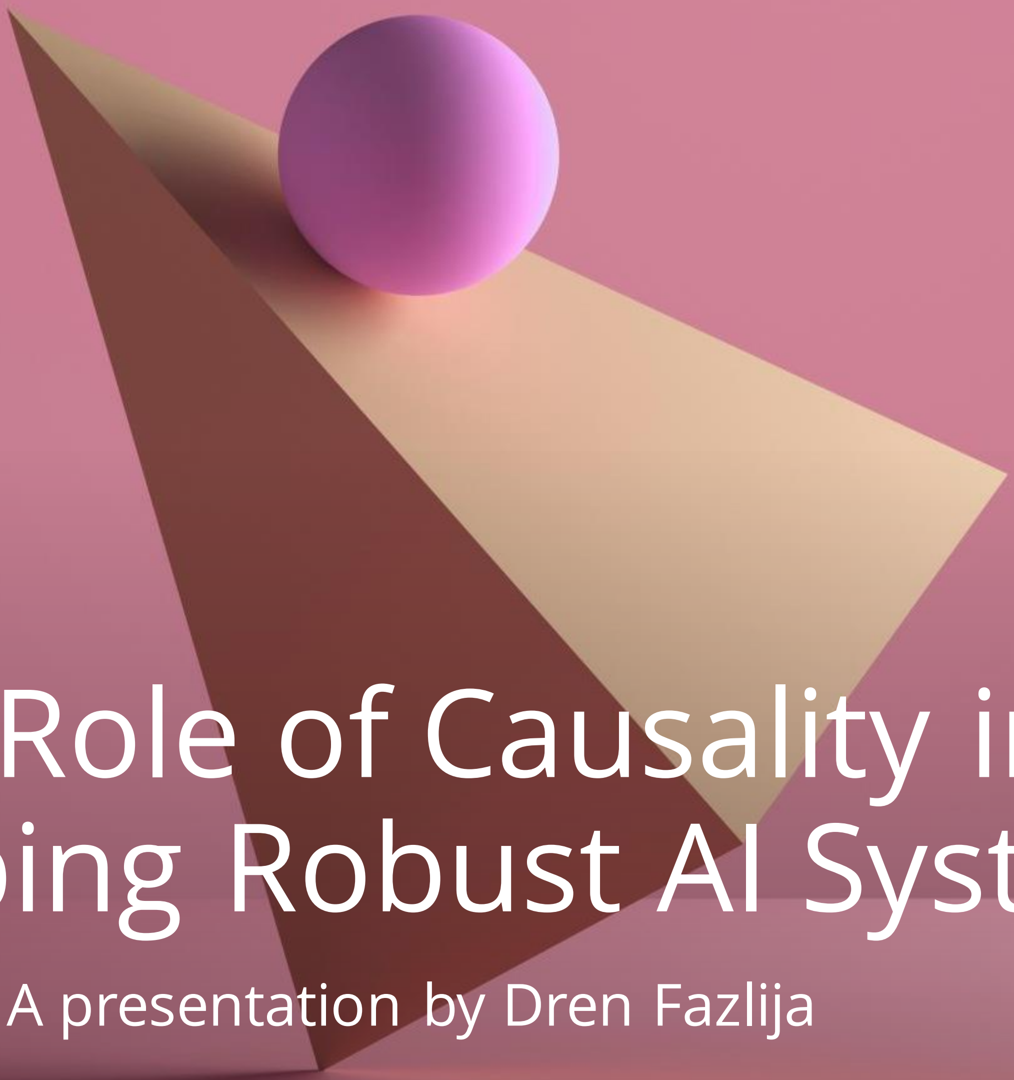


The Role of Causality in Developing Robust AI Systems

A presentation by Dren Fazlija

Robustness and Privacy

- Robustness: Decreasing sensitivity towards input changes

(A) **Cow: 0.99**, Pasture: 0.99, Grass: 0.99, No Person: 0.98, Mammal: 0.98

(B) No Person: 0.99, Water: 0.98, Beach: 0.97, Outdoors: 0.97, Seashore: 0.97

(C) No Person: 0.97, **Mammal: 0.96**, Water: 0.94, Beach: 0.94, Two: 0.94

Source: [Beery2018]

- Privacy: Defending against privacy-evasive attacks
- Causal solutions for both areas overlap significantly
  - Robustness: Methods for centralized learning setting
  - Privacy: Similar methods for decentralized/federated learning setting

