Rashomon Capacity: A Metric for Predictive Multiplicity in Classification

Hsiang Hsu   Flavio P. Calmon

John A. Paulson School of Engineering and Applied Science, Harvard University

hsianghsu@g.harvard.edu, flavio@seas.harvard.edu

Rashomon Capacity: A Metric for Predictive Multiplicity in Classification

Hsiang Hsu and Flavio P. Calmon
John A. Paulson School of Engineering and Applied Sciences, Harvard University
hsianghsu@g.harard.edu, flavio@seas.harvard.edu

Abstract

Predictive multiplicity occurs when classification models with statistically indistinguishable performances assign conflicting predictions to individual samples. When used for decision-making in applications of consequence (e.g., lending, education, criminal justice), models developed without regard for predictive multiplicity may result in unjustified and arbitrary decisions for specific individuals. We introduce a new metric, called Rashomon Capacity, to measure predictive multiplicity in probabilistic classification. Prior metrics for predictive multiplicity focus on classifiers that output thresholded (i.e., 0-1) predicted classes. In contrast, Rashomon Capacity applies to probabilistic classifiers, capturing more nuanced score variations for individual samples. We provide a rigorous derivation for Rashomon Capacity, argue its intuitive appeal, and demonstrate how to estimate it in practice. We show that Rashomon Capacity yields principled strategies for disclosing conflicting models to stakeholders. Our numerical experiments illustrate how Rashomon Capacity captures predictive multiplicity in various datasets and learning models, including neural networks. The tools introduced in this paper can help data scientists measure and report predictive multiplicity prior to model deployment.

1 Introduction

Rashomon effect, introduced by Breiman [1], describes the phenomenon where a multitude of distinct predictive models achieve similar training or test loss. Breiman reported observing the Rashomon effect in several model classes, including linear regression, decision trees, and small neural networks. In a foresighted experiment, Breiman noted that, when retraining a neural network 100 times on three-dimensional data with different random initializations, he "*found 32 distinct minima, each of which gave a different picture, and having about equal test set error*" [1, Section 8]. The set of almost-equally performing models for a given learning problem is called the *Rashomon set* [2, 3].

We focus on a facet of the Rashomon effect in classification problems called *predictive multiplicity*. Predictive multiplicity occurs when competing models in the Rashomon set assign conflicting predictions to individual samples [4]. Fig. 1 presents an updated version of Breiman's neural network experiment and illustrates predictive multiplicity in three classification tasks with different data domains and neural network architectures. Here, models that achieve statistically-indistinguishable performance on a test set assign wildly different predictions to an input sample. If predictive multiplicity is not accounted for, the output for this sample may ultimately depend on arbitrary choices made during training (e.g., parameter initialization).

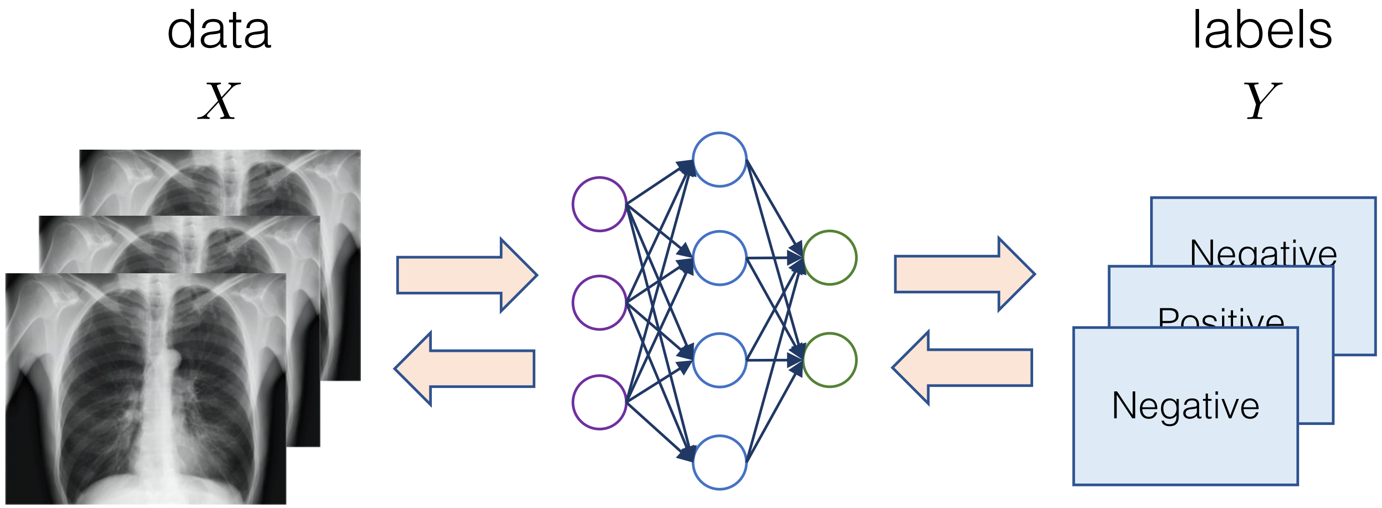
Predictive multiplicity captures the potential individual-level harm introduced by an arbitrary choice of a single model in the Rashomon set. When such a model is used to support automated decision-making in sectors dominated by a few companies or Government—labeled *Algorithmic Leviathans* in [5, Section 3]—predictive multiplicity can lead to unjustified and systemic exclusion of individuals from critical opportunities. For example, an algorithm used for lending may deny a loan to a specific

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

This material is based upon work supported by the National Science Foundation under grants CAREER 1845852, IIS 1926925, and FAI 2040880, and by Meta Ph.D. fellowship.

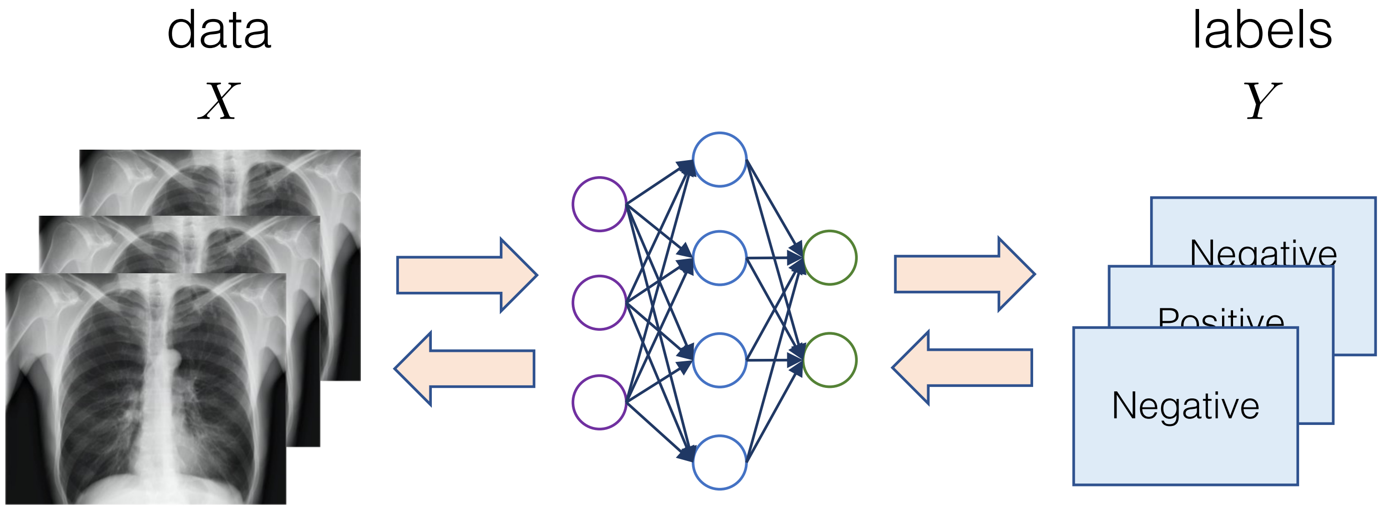
What are the Rashomon Effect and Predictive Multiplicity?

Training Time



What are the Rashomon Effect and Predictive Multiplicity?

Training Time



$h_{\theta_1}(X)$, $\theta_1 \in \Theta$: model parameters (i.e., weights)

What are the Rashomon Effect and Predictive Multiplicity?

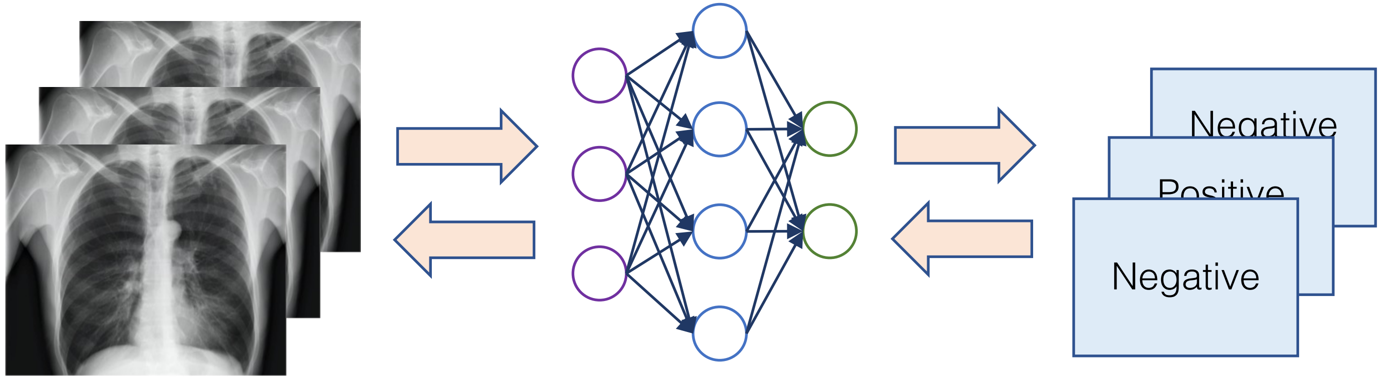
Training Time

data

X

labels

Y

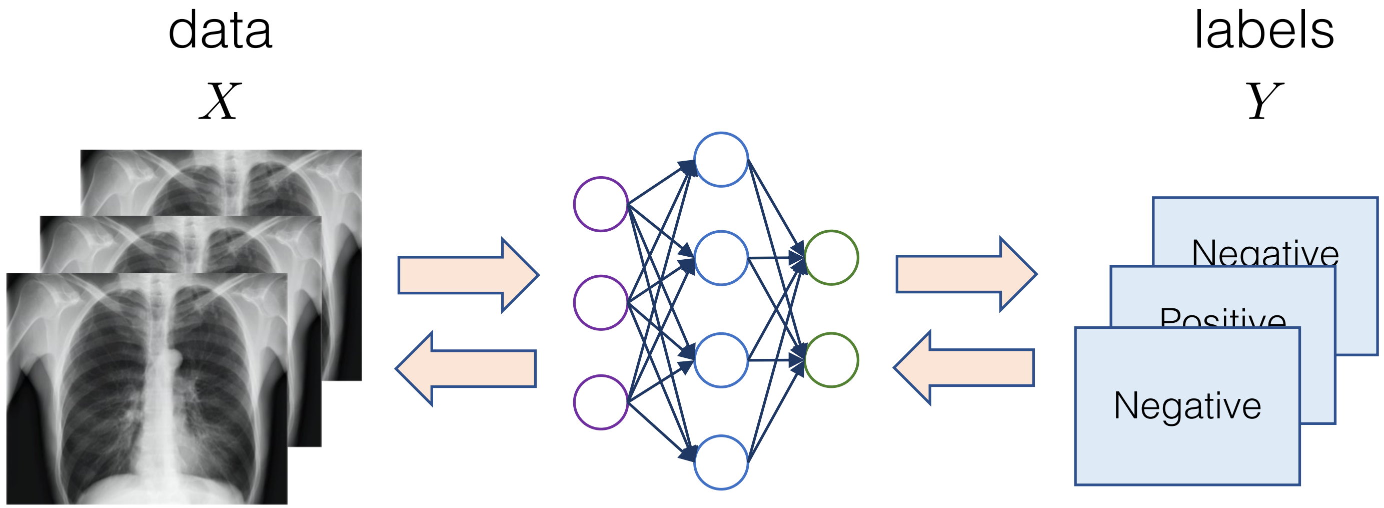


$h_{\theta_1}(X)$, $\theta_1 \in \Theta$: model parameters (i.e., weights)

- Set random seeds
- Initialize weights
- Dropout rates
- Shuffling for batches
- …

What are the Rashomon Effect and Predictive Multiplicity?

