Thoughts on Building Machine Learning Models with Scientific Data

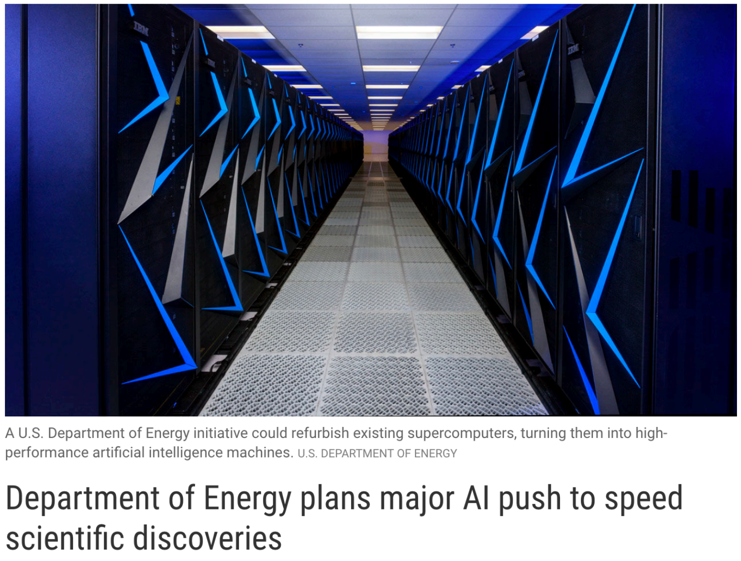
Jay Thiagarajan

CASC
Center for Applied Scientific Computing

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344. Lawrence Livermore National Security, LLC

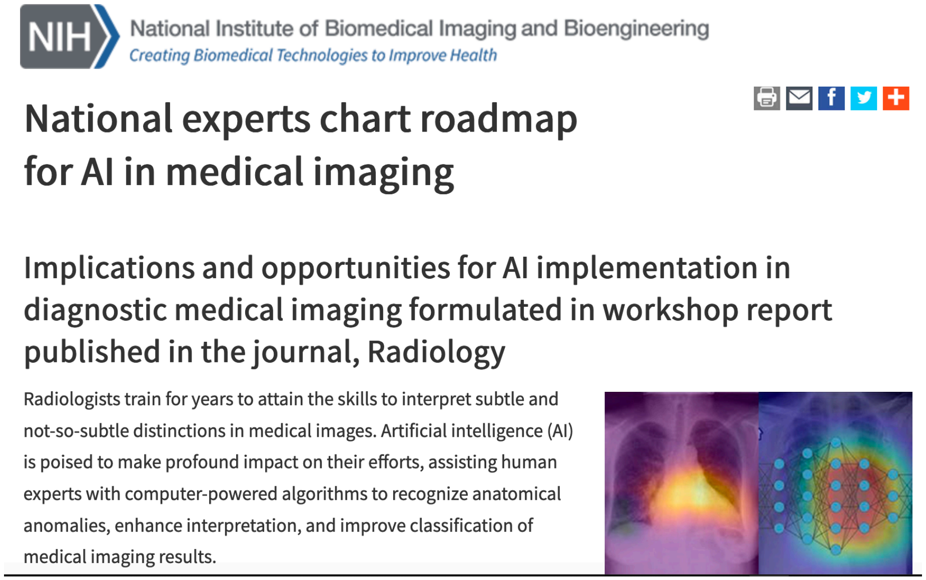
Lawrence Livermore National Laboratory

Data-Driven Methods are Rapidly Transforming Workflows in Science and Engineering



A U.S. Department of Energy initiative could refurbish existing supercomputers, turning them into high-performance artificial intelligence machines. U.S. DEPARTMENT OF ENERGY

Department of Energy plans major AI push to speed scientific discoveries



NIH National Institute of Biomedical Imaging and Bioengineering

Creating Biomedical Technologies to Improve Health

National experts chart roadmap for AI in medical imaging

Implications and opportunities for AI implementation in diagnostic medical imaging formulated in workshop report published in the journal, Radiology

Radiologists train for years to attain the skills to interpret subtle and not-so-subtle distinctions in medical images. Artificial intelligence (AI) is poised to make profound impact on their efforts, assisting human experts with computer-powered algorithms to recognize anatomical anomalies, enhance interpretation, and improve classification of medical imaging results.