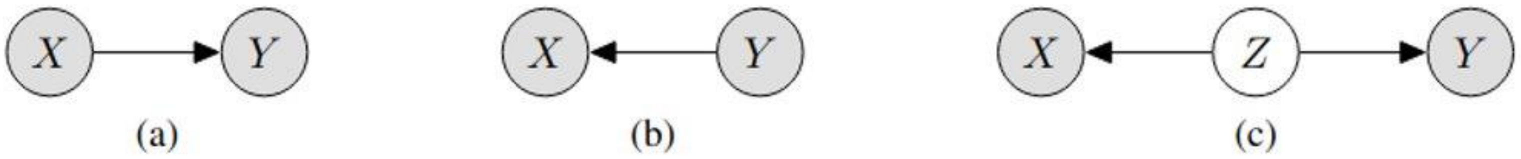
[Beery2018]: Beery, Sara, Grant Van Horn, and Pietro Perona. "Recognition in terra incognita." *Proceedings of the European conference on computer vision (ECCV)*. 2018.

Statistical Machine Learning

- We assume that our data is independent and identically distributed (IID)
- Allows one to infer the performance of models solely through training data
  - Empirical Risk Minimization
- Very unlikely that training data covers all statistical properties of real-world inference data
- Susceptible to distributional shifts caused by unseen data

Enhancing AI with Causality

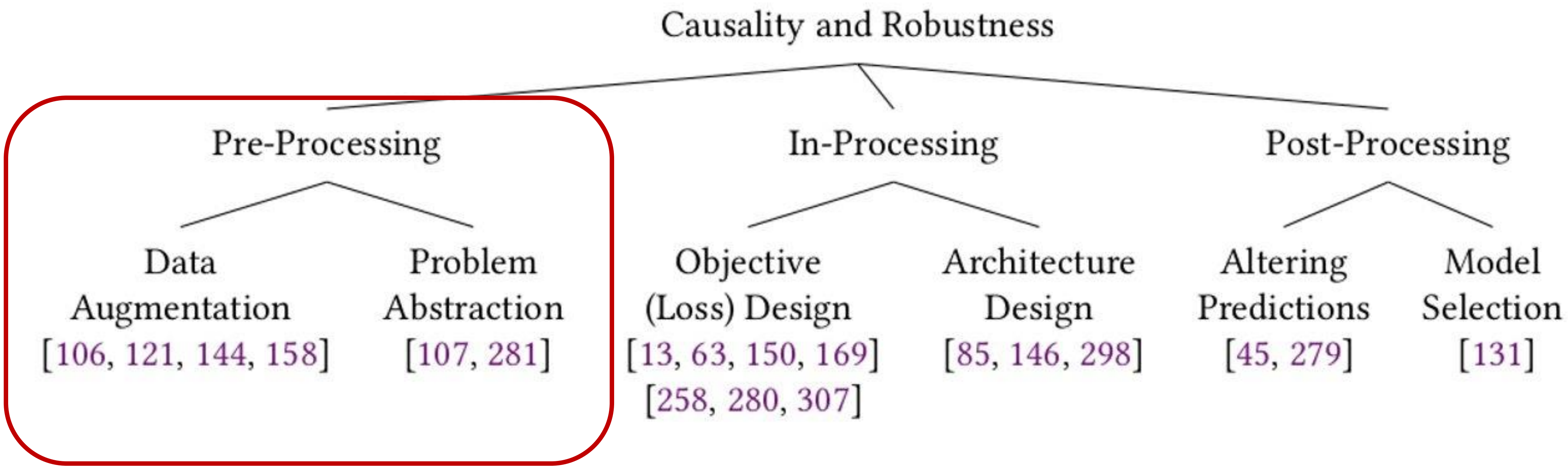
- No definite solution for distributional shifts
- Statistical ML models are not inclined to properly understand causal relationship
  - Simply fall back on observable correlation that works best for the training data
- Causal encodings allow us to constraint this behavior
- Achievable with pre-, in- and post-processing methods



Source: [Schölkopf22]

[Schölkopf22]: Schölkopf, Bernhard, and von Kügelgen, Julius. "From Statistical to Causal Learning." *arXiv*, 2022

Overview of Causal Solutions



Source: [Ganguly23]

[Ganguly23]: Ganguly, Niloy, et al. "A Review of the Role of Causality in Developing Trustworthy AI Systems." *arXiv preprint arXiv:2302.06975* (2023).

