DEEL

DEpendable & Explainable Learning

Domain-Aware DL Model Testing

Houssem Ben Braiek, Ph. D.

IVADO

IRT SAINT EXUPÉRY

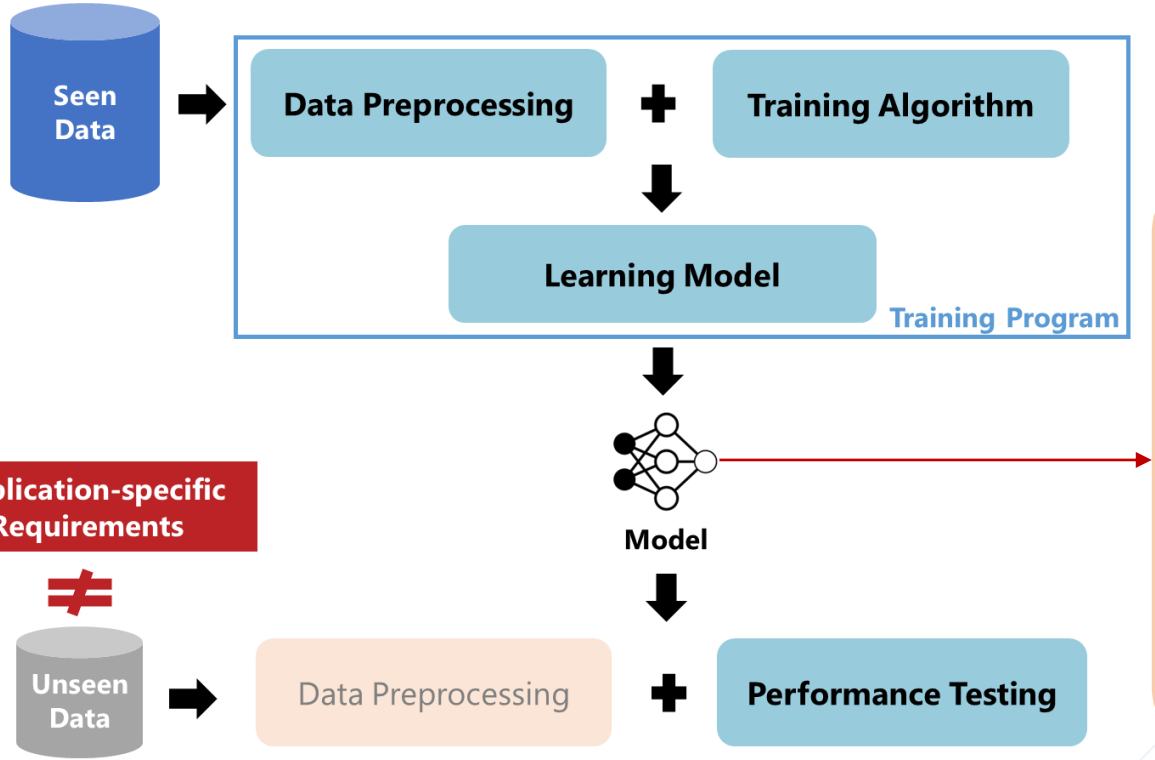
CRIAQ

Consortium de recherche et d'innovation en aérospatiale au Québec

ANITI

Institut intelligence et données

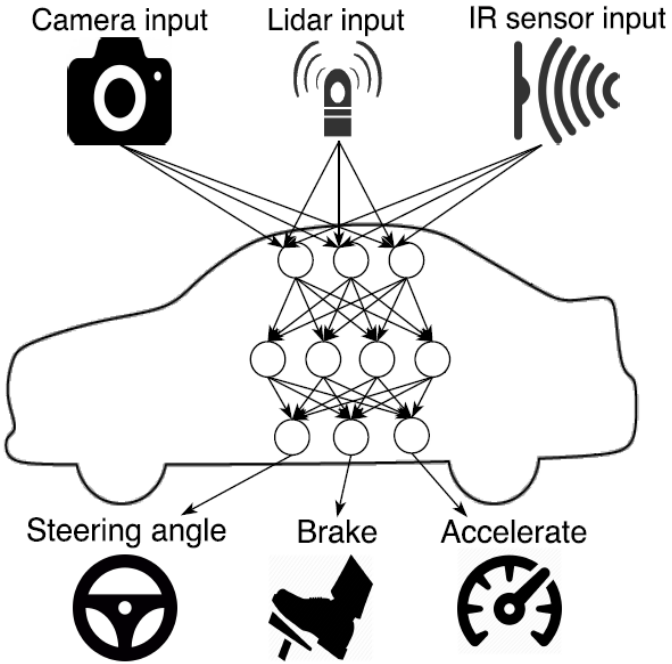
Underspecification Issues of Unseen Datasets

