# Distribution Alignment

David I. Inouye

Purdue University | Elmore Family School of Electrical and Computer Engineering

# Standard ML assumes **all data** is ***relevant***


Give me all your data!

But what if the data is biased or contains spurious correlations?

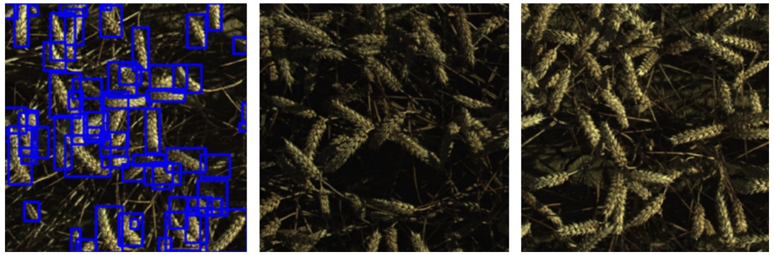
Don't worry,
I'll figure it out.

# What if some data is **assumed (or designed)** to be ***irrelevant***?

- Fair learning (e.g., FNF [Balunović et al., 2022])
  - Sensitive attributes (e.g., race) are *designed* to be irrelevant for social applications (e.g., loan approval)
- Robust learning (e.g., DANN [Ganin et al., 2016], IRM [Arjovsky et al., 2019])
  - The domain of images (e.g., photo vs sketch) is *assumed* to be irrelevant for object detection

Wheat images from Norway

Wheat images from France



Wheat images from Belgium

Image adapted from GlobalWheat dataset images from https://wilds.stanford.edu/datasets/.

# What if some data is **assumed (or designed)** to be ***irrelevant***?

- Unsupervised translation (e.g., CycleGAN [Zhu et al. 2017])
    - The source of images (i.e., real or generated) is *designed* to be irrelevant

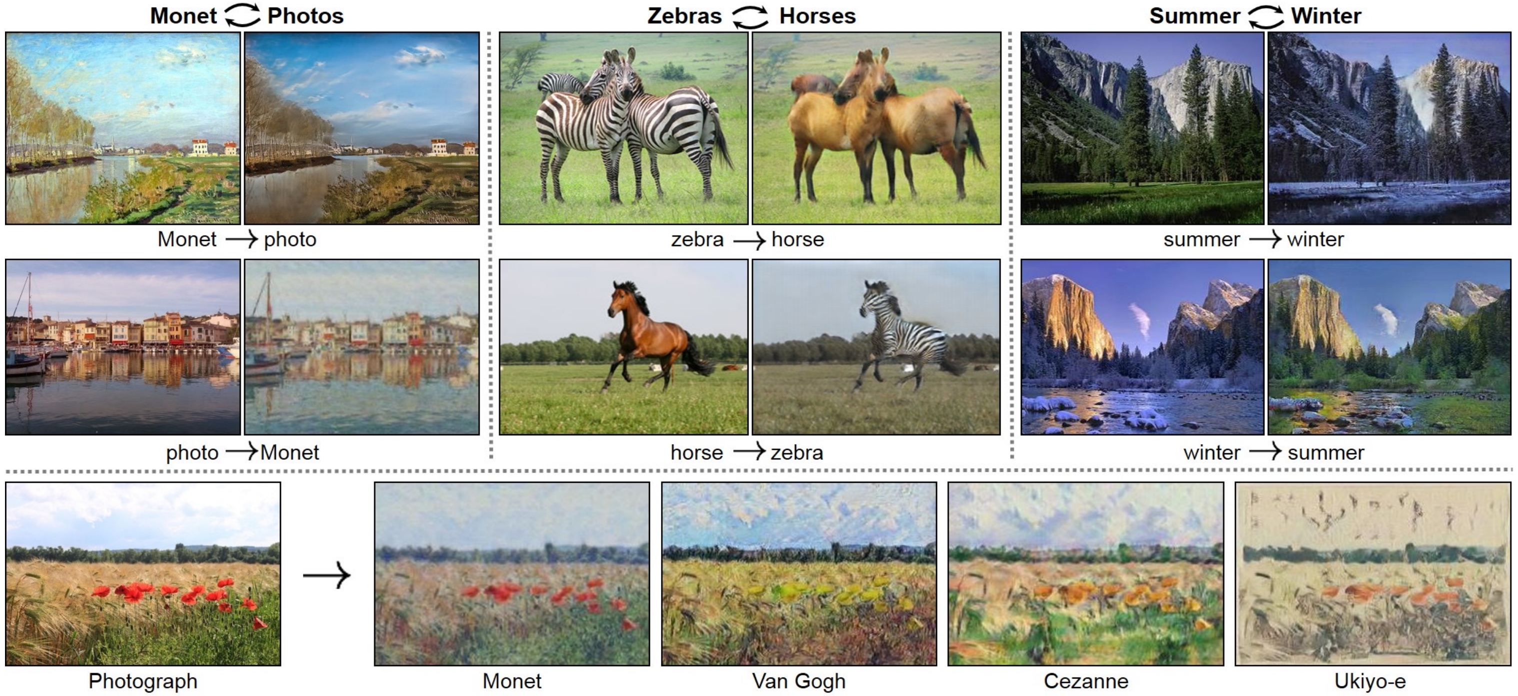
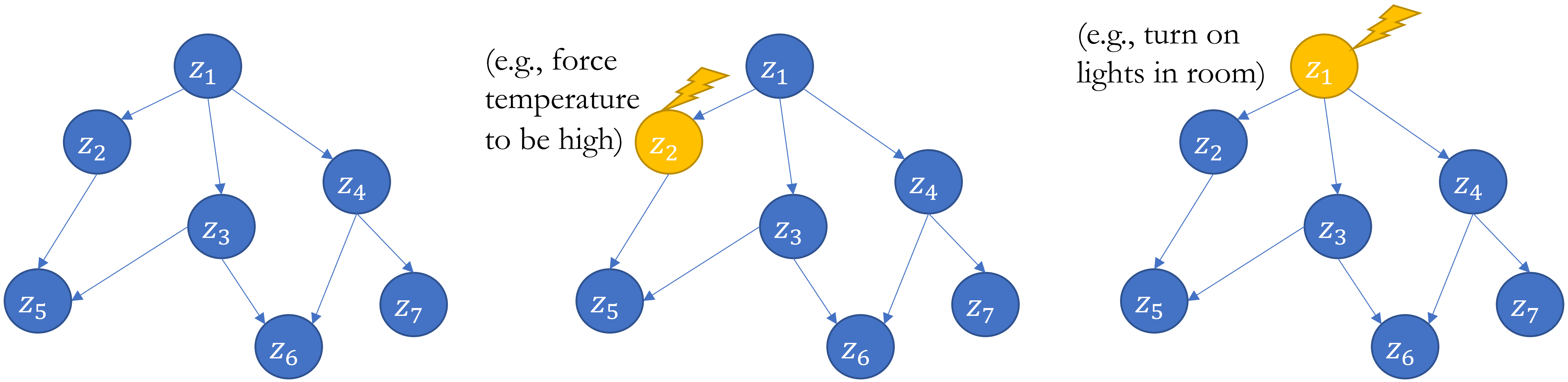



Image from CycleGAN paper: Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 2223-2232).

# What if some data is **assumed (or designed)** to be ***irrelevant***?

- Causal discovery (e.g., ICP [Peters et al., 2016])
  - Interventions are *assumed* to be irrelevant for most causal mechanisms




Observed distribution | Intervened distribution | Another intervened distribution

# How can **known** irrelevant information be used?

- Simply discard irrelevant features
  - However, other features may contain irrelevant information (e.g., while gender is removed, it can be predicted from an applicant's name)
  - Irrelevant features may be *unknown* or *entangled* with relevant features

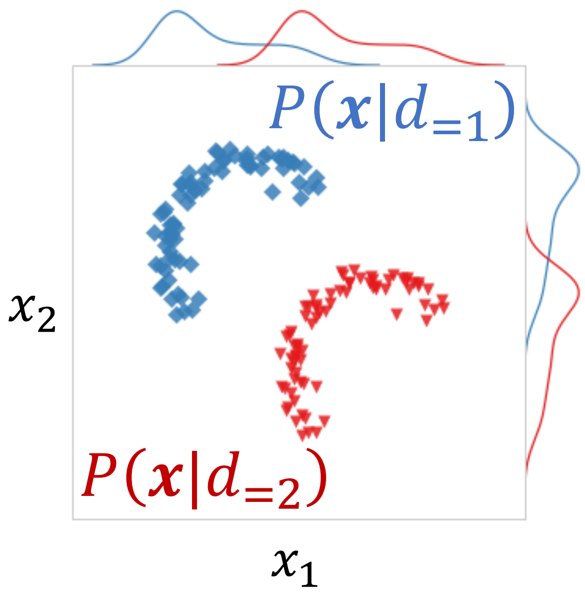
- Model design
  - Hope model *implicitly* ignores irrelevant information (i.e., inductive bias)
  - Design model to *explicitly* ignore easy-to-formalize irrelevant information (e.g., graph models that are invariant to node permutations)

- Distribution alignment (this talk ☺ )
  - *Explicitly* minimize irrelevant information (even if infeasible formalize)

# Alignment Concepts

# **Distribution alignment** is
# the *opposite* objective of classification

Original Space




Optimization Objective

**Classification**

$$\max_g \phi\big(P(g(x)|d_{=1}), P(g(x)|d_{=2})\big)$$

where $g: \mathbb{R}^2 \to \mathbb{R}$ and $\phi$ is a distribution divergence (e.g., KL, JSD, $W_2$)

