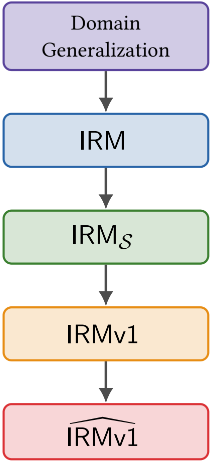
Does Invariant Risk Minimization Capture Invariance?

Pritish Kamath
TTIC

Akilesh Tangella
TTIC

Danica J. Sutherland
TTIC $\longrightarrow$ UBC

Nathan Srebro
TTIC



AISTATS 2021

BIG challenge in Supervised Learning

- Supervised Learning often learns “spurious” correlations.
- Such correlations are easier to detect, but do not hold over data from other sources.

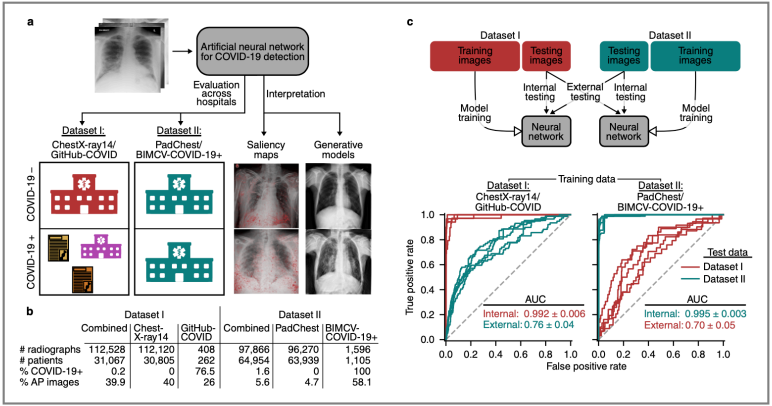
BIG challenge in Supervised Learning

- Supervised Learning often learns “spurious” correlations.
- Such correlations are easier to detect, but do not hold over data from other sources.

AI for radiographic COVID-19 detection selects shortcuts over signal

Alex J. DeGrave[1,2,*], Joseph D. Janizek[1,2,*], and Su-In Lee[1,**]

[www.medrxiv.org/content/10.1101/2020.09.13.20193565v2]



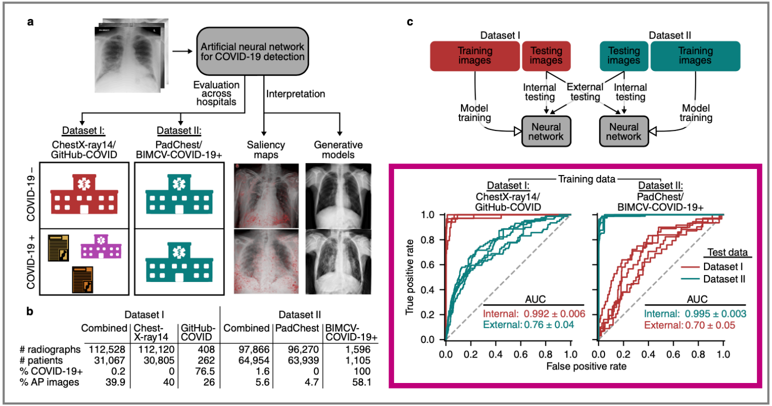
BIG challenge in Supervised Learning

- Supervised Learning often learns “spurious” correlations.
- Such correlations are easier to detect, but do not hold over data from other sources.

AI for radiographic COVID-19 detection selects shortcuts over signal

Alex J. DeGrave[1,2,*], Joseph D. Janizek[1,2,*], and Su-In Lee[1,**]

[www.medrxiv.org/content/10.1101/2020.09.13.20193565v2]



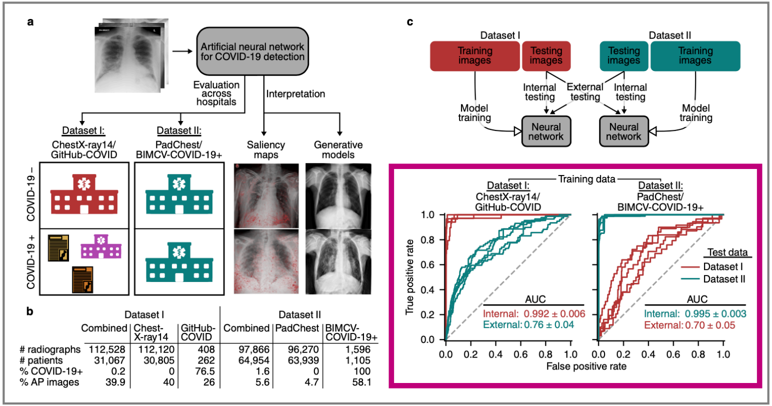
BIG challenge in Supervised Learning

Can we design learning objectives that incentivize our models to only learn correlations that hold over all data sources?