Generative Interventions for Causal Learning

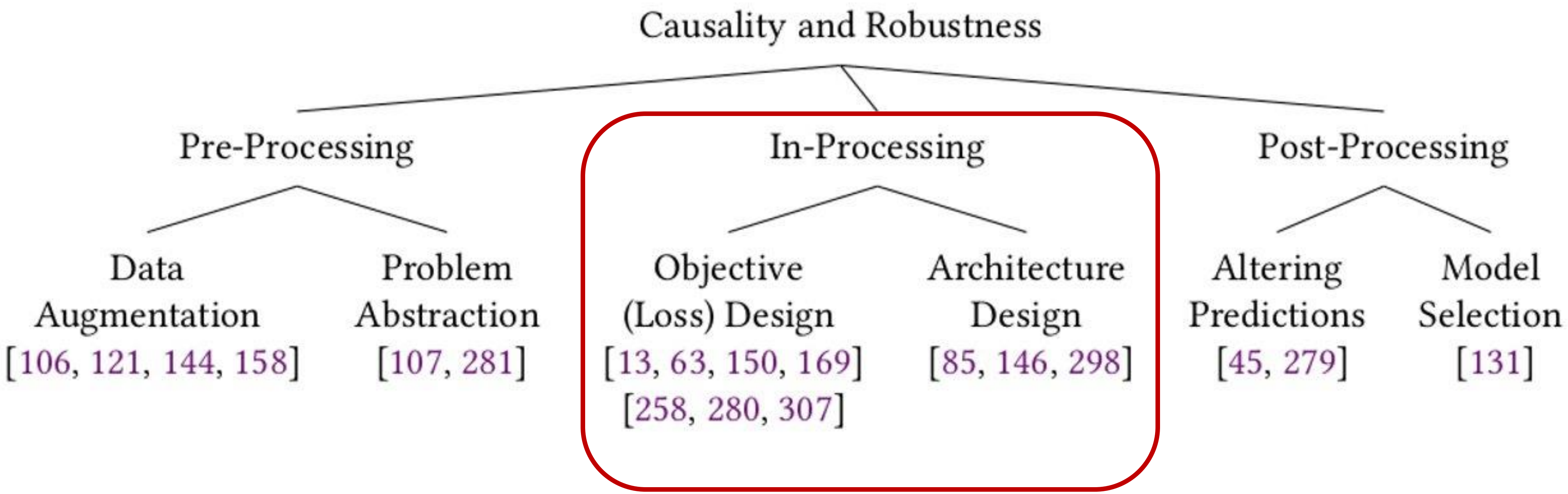
Pre-Processing Method for Robustness [Mao2021]

- Goal: Provide training data whose observable correlations better reflect causal relationships
- Simulate interventions on nuisance factors via GANs

Source: [Mao2021]

[Mao2021] Mao, Chengzhi, et al. "Generative interventions for causal learning." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.

Overview of Causal Solutions



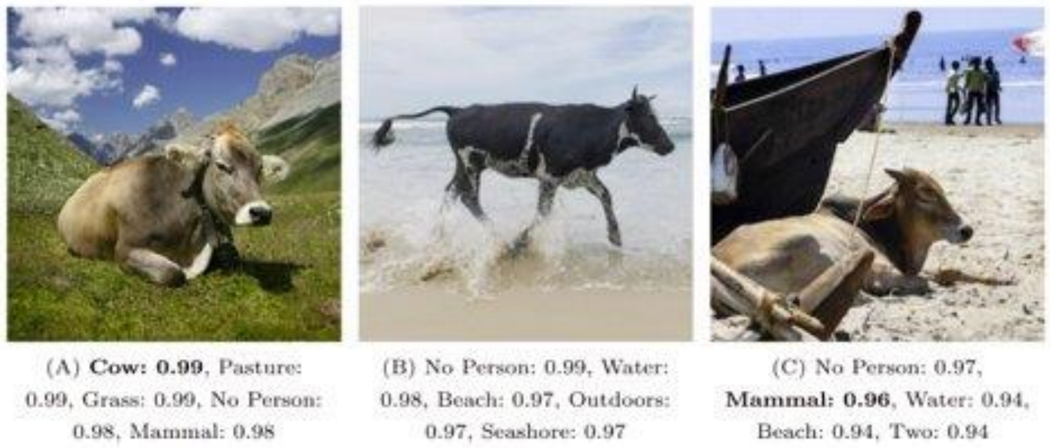
Source: [Ganguly23]

[Ganguly23]: Ganguly, Niloy, et al. "A Review of the Role of Causality in Developing Trustworthy AI Systems." *arXiv preprint arXiv:2302.06975* (2023).