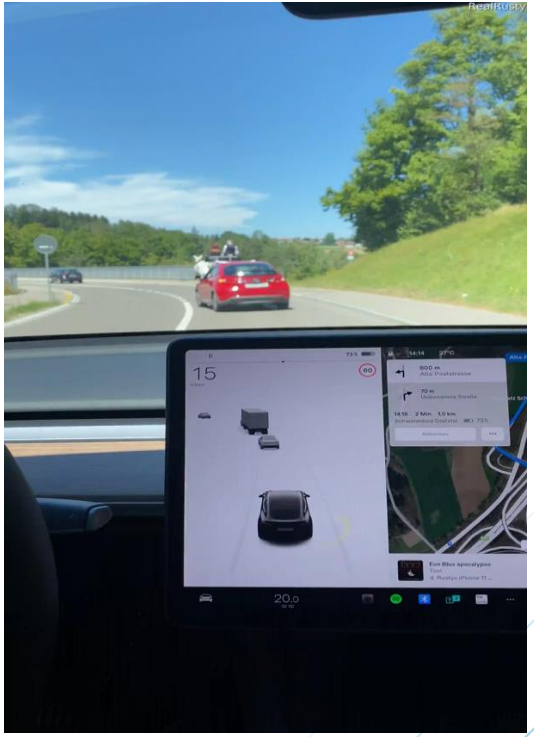


Why do DL practitioners perceive the value of DL testing differently?

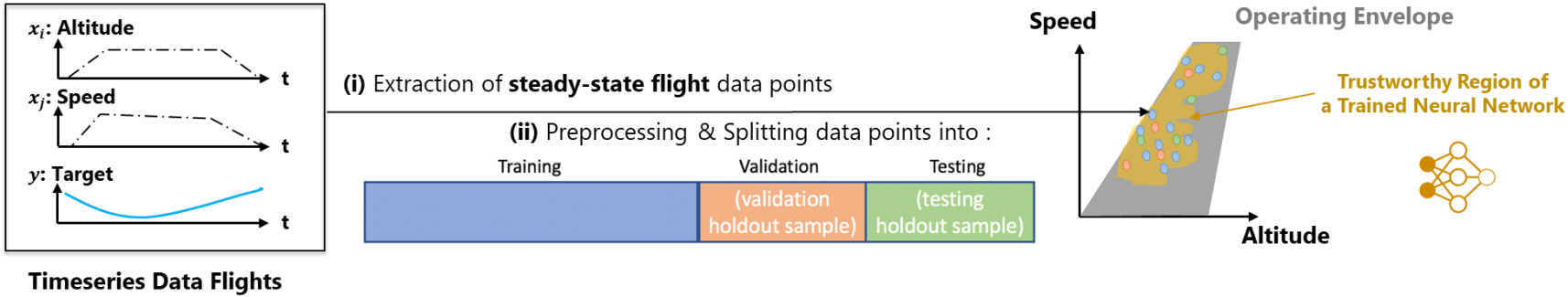
	Low Risk	High Risk
Quantifiable Performance	Outperform the **state-of-the-art** on **testing benchmarks**, e.g., ImageNet, Coco, etc.	Maintain an **acceptable performance** for a **critical function** under **carefully controlled conditions**, e.g., a custom-made cobot that performs repetitive tasks in a manufacturing facility.
Non-Quantifiable Performance	Provide added **value** over **legacy baselines** or **fill a gap**, e.g., filtering ads, recommending movies, etc.	Guarantee an **acceptable performance** for a **critical** function under **all foreseeable operational conditions**, e.g., a generic-purpose cobot that assists the elderly with household duties.

High Risk, Non-Quantifiable Performance …





The Case of Aircraft System Performance Models



➔ A trained NN could illustrate the system performance over the range of included-or-close operational conditions.

'the equipment, systems, and installations must be designed and installed to ensure they perform their intended functions under all foreseeable operating conditions.' U.S Code of Federal Regulations, parts 23, 25, 27, 29

A trustworthy performance model must be qualified to be representative of system behavior throughout the range of foreseeable operational conditions.