AI for radiographic COVID-19 detection selects shortcuts over signal

Alex J. DeGrave[1,2,*], Joseph D. Janizek[1,2,*], and Su-In Lee[1,**]

[www.medrxiv.org/content/10.1101/2020.09.13.20193565v2]



Domain Generalization

Set of environments $\mathcal{E}$. Each $e \in \mathcal{E}$ corresponds to a distribution $\mathcal{D}_e$ over $\mathcal{X} \times \mathcal{Y}$.

Domain Generalization

Set of environments $\mathcal{E}$. Each $e \in \mathcal{E}$ corresponds to a distribution $\mathcal{D}_e$ over $\mathcal{X} \times \mathcal{Y}$.

Goal: $$\min_{f:\mathcal{X}\to\mathbb{R}} \max_{e\in\mathcal{E}} \mathcal{L}_e(f)$$

where, $\mathcal{L}_e(f) := \mathop{\mathbb{E}}_{(X,Y)\sim\mathcal{D}_e} \ell(f(X), Y)$

Example: Square loss $\ell_{\mathrm{sq}}(\widehat{y}, y) := \frac{1}{2}(\widehat{y} - y)^2$ or Logistic loss $\ell_{\log}(\widehat{y}, y) := \log(1 + \exp(-y\widehat{y}))$

Domain Generalization

Set of environments $\mathcal{E}$. Each $e \in \mathcal{E}$ corresponds to a distribution $\mathcal{D}_e$ over $\mathcal{X} \times \mathcal{Y}$.

Goal: $\min\limits_{f:\mathcal{X}\to\mathbb{R}} \max\limits_{e\in\mathcal{E}} \mathcal{L}_e(f)$

where, $\mathcal{L}_e(f) := \mathop{\mathbb{E}}\limits_{(X,Y)\sim\mathcal{D}_e} \ell(f(X), Y)$

Example: Square loss $\ell_{\text{sq}}(\hat{y}, y) := \frac{1}{2}(\hat{y} - y)^2$ or Logistic loss $\ell_{\log}(\hat{y}, y) := \log(1 + \exp(-y\hat{y}))$

What do we have access to?

- Finite set of training environments $\mathcal{E}_{\text{tr}} \subseteq \mathcal{E}$: *not sampled*!
- Training sets S_e sampled from $\mathcal{D}_e$ for $e \in \mathcal{E}_{\text{tr}}$.

Empirical Risk Minimization

Empirical Risk Minimization (baseline): *Mix the training data sources!*

$$\min_{f:\mathcal{X}\to\mathbb{R}} \sum_{e\in\mathcal{E}_{\mathrm{tr}}} \mathcal{L}_e(f)$$

Empirical Risk Minimization

Empirical Risk Minimization (baseline): *Mix the training data sources!*

$$\min_{f:\mathcal{X}\to\mathbb{R}} \sum_{e\in\mathcal{E}_{\mathrm{tr}}} \mathcal{L}_e(f)$$

Fails to generalize to unseen $e \in \mathcal{E}$ if *spurious* correlations that hold over training environments do not hold over $\mathcal{D}_e$.

