香 港 大 學
THE UNIVERSITY OF HONG KONG

ICML
International Conference On Machine Learning

Improving Group Robustness on Spurious Correlation Requires Preciser Group Inference

Yujin Han, Difan Zou

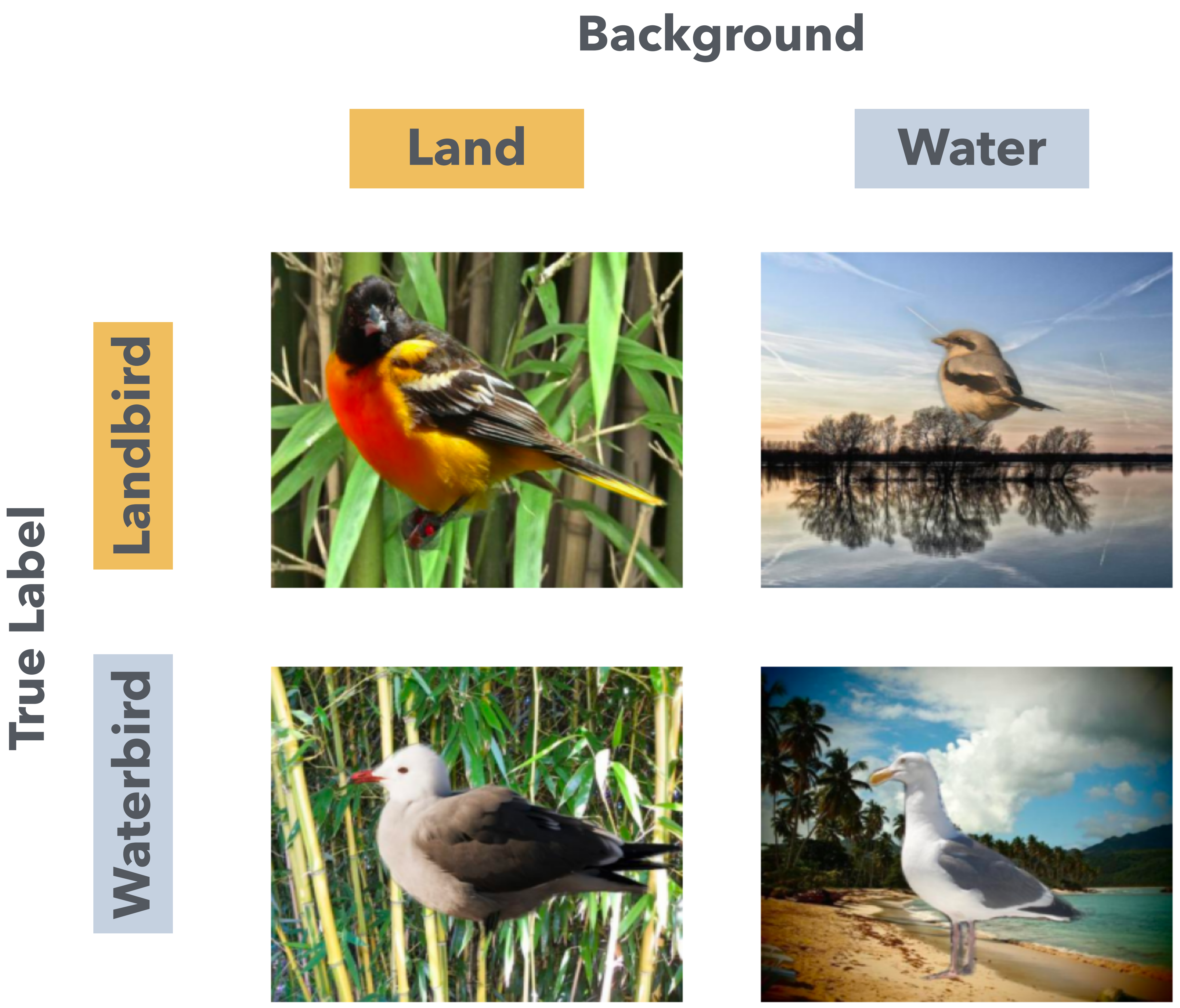
Empirical Risk Minimization (ERM)

$$J_{\mathrm{ERM}}(\theta) = \frac{1}{n} \sum_{i=1}^{n} l(\theta; \mathbf{x}_i, y_i)$$

Minimal average error over the training set

Problem: Low Worst-Group Performance

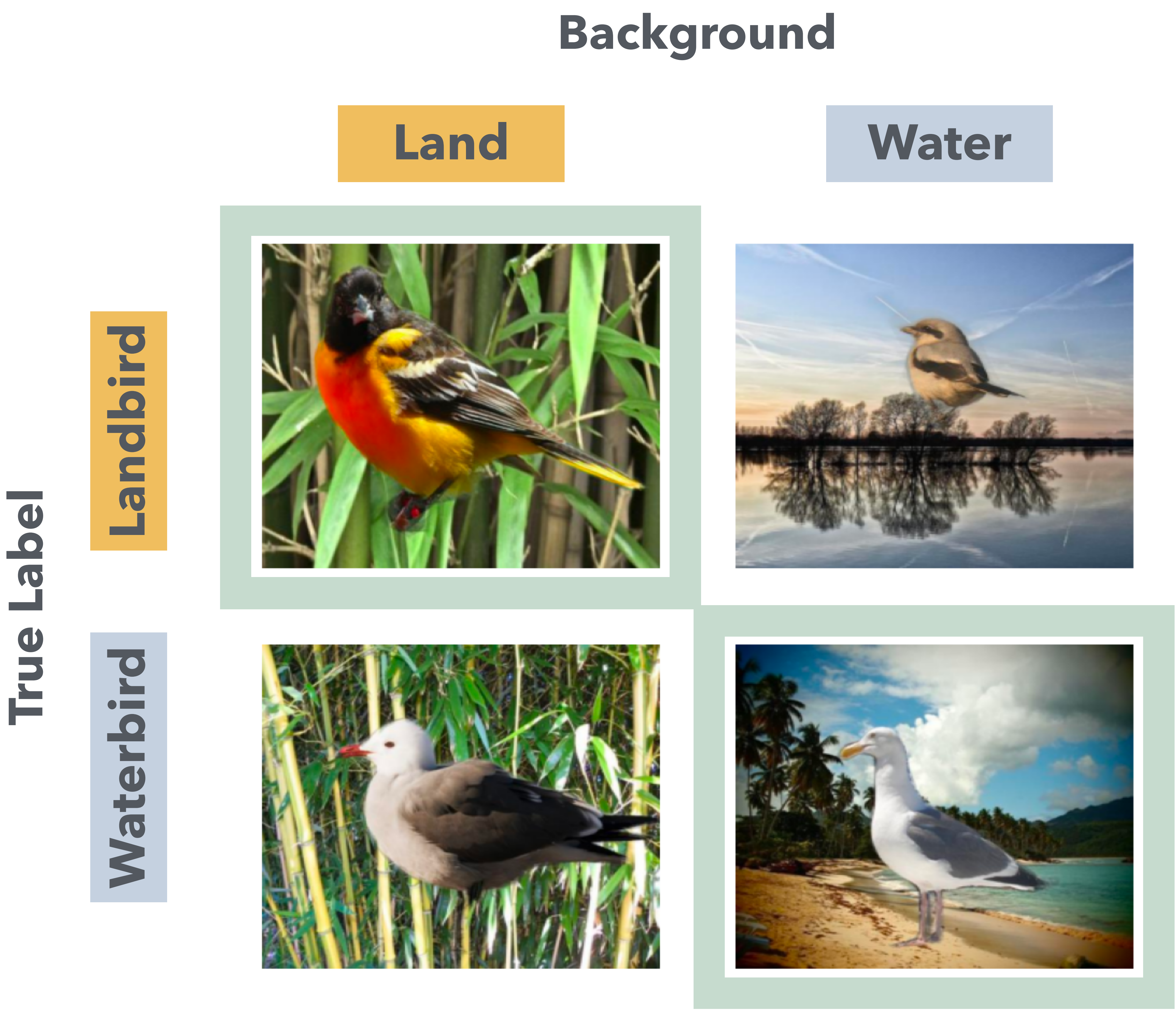
Wildbird image classification (Wah et al., ‘11; Sagawa et al., ‘20)



97.3% average test accuracy

Problem: Low Worst-Group Performance

Wildbird image classification (Wah et al., '11; Sagawa et al., '20)



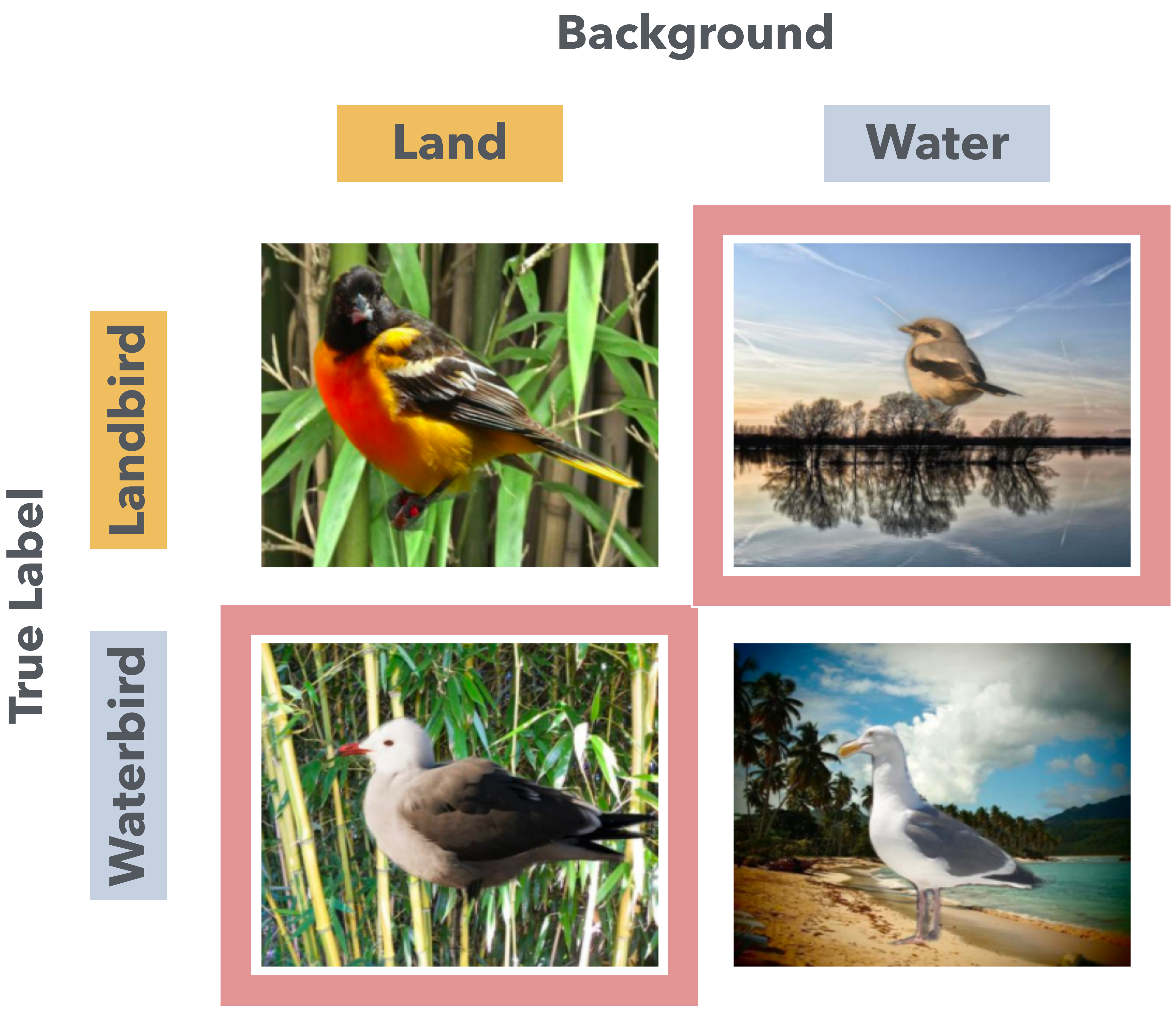
Majority Group

95.0% Sample Size

almost 100% average test accuracy

Problem: Low Worst-Group Performance

Wildbird image classification (Wah et al., ‘11; Sagawa et al., ‘20)



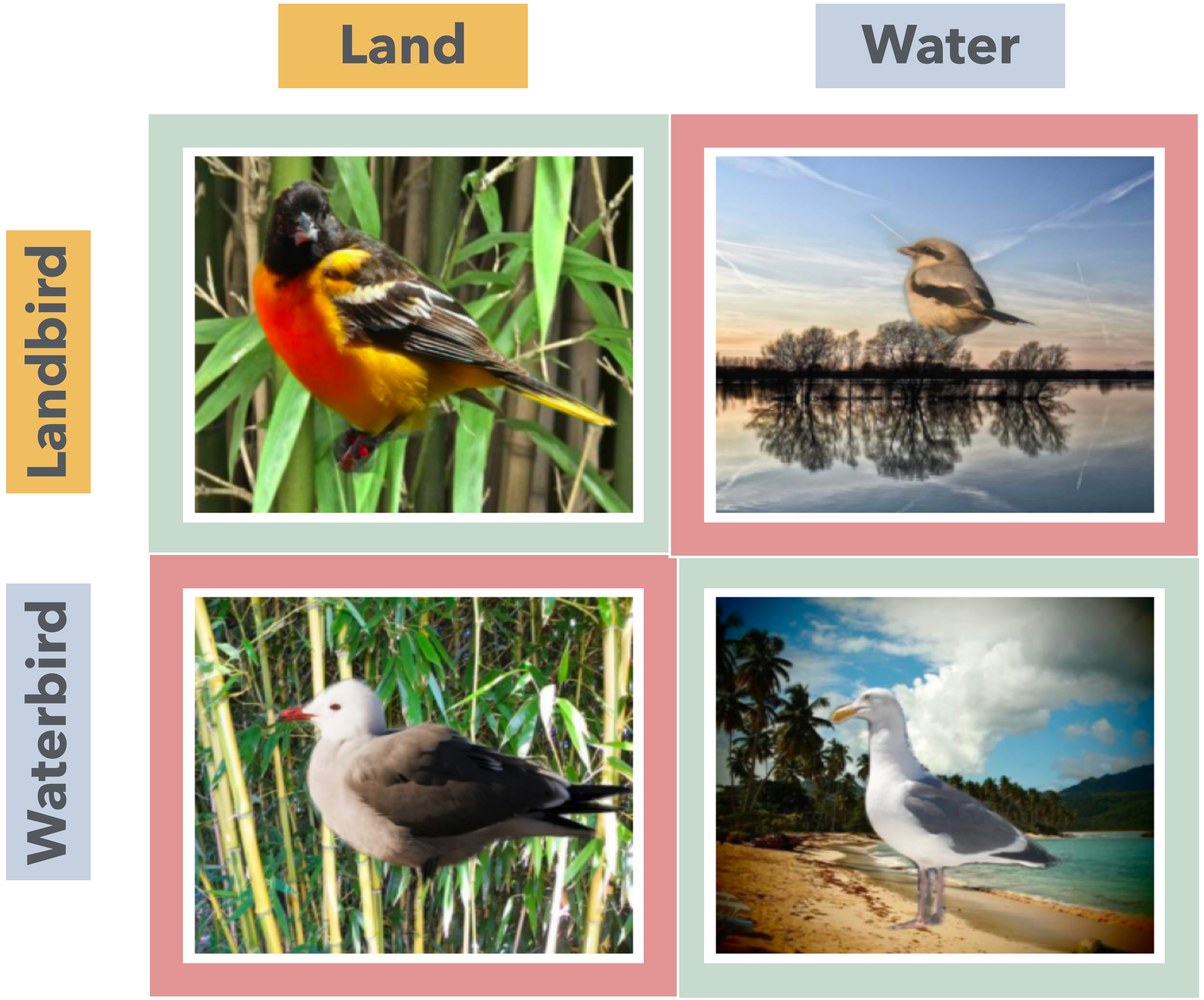
Minority Group

5.0% Sample Size

62.6% average test accuracy

Problem: Low Worst-Group Performance

Wildbird image classification (Wah et al., ‘11; Sagawa et al., ‘20)