Artificial Intelligence and Technology Office

Home » Science & Innovation » Artificial Intelligence and Technology Office

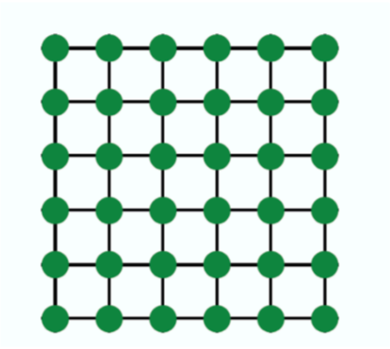
AI at ENERGY.GOV

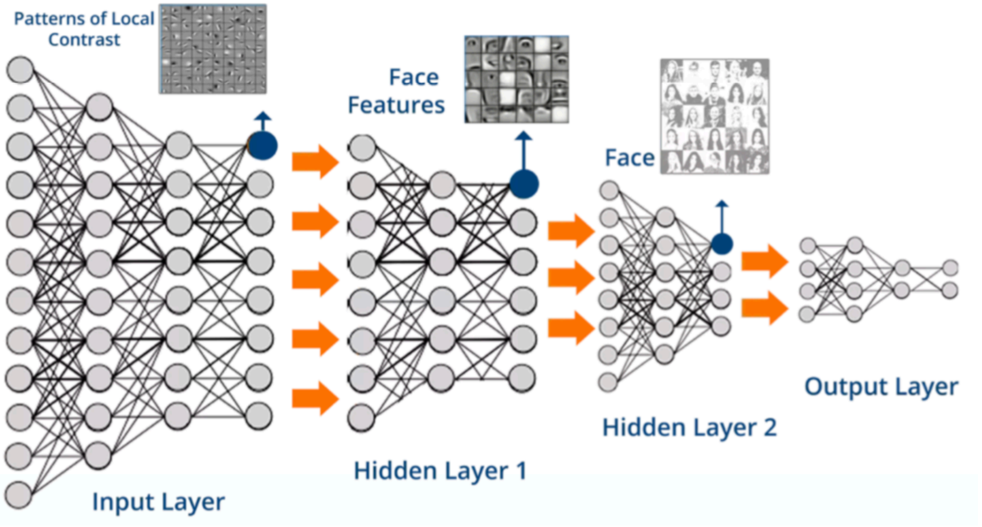
Vision:

Transform DOE into a world-leading AI enterprise by accelerating the research, development, delivery, and adoption of AI.

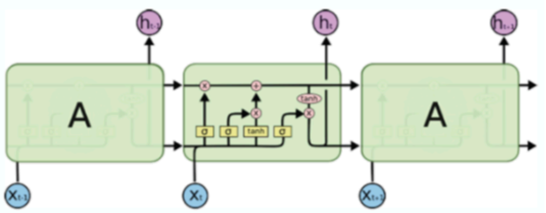
Modern Machine Learning Techniques are Highly Effective in Modeling Complex Data

