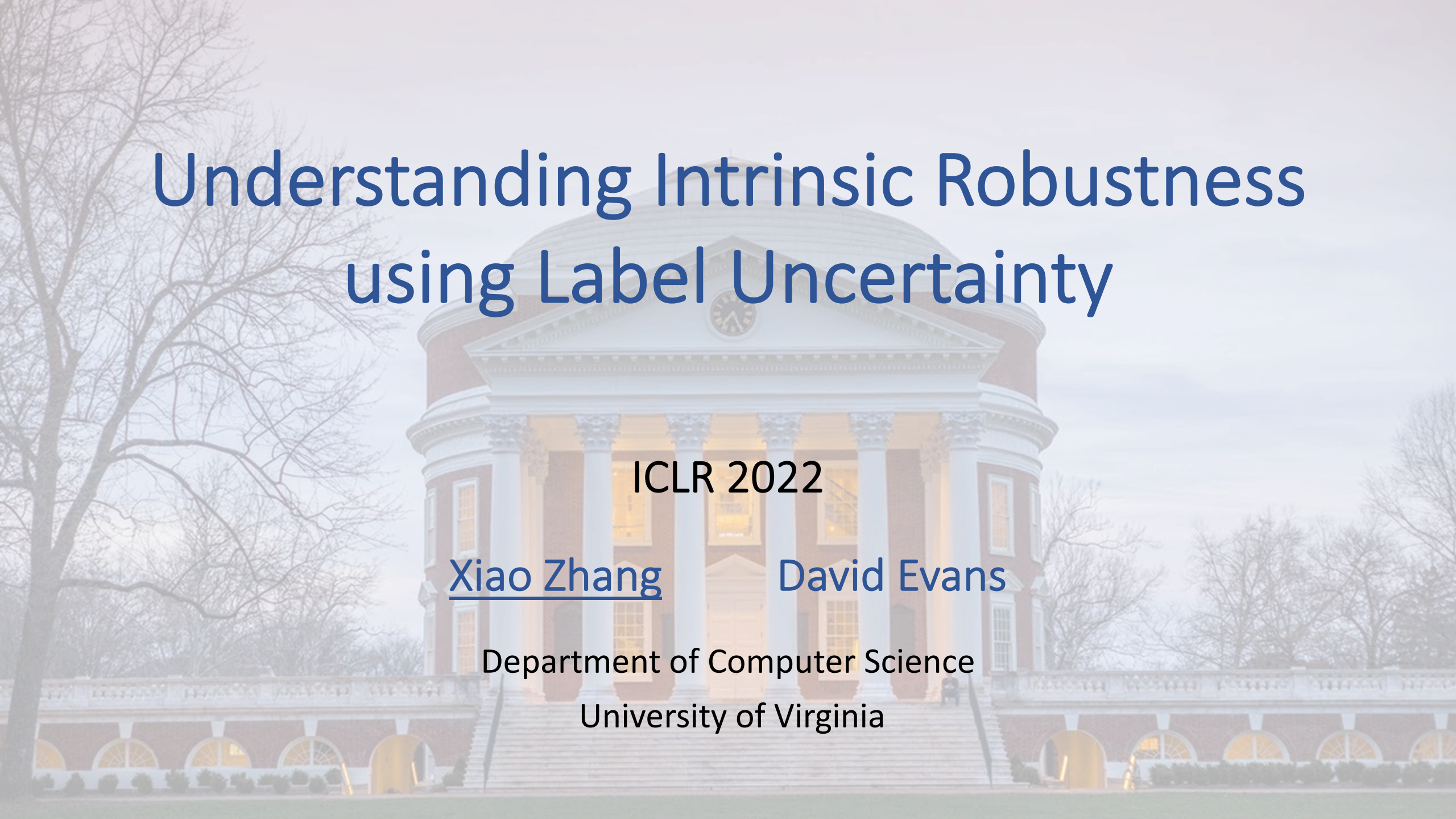
Understanding Intrinsic Robustness using Label Uncertainty