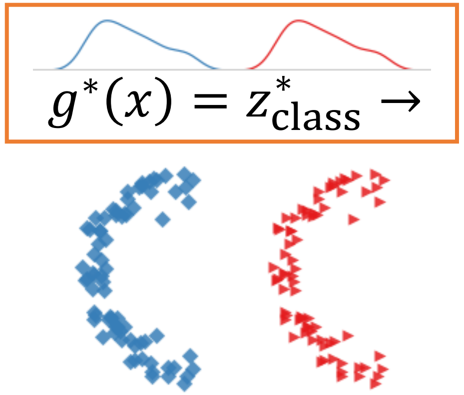
**Distribution alignment**

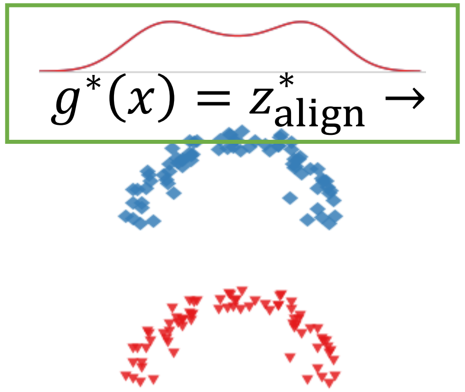
$$\min_g \phi\big(P(g(x)|d_{=1}), P(g(x)|d_{=2})\big)$$

Optimal solution

$$P(g^*(x)|d_{=1}) = P(g^*(x)|d_{=2})$$

Latent Space






David I. Inouye, Purdue University

# Alignment can be with respect to the marginal, conditional, or joint distribution

**Marginal alignment**

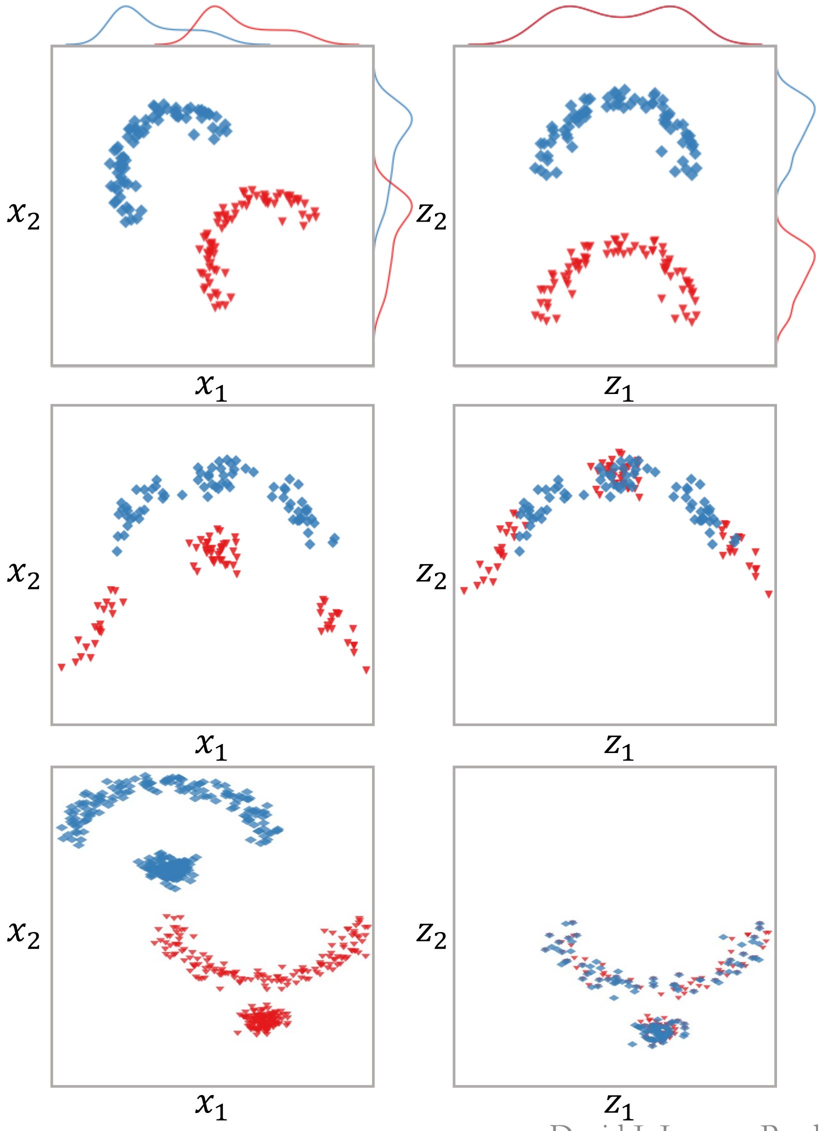
$$P(z_1|d_{=1}) = P(z_1|d_{=2})$$

**Conditional alignment**

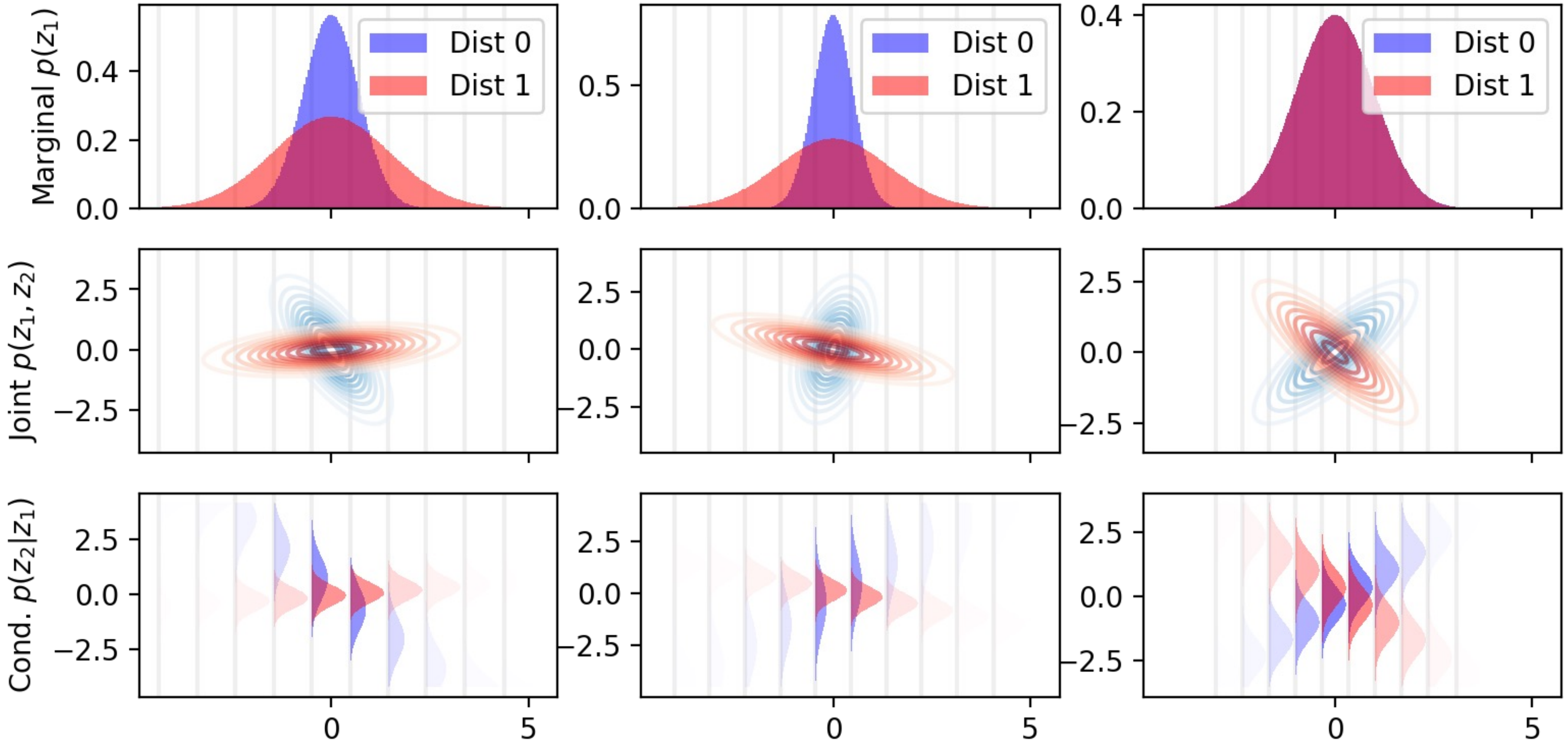
$$P(z_2|z_1, d_{=1}) = P(z_2|z_1, d_{=2})$$

**Joint alignment**

$$P(z_1, z_2|d_{=1}) = P(z_1, z_2|d_{=2})$$



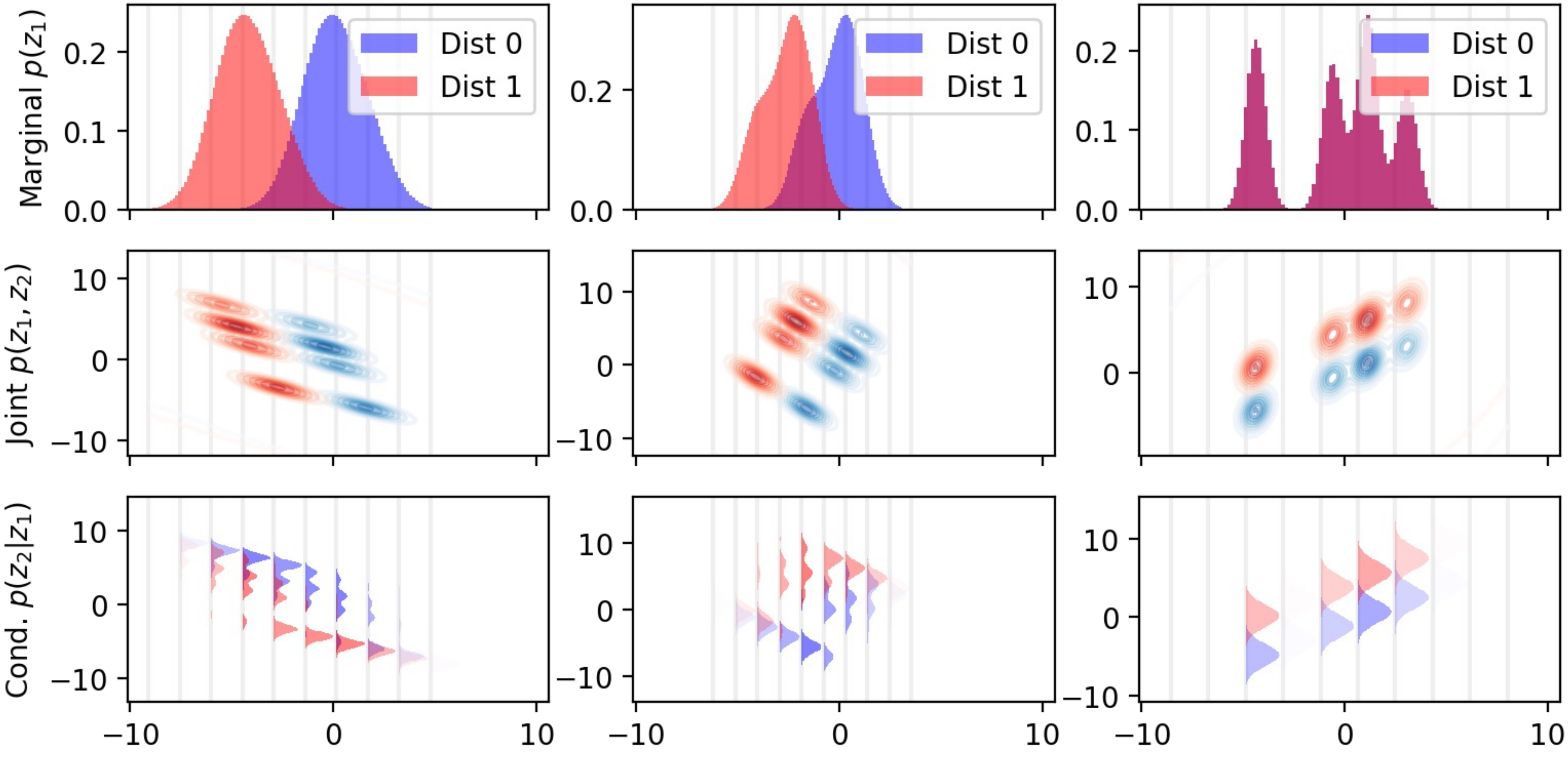
# Example: Marginal alignment without conditional alignment