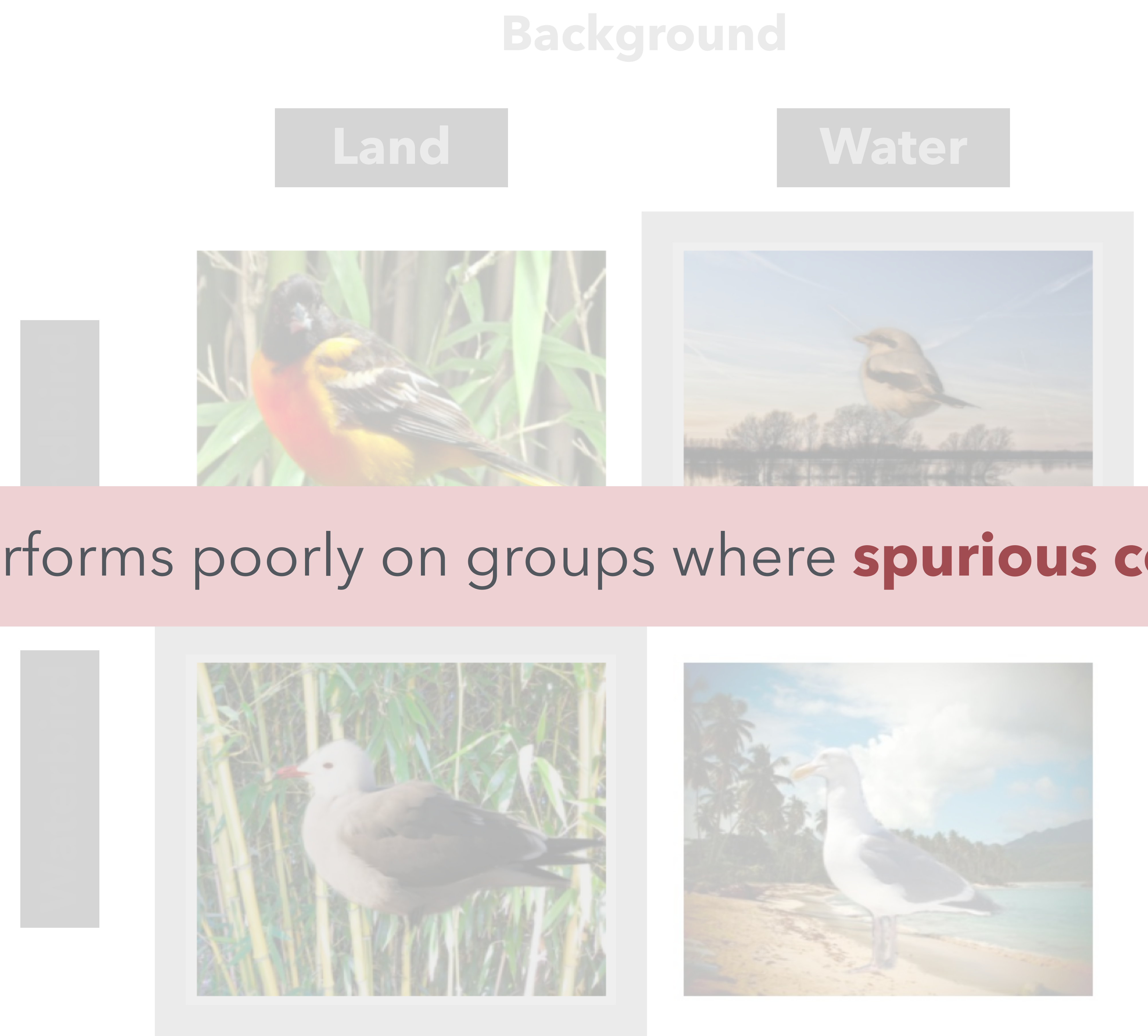
Minority Group

5.0% Sample Size

62.6% average test accuracy

Problem: Low Worst-Group Performance

Wildbird image classification (Wah et al., ‘11; Sagawa et al., ‘20)

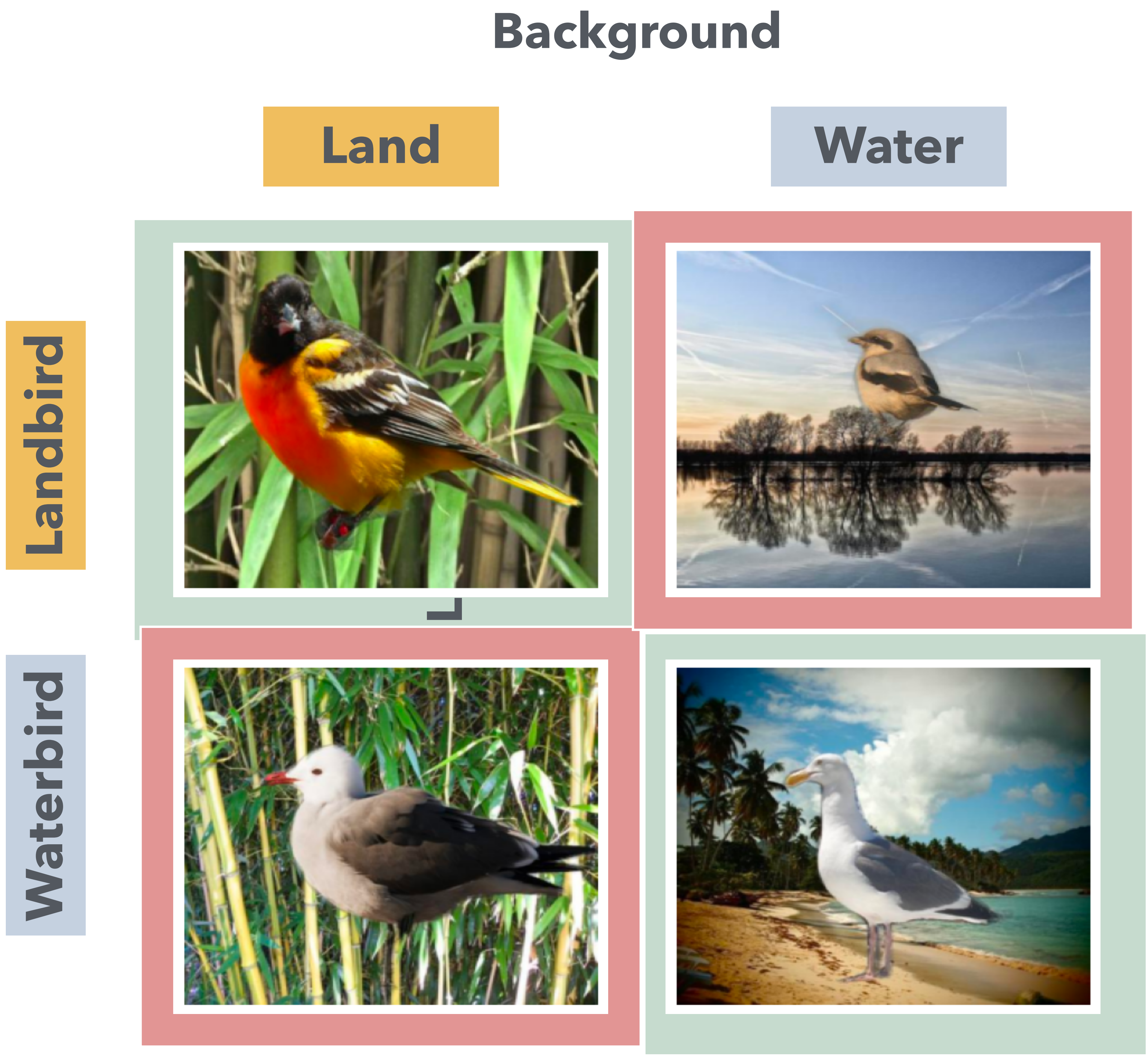


Standard ERM performs poorly on groups where spurious correlations do not hold.

62.6% average test accuracy

Problem: Low Worst-Group Performance

Spurious Correlation

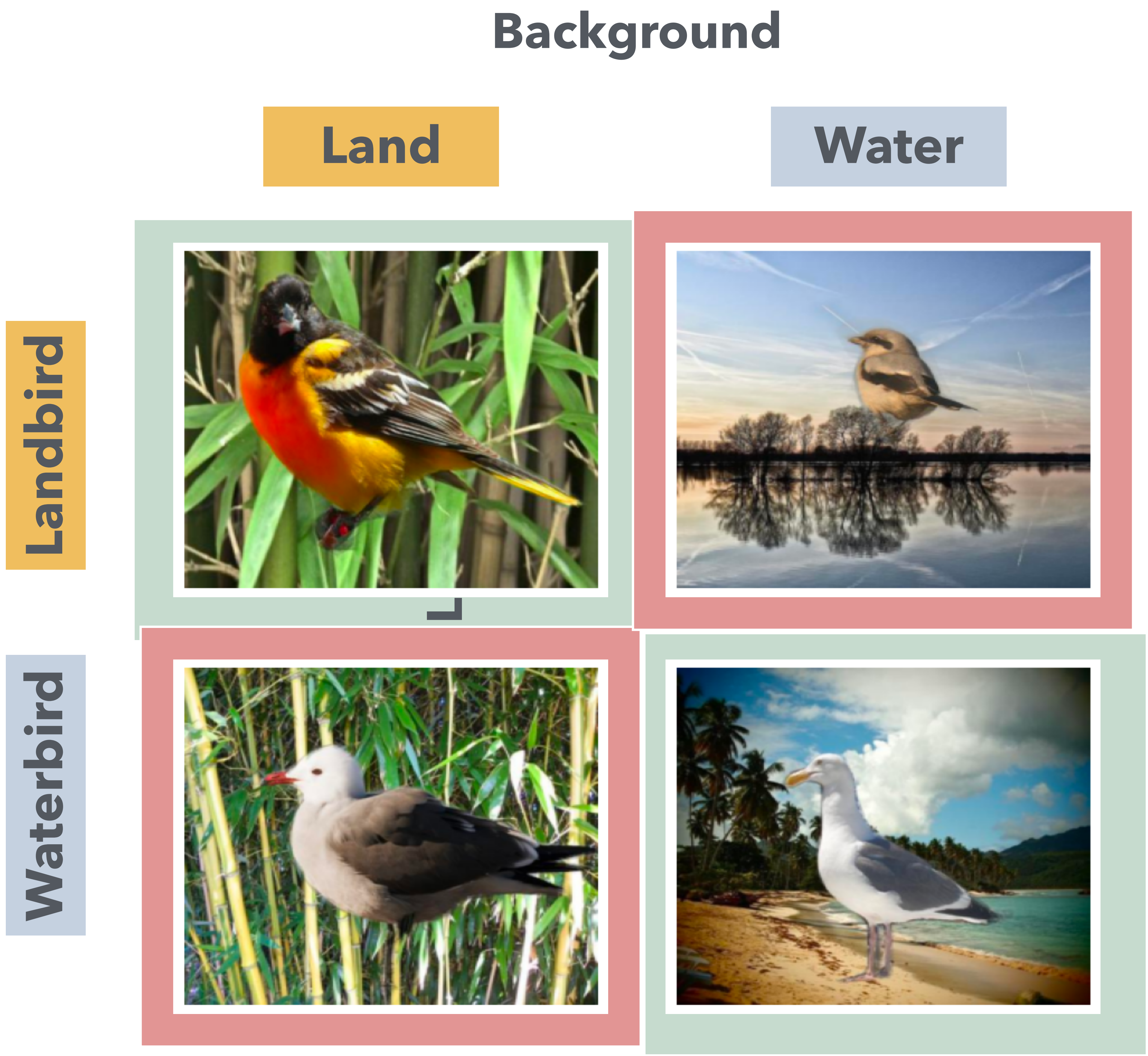


Invariant attribute: **Bird Type** ✓

Spurious attribute: **Background** ✗

Problem: Low Worst-Group Performance

Spurious Correlation



Imbalanced Data

Simplicity Bias

...

Prior Work: Requiring Group Label

Group Reweighting: GroupDRO

ERM

$$\min_{\theta} \mathbb{E}_{(\mathbf{x},y)\sim p^{tr}}[l(\theta; \mathbf{x}, y)]$$

minimal **average** error over the training set

GroupDRO

$$\min_{\theta} \{ \sup_{g\in\mathcal{G}} \mathbb{E}_{(\mathbf{x},y)\sim p^{g}}[l(\theta; \mathbf{x}, y)] \}$$

minimal **worst-case** error over the training set

Prior Work: Requiring Group Label

Group Reweighting: GroupDRO

ERM

$$\min_{\theta} \mathbb{E}_{(\mathbf{x},y)\sim p^{tr}}[l(\theta; \mathbf{x}, y)]$$

minimal **average** error over the training set

GroupDRO

$$\min_{\theta} \{ \sup_{\boxed{g \in \mathcal{G}}} \mathbb{E}_{(\mathbf{x},y)\sim \boxed{p^{g}}}[l(\theta; \mathbf{x}, y)] \}$$

minimal **worst-case** error over the training set

Prior Work: Requiring Group Label

Group Reweighting: GroupDRO

ERM

minimal **average** error over the training set

GroupDRO

minimal **worst-case** error over the training set

Group labels are **expensive** and **labor-intensive**

Prior Work: Inferring Group Label

ERM-Based: Just Train Twice (JTT)