Invariant Risk Minimization

In-Processing Method for Robustness [Arjovsky2019]

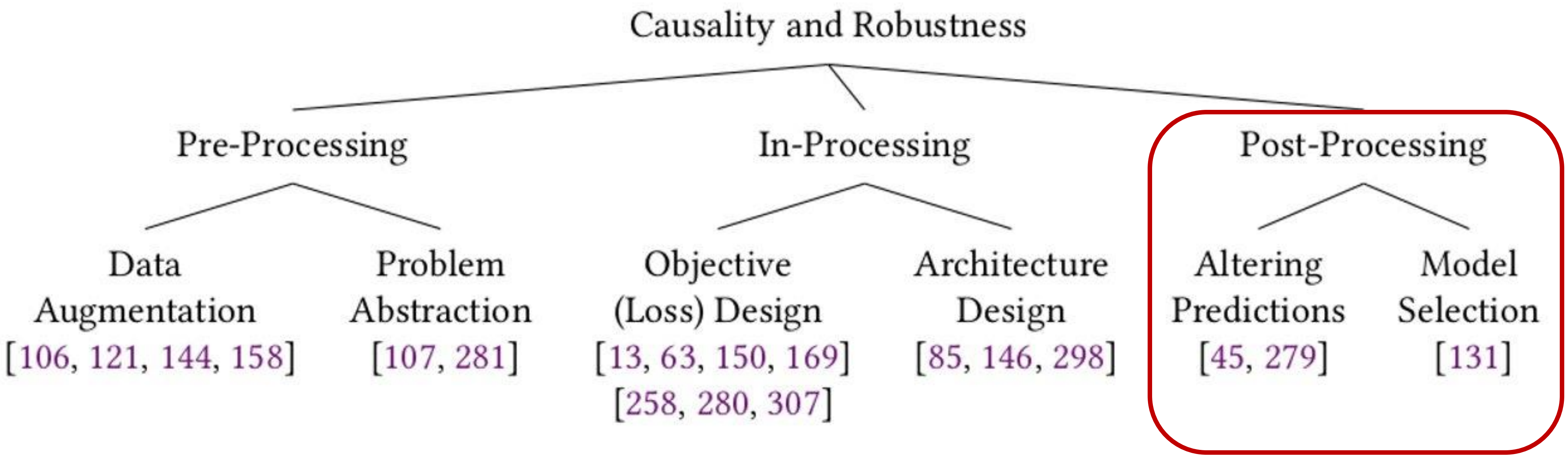
- Feature invariance relates to its causal importance
  - E.g., image background can greatly vary across data points
  - Therefore, it is not important for predicting the label
- Allows one to develop causal models without causal encodings
- Idea: Promote consistent behavior across different environments
- Successful at increasing robustness of image classifiers in the OOD setting



Source: [Beery2018]

[Arjovsky2019]: Arjovsky, Martin, et al. "Invariant risk minimization." *arXiv preprint arXiv:1907.02893* (2019).
[Beery2018]: Beery, Sara, Grant Van Horn, and Pietro Perona. "Recognition in terra incognita." *Proceedings of the European conference on computer vision (ECCV)*. 2018.

Overview of Causal Solutions



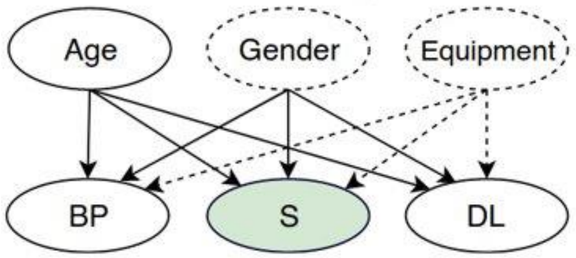
Source: [Ganguly23]

[Ganguly23]: Ganguly, Niloy, et al. "A Review of the Role of Causality in Developing Trustworthy AI Systems." *arXiv preprint arXiv:2302.06975* (2023).