David I. Inouye, Purdue University

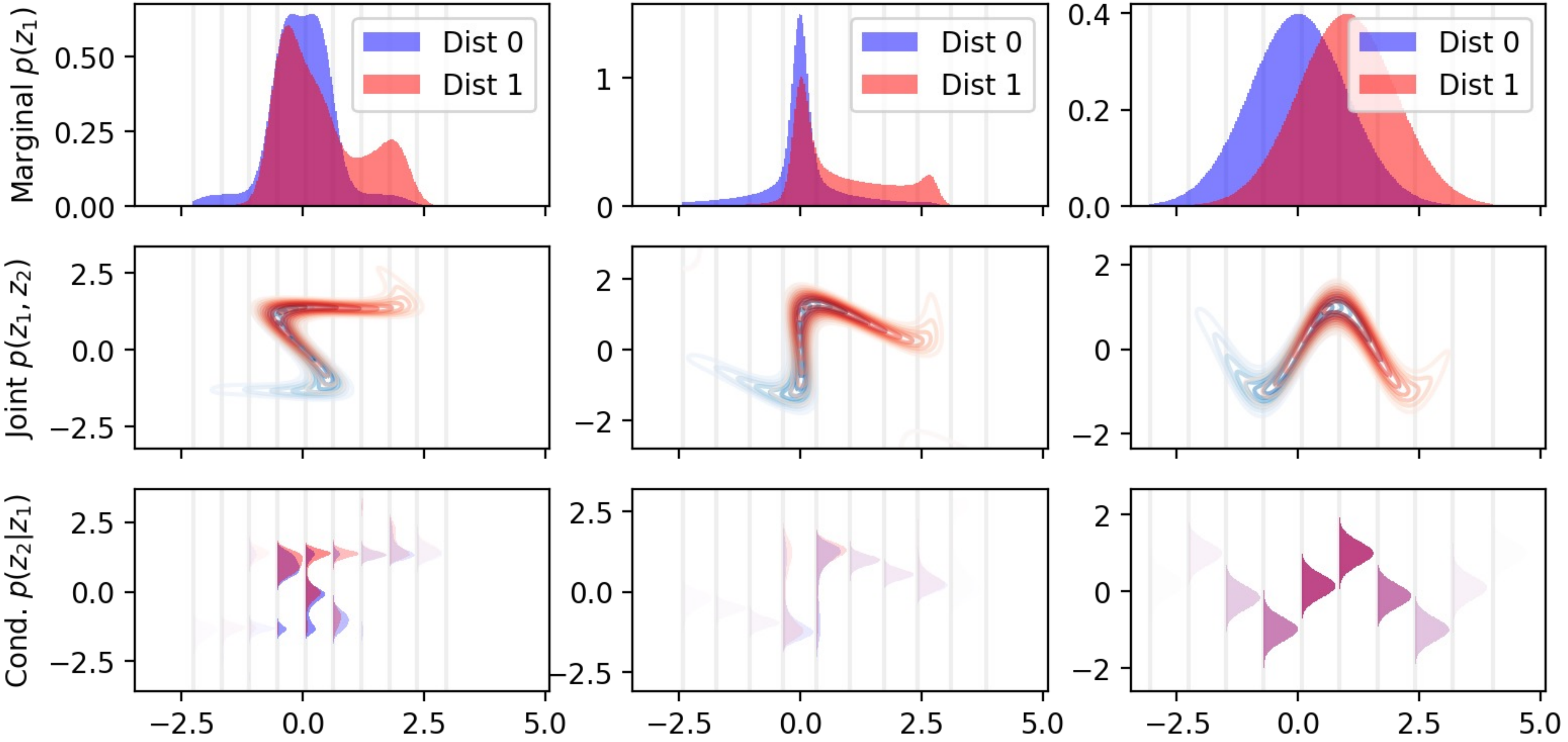
# Example: Marginal alignment without conditional alignment





David I. Inouye, Purdue University

# Example: Conditional alignment without marginal alignment





David I. Inouye, Purdue University

# Distribution alignment minimizes the divergence between two distributions

**Definition 1: Joint Distribution Alignment**

Given samples from the joint distribution $P(\boldsymbol{x}, d)$, *distribution alignment* is the problem of finding an *aligner* $g: \mathcal{X} \times \mathcal{D} \to \mathcal{Z}$ that minimizes a distribution divergence $\phi: \mathcal{P} \times \mathcal{P} \to \mathbb{R}_+$ between the domain-conditional distributions:

Any distribution divergence that satisfies non-negativity and $\phi(P, Q) = 0$ if and only if $P = Q$ (e.g., KL, JSD, $W_2$).

Aligner can depend on domain label *d*

$$\min_{g \in \mathcal{G}} \phi(P(\boldsymbol{z} \mid d_{=1}), P(\boldsymbol{z} \mid d_{=2})), \qquad \text{where} \quad \boldsymbol{z} \equiv g(\boldsymbol{x}, d).$$

**Definition 2: Conditional Distribution Alignment**

Given two variable index sets $\mathcal{A}, \mathcal{B} \in \{1, 2, \dots, m\}$, *conditional alignment* minimizes an aggregation, defined by an aggregator $\Omega_{z_\mathcal{B}}[\cdot]$, over all conditional divergences:

$$\min_{g \in \mathcal{G}} \Omega_{z_\mathcal{B}}[\, \phi(P(\boldsymbol{z}_\mathcal{A} \mid \boldsymbol{z}_\mathcal{B}, d_{=1}), P(\boldsymbol{z}_\mathcal{A} \mid \boldsymbol{z}_\mathcal{B}, d_{=2}))\,], \qquad \text{where} \quad \boldsymbol{z} \equiv g(\boldsymbol{x}, d).$$

Usually this is merely the expectation over $\boldsymbol{z}_\mathcal{B}$, i.e., $\mathbb{E}_{P(\boldsymbol{z}_\mathcal{B})}[\cdot]$



David I. Inouye, Purdue University

# Constraints on aligners can be explicit or implicit

- Explicit constraints
  - *Translation* aligner, i.e., $g(\boldsymbol{x}, d) = \begin{cases} \boldsymbol{x}, & \text{if } d = 1 \\ \tilde{g}(\boldsymbol{x}), & \text{otherwise} \end{cases}$
  - *Shared* aligner between domains, i.e., $g(\boldsymbol{x}, d) = \tilde{g}(\boldsymbol{x})$
  - *Invertible* aligner, i.e., $\exists g^{-1} \text{ s.t. } \forall \boldsymbol{x},\ g^{-1}(g(\boldsymbol{x}, d), d) = \boldsymbol{x}$
    - *Approximately invertible* via cycle consistency $\exists f \text{ s.t. } \forall \boldsymbol{x},\ f(g(\boldsymbol{x}, d), d) \approx \boldsymbol{x}$

- Implicit (soft-)constraints via other optimization terms
  - We will get to this in **alignment applications**