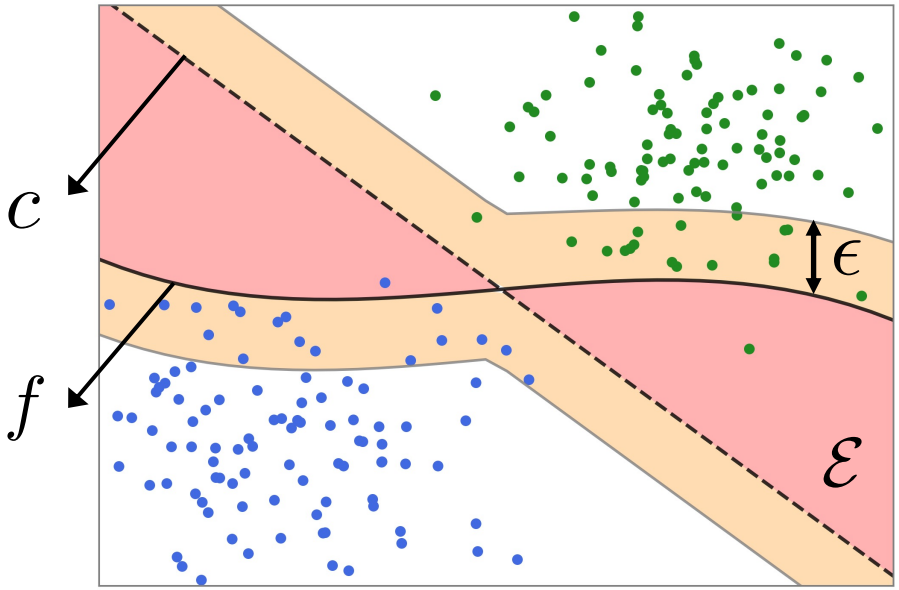
ICLR 2022

Xiao Zhang David Evans

Department of Computer Science

University of Virginia

Intrinsic Robustness [Mahloujifar+, 2019]



$$\mathrm{Risk}(f) = \mu(\mathcal{E})$$

$$\mathrm{AdvRob}_\epsilon(f) = 1 - \mu(\mathcal{E}_\epsilon)$$

Intrinsic Robustness ($\mathcal{F}_\alpha$)

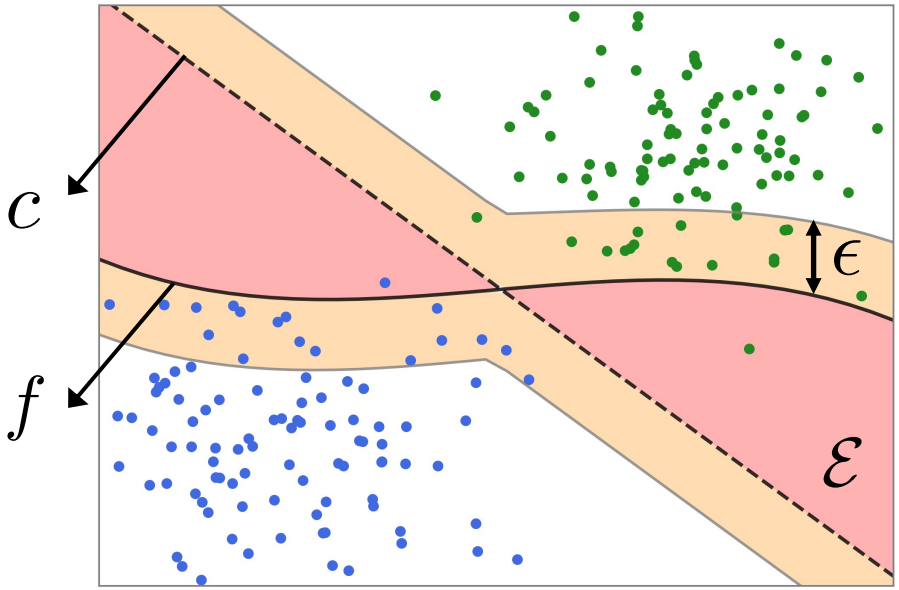
$$\max_f \ \mathrm{AdvRob}_\epsilon(f)$$
$$\text{s.t. } \mathrm{Risk}(f) \geq \alpha$$

