Partial Transportability for Domain Generalization

Kasra Jalaldoust*
Columbia University

Alexis Bellot*
Google DeepMind

Elias Bareinboim
Columbia University

00 Preface

The Research Landscape

Guided by Empirical Evaluations – *Data*

"*There is increasing concern that in modern research, false findings may be the vast majority of published research claims*" – **John P. A. Ioannidis**

"*Well-designed experiments demonstrate the effect only for those specific conditions which the experimental and control group have in common, e.g., same geographical region, historical moment, orientation of the stars, orientation in the magnetic field, etc.*" – **Donald T. Campbell**

"*A learner that makes no a priori assumptions regarding the identity of the target concept has no rational basis for classifying any unseen instances*" – **Tom Mitchell**

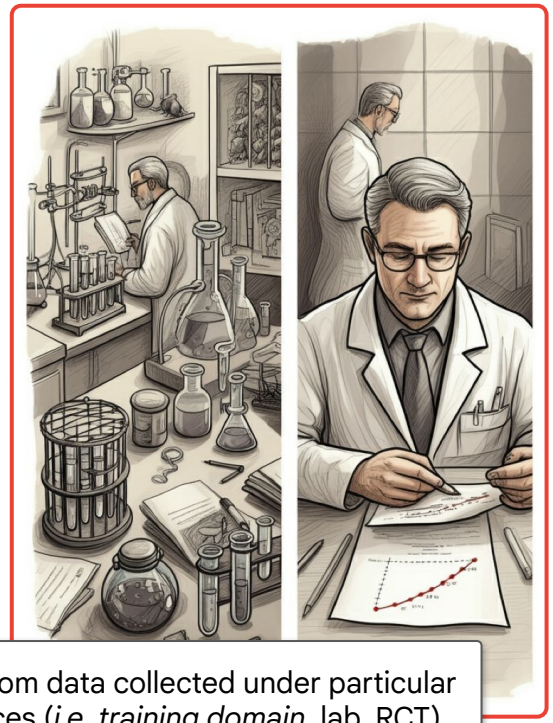
Training and Deploying AI

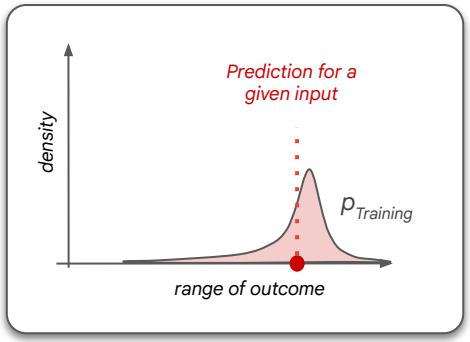
Predicting outcomes for patients (i.e. *test domain*) that may **differ** in multiple aspects from the training domain.

Learning from data collected under particular circumstances (*i.e. training domain*, lab, RCT)

Training and Deploying AI

Predicting outcomes for patients (i.e. *test domain*) that may **differ** in multiple aspects from the training domain.

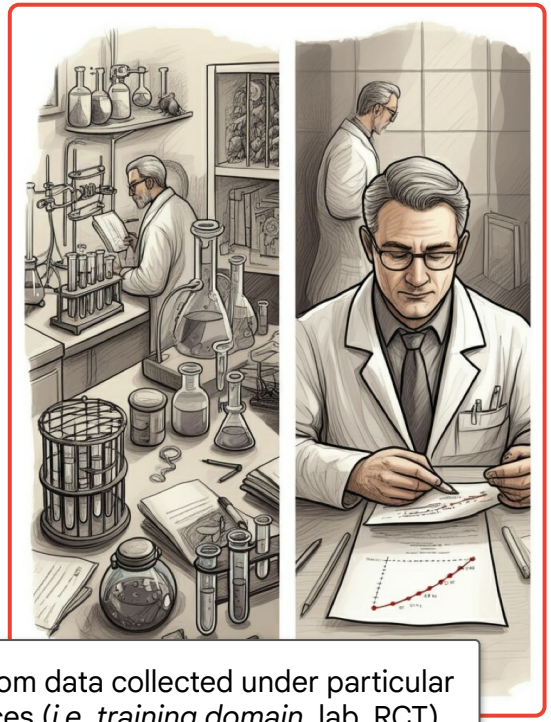


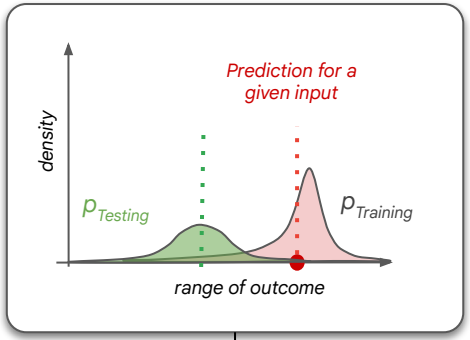


Learning from data collected under particular circumstances (*i.e. training domain*, lab, RCT)

Training and Deploying AI

Predicting outcomes for patients (i.e. *test domain*) that may **differ** in multiple aspects from the training domain.

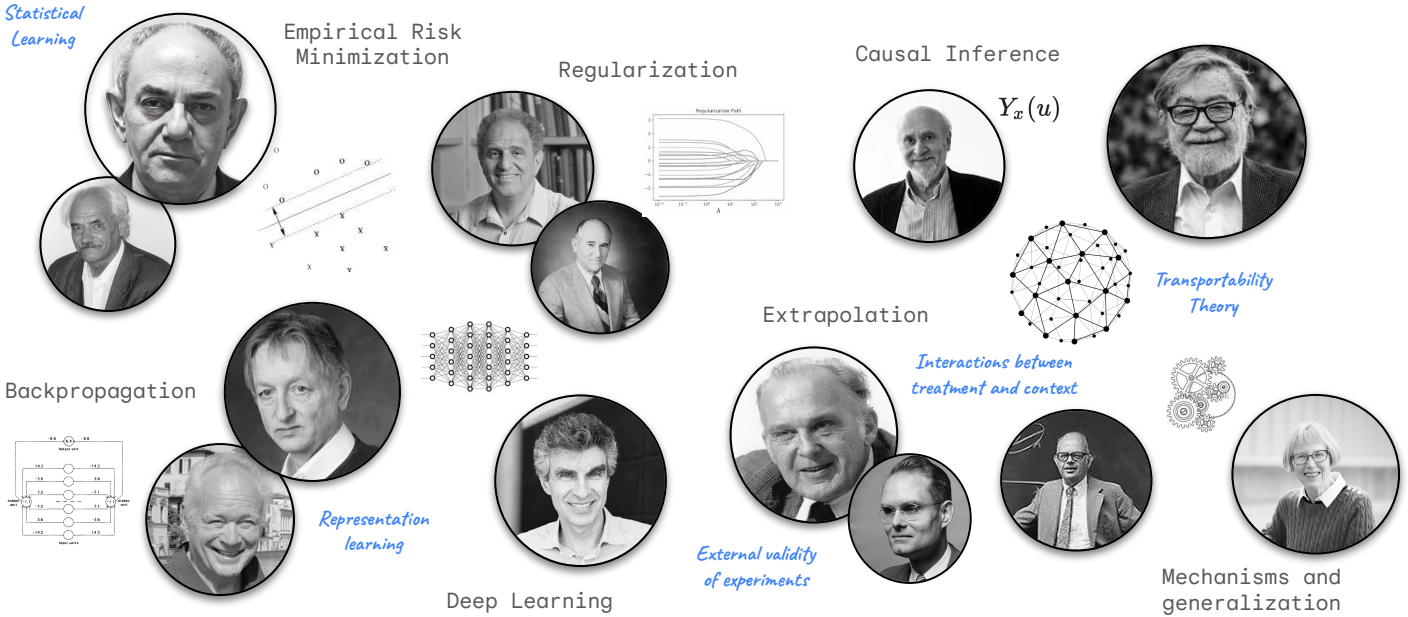




Learning from data collected under particular circumstances (*i.e. training domain*, lab, RCT)

Without an understanding of the relationship between domains, ***extrapolation*** or ***generalization*** is impossible.