Training Time

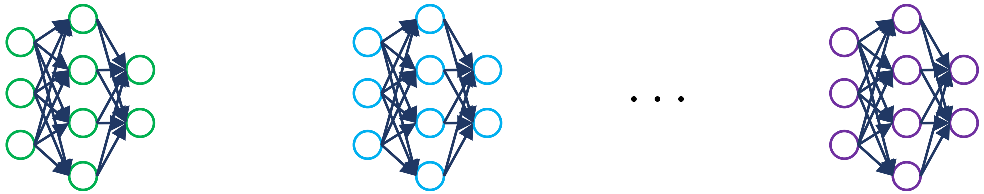


$$h_{\theta_1}(X),\, h_{\theta_2}(X),\, h_{\theta_3}(X),\, \cdots \in \mathcal{H}$$

Random Initializations

What are the Rashomon Effect and Predictive Multiplicity?

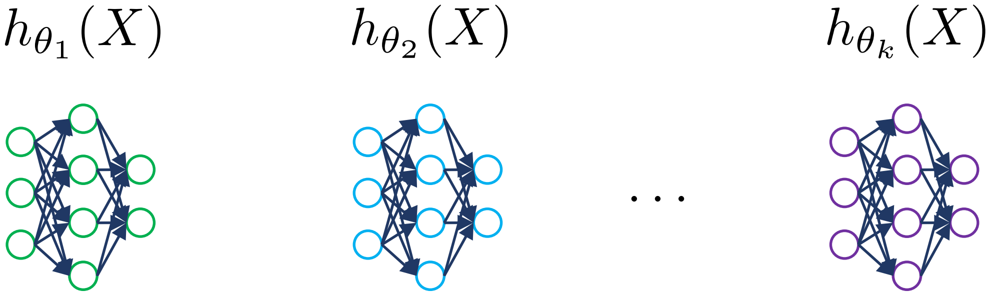
$h_{\theta_1}(X)$ $h_{\theta_2}(X)$ $h_{\theta_k}(X)$



What are the Rashomon Effect and Predictive Multiplicity?

Test Time

test loss: $L(h_\theta) = \mathbb{E}[\ell(h_\theta(X), Y)]$

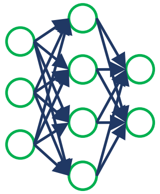


What are the Rashomon Effect and Predictive Multiplicity?

Test Time

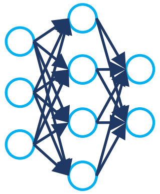
test loss: $L(h_\theta) = \mathbb{E}[\ell(h_\theta(X), Y)]$

$h_{\theta_1}(X)$



$L(h_{\theta_1}) \leq \epsilon$

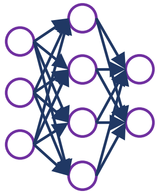
$h_{\theta_2}(X)$



$L(h_{\theta_2}) \leq \epsilon$

$\cdots$

$h_{\theta_k}(X)$

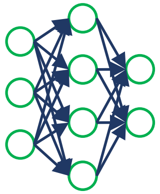
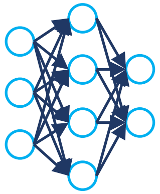
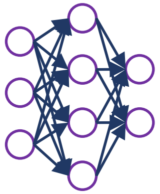


$L(h_{\theta_k}) \leq \epsilon$

What are the Rashomon Effect and Predictive Multiplicity?

Rashomon Effect [Breiman'01]
Many different models have approximately-equal accuracy

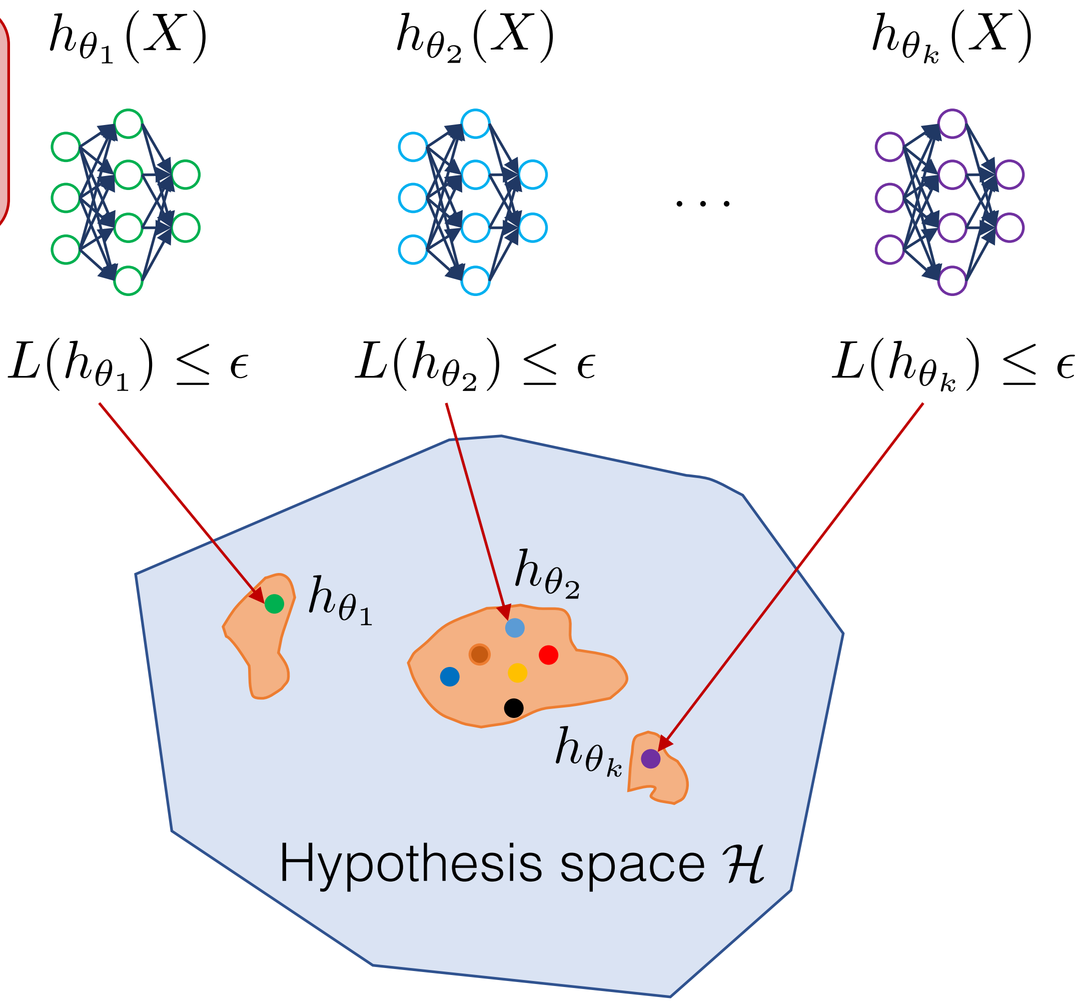
$h_{\theta_1}(X)$ $h_{\theta_2}(X)$ $\dots$ $h_{\theta_k}(X)$

$L(h_{\theta_1}) \leq \epsilon$ $L(h_{\theta_2}) \leq \epsilon$ $L(h_{\theta_k}) \leq \epsilon$

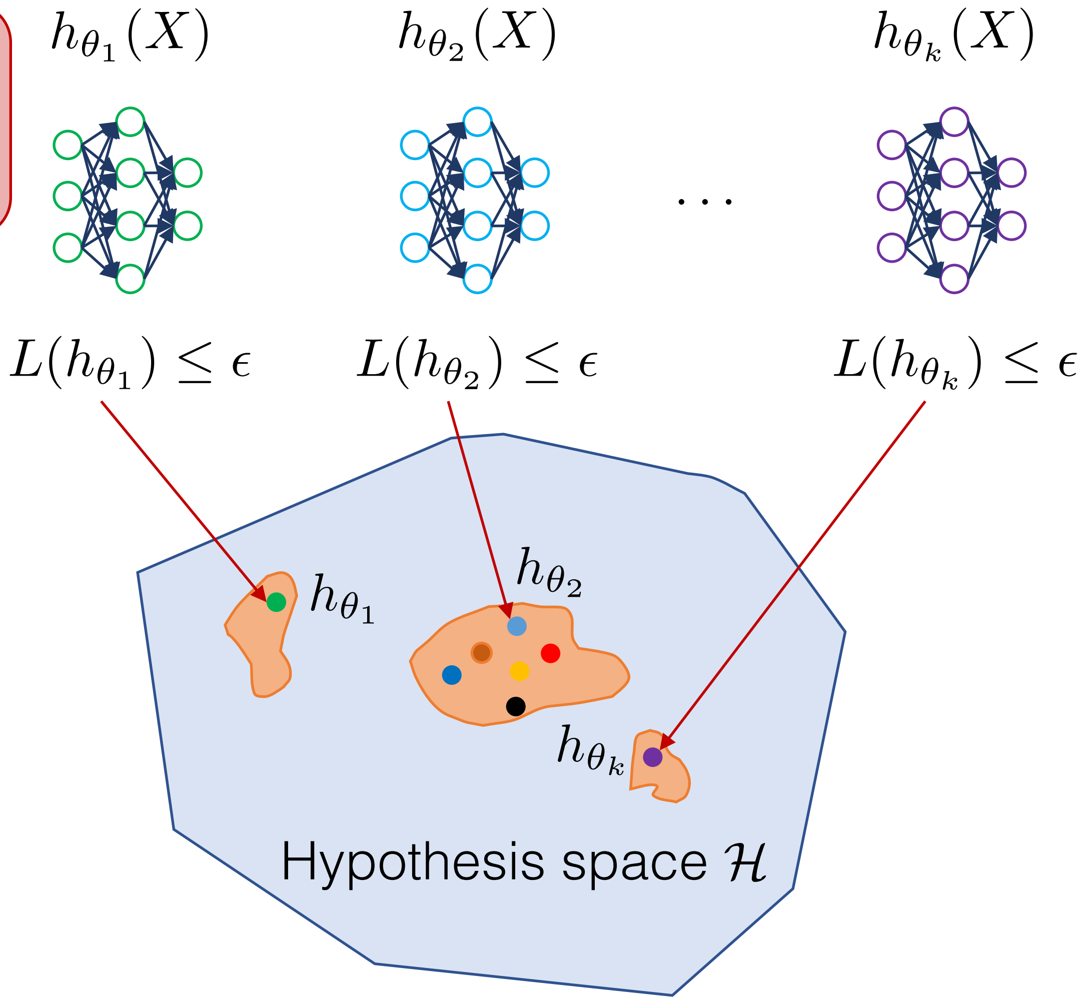
What are the Rashomon Effect and Predictive Multiplicity?

Rashomon Effect [Breiman'01]
Many different models have approximately-equal accuracy