Stage 1: Inferring group labels

1. Train identification model f_{id} via ERM

Prior Work: Inferring Group Label

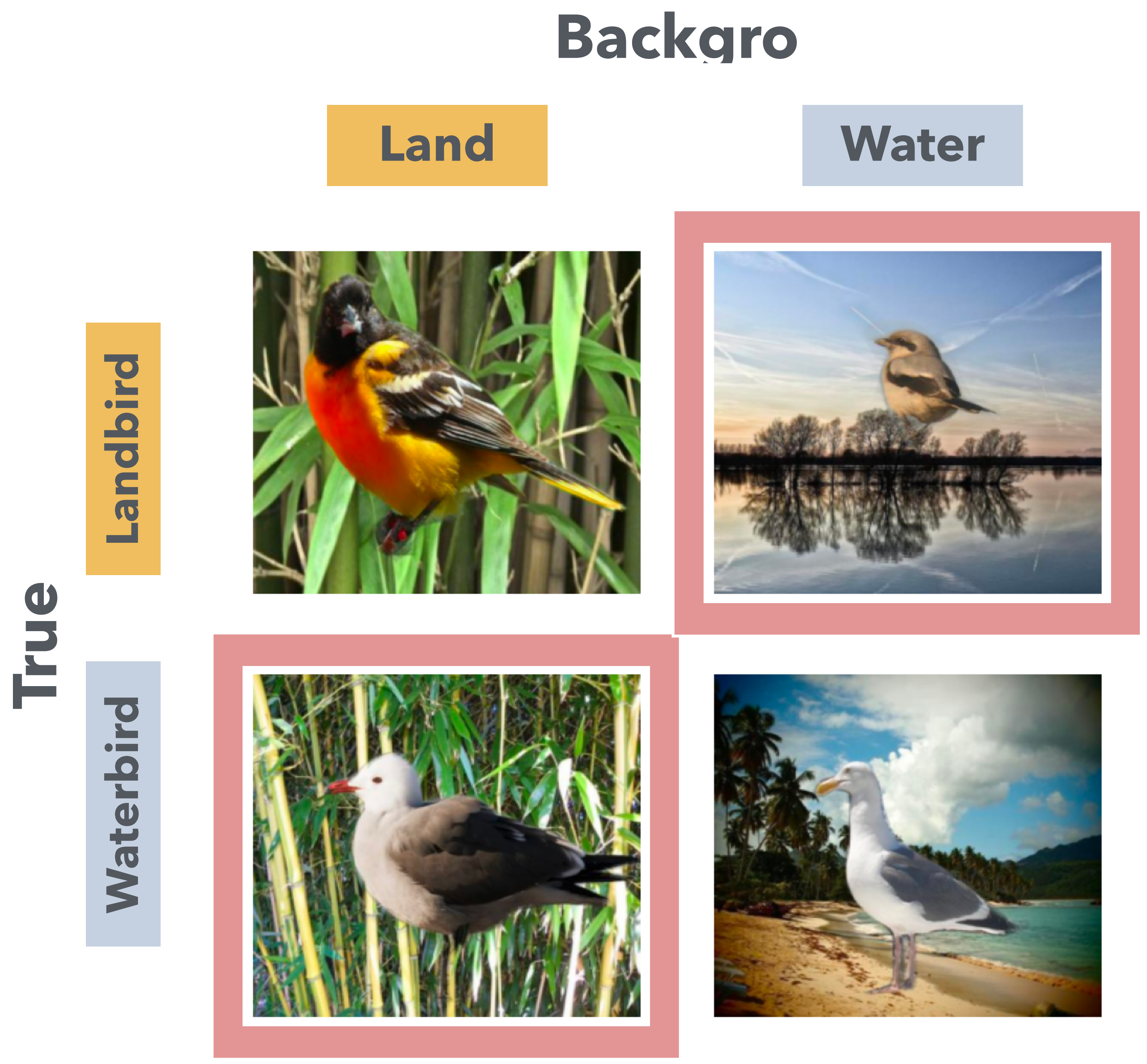
ERM-Based: Just Train Twice (JTT)

Stage 1: Inferring group labels

1. Train identification model f_{id} via ERM
2. Compute **error set E** of misclassified training examples

$$E = \{(\mathbf{x}, y) \mid f_{id}(\mathbf{x}) \neq y\}$$

E is **minority groups** where spurious correlation doesn't hold



Prior Work: Inferring Group Label

ERM-Based: Just Train Twice (JTT)

Stage 1: Inferring group labels

1. Train identification model f_{id} via ERM
2. Compute **error set E** of misclassified training examples

$$E = \{(\mathbf{x}, y) \mid f_{id}(\mathbf{x}) \neq y\}$$

E is **minority groups** where spurious correlation doesn't hold

Stage 2: Invariant learning

3. Upweight identified examples

Prior Work: Inferring Group Label

ERM-Based: Just Train Twice (JTT)

Stage 1: Inferring group labels

1. Train identification model f_{id} via ERM
2. Compute **error set E** of misclassified training examples

$$E = \{(\mathbf{x}, y) \mid f_{id}(\mathbf{x}) \neq y\}$$

E is **minority groups** where spurious correlation doesn't hold

Stage 2: Invariant learning

3. Upweight identified examples
4. Train f_{robust} via ERM on upsampled data

Prior Work: Inferring Group Label

ERM-Based: Just Train Twice (JTT)

Stage 1: Inferring group labels

1. Train identification model f_{id} via ERM
2. C

$$E = \{(\mathbf{x}, y) | f_{id}(\mathbf{x}) \neq y\}$$

E is **minority groups** where spurious correlation doesn't hold

Stage 2: Invariant learning

3. Upweight identified examples

Have **performance gaps** compared to group annotation utilized methods

Prior Work: Inferring Group Label

EI-Based: Environment Inference for Invariant Learning (EIIL)

Stage 1: Inferring group labels

1. Infer pseudo group label via violating invariant principles

Regularization term of IRM:

$$\max C^{EI}(\phi, q) = \max \|\nabla R(\phi, q)\|$$

where $$R^e(\phi, q) = \frac{1}{N} \sum_i q_i(e) l(\phi(\mathbf{x}_i), y_i)$$

Stage 2: Invariant learning

3. Train f_{robust} via GroupDRO

Prior Work: Inferring Group Label

EI-Based: Environment Inference for Invariant Learning (EIIL)

Stage 1: Inferring group labels

1. Infer pseudo group label via violating invariant principles

Reg

where $R^e(\phi, q) = \frac{1}{N}\sum_i q_i(e) l(\phi(\mathbf{x}_i), y_i)$

Stage 2: Invariant learning

3. Train f_{robust} via GroupDRO

Have **performance gaps** compared to group annotation utilized methods

Prior Work: Inferring Group Label

Human Prior

DISC: Discover and Cure

Concept Bank: human-interpretable concepts

Concept category	Concepts
Color	[*blackness, blueness, greenness, redness, whiteness*]
Texture	[*concrete, granite, leather, laminate, metal, blotchy, blurriness, stripes, polka dots, knitted, cracked, frilly, waffled, scaly, lacelike, grooved, stratified, gauzy, marbled, flecked, stained, braided, matted, meshed, cobwebbed, spiralled, dotted, crosshatched, wrinkled, woven, potholed, crystalline, paisley, veined, fibrous, studded, bubbly, pleated, grid, perforated, porous, interlaced, smeared, honeycombed, sprinkled, chequered, lined, banded, bumpy, zigzagged, swirly, pitted, freckled*]
Nature	[*bamboo, beach, bridge, bush, canopy, earth, field, flower, flowerpot, fluorescent, forest, grass, ground, harbor, hill, lake, mountain, muzzle, palm, path, plant, river, sand, sea, snow, tree, water*]
City	[*awning, base, bench, building, earth, fence, field, ground, house, manhole, path, snow, streets*]
Household	[*air-conditioner, apron, armchair, back-pillow, balcony, bannister, bathrooms, bathtub, bed, bedclothes, bedrooms, cabinet, carpet, ceiling, chair, chandelier, chest-of-drawers, countertop, curtain, cushion, desk, dining-rooms, door, door-frame, double-door, drawer, drinking-glass, exhaust-hood, figurine, fireplace, floor, flower, flowerpot, fluorescent, ground, handle, handle-bar, headboard, headlight, house, jar, lamp, light, microwave, mirror, ottoman, oven, pillow, plate, refrigerator, sofa, stairs, toilet*]
Others	[*bird, cat, cow, dog, horse, mouse, paw, arm, back, body, ear, eye, eyebrow, female-face, leg, male-face, foot, hair, hand, head, inside-arm, knob, mouth, neck, nose, outside-arm, ashcan, airplane, bag, bus, beak, bicycle, blind, board, book, bookcase, bottle, bowl, box, brick, basket, bucket, bumper, can, candlestick, cap, car, cardboard, ceramic, chain-wheel, chimney, clock, coach, coffee-table, column, computer, counter, cup, desk, engine, fabric, fan, faucet, flag, floor, food, foot-board, frame, glass, keyboard, lid, loudspeaker, minibike, motorbike, napkin, pack, painted, painting, pane, paper, pedestal, person, pillar, pipe*]

ZIN: auxiliary information z for environmental INference

Auxiliary Information

Built year ; Age; Location; Blond Hair, Eyeglasses …

Prior Work: Inferring Group Label

Human Prior

DISC: Discover and Cure

Concept Bank: human-interpretable concepts

Concept category	Concepts
Color	[blackness, blueness, greenness, redness, whiteness]
Texture	
Nature	hill, lake, mountain, muzzle, palm, path, plant, river, sand, sea, snow, tree, water]
City	[awning, base, bench, building, earth, fence, field, ground, house, manhole, path, snow, streets]
Household	[air-conditioner, apron, armchair, back-pillow, balcony, bannister, bathrooms, bathtub, bed, bedclothes, bedrooms, cabinet, carpet, ceiling, chair, chandelier, chest-of-drawers, countertop, curtain, cushion, desk, dining-rooms, door, door-frame, double-door, drawer, drinking-glass, exhaust-hood, figurine, fireplace, floor, flower, flowerpot, fluorescent, ground, handle, handle-bar, headboard, headlight, house, jar, lamp, light, microwave, mirror, ottoman, oven, pillow, plate, refrigerator, sofa, stairs, toilet]
Others	[bird, cat, cow, dog, horse, mouse, paw, arm, back, body, ear, eye, eyebrow, female-face, leg, male-face, foot, hair, hand, head, inside-arm, knob, mouth, neck, nose, outside-arm, ashcan, airplane, bag, bus, beak, bicycle, blind, board, book, bookcase, bottle, bowl, box, brick, basket, bucket, bumper, can, candlestick, cap, car, cardboard, ceramic, chain-wheel, chimney, clock, coach, coffee-table, column, computer, counter, cup, desk, engine, fabric, fan, faucet, flag, floor, food, foot-board, frame, glass, keyboard, lid, loudspeaker, minibike, motorbike, napkin, pack, painted, painting, pane, paper, pedestal, person, pillar, pipe]

ZIN: auxiliary information z for environmental INference

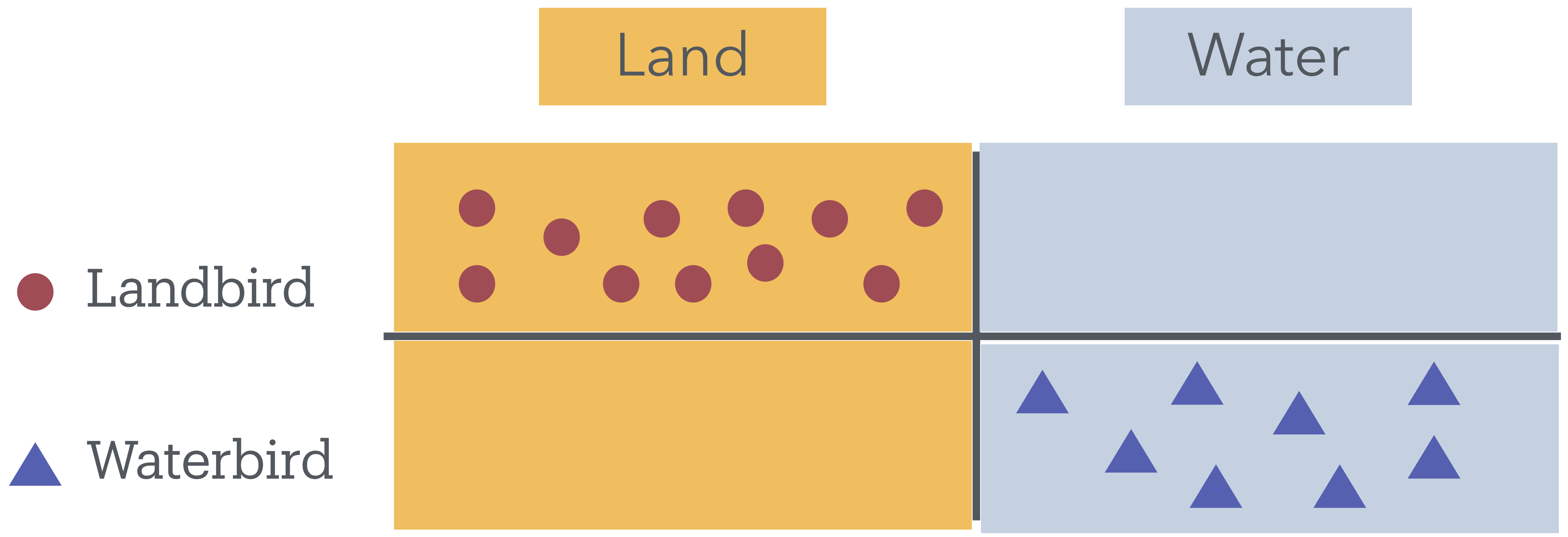
Auxiliary Information

Eyeglasses …

Not be applicable when prior information is **unavailable**

Goal: Inferring Preciser Group Label

GIC: Group **I**nference via data **C**omparison



Training Data

Group Distribution: $g = \{(\bullet, \blacksquare), (\bullet, \blacksquare), (\blacktriangle, \blacksquare), (\blacktriangle, \blacksquare)\} = (10,0,0,8)$

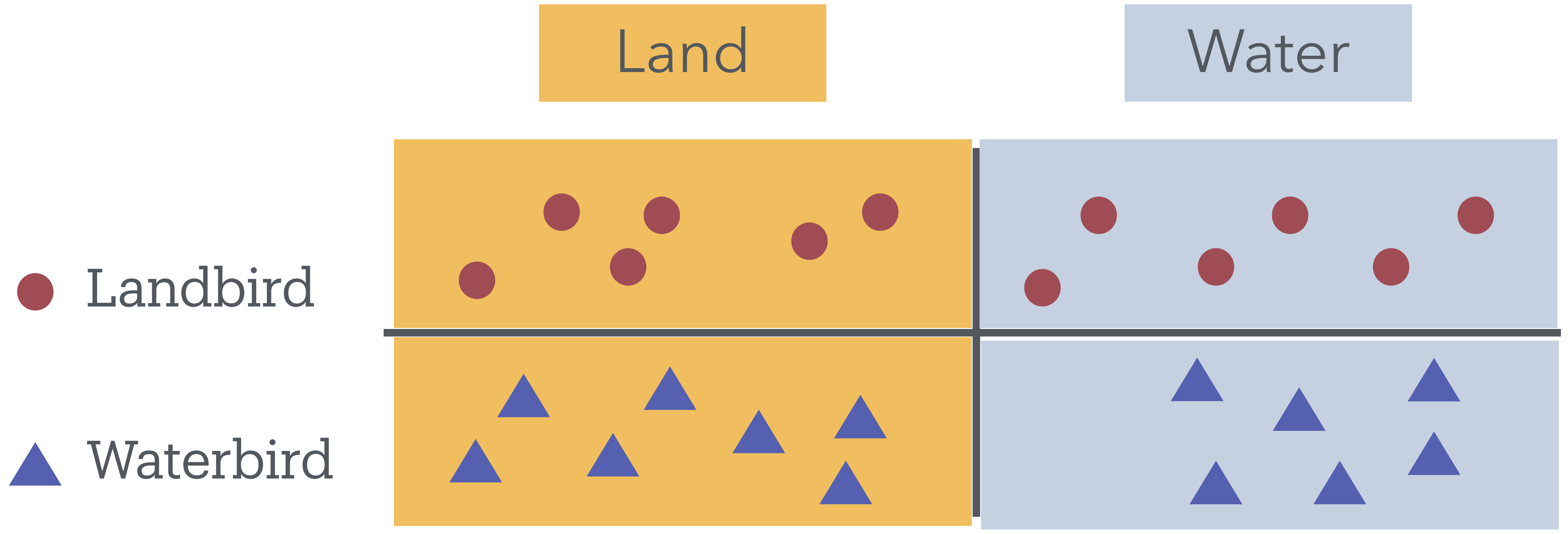
Spurious attribute label and True label: $y_s^{tr} = y^{tr}$ 100%