The Need for Domain-Aware DL Testing Methods

Tradeoff

Statistical Testing

Estimate the **iid performance** of the model for completely **new inputs**.

$$Err = \sum_{i \in D_{test}} \left(\hat{y}^{(i)} - y^{(i)}\right)^2$$

Use unseen test data D_{test} as a proxy for future entries (x_{new}).

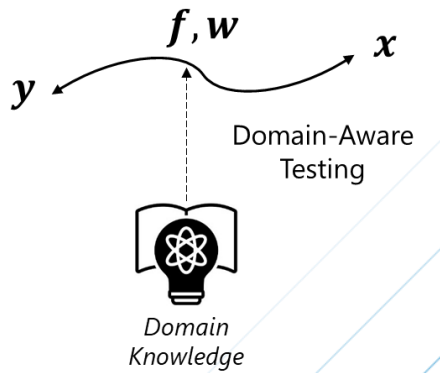
$$D_{test} = \left\{\left(x^{(i)}, y^{(i)}\right)\right\}_{i \in [1,N]}$$

Collection of D_{test} is costly in aircraft industry

Model Testing

Test the **internal logic/mappings** of the model against the prior knowledge on the nature of the relation between x and y.

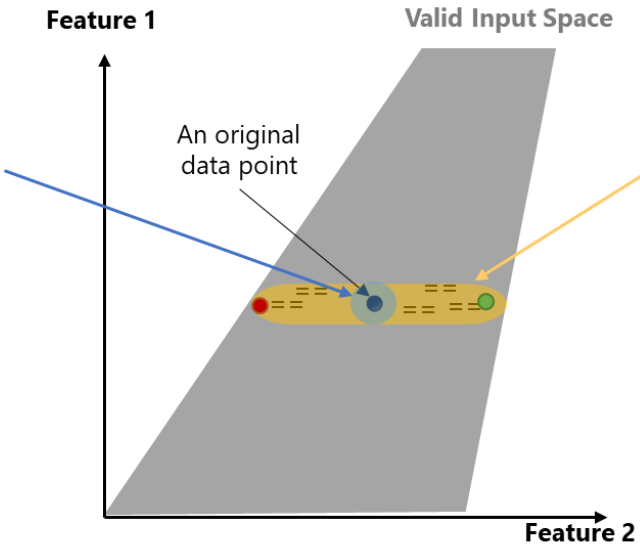
Oracle Problem f^* ? w^*?



Invariance Tests

Perturbation Test Definition

$$\forall\, x' \colon \|x - x'\|_p \leq \delta$$
$$\Rightarrow |f(x) - f(x')| \leq \epsilon$$



Invariance Test Definition

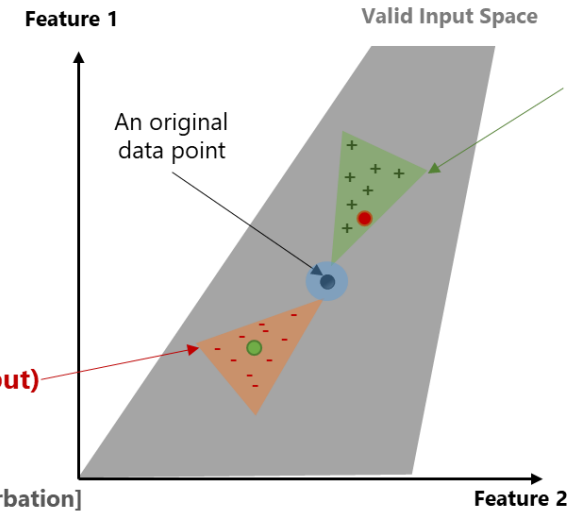
$$x_i \nearrow, x_{i+1} \searrow, \ldots, x_n \leftrightarrow \Rightarrow f \leftrightarrow$$ [Rule Spec]

$$\forall\, x', \forall i \in I_{pr} \colon (x_i - x_i') \leq \delta_i$$ [Signed Perturbation]

$$\Rightarrow |f(x) - f(x')| \leq \epsilon$$ [Test Assertion]

These represent the failed inputs x for which the predictions are not consistent with the derived invariance tests.