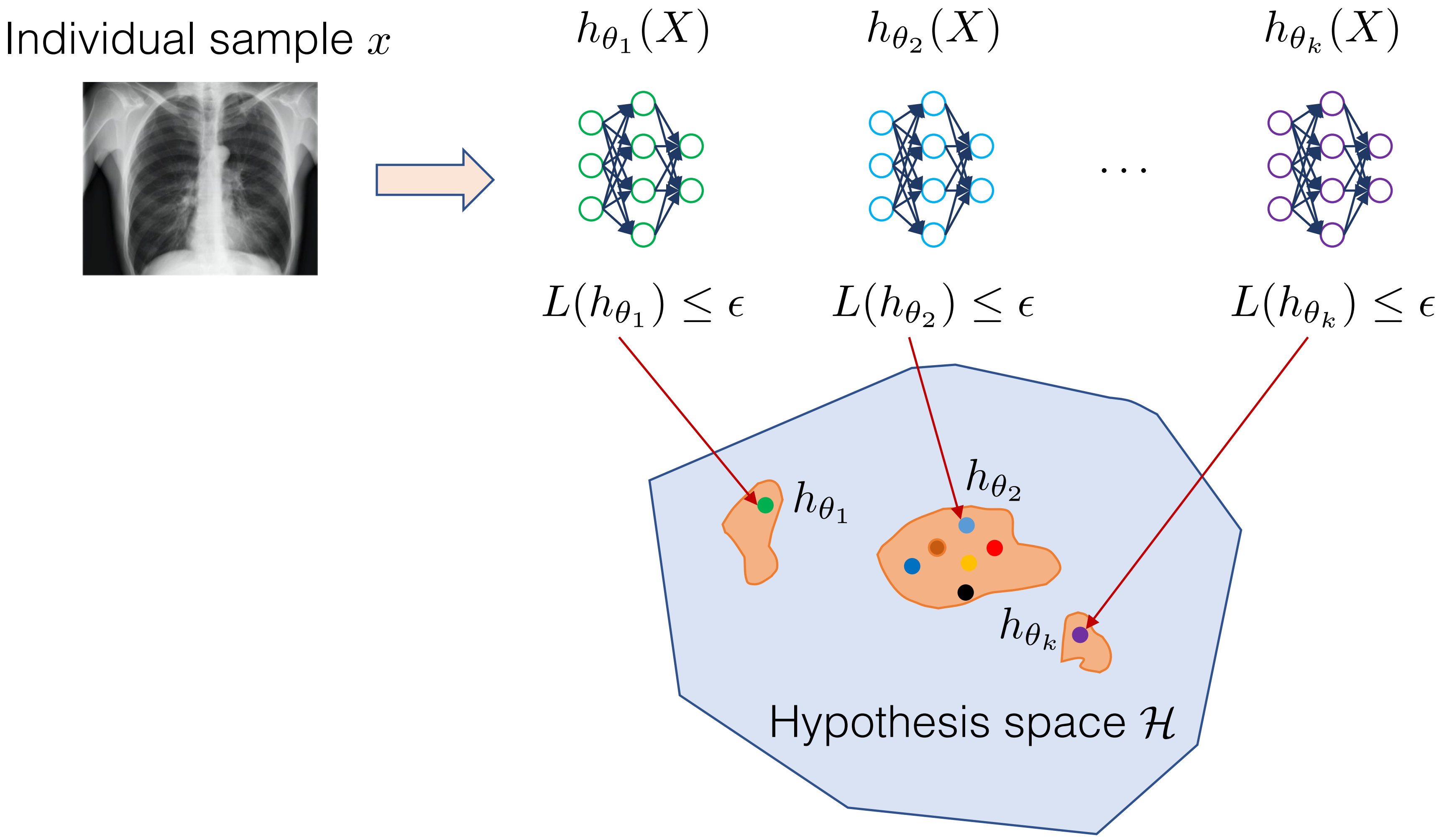
What are the Rashomon Effect and Predictive Multiplicity?

Rashomon Effect [Breiman'01]
Many different models have approximately-equal accuracy



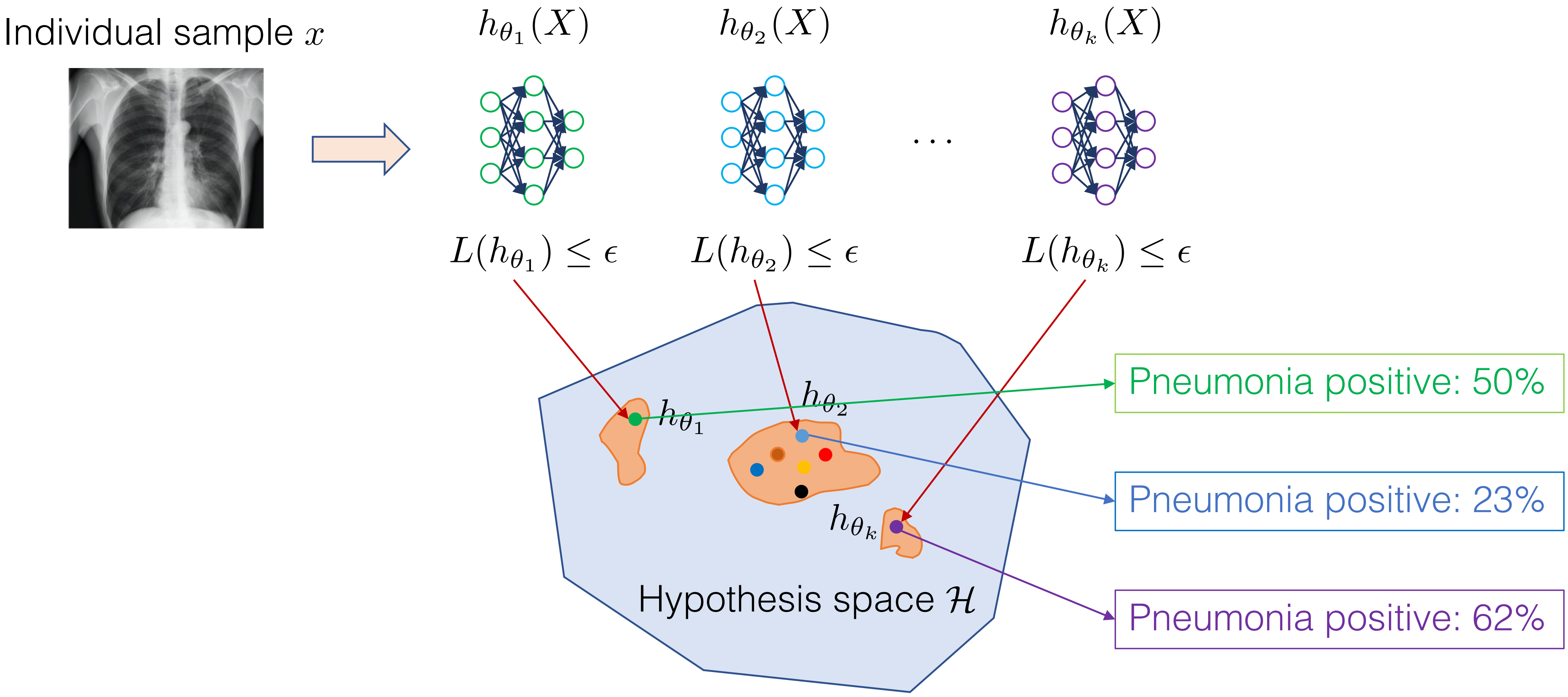
Rashomon set: $\mathcal{R}(\mathcal{H}, \epsilon) \triangleq \{h_\theta \in \mathcal{H}; L(h_\theta) \leq \epsilon\}$

What are the Rashomon Effect and Predictive Multiplicity?



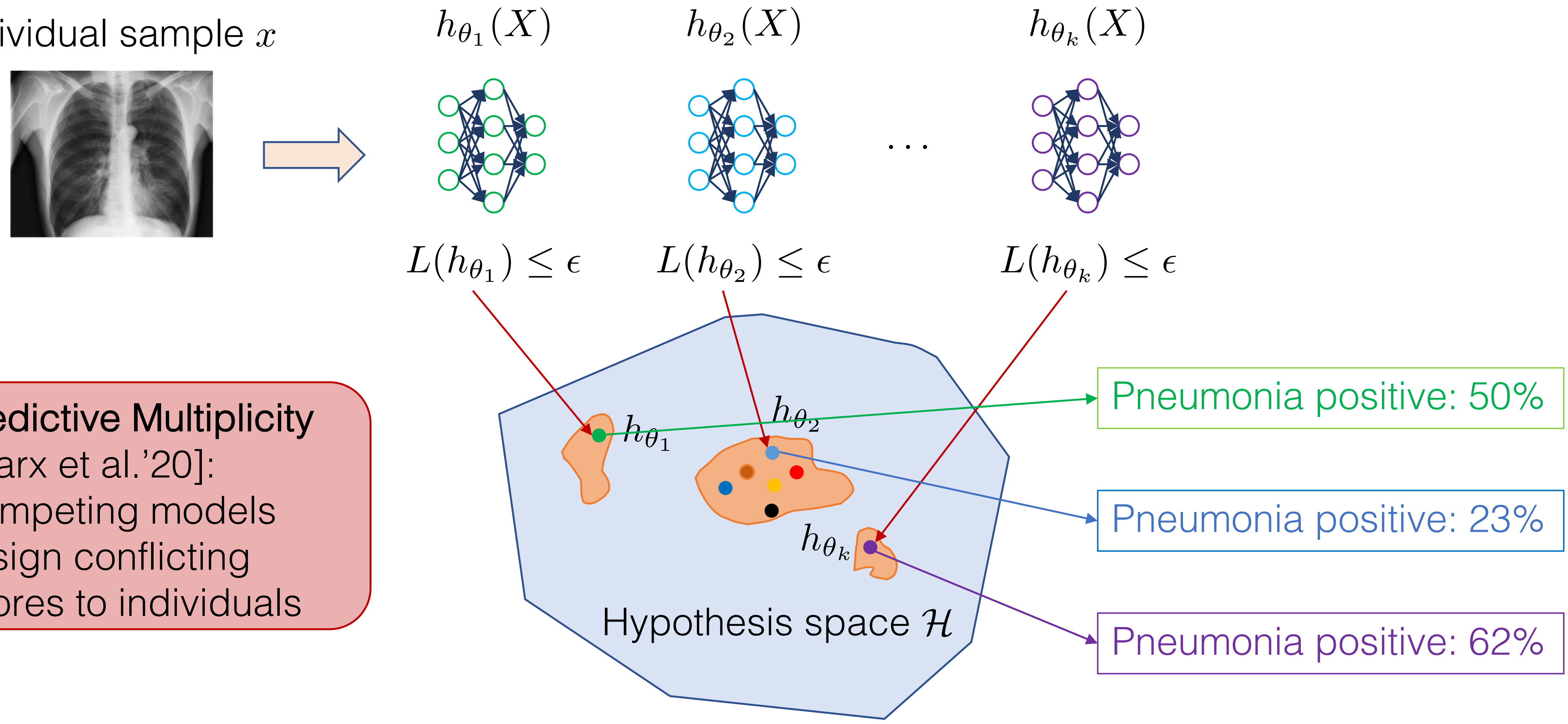
Rashomon set: $\mathcal{R}(\mathcal{H}, \epsilon) \triangleq \{h_\theta \in \mathcal{H}; L(h_\theta) \leq \epsilon\}$

What are the Rashomon Effect and Predictive Multiplicity?



Rashomon set: $\mathcal{R}(\mathcal{H}, \epsilon) \triangleq \{h_\theta \in \mathcal{H}; L(h_\theta) \leq \epsilon\}$

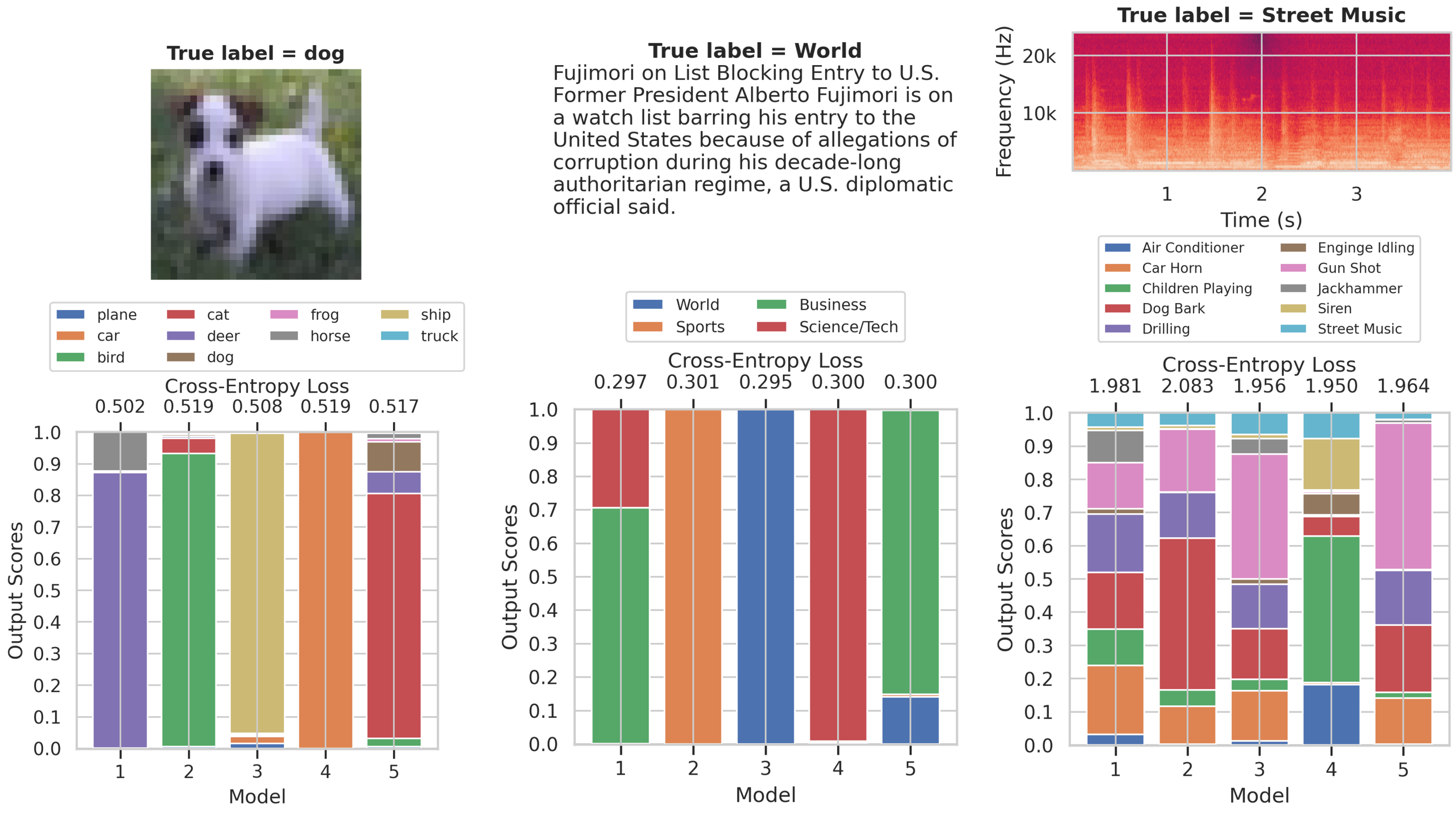
What are the Rashomon Effect and Predictive Multiplicity?