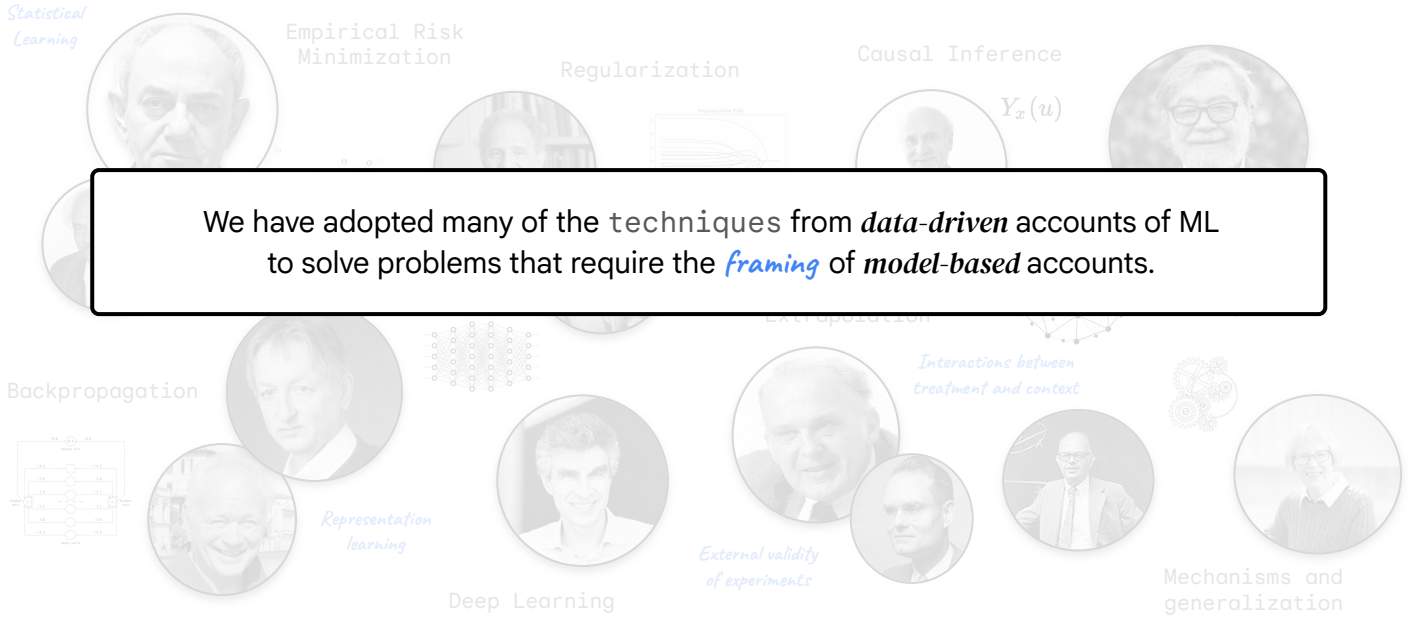
Paradigms in the Data Sciences



Data-driven
1970s - Present

Model-based
1950s - Present

Paradigms in the Data Sciences



We have adopted many of the `techniques` from ***data-driven*** accounts of ML to solve problems that require the *framing* of ***model-based*** accounts.

Data-driven
1970s - Present

Model-based
1950s - Present

Clarifying a Subtle Point…

*Focusing on **optimal training prediction** is worthwhile.*

*There is social, economic, and scientific **benefit** in these pursuits.*

*But neglecting the context in which models are trained is a **limiting perspective**.*

Lab

vs.

Real use case

01

A Causal Perspective

Transportability Problem

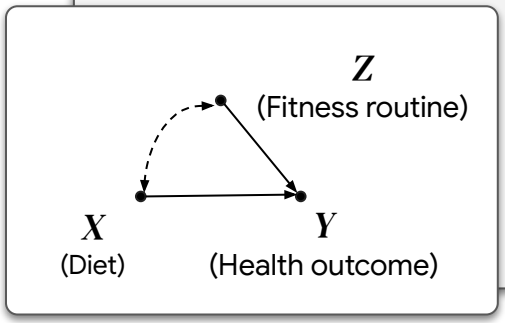
Depends on invariances across domains

Q: What is the error of a given prediction function in a ***target*** domain … given data from a ***source*** domain?

A: Sometimes can be computed.

$$\underbrace{\mathbb{E}_{P^T}}_{\text{Target distribution}}\Big[\overbrace{\mathcal{L}}^{\text{Loss / Error}}(y, \underbrace{h(x)}_{\text{Prediction function}})\Big] \quad \text{given} \quad \underbrace{P^S(x, y, \ldots)}_{\text{Source distribution}} \quad + \quad \ldots$$

Example



Latent

$$M^S = \begin{cases} y \leftarrow f_Y(x, z, u_Y) \\ x \leftarrow f_X(u_{XZ}) \\ z \leftarrow f_Z(u_{XZ}) \\ P(u_Y, u_{XZ}) \end{cases}$$

Transportability Problem

Depends on invariances across domains

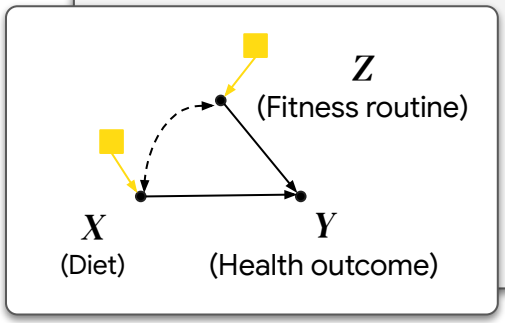
Q: What is the error of a given prediction function in a ***target*** domain … given data from a ***source*** domain?

A: Sometimes can be computed.

