Building Robust Predictive Systems for Tabular Data

Zhipeng (Zippo) He
School of Information Systems

Supervisory team:
A/Prof. Chun Ouyang (QUT)
Prof. Alistair Barros (QUT)
A/Prof. Catarina Moreira (UTS)

Can we trust AI models that are easily deceived? What’s the cost of this fragility?

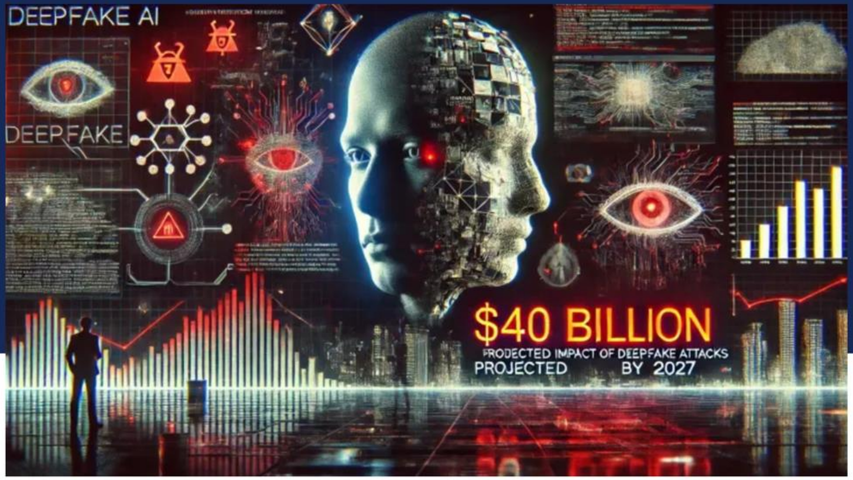
Analysis

Deepfakes will cost $40 billion by 2027 as adversarial AI gains momentum

Louis Columbus
@LouisColumbus

July 1, 2024 3:39 PM

f X in



nature

Explore content ⌄ About the journal ⌄ Publish wi

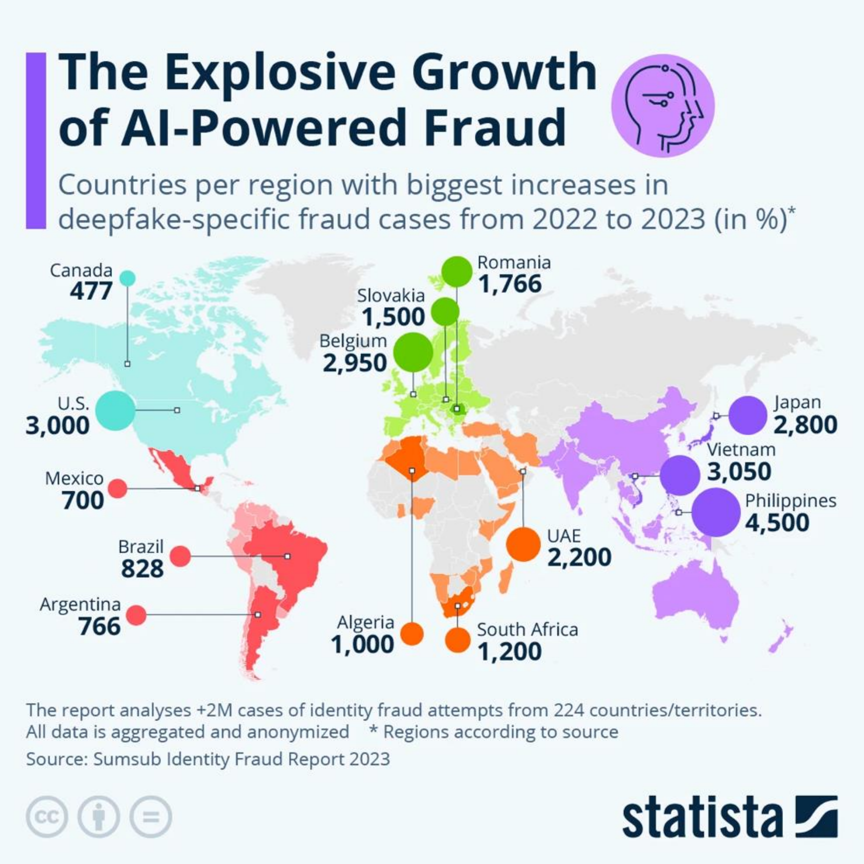
nature › outlook › article

OUTLOOK | 25 July 2024

AI is vulnerable to attack. Can it ever be used safely?

The models that underpin artificial-intelligence systems such as ChatGPT can be subject to attacks that elicit harmful behaviour. Making them safe will not be easy.

By Simon Makin



Research Problem

How can one construct predictive models that are robust to adversarial attacks for tabular data?

