SelectiveNet: A Deep Neural Network with an Integrated Reject Option

Yonatan Geifman, Ran El-Yaniv

Technion - Israel Institute of Technology

Motivation

Uncertainty Landscape

Uncertainty Estimation

Selective Prediction

Active Learning

Probability Calibration

The Story of Alice and Bob

High Level Overview

- Motivation - Test with choice of 5 out of 10 questions
  - Bob - study all 10 chapters, decide at test time
  - Alice - study the easiest 5 chapters.
- Questions:
  - Which one is better?
  - How can Alice know which are easier?

Supervised Learning

- Underlying unknown distribution $P(X, Y)$
- A labeled set $S_m = \{(x, y)\}^m \sim P$
- Our goal is to find $f \in \mathcal{F}$ that minimizes the risk:

$$R(f) \triangleq E_P[\ell(f(x), y)]$$

Selective Classification

- Selective Classifier is a pair (f, g)

$$(f,g)(x) = \begin{cases} f(x), & \text{if } g(x) = 1; \\ \text{don't know}, & \text{if } g(x) = 0. \end{cases}$$

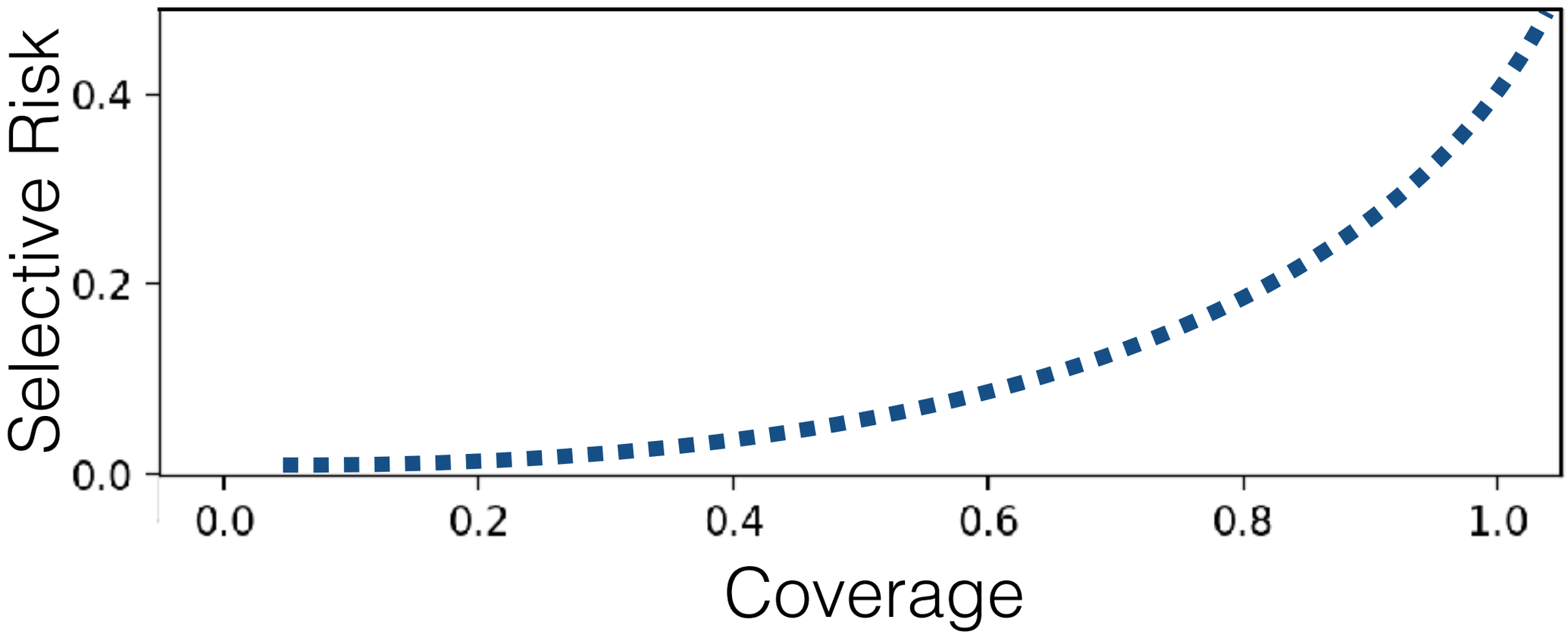
- Coverage:

$$\phi(f,g) \triangleq E_P[g(x)]$$

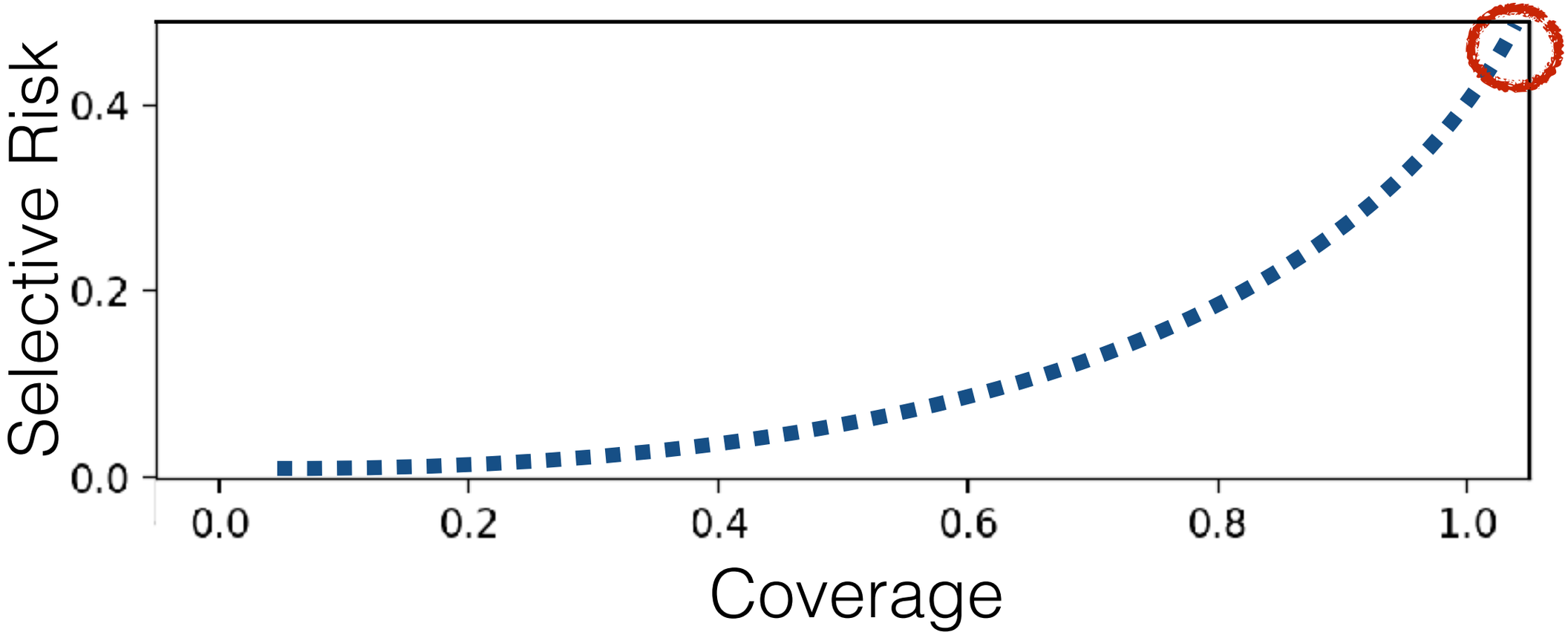
- Selective risk: $R(f,g) \triangleq \dfrac{E_P[\ell(f(x),y)g(x)]}{\phi(f,g)}.$

Ran El-Yaniv, and Yair Wiener. "On the foundations of noise-free selective classification."