Causal Model Selection

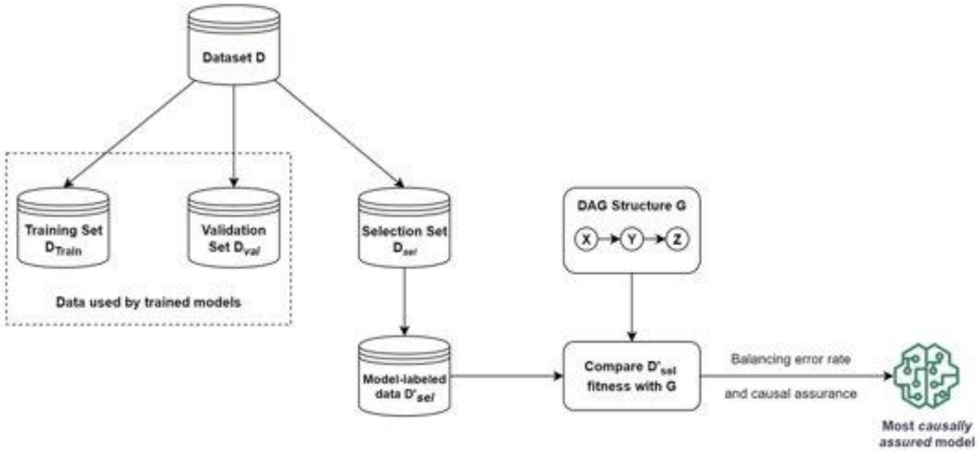
Post-Processing Method for Robustness [Kyono2019]

- Builds on causal invariance assumption
- Goal: pick trained model that best reflects the causality intrinsic to the domain
- Let each model overwrite the labels of the data points with their own prediction
- Resulting dataset should then reflect the very same relationships
- Successfully picks robust models for real-world tabular datasets



(d) Powerlifting dataset

Source: [Kyono2019]

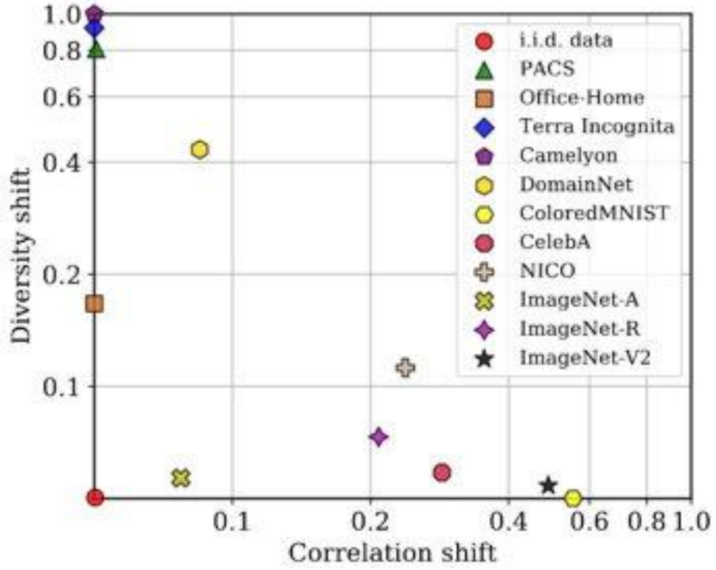


Inspired by [Kyono2019]

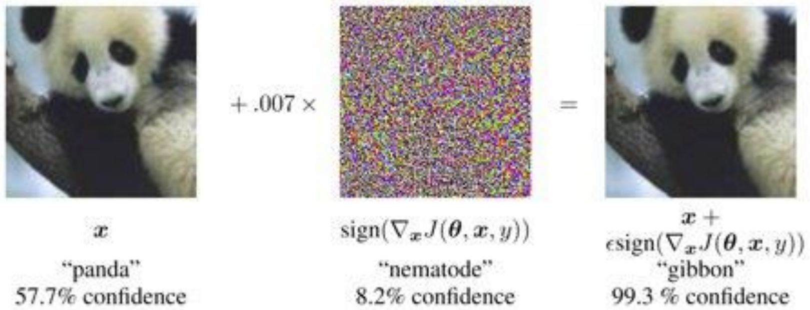
[Kyono2019]: Kyono, Trent, and Mihaela van der Schaar. "Improving model robustness using causal knowledge." *arXiv preprint arXiv:1911.12441* (2019).

Future Work

- Categorize OOD learning abilities of causal solutions
- Explore related fields of Causal ML
  - E.g., Neurosymbolic AI or Object-Centric Learning
- Further explore Adversarial Machine Learning
  - Most solutions are designed for natural OOD data
  - Interesting subarea: Certified robustness



Source: [Ye2022]



Source: [Goodfellow2014]

[Goodfellow2014]: Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." *arXiv preprint arXiv:1412.6572* (2014).
[Ye2022] Ye, Nanyang, et al. "Ood-bench: Quantifying and understanding two dimensions of out-of-distribution generalization." *Proceedings of CVPR* 2022.