Predictive Multiplicity [Marx et al.'20]: Competing models assign conflicting scores to individuals

Rashomon set: $\mathcal{R}(\mathcal{H}, \epsilon) \triangleq \{h_\theta \in \mathcal{H}; L(h_\theta) \leq \epsilon\}$

Predictive Multiplicity occurs in many classification tasks



(a) CIFAR-10 dataset **(b)** AG News dataset **(c)** UrbanSound8k dataset.

Societal Impacts of Predictive Multiplicity

- If predictive multiplicity is not accounted for, decisions supported by ML models may depend on arbitrary and unjustified choices (e.g., model initialization).

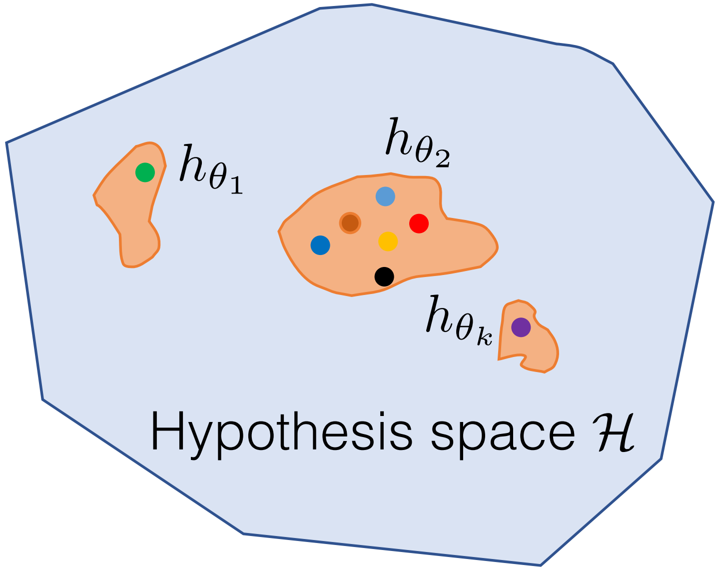
- In sectors dominated by a few algorithms (algorithmic leviathans [Creel&Hellman'21] used in credit scoring, government services), this may lead to arbitrary loss of opportunities to certain individuals:

Medical Service

Education

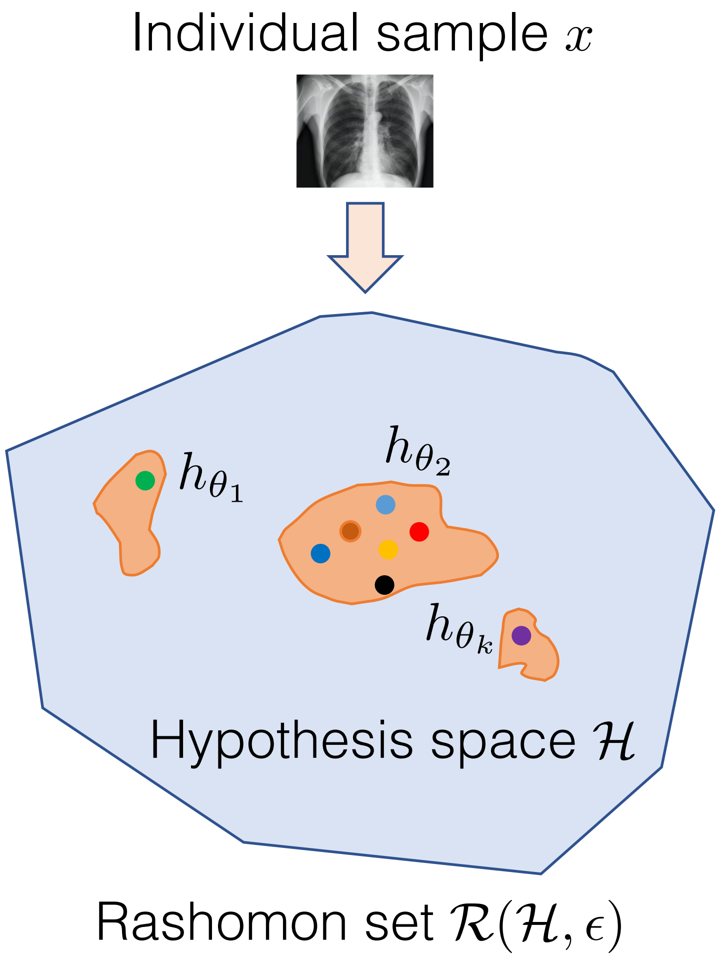
Loans

How to measure predictive multiplicity?



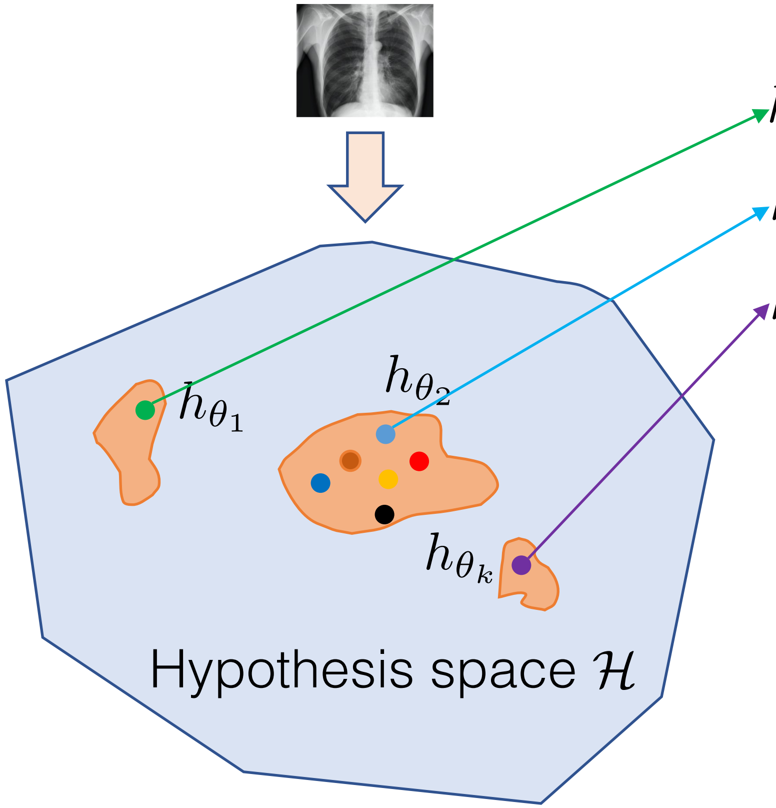
Rashomon set $\mathcal{R}(\mathcal{H}, \epsilon)$

How to measure predictive multiplicity?



How to measure predictive multiplicity?

Individual sample x



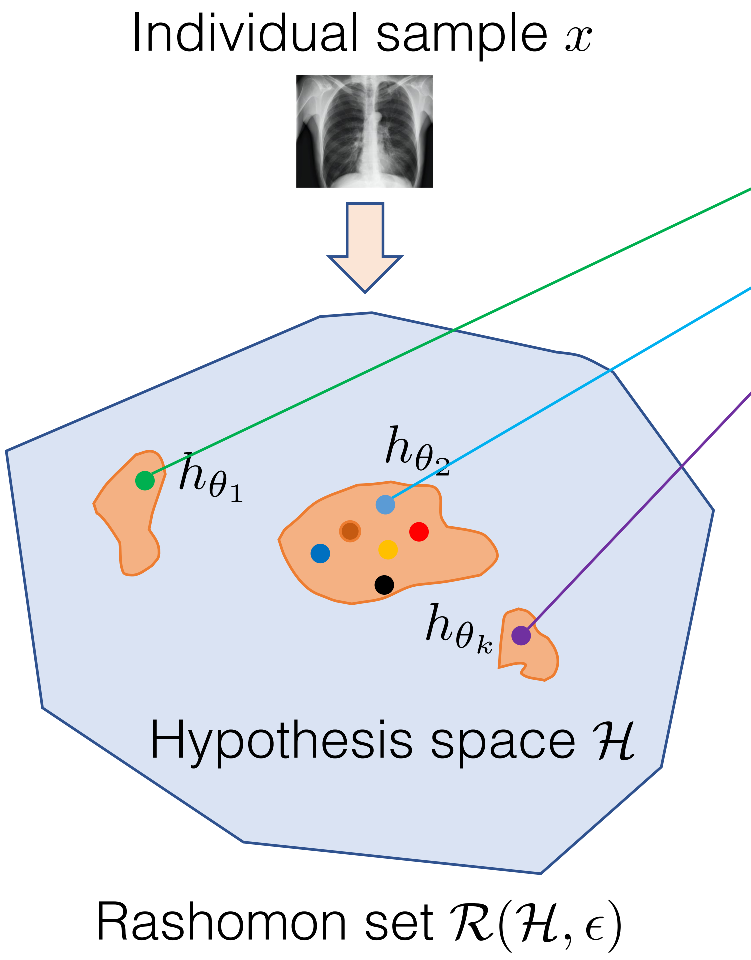
$$h_{\theta_1}(x) = [0.05, 0.90, 0.05] \in \Delta_c$$

$$h_{\theta_2}(x) = [0.62, 0.34, 0.04] \in \Delta_c$$

$$h_{\theta_k}(x) = [0.25, 0.30, 0.45] \in \Delta_c$$

Rashomon set $\mathcal{R}(\mathcal{H}, \epsilon)$

How to measure predictive multiplicity?



$$\mathcal{M}_\epsilon(x) \triangleq \{h_\theta(x); h_\theta \in \mathcal{R}(\mathcal{H}, \epsilon)\}$$

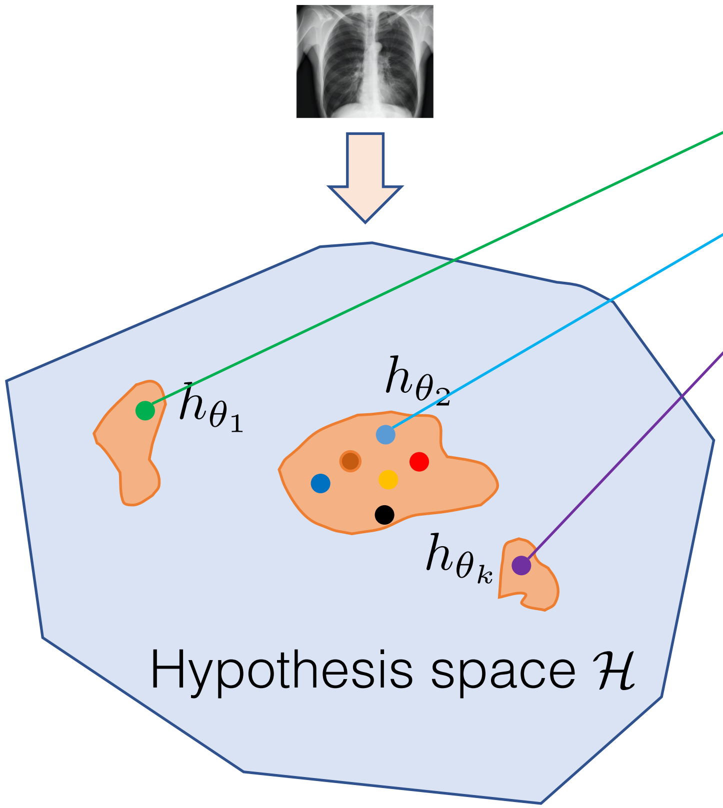
$$h_{\theta_1}(x) = [0.05, 0.90, 0.05] \in \Delta_c$$

$$h_{\theta_2}(x) = [0.62, 0.34, 0.04] \in \Delta_c$$

$$h_{\theta_k}(x) = [0.25, 0.30, 0.45] \in \Delta_c$$

How to measure predictive multiplicity?