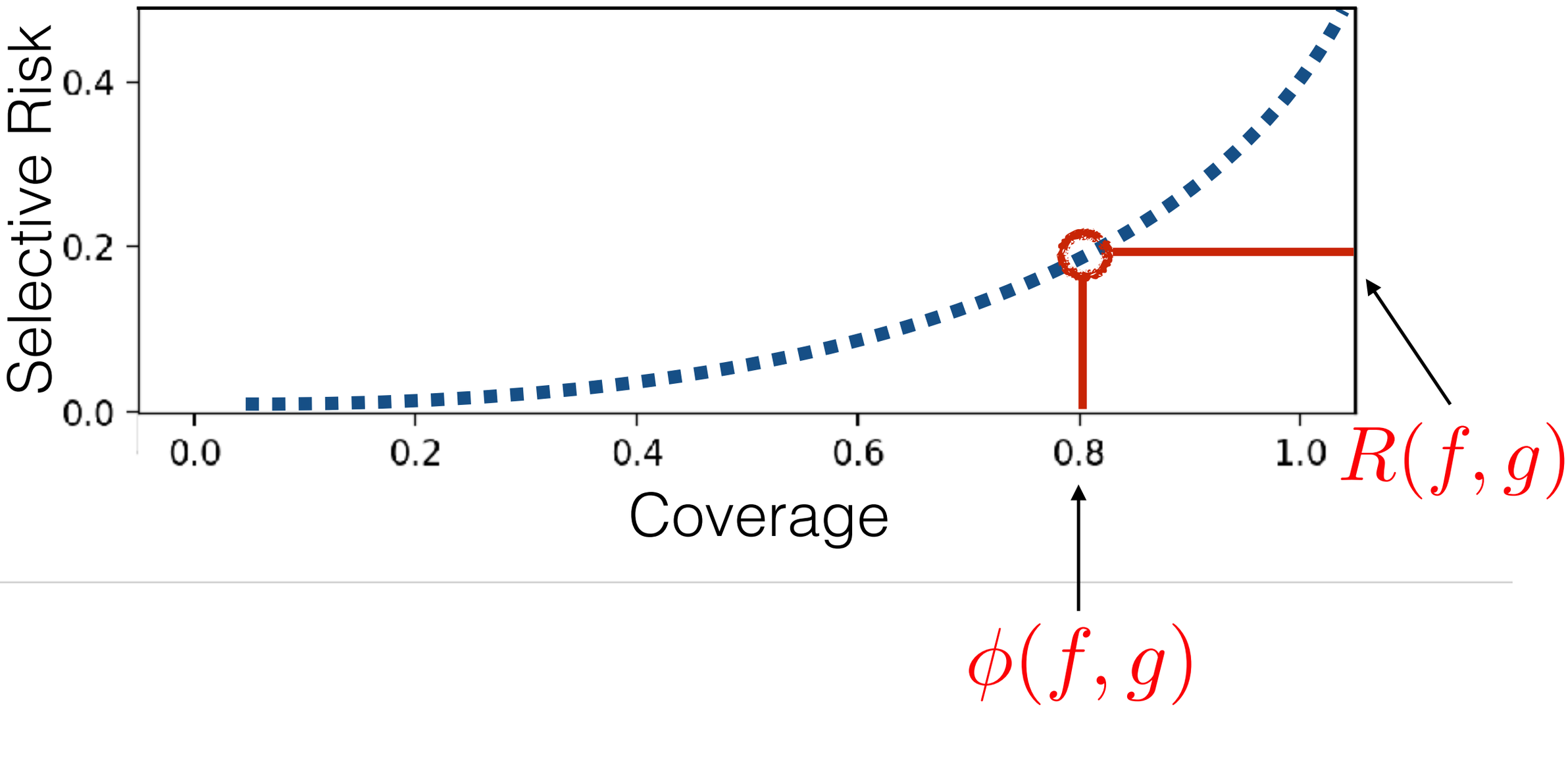
Risk Coverage Tradeoff



Risk Coverage Tradeoff



Risk Coverage Tradeoff



Confidence Rate Functions

- For a classifier f ,We seek for a confidence rate function κ that reflects loss monotonicity

$$\kappa(x_1|f) \leq \kappa(x_2|f) \iff Pr_P[f(x_1) \neq y_1] \geq Pr_P[f(x_2) \neq y_2]$$

- κ is only ranking (not probabilities)

Geifman & El-Yaniv - "Selective classification for deep neural networks” NeurIPS 17

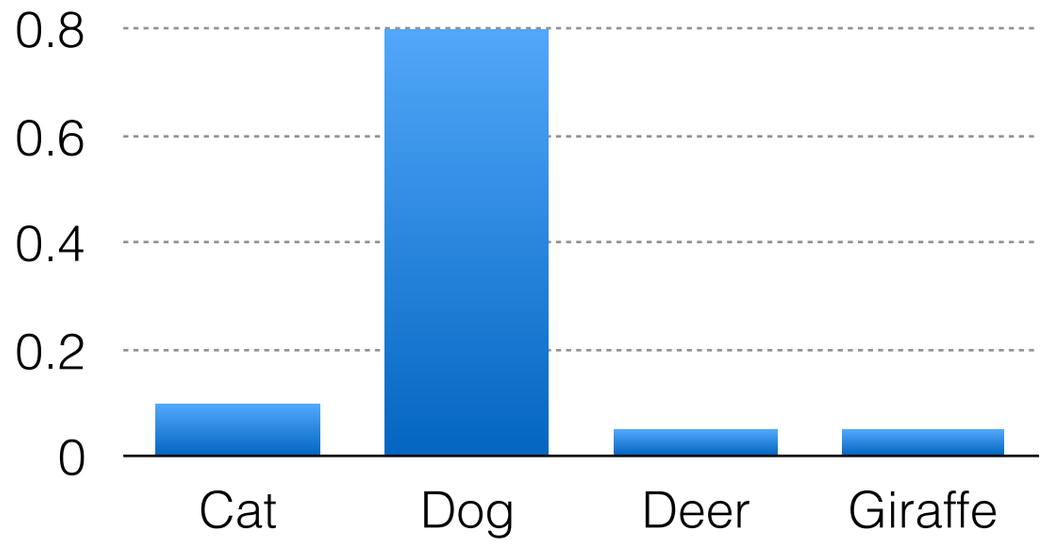
Confidence - Softmax Response

- Simply take κ to be the Softmax Response (SR)