Invariant attribute label and True label: $y_{in}^{tr} = y^{tr}$ 100%

Serious Spurious Correlation

Goal: Inferring Preciser Group Label

GIC: **G**roup **I**nference via data **C**omparison



Comparison Data

Group Distribution: $g = \{(●, ■), (●, ■), (▲, ■), (▲, ■)\} = (6,6,6,6)$

Spurious attribute label and True label: $y_s^c \neq y^c$ → 50%

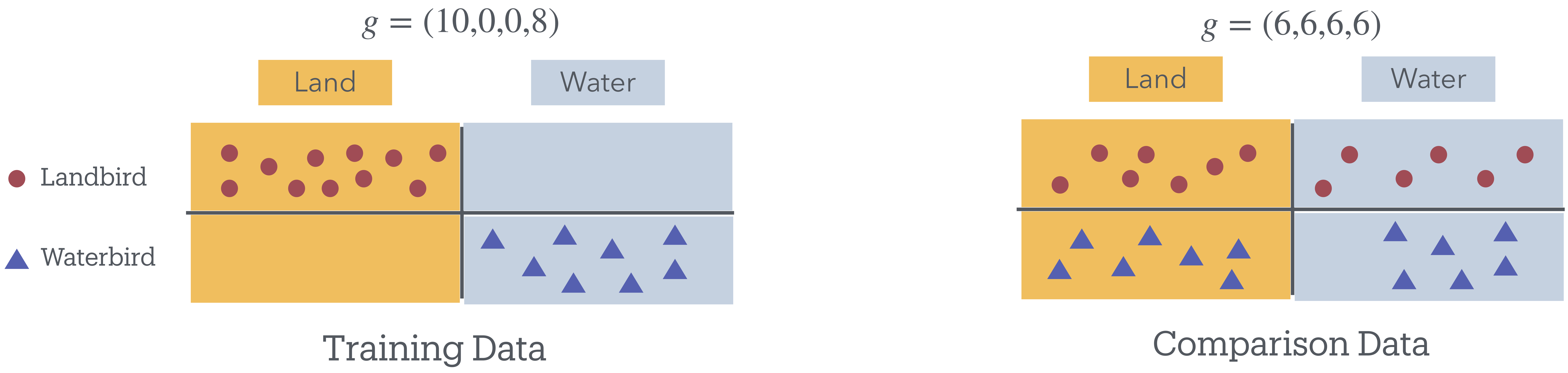
Invariant attribute label and True label: $y_{in}^c = y^c$ → 100%

Slight Spurious Correlation

Goal: Inferring Preciser Group Label

GIC: Group **I**nference via data **C**omparison

Spurious Correlation varies in Datasets with (slight) **different** group distribution



Goal: Inferring Preciser Group Label

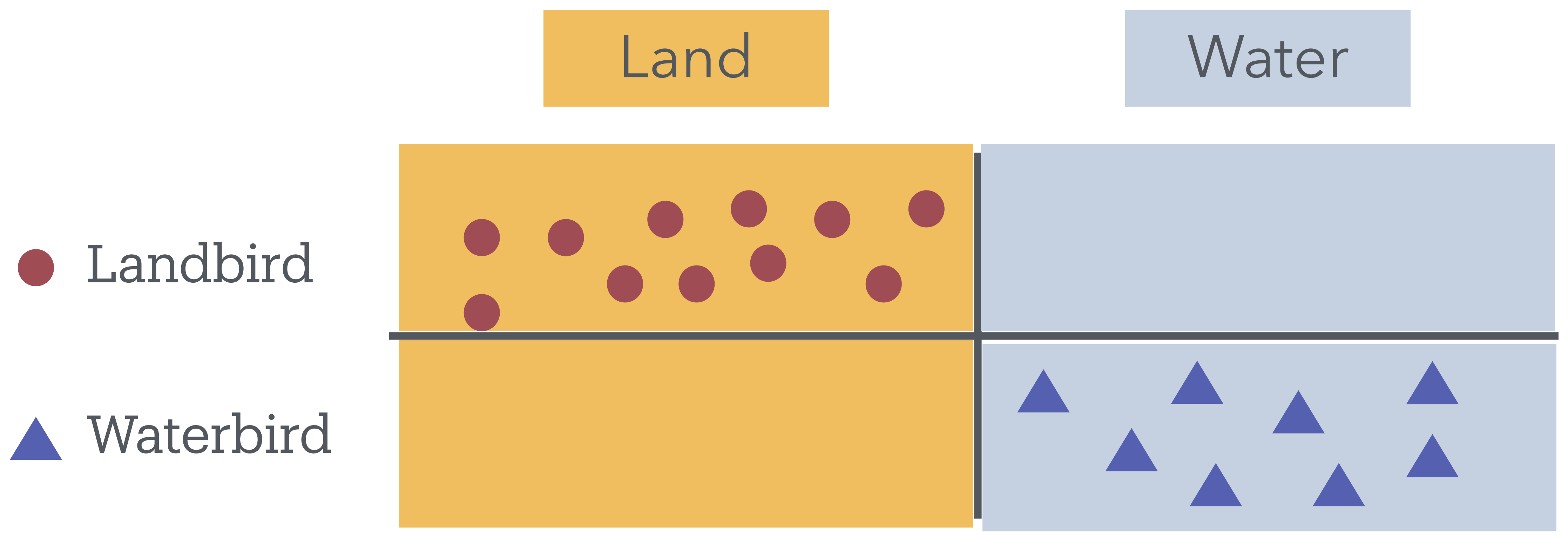
GIC: **G**roup **I**nference via data **C**omparison

Spurious **Correlation** varies in Datasets with (slight) different group distribution

Term 1: **Correlation Term**

Encourage the high correlation between y and y_s in the training set.

$$\max_{\mathbf{w}} I(y^{tr}; \hat{y}^{tr}_{s,\mathbf{w}})$$



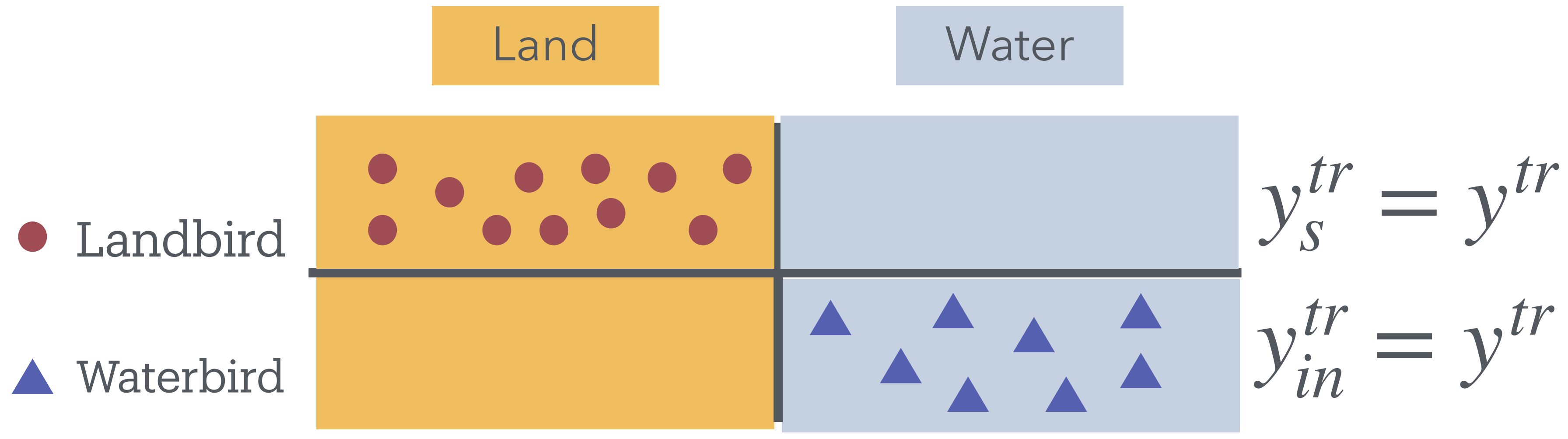
Training Data

Spurious attribute label and True label: $y_s = y$

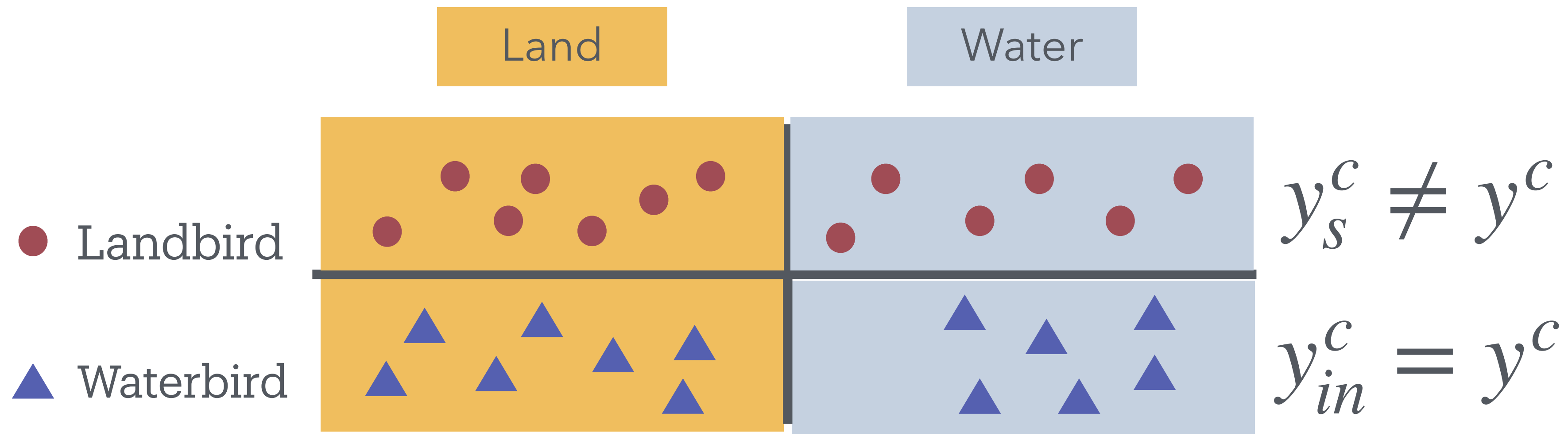
Goal: Inferring Preciser Group Label

GIC: **G**roup **I**nference via data **C**omparison

Spurious Correlation varies in Datasets with (slight) different group distribution



Training Data



Comparison Data

Term 2: **Spurious Term**

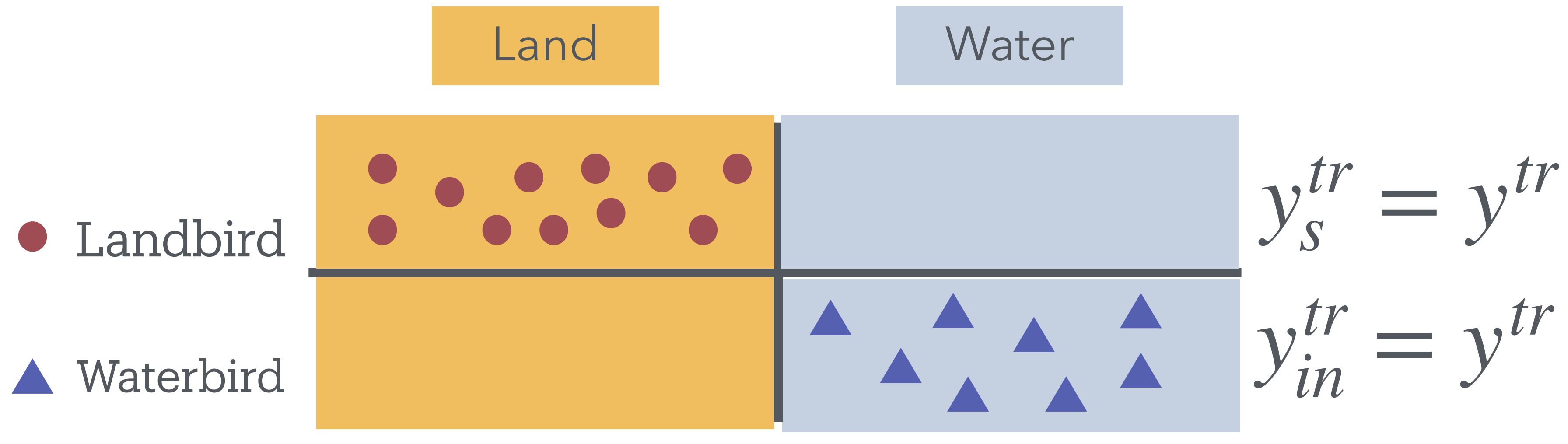
Variability in this correlation between datasets with different group distributions:

$$\max_{\mathbf{w}} \mathrm{KL}(\mathbb{P}(y^{tr} \mid \hat{y}_{s,\mathbf{w}}^{tr}) \mid\mid \mathbb{P}(y^{c} \mid \hat{y}_{s,\mathbf{w}}^{c}))$$

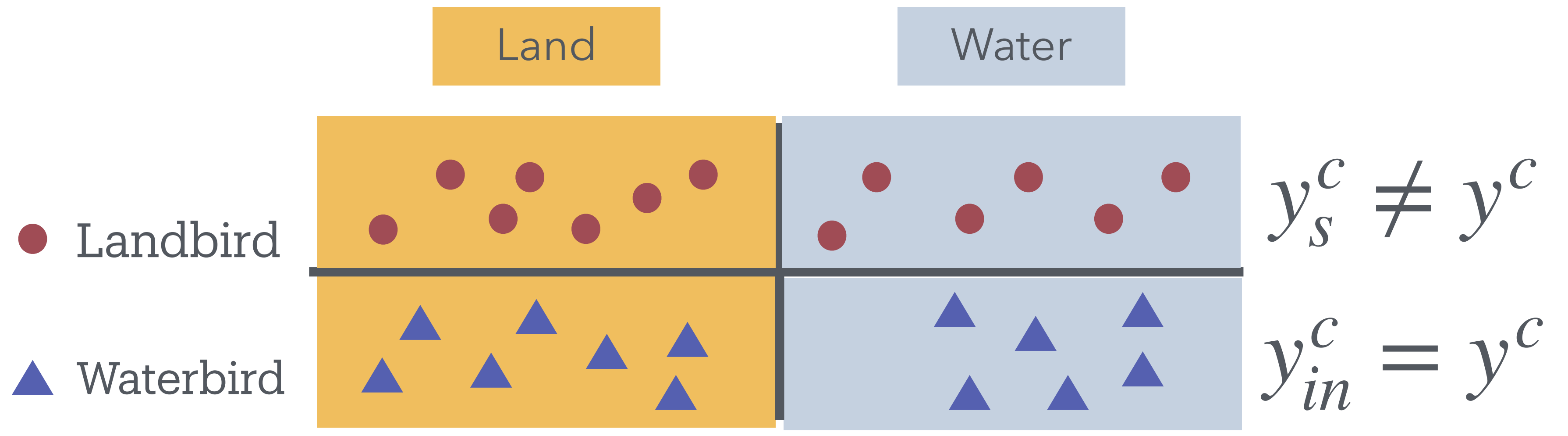
Goal: Inferring Preciser Group Label

GIC: Group Inference via data Comparison

Spurious Correlation varies in Datasets with (slight) different group distribution