# These definitions encompass all alignment types under a unified framework

**Marginal alignment**

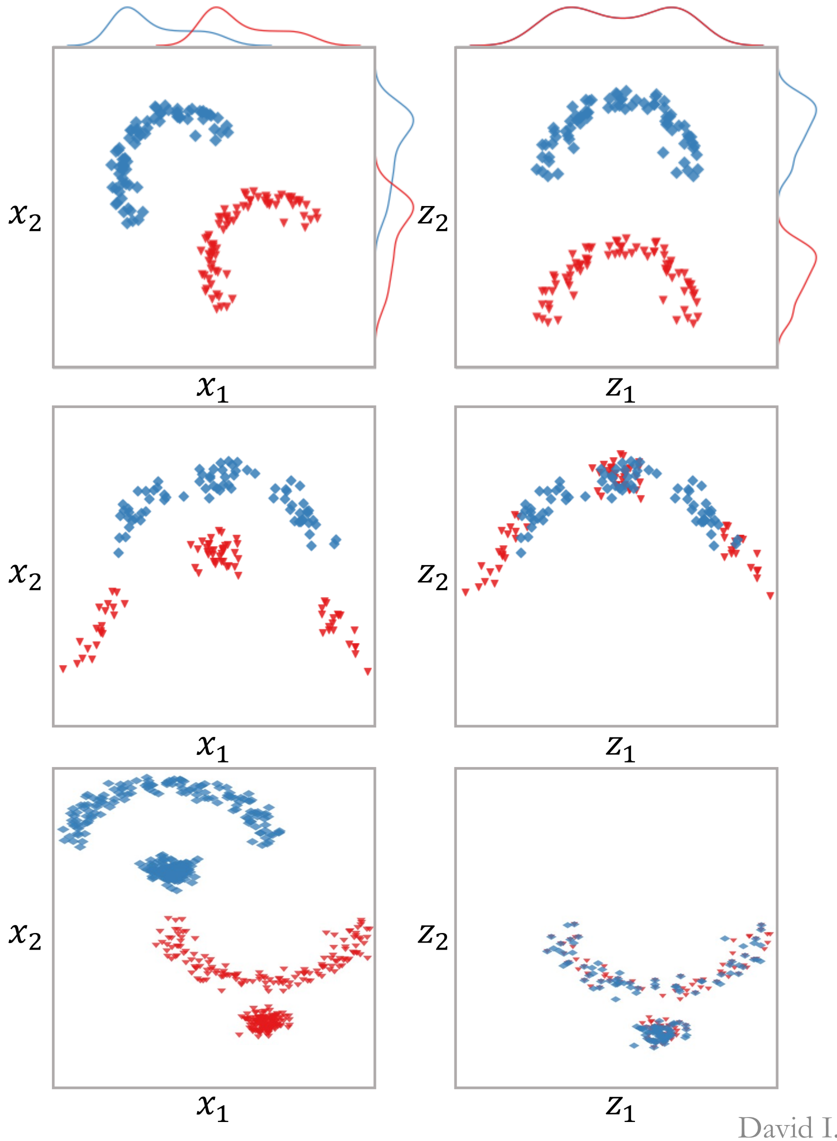
$$P(z_1|d_{=1}) = P(z_1|d_{=2})$$

**Conditional alignment**

$$P(z_2|z_1, d_{=1}) = P(z_2|z_1, d_{=2})$$

**Joint alignment**

$$P(z_1, z_2|d_{=1}) = P(z_1, z_2|d_{=2})$$




Shared aligner
$g(\boldsymbol{x}, d) = Q\boldsymbol{x}$

$z_{\mathcal{A}} = z_1, z_B = \emptyset$

Shift only on y-axis
$g(\boldsymbol{x}, d_{=1}) = \boldsymbol{x}$
$g(\boldsymbol{x}, d_{=2}) = \boldsymbol{x} + [0, a]^T$

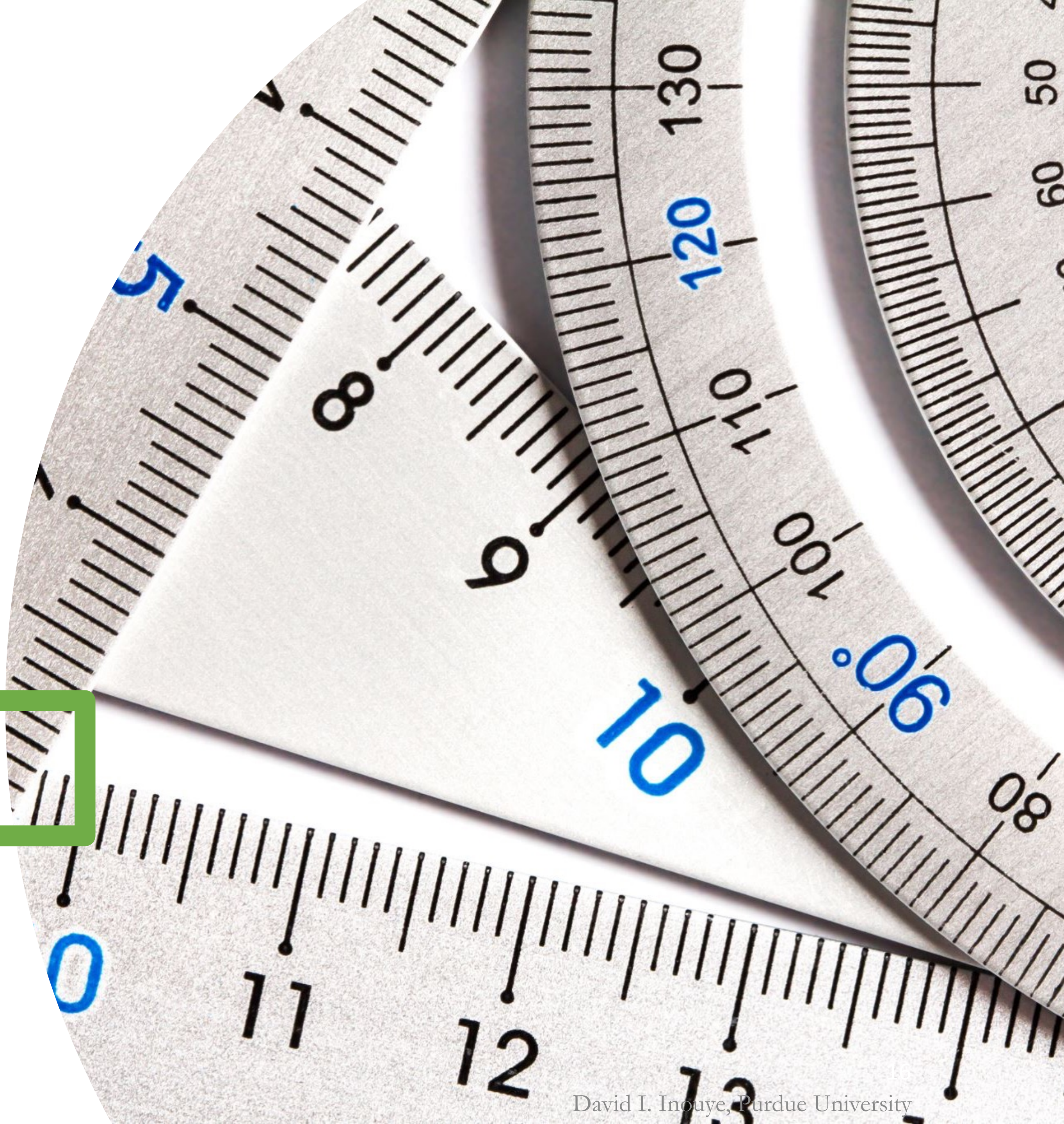
$z_{\mathcal{A}} = z_2, z_B = z_1$

Translation
$g(\boldsymbol{x}, d_{=1}) = \tilde{g}(\boldsymbol{x})$
$g(\boldsymbol{x}, d_{=2}) = \boldsymbol{x}$

$\boldsymbol{z}_{\mathcal{A}} = \boldsymbol{z}, z_B = \emptyset$


David I. Inouye, Purdue University

# Tractable Alignment Measures



David I. Inouye, Purdue University

# *Tractable* alignment measures are needed for evaluation and alignment algorithms

- Two primary uses for alignment measures:
    1. Evaluating or comparing alignment methods
    2. Designing objectives for alignment algorithms
       (i.e., directly minimize alignment measure)

- While theoretic divergences are elegant (e.g., KL, JSD, TV), most of them are intractable to estimate given only samples

    - Thus, this talk focuses on *tractable* alignment measures

# Extrinsic alignment measures have been used for evaluation (but usually not training)

- External task metric
  - Classification accuracy under fair (alignment) constraints
  - Generalization performance on unseen domain (for domain generalization methods that use feature alignment)
  - **Does not measure alignment explicitly**

- Frechet Inception Distance (FID) or Inception Score (IS)
  - Evaluates quality of images from deep generative models based on latent space of Inception v3 network
  - Perceptual measure of image quality and diversity
  - **Inapplicable for applications with limited data or without a well-established semantic latent space**

# Intrinsic measures are used for training but not for evaluation

- Adversarial measures are variational **lower** bounds of divergences

$$\phi_{GAN}(g) = \max_f \mathbb{E}_{P(\boldsymbol{x}|d_{=1})}[\log f(g(\boldsymbol{x},1))] + \mathbb{E}_{P(\boldsymbol{x}|d_{=2})}\left[\log\left(1 - f\left(g(\boldsymbol{x},2)\right)\right)\right]$$

  - If solved perfectly, then $\phi_{GAN}(g) = JSD\left(P(g(\boldsymbol{x},1)|d_{=1}),\ P(g(\boldsymbol{x},2)|d_{=2})\right) + const$
  - If non-optimal, then it is **lower** bound.
  - **Difficult for training (min-max/adversarial) and rarely used for evaluation**

- Other intrinsic measure based on Wasserstein distance
  - Empirical optimal transport algorithms – scales quadratically in number of samples
  - Sliced Wasserstein distance – closed-form solution in 1D via sorting

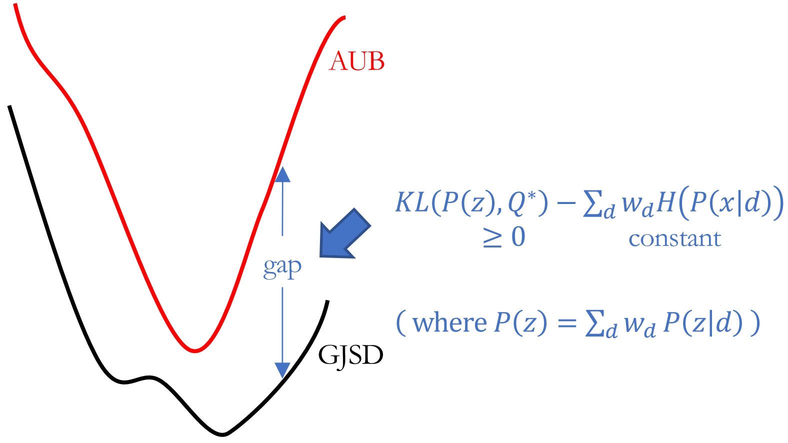
# **Alignment Upper Bound (AUB)** generalizes alignment measures based on ***invertible*** models

- A variational **upper** bound of JSD:

$$\phi_{AUB}(g) = \min_{Q \in \mathcal{Q}} \sum_{d=1}^{k} \mathbb{E}_{P(\boldsymbol{x}|d)}\left[-\log\left|J_{g_d}\right| Q\left(g(\boldsymbol{x}, d)\right)\right]$$

