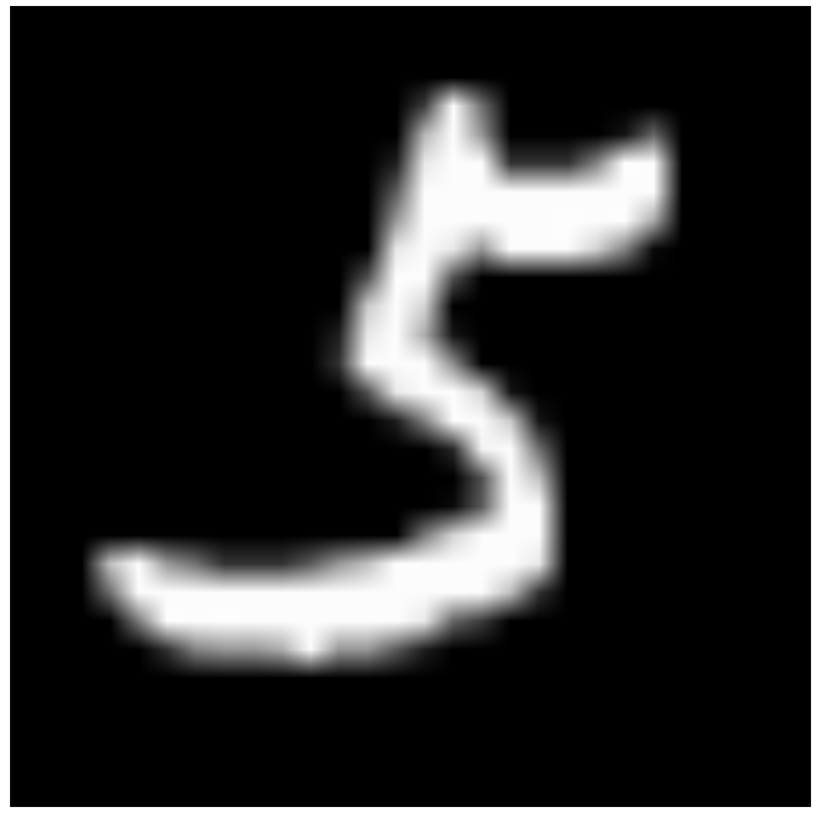
THE STATE OF THE ART

TOOL DLFUZZ?