$$\kappa_f \triangleq \max_{j \in \mathcal{Y}} (f(x|j))$$

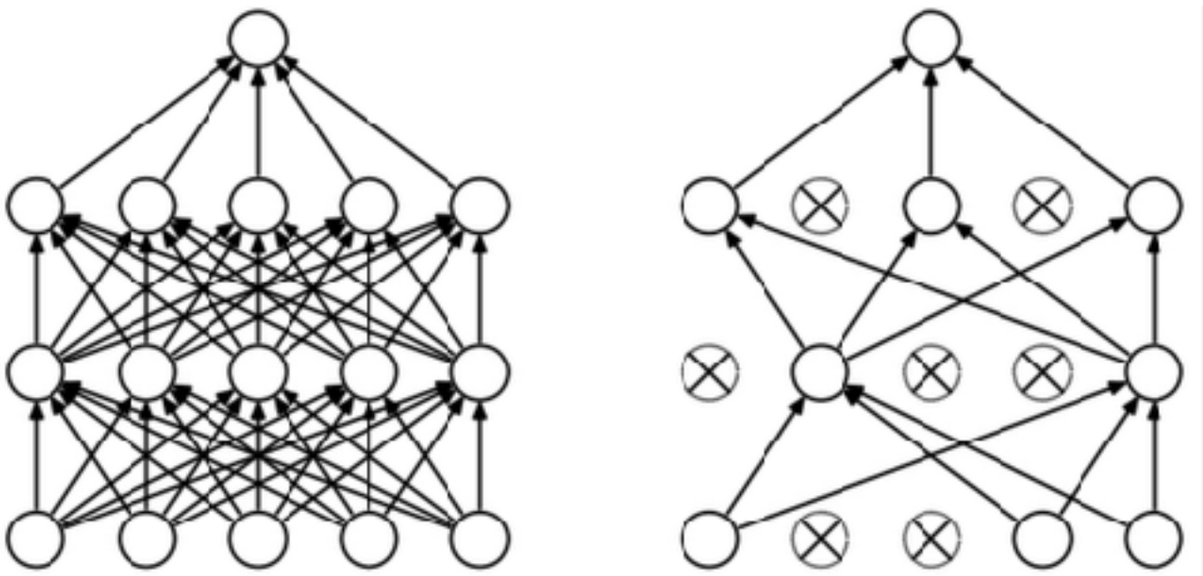
- Reflects the classification margin

- Other variants - Entropy, max minus 2nd



Confidence - MC-Dropout

- Apply dropout at inference
- Estimate prediction statistics over numerous forward passes with dropout



(a) Standard Neural Net (b) After applying dropout.

Gal & Ghahramani. "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning."

From Uncertainty to Selective Classifier

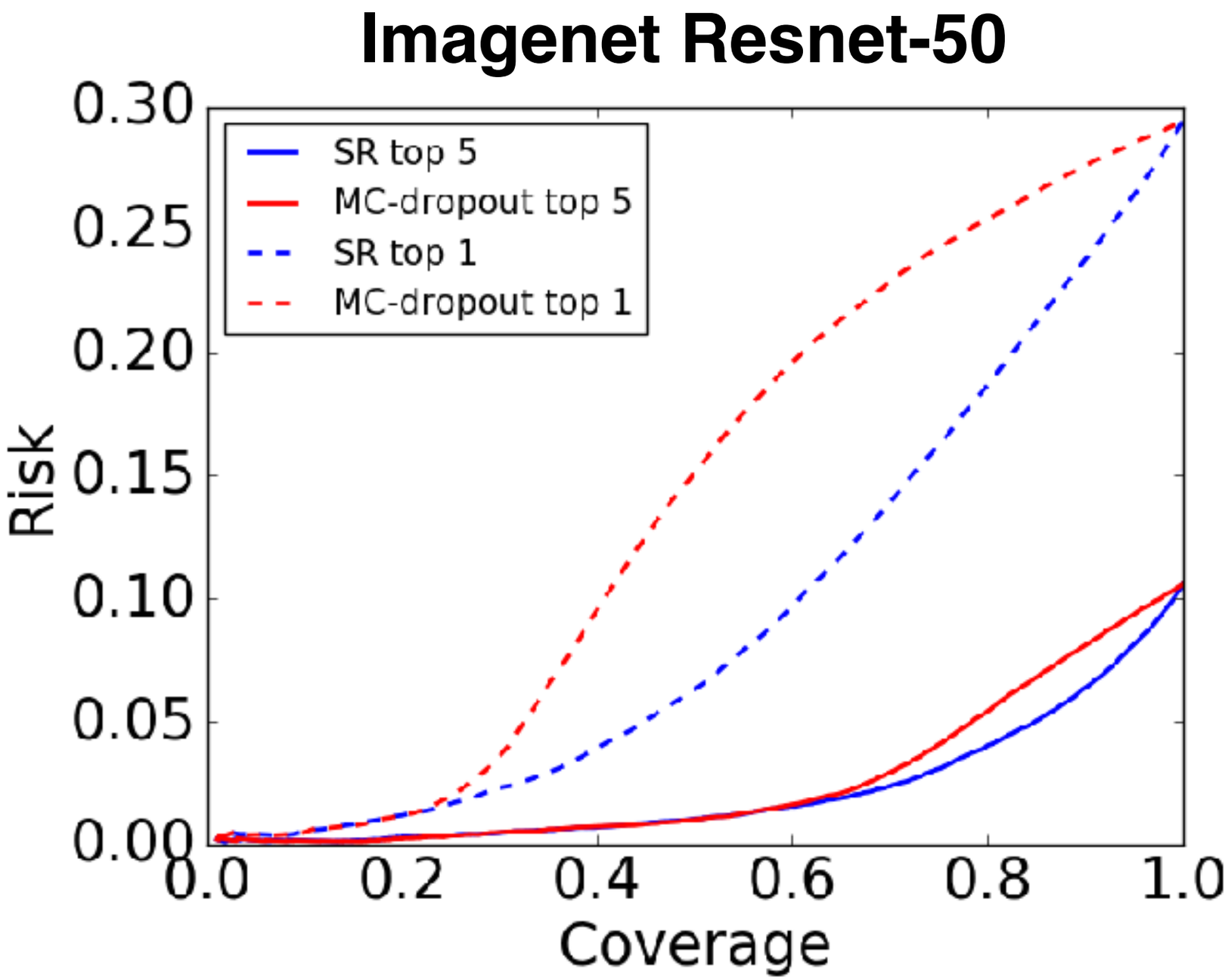
- A selective classifier can be obtained by thresholding the confidence rate function

$$g_\theta(x) \triangleq \begin{cases} 1, & \text{if } \kappa_f(x) \geq \theta; \\ 0, & \text{otherwise.} \end{cases}$$

- Selection of θ is based on a validation set

Geifman & El-Yaniv - "Selective classification for deep neural networks" NeurIPS 17

