Probabilistically Robust Learning

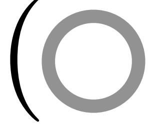
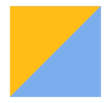
Balancing Average- and Worst-case Performance

Alex Robey, Luiz F. O. Chamon, George J. Pappas, Hamed Hassani

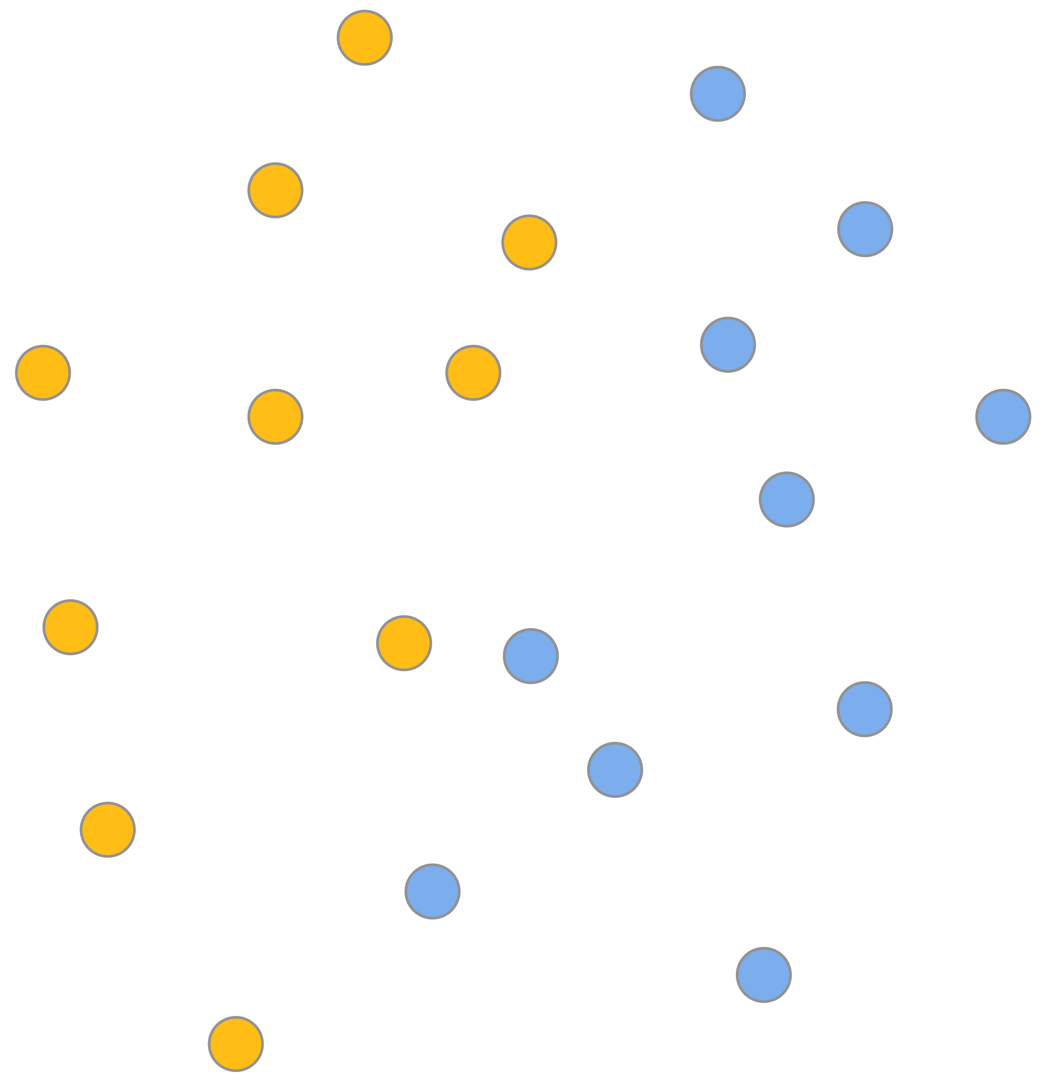
How should we learn from data?

How should we learn from data?

$(x, y) = ($  $,$  $) \sim \mathbb{P}(X, Y)$

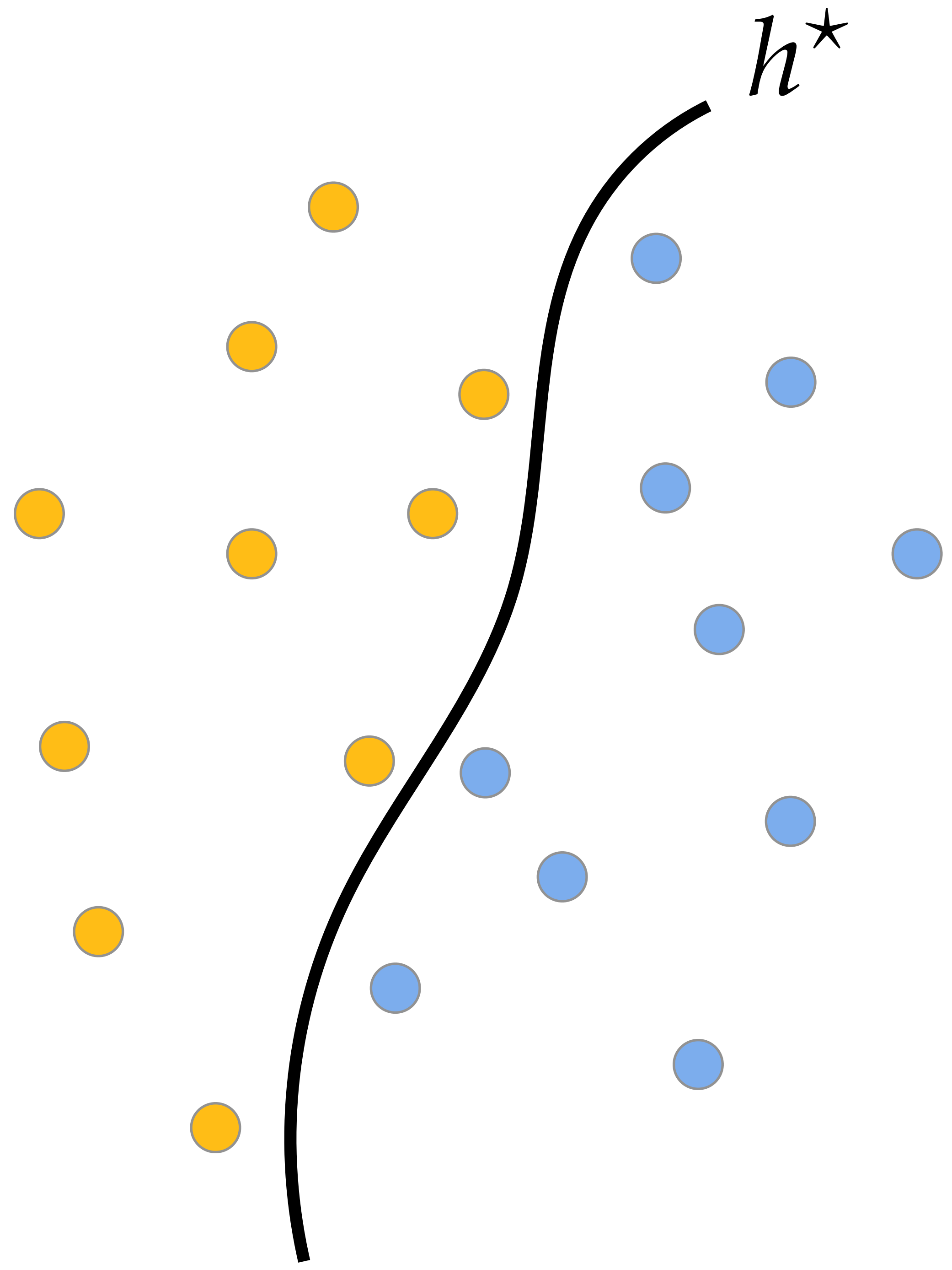
How should we learn from data?

$$(x, y) = (\bigcirc, \blacksquare) \sim \mathbb{P}(X, Y)$$



How should we learn from data?

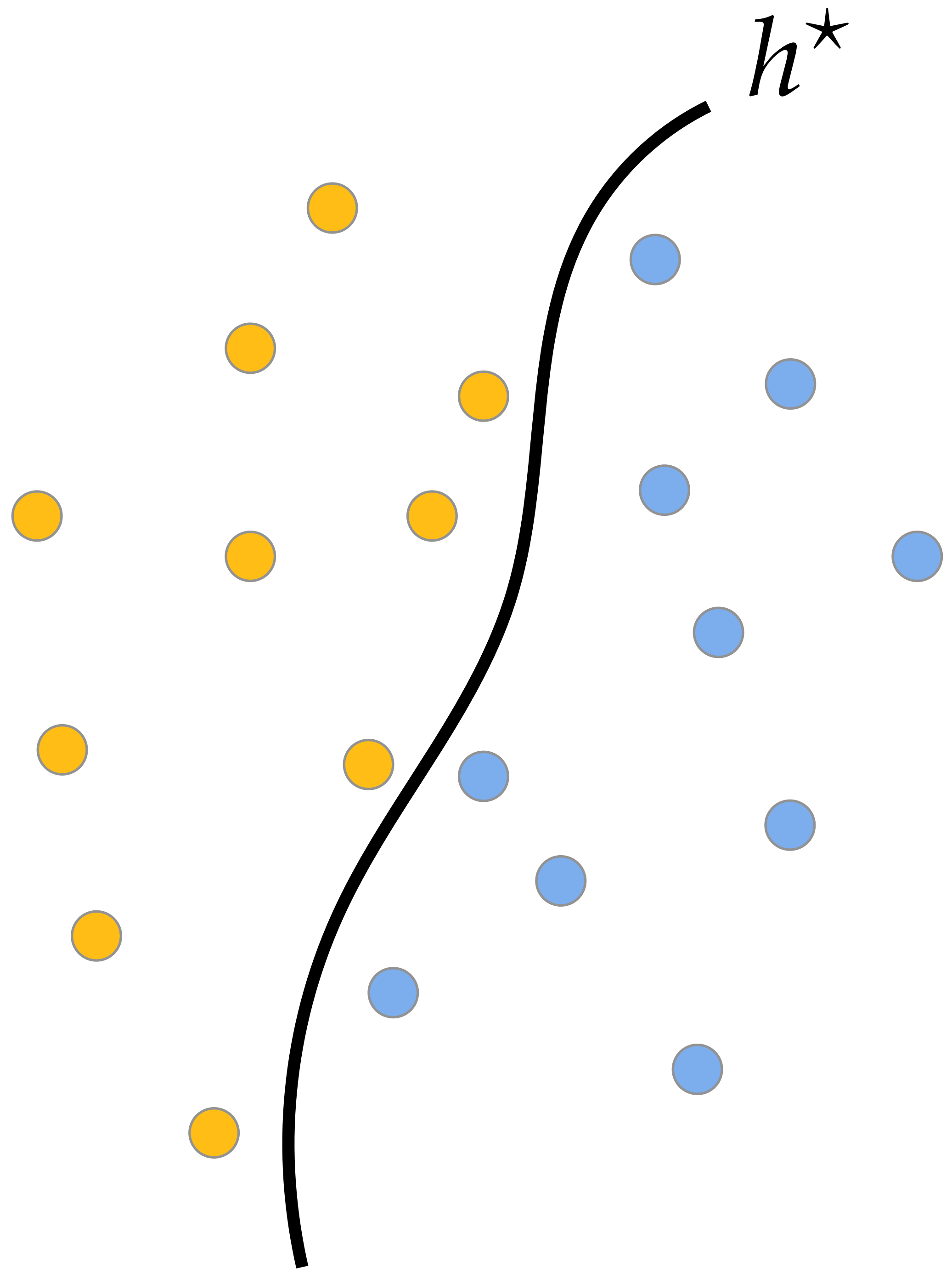
$$(x, y) = (\bigcirc, \blacksquare) \sim \mathbb{P}(X, Y)$$



$$\min_{h \in \mathcal{H}} \mathrm{SR}(h) \triangleq \mathbb{E}_{(x,y)} \left[\ell(h(x), y) \right]$$

How should we learn from data?

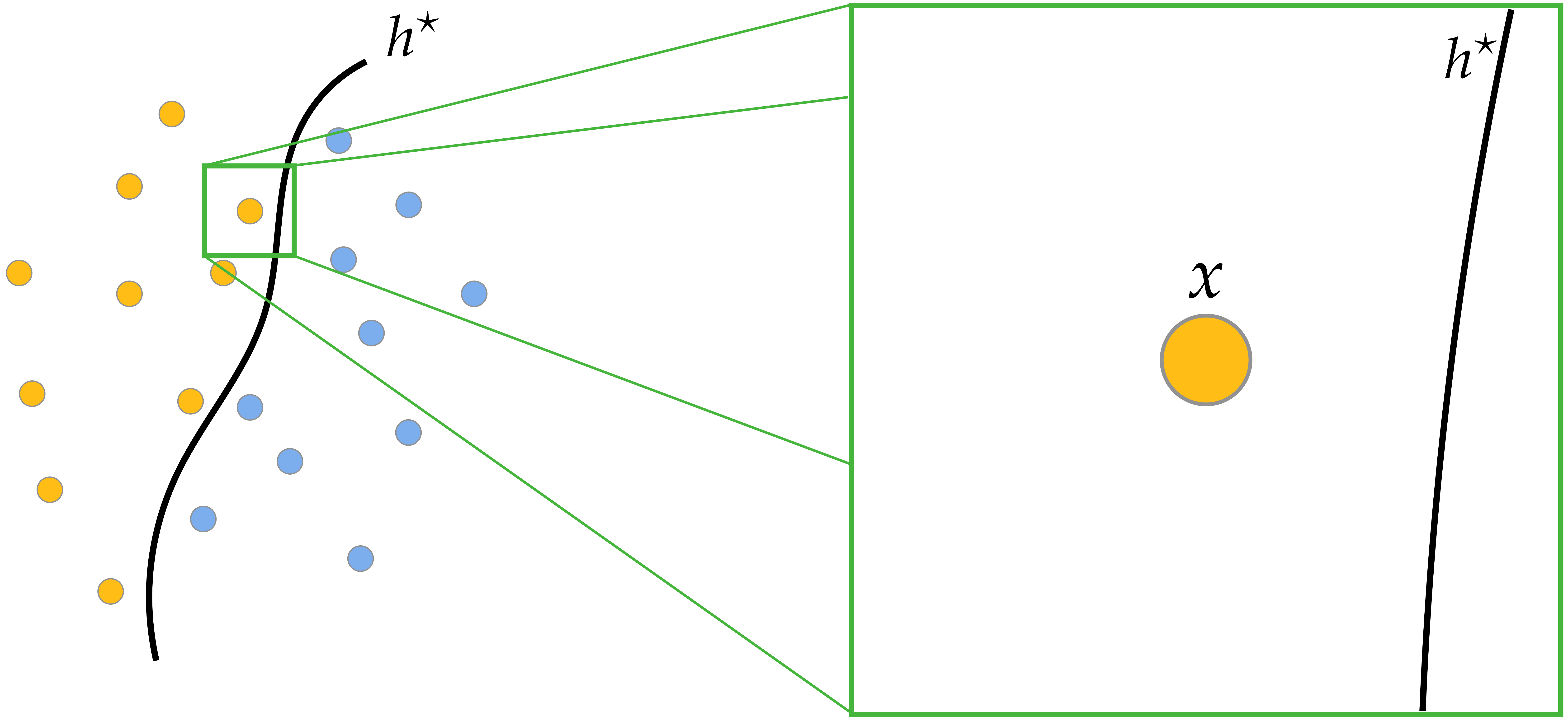
$$(x, y) = (\bigcirc, \blacksquare) \sim \mathbb{P}(X, Y)$$



$$\min_{h \in \mathcal{H}} \mathrm{SR}(h) \triangleq \mathbb{E}_{(x,y)} \left[\ell(h(x), y) \right]$$

How should we learn from data?