↕

Concentration Problem

$$\min_{\mathcal{E}} \ \mu(\mathcal{E}_\epsilon)$$
$$\text{s.t. } \mu(\mathcal{E}) \geq \alpha$$

Previous works ignore labels



$$\mathrm{Risk}(f) = \mu(\mathcal{E})$$

$$\mathrm{AdvRob}_\epsilon(f) = 1 - \mu(\mathcal{E}_\epsilon)$$

Intrinsic Robustness ($\mathcal{F}_\alpha$)

$$\max_f \ \mathrm{AdvRob}_\epsilon(f)$$
$$\text{s.t. } \mathrm{Risk}(f) \geq \alpha$$

Concentration Problem

$$\min_{\mathcal{E}} \ \mu(\mathcal{E}_\epsilon)$$
$$\text{s.t. } \mu(\mathcal{E}) \geq \alpha$$

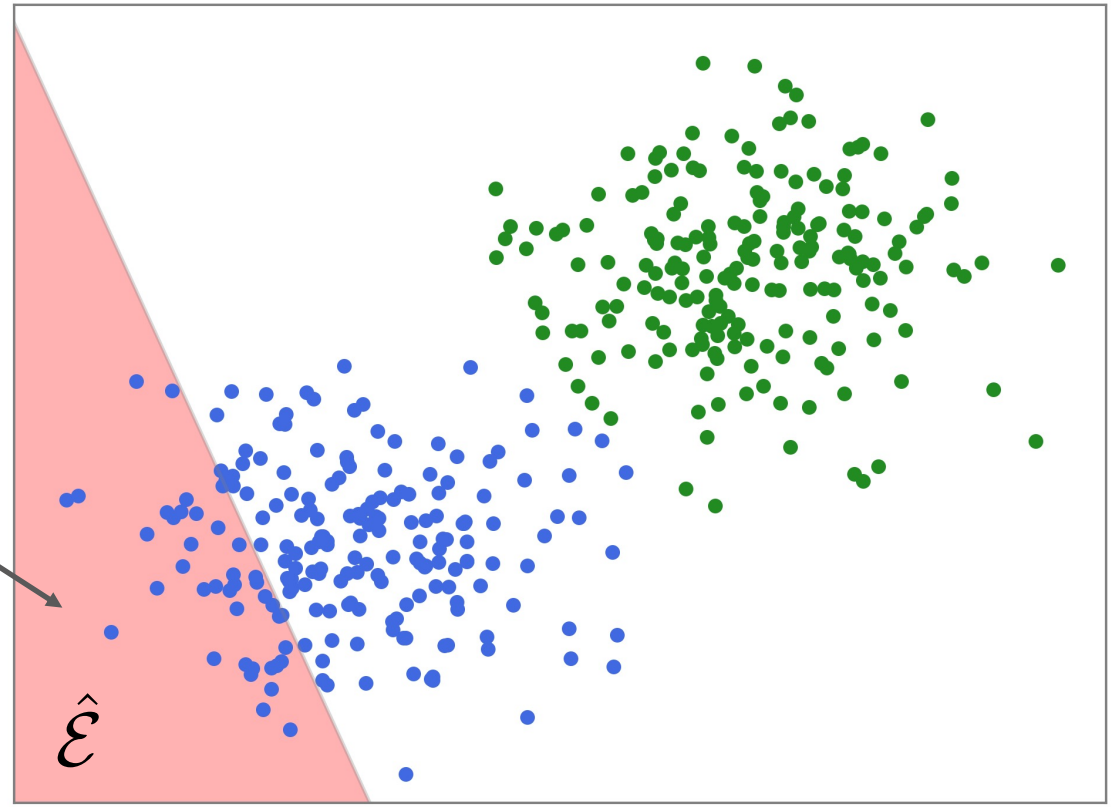
Label assignment does not affect Intrinsic robustness ($\mathcal{F}_\alpha$)