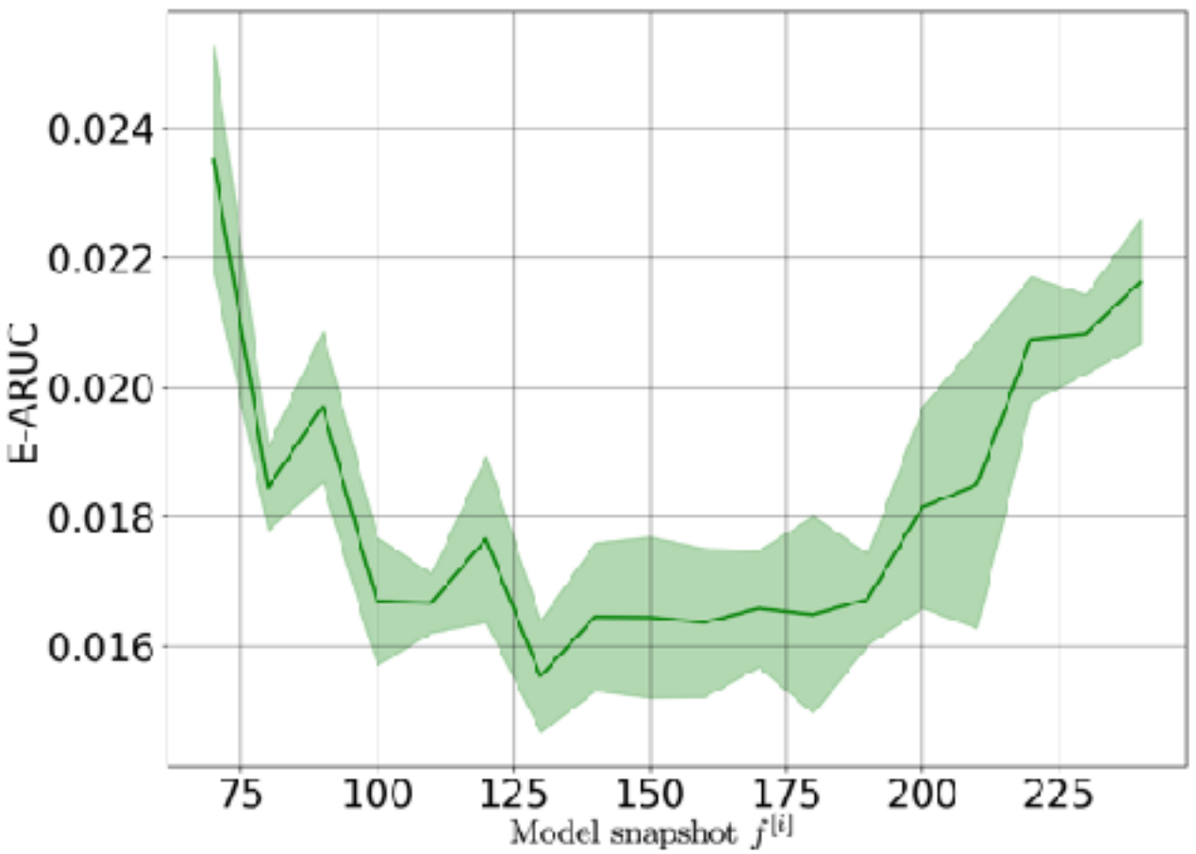
SR or MC-dropout?



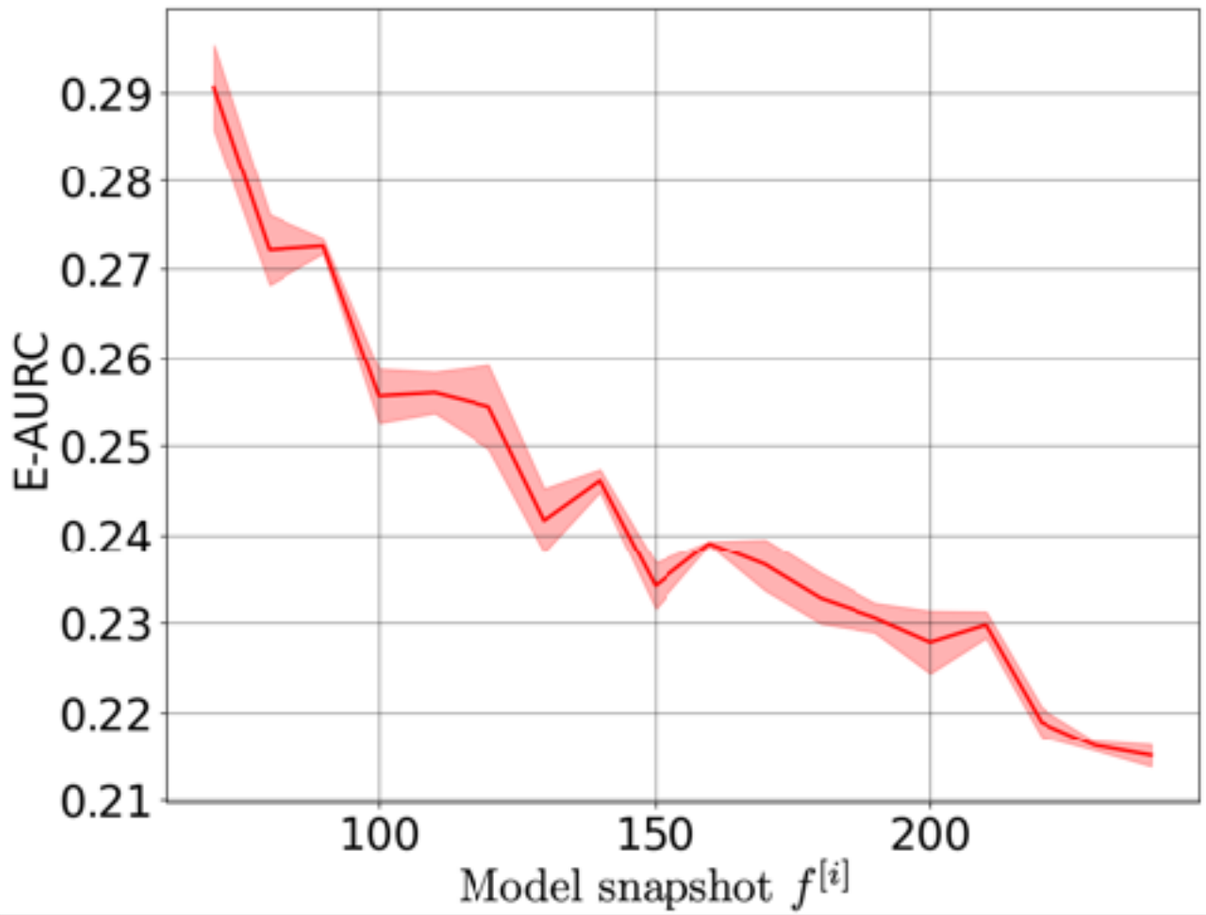
Geifman & El-Yaniv - "Selective classification for deep neural networks" NeurIPS 17