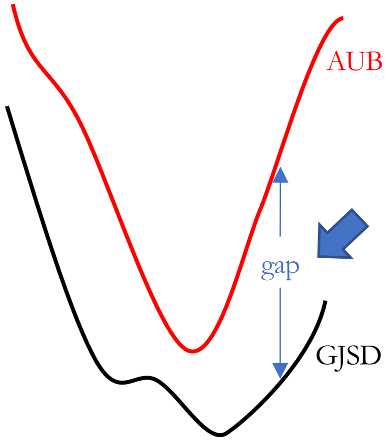
  - $Q$ is a density model ***shared*** among domains
  - $g$ is ***invertible*** and $\left|J_{g_d}\right|$ is the determinant Jacobian of $g(\cdot, d)$

- ***Bound gap*** is exactly $KL\left(\sum_d w_d P(z|d),\; Q(\mathbf{z})\right)$
- ***Any*** $Q$ provides an **upper** bound on JSD + const
- Alignment is ***cooperative***: $\min_g \phi_{AUB}(g) = \min_g \min_Q \ldots$
  - The optimal solution aligns the distributions **regardless of $Q$**




Cho, W., Gong, Z., & Inouye, D. I. (2022). Cooperative Distribution Alignment via JSD Upper Bound. Accepted to *Neural Information Processing Systems (NeurIPS)*. Preprint: https://arxiv.org/abs/2207.02286

# AUB(1): JSD as entropy of mixture minus mixture of entropies

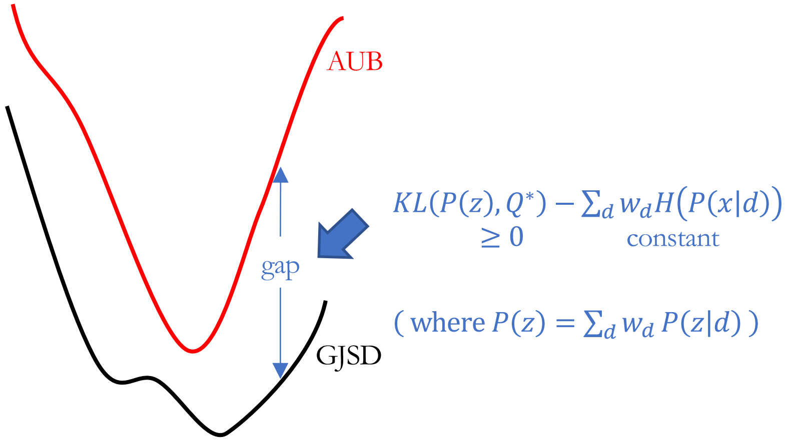



$KL(P(z), Q^*) - \sum_d w_d H(P(x|d))$
$\geq 0$ constant

( where $P(z) = \sum_d w_d \, P(z|d)$ )

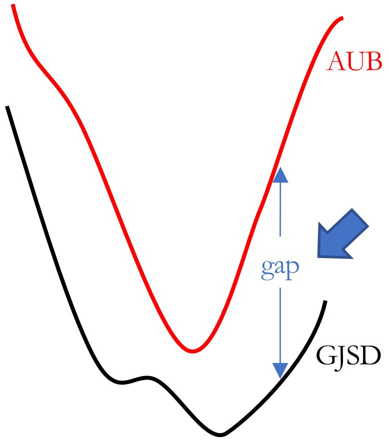
- $JSD\big(P(z|d_{=1}), P(z|d_{=2})\big)$
- $= \sum_d \frac{1}{2} KL\big(P(z|d), P(z)\big) \qquad \big(\text{Let } P(z) = \sum_d \frac{1}{2} P(z|d), \text{i.e., a mixture}\big)$
- $= \sum_d \frac{1}{2} \mathbb{E}_{P(z|d)} \left[\log \frac{P(z|d)}{P(z)}\right]$
- $= \sum_d \frac{1}{2} \mathbb{E}_{P(z|d)}[-\log P(z)] - \sum_d \frac{1}{2} \mathbb{E}_{P(z|d)}[-\log P(z|d)]$
- $= \sum_d \frac{1}{2} \int_{\mathcal{Z}} P(z|d)(-\log P(z))dz - \sum_d \frac{1}{2} H\big(P(z|d)\big)$
- $= \int_{\mathcal{Z}} \sum_d \frac{1}{2} P(z|d) \big(-\log P(z)\big)dz - \sum_d \frac{1}{2} H\big(P(z|d)\big)$
- $= \int_{\mathcal{Z}} P(z) \big(-\log P(z)\big)dz - \sum_d \frac{1}{2} H\big(P(z|d)\big)$
- $= H\big(P(z)\big) - \sum_d \frac{1}{2} H\big(P(z|d)\big)$

# AUB(2): Latent entropy is observed entropy + log determinant term




- $H\big(P(z|d)\big)$
- $= \mathbb{E}_{P(z|d)}[-\log P(z|d)]$
- $= \mathbb{E}_{P(x|d)}[-\log P(z = g(x, d)|d)]$
- $= \mathbb{E}_{P(x|d)}\left[-\log P(x|d)\left|J_{g_d}(x)\right|^{-1}\right]$
- $= \mathbb{E}_{P(x|d)}[-\log P(x|d)] + \mathbb{E}_{P(x|d)}\left[-\log\left|J_{g_d}(x)\right|^{-1}\right]$
- $= H\big(P(x|d)\big) + \mathbb{E}_{P(x|d)}\left[-\log\left|J_{g_d}(x)\right|^{-1}\right]$

# AUB(3): Latent cross entropy is weighted observed cross entropy

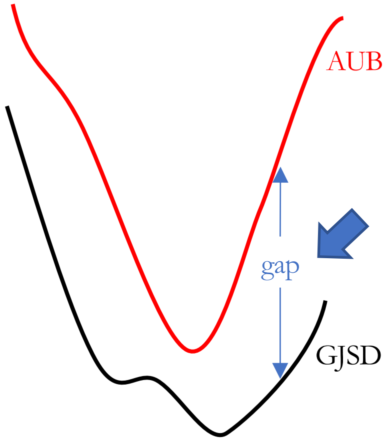



$KL(P(z), Q^*) - \sum_d w_d H(P(x|d))$
$\geq 0$ constant

( where $P(z) = \sum_d w_d \, P(z|d)$ )

- $H_c\big(P(z), Q(z)\big) \equiv \mathbb{E}_{P(z)}\big[-\log\big(Q(z)\big)\big]$
  (Note that: $KL(P, Q) = H_c(P, Q) - H(P)$)
- $= -\int_{\mathcal{Z}} \sum_d \frac{1}{2} P(z|d) \log\big(Q(z)\big)\, dz$
- $= -\sum_d \frac{1}{2} \int_{\mathcal{Z}} P(z|d) \log\big(Q(z)\big)\, dz$
- $= \sum_d \frac{1}{2} \mathbb{E}_{P(z|d)}\big[-\log\big(Q(z)\big)\big]$
- $= \sum_d \frac{1}{2} \mathbb{E}_{P(x|d)}\Big[-\log\Big(Q\big(g(x, d)\big)\Big)\Big]$

# AUB(4): AUB is upper bound on JSD + const




$KL(P(z), Q^*) - \sum_d w_d H(P(x|d))$

$\geq 0$ constant

( where $P(z) = \sum_d w_d \, P(z|d)$ )

- $JSD\big(P(z|d_{=1}), P(z|d_{=2})\big)$
- $= H\big(P(z)\big) - \sum_d \frac{1}{2} H\big(P(z|d)\big)$
- $= H_c\big(P(z), Q(z)\big) - H_c\big(P(z), Q(z)\big) + H\big(P(z)\big) - \sum_d \frac{1}{2} H\big(P(z|d)\big)$
- $= H_c\big(P(z), Q(z)\big) - KL\big(P(z), Q(z)\big) - \sum_d \frac{1}{2} H\big(P(z|d)\big)$
- $\leq H_c\big(P(z), Q(z)\big) - \sum_d \frac{1}{2} H\big(P(z|d)\big)$
- $= \sum_d \frac{1}{2} \mathbb{E}_{P(x|d)}\left[-\log\Big(Q\big(g(x,d)\big)\Big)\right] - \sum_d \frac{1}{2}\Big(\mathbb{E}_{P(x|d)}\big[\log|J_{g_d}(x)|\big] + H\big(P(x|d)\big)\Big)$
- $= \sum_d \frac{1}{2} \mathbb{E}_{P(x|d)}\left[-\log\Big(|J_{g_d}(x)| Q\big(g(x,d)\big)\Big)\right] - \sum_d \frac{1}{2} H\big(P(x|d)\big)$

Constant w.r.t $g$

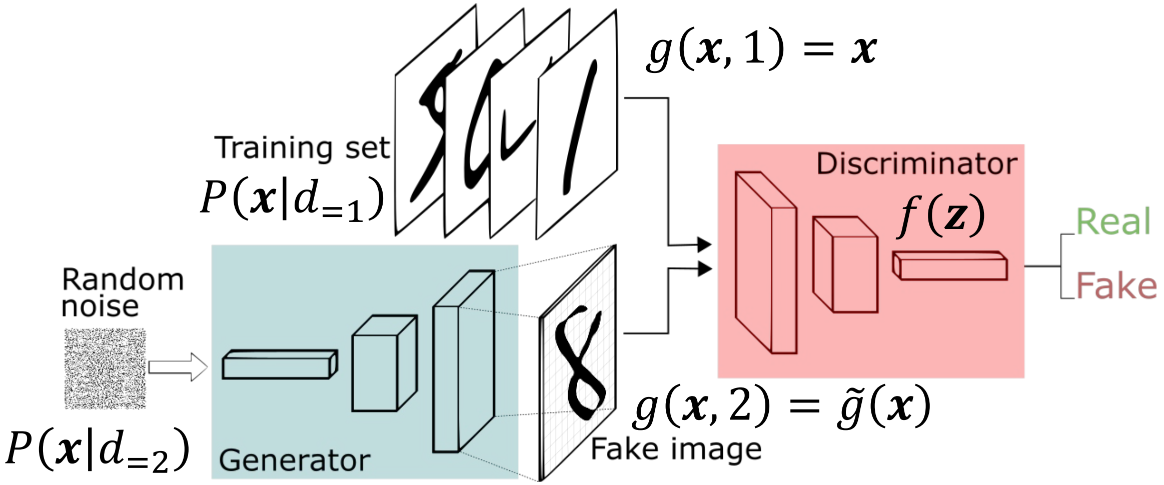
# Alignment Algorithms

# Adversarial optimization (GAN-inspired) is the standard approach to alignment

- Intuition – Competitive game
  - Counterfeiter is trying to avoid getting caught
  - Police is trying to catch counterfeiter
- Algorithm – Usually alternating optimization between min and max
- Benefits
  - No constraints on generator and discriminator models
- Drawbacks
  - Lacks domain-agnostic evaluation metrics (e.g., unable to check for overfitting)
  - Unstable or poorly conditioned optimization