Invariant Risk Minimization [Arjovsky, Bottou, Gulrajani, Lopez-Paz '19]

Invariant Risk Minimization (IRM): Only allow predictors that are "invariant" over training environments.

$$\min_{f:\mathcal{X}\to\mathbb{R}} \sum_{e\in\mathcal{E}_{\mathrm{tr}}} \mathcal{L}_e(f)$$

subject to:

- $f = w \circ \varphi$ for $\begin{cases} \text{representation} & \varphi : \mathcal{X} \to \mathcal{Z} \\ \text{predictor} & w : \mathcal{Z} \to \mathbb{R} \end{cases}$
- w is simultaneously optimal for all training environments.

$$\forall\, e \in \mathcal{E}_{\mathrm{tr}} \quad : \qquad w \in \operatorname*{argmin}_{\overline{w}:\mathcal{Z}\to\mathbb{R}} \mathcal{L}_e(\overline{w} \circ \varphi)$$

Invariant Risk Minimization [Arjovsky, Bottou, Gulrajani, Lopez-Paz '19]

Invariant Risk Minimization (IRM): Only allow predictors that are "invariant" over training environments.

$$\min_{f:\mathcal{X}\to\mathbb{R}} \sum_{e\in\mathcal{E}_{\text{tr}}} \mathcal{L}_e(f)$$

subject to:

- $f = w \circ \varphi$ for $\begin{cases} \text{representation} & \varphi : \mathcal{X} \to \mathcal{Z} \\ \text{predictor} & w : \mathcal{Z} \to \mathbb{R} \end{cases}$
- w is simultaneously optimal for all training environments.

$$\forall\, e \in \mathcal{E}_{\text{tr}} \quad : \qquad w \in \operatorname*{argmin}_{\overline{w}:\mathcal{Z}\to\mathbb{R}} \mathcal{L}_e(\overline{w} \circ \varphi)$$

in short:

$$f \in \mathcal{I}(\mathcal{E}_{\text{tr}})$$

Invariant Risk Minimization [Arjovsky, Bottou, Gulrajani, Lopez-Paz '19]

Invariant Risk Minimization (IRM): Only allow predictors that are "invariant" over training environments.

$$\min_{f:\mathcal{X}\to\mathbb{R}} \sum_{e\in\mathcal{E}_{\mathrm{tr}}} \mathcal{L}_e(f)$$

subject to:

- $f = w \circ \varphi$ for $\begin{cases} \text{representation} & \varphi : \mathcal{X} \to \mathcal{Z} \\ \text{predictor} & w : \mathcal{Z} \to \mathbb{R} \end{cases}$
- w is simultaneously optimal for all training environments.

$$\forall\, e \in \mathcal{E}_{\mathrm{tr}} \;:\qquad w \in \operatorname*{argmin}_{\overline{w}:\mathcal{Z}\to\mathbb{R}} \mathcal{L}_e(\overline{w} \circ \varphi)$$

in short:

$$f \in \mathcal{I}(\mathcal{E}_{\mathrm{tr}})$$

Challenging Bi-Level Optimization Problem!

IRM with Linear w [Arjovsky, Bottou, Gulrajani, Lopez-Paz '19]

Invariant Risk Minimization (IRM-Linear): Constrain w to be a linear predictor.

$$\min_{f:\mathcal{X}\to\mathbb{R}} \sum_{e\in\mathcal{E}_{\text{tr}}} \mathcal{L}_e(f)$$

subject to:

- $f = w \circ \varphi$ for $\begin{cases} \text{representation} & \varphi : \mathcal{X} \to \mathbb{R}^d \\ \textit{linear} \text{ predictor} & w : \mathbb{R}^d \to \mathbb{R} \end{cases}$
- w is simultaneously optimal for all training environments.

$$\forall\, e \in \mathcal{E}_{\text{tr}} \quad : \qquad w \in \operatorname*{argmin}_{\overline{w}\in\mathbb{R}^d} \mathcal{L}_e(\langle \overline{w}, \varphi\rangle)$$

IRM with Linear w [Arjovsky, Bottou, Gulrajani, Lopez-Paz '19]

Invariant Risk Minimization (IRM-Linear): Constrain w to be a linear predictor.

$$\min_{f:\mathcal{X}\to\mathbb{R}} \sum_{e\in\mathcal{E}_{\text{tr}}} \mathcal{L}_e(f)$$

subject to:

- $f = w \circ \varphi$ for $\begin{cases} \text{representation} & \varphi : \mathcal{X} \to \mathbb{R}^d \\ \textit{linear} \text{ predictor} & w : \mathbb{R}^d \to \mathbb{R} \end{cases}$
- w is simultaneously optimal for all training environments.