Individual sample x



Rashomon set $\mathcal{R}(\mathcal{H}, \epsilon)$

$$\mathcal{M}_\epsilon(x) \triangleq \{h_\theta(x); h_\theta \in \mathcal{R}(\mathcal{H}, \epsilon)\}$$

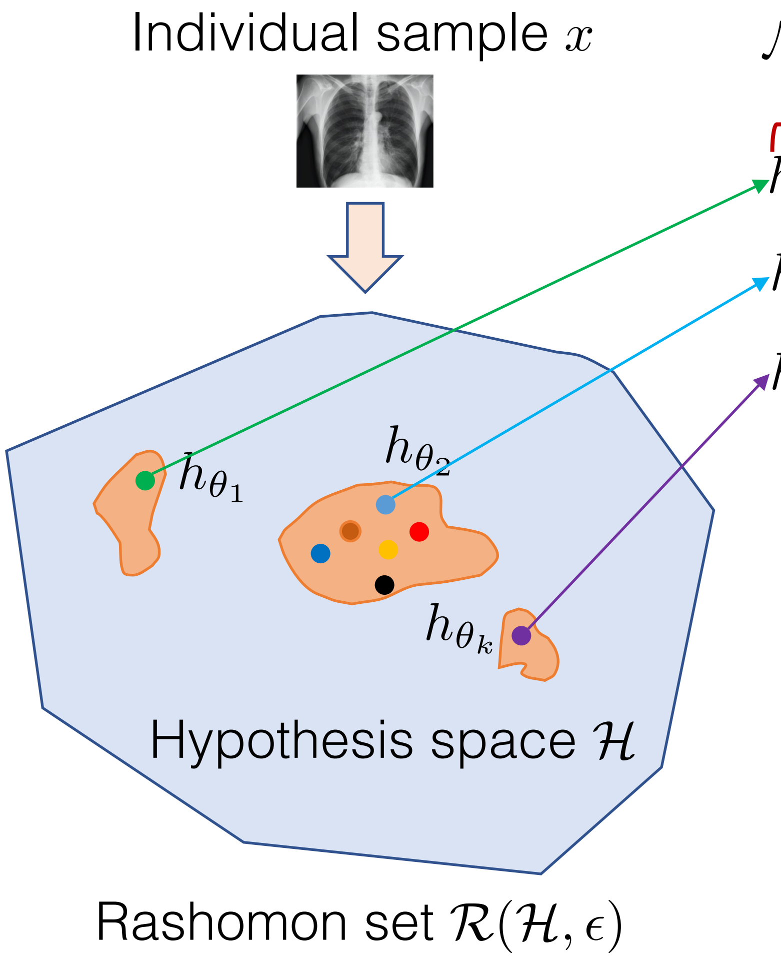
$$h_{\theta_1}(x) = [0.05, 0.90, 0.05] \in \Delta_c$$

$$h_{\theta_2}(x) = [0.62, 0.34, 0.04] \in \Delta_c$$

$$h_{\theta_k}(x) = [0.25, 0.30, 0.45] \in \Delta_c$$

How do we measure the score variations?

How to measure predictive multiplicity?



$$\mathcal{M}_\epsilon(x) \triangleq \{h_\theta(x); h_\theta \in \mathcal{R}(\mathcal{H}, \epsilon)\}$$

$$h_{\theta_1}(x) = [0.05, 0.90, 0.05] \in \Delta_c$$

$$h_{\theta_2}(x) = [0.62, 0.34, 0.04] \in \Delta_c$$

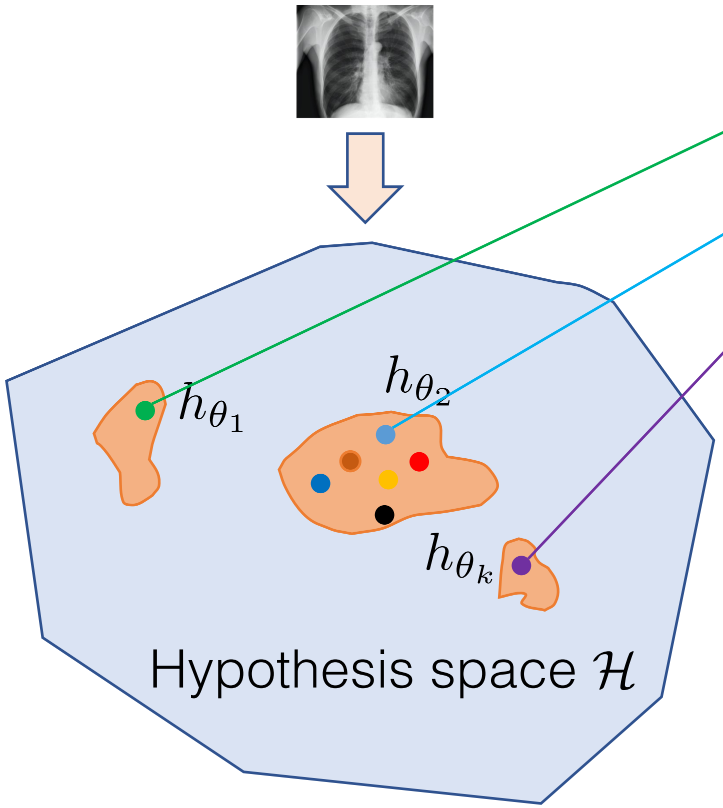
$$h_{\theta_k}(x) = [0.25, 0.30, 0.45] \in \Delta_c$$

How do we measure the score variations?

$$m : \mathcal{M}_\epsilon(x) \to \mathbb{R}^+$$

How to measure predictive multiplicity?

Individual sample x



Rashomon set $\mathcal{R}(\mathcal{H}, \epsilon)$

$$\mathcal{M}_\epsilon(x) \triangleq \{h_\theta(x); h_\theta \in \mathcal{R}(\mathcal{H}, \epsilon)\}$$

$$h_{\theta_1}(x) = [0.05, 0.90, 0.05] \in \Delta_c$$

$$h_{\theta_2}(x) = [0.62, 0.34, 0.04] \in \Delta_c$$

$$h_{\theta_k}(x) = [0.25, 0.30, 0.45] \in \Delta_c$$

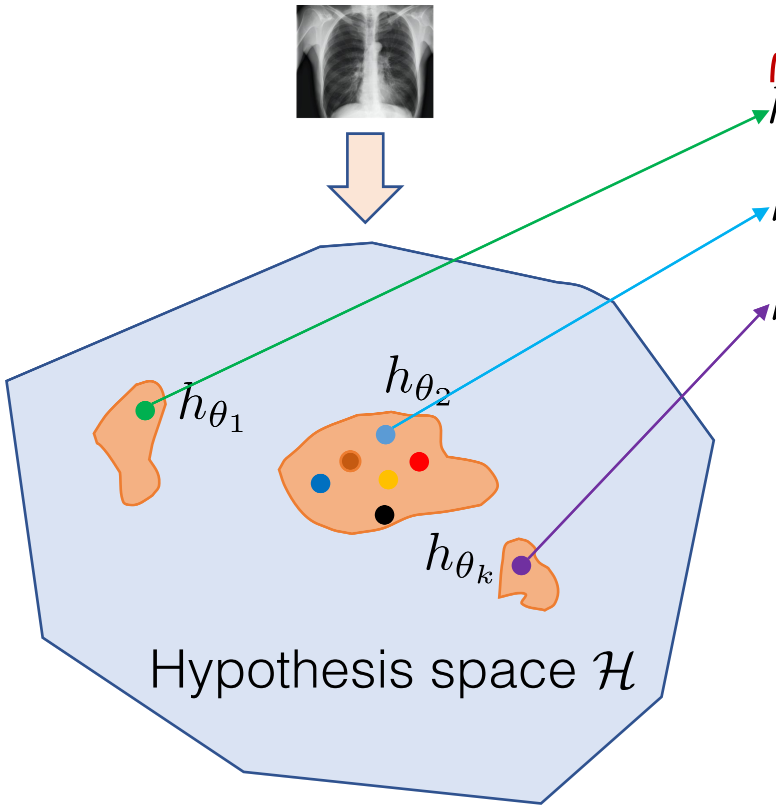
How do we measure the score variations?

$$m : \mathcal{M}_\epsilon(x) \to \mathbb{R}^+$$

Desirable Properties

How to measure predictive multiplicity?

Individual sample x



Rashomon set $\mathcal{R}(\mathcal{H}, \epsilon)$

$$\mathcal{M}_\epsilon(x) \triangleq \{h_\theta(x); h_\theta \in \mathcal{R}(\mathcal{H}, \epsilon)\}$$

$$h_{\theta_1}(x) = [0.05, 0.90, 0.05] \in \Delta_c$$

$$h_{\theta_2}(x) = [0.62, 0.34, 0.04] \in \Delta_c$$

$$h_{\theta_k}(x) = [0.25, 0.30, 0.45] \in \Delta_c$$

How do we measure the score variations?

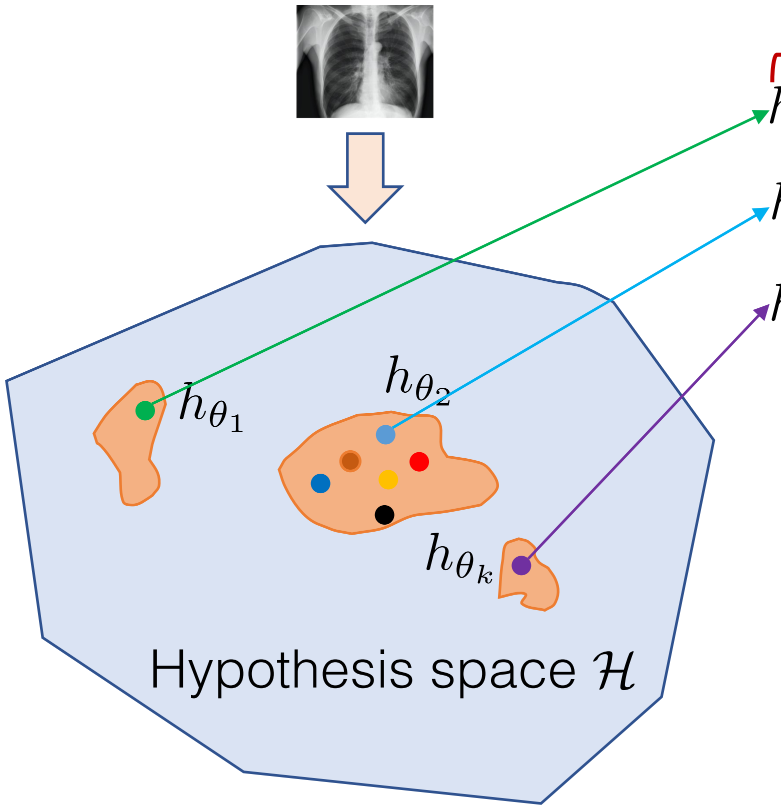
$$m : \mathcal{M}_\epsilon(x) \rightarrow \mathbb{R}^+$$

Desirable Properties

1. $1 \leq m(x) \leq c$

How to measure predictive multiplicity?

Individual sample x



Rashomon set $\mathcal{R}(\mathcal{H}, \epsilon)$

$$\mathcal{M}_\epsilon(x) \triangleq \{h_\theta(x); h_\theta \in \mathcal{R}(\mathcal{H}, \epsilon)\}$$

$$h_{\theta_1}(x) = [0.05, 0.90, 0.05] \in \Delta_c$$

$$h_{\theta_2}(x) = [0.05, 0.90, 0.05] \in \Delta_c$$

$$h_{\theta_k}(x) = [0.05, 0.90, 0.05] \in \Delta_c$$

How do we measure the score variations?