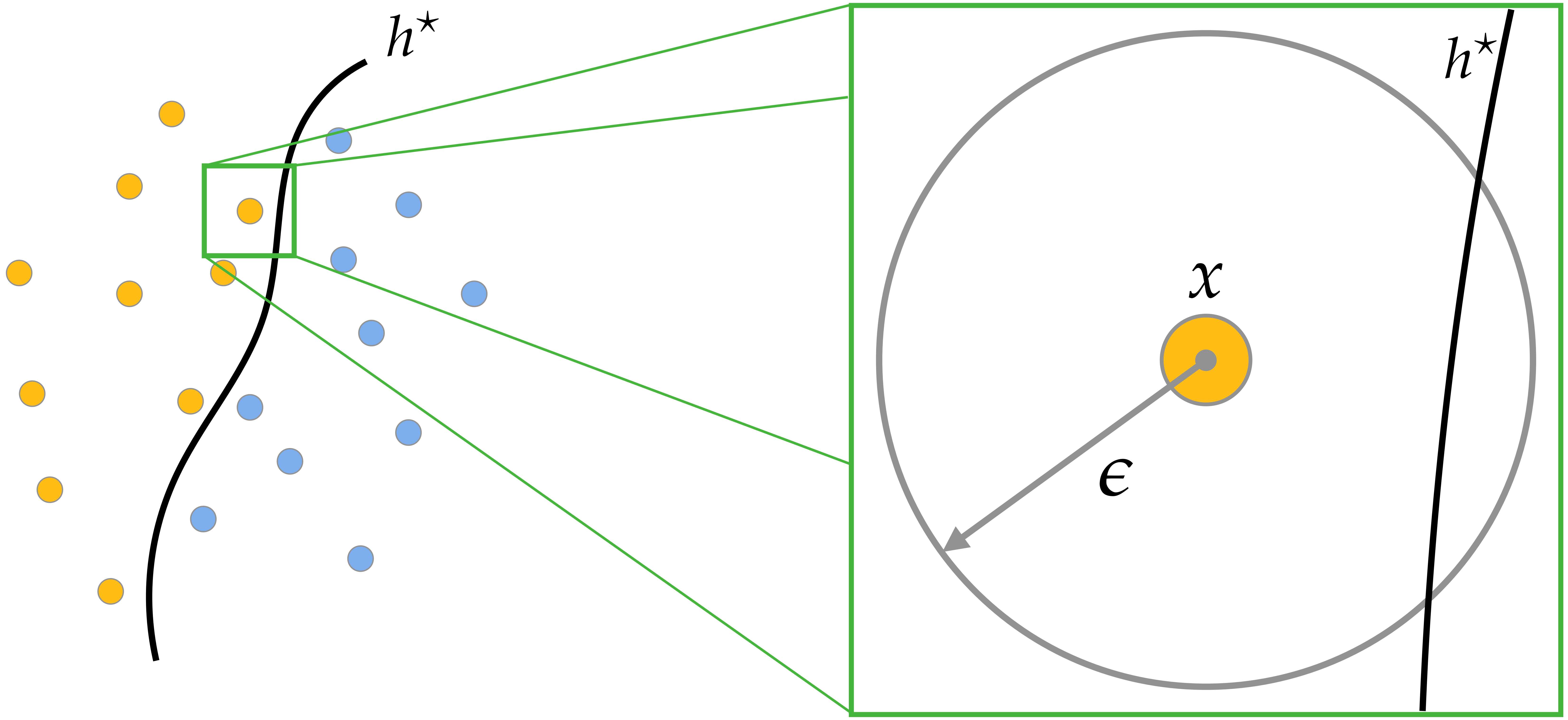
$$(x, y) = (\bigcirc, \blacksquare) \sim \mathbb{P}(X, Y)$$



$$\min_{h \in \mathcal{H}} \mathrm{SR}(h) \triangleq \mathbb{E}_{(x,y)} \left[\ell(h(x), y) \right]$$

How should we learn from data?

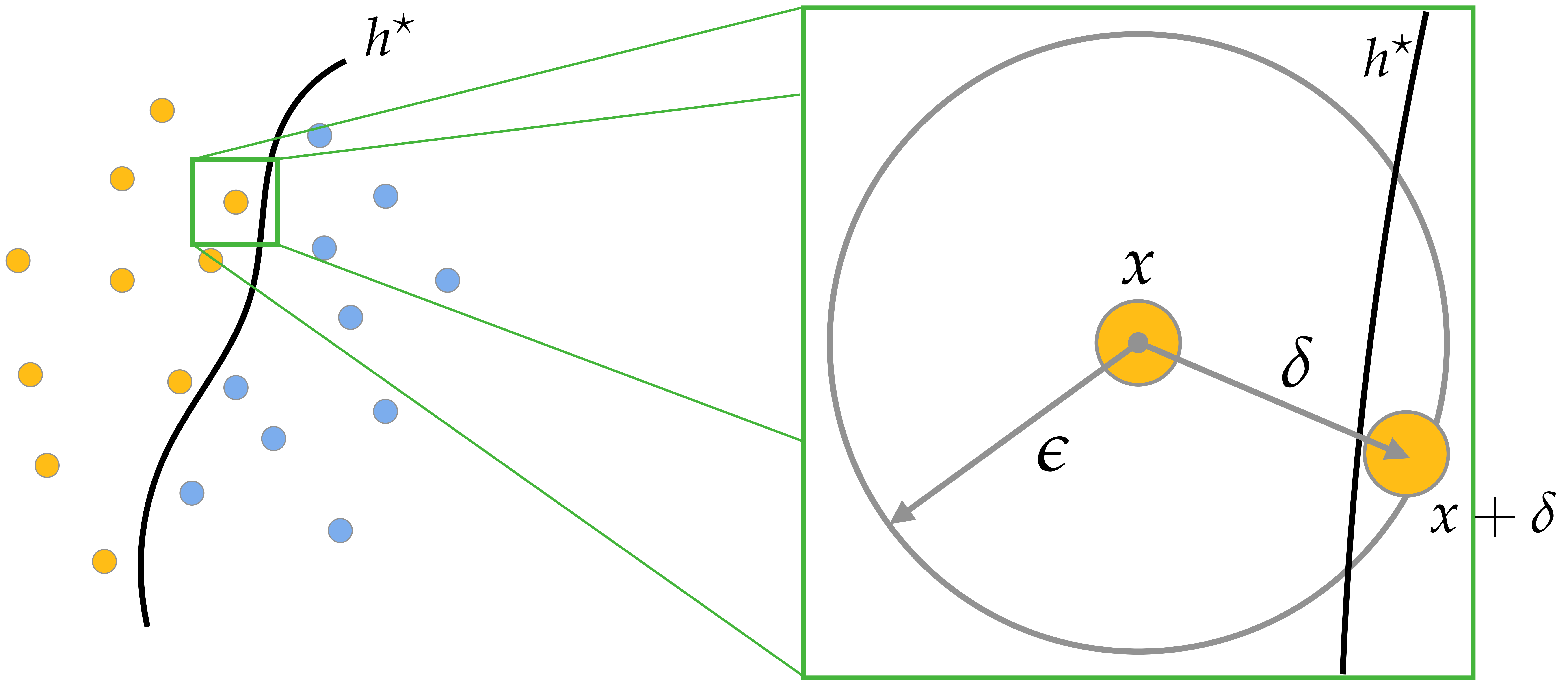
$$(x, y) = (\bigcirc, \blacksquare) \sim \mathbb{P}(X, Y)$$



$$\min_{h \in \mathcal{H}} \mathrm{SR}(h) \triangleq \mathbb{E}_{(x,y)}\left[\ell(h(x), y)\right]$$

How should we learn from data?

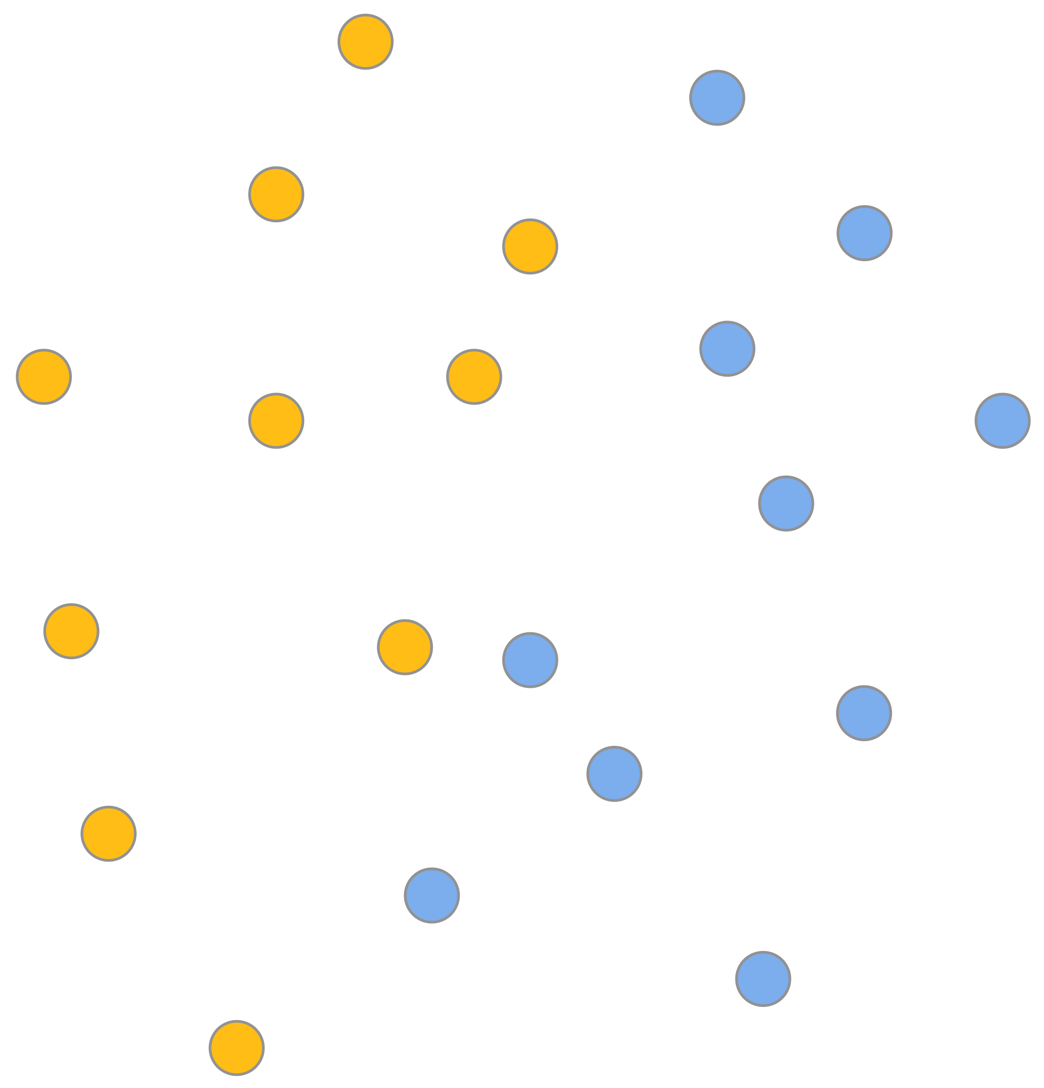
$$(x, y) = (\bigcirc, \blacksquare) \sim \mathbb{P}(X, Y)$$