Zhipeng (Zippo) He
School of Information Systems

Test the ML Models Like Software

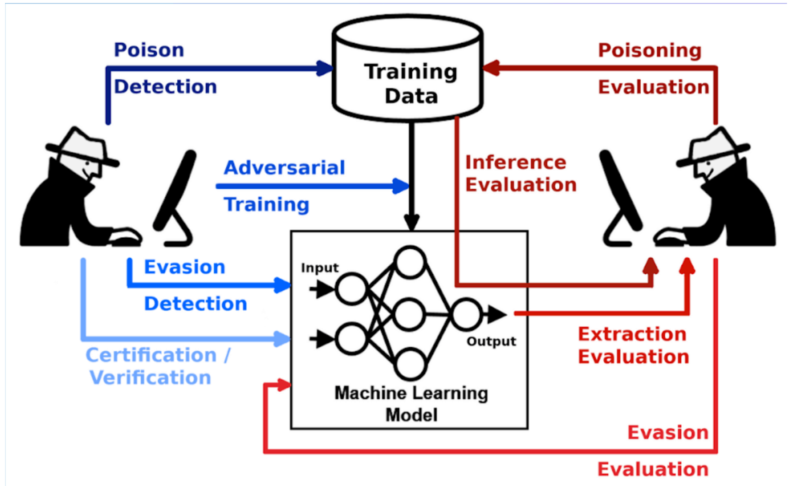
Software Testing

Purpose: Identify bugs and vulnerabilities.

Method: Test edge cases and unexpected inputs.

Goal: Ensure software is robust and reliable.

Adversarial Attacks in Machine Learning



Purpose: Identify weaknesses in ML models.

Method: Craft inputs to exploit vulnerabilities.

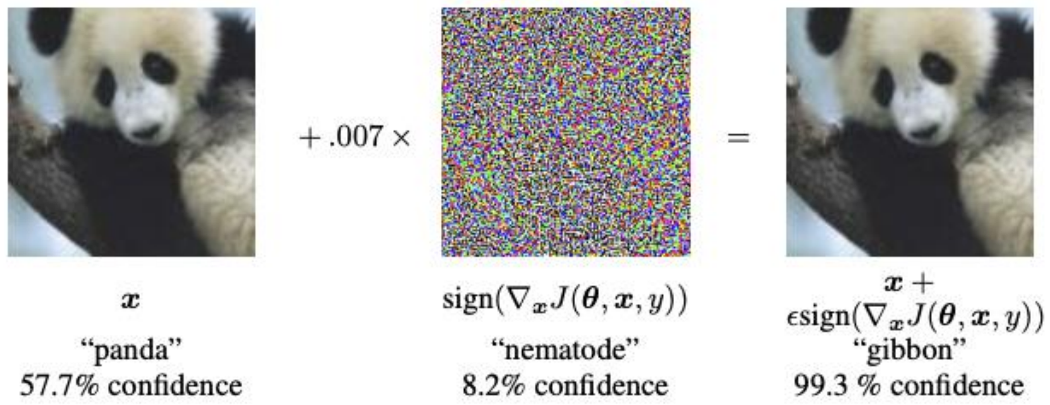
Goal: Improve model robustness.

Zhipeng (Zippo) He
School of Information Systems

What are Adversarial Attacks?

An adversarial attack is a method to generate adversarial examples.

*“Adversarial examples are **specialised inputs created with the purpose of confusing a neural network, resulting in the misclassification of a given input**. These notorious inputs are **indistinguishable** to the human eye but cause the network to fail to identify the contents of the image.” [1]*

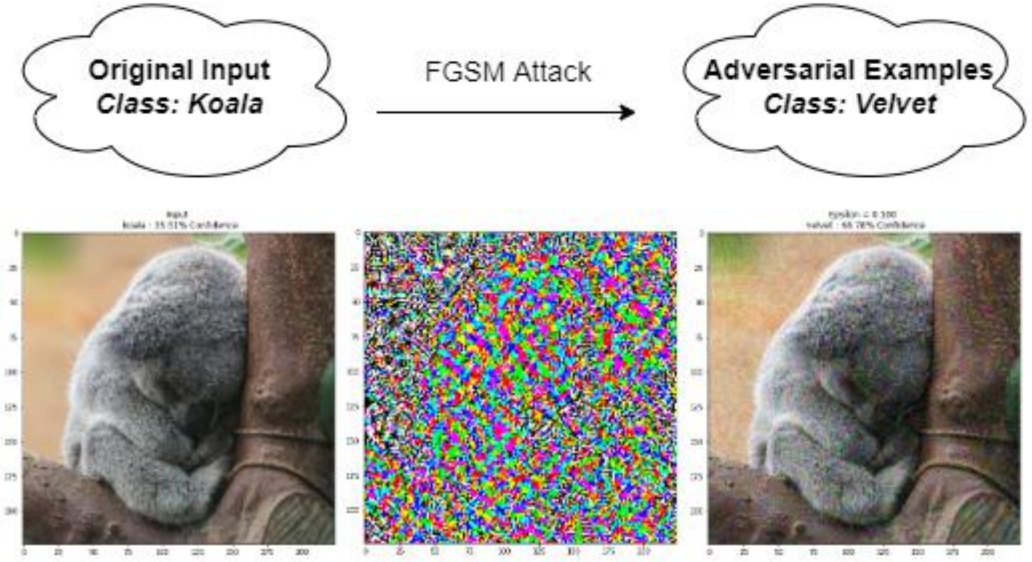


Zhipeng (Zippo) He
School of Information Systems

[1] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Different Concepts of Imperceptibility

The perturbation on tabular data is more noticeable than images.