Loss / Error

$$\underbrace{\mathbb{E}_{P^T}}_{\text{Target distribution}}\left[\underbrace{\mathcal{L}(y, h(x))}_{\text{Prediction function}}\right] \quad \text{given} \quad \underbrace{P^S(x, y, \ldots)}_{\text{Source distribution}} \quad + \quad \ldots$$

Example



Latent

$$M^S = \begin{cases} y \leftarrow f_Y(x, z, u_Y) \\ x \leftarrow f_X(u_{XZ}) \\ z \leftarrow f_Z(u_{XZ}) \\ P(u_Y, u_{XZ}) \end{cases}$$

Differences across domains captured in Selection diagram

Transportability Problem

Depends on invariances across domains

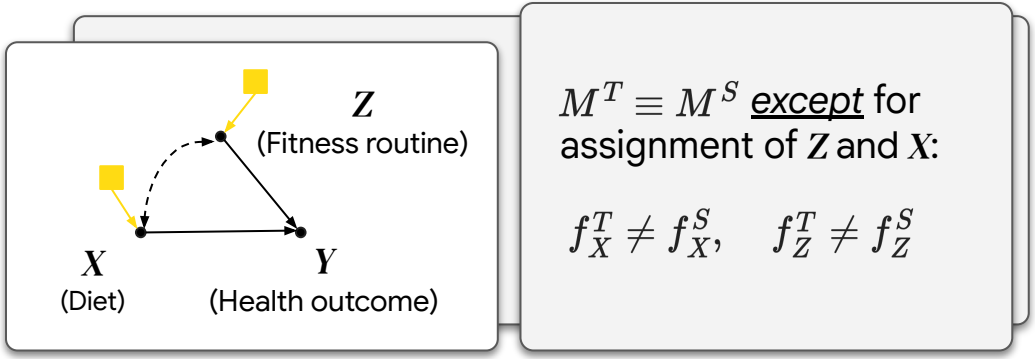
Q: What is the error of a given prediction function in a ***target*** domain … given data from a ***source*** domain?

A: Sometimes can be computed.

Loss / Error

$$\underbrace{\mathbb{E}_{P^T}}_{\text{Target distribution}}\left[\underbrace{\mathcal{L}(y, h(x))}_{\text{Prediction function}}\right] \quad \text{given} \quad \underbrace{P^S}_{\text{Source distribution}}(x, y, \ldots) \quad + \quad \ldots$$

Example



$M^T \equiv M^S$ except for assignment of Z and X:

$$f_X^T \neq f_X^S, \quad f_Z^T \neq f_Z^S$$

Differences across domains captured in Selection diagram

Transportability Problem

Depends on invariances across domains

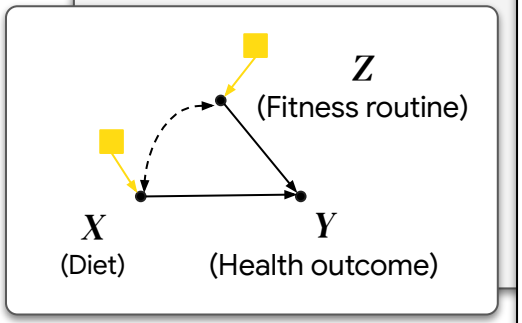
Q: What is the error of a given prediction function in a ***target*** domain … given data from a ***source*** domain?

A: Sometimes can be computed.

Loss / Error

$$\underbrace{\mathbb{E}_{P^T}}_{\text{Target distribution}}[\underbrace{\mathcal{L}(y, \underbrace{h(x)}_{\text{Prediction function}})}_{\text{Loss / Error}}] \quad \text{given} \quad \underbrace{P^S(x, y, \ldots)}_{\text{Source distribution}} \quad + \quad \ldots$$

Example



Differences across domains captured in Selection diagram

Compute error of $h(x)$ with access to $P^S(y, x, z), P^T(x, z)$

$$\mathbb{E}_{P^T}[\mathcal{L}(y, h(x))] = \sum_{x,y} \mathcal{L}(y, h(x)) P^T(x, y)$$ by definition

Transportability Problem

Depends on invariances across domains

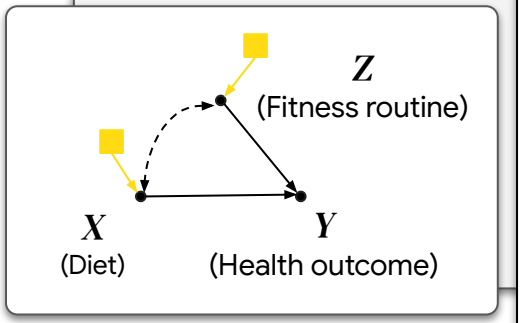
Q: What is the error of a given prediction function in a ***target*** domain … given data from a ***source*** domain?

A: Sometimes can be computed.

Loss / Error

$$\underbrace{\mathbb{E}_{P^T}}_{\text{Target distribution}}[\underbrace{\mathcal{L}(y, \underbrace{h(x)}_{\text{Prediction function}})}_{\text{Loss / Error}}] \quad \text{given} \quad \underbrace{P^S(x, y, \ldots)}_{\text{Source distribution}} \quad + \quad \ldots$$

Example



Compute error of $h(x)$ with access to $P^S(y, x, z), P^T(x, z)$

$$\mathbb{E}_{P^T}[\mathcal{L}(y, h(x))] = \sum_{x,y} \mathcal{L}(y, h(x)) P^T(x, y) \quad \text{by definition}$$

$$= \sum_{x,y,z} \mathcal{L}(y, h(x)) P^T(y \mid x, z) P^T(x, z) \quad \text{by marginal.}$$

Differences across domains captured in Selection diagram

Transportability Problem

Depends on invariances across domains

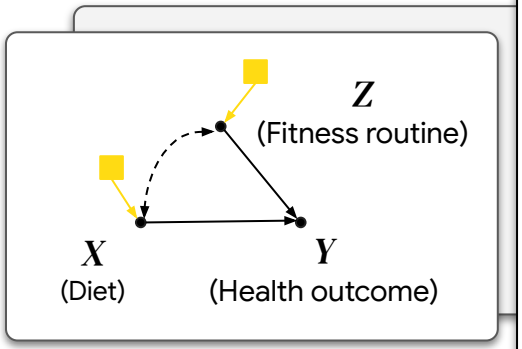
Q: What is the error of a given prediction function in a ***target*** domain … given data from a ***source*** domain?

A: Sometimes can be computed.

Loss / Error

$$\underbrace{\mathbb{E}_{P^T}}_{\text{Target distribution}}[\underbrace{\mathcal{L}(y, h(x))}_{\text{Prediction function}}] \quad \text{given} \quad \underbrace{P^S(x, y, \ldots)}_{\text{Source distribution}} \quad + \quad \ldots$$