$$\forall\, e \in \mathcal{E}_{\text{tr}} \quad : \qquad w \in \operatorname*{argmin}_{\overline{w}\in\mathbb{R}^d} \mathcal{L}_e(\langle \overline{w}, \varphi\rangle)$$

in short:

$$f \in \mathcal{I}_{\mathcal{S}}(\mathcal{E}_{\text{tr}})$$

IRM with Linear w [Arjovsky, Bottou, Gulrajani, Lopez-Paz '19]

Invariant Risk Minimization (IRM-Linear): Constrain w to be a linear predictor.

$$\min_{f:\mathcal{X}\to\mathbb{R}} \sum_{e\in\mathcal{E}_{\text{tr}}} \mathcal{L}_e(f)$$

subject to:

- $f = w \circ \varphi$ for $\begin{cases} \text{representation} & \varphi : \mathcal{X} \to \mathbb{R}^d \\ \textit{linear} \text{ predictor} & w : \mathbb{R}^d \to \mathbb{R} \end{cases}$
- w is simultaneously optimal for all training environments.

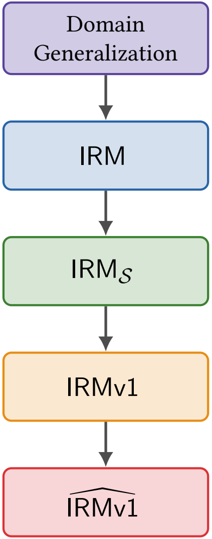
$$\forall\, e \in \mathcal{E}_{\text{tr}} \ : \qquad w \in \operatorname*{argmin}_{\overline{w}\in\mathbb{R}^d} \mathcal{L}_e(\langle \overline{w}, \varphi\rangle)$$

in short:

$$f \in \mathcal{I}_{\mathcal{S}}(\mathcal{E}_{\text{tr}})$$

$$\min_{f:\mathcal{X}\to\mathbb{R}} \sum_{e\in\mathcal{E}_{\text{tr}}} \mathcal{L}_e(f) + \lambda \cdot \left| \frac{\partial}{\partial w} \mathcal{L}_e(w\cdot f)|_{w=1} \right|^2 \qquad \ldots\text{(IRMv1)}$$

Our Contributions



$$\min_f \max_{e\in\mathcal{E}} \mathcal{L}_e(f)$$

$$\min_{f\in\mathcal{I}(\mathcal{E}_{\mathrm{tr}})} \sum_{e\in\mathcal{E}_{\mathrm{tr}}} \mathcal{L}_e(f)$$

$$\min_{f\in\mathcal{I}_{\mathcal{S}}(\mathcal{E}_{\mathrm{tr}})} \sum_{e\in\mathcal{E}_{\mathrm{tr}}} \mathcal{L}_e(f)$$

$$\min_f \sum_{e\in\mathcal{E}_{\mathrm{tr}}} \mathcal{L}_e(f) + \lambda \cdot \left| \frac{\partial}{\partial w} \mathcal{L}_e(w\cdot f)|_{w=1} \right|^2$$

- Empirical loss $\widehat{\mathcal{L}}_e$
- f = neural network
- Training with (S)GD
- Choose λ in IRMv1

Our Contributions

Most subsequent work interchangeably use IRM, $\text{IRM}_{\mathcal{S}}$, IRMv1 and $\widehat{\text{IRMv1}}$.

Domain Generalization $\quad \min_{f} \max_{e \in \mathcal{E}} \mathcal{L}_e(f)$

↓

IRM $\quad \min_{f \in \mathcal{I}(\mathcal{E}_{\text{tr}})} \sum_{e \in \mathcal{E}_{\text{tr}}} \mathcal{L}_e(f)$

↓

$\text{IRM}_{\mathcal{S}}$ $\quad \min_{f \in \mathcal{I}_{\mathcal{S}}(\mathcal{E}_{\text{tr}})} \sum_{e \in \mathcal{E}_{\text{tr}}} \mathcal{L}_e(f)$

↓

IRMv1 $\quad \min_{f} \sum_{e \in \mathcal{E}_{\text{tr}}} \mathcal{L}_e(f) + \lambda \cdot \left| \frac{\partial}{\partial w} \mathcal{L}_e(w \cdot f)|_{w=1} \right|^2$

↓

$\widehat{\text{IRMv1}}$

- Empirical loss $\widehat{\mathcal{L}}_e$
- f = neural network
- Training with (S)GD
- Choose λ in IRMv1

Our Contributions

Most subsequent work interchangeably use IRM, $\text{IRM}_{\mathcal{S}}$, IRMv1 and $\widehat{\text{IRMv1}}$.