Training Data



Comparison Data

Term 2: **Spurious Term**

$$\max_{\mathbf{w}} \mathrm{KL}(\mathbb{P}(y^{tr} \,|\, \hat{y}^{tr}_{s,\mathbf{w}}) \,||\, \mathbb{P}(y^{c} \,|\, \hat{y}^{c}_{s,\mathbf{w}}))$$

Volating the invariant learning principle

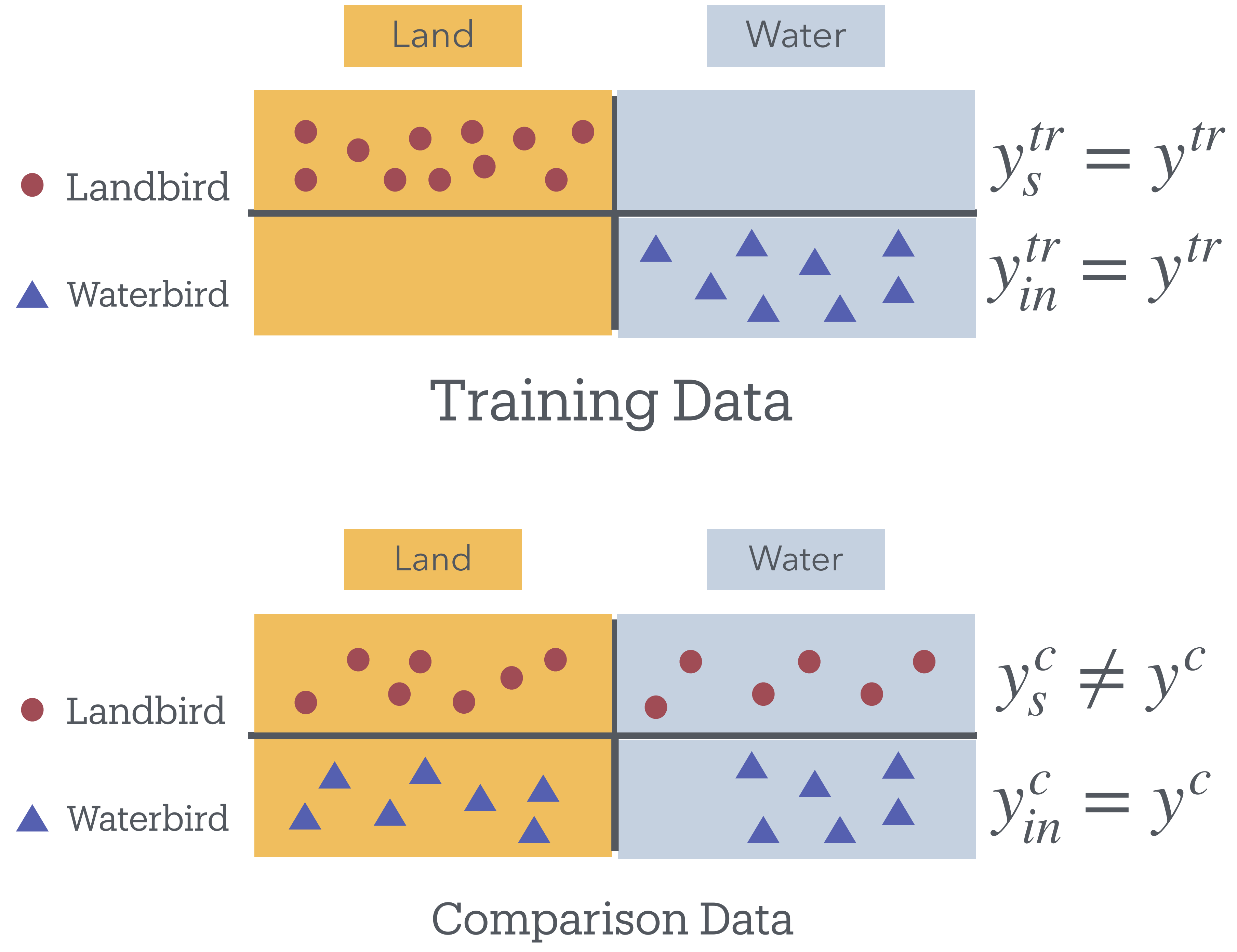
High Correlation: **Invariant attribute**

$$\mathrm{KL}(\mathbb{P}(y^{tr} \,|\, \hat{y}^{tr}_{s,\mathbf{w}}) \,||\, \mathbb{P}(y^{c} \,|\, \hat{y}^{c}_{s,\mathbf{w}})) = 0$$

Goal: Inferring Preciser Group Label

GIC: **G**roup **I**nference via data **C**omparison

Spurious Correlation varies in Datasets with (slight) different group distribution



Term 2: **Spurious Term**

$$\max_{\mathbf{w}} \mathrm{KL}(\mathbb{P}(y^{tr} \,|\, \hat{y}^{tr}_{s,\mathbf{w}}) \,||\, \mathbb{P}(y^{c} \,|\, \hat{y}^{c}_{s,\mathbf{w}}))$$

Volating the invariant learning principle

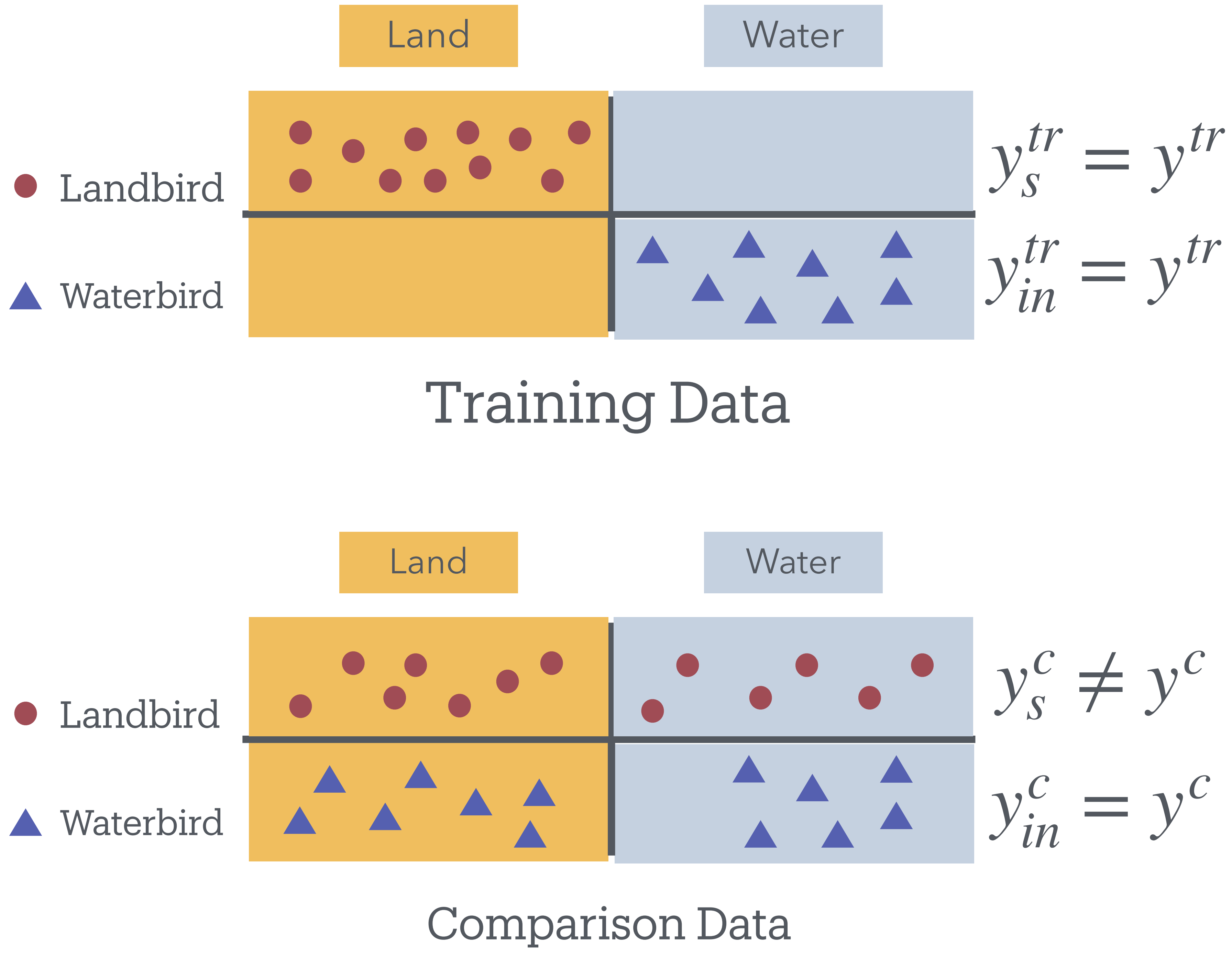
High Correlation: **Spurious attribute**

$$\mathrm{KL}(\mathbb{P}(y^{tr} \,|\, \hat{y}^{tr}_{s,\mathbf{w}}) \,||\, \mathbb{P}(y^{c} \,|\, \hat{y}^{c}_{s,\mathbf{w}})) \geq 0$$

Goal: Inferring Preciser Group Label

GIC: Group **I**nference via data **C**omparison

Spurious Correlation varies in Datasets with (slight) different group distribution



Term 2: **Correlation Term**

Variability in this correlation between datasets with different group distributions:

$$\max_{\mathbf{w}} \mathrm{KL}(\mathbb{P}(y^{tr} \,|\, \hat{y}^{tr}_{s,\mathbf{w}}) \,||\, \mathbb{P}(y^{c} \,|\, \hat{y}^{c}_{s,\mathbf{w}}))$$

Volating the invariant learning principle

Goal: Inferring Preciser Group Label

GIC: **G**roup **I**nference via data **C**omparison

Term 1: **Correlation Term**

Encourage the high correlation between y and y_s in the training set.

$$\max_{\mathbf{w}} I(y^{tr}; \hat{y}^{tr}_{s,\mathbf{w}})$$

Term 2: **Correlation Term**

Variability in this correlation between datasets with different group distributions:

$$\max_{\mathbf{w}} \mathrm{KL}(\mathbb{P}(y^{tr} \mid \hat{y}^{tr}_{s,\mathbf{w}}) \,||\, \mathbb{P}(y^{c} \mid \hat{y}^{c}_{s,\mathbf{w}}))$$

Training Objective: $\max_{\mathbf{w}} I(y^{tr}; \hat{y}^{tr}_{s,\mathbf{w}}) + \gamma \mathrm{KL}(\mathbb{P}(y^{tr} \mid \hat{y}^{tr}_{s,\mathbf{w}}) \,||\, \mathbb{P}(y^{c} \mid \hat{y}^{c}_{s,\mathbf{w}}))$

where $\gamma \geq 0$ is a weighting parameter used to balance Correlation Term and Spurious Term.

Goal: Inferring Preciser Group Label

GIC: **G**roup **I**nference via data **C**omparison

Term 1: **Correlation Term**

Encourage the high correlation between y and y_s in the training set.

$$\max_{\mathbf{w}} I(y^{tr}; \hat{y}^{tr}_{s,\mathbf{w}})$$

Term 2: **Correlation Term**

Variability in this correlation between datasets with different group distributions:

$$\max_{\mathbf{w}} \mathrm{KL}(\mathbb{P}(y^{tr} \mid \hat{y}^{tr}_{s,\mathbf{w}}) \mid\mid \mathbb{P}(y^{c} \mid \hat{y}^{c}_{s,\mathbf{w}}))$$

Training Objective: $\max_{\mathbf{w}} I(y^{tr}; \hat{y}^{tr}_{s,\mathbf{w}}) + \gamma \mathrm{KL}(\mathbb{P}(y^{tr} \mid \hat{y}^{tr}_{s,\mathbf{w}}) \mid\mid \mathbb{P}(y^{c} \mid \hat{y}^{c}_{s,\mathbf{w}}))$

where $\gamma \geq 0$ is a weighting parameter used to balance Correlation Term and Spurious Term.

Goal: Inferring Preciser Group Label

GIC: **G**roup **I**nference via data **C**omparison

Training Objective: $\max_{\mathbf{w}} I(y^{tr}; \hat{y}^{tr}_{s,\mathbf{w}}) + \gamma \mathrm{KL}(\mathbb{P}(y^{tr} \mid \hat{y}^{tr}_{s,\mathbf{w}}) \,||\, \mathbb{P}(y^{c} \mid \hat{y}^{c}_{s,\mathbf{w}}))$

Mutual information is difficult to accurately estimate (Paninski, 2003; Belghazi et al., 2018)

Goal: Inferring Preciser Group Label

GIC: **G**roup **I**nference via data **C**omparison

Training Objective: $\max_{\mathbf{w}} I(y^{tr}; \hat{y}^{tr}_{s,\mathbf{w}}) + \gamma \mathrm{KL}(\mathbb{P}(y^{tr} \mid \hat{y}^{tr}_{s,\mathbf{w}}) \,||\, \mathbb{P}(y^{c} \mid \hat{y}^{c}_{s,\mathbf{w}}))$

Mutual information is difficult to accurately estimate (Paninski, 2003; Belghazi et al., 2018)

Term 1: **Correlation Term**

Encourage the high correlation between y and y_s in the training set.

$$\max_{\mathbf{w}} I(y^{tr}; \hat{y}^{tr}_{s,\mathbf{w}})$$

Replace

$$\min_{\mathbf{w}} H(y^{tr}, \hat{y}^{tr}_{s,\mathbf{w}})$$

Goal: Inferring Preciser Group Label

GIC: **G**roup **I**nference via data **C**omparison

Training Objective: $\max_{\mathbf{w}} I(y^{tr}; \hat{y}^{tr}_{s,\mathbf{w}}) + \gamma \mathrm{KL}(\mathbb{P}(y^{tr} \,|\, \hat{y}^{tr}_{s,\mathbf{w}}) \,||\, \mathbb{P}(y^{c} \,|\, \hat{y}^{c}_{s,\mathbf{w}}))$

Cannot handle the situation where the comparison data is unlabeled (y^c).

Goal: Inferring Preciser Group Label

GIC: **G**roup **I**nference via data **C**omparison

Training Objective: $\max_{\mathbf{w}} I(y^{tr}; \hat{y}^{tr}_{s,\mathbf{w}}) + \gamma \mathrm{KL}(\mathbb{P}(y^{tr} \,|\, \hat{y}^{tr}_{s,\mathbf{w}}) \,||\, \mathbb{P}(y^{c} \,|\, \hat{y}^{c}_{s,\mathbf{w}}))$

Cannot handle the situation where the comparison data is unlabeled (y^c).