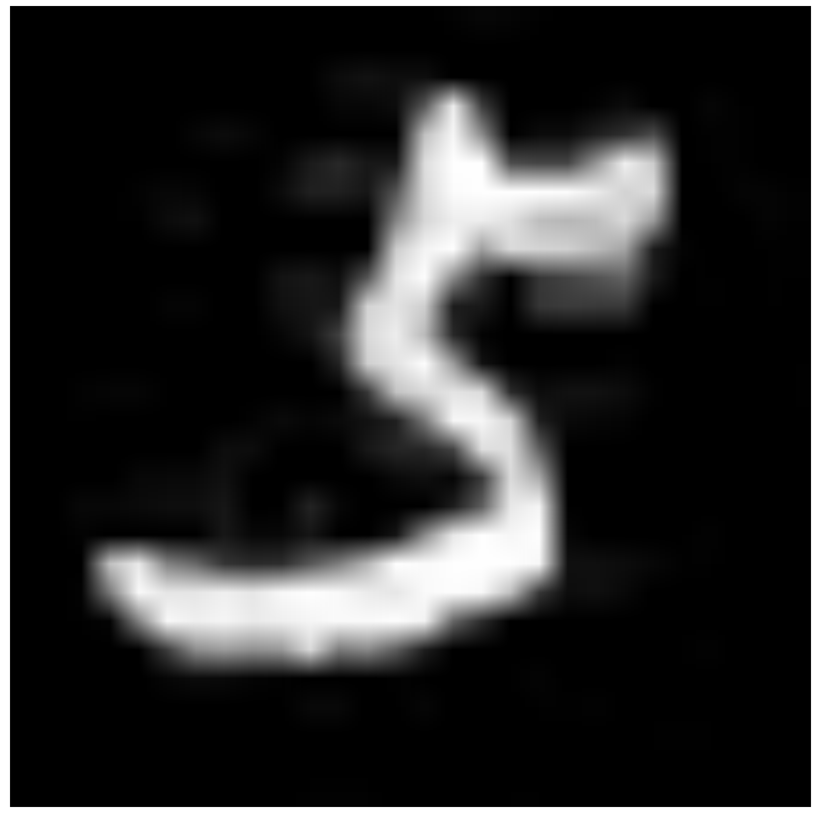


DEEPJANUS EXPLORES A SIGNIFICANTLY LARGER FRONTIER THAN DLFUZZ

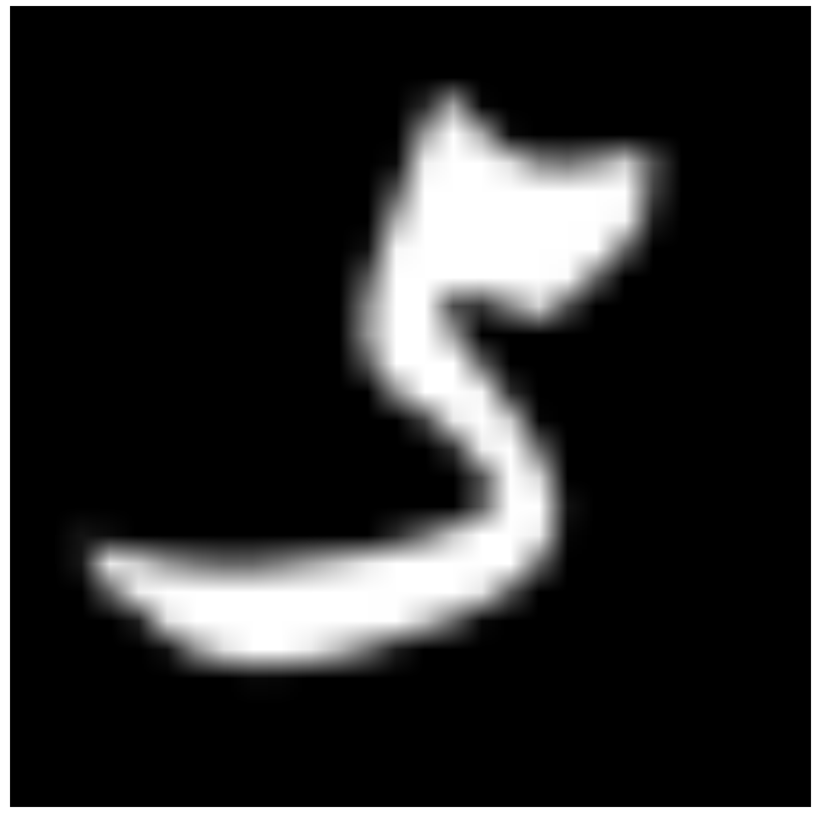
Original Seed



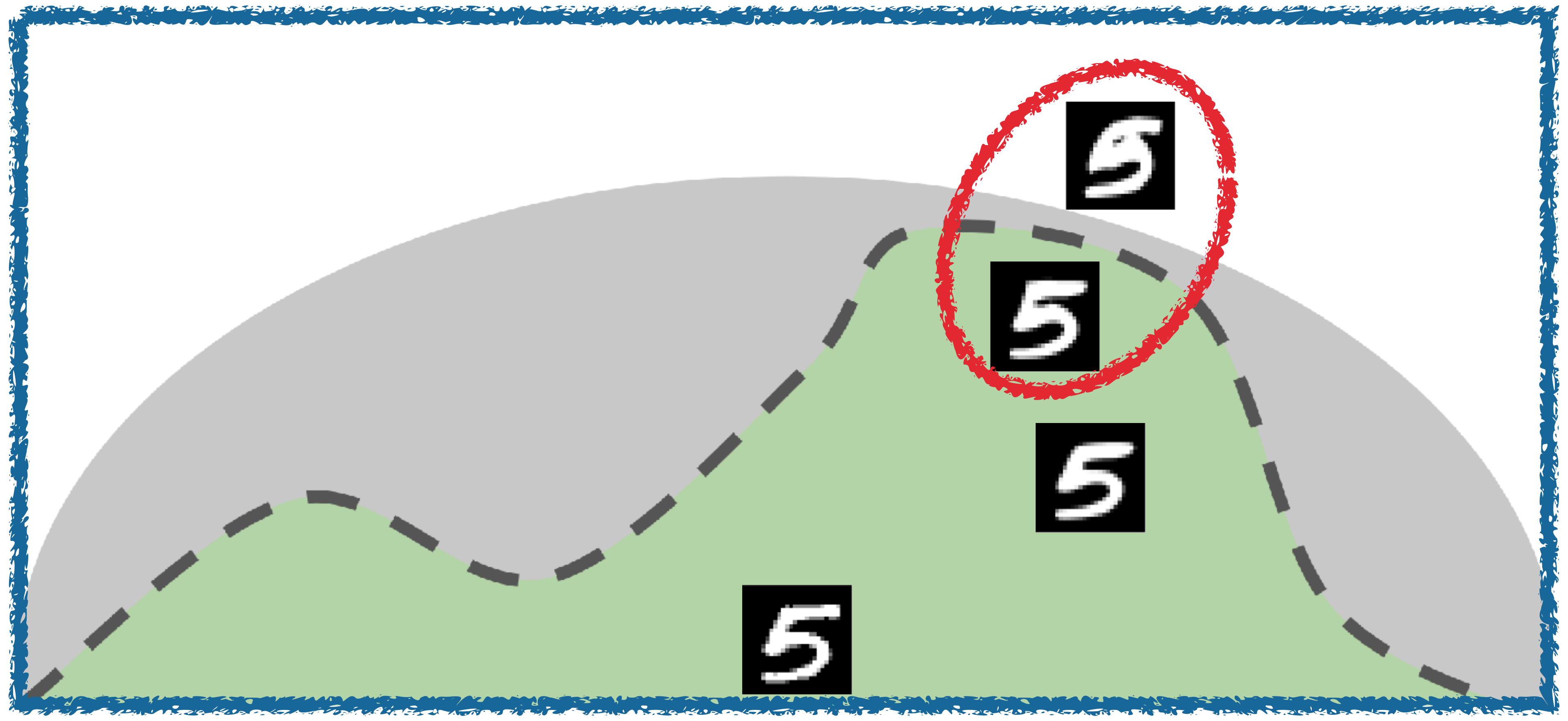