Example

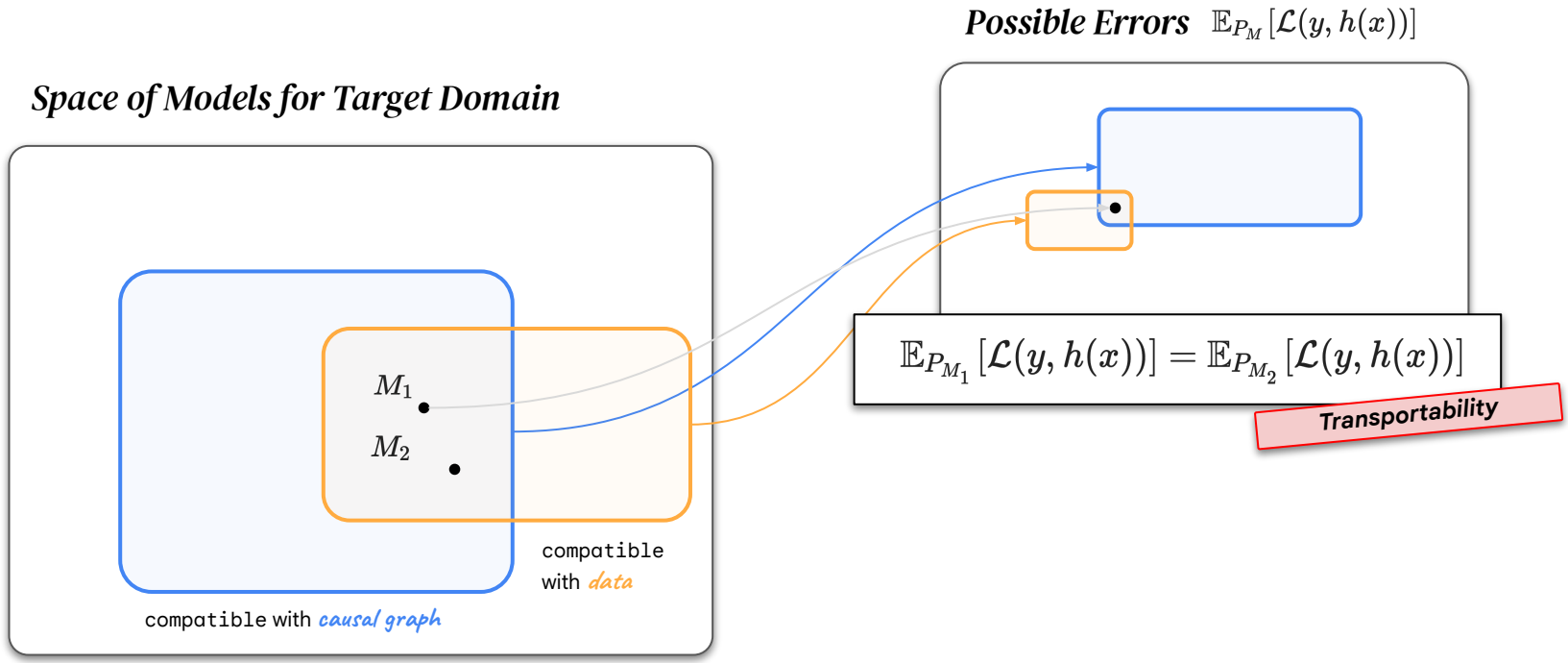


Differences across domains captured in Selection diagram

Compute error of $h(x)$ with access to $P^S(y, x, z), P^T(x, z)$

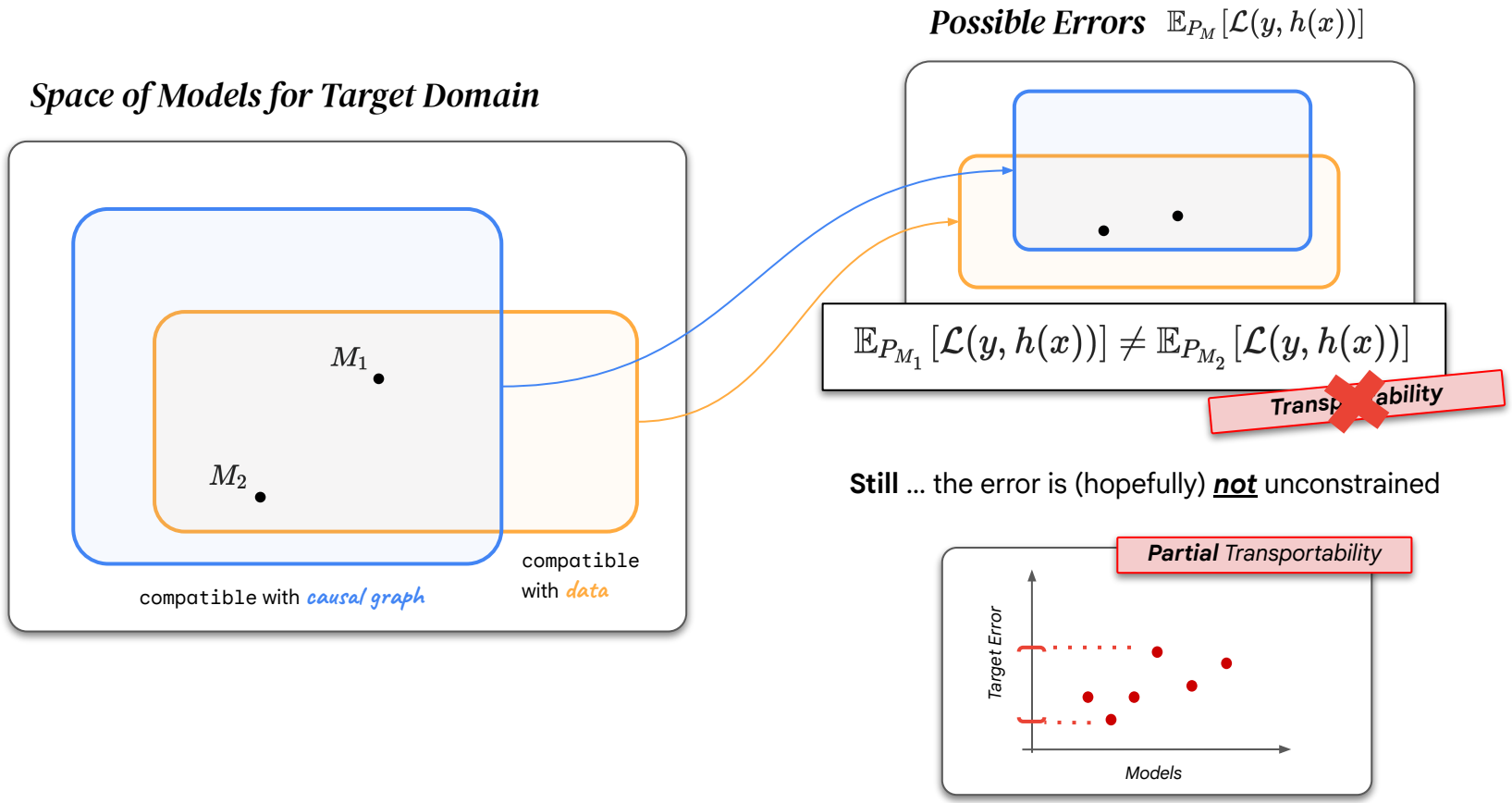
$$\begin{aligned}
\mathbb{E}_{P^T}[\mathcal{L}(y, h(x))] &= \sum_{x,y} \mathcal{L}(y, h(x)) P^T(x, y) && \text{by definition} \\
&= \sum_{x,y,z} \mathcal{L}(y, h(x)) P^T(y \mid x, z) P^T(x, z) && \text{by marginal.} \\
&= \sum_{x,y,z} \mathcal{L}(y, h(x)) P^S(y \mid x, z) P^T(x, z) && \text{by invariance}
\end{aligned}$$

Thinking in terms of Models …



Causal Inference and The Data-Fusion Problem. E. Bareinboim, J. Pearl. PNAS-16.
From Statistical Transportability to Estimating the Effect of Stochastic Interventions. J. Correa, E. Bareinboim. IJCAI-19.
General Identifiability with Arbitrary Surrogate Experiments. S. Lee, J. Correa, E. Bareinboim. UAI-19.