Term 2: **Correlation Term**

Variability in this correlation between datasets with different group distributions:

$$\max_{\mathbf{w}} \mathrm{KL}(\mathbb{P}(y^{tr} \,|\, \hat{y}^{tr}_{s,\mathbf{w}}) \,||\, \mathbb{P}(y^{c} \,|\, \hat{y}^{c}_{s,\mathbf{w}}))$$

Theorem 3.1

$$\mathrm{KL}(\mathbb{P}(y^{tr} \,|\, \hat{y}^{tr}_{s,\mathbf{w}}) \,||\, \mathbb{P}(y^{c} \,|\, \hat{y}^{c}_{s,\mathbf{w}}) \geq \mathrm{KL}(\mathbb{P}(\mathbf{z}^{tr} \,|\, \hat{y}^{tr}_{s,\mathbf{w}}) \,||\, \mathbb{P}(\mathbf{z}^{c} \,|\, \hat{y}^{c}_{s,\mathbf{w}}))$$

Goal: Inferring Preciser Group Label

GIC: **G**roup **I**nference via data **C**omparison

Training Objective: $\max_{\mathbf{w}} I(y^{tr}; \hat{y}^{tr}_{s,\mathbf{w}}) + \gamma \mathrm{KL}(\mathbb{P}(y^{tr} \mid \hat{y}^{tr}_{s,\mathbf{w}}) \mid\mid \mathbb{P}(y^{c} \mid \hat{y}^{c}_{s,\mathbf{w}}))$

Better Training Objective:

$\downarrow$

Labeled Comparison Data (GIC_{c_y})

$$\min_{\mathbf{w}} H(y^{tr}, \hat{y}^{tr}_{s,\mathbf{w}}) - \gamma \mathrm{KL}(\mathbb{P}(\mathbf{y}^{tr} \mid \hat{y}^{tr}_{s,\mathbf{w}}) \mid\mid \mathbb{P}(\mathbf{y}^{c} \mid \hat{y}^{c}_{s,\mathbf{w}}))$$

Unlabeled Comparison Data (GIC_{c})

$$\min_{\mathbf{w}} H(y^{tr}, \hat{y}^{tr}_{s,\mathbf{w}}) - \gamma \mathrm{KL}(\mathbb{P}(\mathbf{z}^{tr} \mid \hat{y}^{tr}_{s,\mathbf{w}}) \mid\mid \mathbb{P}(\mathbf{z}^{c} \mid \hat{y}^{c}_{s,\mathbf{w}}))$$

Goal: Inferring Preciser Group Label

GIC: Group Inference via data Comparison

Better Training Objective:

Labeled Comparison Data (GIC_{c_y})

$$\min_{\mathbf{w}} H(y^{tr}, \hat{y}^{tr}_{s,\mathbf{w}}) - \gamma \mathrm{KL}(\mathbb{P}(\mathbf{y}^{tr} \mid \hat{y}^{tr}_{s,\mathbf{w}}) \,||\, \mathbb{P}(\mathbf{y}^{c} \mid \hat{y}^{c}_{s,\mathbf{w}}))$$

Unlabeled Comparison Data (GIC_c)

$$\min_{\mathbf{w}} H(y^{tr}, \hat{y}^{tr}_{s,\mathbf{w}}) - \gamma \mathrm{KL}(\mathbb{P}(\mathbf{z}^{tr} \mid \hat{y}^{tr}_{s,\mathbf{w}}) \,||\, \mathbb{P}(\mathbf{z}^{c} \mid \hat{y}^{c}_{s,\mathbf{w}}))$$

The connection with ERM-based method:

When there is **no** group difference or $\gamma = 0$

GIC **degenerates** to ERM

GIC: Regularized ERM

ERM-based inference should serve as the performance baseline for GIC.

Goal: Inferring Preciser Group Label

GIC: Group Inference via data Comparison

Better Training Objective:

Labeled Comparison Data (GIC_{c_y})

$$\min_{\mathbf{w}} H(y^{tr}, \hat{y}^{tr}_{s,\mathbf{w}}) - \gamma \mathrm{KL}(\mathbb{P}(\mathbf{y}^{tr} \mid \hat{y}^{tr}_{s,\mathbf{w}}) \mid\mid \mathbb{P}(\mathbf{y}^{c} \mid \hat{y}^{c}_{s,\mathbf{w}}))$$

Unlabeled Comparison Data (GIC_c)

$$\min_{\mathbf{w}} H(y^{tr}, \hat{y}^{tr}_{s,\mathbf{w}}) - \gamma \mathrm{KL}(\mathbb{P}(\mathbf{z}^{tr} \mid \hat{y}^{tr}_{s,\mathbf{w}}) \mid\mid \mathbb{P}(\mathbf{z}^{c} \mid \hat{y}^{c}_{s,\mathbf{w}}))$$

Algorithm 1 GIC

Input: Training data $\mathcal{D}$; comparison data $\mathcal{C}$; feature extractor $\Phi(\cdot)$; weighting parameters γ; training epochs K of GIC

Stage 1: Extracting feature representations

Obtain $\mathbf{z}^{tr} = \Phi(\mathbf{x}^{tr}), \mathbf{z}^{c} = \Phi(\mathbf{x}^{c})$ where $\mathbf{x}^{tr} \in \mathcal{D}, \mathbf{x}^{c} \in \mathcal{C}$.

Stage 2: Inferring group labels

Initialize the parameters $\mathbf{w}$ for spurious attribute classifier f_{GIC}.

for epoch 1 to K **do**

  if the true label of $\mathcal{C}$ is available **then**

    Optimizing Equation (11) to update $\mathbf{w}$.

  else

    Optimizing Equation (12) to update $\mathbf{w}$.

  end if

end for

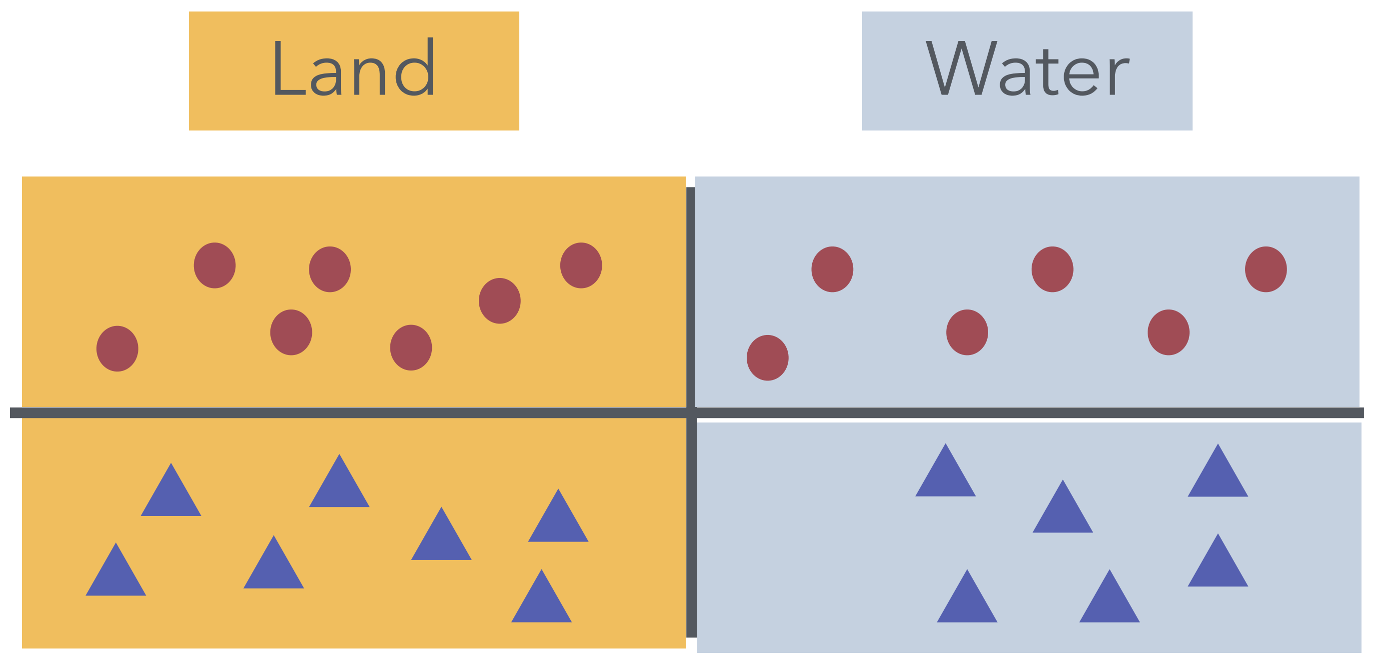
Infer spurious attribute labels $\hat{y}^{tr}_{s,\mathbf{w}} = f_{\mathrm{GIC}}(\mathbf{z}^{tr}; \mathbf{w})$.

Return: Pseudo group labels $\hat{g} = (y^{tr}, \hat{y}^{tr}_{s,\mathbf{w}})$.