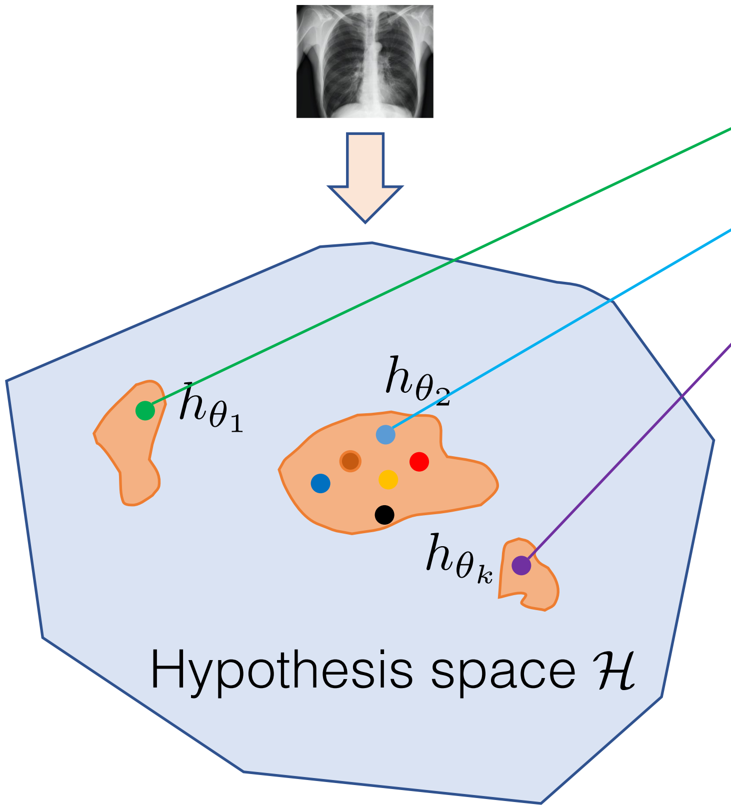
$$m : \mathcal{M}_\epsilon(x) \to \mathbb{R}^+$$

Desirable Properties

1. $1 \leq m(x) \leq c$
2. $m(x) = 1 \Rightarrow$ predictions from all models match

How to measure predictive multiplicity?

Individual sample x



Rashomon set $\mathcal{R}(\mathcal{H}, \epsilon)$

$$\mathcal{M}_\epsilon(x) \triangleq \{h_\theta(x); h_\theta \in \mathcal{R}(\mathcal{H}, \epsilon)\}$$

$$h_{\theta_1}(x) = [\underline{1.00}, 0.00, 0.00] \in \Delta_c$$

$$h_{\theta_2}(x) = [0.00, \underline{1.00}, 0.00] \in \Delta_c$$

$$h_{\theta_k}(x) = [0.00, 0.00, \underline{1.00}] \in \Delta_c$$

How do we measure the score variations?

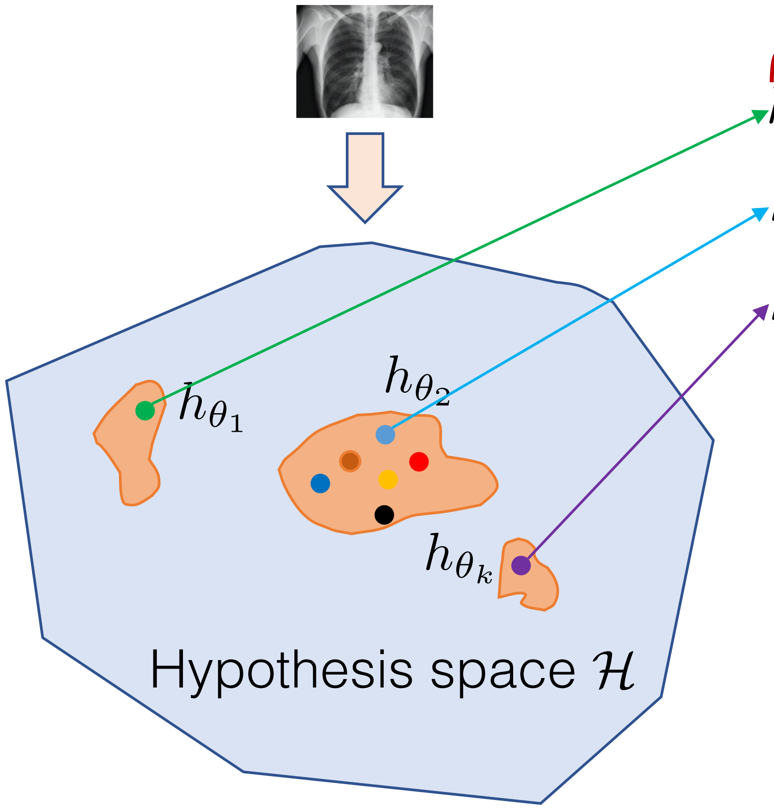
$$m : \mathcal{M}_\epsilon(x) \to \mathbb{R}^+$$

Desirable Properties

1. $1 \leq m(x) \leq c$
2. $m(x) = 1 \Rightarrow$ predictions from all models match
3. $m(x) = c \Rightarrow$ there are models in the Rashomon Set; that assign each of the c classes

How to measure predictive multiplicity?

Individual sample x



Rashomon set $\mathcal{R}(\mathcal{H}, \epsilon)$

$$\mathcal{M}_\epsilon(x) \triangleq \{h_\theta(x); h_\theta \in \mathcal{R}(\mathcal{H}, \epsilon)\}$$

$$h_{\theta_1}(x) = [0.05, 0.90, 0.05] \in \Delta_c$$

$$h_{\theta_2}(x) = [0.62, 0.34, 0.04] \in \Delta_c$$

$$h_{\theta_k}(x) = [0.25, 0.30, 0.45] \in \Delta_c$$

How do we measure the score variations?

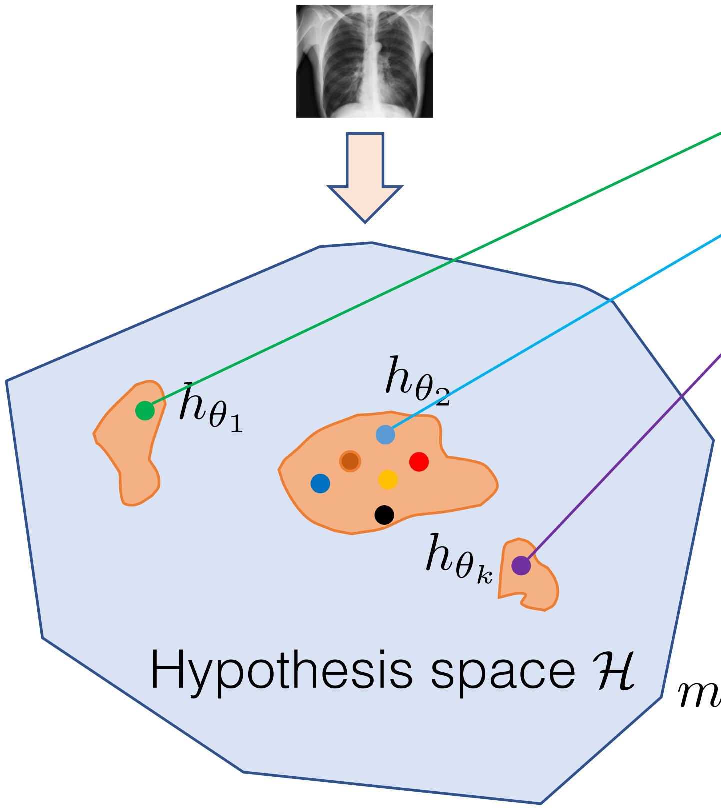
$$m : \mathcal{M}_\epsilon(x) \to \mathbb{R}^+$$

Desirable Properties

1. $1 \leq m(x) \leq c$
2. $m(x) = 1 \Rightarrow$ predictions from all models match
3. $m(x) = c \Rightarrow$ there are models in the Rashomon Set; that assign each of the c classes
4. Monotonic in $|\mathcal{R}(\mathcal{H}, \epsilon)|$

How to measure predictive multiplicity?

Individual sample x



Rashomon set $\mathcal{R}(\mathcal{H}, \epsilon)$

$$\mathcal{M}_\epsilon(x) \triangleq \{h_\theta(x); h_\theta \in \mathcal{R}(\mathcal{H}, \epsilon)\}$$

$$h_{\theta_1}(x) = [0.05, 0.90, 0.05] \in \Delta_c$$

$$h_{\theta_2}(x) = [0.62, 0.34, 0.04] \in \Delta_c$$

$$h_{\theta_k}(x) = [0.25, 0.30, 0.45] \in \Delta_c$$

$$h_{\theta_{k+1}}(x) = [0.83, 0.08, 0.09] \in \Delta_c$$

$$m(\{h_{\theta_1}, \cdots, h_{\theta_k}\}) \leq m(\{h_{\theta_1}, \cdots, h_{\theta_{k+1}}\})$$

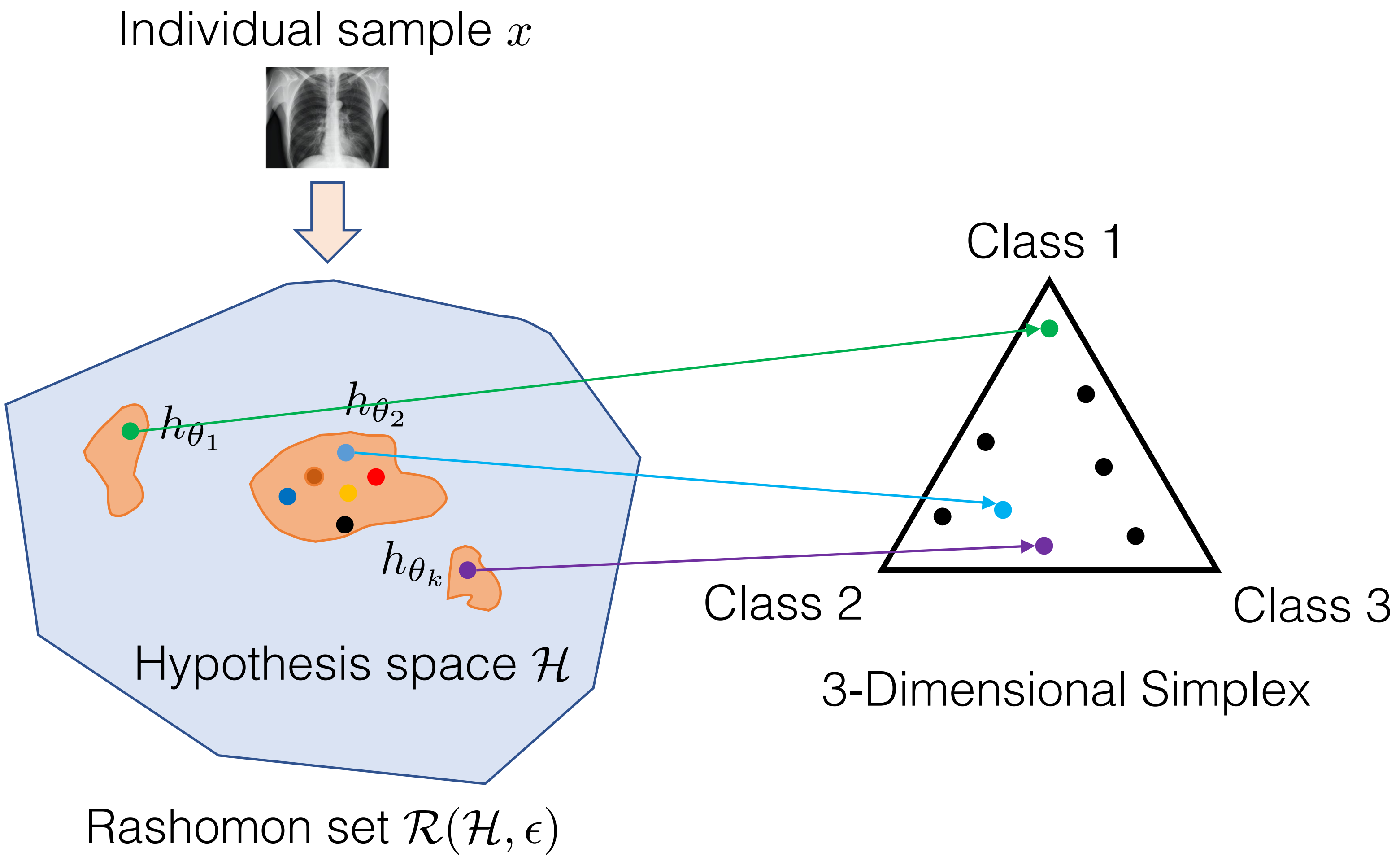
How do we measure the score variations?

$$m : \mathcal{M}_\epsilon(x) \rightarrow \mathbb{R}^+$$

Desirable Properties

1. $1 \leq m(x) \leq c$
2. $m(x) = 1 \Rightarrow$ predictions from all models match
3. $m(x) = c \Rightarrow$ there are models in the Rashomon Set; that assign each of the c classes
4. Monotonic in $|\mathcal{R}(\mathcal{H}, \epsilon)|$