Softmax Response is Not Optimal

“Easy" Points

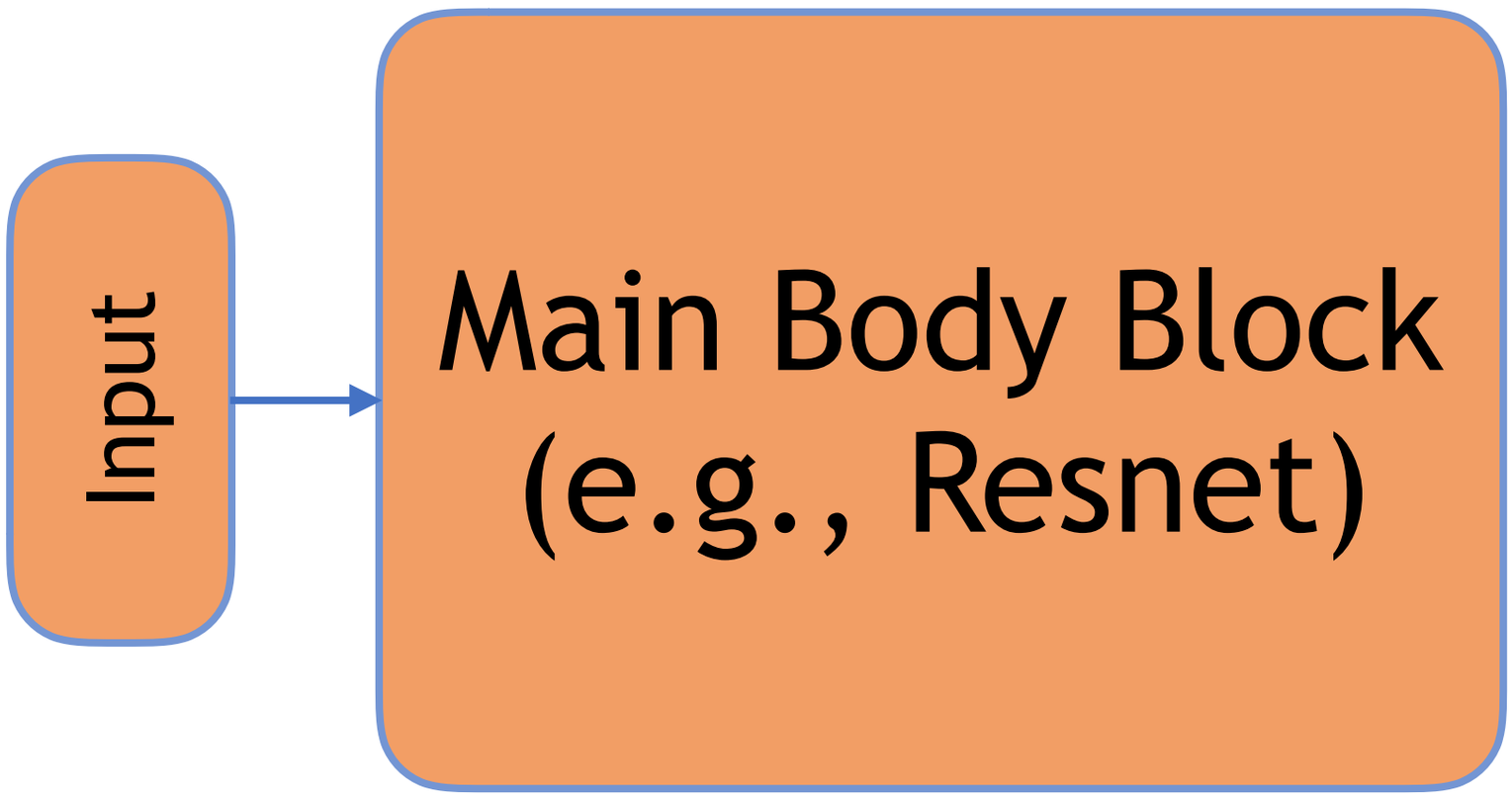


“Hard” Points

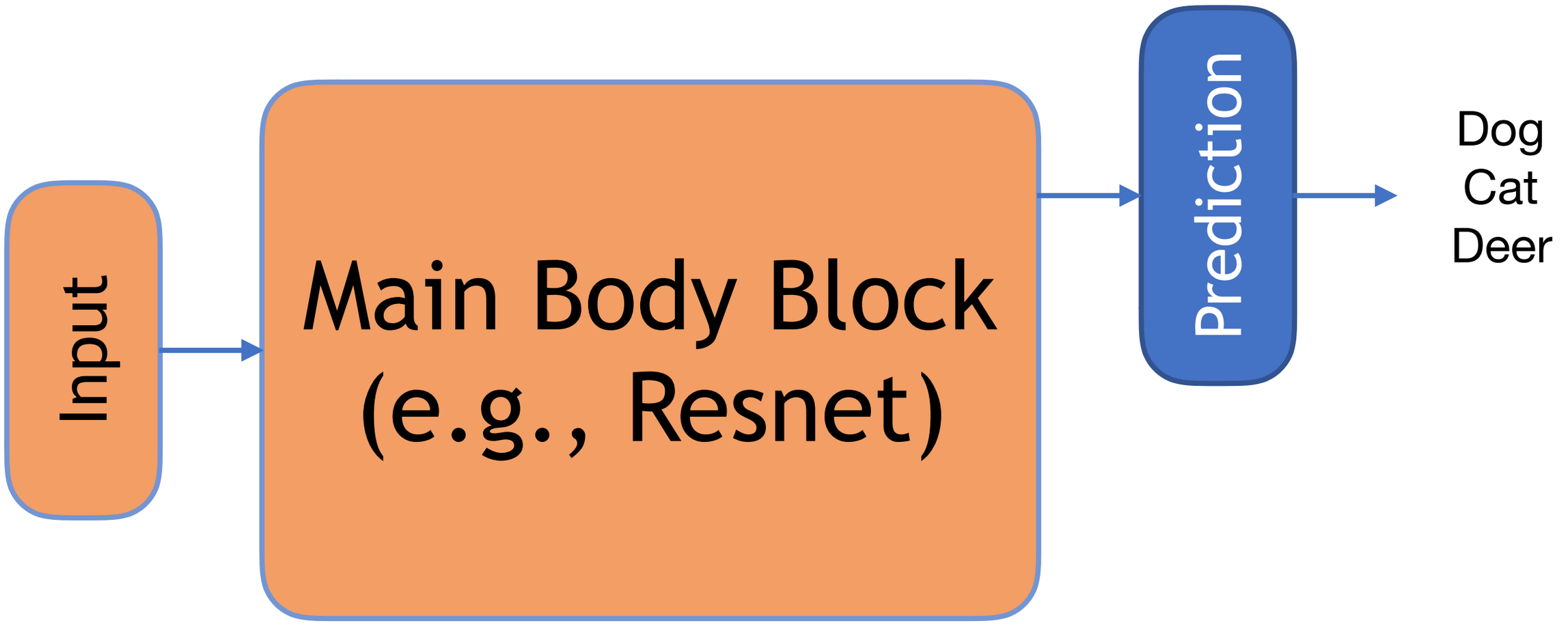


Geifman & El-Yaniv - “Bias-Reduced Uncertainty Estimation for deep neural classifiers” ICLR 19