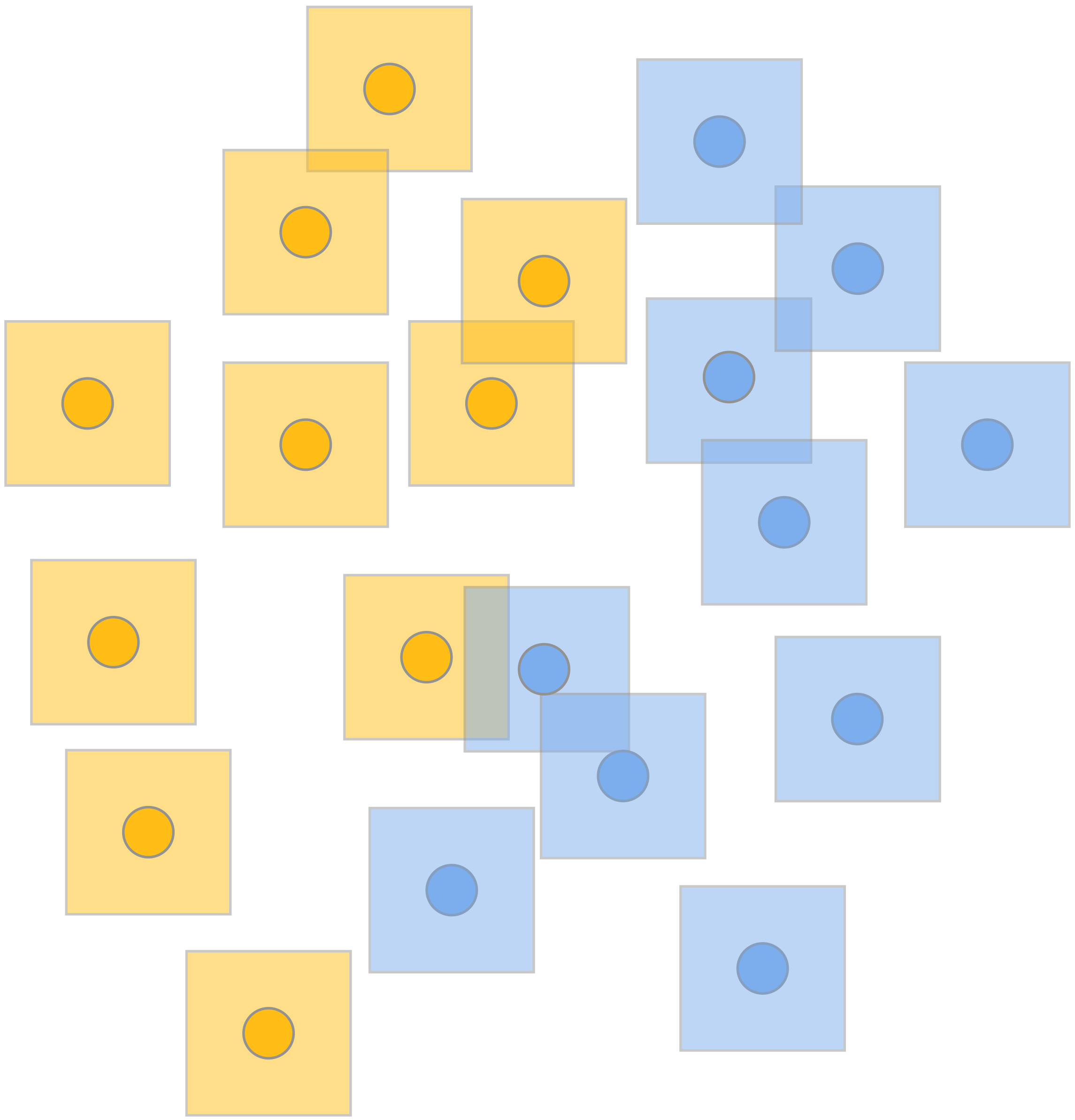
$$\min_{h \in \mathcal{H}} \mathrm{SR}(h) \triangleq \mathbb{E}_{(x,y)} \left[\ell(h(x), y) \right]$$

How should we learn from data?

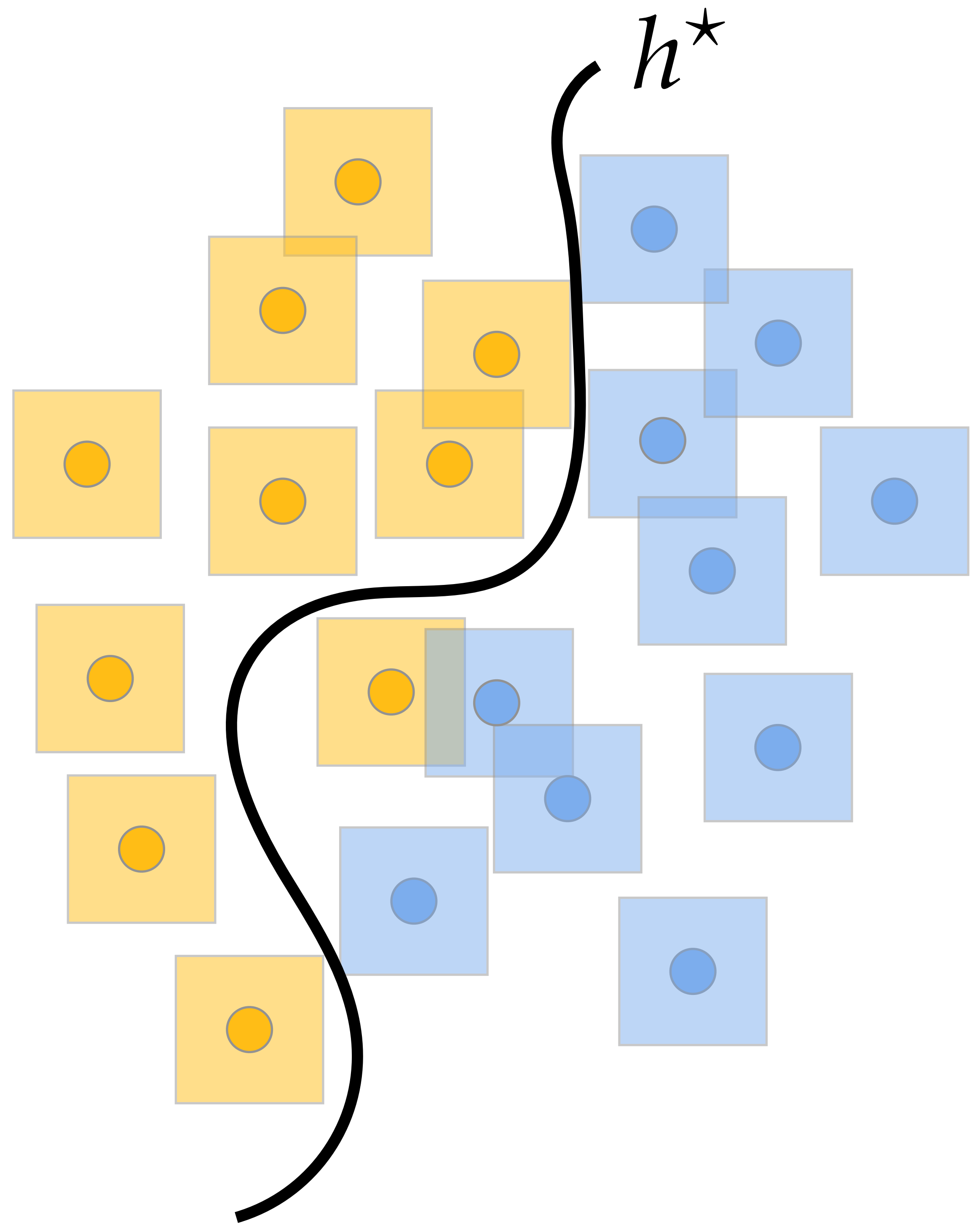
How should we learn from data?



How should we learn from data?

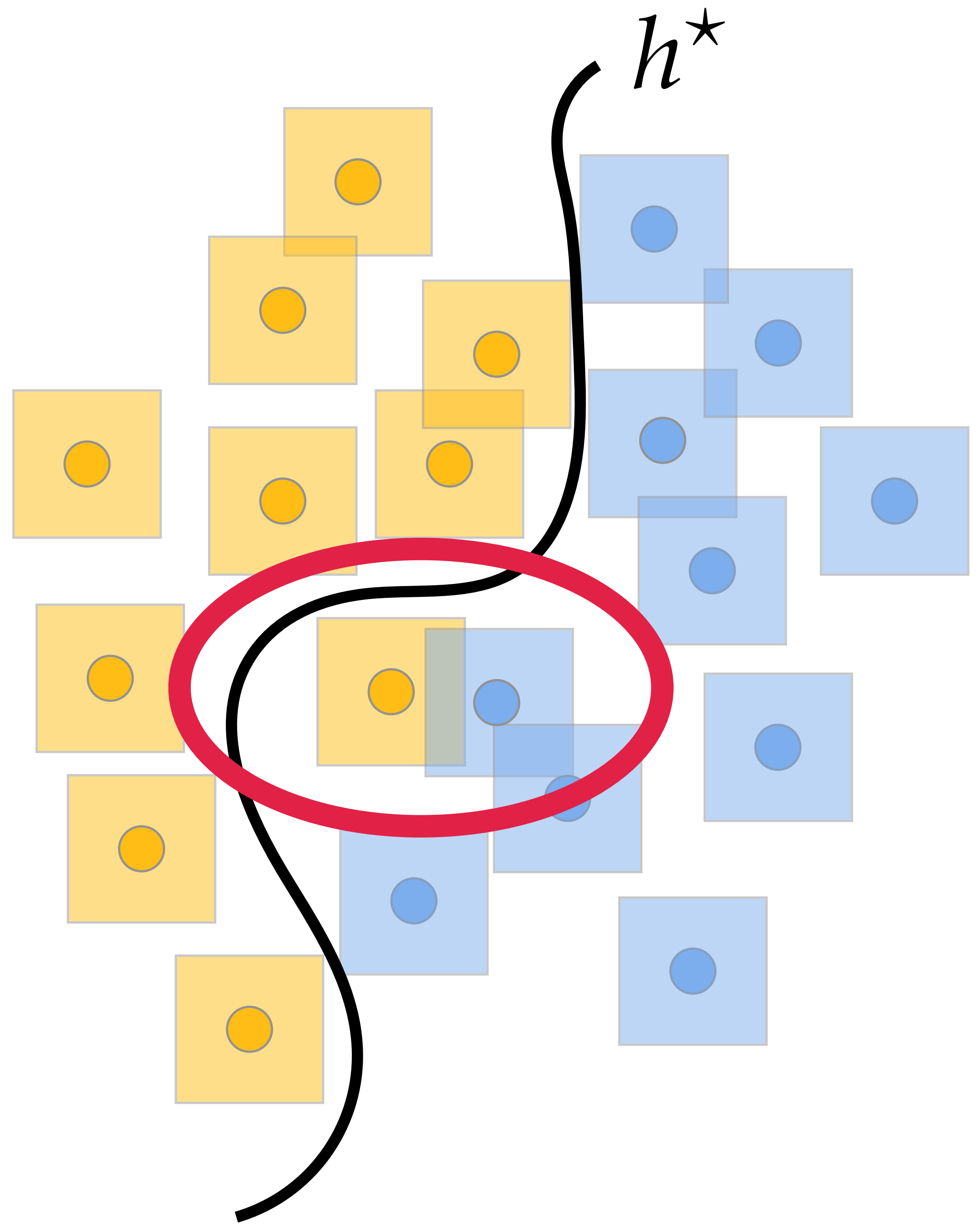


How should we learn from data?



$$\min_{h\in\mathcal{H}} \mathrm{AR}(h) \triangleq \mathbb{E}_{(x,y)}\left[\max_{\delta\in\Delta} \ell(h(x+\delta),y)\right]$$

How should we learn from data?

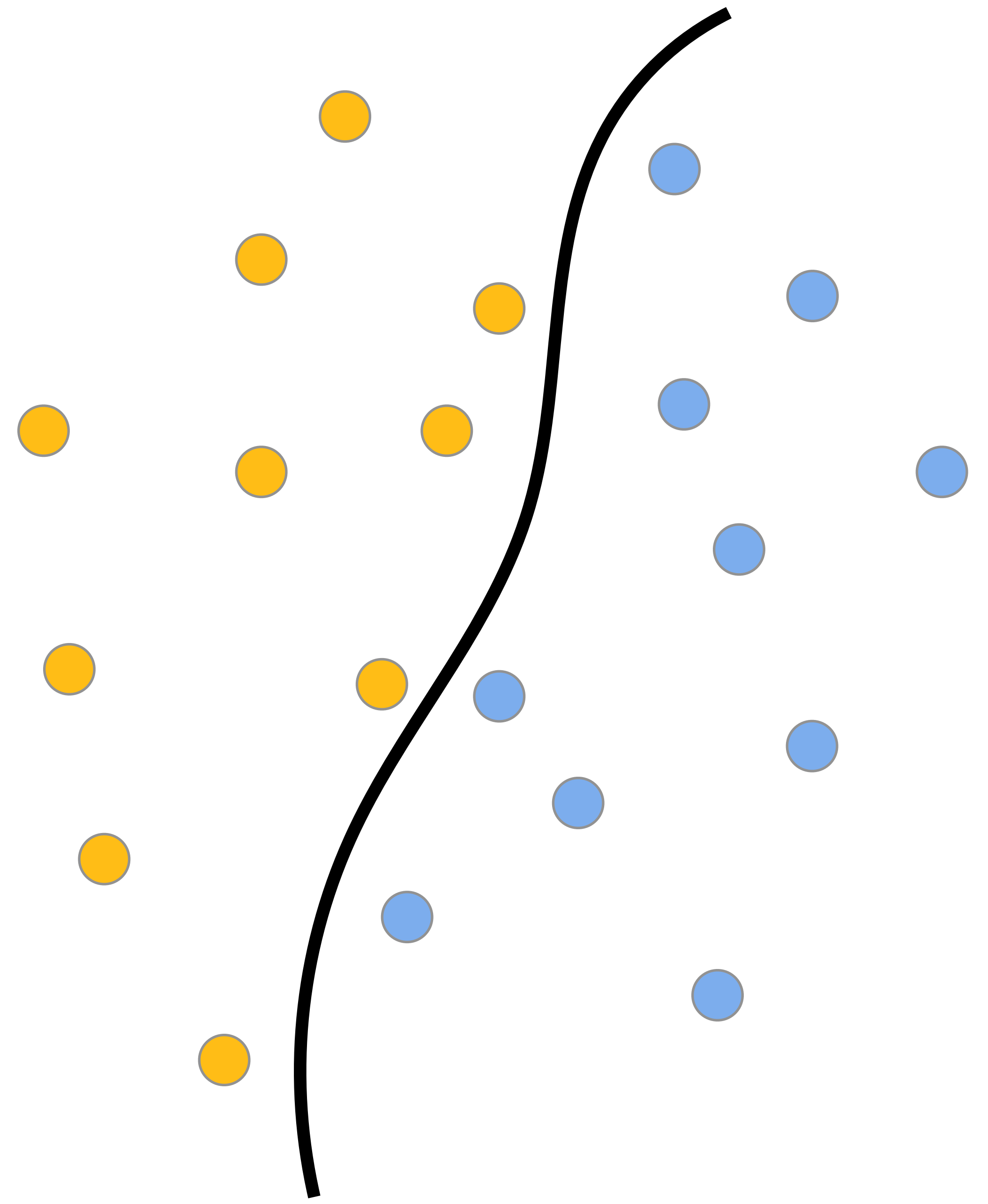


$$\min_{h\in\mathcal{H}} \mathrm{AR}(h) \triangleq \mathbb{E}_{(x,y)}\left[\max_{\delta\in\Delta} \ell(h(x+\delta),y)\right]$$

How should we learn from data?