Thinking in terms of Models …



Thinking in terms of Models …

Possible Errors $\mathbb{E}_{P_M}[\mathcal{L}(y, f(x))]$

Space of Models for Target Domain

How to (*tightly*) **estimate** the *maximum* target error?

M_1

M_2

compatible with *causal graph*

compatible with *data*

From finite samples

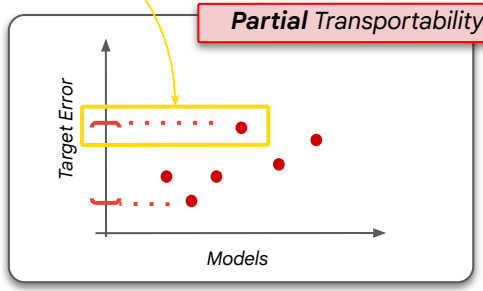
$\mathbb{E}_{P_{M_1}}[\mathcal{L}(y, f(x))] \neq \mathbb{E}_{P_{M_2}}[\mathcal{L}(y, f(x))]$

Identifiability

$$\max_{M \in \text{ All Causal Models}} \mathbb{E}_{P_M}[\mathcal{L}(y, h(x))]$$

not unconstrained

***Partial** Transportability*



02

Evaluating (Maximum) Target Error

*Possible **if** observed variables are discrete :* $P^T(Y = y, \boldsymbol{X} = \boldsymbol{x})$

Non-parametric regime, only assume data and graph!

All discrete probabilities and constraints (selection diagrams) induced by SCMs

... may be **equivalently** generated by a special subset of SCMs.

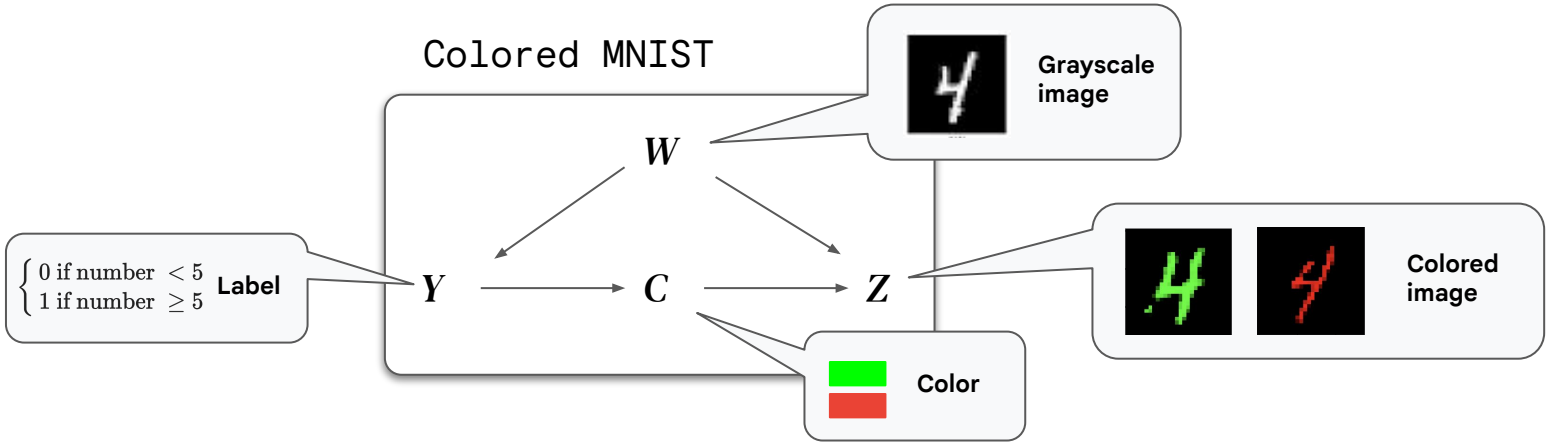
*Powerful because we can search over NCMs with **gradient**-descent*

$\{f_V : V \in \boldsymbol{V}\}$ **NN** $\{P(U_V) : U_V \in \boldsymbol{U}\}$ **Uniform**

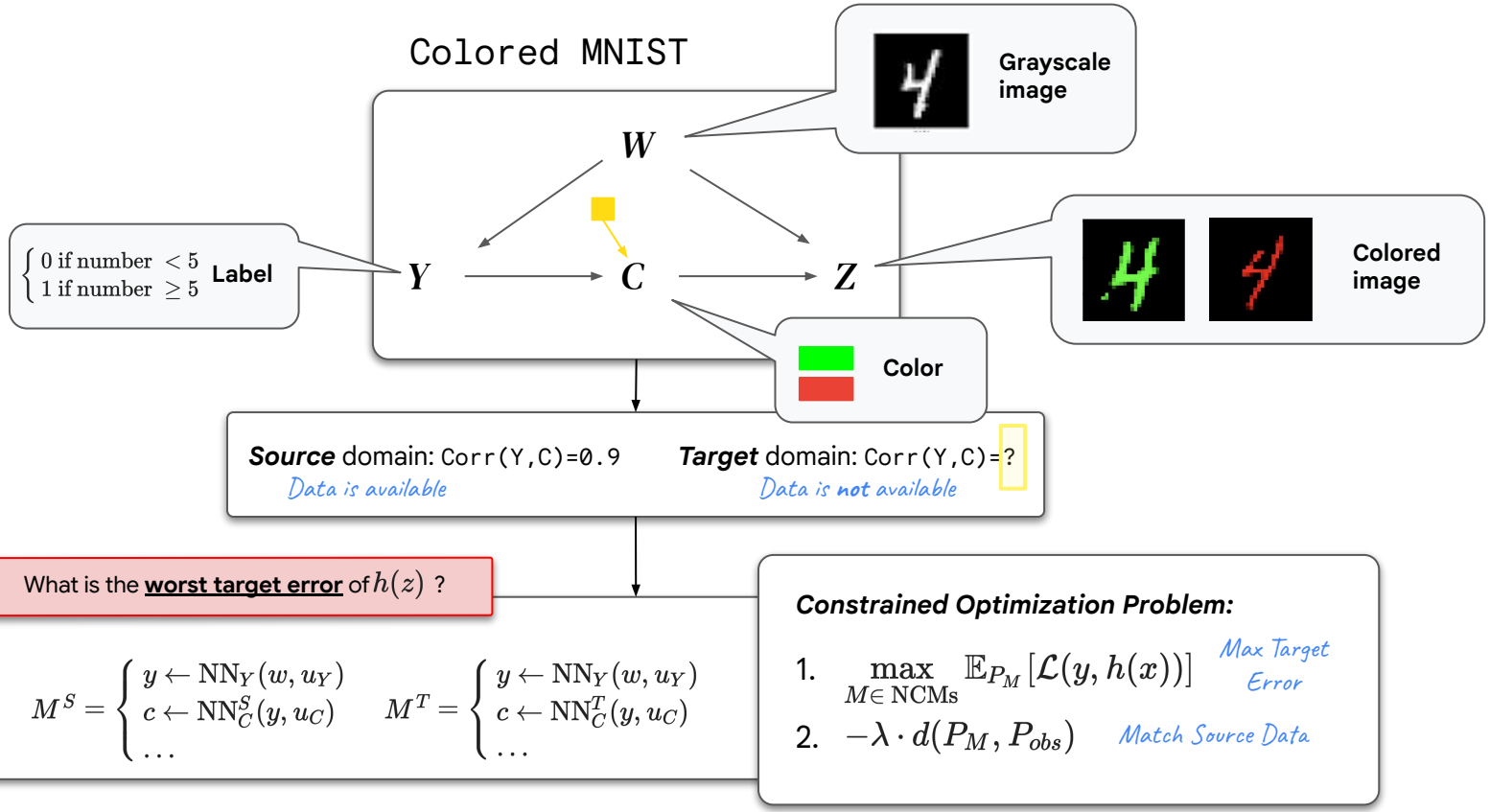
$$\max_{M \in \text{ Neural Causal Models}} \mathbb{E}_{P_M}[\mathcal{L}(y, h(x))] = \max_{M \in \text{ All Causal Models}} \mathbb{E}_{P_M}[\mathcal{L}(y, h(x))]$$