Original Input
No Diabetes

FGSM Attack

Adversarial Example
Has Diabetes

Pregnancies	1
Glucose	71
Blood pressure	48
Skin thickness	18
Insulin	76
BMI	20.4
Diabetes Pedigree function	0.323
Age	22

Pregnancies	6
Glucose	130
Blood pressure	11
Skin thickness	0
Insulin	329
BMI	40.5
Diabetes Pedigree function	1.026
Age	40

Zhipeng (Zippo) He
School of Information Systems

How existing works evaluate attacks?

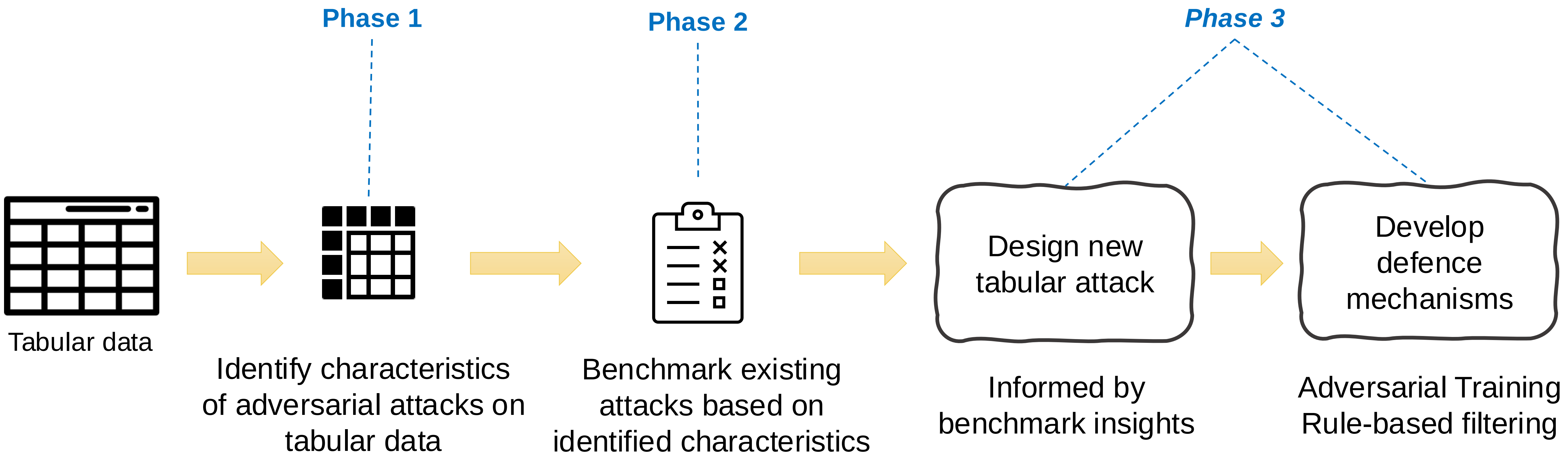
Most attacks are designed for images.

Benchmark/Paper	Data Type		Evaluation Metric	Main Focus
Benchmarking Transferable Adversarial Attacks [2]	**Image**		Attack Transferability Score	Evaluates transferability of adversarial attacks across different architectures
Benchmarking Adversarial Robustness on Image Classification [3]	**Image**	White-box, Black-box Attacks	**Robust Accuracy**, L_∞ Norm	Benchmark for adversarial robustness ag
BlackboxBench: A Comprehensive Benchmark [4]	**Image**	Black-box Adversarial Attacks	**Attack Success Rate**, Query Count	Ev bo
RobustBench: Adversarial Robustness Benchmark [5]	**Image**	L_∞, L_2 Norm-based Attacks	**Robust Accuracy**	St robustness and common corruption robustness
REAP: Realistic Adversarial Patch Benchmark [6]	**Image**	Patch-based Adversarial Attacks	**Patch Success Rate**, Realism Score	Evaluates realistic adversarial patches in real-world conditions
AttackBench: Gradient-based Attack Evaluation [7]	**Image**	Gradient-based Attacks	**Adversarial Success Rate**	Focuses on gradient-based attacks for generating adversarial examples
Graph Robustness Benchmark [8]	Graph Data	Adversarial Attacks on Graphs	**Robust Accuracy**	Benchmarks adversarial robustness of graph machine learning models
Adversarial VQA Benchmark [9]	VQA	Adversarial Attacks on VQA	**Robust Accuracy**	Evaluates robustness of visual question answering models to adversarial inputs
Benchmarking Adversarial Attacks and Defenses for Time-Series Data [10]	Time-series	Adversarial Attacks on Time-Series	**Attack Success Rate**	Evaluates adversarial attacks and defenses specifically for time-series data
From Hero to Zeroe: A Benchmark of Low-Level Adversarial Attacks [11]	Low-Level Text	Low-Level Adversarial Attacks on NLP	**Attack Success Rate**, Perturbation Size, Visual and Phonetic Similarity	Benchmarks adversarial attacks targeting low-level data manipulations (character-level)

Most benchmarks assess the effectiveness of attacks only.