Directional Expectation Tests



Directional Expectation Test (Increasing output)

$x_i \nearrow, x_{i+1} \searrow, \ldots, x_n \leftrightarrow \Rightarrow f \nearrow$ [Rule Spec]

$\forall\, x', \forall i \in I_{inc}, (x_i - x_i') \leq \delta_i$ [Signed Perturbation]

$\Rightarrow f(x) \geq f(x')$ [Test Assertion]

Directional Expectation Test (Decreasing output)

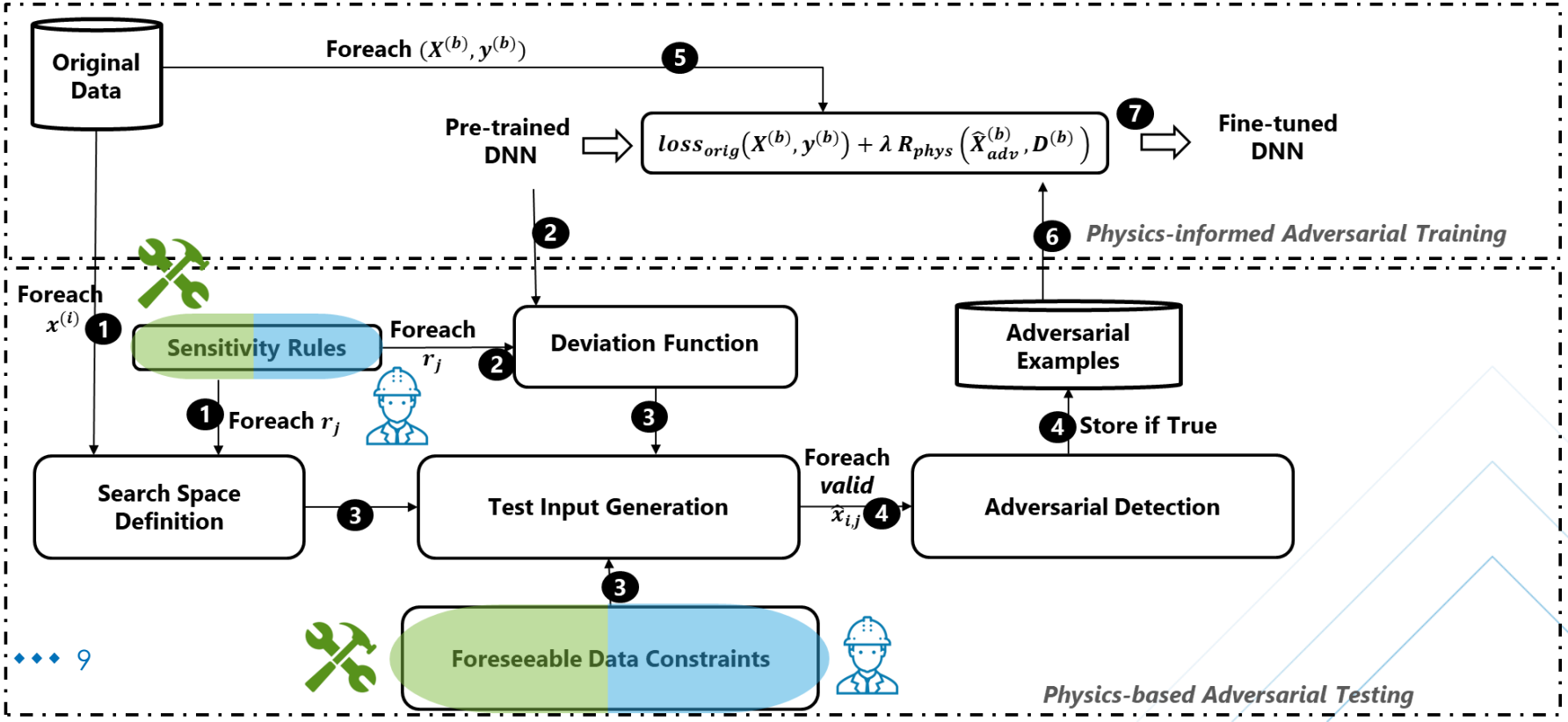
$x_i \nearrow, x_{i+1} \searrow, \ldots, x_n \leftrightarrow \Rightarrow f \searrow$ [Rule Spec]

$\forall\, x', \forall i \in I_{dec}, (x_i - x_i') \leq \delta_i$ [Signed Perturbation]

[Test Assertion]