Data defined on spatial grids

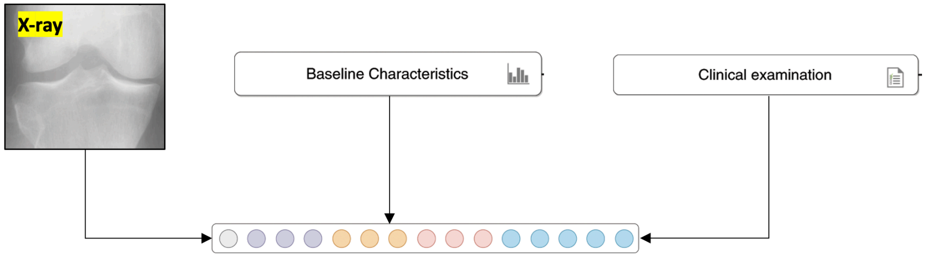




Time-varying data



Multi-modal data



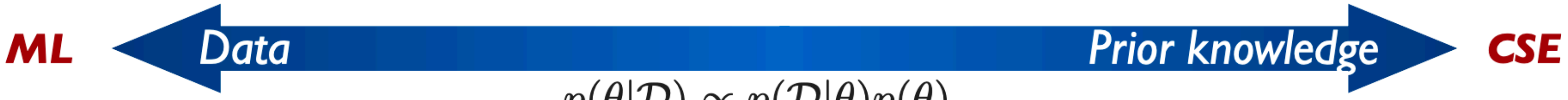
Baseline Characteristics

Clinical examination

What is Different About Applying Machine Learning Tools to Scientific Data?

Scientific problems present unique challenges, often requiring custom solutions:

- High cost of data acquisition
- High-dimensional observations and large parameter spaces
- Complex noise processes
- Uncertainty quantification
- Robust design/control



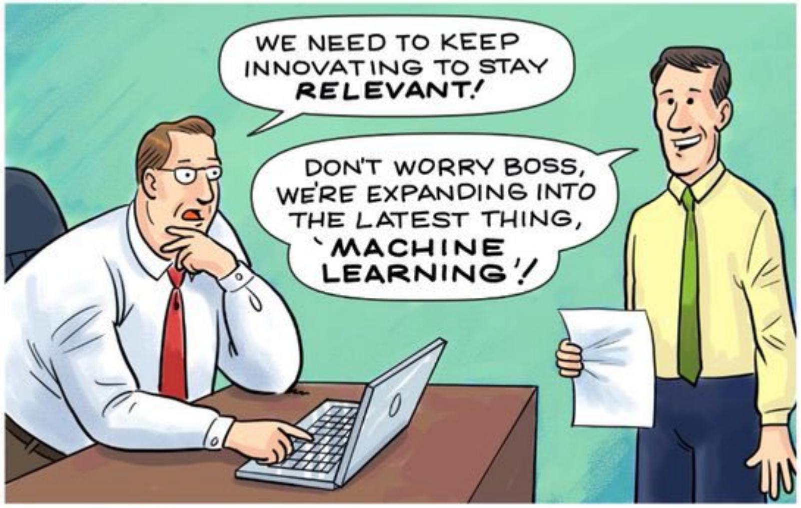
ML Data Prior knowledge CSE

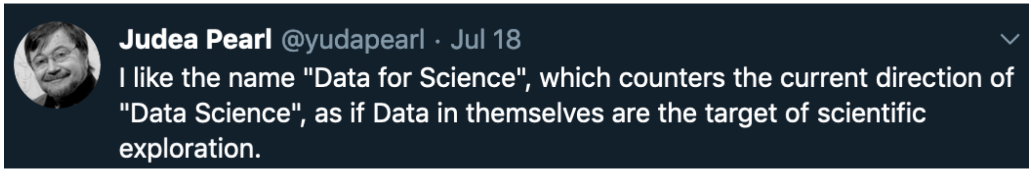
$$p(\theta|\mathcal{D}) \propto p(\mathcal{D}|\theta)p(\theta)$$

A key step forward is to bridge conventional domain-informed modeling and machine learning to tackle these challenges

This Talk…

Insights from our experience in building predictive models for scientific data.


WE NEED TO KEEP INNOVATING TO STAY RELEVANT!
DON'T WORRY BOSS, WE'RE EXPANDING INTO THE LATEST THING, 'MACHINE LEARNING'!