Refer to reference list in the end of slides

Research Roadmap



Zhipeng (Zippo) He
School of Information Systems

Characteristics of Adversarial Attacks on Tabular Data

Effectiveness
- Model Accuracy
- Attack Success Rate

Imperceptibility
- No comprehensive definition for tabular data

Transferability
- Models
- Datasets

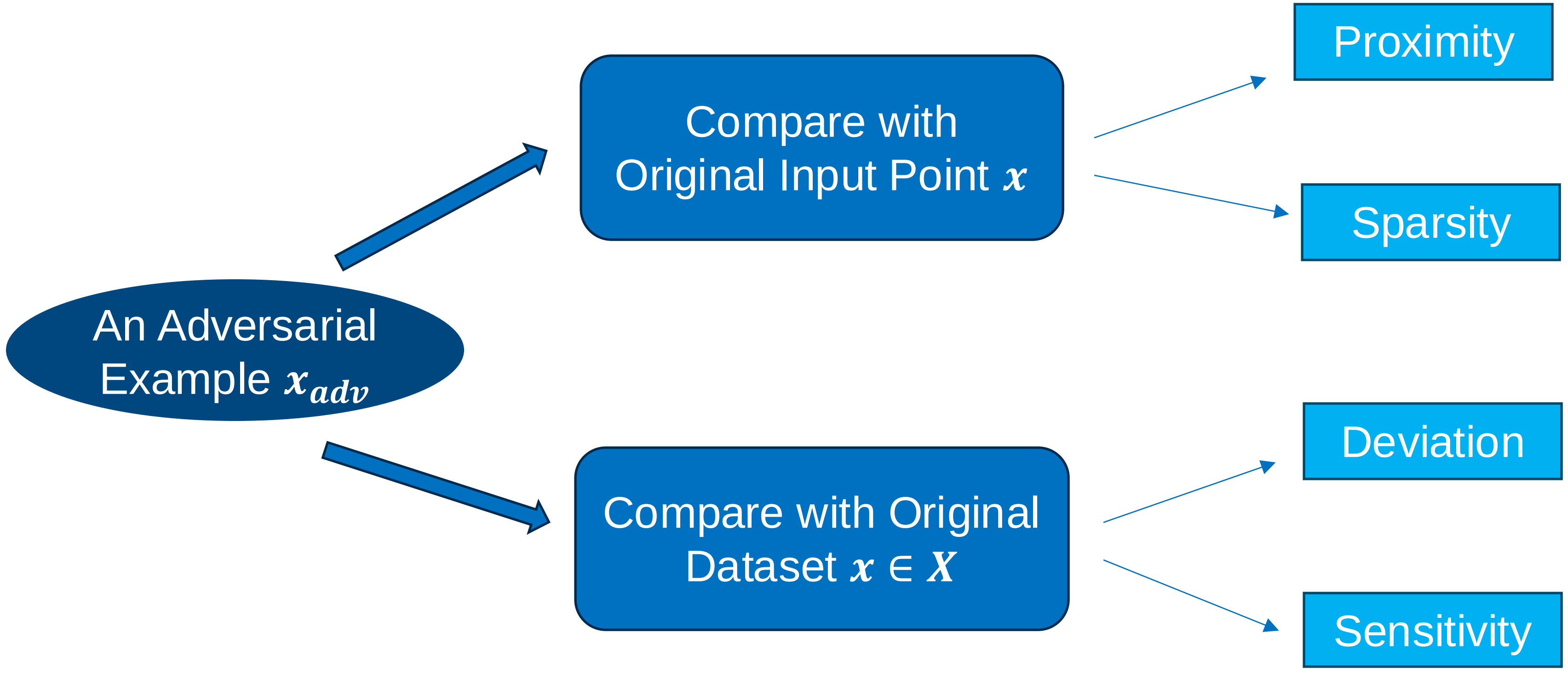
(Not in research scope)

Research Question 1

What properties can be used to define the imperceptibility of adversarial attacks on tabular data?

Zhipeng (Zippo) He
School of Information Systems

Quantitative Imperceptibility Properties [12]



$$\ell_2(\boldsymbol{x}^{adv}, \boldsymbol{x}) = \sqrt{\sum_{i=1}^{n} (x_i^{adv} - x_i)^2}$$

$$\ell_\infty(\boldsymbol{x}^{adv}, \boldsymbol{x}) = \|\boldsymbol{x}^{adv} - \boldsymbol{x}\|_\infty = \sup_n |x_n^{adv} - x_n|$$

$$Spa(\boldsymbol{x}^{adv}, \boldsymbol{x}) = \ell_0(\boldsymbol{x}^{adv}, \boldsymbol{x}) = \sum_{i=1}^{n} \mathbb{1}(x_i^{adv} - x_i)$$

Mahalanobis distance (MD)

$$\mathrm{MD}(\boldsymbol{x}^{adv}, \boldsymbol{x}) = \sqrt{(\boldsymbol{x}^{adv} - \boldsymbol{x}) V^{-1} (\boldsymbol{x}^{adv} - \boldsymbol{x})^T}$$

V is the covariance matrix of Dataset X

$$\mathrm{SDV}(x_i) = \sqrt{\frac{\sum_j^m (x_{i,j} - \bar{x}_i)^2}{m}}$$ Standard Deviation

$$\mathrm{SEN}(\boldsymbol{x}, \boldsymbol{x}^{adv}) = \sum_{i=1}^{n} \frac{\|x_i^{adv} - x_i\|_2}{\mathrm{SDV}(x_i)}$$