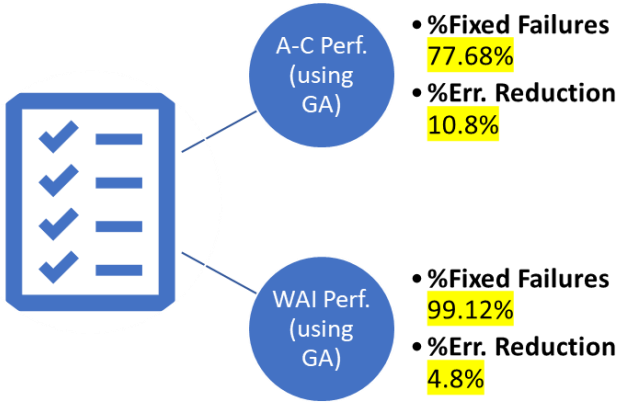
These represent the failed inputs x for which the predictions are not consistent with the derived directional expectation tests.

End-to-End Workflow of the Proposed Method

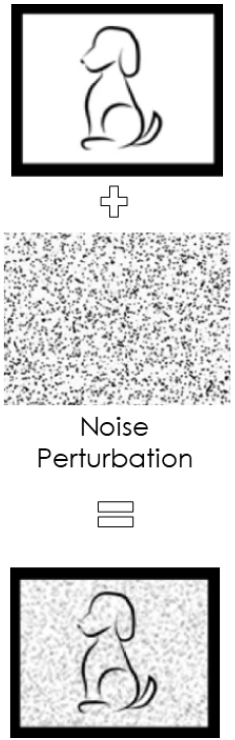


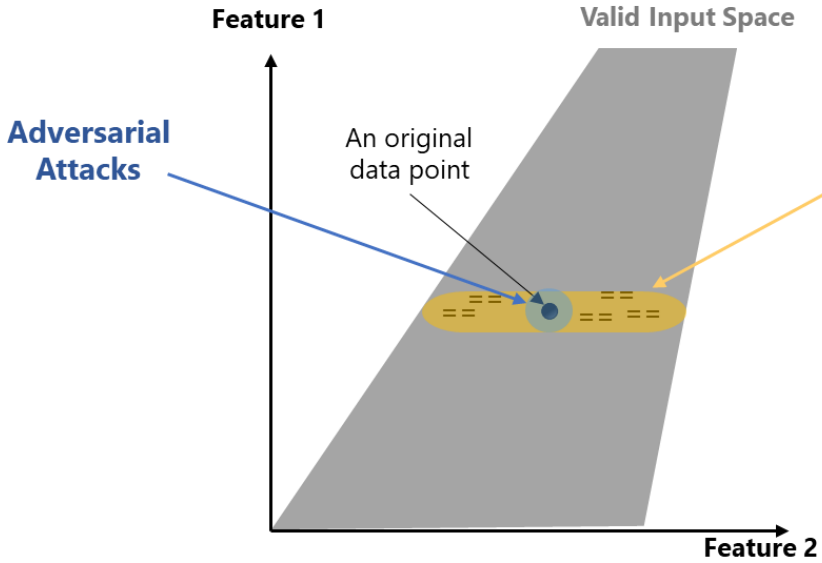
Evaluation Models & Results

Model	Predicted Target	Description
Aircraft(A-C) Performance Model	α: angle of attack	The model maps steady-state angle of attack (α) to features related to flight conditions and wing configurations.
Wing Anti-Icing (WAI) Performance Model	T^b_{skin}: A-wing leading-edge skin temperature T^b_{skin}: B-wing leading-edge skin temperature	The model maps the states of skin temperature sensors to features related to flight conditions, wing configurations, and high-pressure pneumatic system conditions at the wing root.