Judea Pearl @yudapearl · Jul 18

I like the name "Data for Science", which counters the current direction of "Data Science", as if Data in themselves are the target of scientific exploration.

"Though model choices, learning methods and constraints are highly specific to the application and data characteristics, there are useful ideas to share!

Use representation learning to explore and study complex relationships in scientific data!

Identifying the True Generative Factors of Data Can Help Build Reliable Predictive Models

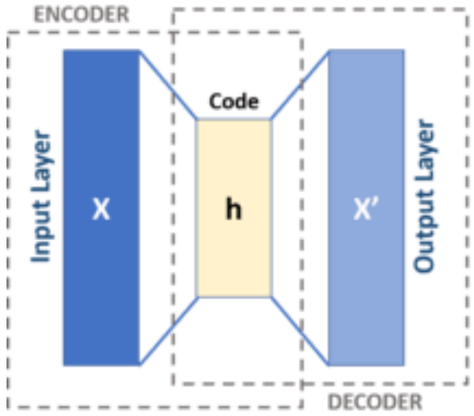
Representation learning allows us to infer latent features that succinctly describe the governing physical process.

Desired properties: (i) task-agnostic; (ii) low-dimensional; (iii) robust to data noise; (iv) preserves key relationships; (v) disentangled factors; (vi) known generative process.

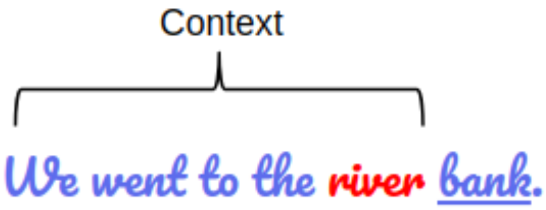
A long-standing problem in machine learning – since the advent of principal component analysis.

Representation Learning Methods can be Grouped into Three Categories

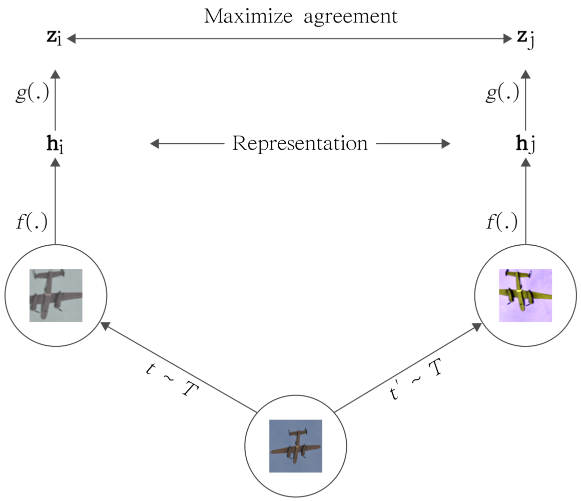
(i) Generative – Ability to sample data using a parametric/non-parametric generative model with low number of latent factors.



(ii) Context prediction – Predict missing information using its context.