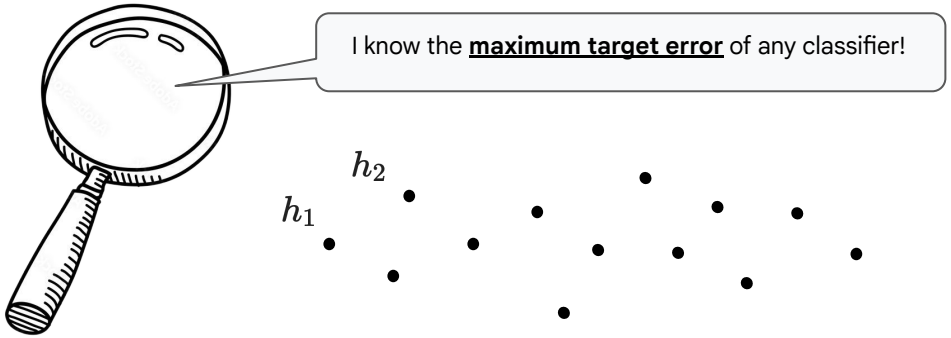
The causal-neural connection: Expressiveness, learnability, and inference. K. Xia, K. Lee, Y. Bengio, and E. Bareinboim. NeurIPS-21
Partial counterfactual identification from observational and experimental data. J. Zhang, J. Tian, and E. Bareinboim. ICML-22

Example of how this works …

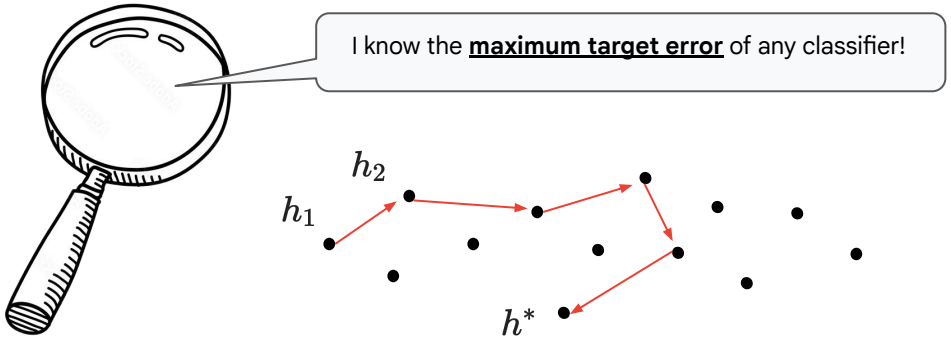


Example of how this works …





But … which classifier has the ***lowest*** maximum target error ?



But ... which classifier has the ***lowest*** maximum target error ?

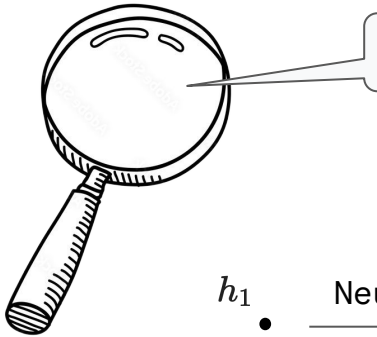
Optimization / Search

$$\arg\min_{h} \max_{M \in \text{All Causal Models}} \mathbb{E}_{P_M}\left[\mathcal{L}(y, h(x))\right]$$

Evaluation Max Target Error (Neural-TR)

03

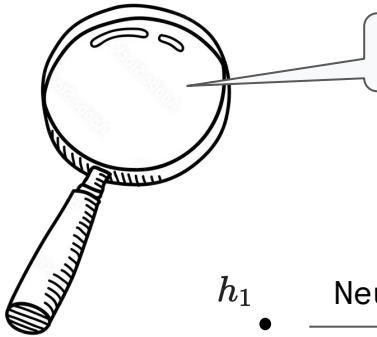
Optimizing for best (Maximum) Target Error



I can **generate adversarial data** for any prediction function !

h_1 • Neural-TR → SCM that entails *worst-case performance* of h_1 → *Sample data* → $D^1 \sim P_{M^1}$

$h_2 \leftarrow \arg\min_h \mathbb{E}_{D_1}[\mathcal{L}(y, h(x))]$ •



I can **generate adversarial data** for any prediction function !

h_1 • Neural-TR → SCM that entails *worst-case performance* of h_1 → *Sample data* → $D^1 \sim P_{M^1}$

h_2 • Neural-TR → SCM that entails *worst-case performance* of h_2 → *Sample data* → $D^2 \sim P_{M^2}$

$h_3 \leftarrow \arg\min_h \max_{D \in \mathbb{D}^*} \mathbb{E}_D[\mathcal{L}(y, h(x))] \qquad \mathbb{D}^* = D_1 \cup D_2$ •

***Best** worst-case predictor across collection of datasets*

I can **generate adversarial data** for any prediction function !