Labels Matter for Intrinsic Robustness

Concentration Problem

$$\min_{\mathcal{E}} \ \mu(\mathcal{E}_\epsilon)$$

$$\text{s.t.} \ \ \mu(\mathcal{E}) \geq \alpha$$

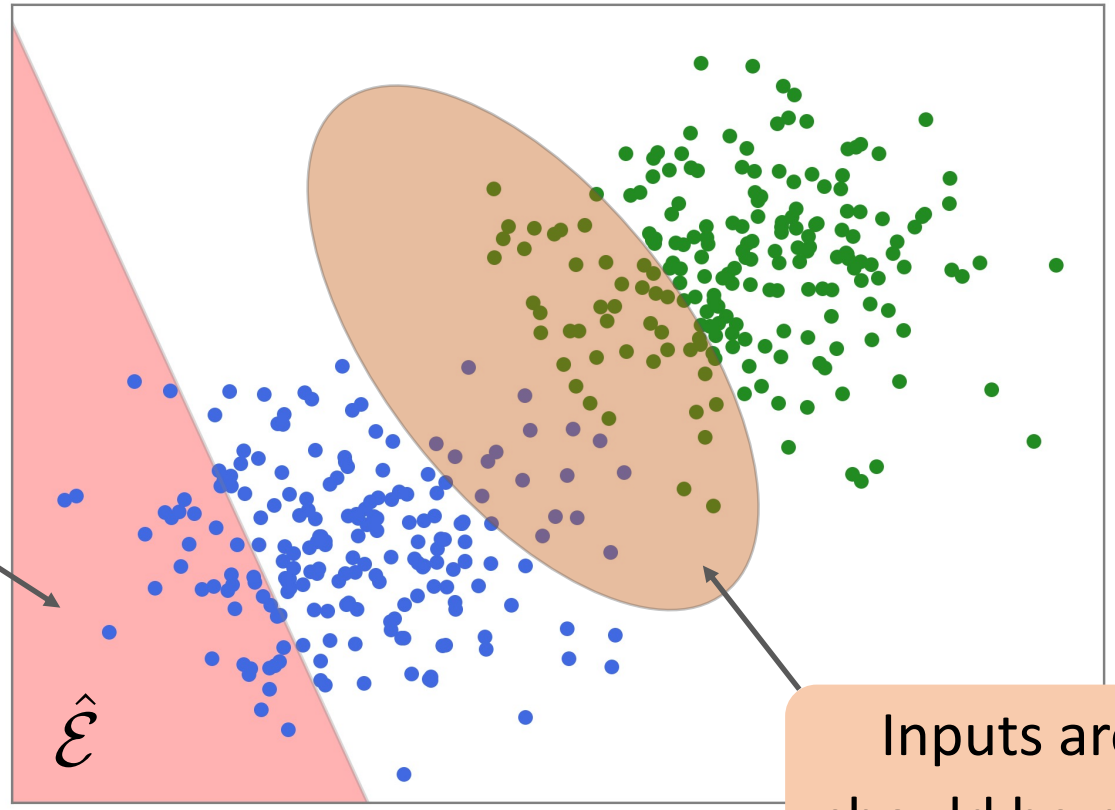


Labels Matter for Intrinsic Robustness

Concentration Problem

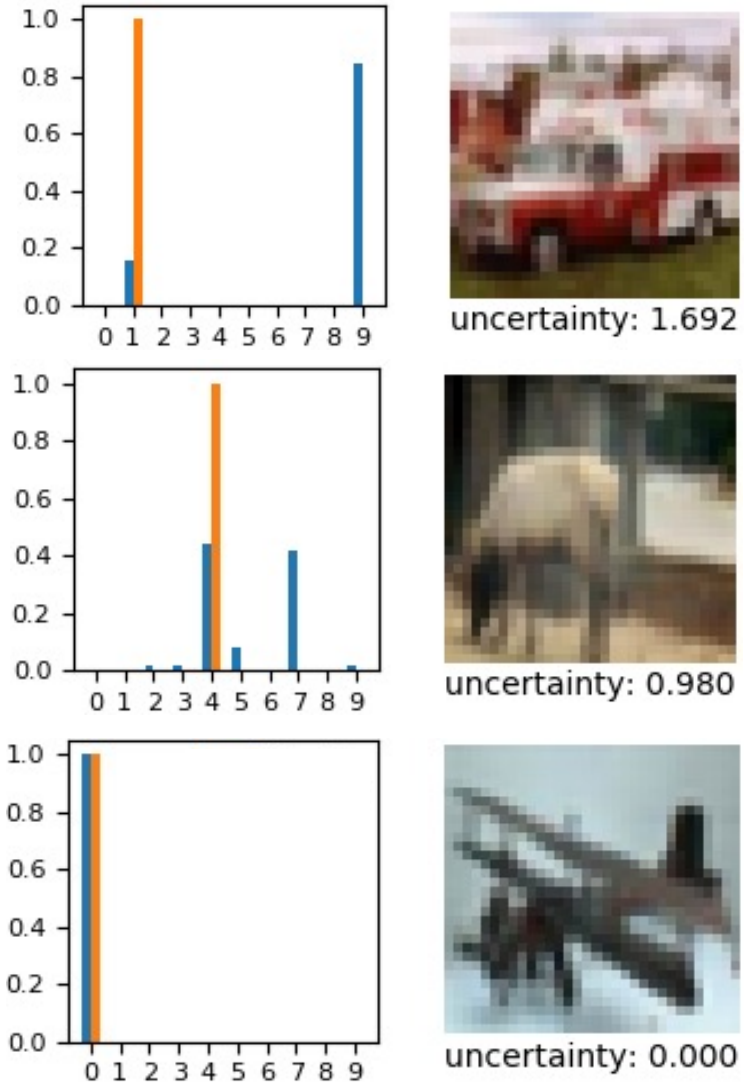
$$\min_{\mathcal{E}} \ \mu(\mathcal{E}_\epsilon)$$

$$\text{s.t.} \ \ \mu(\mathcal{E}) \geq \alpha$$



Inputs around the boundary should be more likely to be error

Formalizing Label Uncertainty



$$1 - [\eta(\boldsymbol{x})]_{c(\boldsymbol{x})} + \max_{y' \neq c(\boldsymbol{x})} [\eta(\boldsymbol{x})]_{y'}$$