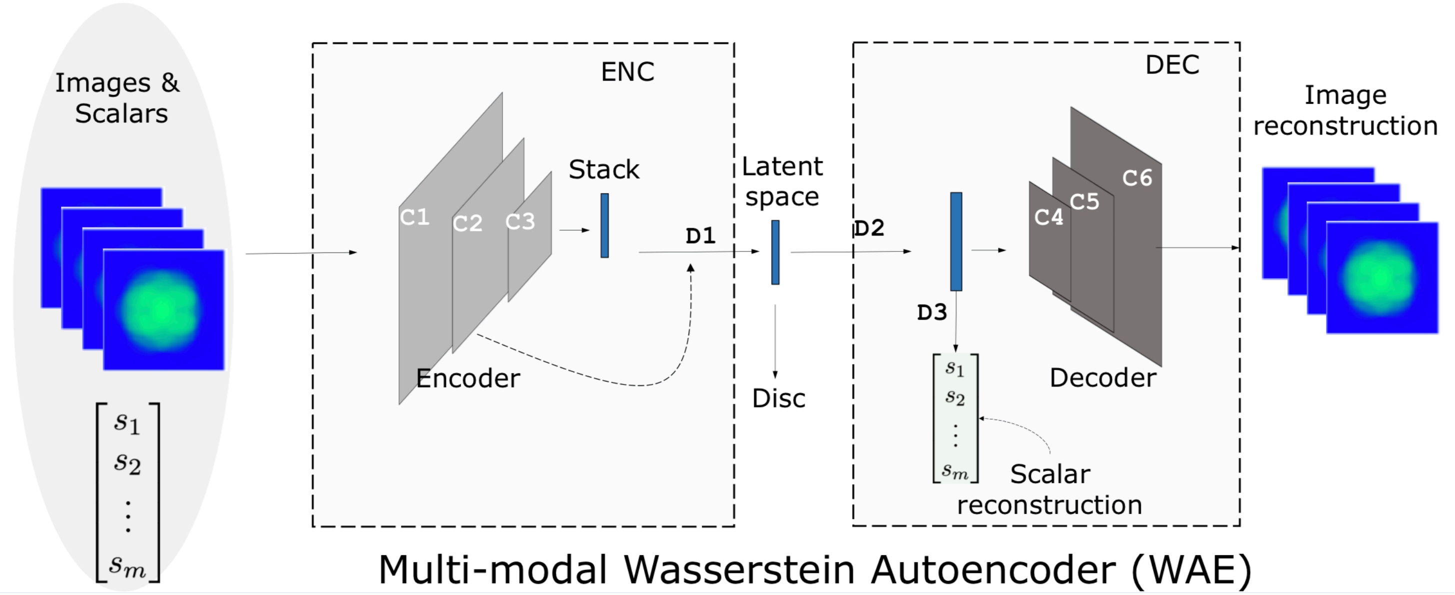


(iii) Contrastive – Build representations by learning to contrast.



Explore Learned Representations using Scientific Priors

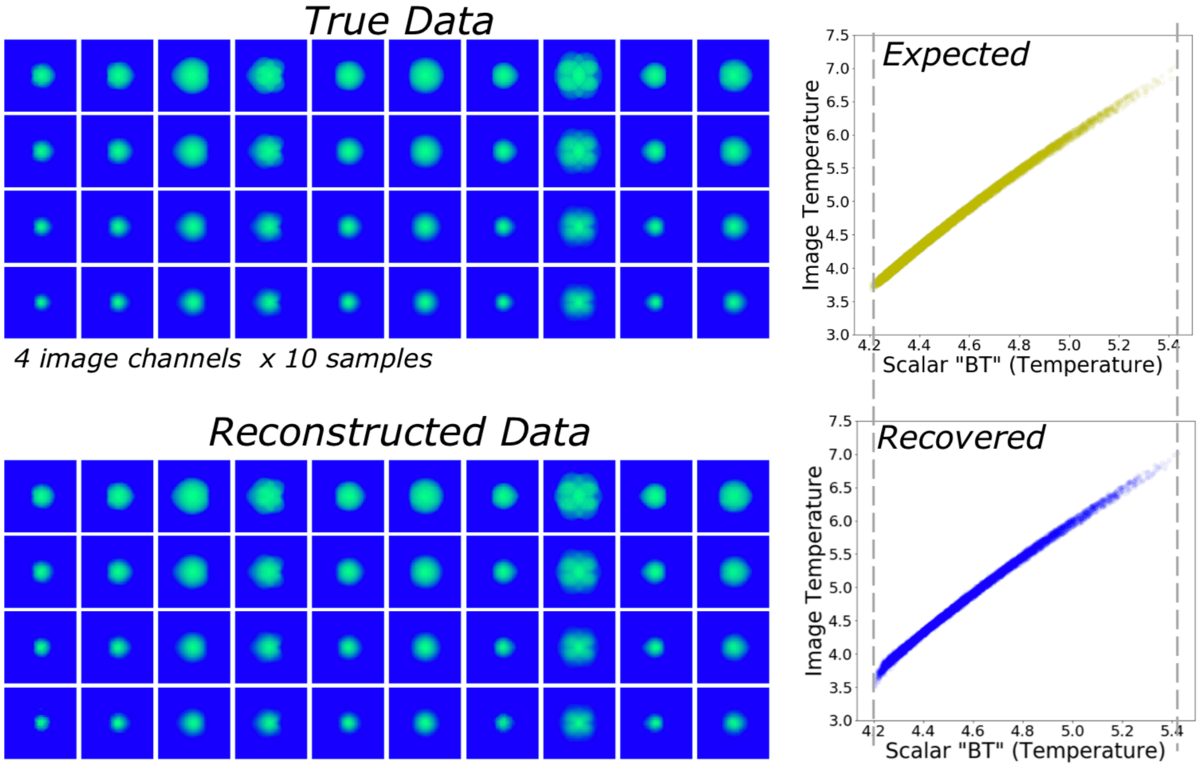
Example 1: Multi-modal measurements from an inertial confinement fusion simulator