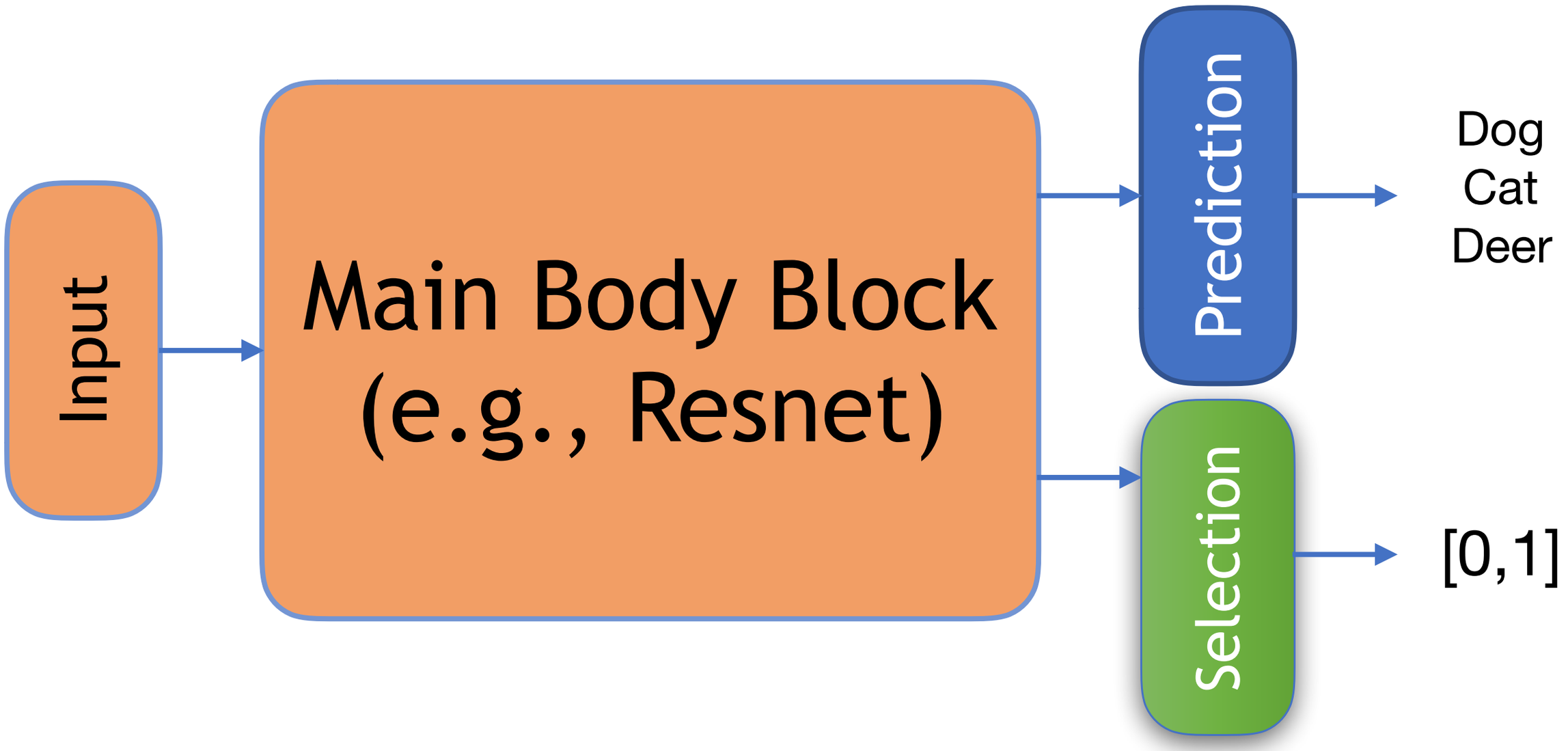
SelectiveNet



SelectiveNet



SelectiveNet



Optimization

- Inspired by interior point methods (IPM)
- Constrained optimization problem:

$$\theta^* = \arg\min_{\theta \in \Theta}(R(f_\theta, g_\theta))$$

$$s.t.\ \phi(g_\theta) \geq c.$$

- Unconstrained objective:

$$\mathcal{L}_{(f,g)} = \hat{r}(f, g|S_m) + \lambda\Psi(c - \hat{\phi}(g|S_m))$$

$$\Psi(a) = \max(0, a)^2$$

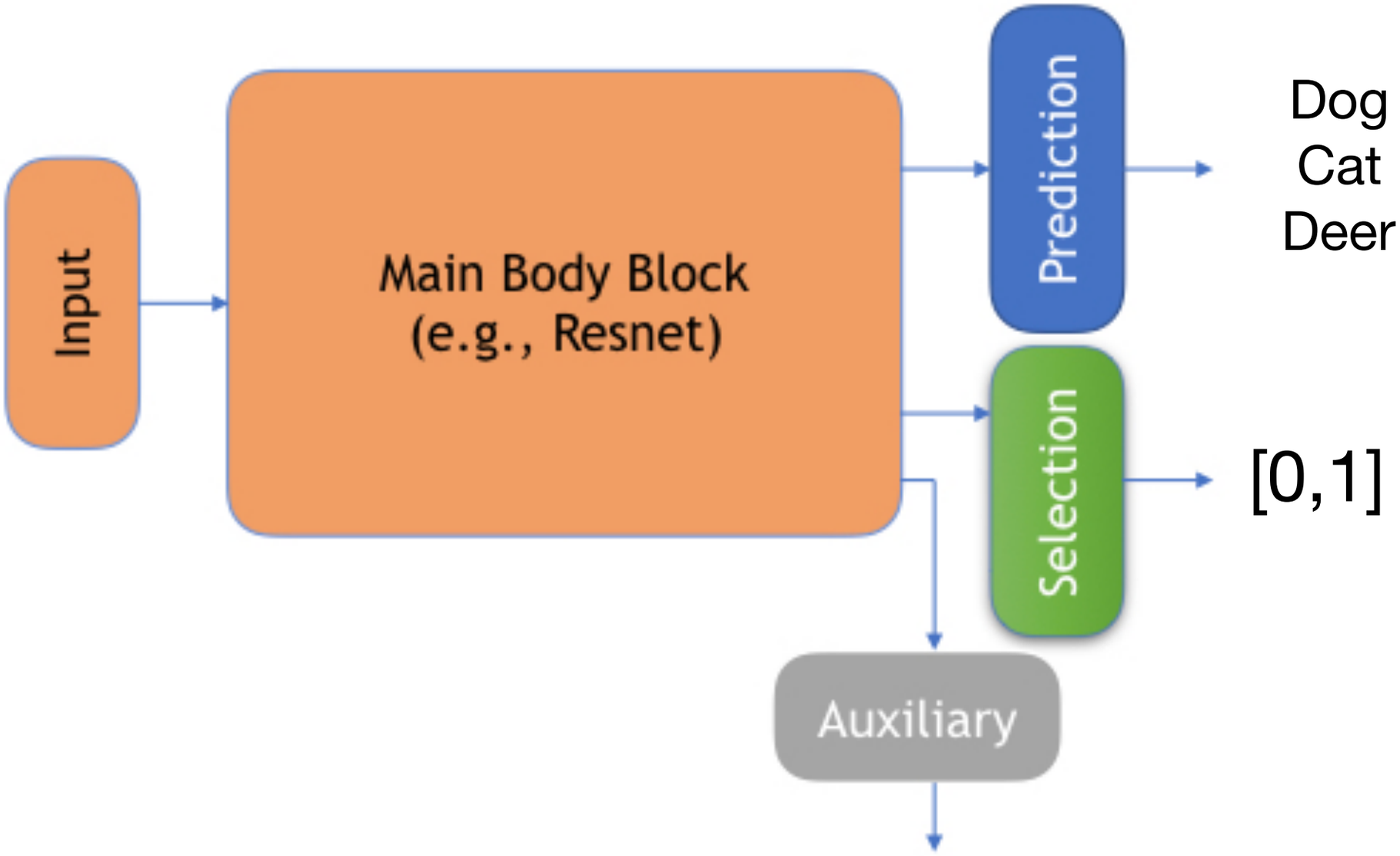
$$\hat{r}(f, g|S_m) = \frac{\frac{1}{m}\sum_{i=1}^{m} \ell(f(x_i), y_i)g(x_i)}{\hat{\phi}(g|S_m)}$$

The Story of Alice and Bob

Back to Bob and Alice

- How can we learn g before f is converged?
- How can Alice choose which 5 chapters to study?
- Can we get "world knowledge" from the rejected part?