How should we learn from data?

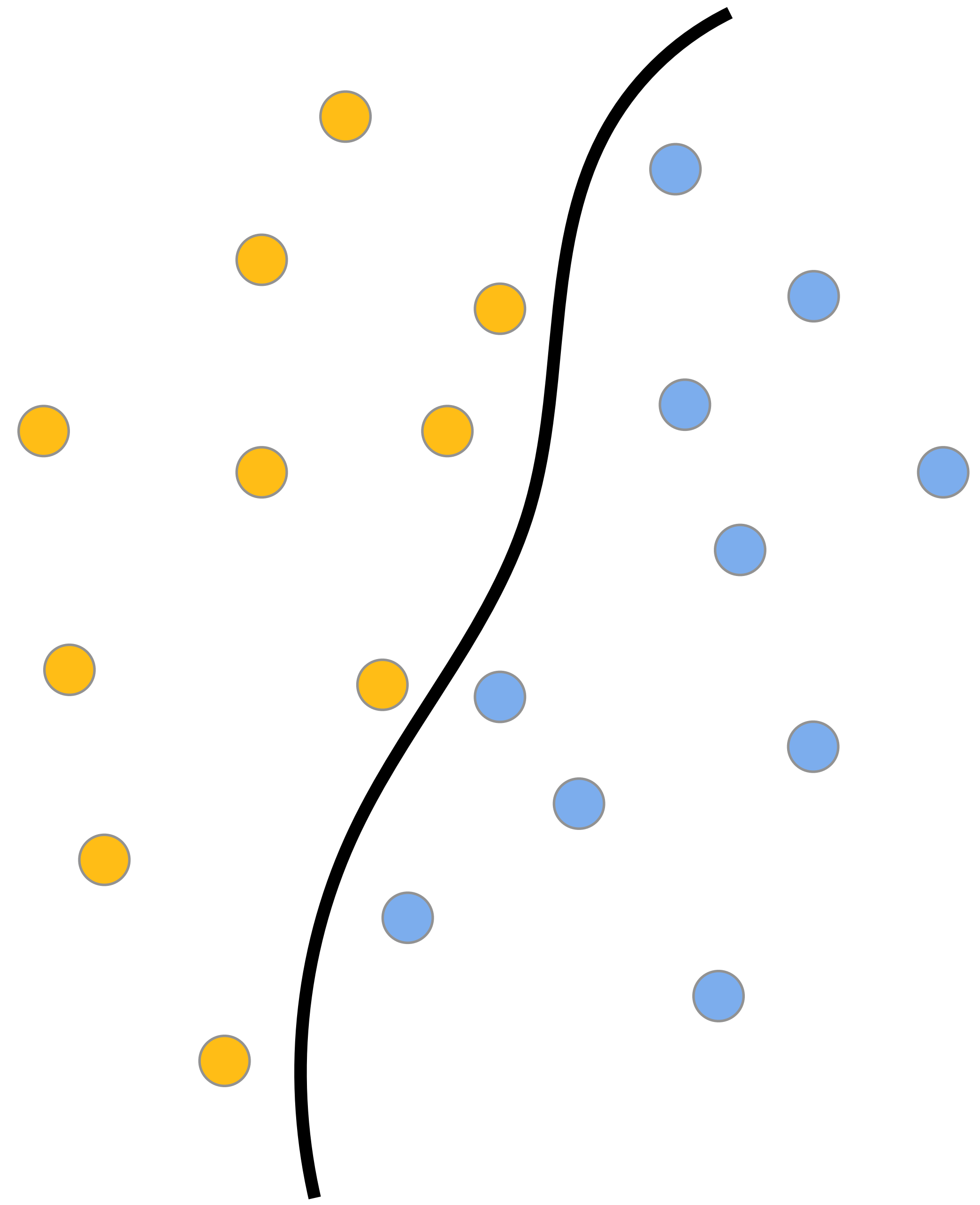
Standard risk minimization



“Accurate, yet brittle”

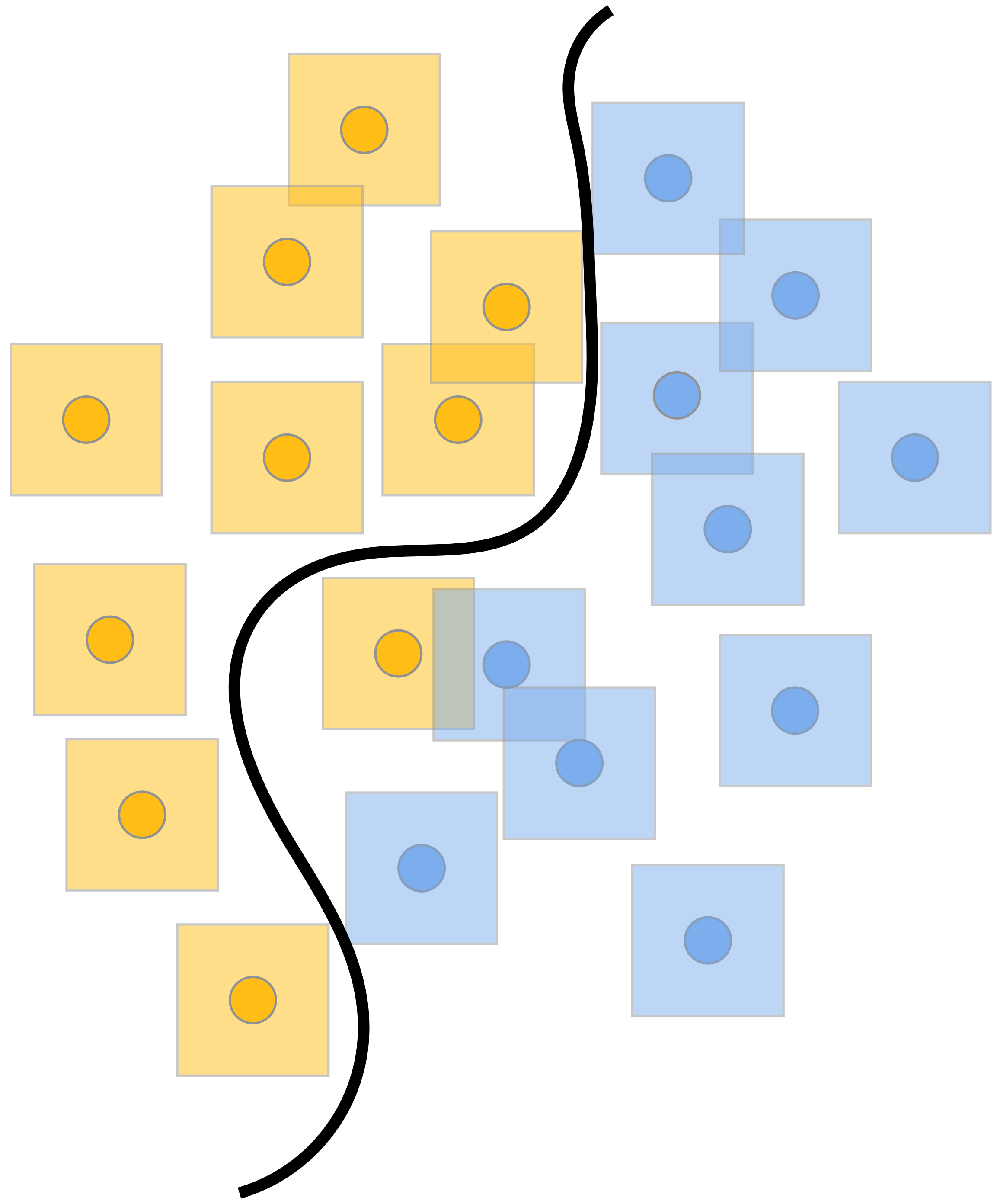
How should we learn from data?

Standard risk minimization



“Accurate, yet brittle”

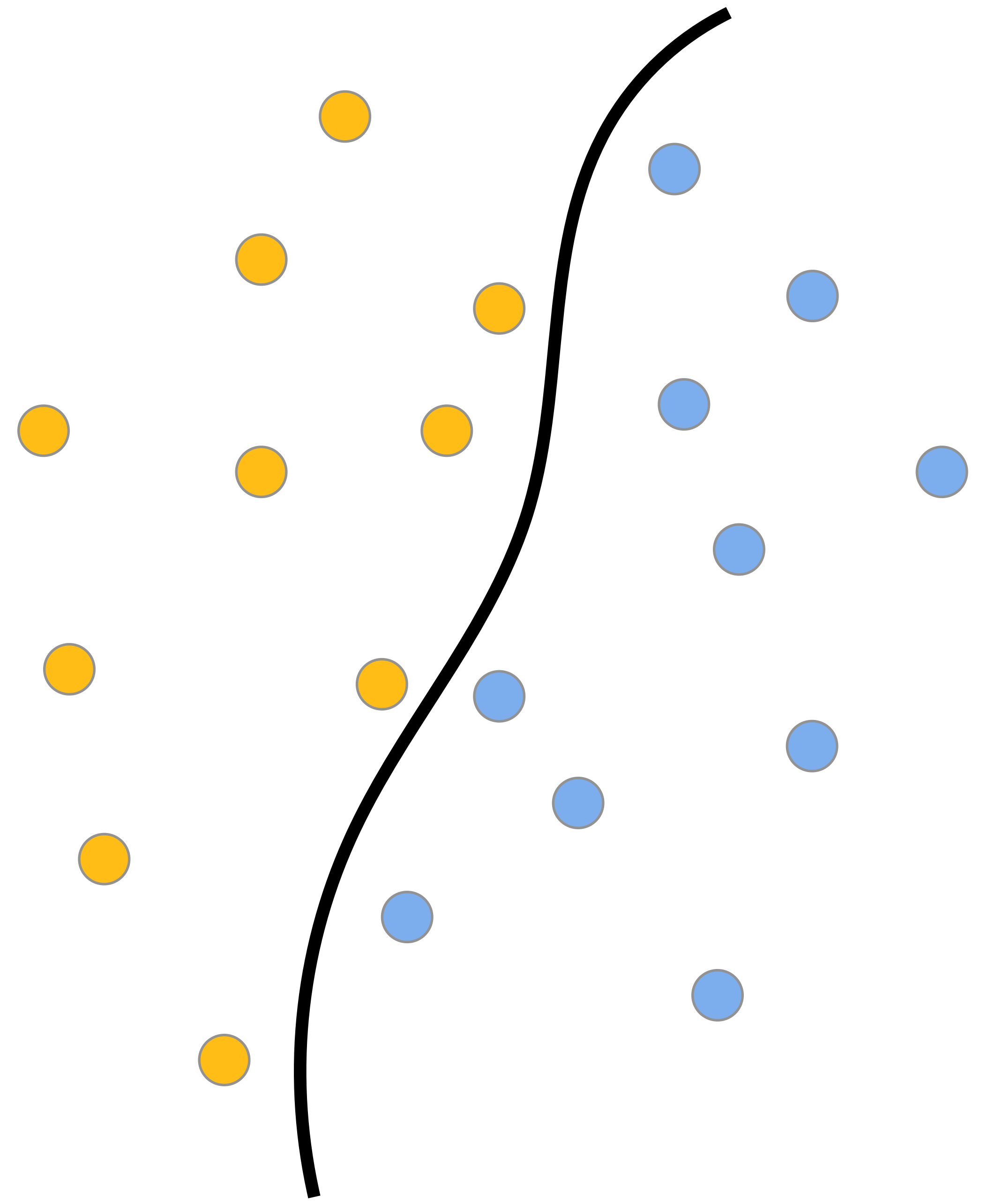
Adversarial training



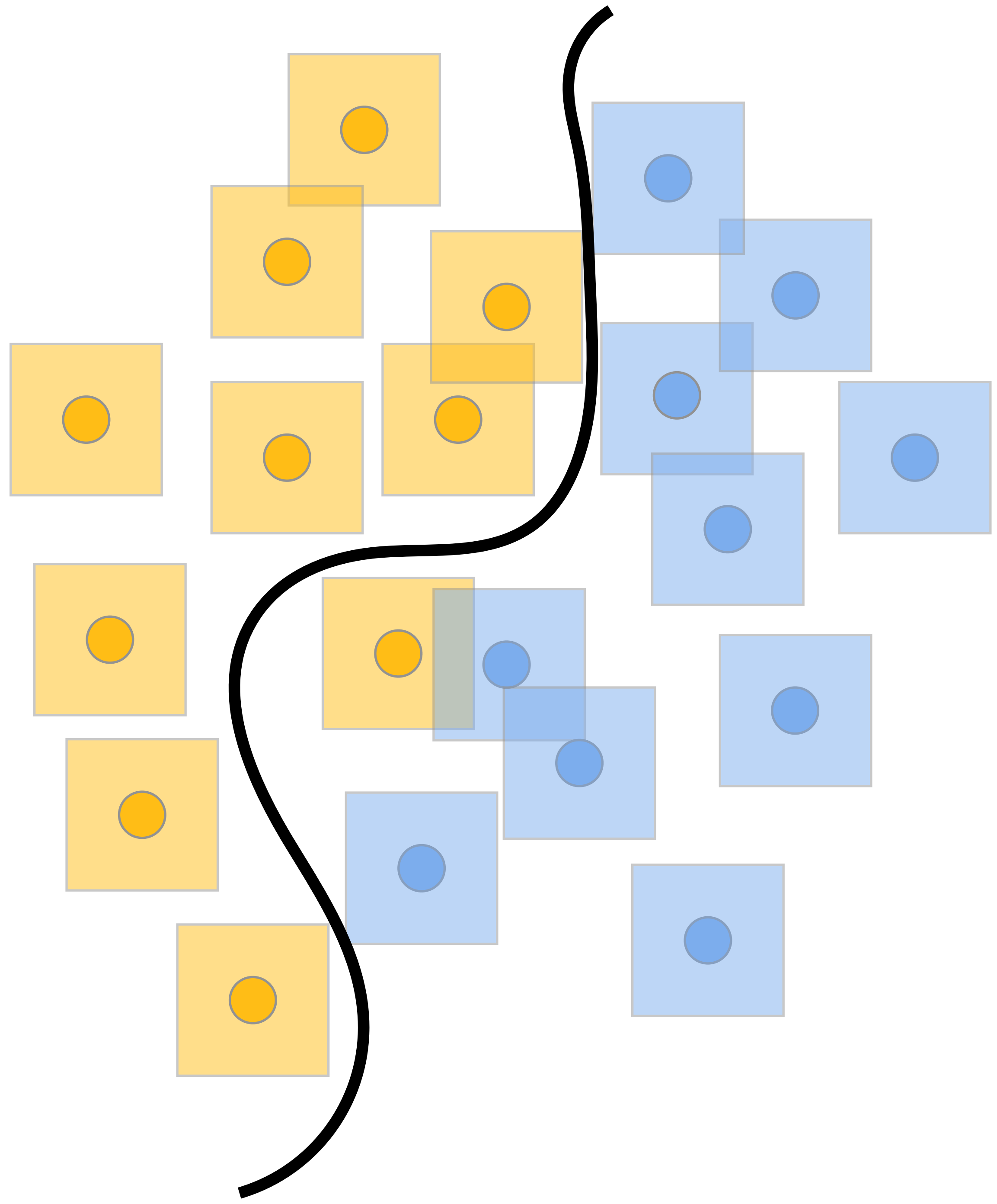
“Robust, yet conservative”

How should we learn from data?