label distribution **ground-truth classifier**

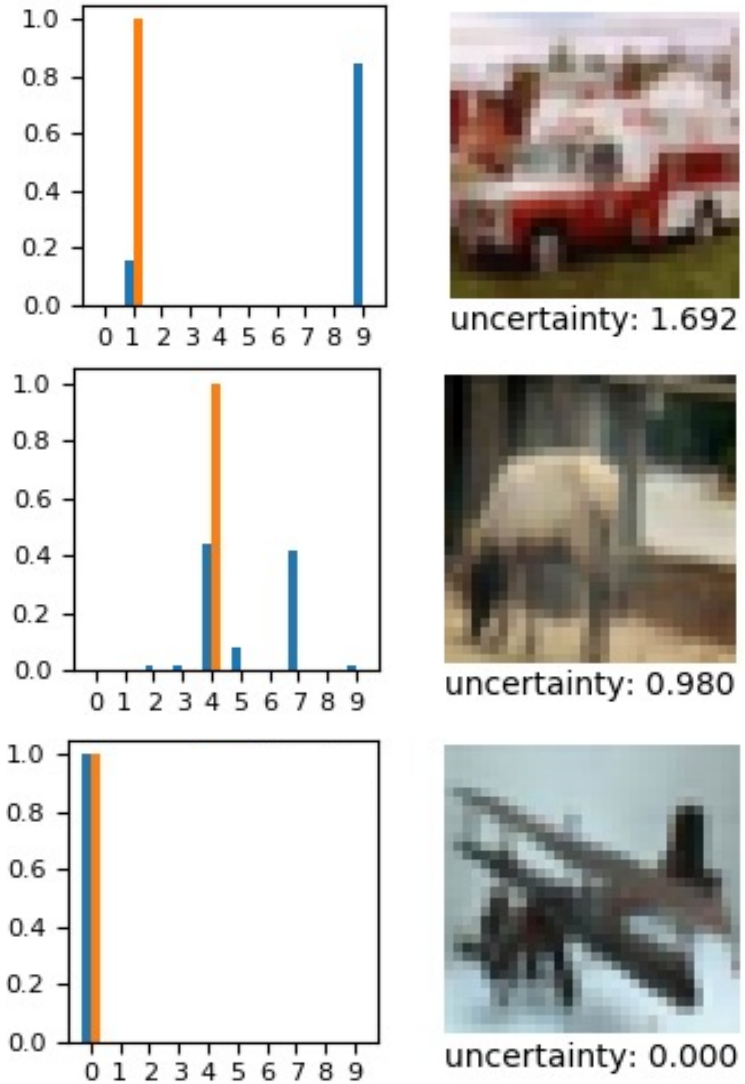
Capture data intrinsic uncertainty with respect to label assignment

original CIFAR-10 label

CIFAR-10H soft human labels [Peterson+, 2019]