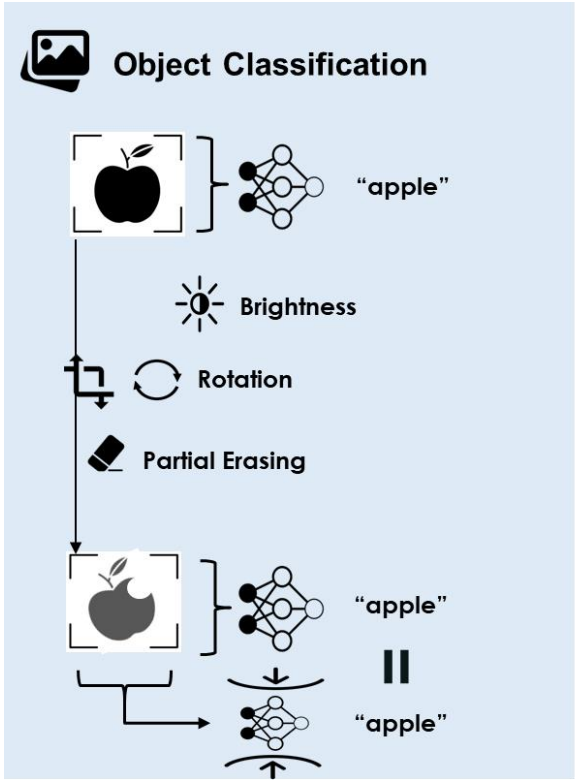
Analogies with other DL Applications

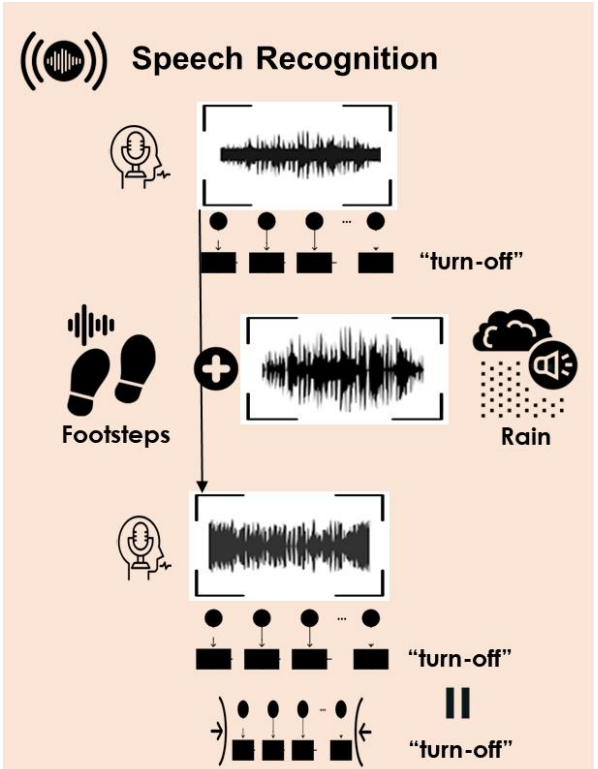


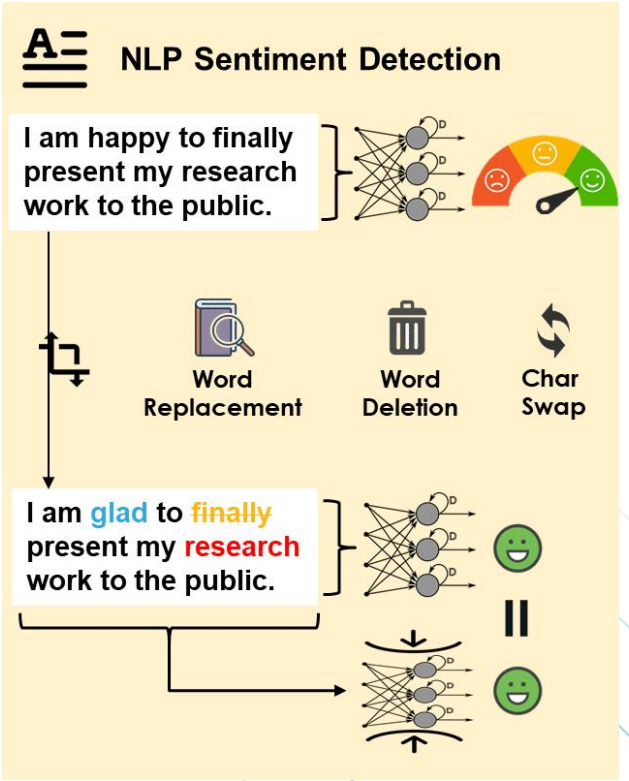


Semantically-preserving Data Transformations

Semantically-preserving Data Transformations







Semantically-preserving Data Transformations

For Images:

Pixel-value Transformations:

- Brightness
- Contrast
- Blurring
- Partial Erasing
- Pixel Perturbation

Affine Transformations:

- Translation
- Shearing
- Scaling
- Rotation

For Audio Speeches:

Signal-wise Conversions:

- Speed
- Pitch
- Loudness

Additive Noise Signals:

- Random noisy perturbations
- **Colored noises:** white, pink, brown.
- **Indoor Noises:** breathing, footsteps, laughing, clock-tick, etc.
- **Outdoor Noises:** Engine, Fireworks, Rain, Train, etc.