Standard risk minimization



Adversarial training



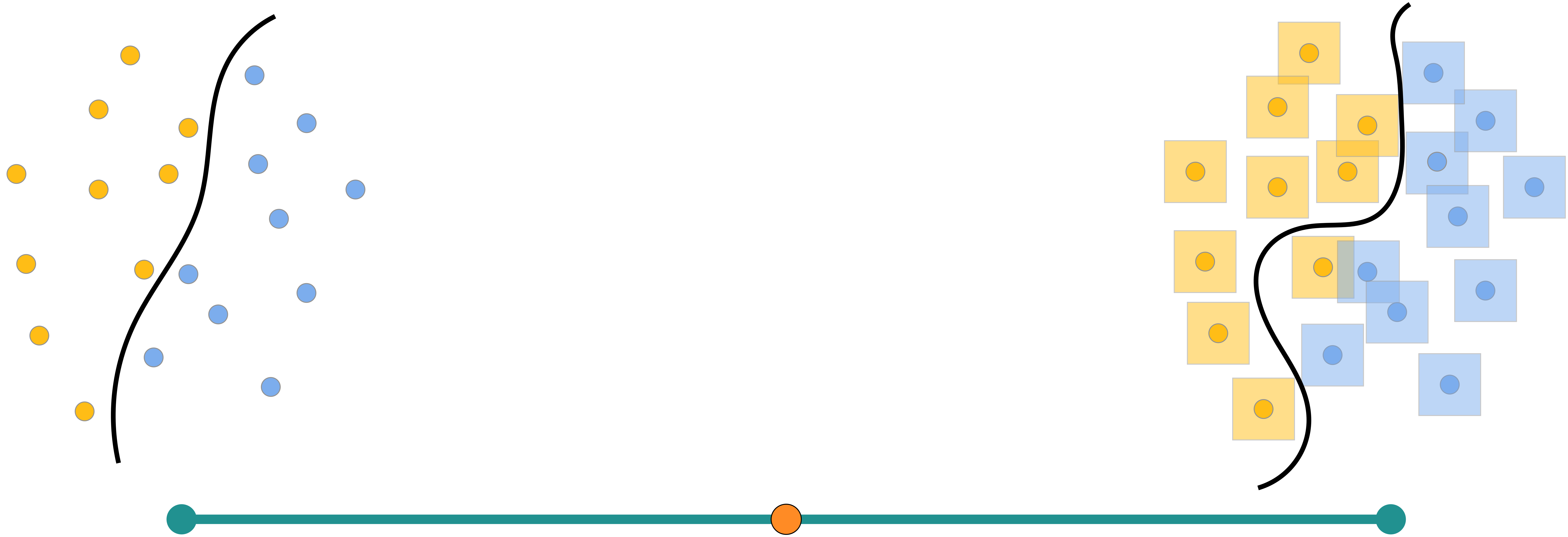
“Accurate, yet brittle”

“Robust, yet conservative”

Question: How can we balance average- and worst-case performance?

Standard risk minimization

Adversarial training

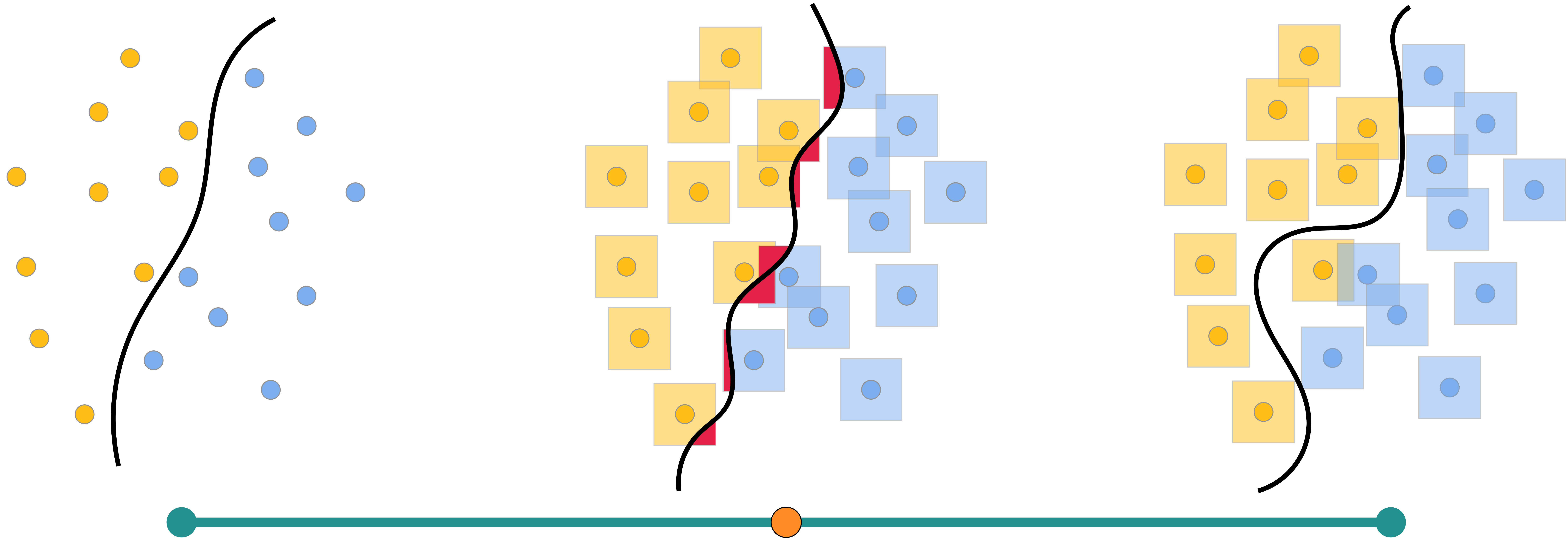


"Accurate, yet brittle"

"Robust, yet conservative"

Question: How can we balance average- and worst-case performance?

Standard risk minimization | PRL | Adversarial training

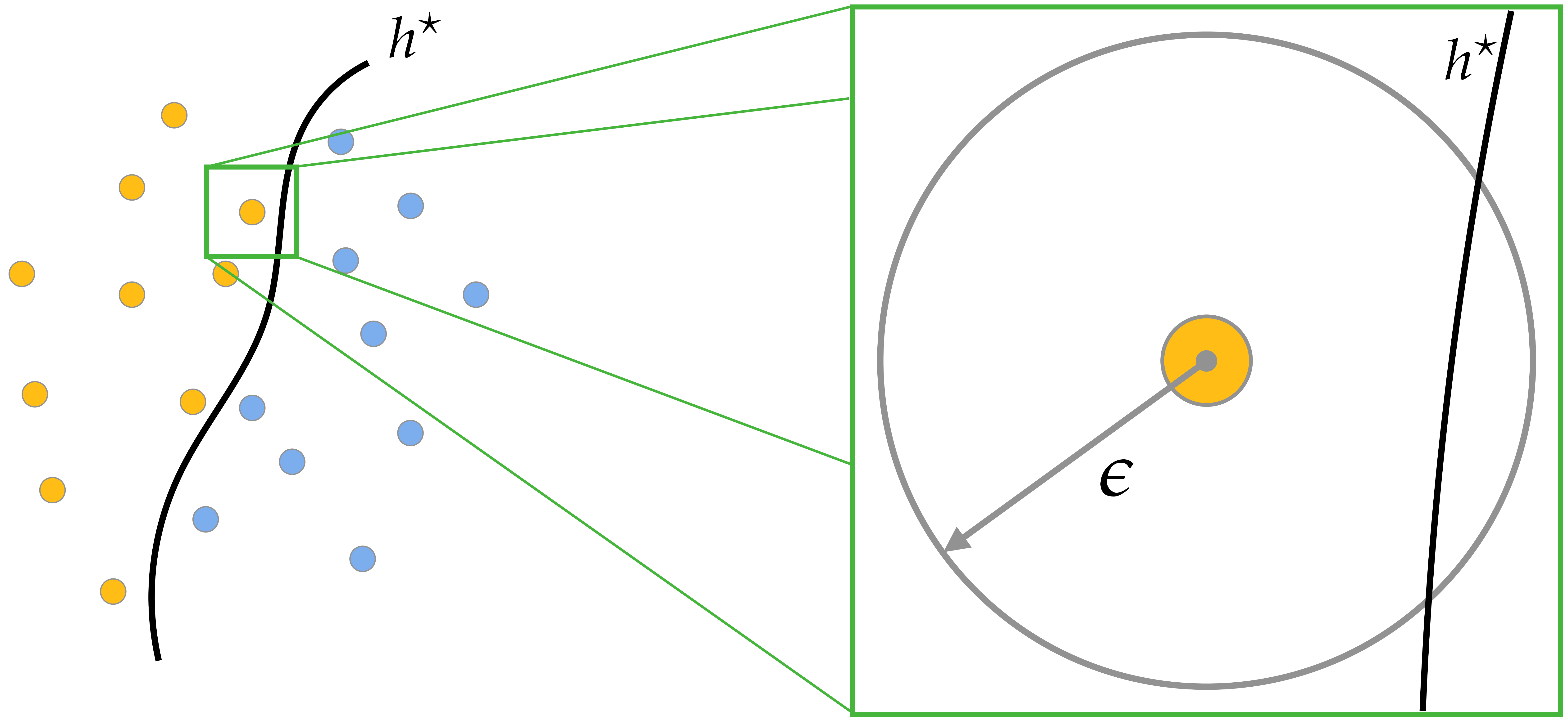


"Accurate, yet brittle" | "Robust, yet conservative"

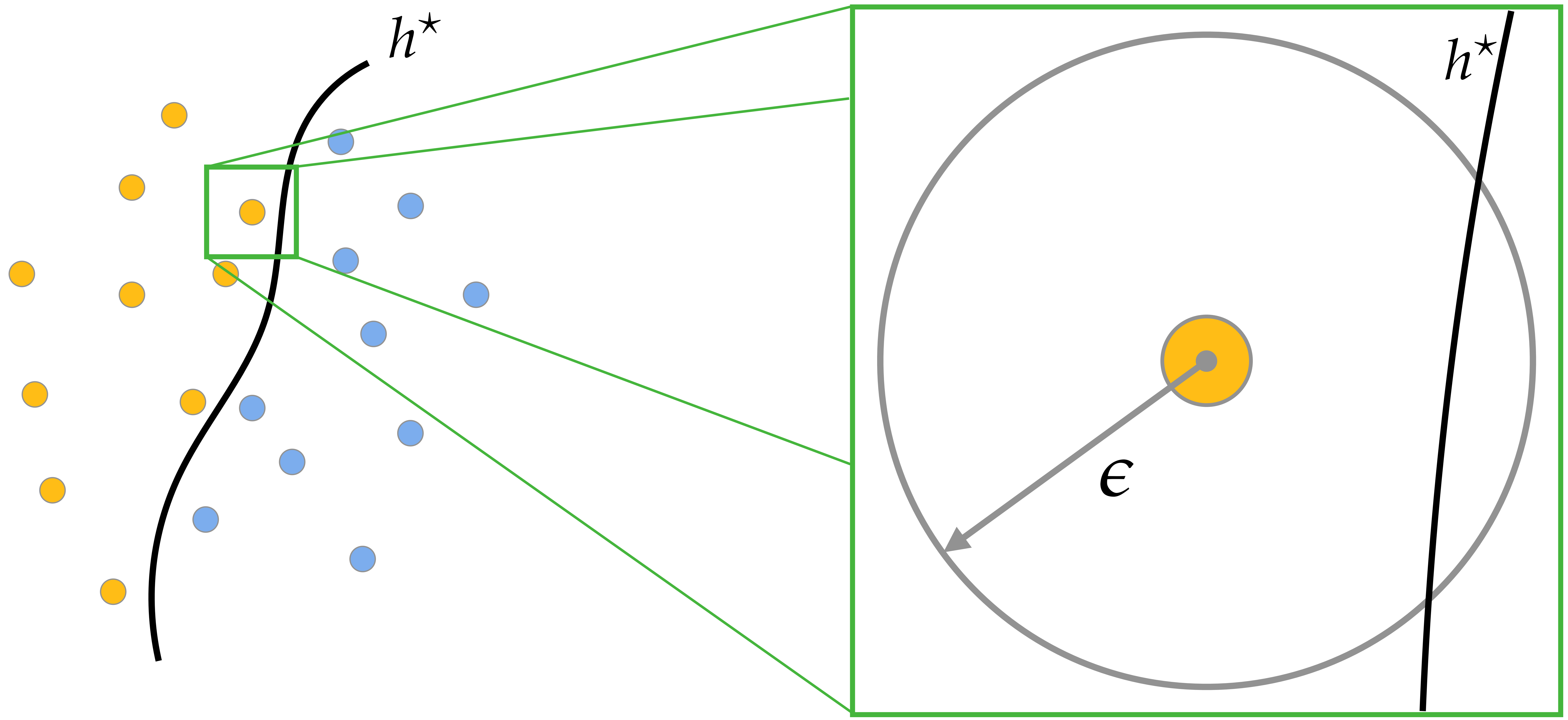
Our solution: ***Probabilistically Robust Learning (PRL)***