Iterate this process … converge to …

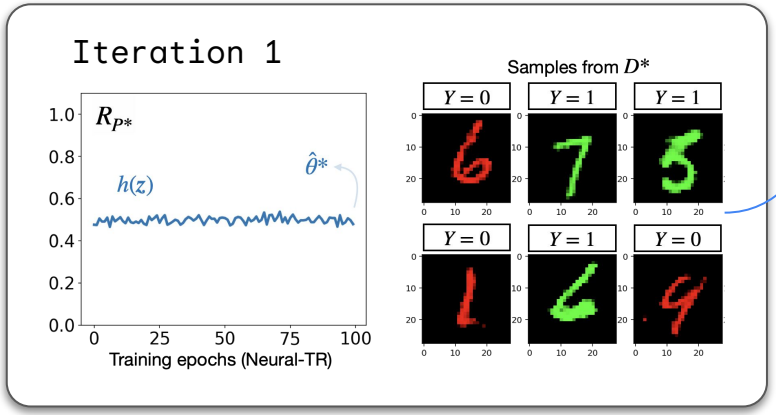
$$\arg\min_{h} \max_{M \in \text{ All Causal Models}} \mathbb{E}_{P_M}\left[\mathcal{L}(y, h(x))\right]$$

worst-case performance of h_2

$$h_3 \leftarrow \arg\min_{h} \max_{D \in \mathbb{D}^*} \mathbb{E}_D\left[\mathcal{L}(y, h(x))\right] \qquad \mathbb{D}^* = D_1 \cup D_2$$

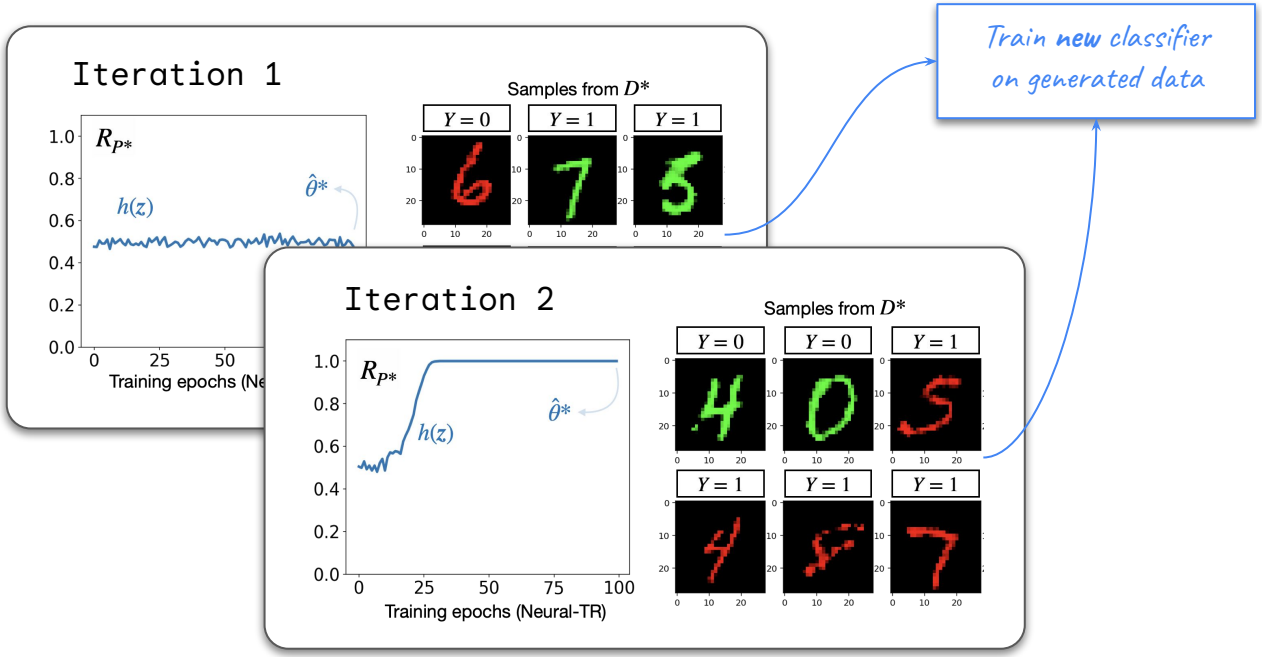
Best *worst-case predictor across collection of datasets*

What is the best (worst-case) classifier for CMNIST?