Zhipeng (Zippo) He
School of Information Systems

[12] He, Z., Ouyang, C., Alzubaidi, L., Barros, A., & Moreira, C. (2024). Investigating Imperceptibility of Adversarial Attacks on Tabular Data: An Empirical Analysis. *arXiv preprint arXiv:2407.11463*. [Accepted, In Press]

Qualitative Imperceptibility Properties [12]

Feature interdependencies

Age	Education	Marital Status	Occupation	Relationship	Race	…	Income
43	Masters	Married-civ-spouse	Exec-managerial	Husband	White	…	>50K
28	HS-grad	Never-married	Craft-repair	Other-relative	Asian	…	<=50K

Feasibility: Feasible feature range

Require Domain Knowledge of Tabular Data

Immutability

Zhipeng (Zippo) He
School of Information Systems

[12] He, Z., Ouyang, C., Alzubaidi, L., Barros, A., & Moreira, C. (2024). Investigating Imperceptibility of Adversarial Attacks on Tabular Data: An Empirical Analysis. *arXiv preprint arXiv:2407.11463*. [Accepted, In Press]

Research Question 2

Which attacks can generate adversarial examples that are both effective and imperceptible?

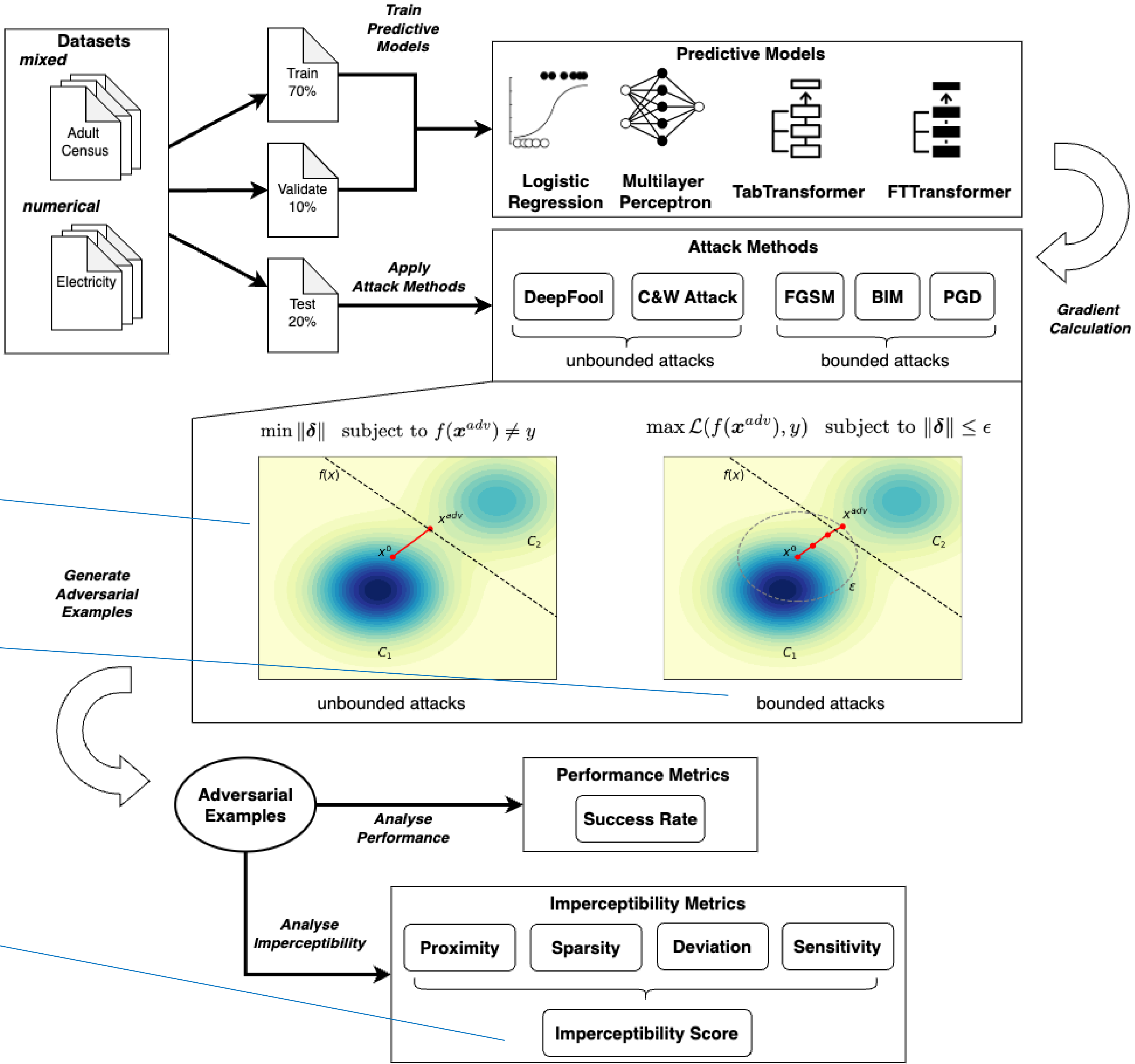
Phase 2

Benchmark Design