0-airplane, 1-automobile, 2-bird, 3-cat, 4-deer, 5-dog, 6-frog, 7-horse, 8-ship, 9-truck

Formalizing Label Uncertainty



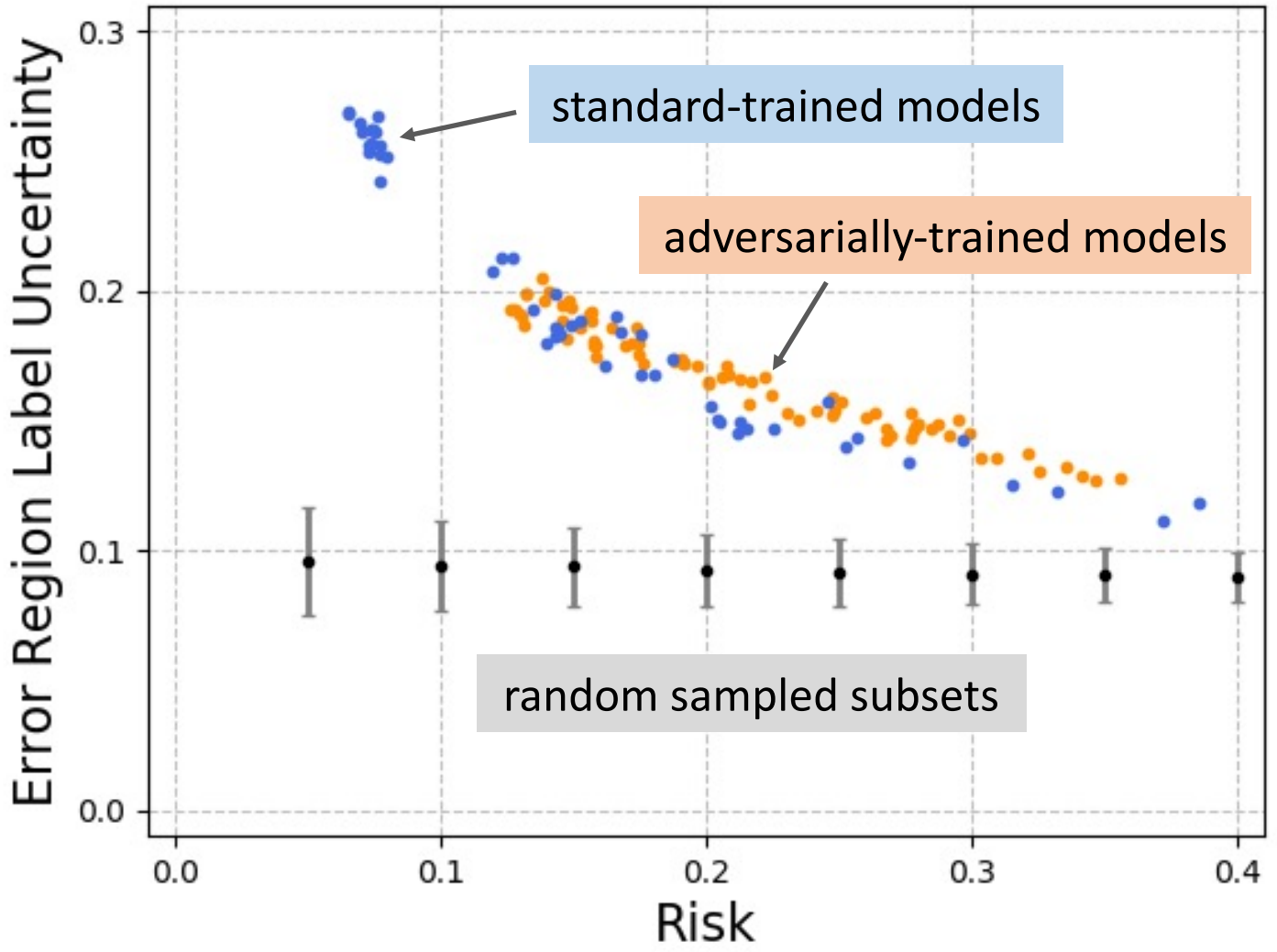
$$1 - [\eta(\boldsymbol{x})]_{c(\boldsymbol{x})} + \max_{y' \neq c(\boldsymbol{x})} [\eta(\boldsymbol{x})]_{y'}$$