Our solution: *Probabilistically Robust Learning (PRL)*

Our solution: *Probabilistically Robust Learning (PRL)*

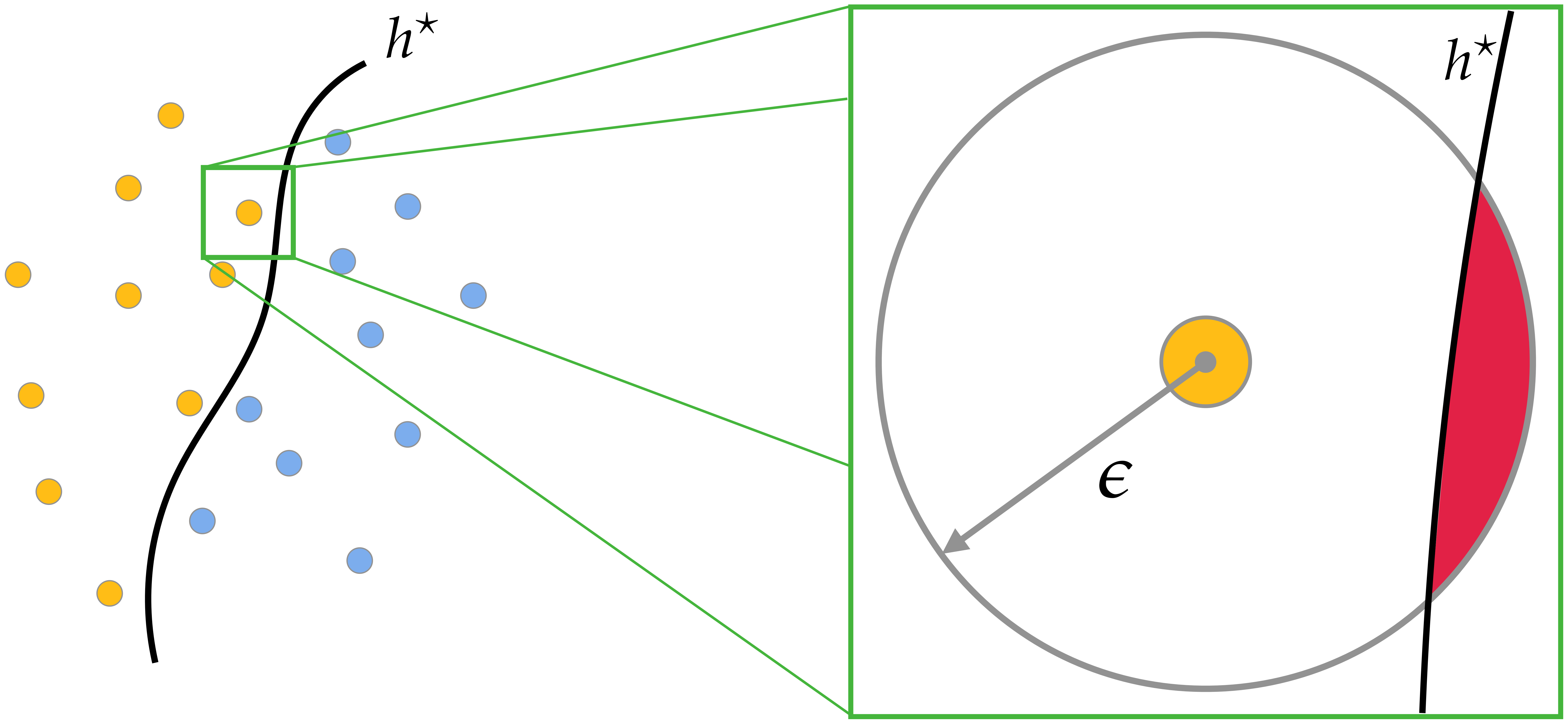


Our solution: *Probabilistically Robust Learning (PRL)*



Core idea: Enforce robustness to most — not all — perturbations.

Our solution: *Probabilistically Robust Learning (PRL)*



Core idea: Enforce robustness to most — not all — perturbations.

Our solution: *Probabilistically Robust Learning (PRL)*

Theoretical

Algorithmic

Our solution: *Probabilistically Robust Learning (PRL)*

Theoretical

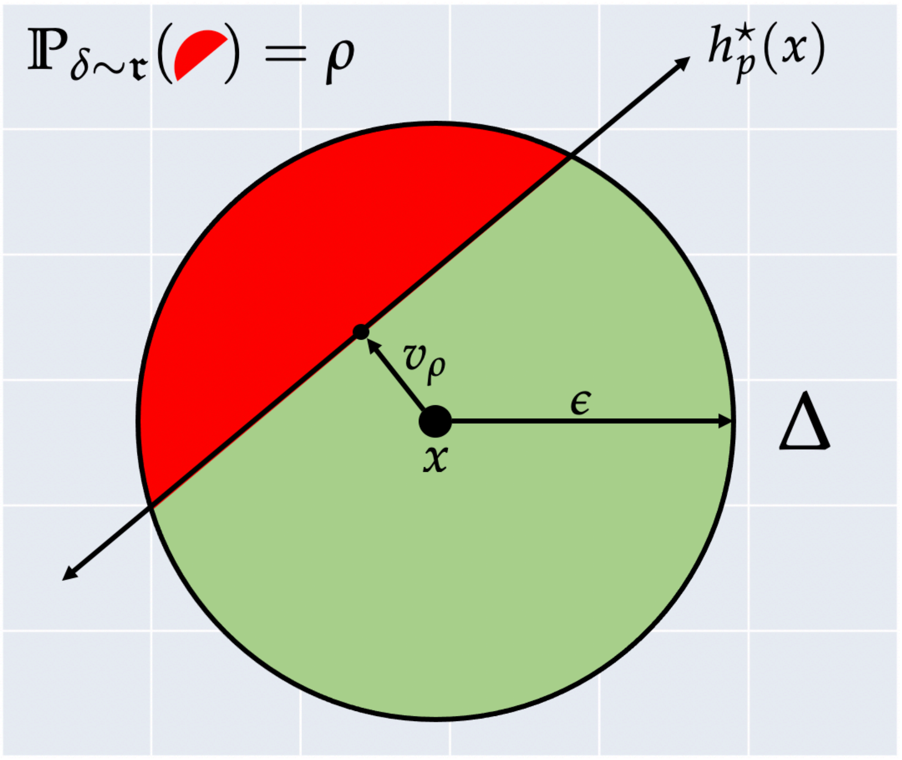
- ***(Lack of) Provable tradeoffs***: Probabilistic robustness is **not** at odds with accuracy

Algorithmic

Our solution: *Probabilistically Robust Learning (PRL)*

Theoretical

- ***(Lack of) Provable tradeoffs***: Probabilistic robustness is **not** at odds with accuracy
  - Linear regression
  - Mixture-of-Gaussians classification

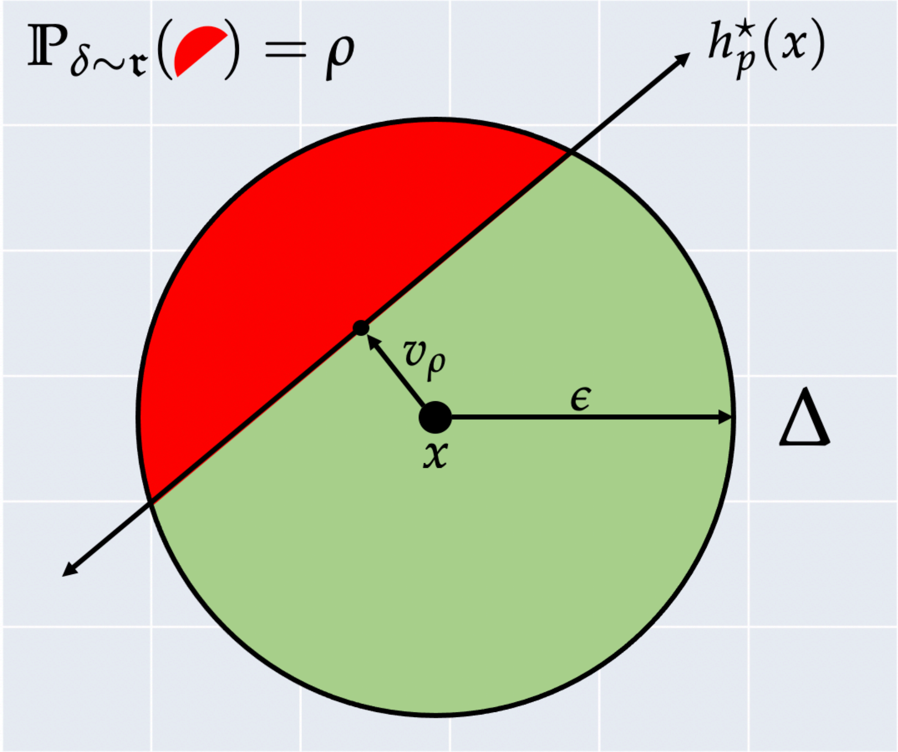


Algorithmic

Our solution: *Probabilistically Robust Learning (PRL)*

Theoretical

- ***(Lack of) Provable tradeoffs***: Probabilistic robustness is **not** at odds with accuracy
  - Linear regression
  - Mixture-of-Gaussians classification



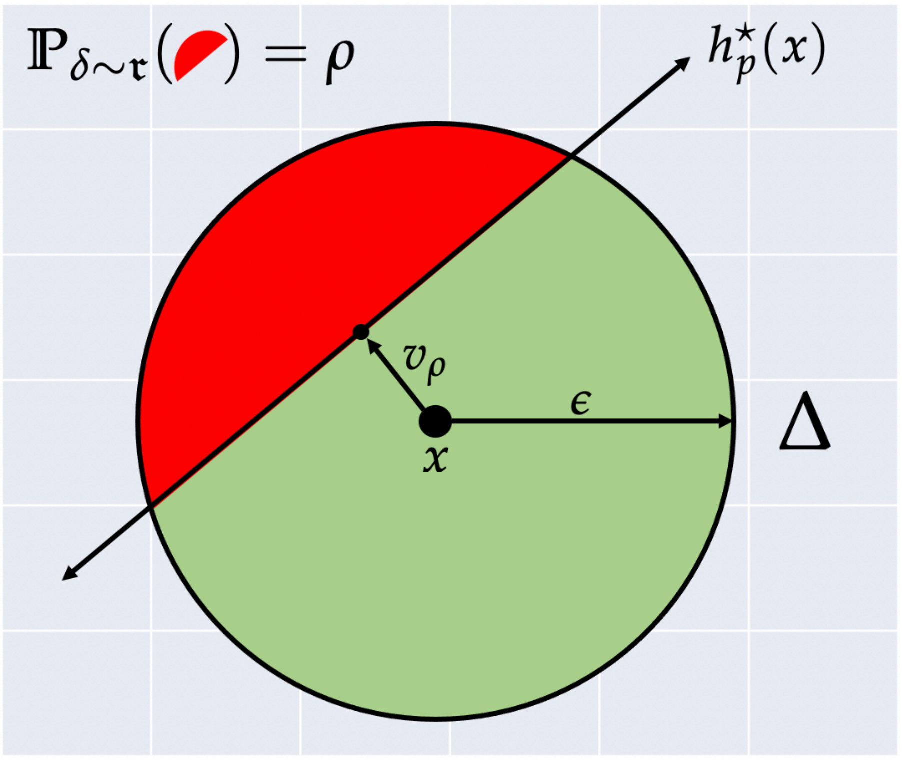
- ***Sample complexity:*** PR can

Algorithmic

Our solution: *Probabilistically Robust Learning (PRL)*

Theoretical

- ***(Lack of) Provable tradeoffs***: Probabilistic robustness is **not** at odds with accuracy
  - Linear regression
  - Mixture-of-Gaussians classification



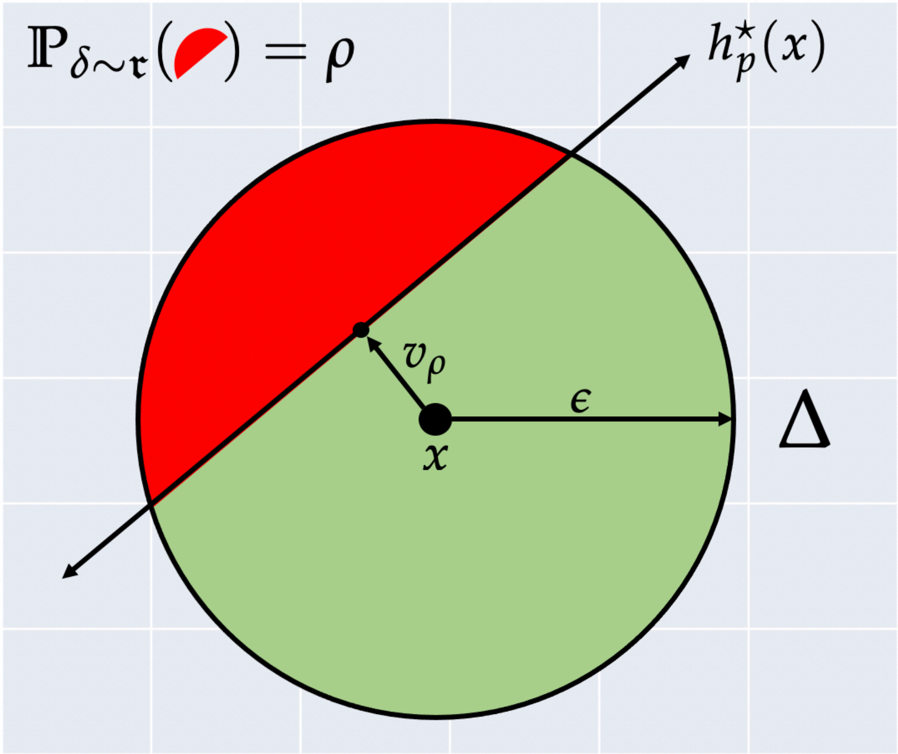
- ***Sample complexity:*** PR can
  - **match** the sample complexity of **ERM**