If $\widehat{\text{IRMv1}}$ does not work in some example, which step is to be blamed?

Domain Generalization $\quad \min_f \max_{e \in \mathcal{E}} \mathcal{L}_e(f)$

$\downarrow$

IRM $\quad \min_{f \in \mathcal{I}(\mathcal{E}_{\text{tr}})} \sum_{e \in \mathcal{E}_{\text{tr}}} \mathcal{L}_e(f)$

$\downarrow$

$\text{IRM}_{\mathcal{S}}$ $\quad \min_{f \in \mathcal{I}_{\mathcal{S}}(\mathcal{E}_{\text{tr}})} \sum_{e \in \mathcal{E}_{\text{tr}}} \mathcal{L}_e(f)$

$\downarrow$

IRMv1 $\quad \min_f \sum_{e \in \mathcal{E}_{\text{tr}}} \mathcal{L}_e(f) + \lambda \cdot \left| \frac{\partial}{\partial w} \mathcal{L}_e(w \cdot f)|_{w=1} \right|^2$

$\downarrow$

$\widehat{\text{IRMv1}}$

- Empirical loss $\widehat{\mathcal{L}}_e$
- f = neural network
- Training with (S)GD
- Choose λ in IRMv1

Our Contributions

Domain Generalization $\quad \min_f \max_{e\in\mathcal{E}} \mathcal{L}_e(f)$

$\downarrow$

IRM $\quad \min_{f\in\mathcal{I}(\mathcal{E}_{\text{tr}})} \sum_{e\in\mathcal{E}_{\text{tr}}} \mathcal{L}_e(f)$

$\downarrow$

$\text{IRM}_{\mathcal{S}}$ $\quad \min_{f\in\mathcal{I}_{\mathcal{S}}(\mathcal{E}_{\text{tr}})} \sum_{e\in\mathcal{E}_{\text{tr}}} \mathcal{L}_e(f)$

$\downarrow$

IRMv1 $\quad \min_f \sum_{e\in\mathcal{E}_{\text{tr}}} \mathcal{L}_e(f) + \lambda \cdot \left| \frac{\partial}{\partial w} \mathcal{L}_e(w\cdot f)|_{w=1} \right|^2$

$\downarrow$

$\widehat{\text{IRMv1}}$

- Empirical loss $\widehat{\mathcal{L}}_e$
- f = neural network
- Training with (S)GD
- Choose λ in IRMv1

Result #1: In a simple theoretical example (abstraction of `Colored-MNIST` studied by [Arjovsky et al.]):

- IRM succeeds, $\text{IRM}_{\mathcal{S}}$ fails.
- Even with *infinitely* many training environments!
- Because: $\mathcal{I}_{\mathcal{S}}(\mathcal{E}_{\text{tr}}) \supsetneq \mathcal{I}(\mathcal{E}_{\text{tr}})$.

Our Contributions

Domain Generalization $\quad \min_{f} \max_{e \in \mathcal{E}} \mathcal{L}_e(f)$

IRM $\quad \min_{f \in \mathcal{I}(\mathcal{E}_{\text{tr}})} \sum_{e \in \mathcal{E}_{\text{tr}}} \mathcal{L}_e(f)$

$\text{IRM}_{\mathcal{S}}$ $\quad \min_{f \in \mathcal{I}_{\mathcal{S}}(\mathcal{E}_{\text{tr}})} \sum_{e \in \mathcal{E}_{\text{tr}}} \mathcal{L}_e(f)$

IRMv1 $\quad \min_{f} \sum_{e \in \mathcal{E}_{\text{tr}}} \mathcal{L}_e(f) + \lambda \cdot \left| \frac{\partial}{\partial w} \mathcal{L}_e(w \cdot f)|_{w=1} \right|^2$

$\widehat{\text{IRMv1}}$

- Empirical loss $\widehat{\mathcal{L}}_e$
- f = neural network
- Training with (S)GD
- Choose λ in IRMv1

Result #2: In another simple theoretical example:

- Training Envs correctly identify invariances $\mathcal{I}(\mathcal{E}_{\text{tr}}) = \mathcal{I}(\mathcal{E})$.
- Yet, IRM chooses sub-optimal invariant predictor.
- Because: Loss of an invariant predictor need not be invariant.