label distribution **ground-truth classifier**

Capture data intrinsic uncertainty with respect to label assignment

Do errors of machine learning classifiers also have high label uncertainty?

Most Errors in High-Uncertainty Regions



strong correlation

constrain label uncertainty for measuring concentration

Concentration with label uncertainty

Intrinsic Robustness ($\mathcal{F}_\alpha$) ⟷ Standard Concentration

↓ ↓

Intrinsic Robustness ($\mathcal{F}_{\alpha,\gamma}$) ⟷ Concentration with label constraint

$$\max_f \ \mathrm{AdvRob}_\epsilon(f)$$

$$\text{s.t. } \mathrm{Risk}(f) \geq \alpha$$

$$\boxed{\mathrm{LU}(\mathcal{E}(f)) \geq \gamma}$$

classifiers with large error region label uncertainty

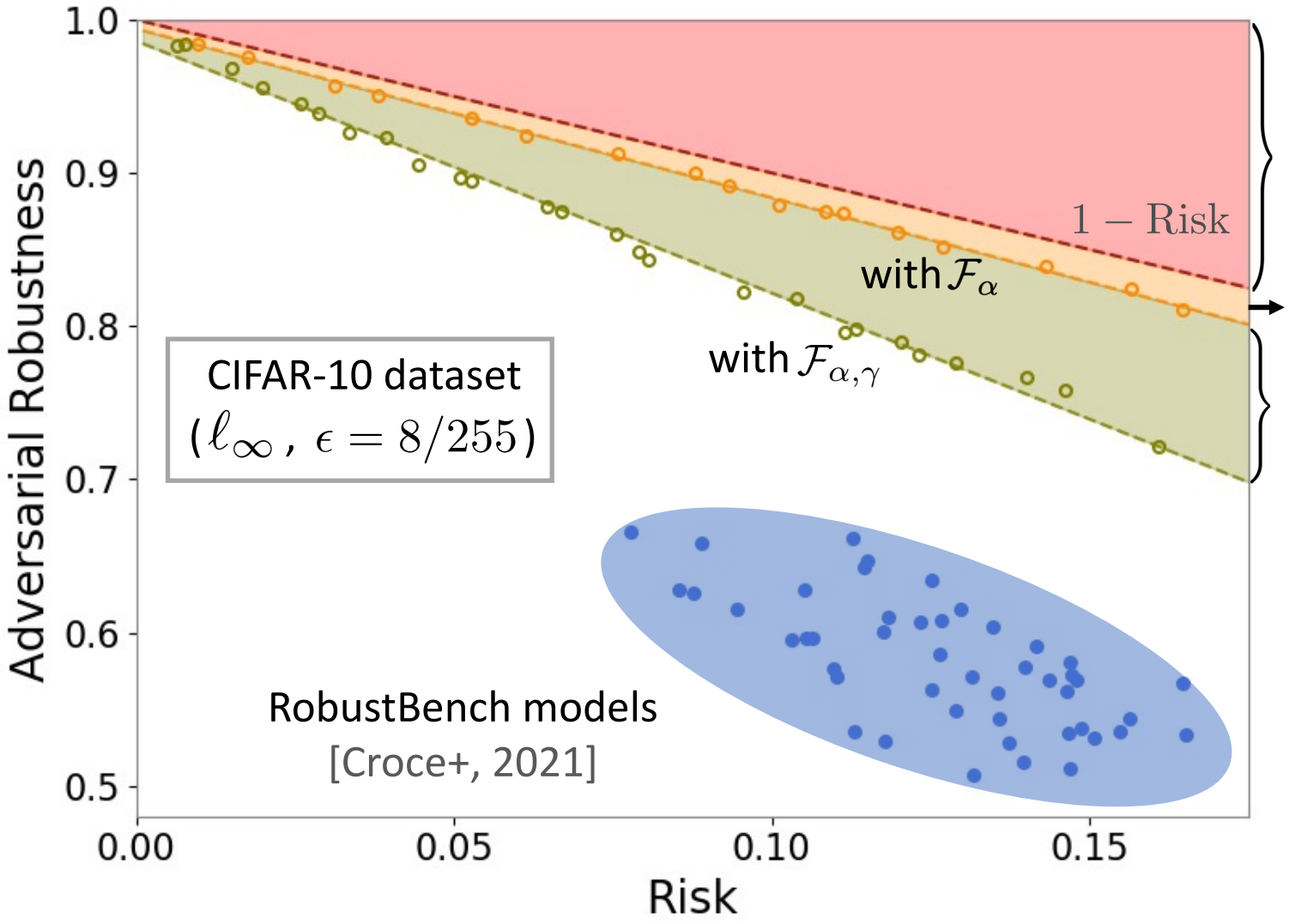
$$\min_{\mathcal{E}} \ \mu(\mathcal{E}_\epsilon)$$

$$\text{s.t. } \mu(\mathcal{E}) \geq \alpha$$

$$\boxed{\mathrm{LU}(\mathcal{E}) \geq \gamma}$$

subsets with high label uncertainty

Understanding the Cause of Vulnerability



Cause of Vulnerability

imperfect risk

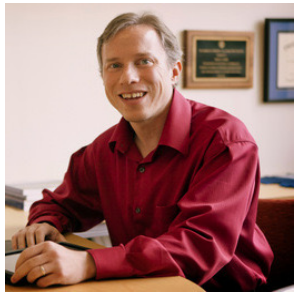
concentration

uncertain inputs

more fundamental!

Thank you for listening!

Understanding Intrinsic Robustness using Label Uncertainty



Summary: Standard concentration is not sufficient; the label matters.

Xiao Zhang

David Evans

Email: shawn@virginia.edu

UNIVERSITY of VIRGINIA

Code Repo: github.com/xiaozhanguva/intrinsic_rob_lu