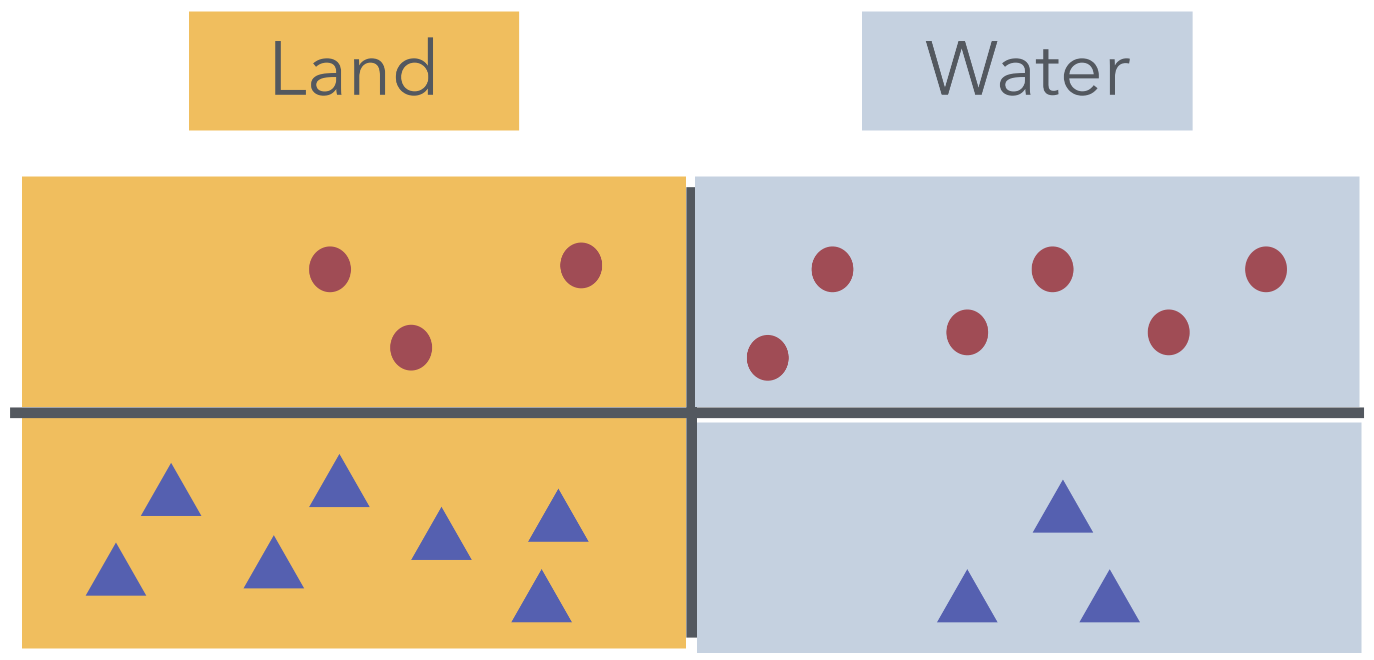
Goal: Inferring Preciser Group Label

GIC: **G**roup **I**nference via data **C**omparison

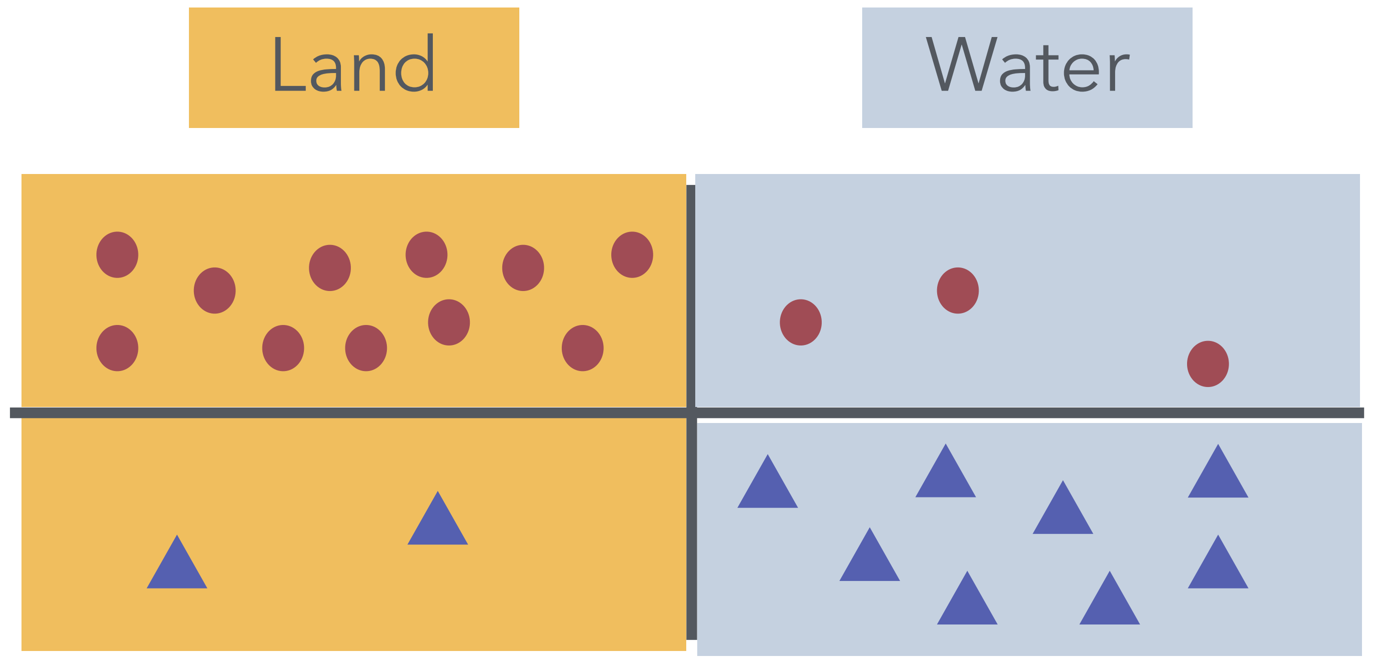
Possible **Source** of Comparison Data



Validation Data



Test Data



Non-uniform from Training Data

Goal: Inferring Preciser Group Label

GIC: Group **I**nference via data **C**omparison

Possible **Source** of Comparison Data

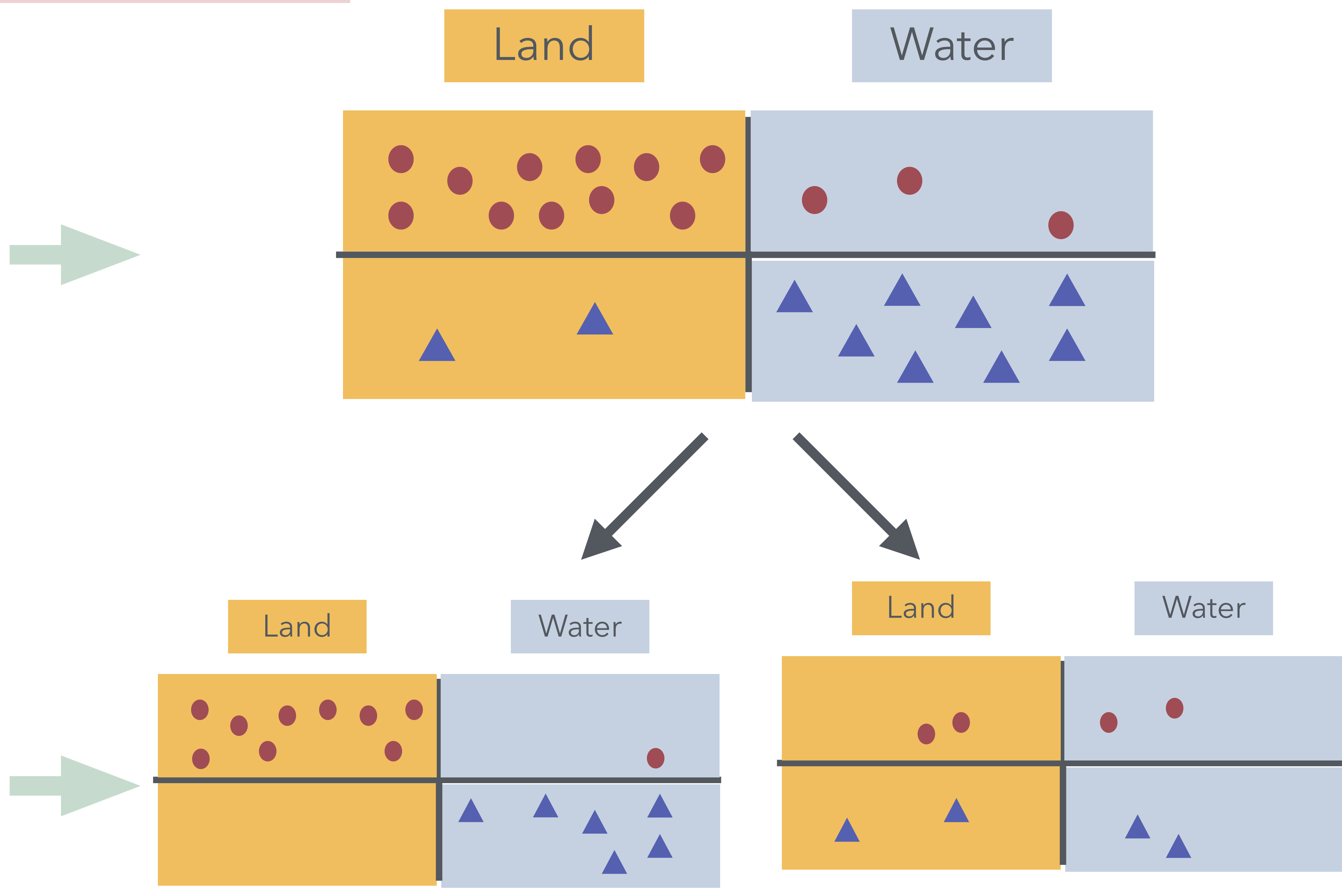
Non-uniform from Training Data

Step 1: ERM-based group inference

Infer group labels of training data via ERM-based method.

Step 2: Sampling in non-uniform manner

1. Train identification model f_{id} via ERM
2. Compute **error set E** of misclassified training examples



Training Data

Comparison Data

Goal: Inferring Preciser Group Label

Experiments on GIC: Datasets

Visualization of evaluated datasets with minority groups marked by red boxes

Comparison Data: (unlabeled) validation data

Goal: Inferring Preciser Group Label

Experiments on GIC: Performance in Mitigating Spurious Correlation

Method	Group Labels Train / Val	CMNIST Avg.	CMNIST Worst	Waterbirds Avg.	Waterbirds Worst	CelebA Avg.	CelebA Worst	CivilComments Avg.	CivilComments Worst
Oracle Group labels are required									
GroupDRO	✓/×	74.4±0.5	69.8±2.6	92.0±0.6	89.9±0.6	91.2±0.4	87.2±1.6	89.9±0.5	70.0±2.0
LISA	✓/×	74.0±0.1	73.3±0.2	91.8±0.3	89.2±0.6	92.4±0.4	89.3±1.1	89.2±0.9	72.6±0.1
DFR	×/✓	72.2±1.1	70.6±1.1	94.2±0.4	92.9±0.2	91.3±0.3	88.3±1.1	87.2±0.3	70.1±0.8
SSA	×/✓	75.0±0.3	71.1±0.4	92.2±0.9	89.0±0.6	92.8±0.1	89.8±1.3	88.2±2.0	69.9±2.0
Oracle Group labels are not required									
ERM	×/×	12.9±0.8	3.4±0.9	97.3±1.0	62.6±0.3	94.9±0.3	47.7±2.1	92.1±0.4	58.6±1.7
JTT	×/×	76.4±3.3	67.3±5.1	89.3±0.7	83.8±1.2	88.1±0.3	81.5±1.7	91.1	69.3
EIIL	×/×	74.1±0.2	65.5±5.1	96.5±0.2	77.2±1.0	85.7±0.1	81.7±0.8	90.5±0.2	67.0±2.4
CnC	×/×	-	-	90.9±0.1	**88.5**±0.3	89.9±0.5	88.8±0.9	81.7±0.5	68.9±2.1
GIC$_{\mathcal{C}_y}$-M	×/×	73.2±0.2	**72.2**±0.5	89.6±1.3	86.3±0.1	91.9±0.1	89.4±0.2	90.0±0.2	**72.5**±0.3
GIC$_{\mathcal{C}}$-M	×/×	73.1±0.5	71.7±0.3	89.3±0.8	85.4±0.1	92.1±0.1	**89.5**±0.0	89.7±0.0	72.3±0.2

Invariant learning
methods: Mixup

Better Worst-group Accuracy

Goal: Inferring Preciser Group Label

Experiments on GIC: Performance in Mitigating Spurious Correlation

Method	Group Labels Train / Val	CMNIST Avg.	CMNIST Worst	Waterbirds Avg.	Waterbirds Worst	CelebA Avg.	CelebA Worst	CivilComments Avg.	CivilComments Worst
Oracle Group labels are required									
GroupDRO	✓/×	74.4±0.5	69.8±2.6	92.0±0.6	89.9±0.6	91.2±0.4	87.2±1.6	89.9±0.5	70.0±2.0
LISA	✓/×	74.0±0.1	73.3±0.2	91.8±0.3	89.2±0.6	92.4±0.4	89.3±1.1	89.2±0.9	72.6±0.1
DFR	×/✓	72.2±1.1	70.6±1.1	94.2±0.4	92.9±0.2	91.3±0.3	88.3±1.1	87.2±0.3	70.1±0.8
SSA	×/✓	75.0±0.3	71.1±0.4	92.2±0.9	89.0±0.6	92.8±0.1	89.8±1.3	88.2±2.0	69.9±2.0
Oracle Group labels are not required									
ERM	×/×	12.9±0.8	3.4±0.9	97.3±1.0	62.6±0.3	94.9±0.3	47.7±2.1	92.1±0.4	58.6±1.7
JTT	×/×	76.4±3.3	67.3±5.1	89.3±0.7	83.8±1.2	88.1±0.3	81.5±1.7	91.1	69.3
EIIL	×/×	74.1±0.2	65.5±5.1	96.5±0.2	77.2±1.0	85.7±0.1	81.7±0.8	90.5±0.2	67.0±2.4
CnC	×/×	-	-	90.9±0.1	**88.5**±0.3	89.9±0.5	88.8±0.9	81.7±0.5	68.9±2.1
GIC$_{\mathcal{C}_y}$-M	×/×	73.2±0.2	**72.2**±0.5	89.6±1.3	86.3±0.1	91.9±0.1	89.4±0.2	90.0±0.2	**72.5**±0.3
GIC$_{\mathcal{C}}$-M	×/×	73.1±0.5	71.7±0.3	89.3±0.8	85.4±0.1	92.1±0.1	**89.5**±0.0	89.7±0.0	72.3±0.2

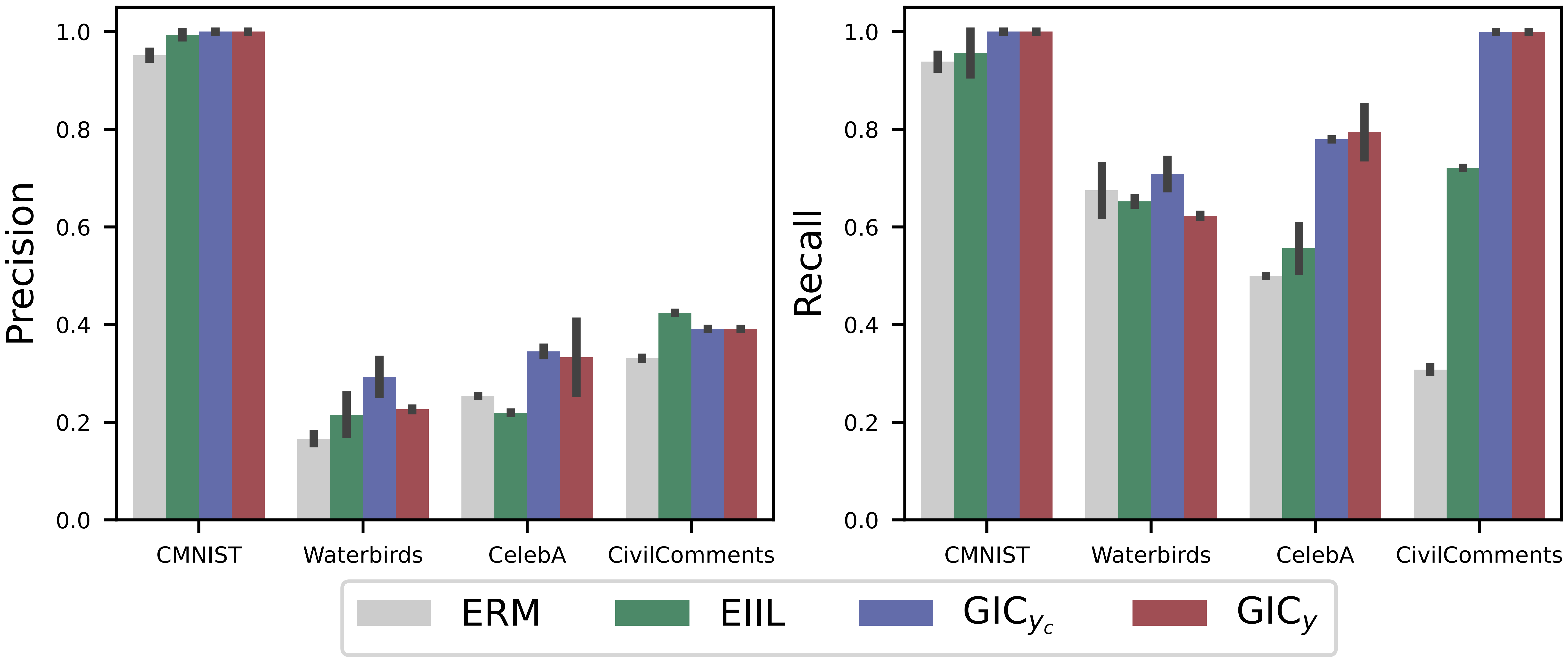
Even competing with methods with group labels on certain datasets

Goal: Inferring Preciser Group Label

Experiments on GIC: Performance in Inferring Group Labels

Precision: proportion of correctly inferred samples belonging to the true minority group (**higher**)

Recall: proportion of samples from the true minority group correctly inferred (**diversity**)



Inferring **Preciser** Group Label

Goal: Inferring Preciser Group Label

Experiments on GIC: Performance in Inferring Group Labels

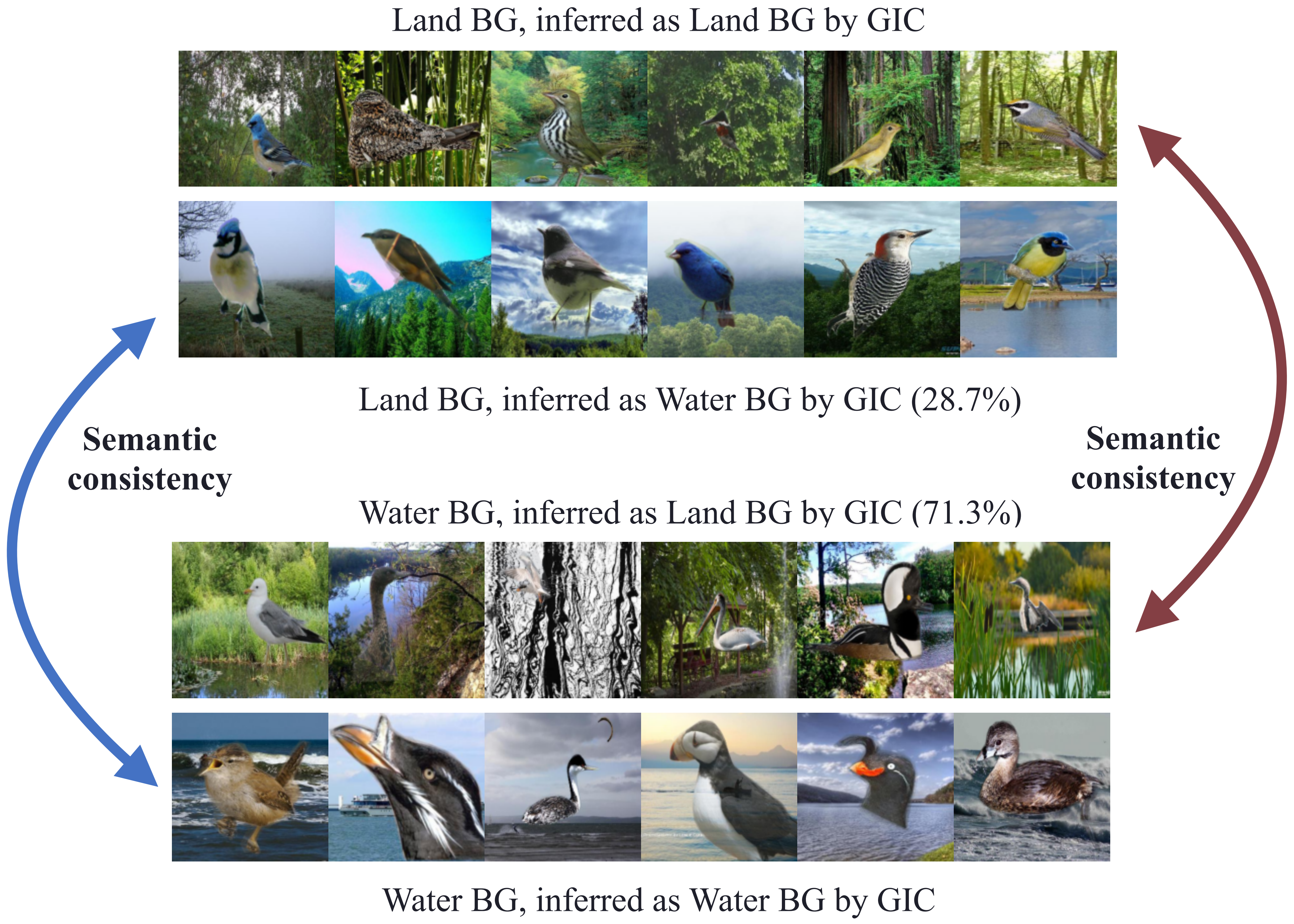
Consider 4 different invariant learning methods