Find the smallest possible change that flips a model's prediction

Find a data point within attack budget ϵ that maximizes the loss

Harmonic mean of four quantitative properties



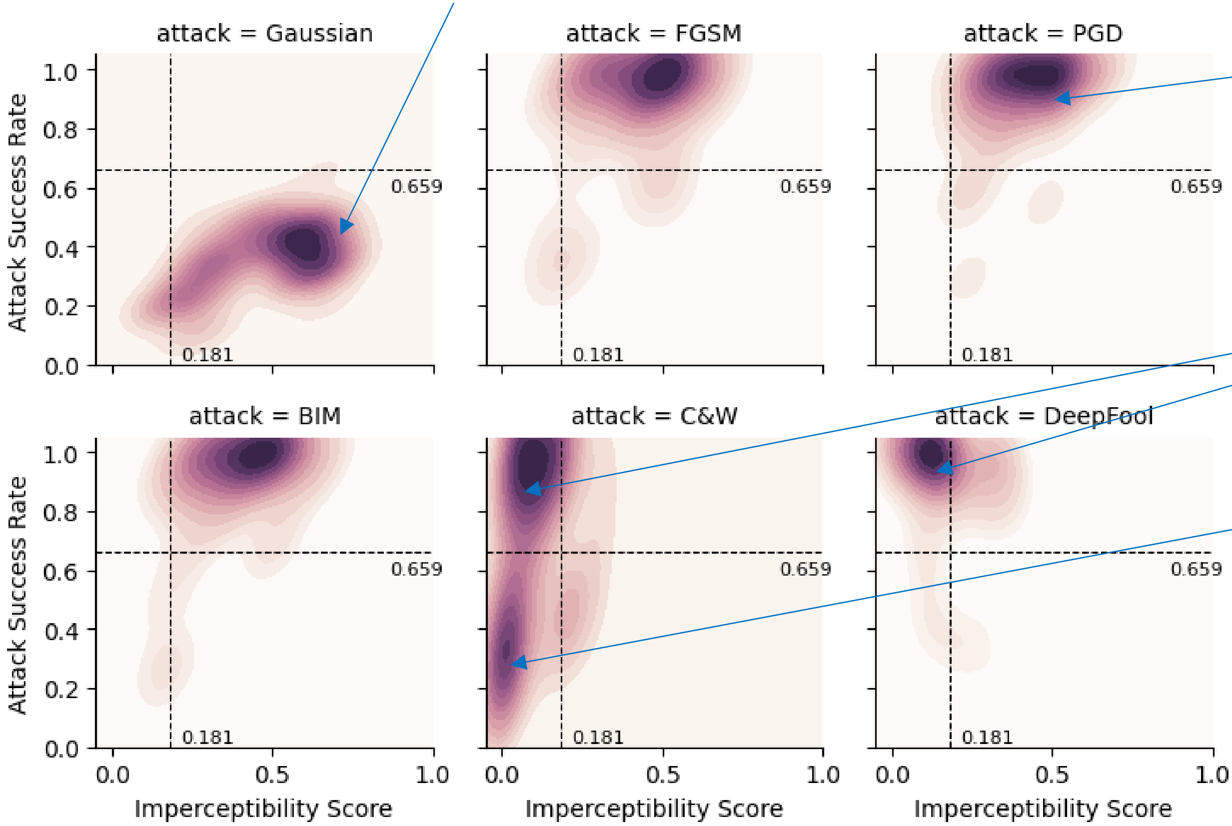
Overall: Effectiveness vs Imperceptibility

Ineffective and perceptible

Effective but perceptible

Effective and imperceptible

Ineffective but imperceptible



Finding

Only DeepFool can generate both effective and imperceptible adversarial examples

Divided into four sectors by maximum ASR value (0.659) and the minimum IS value (0.181) of Gaussian Noise. Higher attack success rate is better. Lower imperceptibility score is better.

Imperceptibility Insights

- ***Sparsity***: Any attack can perturb numerical features. Only PGD can change categorical features on all models.
- ***Proximity***: *Unbounded attacks* (DeepFool and C&W) generally make less changes that *bounded attacks* (FGSM, PGD & BIM) in proximity metrics
- ***Deviation***: *Unbounded attacks* (DeepFool and C&W) more likely generate in-distribution attack examples than *bounded attacks* (FGSM, PGD & BIM)