**Adversarial alignment problem**

$$\min_g \max_f \mathbb{E}_{P(\boldsymbol{x}|d_{=1})}[\log f(g(\boldsymbol{x},1))] + \mathbb{E}_{P(\boldsymbol{x}|d_{=2})}\left[\log\left(1 - f(g(\boldsymbol{x},2))\right)\right]$$

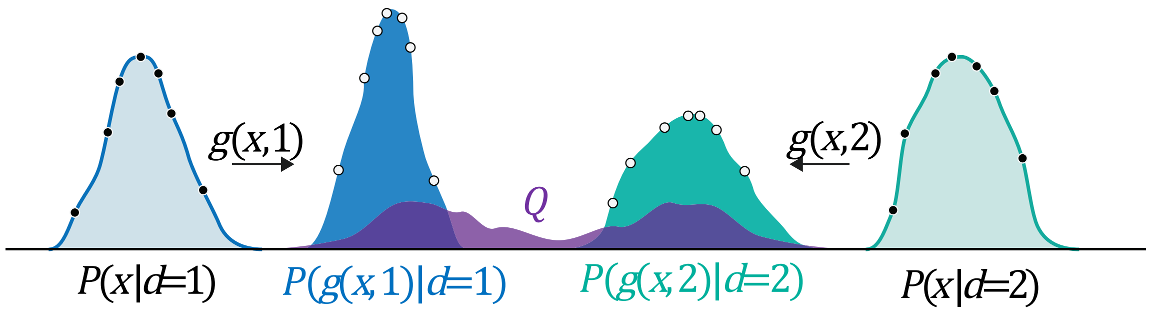


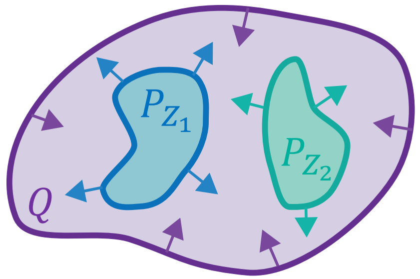

https://www.freecodecamp.org/news/an-intuitive-introduction-to-generative-adversarial-networks-gans-7a2264a81394/

# AUB optimization provides a **cooperative** alternative to adversarial alignment

**AUB cooperative alignment problem**

$$\min_{g} \min_{Q \in \mathcal{Q}} \sum_{j=1}^{k} \mathbb{E}_{P(\boldsymbol{x}|d)}\left[\log|J_{g_d}| Q(g(\boldsymbol{x}, d))\right]$$



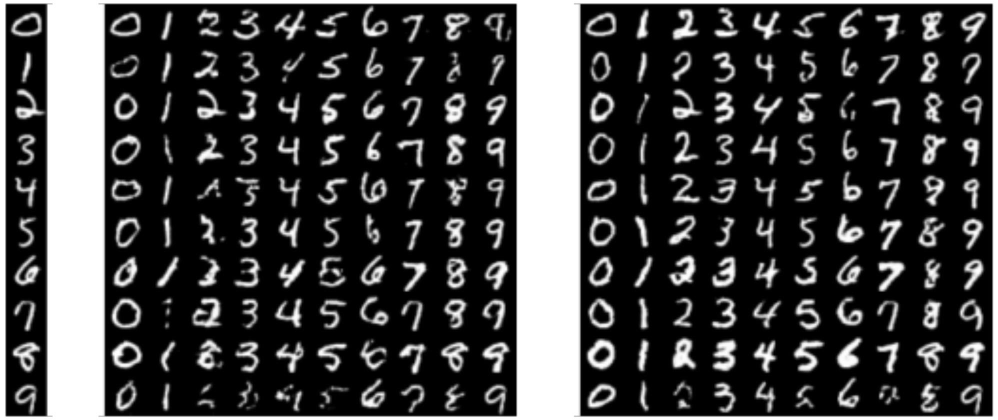

- Minimizing $g$ makes distributions closer to current $Q$ (left)
- Minimizing $Q$ tightens bound by getting closer to the latent mixture, i.e., $\sum_d P(g(x,d)|d)$ (right)


Cho, W., Gong, Z., & Inouye, D. I. (2022). Cooperative Distribution Alignment via JSD Upper Bound. Accepted to *Neural Information Processing Systems (NeurIPS)*. Preprint: https://arxiv.org/abs/2207.02286

# AUB can perform alignment on tabular data and between multiple domains

| | MINIBOONE (42) | GAS (7) | HEPMASS (20) | POWER (5) |
|---|---|---|---|---|
| LRMF | 12.79 | -6.17 | 18.49 | -0.93 |
| AF (MLE) | 14.08 | -6.52 | 19.37 | -0.77 |
| AF (Adv. only) | 18.18 | -3.15 | 21.70 | -0.39 |
| AF (hybrid) | 19.49 | -3.76 | 21.42 | -0.43 |
| Ours | **12.11** | **-7.09** | **18.26** | **-1.19** |

These results on 4 benchmark tabular datasets demonstrate that our algorithm can improve the AUB alignment measure on test data.



AlignFlow (MLE) Ours

Our AUB algorithm can translate between 10 domains (MNIST digits here) better than the closest competitor (AlignFlow) for invertible models. (Original real digits are far left and grid is translations to all other digits.)

Cho, W., Gong, Z., & Inouye, D. I. (2022). Cooperative Distribution Alignment via JSD Upper Bound. Accepted to *Neural Information Processing Systems (NeurIPS)*. Preprint: [https://arxiv.org/abs/2207.02286](https://arxiv.org/abs/2207.02286)

# Iterative alignment flows iteratively solve 1D alignment problems to create deep aligner

1. Find 1D projection that is **maximally misaligned** (i.e., max sliced Wasserstein distance)

$$\max_{\theta} W_2\left(P(\theta^T \boldsymbol{x}|d_{=1}), P(\theta^T \boldsymbol{x}|d_{=2})\right)$$

2. Align along this 1D projection by mapping to barycenter distribution

$$\min_{g} \mathbb{E}_{P(\tilde{x},d)}[\|g(\tilde{x},d) - \tilde{x}\|^2]$$
$$\text{s.t.} \quad \tilde{x} = \theta^T \boldsymbol{x}, \qquad P(g(\tilde{x},1)|d_{=1}) = P(g(\tilde{x},2)|d_{=2})$$

3. Update aligner (add one layer) and repeat

$$\tilde{g}(\boldsymbol{x}) = g(\theta^T \boldsymbol{x}, d)\theta + \boldsymbol{x}_{\boldsymbol{\theta}}^{\perp}$$

$$\tilde{g}_{\text{global}}^{\text{new}} = \tilde{g} \circ \tilde{g}_{\text{global}}^{\text{old}}$$

$$x^{new} = \tilde{g}(\boldsymbol{x})$$

Zhou, Z., Gong, Z., Ravikumar, P., & Inouye, D. I. (2022, May). Iterative Alignment Flows. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*. https://proceedings.mlr.press/v151/zhou22b/zhou22b.pdf

# INB is significantly faster than the closest invertible model baselines

| | Model | WD | FID | TC | Time(s) |
|---|---|---|---|---|---|
| Ours | NB | $60.010 \pm 0.000$ | $229.551 \pm 0.000$ | $\mathbf{28.115 \pm 0.000}$ | **25** |
| | INB ($L = 20$) | $23.481 \pm 0.161$ | $43.196 \pm 0.588$ | $31.671 \pm 0.056$ | 430 |
| | INB ($L = 250$) | $\mathbf{23.183 \pm 0.095}$ | $\mathbf{37.480 \pm 0.008}$ | $32.841 \pm 0.097$ | 2200 |
| Iterative Baselines | DD | $39.079 \pm 0.000$ | $166.320 \pm 0.000$ | $235.164 \pm 0.000$ | 360 |
| | SINF-Align($0 \Rightarrow 1$) | $50.151 \pm 0.950$ | $247.142 \pm 0.972$ | — | 50 |
| | SINF-Align($1 \Rightarrow 0$) | $42.658 \pm 1.253$ | $202.058 \pm 1.716$ | — | 50 |
| Deep Baselines | AlignFlow($\lambda$ =1e-4) | 56.386 | 158.654 | 392.578 | 220000 |
| | AlignFlow($\lambda$ =1e-5) | 60.452 | 191.983 | 412.531 | 220000 |

Zhou, Z., Gong, Z., Ravikumar, P., & Inouye, D. I. (2022, May). Iterative Alignment Flows. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*. https://proceedings.mlr.press/v151/zhou22b/zhou22b.pdf

# Alignment Applications

Alignment applications can be unified as a task objective + ***(soft) alignment constraints***

**Task objective**

- "What we want"
- Relevant information

**Alignment constraints**

- "What we don't want"
- Irrelevant information


David I. Inouye, Purdue University

# Fair classification aims to classify correctly while controlling for sensitive attributes

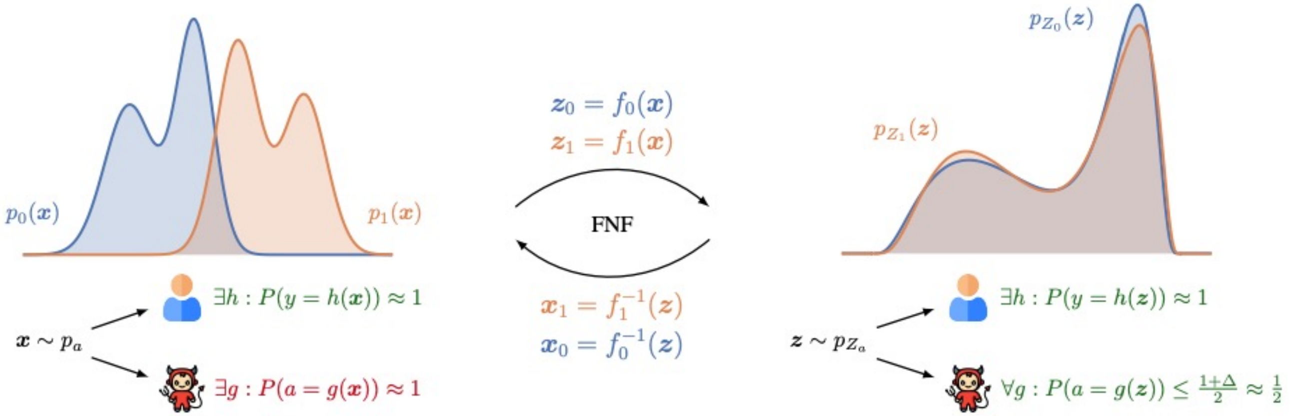
**Task objective:**
**"What we want" / "relevant"**