Auxiliary Output



Auxiliary output

- In low coverage the effective training set size is reduced
- We can learn “world knowledge” from rejected instances
- Auxiliary output is added as regularisation for representation learning based on the entire domain

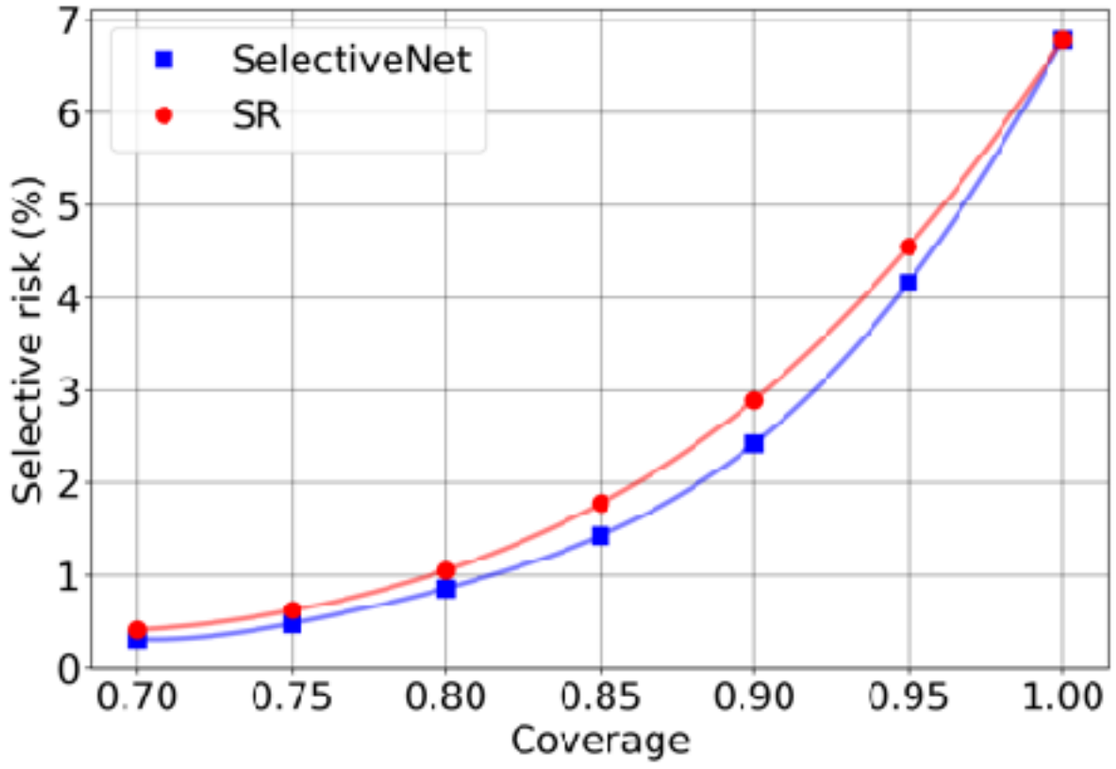
$$\mathcal{L}_h = \hat{r}(h|S_m) = \frac{1}{m}\sum_{i=1}^{m} \ell(h(x_i), y_i)$$

- Combined with the selective risk:

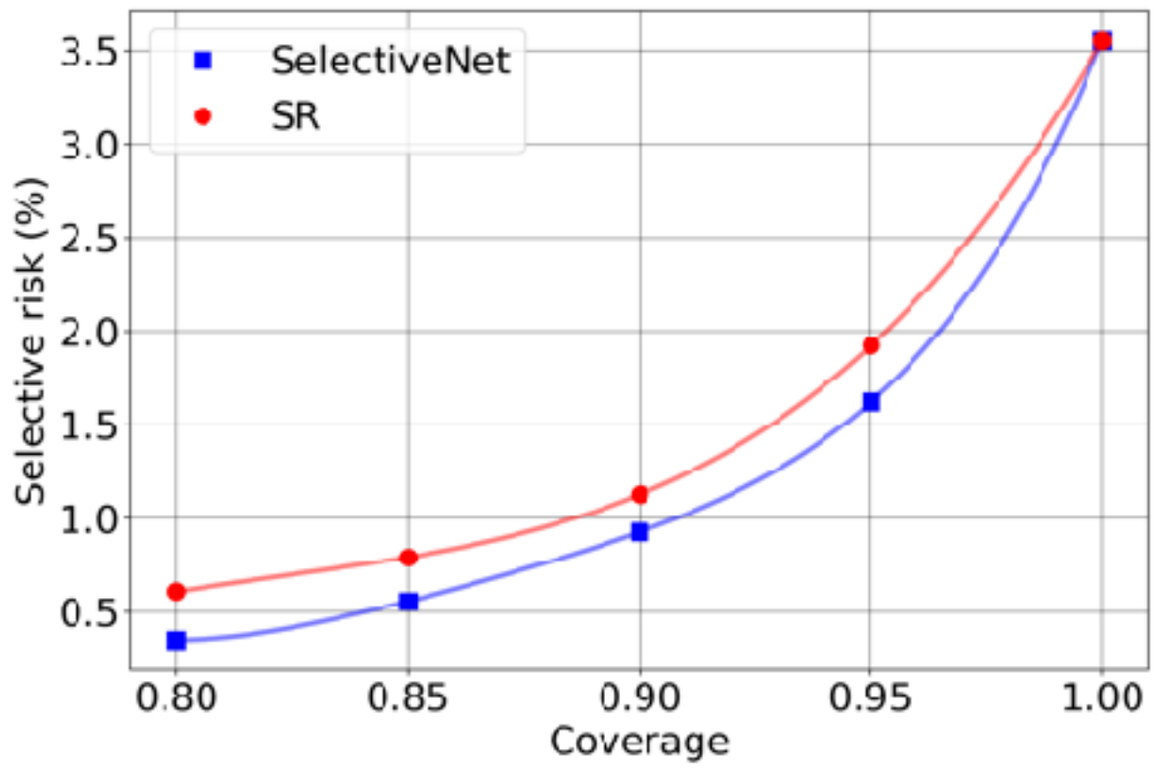
$$\mathcal{L} = \alpha\mathcal{L}_{(f,g)} + (1-\alpha)\mathcal{L}_h$$