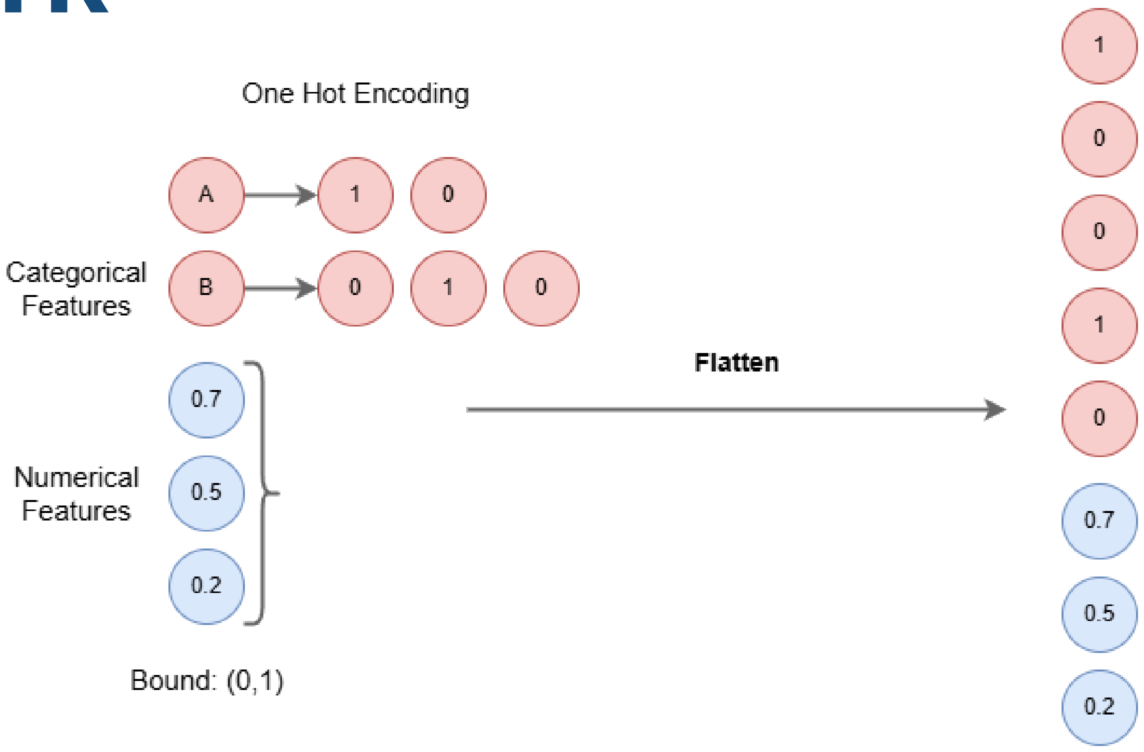
Unbounded Attacks are more promising in generating imperceptible adversarial examples than bounded attack

Zhipeng (Zippo) He
School of Information Systems

Limitation in Benchmark

Is one-hot encoding suitable for adversarial attacks on tabular data?

- *While one-hot encoding simplifies the handling of categorical features by making them compatible with <u>standard distance measurements</u> (such as Lp norms) used for continuous features, it can introduce **more sparse feature space**.*
- *Changing one categorical feature requires perturbation on <u>at least two</u> encoded features.*



Proximity of perturbing one numerical feature from 0 to 1 →

$$\ell_2 = \sqrt{(1-0)^2} = 1$$
$$\ell_\infty = 1$$

Proximity of perturbing categorical feature A from True to False →

$$\ell_2 = \sqrt{(0-1)^2 + (1-0)^2} = \sqrt{2}$$
$$\ell_\infty = 1$$

Research Question 3

How can new adversarial attacks on tabular data be designed to generate both effective and imperceptible adversarial examples?

How to design new tabular attacks

What to do

- Use Unbounded Attacks
- Address properties of imperceptibility

What to avoid

- Make perturbation in original feature space

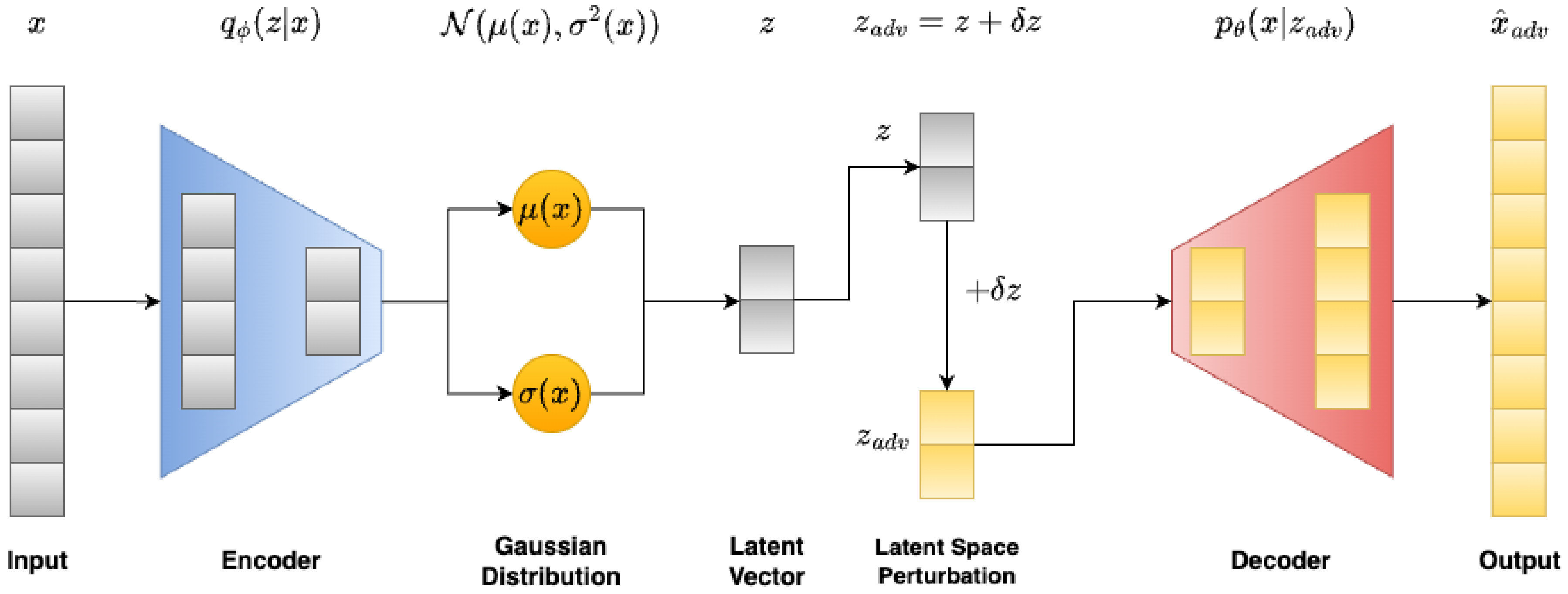
Zhipeng (Zippo) He
School of Information Systems