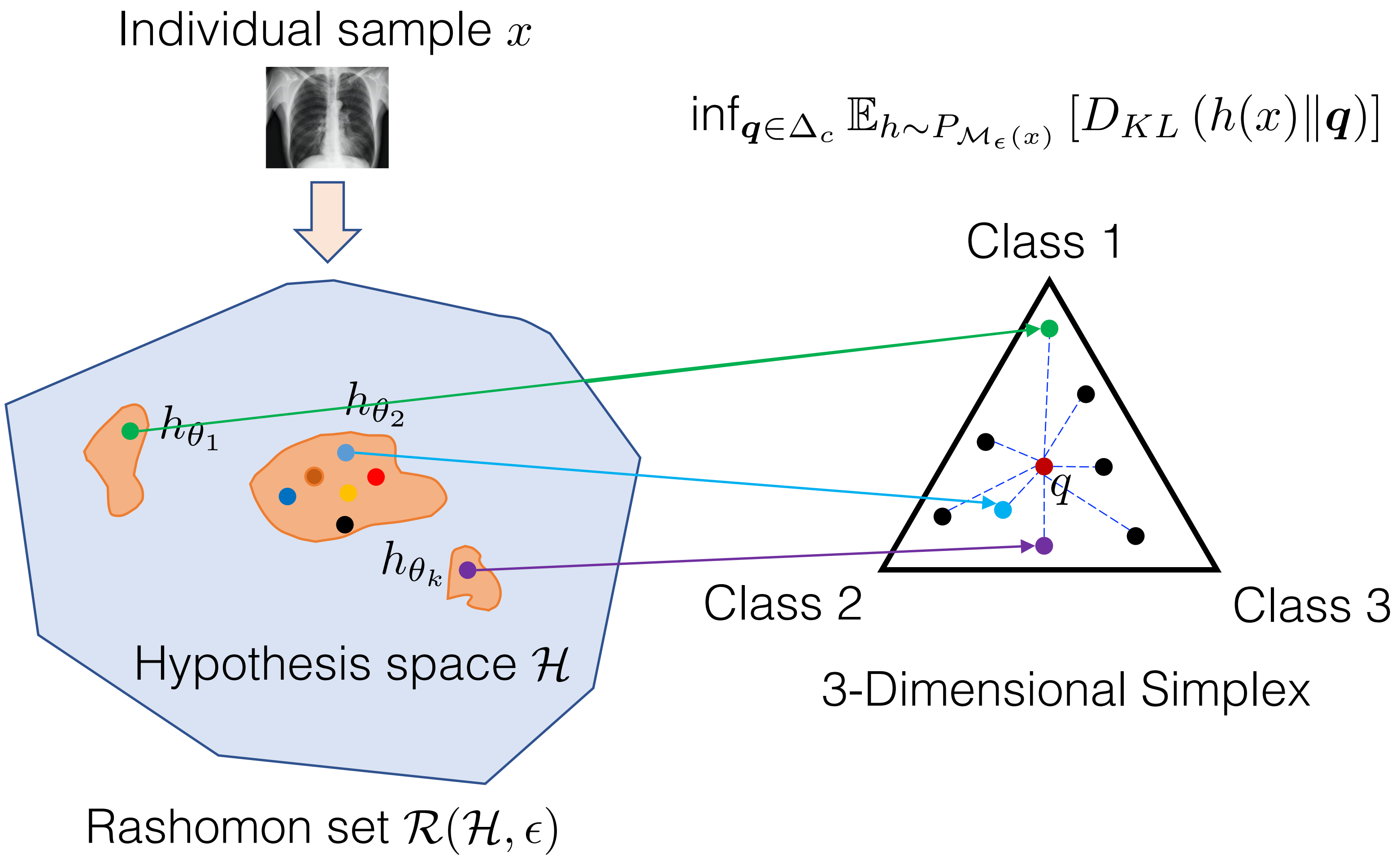
How to measure predictive multiplicity?



Desirable Properties

1. $1 \leq m(x) \leq c$
2. $m(x) = 1 \Rightarrow$ predictions from all models match
3. $m(x) = c \Rightarrow$ there are models in the Rashomon Set; that assign each of the c classes
4. Monotonic in $|\mathcal{R}(\mathcal{H}, \epsilon)|$

How to measure predictive multiplicity?

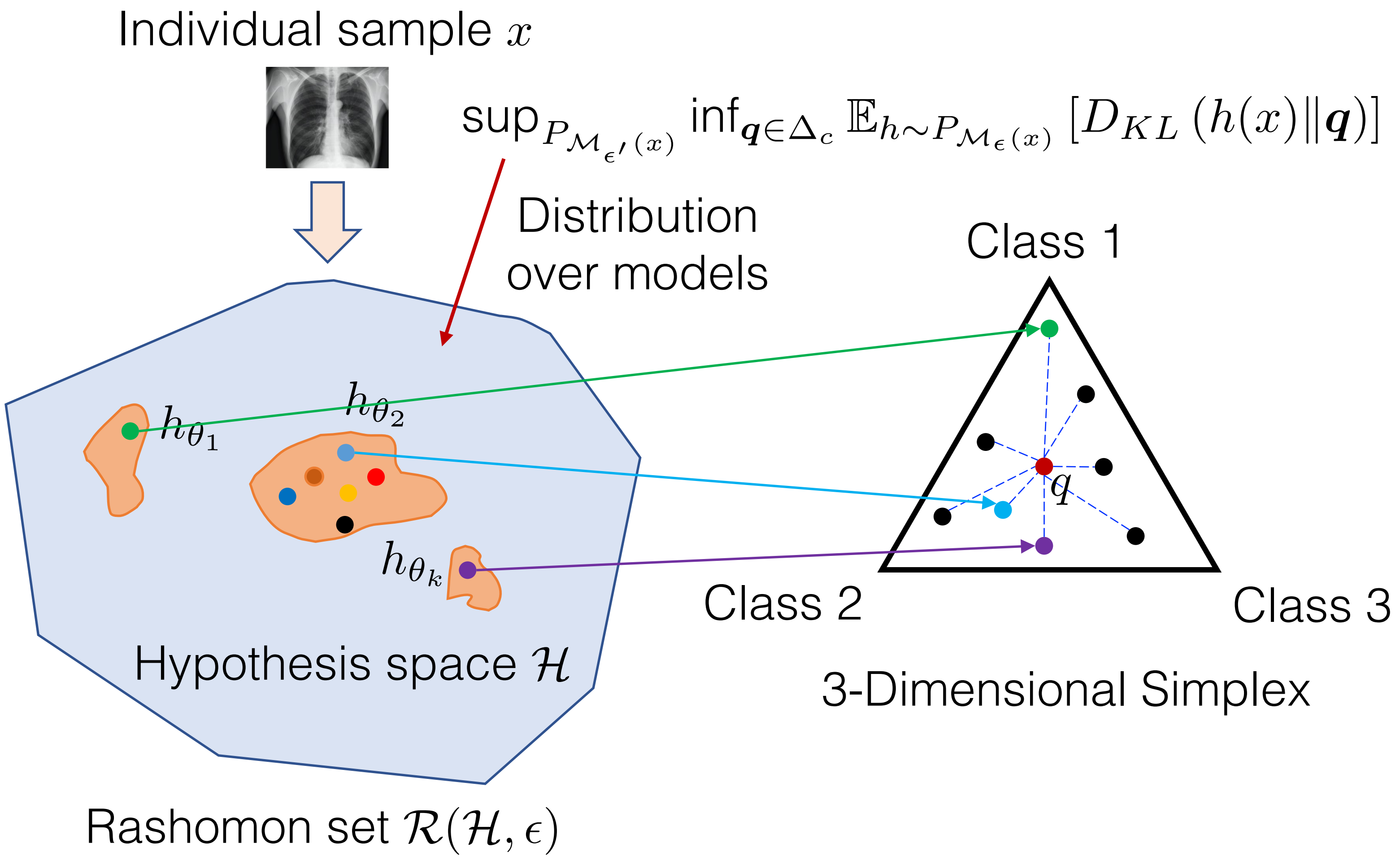


Desirable Properties

1. $1 \le m(x) \le c$
2. $m(x) = 1 \Rightarrow$ predictions from all models match
3. $m(x) = c \Rightarrow$ there are models in the Rashomon Set; that assign each of the c classes
4. Monotonic in $|\mathcal{R}(\mathcal{H},\epsilon)|$

How to measure predictive multiplicity?

$$\sup_{P_{\mathcal{M}_{\epsilon'}(x)}} \inf_{\boldsymbol{q}\in\Delta_c} \mathbb{E}_{h\sim P_{\mathcal{M}_{\epsilon}(x)}} \left[D_{KL}\left(h(x)\|\boldsymbol{q}\right)\right]$$



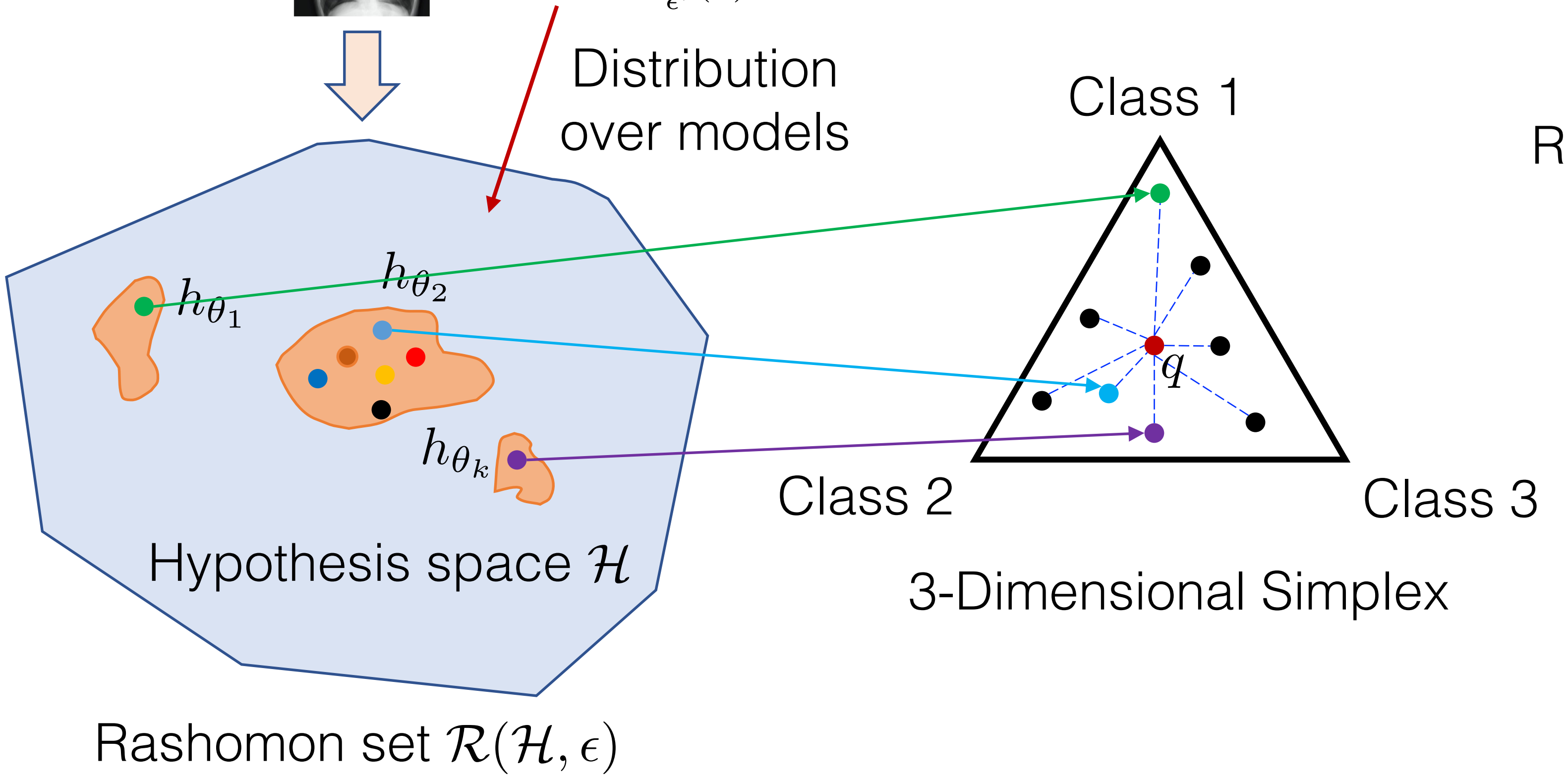
Desirable Properties

1. $1 \leq m(x) \leq c$
2. $m(x) = 1 \Rightarrow$ predictions from all models match
3. $m(x) = c \Rightarrow$ there are models in the Rashomon Set; that assign each of the c classes
4. Monotonic in $|\mathcal{R}(\mathcal{H},\epsilon)|$

How to measure predictive multiplicity?