Method	+*GroupDRO*		+*Subsample*		+*Upsample*		+*Mixup*	
	Waterbirds	CelebA	Waterbirds	CelebA	Waterbirds	CelebA	Waterbirds	CelebA
ERM	75.6±0.4	77.2±0.1	79.4±0.3	78.5±0.1	83.8±1.2	81.5±1.7	82.1±0.8	80.6±1.7
EI	77.2±1.0	81.7±0.8	81.9±1.4	82.8±0.5	81.3±0.7	84.8±0.2	85.7±0.4	84.9±3.7
$\boxed{\mathrm{GIC}_{\mathcal{C}_y}}$	**80.2**±0.1	**82.1**±0.3	**83.5**±0.8	**86.1**±2.2	**84.1**±0.0	87.2±0.0	**86.3**±0.1	89.4±0.2
$\boxed{\mathrm{GIC}_{\mathcal{C}}}$	79.2±0.4	79.7±0.6	82.1±1.1	83.1±0.3	82.1±0.7	**87.8**±1.1	85.4±0.1	**89.5**±0.0

Preciser Group Label makes higher worst-group accuracy

Goal: Inferring Preciser Group Label

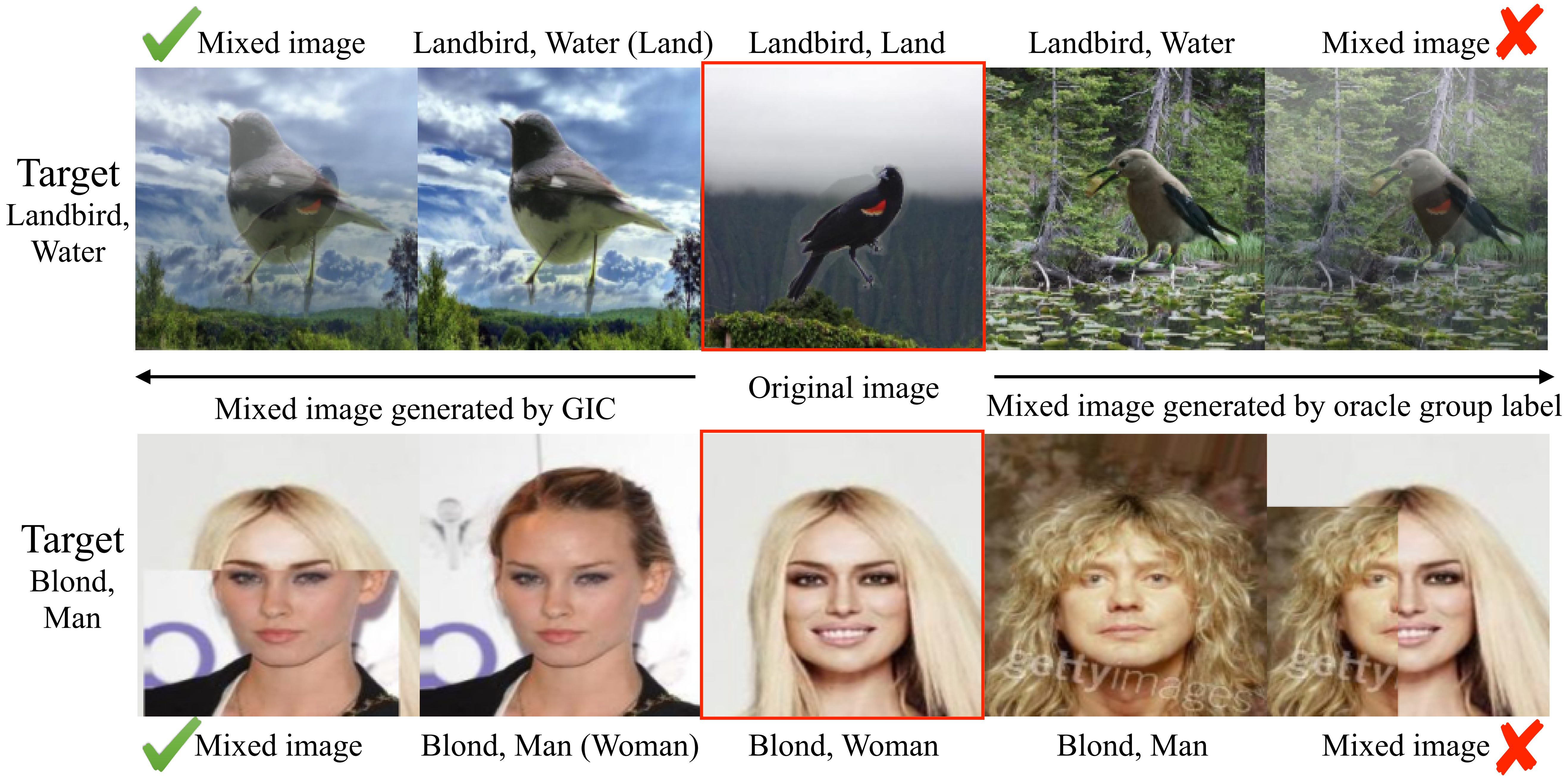
Experiments on GIC: Error Cases Analyse



Semantic Consistency

Goal: Inferring Preciser Group Label

Experiments on GIC: Error Cases Analyse



High semantic consistency benefits methods achieving invariant learning by disrupting image semantics.

Summary

1. Standard ERM may prioritize learning spurious correlations, leading to **poor accuracy** on groups where these correlations do not hold.

Summary

1. Standard ERM may prioritize learning spurious correlations, leading to **poor accuracy** on groups where these correlations do not hold.

2. Improving worst-group accuracy requires group labels: **performs well** but is **expensive**.

Summary

1. Standard ERM may prioritize learning spurious correlations, leading to **poor accuracy** on groups where these correlations do not hold.

2. Improving worst-group accuracy requires group labels: **performs well** but is **expensive**.

3. Group inferred methods have **performance gaps** compared to group annotation utilized methods and may not be applicable when **prior information** is **unavailable**.

Summary

1. Standard ERM may prioritize learning spurious correlations, leading to **poor accuracy** on groups where these correlations do not hold.

2. Improving worst-group accuracy requires group labels: **performs well** but is **expensive**.

3. Group inferred methods have **performance gaps** compared to group annotation utilized methods and may not be applicable when **prior information** is **unavailable**.

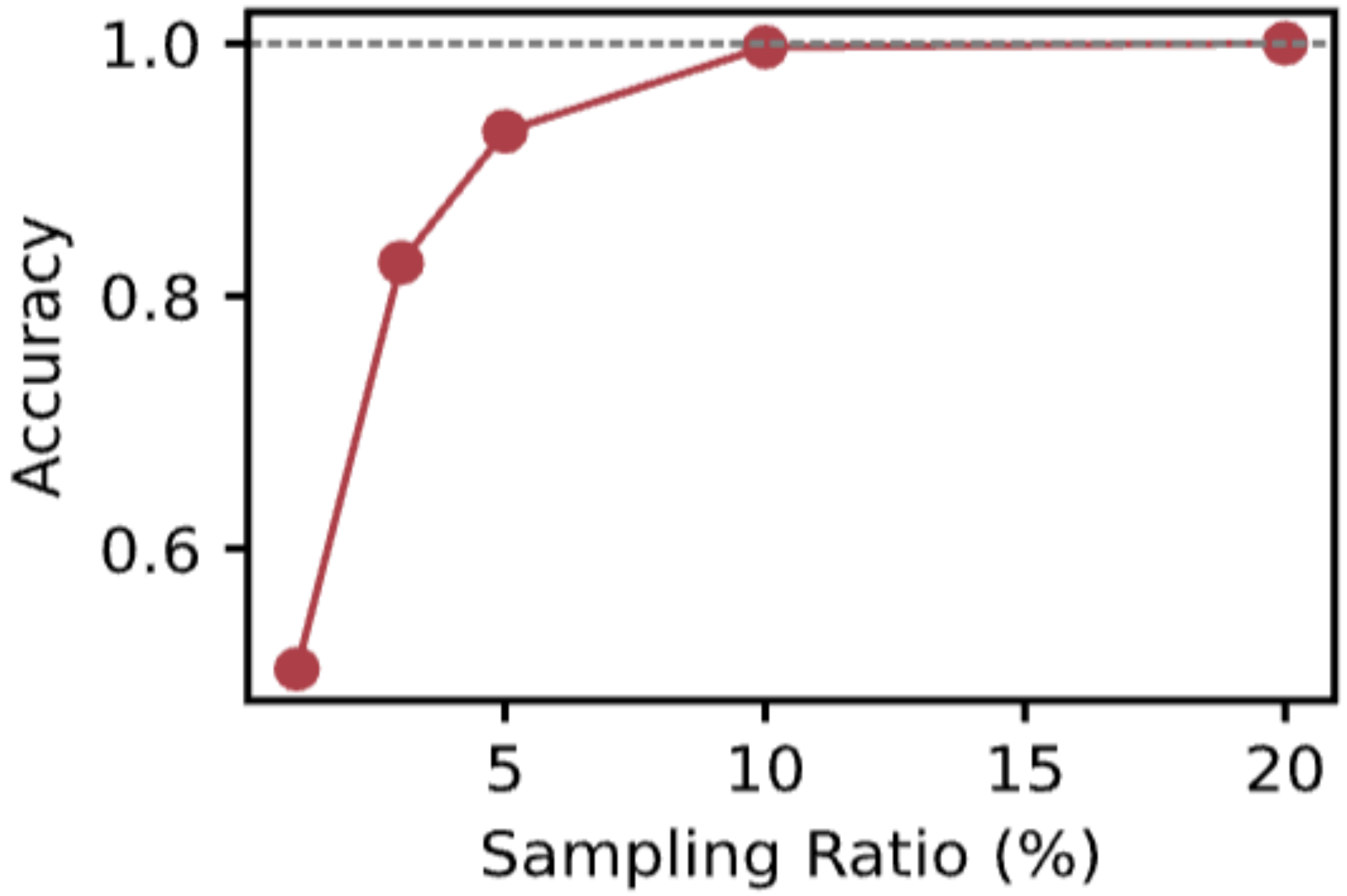
4. GIC: more **preciser** inferring group labels to improve the worst-group performance; **semantic consistency** aids in mitigating spurious correlations.

Thanks

Appendix: Inferring Preciser Group Label

More Experiments on GIC

Comparison Data: Sampling non-uniformly from training data

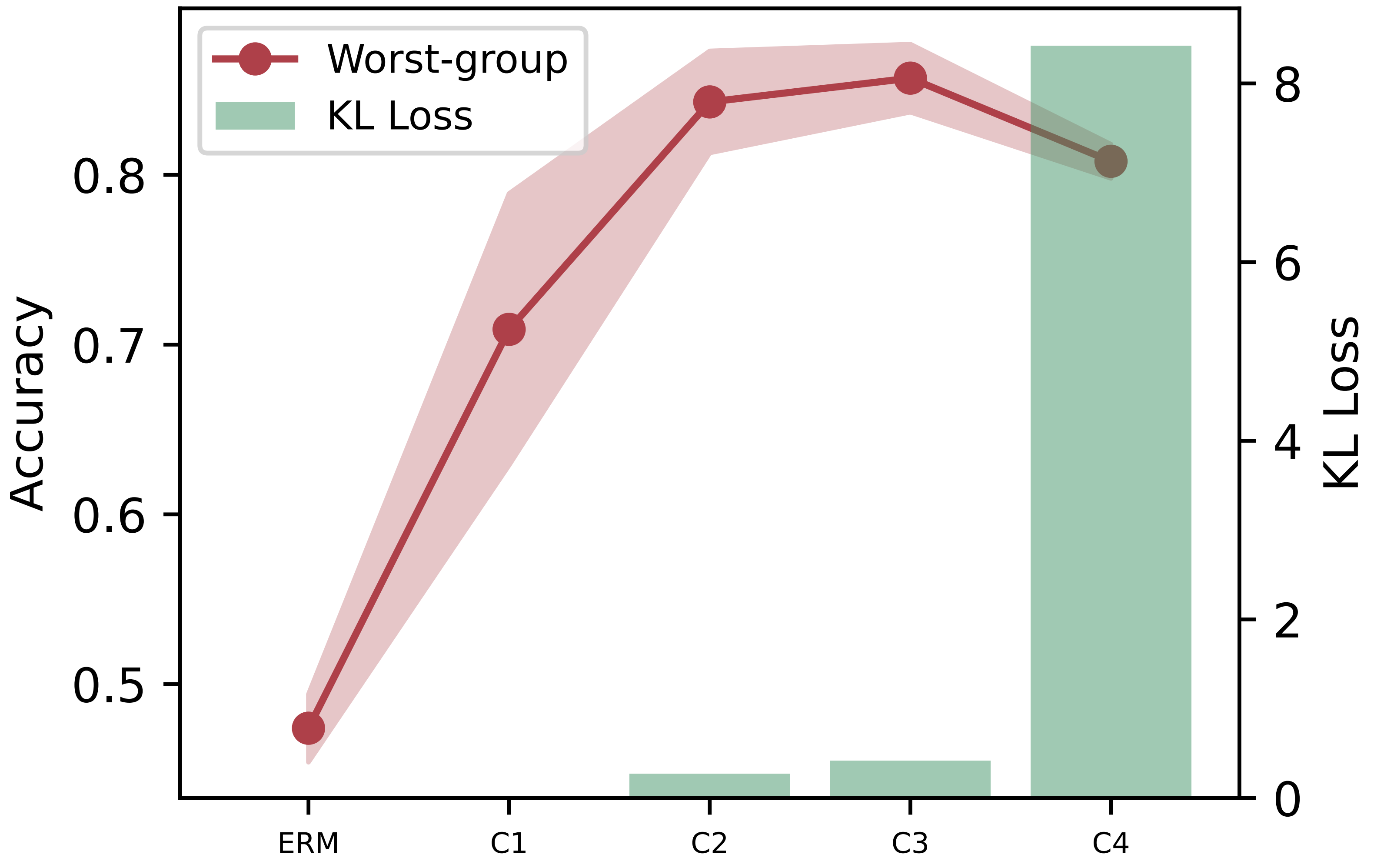


Method	CMNIST	Waterbirds	CelebA
validation			
GIC$_{\mathcal{C}_y}$-M	72.2±0.5	86.3±0.1	89.4±0.2
GIC$_{\mathcal{C}}$-M	71.7±0.3	85.4±0.1	89.5±0.0
non-uniform sampling.			
GIC$_{\mathcal{C}_t}$-M	72.2 ±0.1	85.7±0.3	87.5±0.7

Appendix: Inferring Preciser Group Label

More Experiments on GIC

Slight group difference is enough for GIC-based group inference



Dataset		g1	g2	g3	g4
Training Data		71629	66874	22880	1387
Comparison Data	C1	7163	6687	2288	139
	C2	8535	8276	2874	182
	C3	8535	8276	2874	2874
	C4	2874	2874	2874	2874

Difference increased