Ongoing work

To find an adversarial example in latent space,

$$Attack\ Loss = L_{model}(x, \hat{x}_{adv}) + L_{dist}(z, z_{adv}) + L_{spa}$$

Generate adversarial example with a trained Variational Autoencoder (VAE)



Key Takeaways

Proposing a set of imperceptibility properties and metrics for adversarial attacks on tabular data

Benchmarking existing tabular attack on both effectiveness and imperceptibility

Unbounded attacks are more promising in generating both effective and imperceptible adversarial examples

Using VAE to map datasets into latent space and generating adversarial examples in latent space

Zhipeng (Zippo) He
School of Information Systems

Reference:

[1] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

[2] Jin, Z., Zhang, J., Zhu, Z. & Chen, H. (2024). Benchmarking Transferable Adversarial Attacks. *Workshop on AI Systems with Confidential Computing (AISCC) 2024*.

[3] Dong, Y., Fu, Q. A., Yang, X., Pang, T., Su, H., Xiao, Z., & Zhu, J. (2020). Benchmarking adversarial robustness on image classification. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 321-331).

[4] Zheng, M., Yan, X., Zhu, Z., Chen, H., & Wu, B. (2023). BlackboxBench: A Comprehensive Benchmark of Black-box Adversarial Attacks. *arXiv preprint arXiv:2312.16979*.

[5] Croce, F., Andriushchenko, M., Sehwag, V., Debenedetti, E., Flammarion, N., Chiang, M., ... & Hein, M. (2020). Robustbench: a standardized adversarial robustness benchmark. *arXiv preprint arXiv:2010.09670*.

[6] Hingun, N., Sitawarin, C., Li, J., & Wagner, D. (2023). REAP: a large-scale realistic adversarial patch benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 4640-4651).

[7] Cinà, A. E., Rony, J., Pintor, M., Demetrio, L., Demontis, A., Biggio, B., ... & Roli, F. (2024). AttackBench: Evaluating Gradient-based Attacks for Adversarial Examples. *arXiv preprint arXiv:2404.19460*.

[8] Zheng, Q., Zou, X., Dong, Y., Cen, Y., Yin, D., Xu, J., ... & Tang, J. (2021). Graph robustness benchmark: Benchmarking the adversarial robustness of graph machine learning. *arXiv preprint arXiv:2111.04314*.

[9] Li, L., Lei, J., Gan, Z., & Liu, J. (2021). Adversarial vqa: A new benchmark for evaluating the robustness of vqa models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 2042-2051).

Reference:

[10] Siddiqui, S. A., Dengel, A., & Ahmed, S. (2020). Benchmarking adversarial attacks and defenses for time-series data. In *International Conference on Neural Information Processing* (pp. 544-554). Cham: Springer International Publishing.

[11] Eger, S., & Benz, Y. (2020). From hero to zéroe: A benchmark of low-level adversarial attacks. In *Proceedings of the 1st conference of the Asia-Pacific chapter of the association for computational linguistics and the 10th international joint conference on natural language processing* (pp. 786-803).

[12] He, Z., Ouyang, C., Alzubaidi, L., Barros, A., & Moreira, C. (2024). Investigating Imperceptibility of Adversarial Attacks on Tabular Data: An Empirical Analysis. *arXiv preprint arXiv:2407.11463*. [Accepted by ISWA]

[13] Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B., & Swami, A. (2016). The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy (EuroS&P)* (pp. 372-387).

[14] Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data?. *Advances in neural information processing systems*, *35*, 507-520.

[15] Huang, X., Khetan, A., Cvitkovic, M., & Karnin, Z. (2020). Tabtransformer: Tabular data modeling using contextual embeddings. *arXiv preprint arXiv:2012.06678*.

[16] Gorishniy, Y., Rubachev, I., Khrulkov, V., & Babenko, A. (2021). Revisiting deep learning models for tabular data. *Advances in Neural Information Processing Systems*, *34*, 18932-18943.

ACKNOWLEDGEMENT OF TRADITIONAL OWNERS

QUT acknowledges the Turrbal and Yugara, as the First Nations owners of the lands where QUT now stands. We pay respect to their Elders, lores, customs and creation spirits. We recognise that these lands have always been places of teaching, research and learning.

QUT acknowledges the important role Aboriginal and Torres Strait Islander people play within the QUT community.