Algorithmic

Our solution: *Probabilistically Robust Learning (PRL)*

Theoretical

- ***(Lack of) Provable tradeoffs***: Probabilistic robustness is **not** at odds with accuracy
  - Linear regression
  - Mixture-of-Gaussians classification



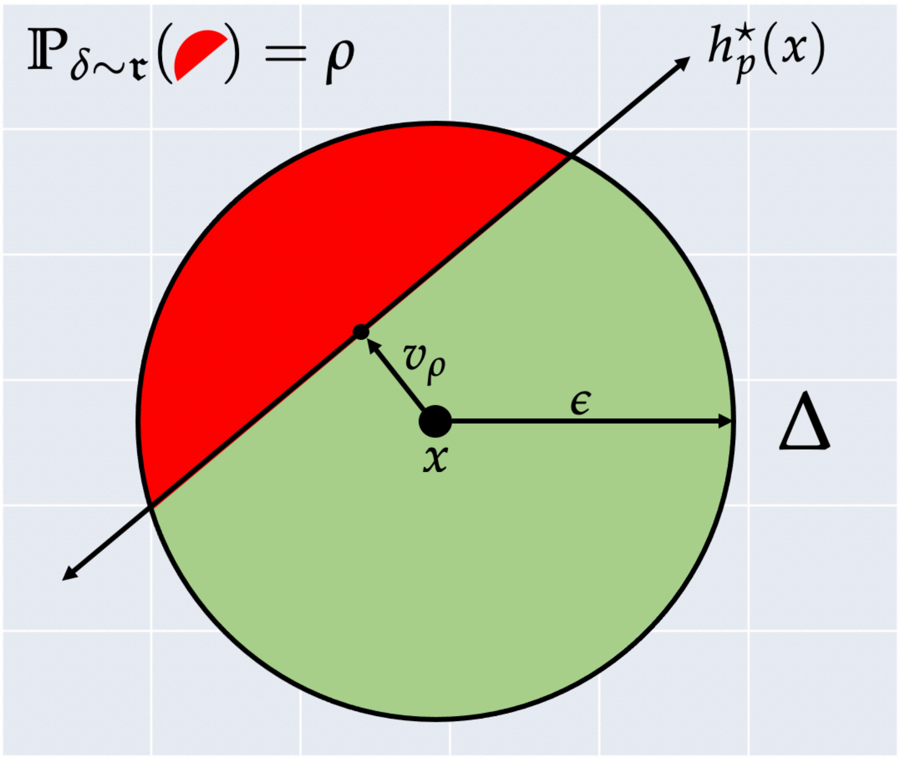
- ***Sample complexity:*** PR can
  - **match** the sample complexity of **ERM**
  - be **exponentially smaller** than the sample complexity of **adversarial training**

Algorithmic

Our solution: *Probabilistically Robust Learning (PRL)*

Theoretical

- ***(Lack of) Provable tradeoffs***: Probabilistic robustness is **not** at odds with accuracy
  - Linear regression
  - Mixture-of-Gaussians classification



- ***Sample complexity:*** PR can
  - **match** the sample complexity of **ERM**
  - be **exponentially smaller** than the sample complexity of **adversarial training**

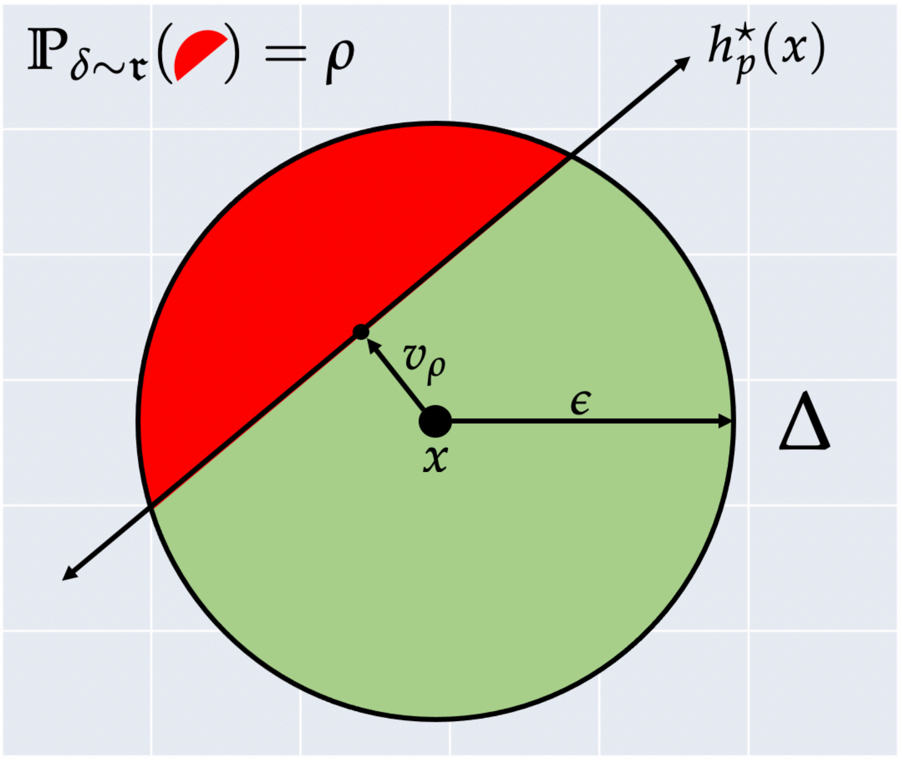
Algorithmic

- ***Outperform baselines***:
  - MNIST, CIFAR-10, SVHN

Our solution: *Probabilistically Robust Learning (PRL)*

Theoretical

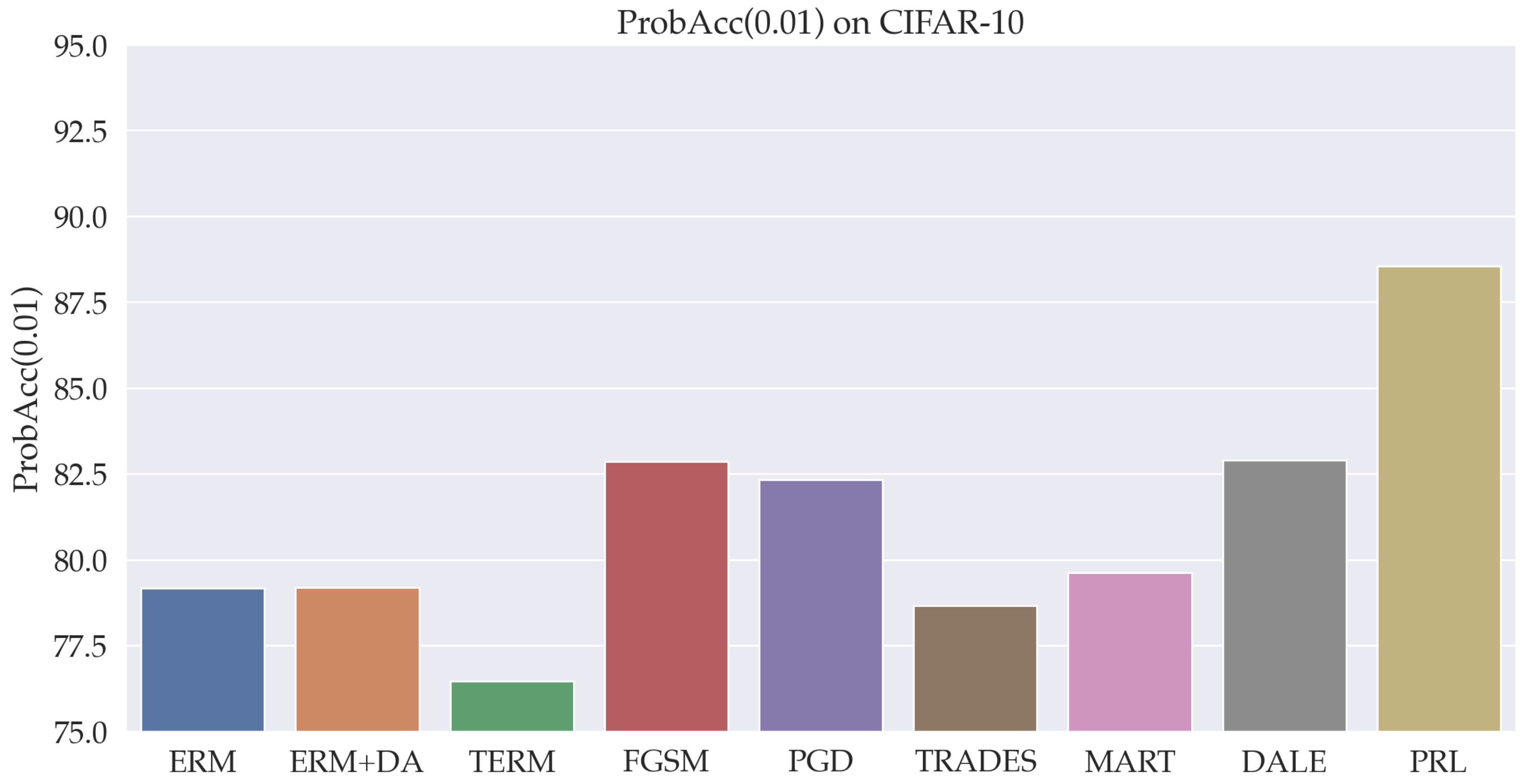
- ***(Lack of) Provable tradeoffs***: Probabilistic robustness is **not** at odds with accuracy
  - Linear regression
  - Mixture-of-Gaussians classification



- ***Sample complexity:*** PR can
  - **match** the sample complexity of **ERM**
  - be **exponentially smaller** than the sample complexity of **adversarial training**

Algorithmic

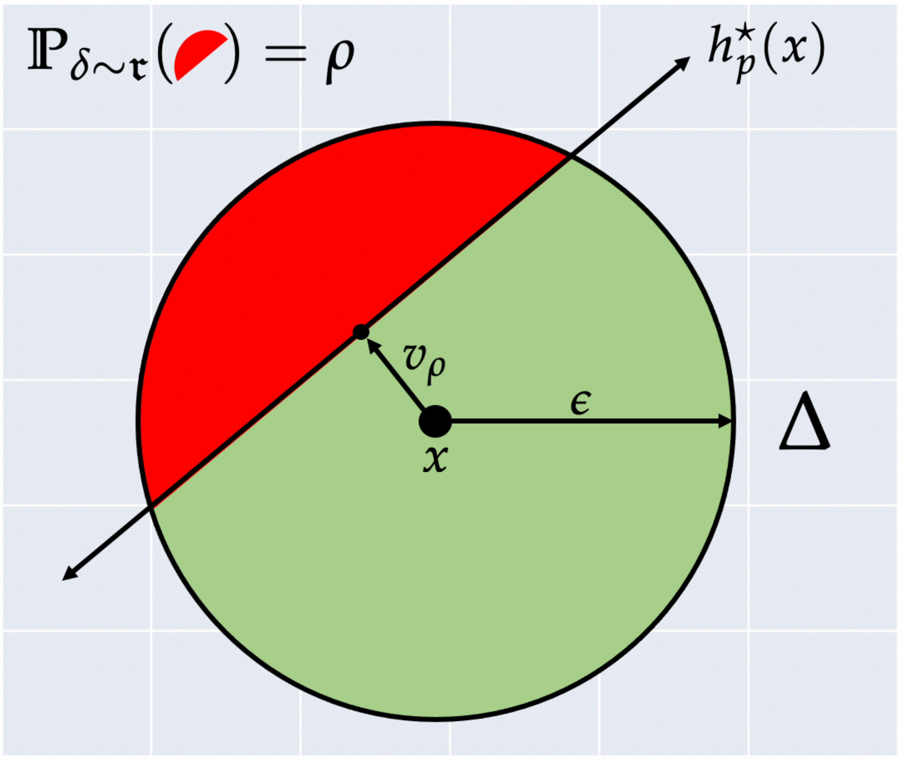
- ***Outperform baselines***:
  - MNIST, CIFAR-10, SVHN



Our solution: *Probabilistically Robust Learning (PRL)*

Theoretical

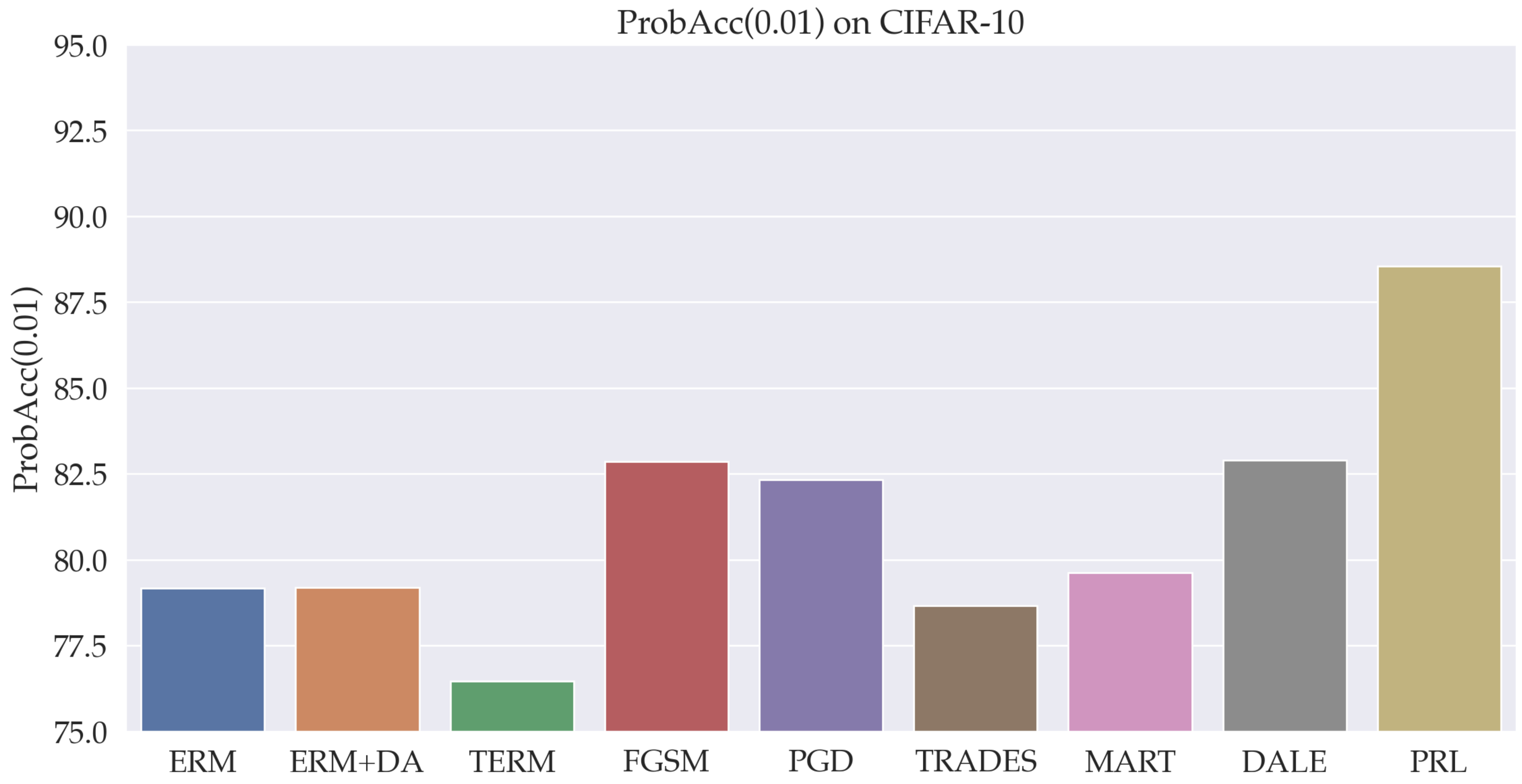
- ***(Lack of) Provable tradeoffs***: Probabilistic robustness is **not** at odds with accuracy
  - Linear regression
  - Mixture-of-Gaussians classification