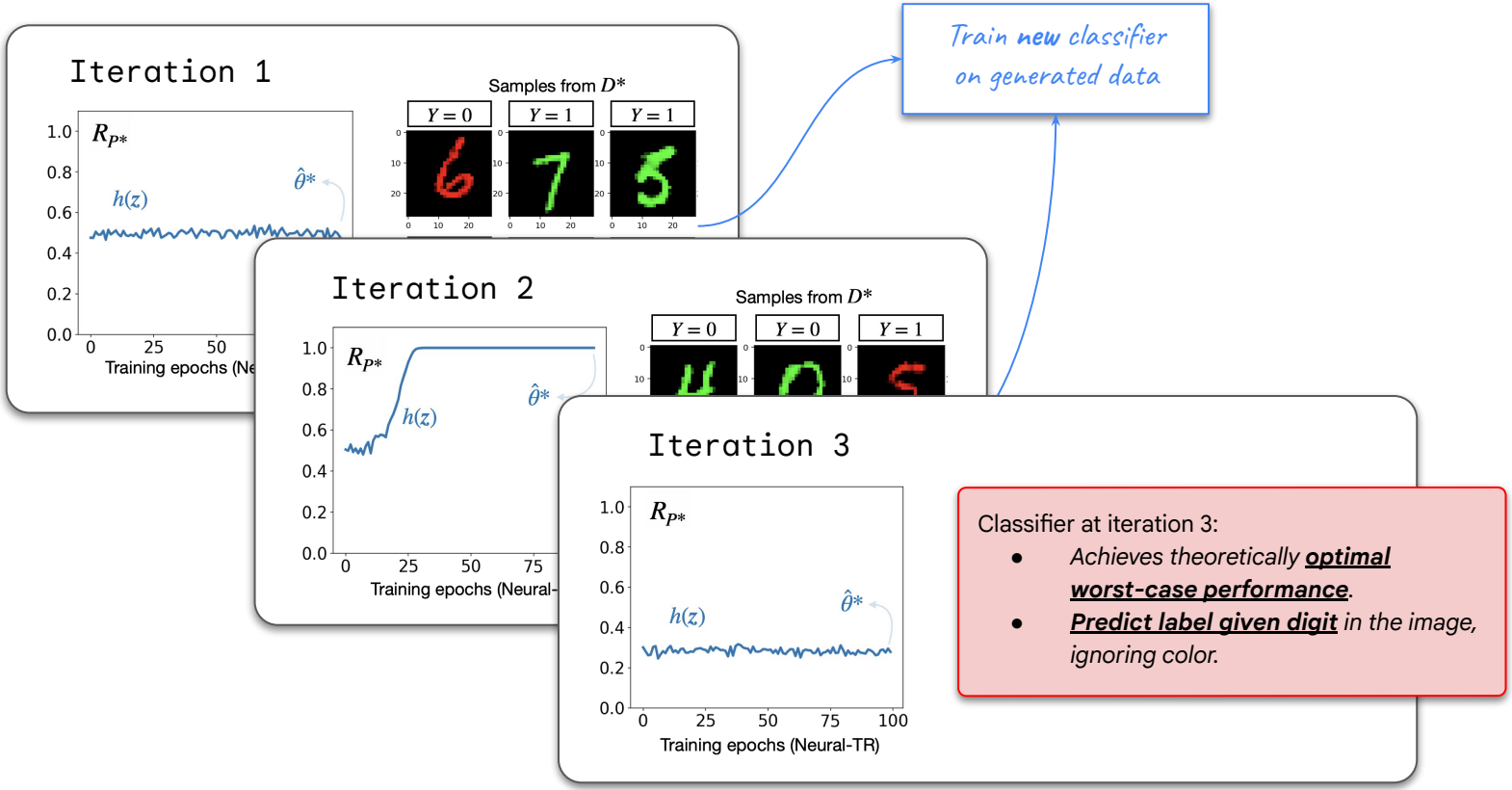


*Train **new** classifier on generated data*

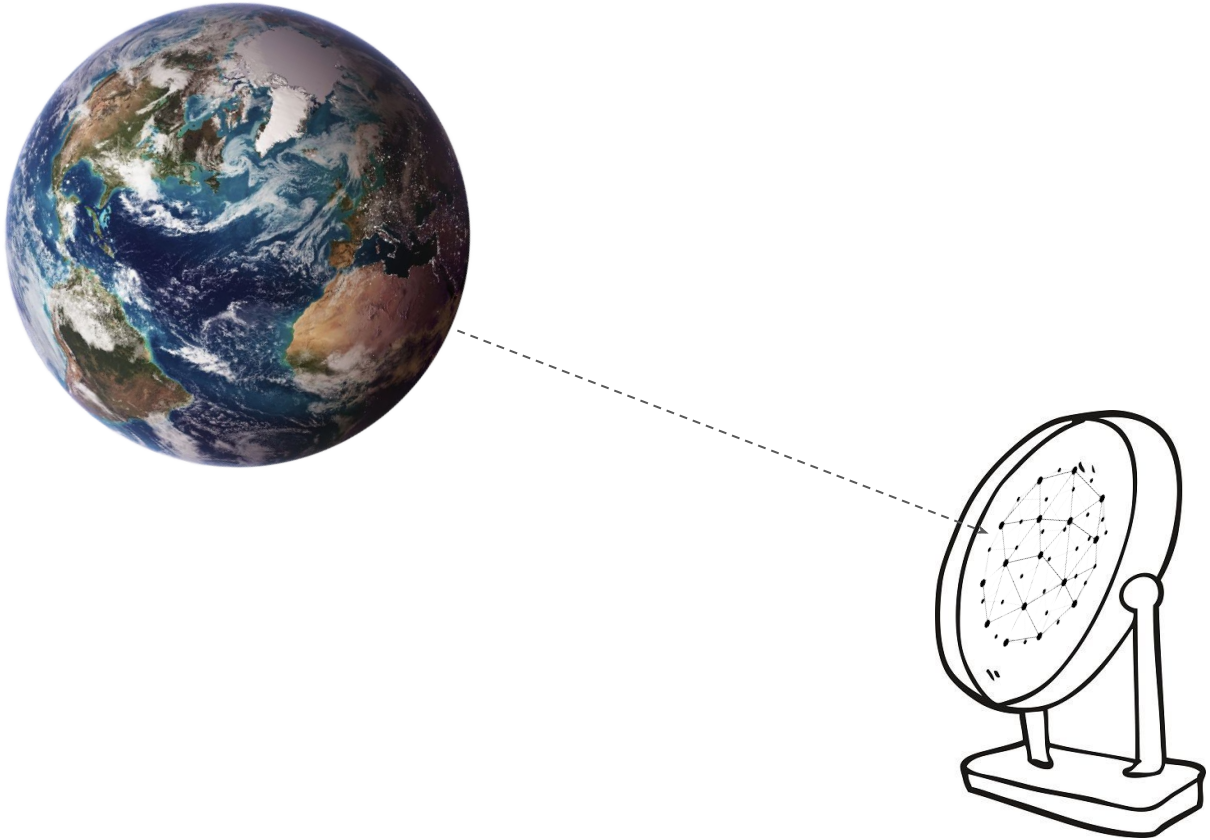
What is the best (worst-case) classifier for CMNIST?

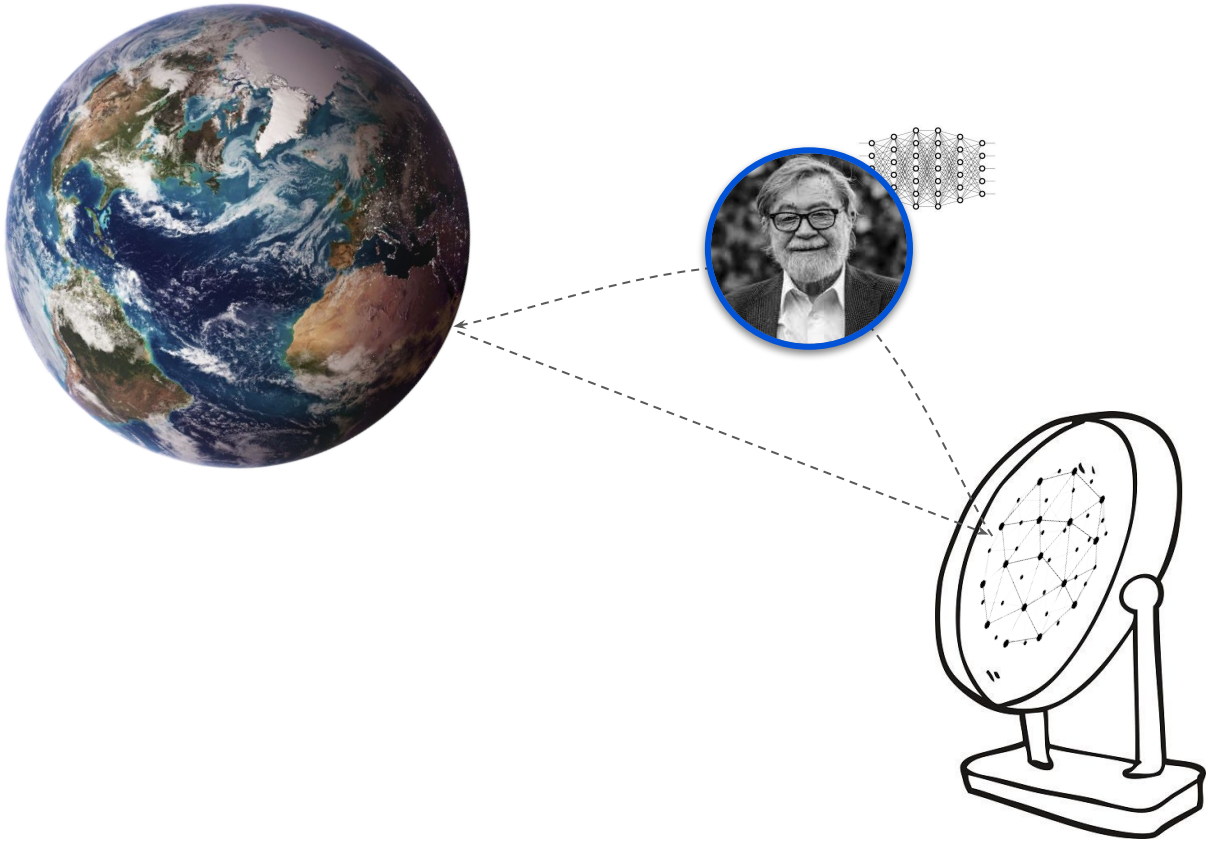


What is the best (worst-case) classifier for CMNIST?



04 Conclusion





Questions?