- Accurately predict whether a loan application should be approved
- Standard classification loss
  $$\mathbb{E}_{P(\boldsymbol{x},d)}\left[\ell\left(f\left(g(x,d)\right),y\right)\right]$$

**(Soft) alignment constraints:**
**"What we don't want" / "irrelevant"**

- The prediction must be *independent* of sensitive attribute $\boldsymbol{d}$
- Alignment constraint loss
  $$\phi\left(P(g(\boldsymbol{x},d)|d_{=1}),P(g(\boldsymbol{x},d)|d_{=2})\right)$$

Raw representation is good for task but sensitive attribute can be determined




Aligned representation is good for task but sensitive attribute **cannot** be determined

Illustration from: Balunovic, M., Ruoss, A., & Vechev, M. (2021, September). Fair normalizing flows. In *International Conference on Learning Representations.*

# Unsupervised image-to-image translation aims to preserve content while changing domains

**Task objective:**
**"What we want" / "relevant"**

- Preserve semantic image content
- Both explicit and implicit methods (e.g., CycleGAN)
  - Cycle consistency loss (explicit)
  - Identity regularization (explicit)
  - CNN architecture (implicit)

**(Soft) alignment constraints:**
**"What we don't want" / "irrelevant"**

- Change the style (or domain) of the image
  - Translated image should "look like" images from the other domain
- Alignment constraint loss
  $$\phi\big(P(g(\boldsymbol{x},d)|d_{=1}),P(\boldsymbol{x}|d_{=2})\big)$$

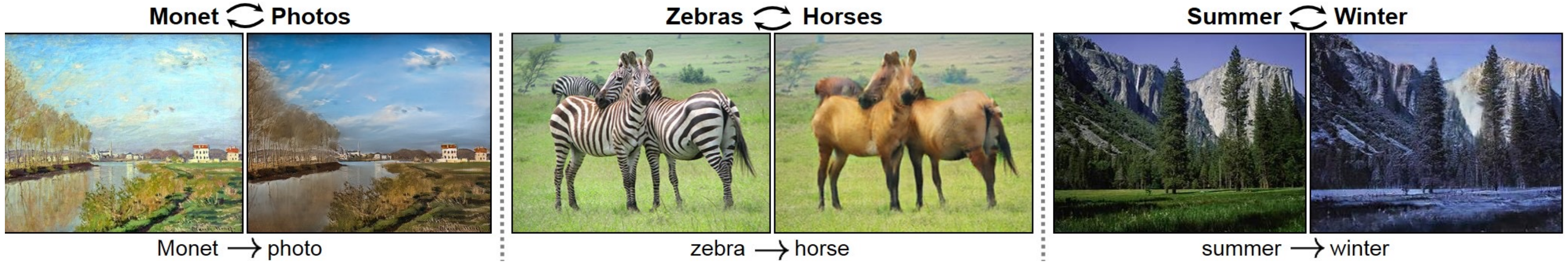


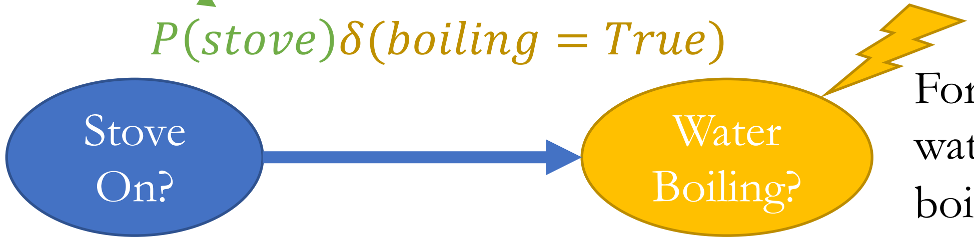

Image from CycleGAN paper: Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 2223-2232).

# Background: Causal probabilistic models *implicitly* encode the effect of **interventions**

The stove distribution is the same, i.e., **aligned**!

Implied factorization
$P(stove)P(boiling|stove)$


Intervened distribution
$P(stove)\delta(boiling = True)$




Force the water to boil

Both are valid factorizations.
But which factorization is ***causal***?

One idea: The factorization that changes the least under an intervention.


Implied factorization
$P(boiling)P(stove|boiling)$


Force the water to boil

Intervened distribution
$\delta(boiling = True)\tilde{P}(stove|boiling)$

The stove distribution is different under intervention. $\tilde{P}(stove|boiling) \equiv P(stove)$

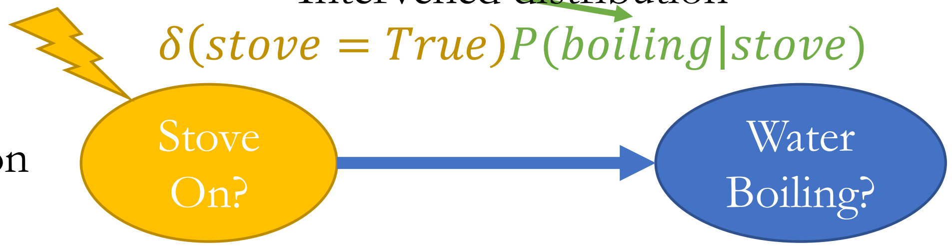
# Background: Causal probabilistic models *implicitly* encode the effect of **interventions**

The boiling distribution is the same, i.e., **aligned**!

Implied factorization
$P(stove)P(boiling|stove)$


Force stove on

Intervened distribution
$\delta(stove = True)P(boiling|stove)$




Both are valid factorizations.
But which factorization is ***causal***?

One idea: The factorization that changes the least under an intervention.

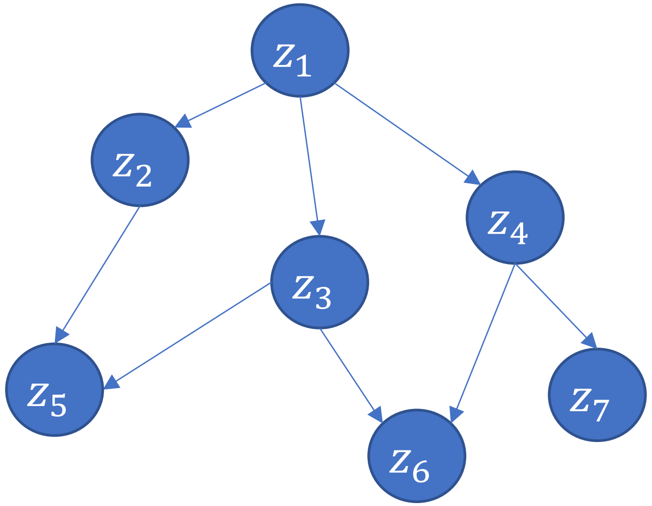

Implied factorization
$P(boiling)P(stove|boiling)$

Force stove on

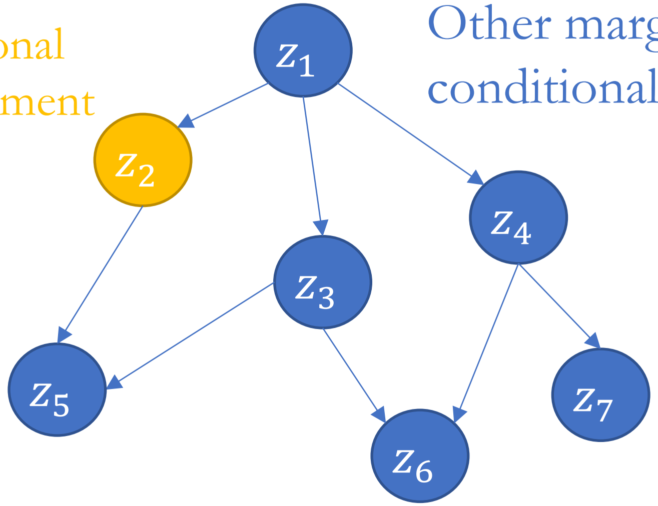

Intervened distribution
$\tilde{\mathrm{P}}(\mathrm{boiling})\delta(stove = True)$

The boiling distribution is different under intervention.

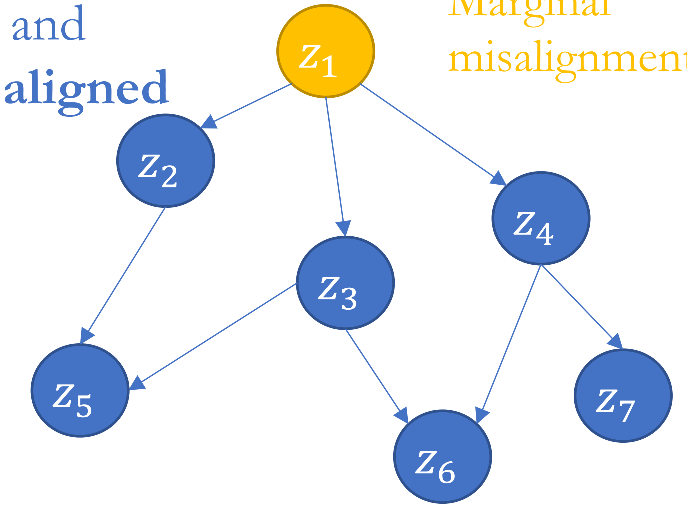
# Different domains can be viewed as ***unknown*** interventions in a ***latent*** causal space




Conditional misalignment




Other marginals and conditionals are **aligned**




Marginal misalignment

Latent space $\boldsymbol{z}|d_{=0} \sim \text{CausalModel}$ $\boldsymbol{z}|d_{=1} \sim \text{IntervenedCausalModel}$ $\boldsymbol{z}|d_{=2} \sim \text{IntervenedCausalModel}$

Observed space $\boldsymbol{x} = g^{-1}(\boldsymbol{z})$

Image adapted from GlobalWheat dataset images from https://wilds.stanford.edu/datasets/.