DLFuzz

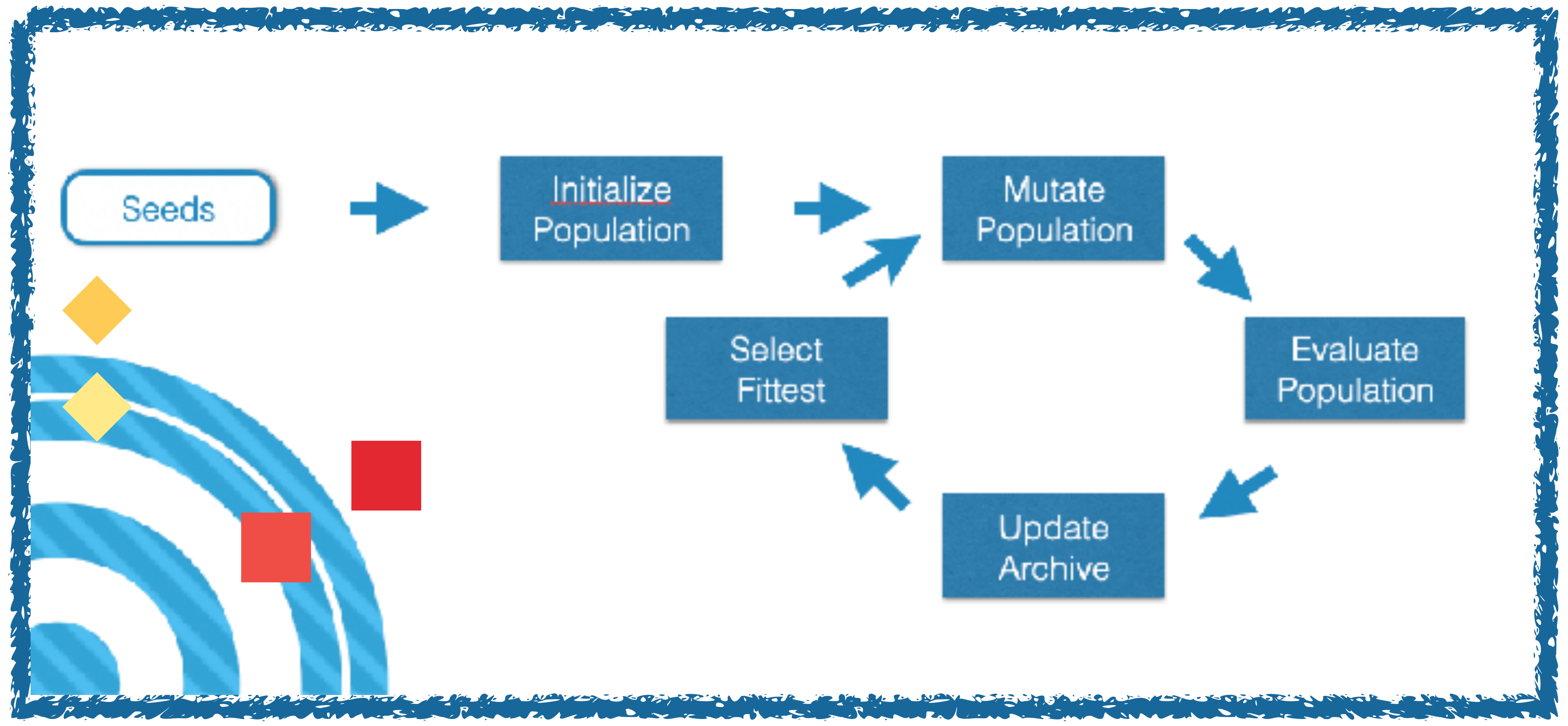


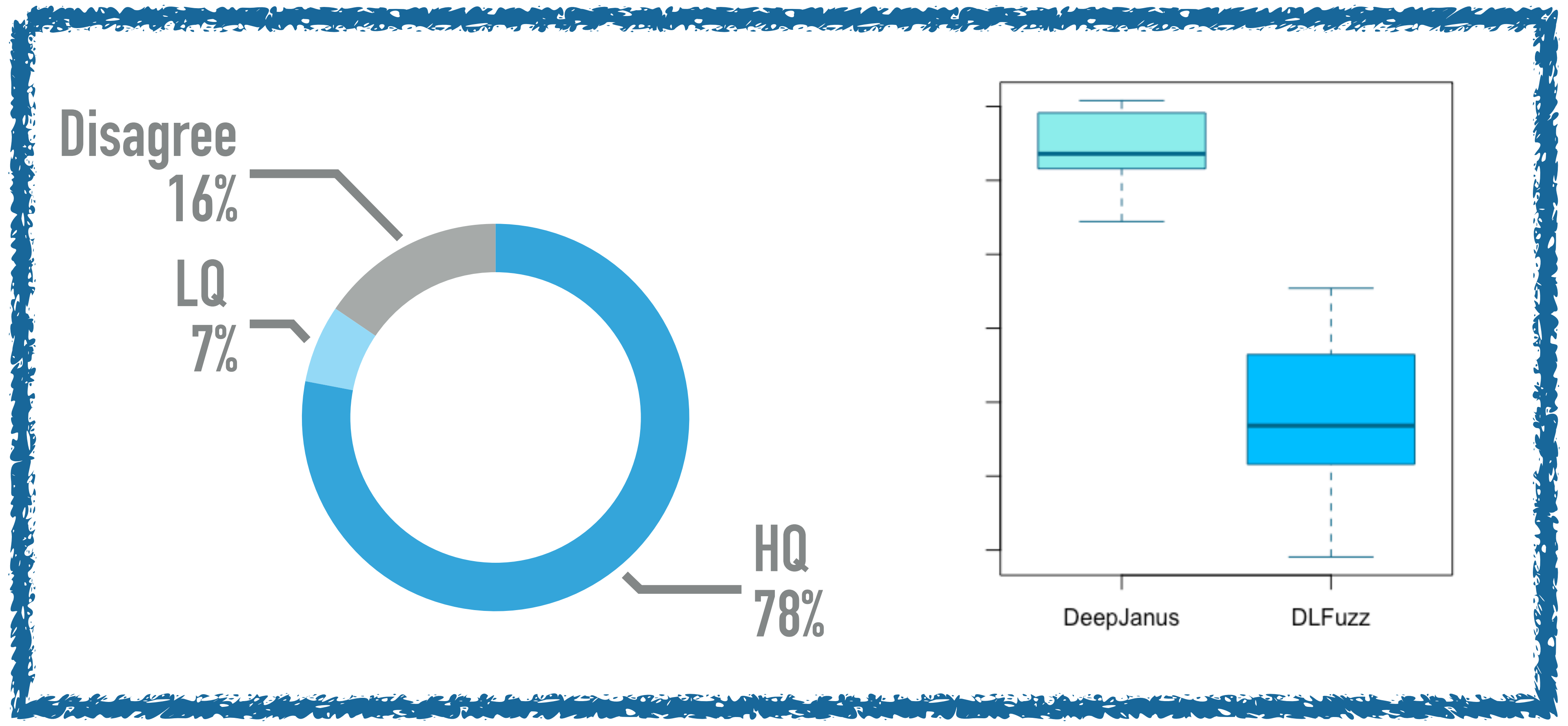
DeepJanus



INPUTS GENERATED BY DEEPJANUS ARE MORE REALISTIC THAN THE ONES OF DLFUZZ




Seeds
Initialize Population
Mutate Population
Evaluate Population
Update Archive
Select Fittest


Disagree 16%
LQ 7%
HQ 78%
DeepJanus
DLFuzz

Artifacts Available
acm
Artifacts Evaluated
acm
Functional