- ***Sample complexity:*** PR can
  - **match** the sample complexity of **ERM**
  - be **exponentially smaller** than the sample complexity of **adversarial training**

Algorithmic

- ***Outperform baselines***:
  - MNIST, CIFAR-10, SVHN



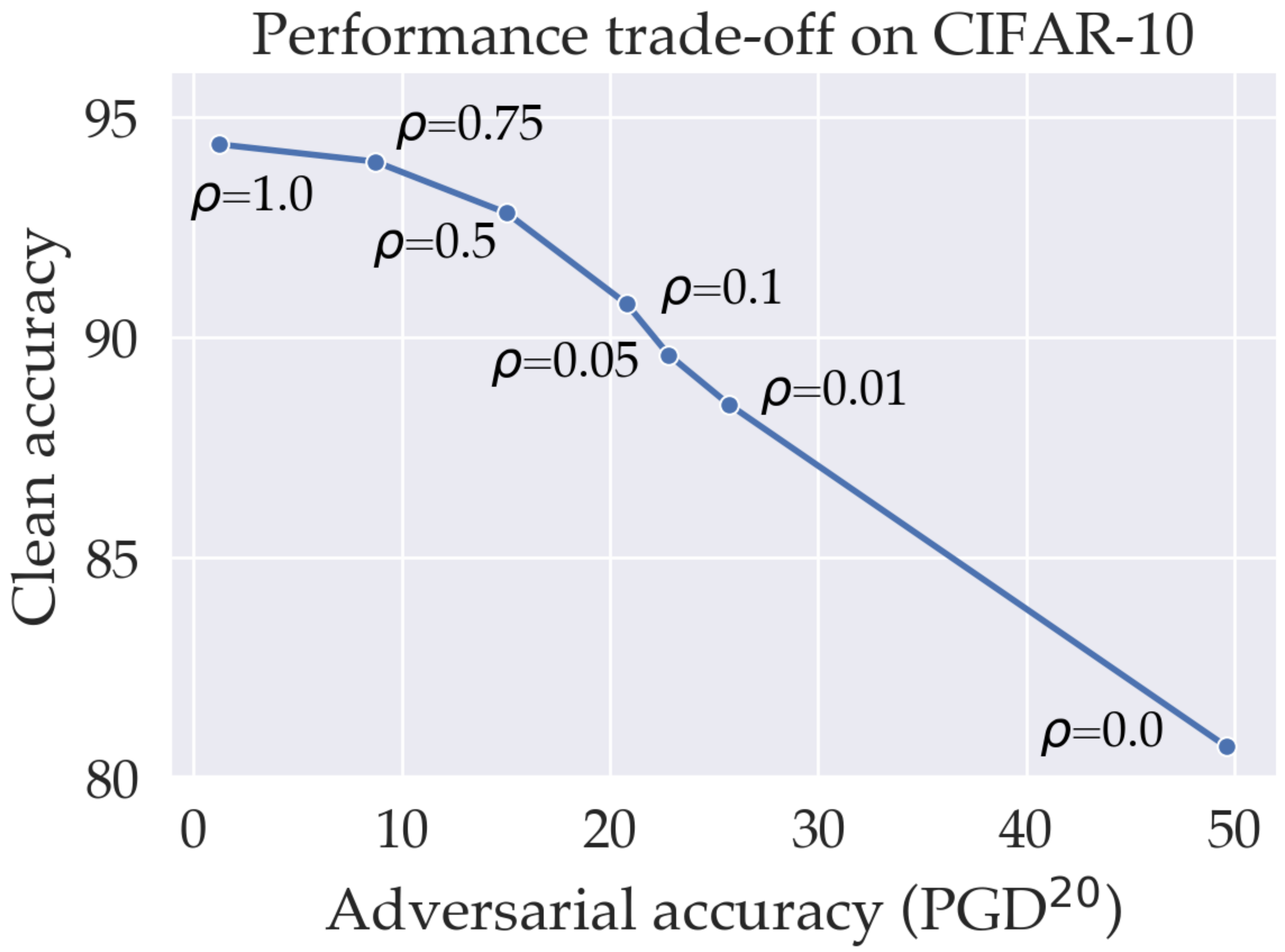
- ***Interpolate between ERM and adv. training:***

Our solution: *Probabilistically Robust Learning (PRL)*

- ***Interpolate between ERM and adv. training:***

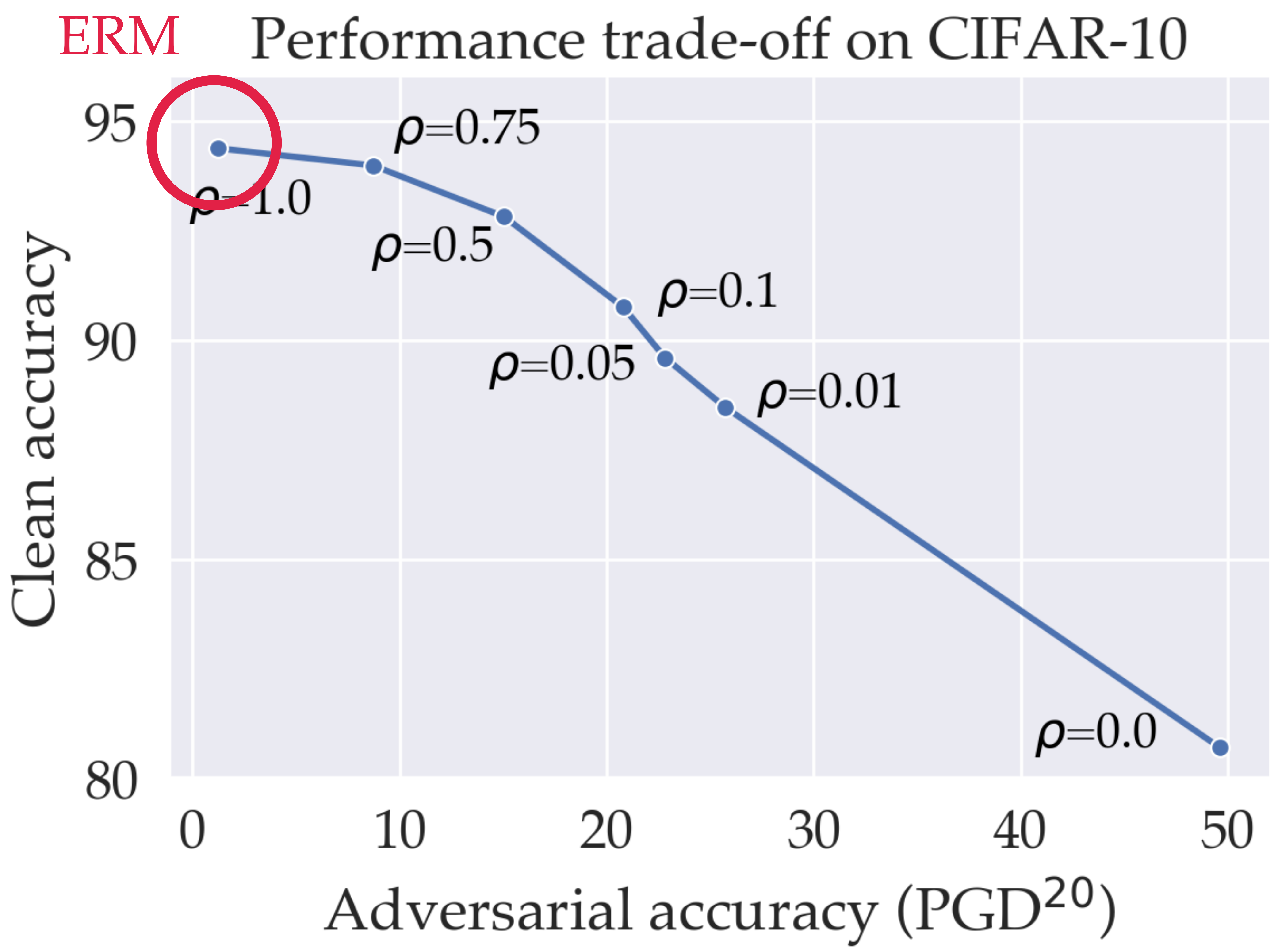
Our solution: *Probabilistically Robust Learning (PRL)*

- *Interpolate between ERM and adv. training:*



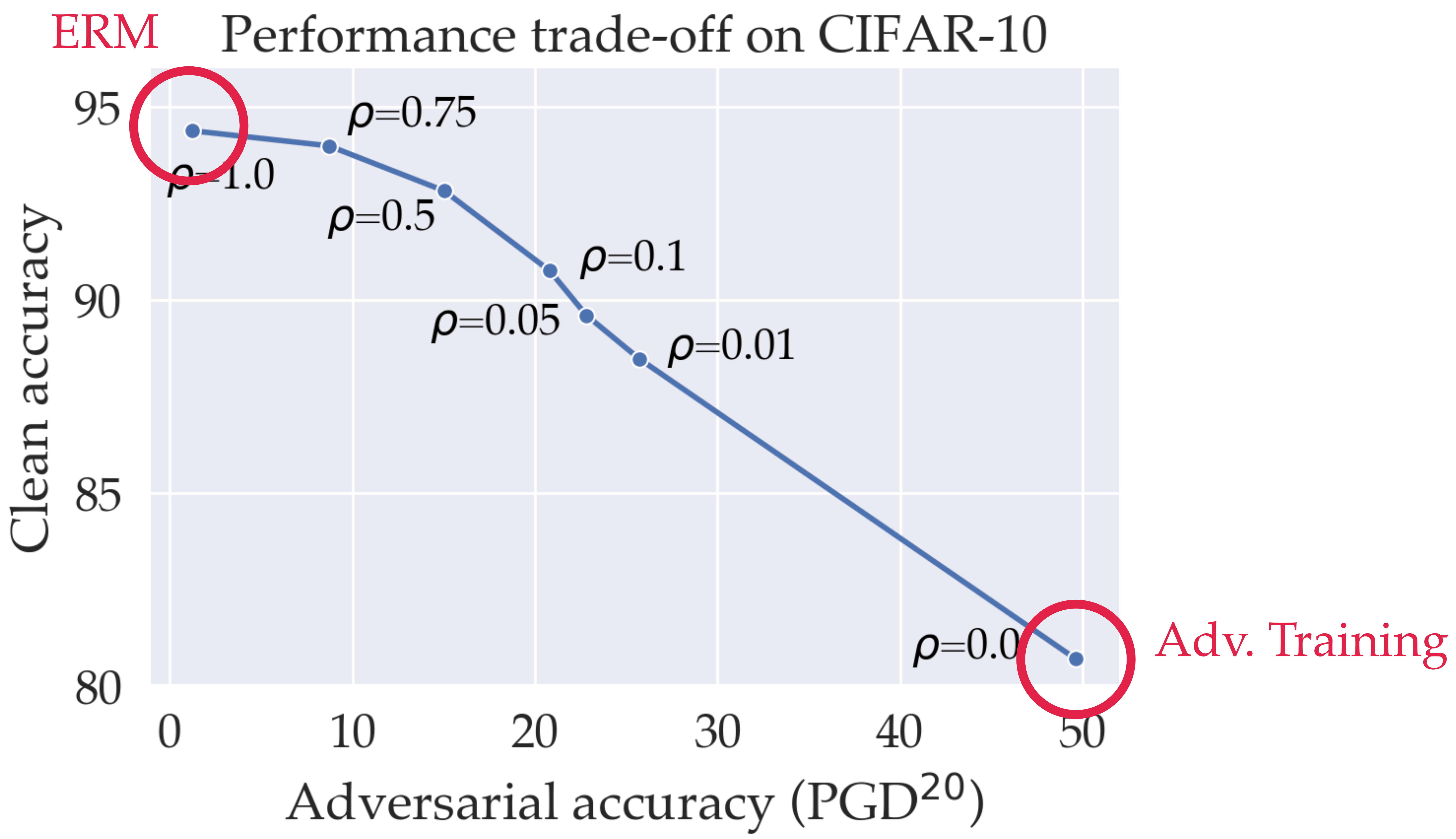
Our solution: *Probabilistically Robust Learning (PRL)*

- *Interpolate between ERM and adv. training:*



Our solution: *Probabilistically Robust Learning (PRL)*

- *Interpolate between ERM and adv. training:*



Any questions?

Alex Robey
arobey1@seas.upenn.edu