Result #1: In a simple theoretical example (abstraction of `Colored-MNIST` studied by [Arjovsky et al.]):

- IRM succeeds, $\text{IRM}_{\mathcal{S}}$ fails.
- Even with *infinitely* many training environments!
- Because: $\mathcal{I}_{\mathcal{S}}(\mathcal{E}_{\text{tr}}) \supsetneq \mathcal{I}(\mathcal{E}_{\text{tr}})$.

Our Contributions

Result #3: Establish (sufficient) conditions under which, finite training environments capture right invariances:

$$\mathcal{I}(\mathcal{E}_{\text{tr}}) = \mathcal{I}(\mathcal{E})$$

Result #2: In another simple theoretical example:

- Training Envs correctly identify invariances $\mathcal{I}(\mathcal{E}_{\text{tr}}) = \mathcal{I}(\mathcal{E})$.
- Yet, IRM chooses sub-optimal invariant predictor.
- Because: Loss of an invariant predictor need not be invariant.

Result #1: In a simple theoretical example (abstraction of `Colored-MNIST` studied by [Arjovsky et al.]):

- IRM succeeds, $\text{IRM}_{\mathcal{S}}$ fails.
- Even with *infinitely* many training environments!
- Because: $\mathcal{I}_{\mathcal{S}}(\mathcal{E}_{\text{tr}}) \supsetneq \mathcal{I}(\mathcal{E}_{\text{tr}})$.

Domain Generalization: $\min_f \max_{e \in \mathcal{E}} \mathcal{L}_e(f)$

IRM: $\min_{f \in \mathcal{I}(\mathcal{E}_{\text{tr}})} \sum_{e \in \mathcal{E}_{\text{tr}}} \mathcal{L}_e(f)$

$\text{IRM}_{\mathcal{S}}$: $\min_{f \in \mathcal{I}_{\mathcal{S}}(\mathcal{E}_{\text{tr}})} \sum_{e \in \mathcal{E}_{\text{tr}}} \mathcal{L}_e(f)$

IRMv1: $\min_f \sum_{e \in \mathcal{E}_{\text{tr}}} \mathcal{L}_e(f) + \lambda \cdot \left| \frac{\partial}{\partial w} \mathcal{L}_e(w \cdot f)|_{w=1} \right|^2$

$\widehat{\text{IRMv1}}$:

- Empirical loss $\widehat{\mathcal{L}}_e$
- f = neural network
- Training with (S)GD
- Choose λ in IRMv1

Our Contributions

Result #3: Establish (sufficient) conditions under which, finite training environments capture right invariances:

$$\mathcal{I}(\mathcal{E}_{\text{tr}}) = \mathcal{I}(\mathcal{E})$$

Result #2: In an

- Training Envs
- Yet, IRM choos
- Because: Loss

Result #1: In a simple theoretical example (abstraction of `Colored-MNIST` studied by [Arjovsky et al.]):

- IRM succeeds, IRM_S fails.
- Even with *infinitely* many training environments!
- Because: $\mathcal{I}_S(\mathcal{E}_{\text{tr}}) \supsetneq \mathcal{I}(\mathcal{E}_{\text{tr}})$.

Domain Generalization $\quad \min_f \max_{e \in \mathcal{E}} \mathcal{L}_e(f)$

IRM $\quad \min_{f \in \mathcal{I}(\mathcal{E}_{\text{tr}})} \sum_{e \in \mathcal{E}_{\text{tr}}} \mathcal{L}_e(f)$

$\sum_{e \in \mathcal{E}_{\text{tr}}} \mathcal{L}_e(f)$

IRMv1 $\quad \min_f \sum_{e \in \mathcal{E}_{\text{tr}}} \mathcal{L}_e(f) + \lambda \cdot \left| \frac{\partial}{\partial w} \mathcal{L}_e(w \cdot f)|_{w=1} \right|^2$

$\widehat{\text{IRMv1}}$

- Empirical loss $\widehat{\mathcal{L}}_e$
- f = neural network
- Training with (S)GD
- Choose λ in IRMv1

Key Takeaway

IRM is a potentially powerful paradigm.
Making it practical needs further investigation.