Individual sample x

$$\sup_{P_{\mathcal{M}_{\epsilon'}(x)}} \inf_{\boldsymbol{q}\in\Delta_c} \mathbb{E}_{h\sim P_{\mathcal{M}_\epsilon(x)}} \left[D_{KL}\left(h(x)\|\boldsymbol{q}\right)\right] \triangleq C_\epsilon(x)$$

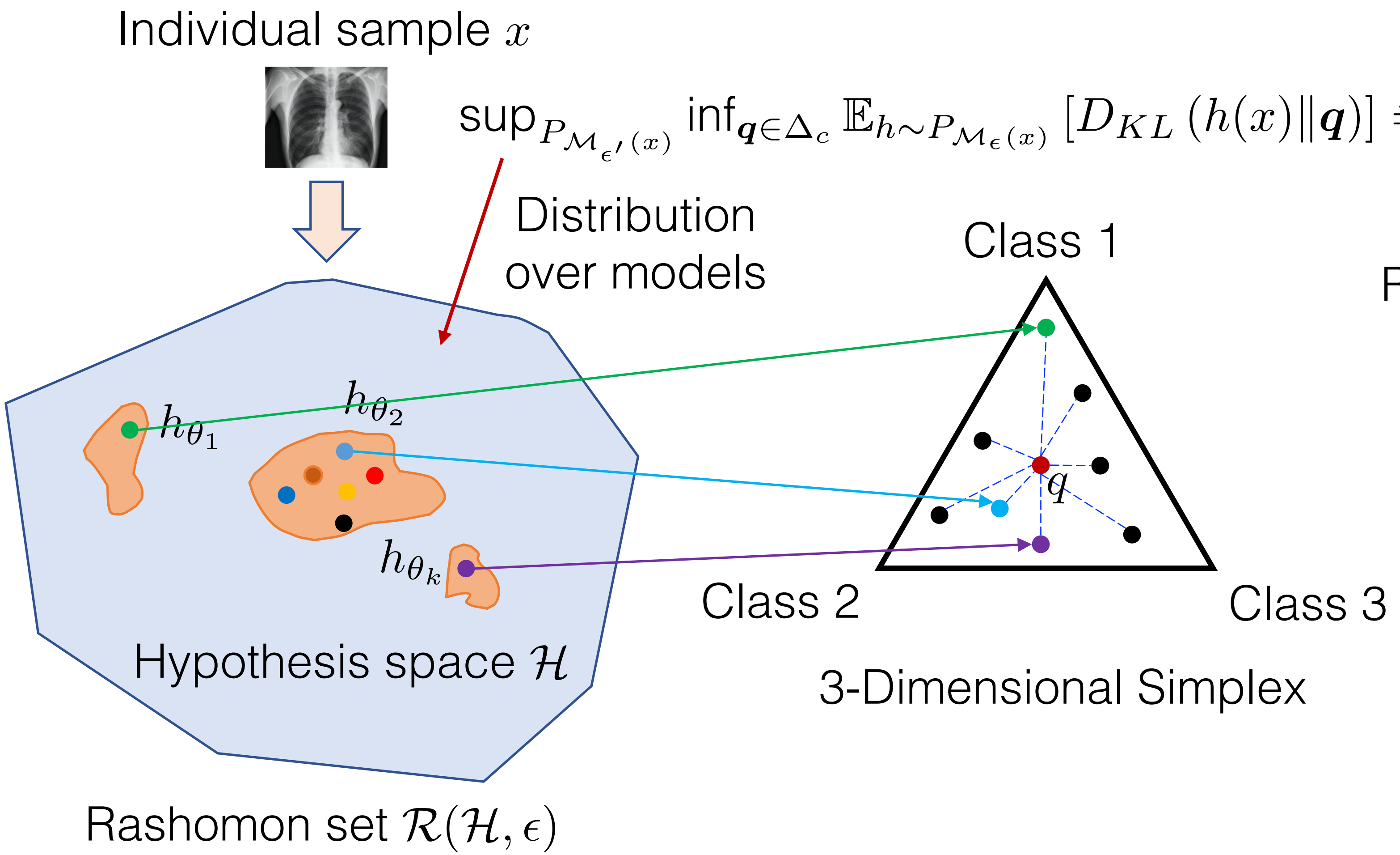


Rashomon Capacity: $m_\epsilon(x) \triangleq 2^{C_\epsilon(x)}$

Desirable Properties

1. $1 \leq m(x) \leq c$
2. $m(x) = 1 \Rightarrow$ predictions from all models match
3. $m(x) = c \Rightarrow$ there are models in the Rashomon Set; that assign each of the c classes
4. Monotonic in $|\mathcal{R}(\mathcal{H},\epsilon)|$

How to measure predictive multiplicity?



$$\sup_{P_{\mathcal{M}_{\epsilon'}(x)}} \inf_{\boldsymbol{q} \in \Delta_c} \mathbb{E}_{h \sim P_{\mathcal{M}_\epsilon(x)}} \left[D_{KL}\left(h(x) \| \boldsymbol{q}\right)\right] \triangleq C_\epsilon(x)$$

Rashomon Capacity: $m_\epsilon(x) \triangleq 2^{C_\epsilon(x)}$

Desirable Properties

1. $1 \leq m(x) \leq c$
2. $m(x) = 1 \Rightarrow$ predictions from all models match
3. $m(x) = c \Rightarrow$ there are models in the Rashomon Set; that assign each of the c classes
4. Monotonic in $|\mathcal{R}(\mathcal{H}, \epsilon)|$

How to approximate Rashomon Capacity in practice?

Two core challenges: **Rashomon set**: $\mathcal{R}(\mathcal{H}, \epsilon) \triangleq \{h_\theta \in \mathcal{H}; L(h_\theta) \leq \epsilon\}$

How to approximate Rashomon Capacity in practice?

Two core challenges: **Rashomon set**: $\mathcal{R}(\mathcal{H}, \epsilon) \triangleq \{h_\theta \in \mathcal{H}; L(h_\theta) \leq \epsilon\}$

1. How to choose ϵ?

How to approximate Rashomon Capacity in practice?

Two core challenges: **Rashomon set**: $\mathcal{R}(\mathcal{H}, \epsilon) \triangleq \{h_\theta \in \mathcal{H}; L(h_\theta) \leq \epsilon\}$

1. How to choose ϵ?
2. Approximating the Rashomon set without exhaustively searching?

How to approximate Rashomon Capacity in practice?

Two core challenges: **Rashomon set**: $\mathcal{R}(\mathcal{H}, \epsilon) \triangleq \{h_\theta \in \mathcal{H}; L(h_\theta) \leq \epsilon\}$

1. How to choose ϵ?
2. Approximating the Rashomon set without exhaustively searching?

Using a reference model and approximating the true Rashomon set by a Rashomon subset

$$\widetilde{\mathcal{R}}(\mathcal{H}, \epsilon') \triangleq \{h_{\theta_i} \in \mathcal{H}; L(h_{\theta_i}) \leq \hat{L}(h_{\theta^*}) + \epsilon'\}_{i=1}^{K} \subseteq \mathcal{R}(\mathcal{H}, \hat{L}(h_{\theta^*}) + \epsilon')$$

How to approximate Rashomon Capacity in practice?

Two core challenges: **Rashomon set**: $\mathcal{R}(\mathcal{H}, \epsilon) \triangleq \{h_\theta \in \mathcal{H}; L(h_\theta) \leq \epsilon\}$

1. How to choose ϵ?
2. Approximating the Rashomon set without exhaustively searching?

Using a reference model and approximating the true Rashomon set by a Rashomon subset

$$\widetilde{\mathcal{R}}(\mathcal{H}, \epsilon') \triangleq \{h_{\theta_i} \in \mathcal{H}; L(h_{\theta_i}) \leq \hat{L}(h_{\theta^*}) + \epsilon'\}_{i=1}^{K} \subseteq \mathcal{R}(\mathcal{H}, \hat{L}(h_{\theta^*}) + \epsilon')$$

Rashomon Capacity: $C_\epsilon(x) \triangleq \sup_{P_{\widetilde{\mathcal{M}}_{\epsilon'}(x)}} \inf_{\boldsymbol{q} \in \Delta_c} \mathbb{E}_{h \sim P_{\widetilde{\mathcal{M}}_{\epsilon'}(x)}} [D_{KL}(h(x) \| \boldsymbol{q})] \Rightarrow$ Blahut-Arimoto algorithms

How to approximate Rashomon Capacity in practice?

Two core challenges: **Rashomon set**: $\mathcal{R}(\mathcal{H}, \epsilon) \triangleq \{h_\theta \in \mathcal{H}; L(h_\theta) \leq \epsilon\}$

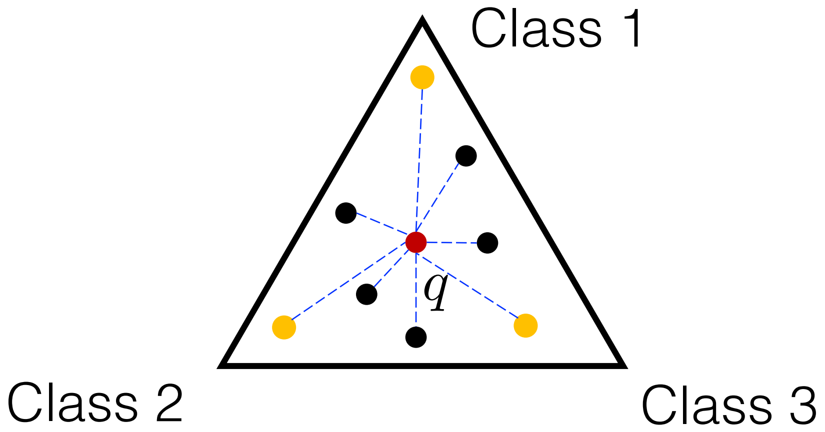
1. How to choose ϵ?
2. Approximating the Rashomon set without exhaustively searching?

Using a reference model and approximating the true Rashomon set by a Rashomon subset

$$\widetilde{\mathcal{R}}(\mathcal{H}, \epsilon') \triangleq \{h_{\theta_i} \in \mathcal{H}; L(h_{\theta_i}) \leq \hat{L}(h_{\theta^*}) + \epsilon'\}_{i=1}^{K} \subseteq \mathcal{R}(\mathcal{H}, \hat{L}(h_{\theta^*}) + \epsilon')$$

Rashomon Capacity: $C_\epsilon(x) \triangleq \sup_{P_{\widetilde{\mathcal{M}}_{\epsilon'}(x)}} \inf_{\boldsymbol{q}\in\Delta_c} \mathbb{E}_{h\sim P_{\widetilde{\mathcal{M}}_{\epsilon'}(x)}} \left[D_{KL}\left(h(x)\|\boldsymbol{q}\right)\right]$ $\Rightarrow$ Blahut-Arimoto algorithms

For each sample x, here are at most c models in a Rashomon subset $\widetilde{\mathcal{R}}(\mathcal{H}, \epsilon)$ whose output scores yield the same Rashomon Capacity for x as the entire Rashomon set.



Takeaway

Rashomon Capacity: A Metric for Predictive Multiplicity in Classification

Takeaway

Rashomon Capacity: A Metric for Predictive Multiplicity in Classification

1. How to **measure** predictive multiplicity?
   - ➢ Rashomon Capacity, first time evaluated on neural networks

Takeaway

Rashomon Capacity: A Metric for Predictive Multiplicity in Classification

1. How to **measure** predictive multiplicity?
   - Rashomon Capacity, first time evaluated on neural networks

2. How to **approximate** a predictive multiplicity metric?
   - Select a Rashomon subset with the most score variation

Takeaway

Rashomon Capacity: A Metric for Predictive Multiplicity in Classification

1. How to **measure** predictive multiplicity?
   - ➢ Rashomon Capacity, first time evaluated on neural networks

2. How to **approximate** a predictive multiplicity metric?
   - ➢ Select a Rashomon subset with the most score variation

3. How to **report** and "**resolve**" predictive multiplicity?
   - ➢ Please check empirical studies in the paper!

Takeaway

Rashomon Capacity: A Metric for Predictive Multiplicity in Classification

1. How to **measure** predictive multiplicity?
   - Rashomon Capacity, first time evaluated on neural networks

2. How to **approximate** a predictive multiplicity metric?
   - Select a Rashomon subset with the most score variation

3. How to **report** and “**resolve**” predictive multiplicity?
   - Please check empirical studies in the paper!

Full paper

Thank you for listening!
Please stop by our poster if you are interested!