# Sparse intervention assumption => misalignment sparsity (Only a few conditionals are misaligned)

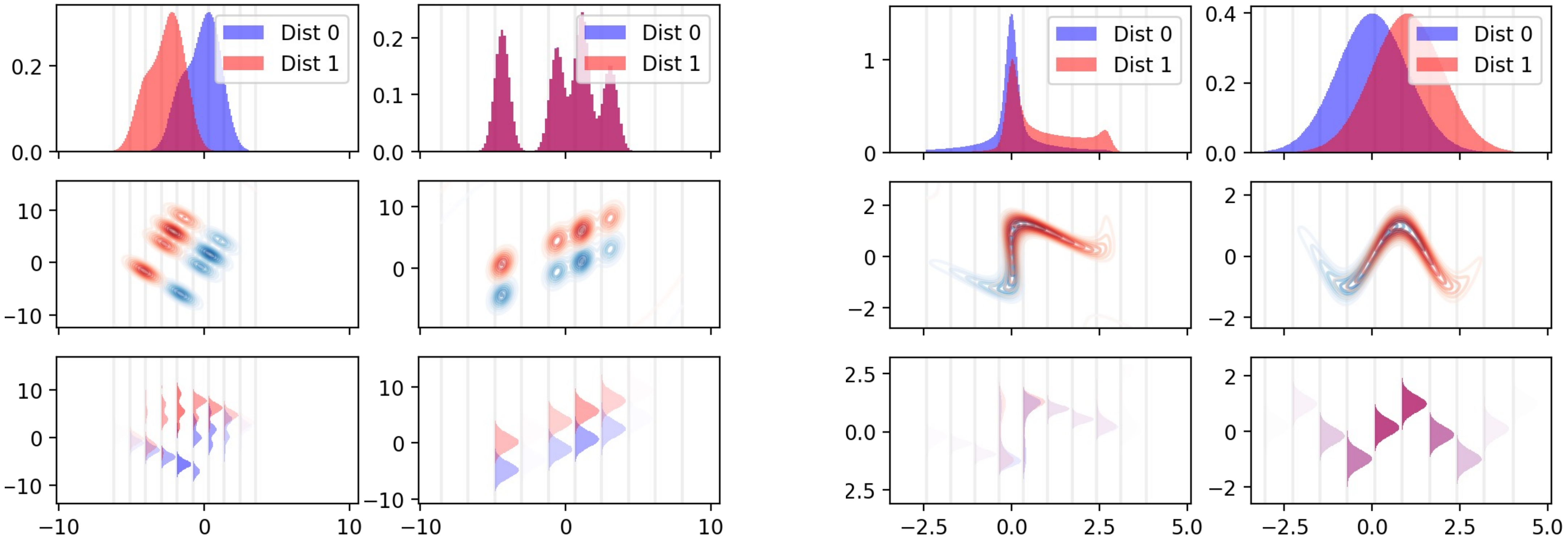
In 2D this means that either the marginal or conditionals are misaligned but **not both**.

Nothing is aligned (non-sparse)

Marginal $p(z_1)$ is aligned (sparse)

Nothing is aligned (non-sparse)

Conditional $p(z_2|z_1)$ is aligned (sparse)




Ongoing work with Prof. Murat Kocaoglu and Sean Kulinski. Vijay Prasad contributed to initial ideas.

# Future Vision: Alignment as (soft) constraints to combat **underspecification** in deep learning

**Underspecification Presents Challenges for Credibility in Modern Machine Learning**


**Alexander D'Amour*** ALEXDAMOUR@GOOGLE.COM
**Katherine Heller*** KHELLER@GOOGLE.COM
**Dan Moldovan*** MDAN@GOOGLE.COM
**Ben Adlam** ADLAM@GOOGLE.COM
**Babak Alipanahi** BABAKA@GOOGLE.COM
**Alex Beutel** ALEXBEUTEL@GOOGLE.COM
**Christina Chen** CHRISTINIUM@GOOGLE.COM
**Jonathan Deaton** JDEATON@GOOGLE.COM
**Jacob Eisenstein** JEISENSTEIN@GOOGLE.COM
**Matthew D. Hoffman** MHOFFMAN@GOOGLE.COM
**Farhad Hormozdiari** FHORMOZ@GOOGLE.COM
**Neil Houlsby** NEILHOULSBY@GOOGLE.COM
**Shaobo Hou** SHAOBOHOU@GOOGLE.COM
**Ghassen Jerfel** GHASSEN@GOOGLE.COM
**Alan Karthikesalingam** ALANKARTHI@GOOGLE.COM
**Mario Lucic** LUCIC@GOOGLE.COM
**Yian Ma** YIANMA@UCSD.EDU
**Cory McLean** CYM@GOOGLE.COM
**Diana Mincu** DMINCU@GOOGLE.COM
**Akinori Mitani** AMITANI@GOOGLE.COM
**Andrea Montanari** MONTANARI@STANFORD.EDU
**Zachary Nado** ZNADO@GOOGLE.COM
**Vivek Natarajan** NATVIV@GOOGLE.COM
**Christopher Nielson†** CHRISTOPHER.NIELSON@VA.GOV
**Thomas F. Osborne†** THOMAS.OSBORNE@VA.GOV
**Rajiv Raman** DRRRN@SNMAIL.ORG
**Kim Ramasamy** KIM@ARAVIND.ORG
**Rory Sayres** SAYRES@GOOGLE.COM
**Jessica Schrouff** SCHROUFF@GOOGLE.COM
**Martin Seneviratne** MARTSEN@GOOGLE.COM
**Shannon Sequeira** SHNNN@GOOGLE.COM
**Harini Suresh** HSURESH@MIT.EDU
**Victor Veitch** VICTORVEITCH@GOOGLE.COM
**Max Vladymyrov** MXV@GOOGLE.COM
**Xuezhi Wang** XUEZHIW@GOOGLE.COM
**Kellie Webster** WEBSTERK@GOOGLE.COM
**Steve Yadlowsky** YADLOWSKY@GOOGLE.COM
**Taedong Yun** TEDYUN@GOOGLE.COM
**Xiaohua Zhai** XZHAI@GOOGLE.COM
**D. Sculley** DSCULLEY@GOOGLE.COM

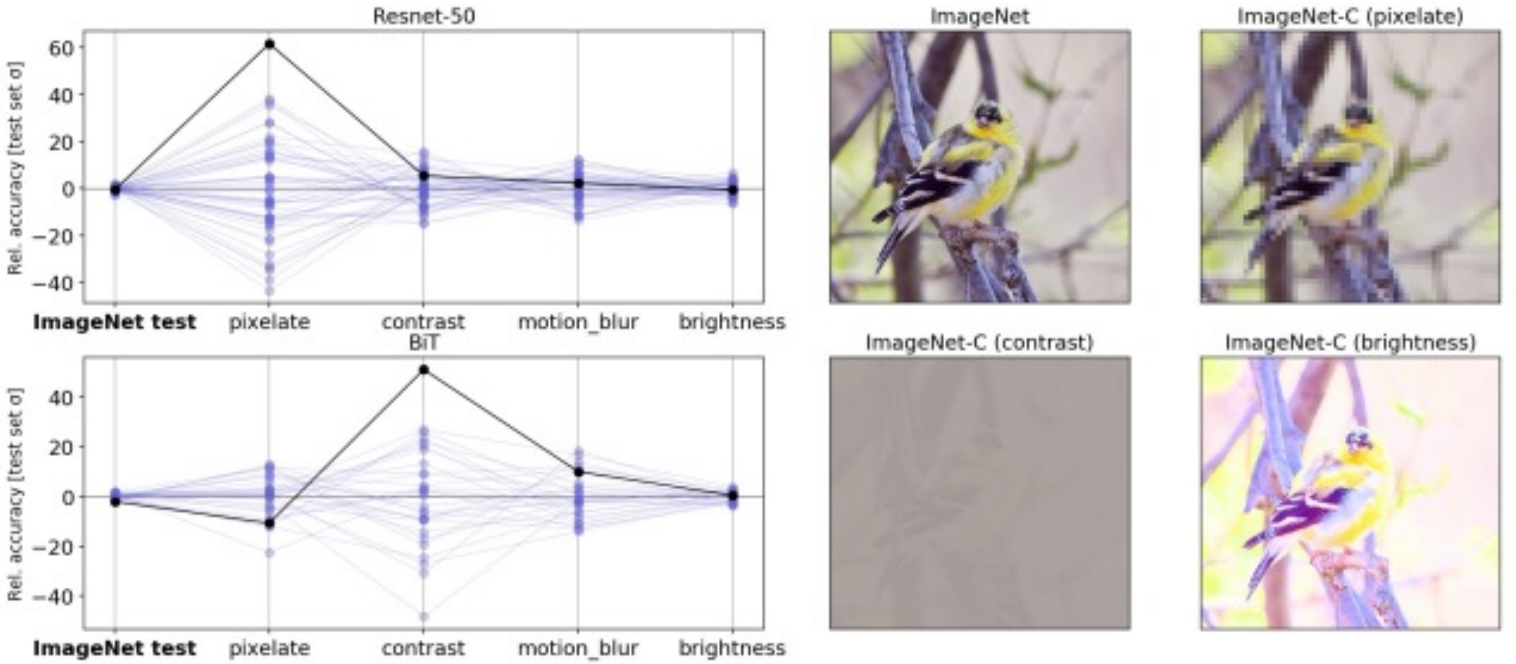




Figure 4: **Image classification model performance on stress tests is sensitive to random initialization in ways that are not apparent in iid evaluation. (Top Left)**


D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., ... & Sculley, D. (2020). Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395*.

# Future research opportunities in all areas of distribution alignment

- Alignment concepts
  - Conditional alignment in particular
- Alignment measures
  - More application-agnostic measures
  - Rigorous evaluation protocols
- Alignment algorithms
  - Beyond adversarial
  - More stable optimization
- Alignment applications
  - What robustness can we achieve?
  - Can we make this more general?