Empirical Results

Cifar-10

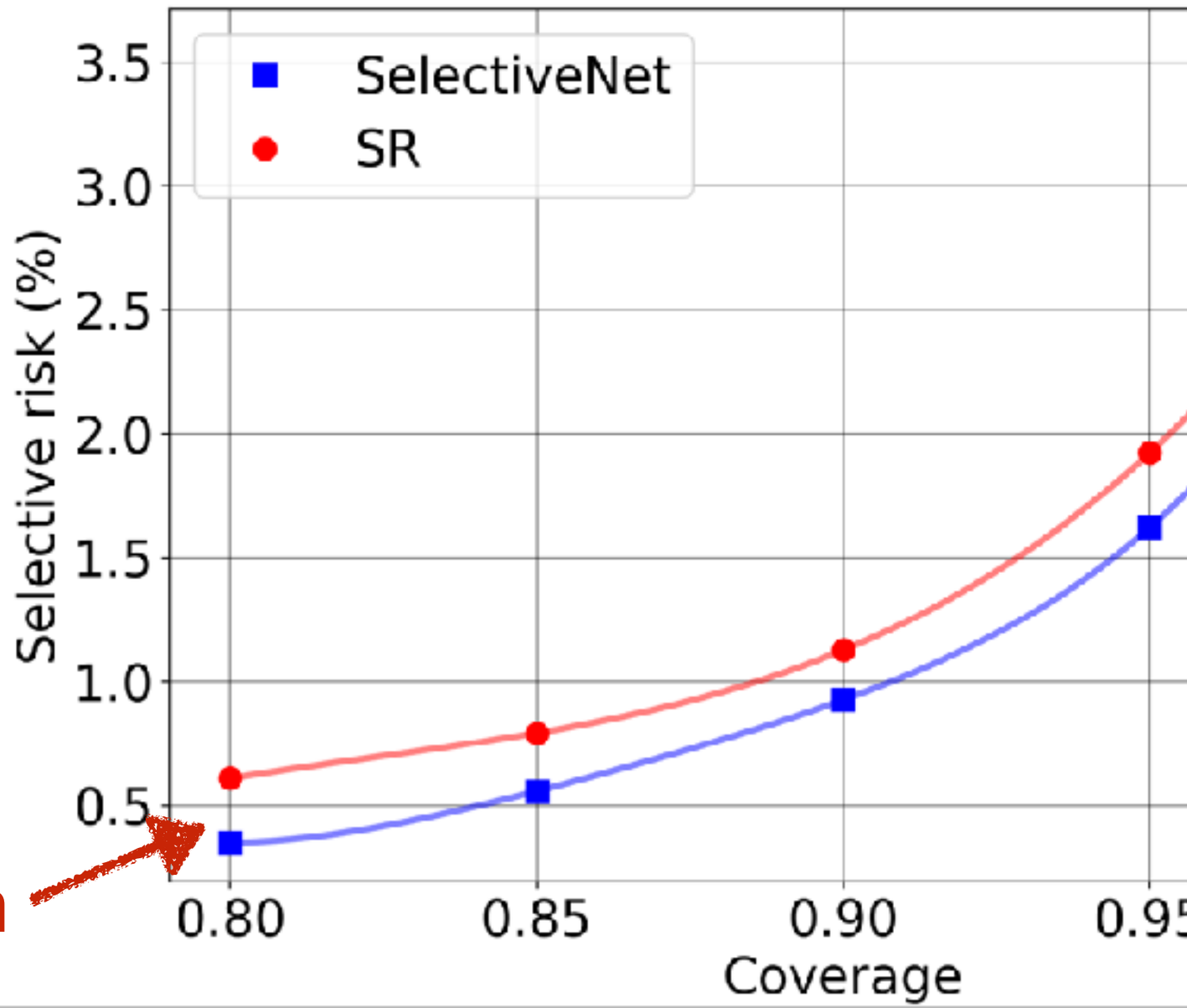


Cats vs. Dogs



Empirical Results

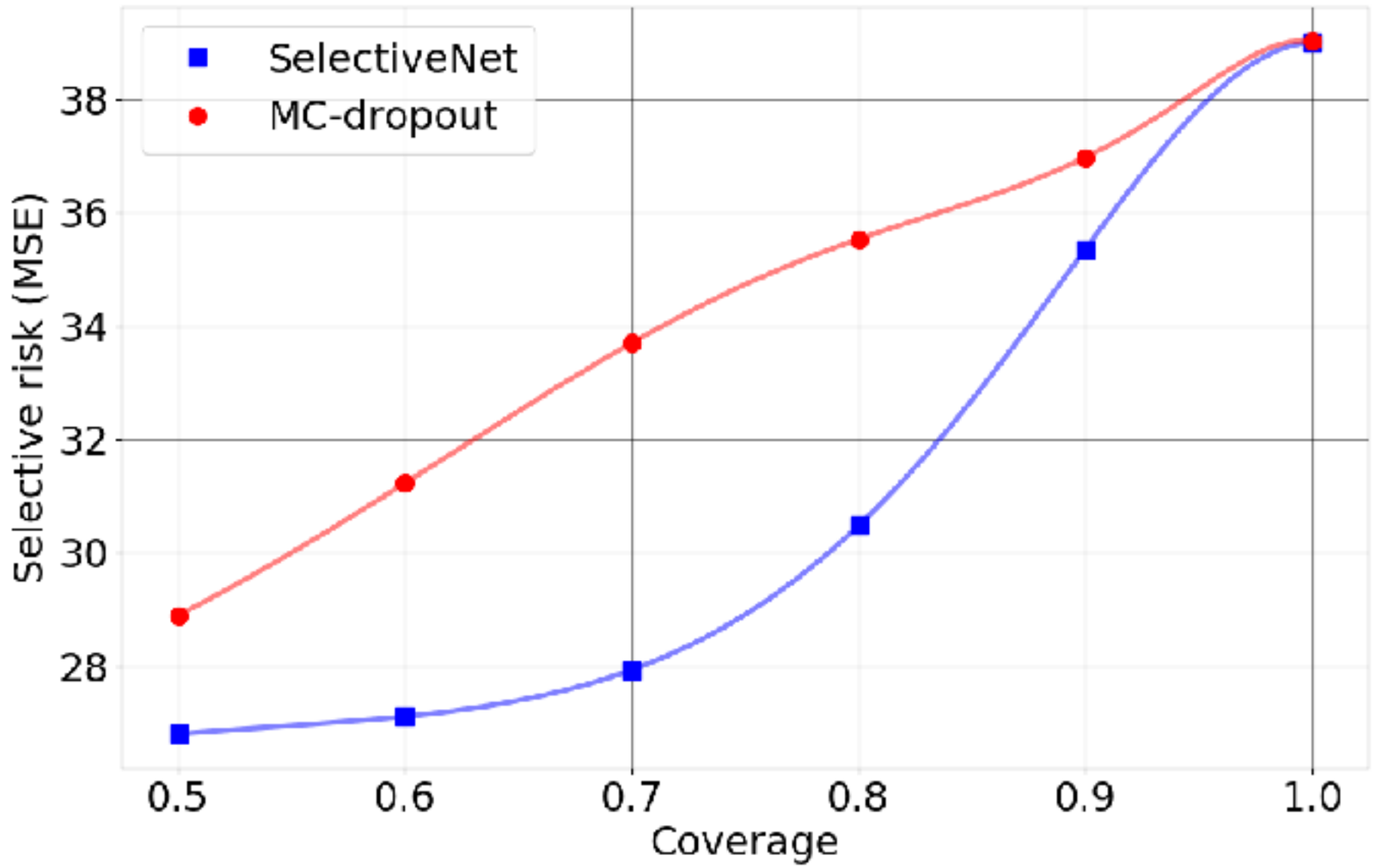
Cats vs. Dogs



40% risk reduction

Empirical Results - Regression

Concrete Compressive Strength (UCI)



Embedding Analysis

SelectiveNet

Threshold

- SelectiveNet does not “invest” representational capacity on rejected instances

Conclusion

- Improved results compared to SR and MC-dropout
- No need of additional training set for calibration
- Selective regression with fast inference

Thank You

Visit us at poster #1 Grand ball room 6:30 PM