Autoencoding with a reconstruction loss

Explore Learned Representations using Scientific Priors

Example 1: Multi-modal measurements from an inertial confinement fusion simulator

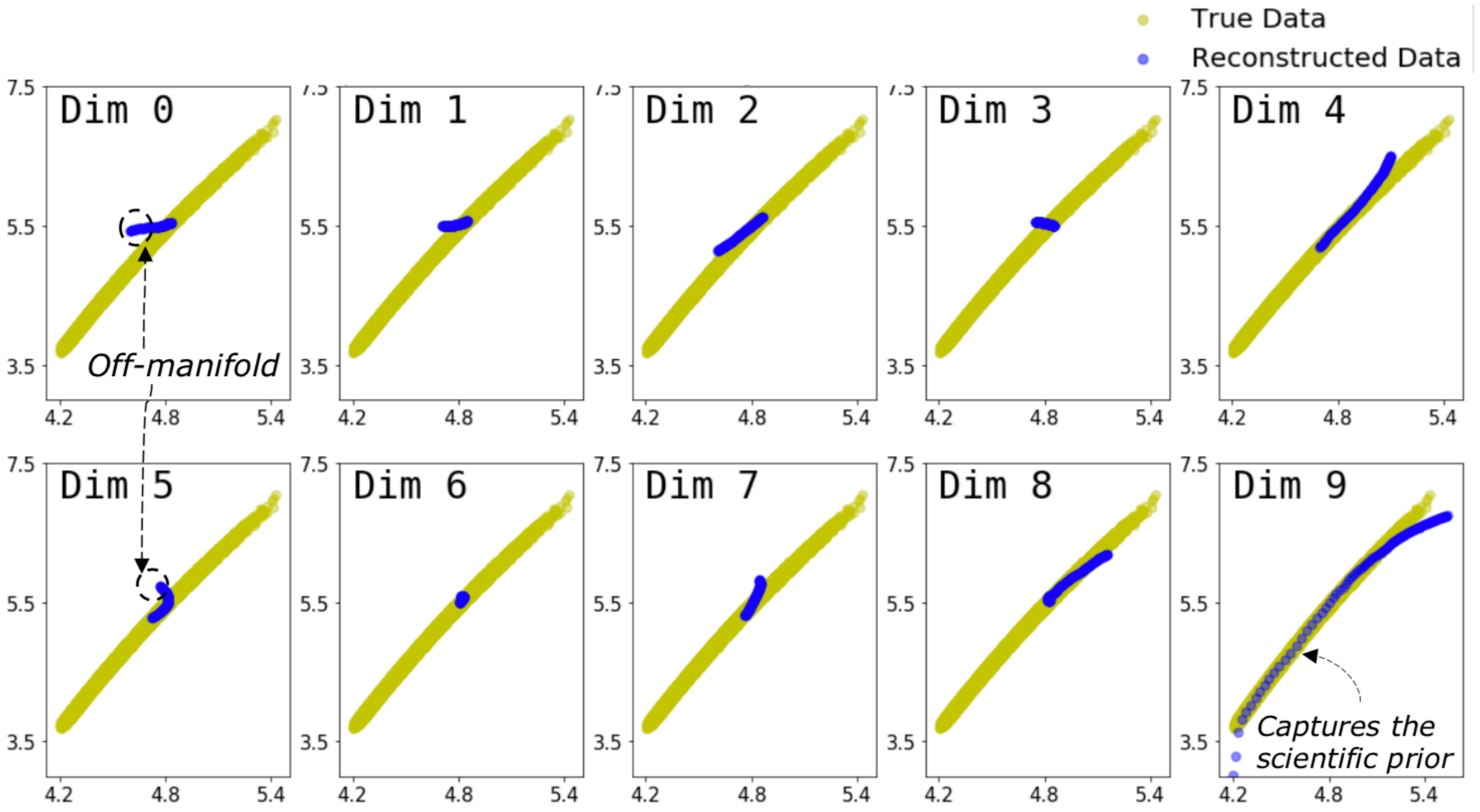


Evaluating the scientific prior on reconstructed and true samples

We use notions of thermal equilibrium to relate predicted ion temperatures to estimates of electron temperature formed by ratios of x-ray image brightness.

Explore Learned Representations using Scientific Priors

Example 1: Multi-modal measurements from an inertial confinement fusion simulator