For Natural Language Texts :

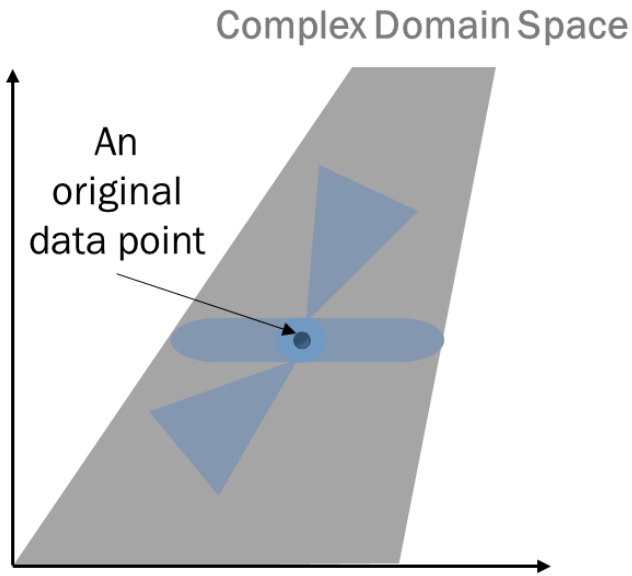
Char-level Transformations:

- Random Insertion
- Random Swap
- Random Deletion

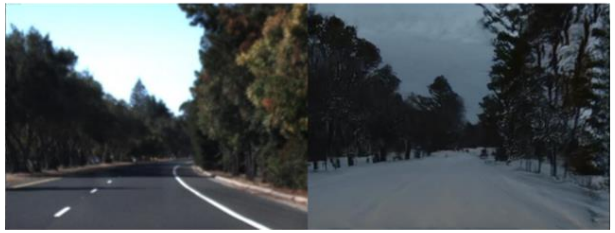
Word-level Transformations:

- Synonym/Embedding Replacement
- Random Insertion
- Random Swap
- Random Deletion

How can we generate valid inputs from complex domains?

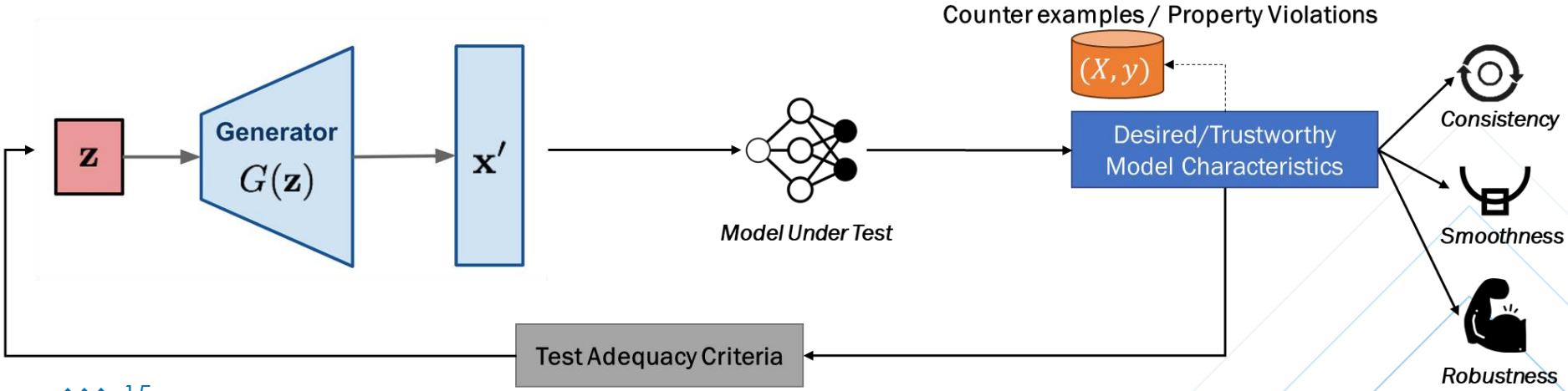


As software tests are written in code, DL tests can be produced by DL models !



DeepRoad [1] use GANs:

- map image from source domain to latent domain.
- generate image in the new domain from latent domain.



[1] Zhang et al. DeepRoad: GAN-based Metamorphic Autonomous Driving System Testing. arXiv:1802.02295]

DEEL

DEpendable & Explainable Learning