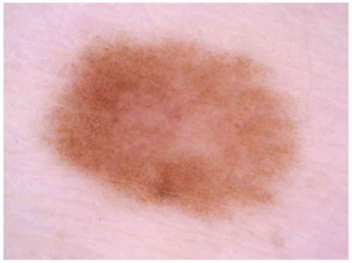


Walking along different directions on the physics manifold → evaluating the scientific prior

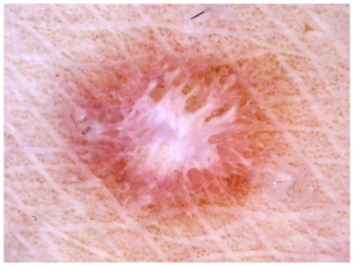
Explore Learned Representations using Scientific Priors

Example 2: Dermoscopy images from subjects diagnosed with different types of skin lesions

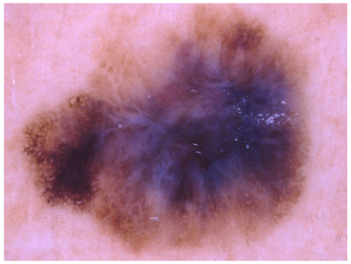
Nevus



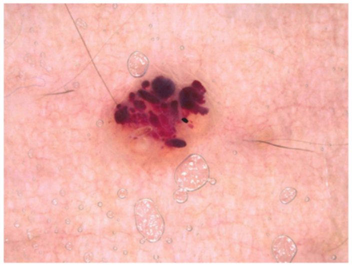
Dermatofibroma



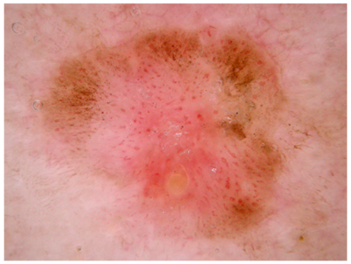
Melanoma



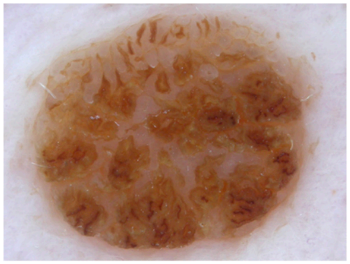
Vascular



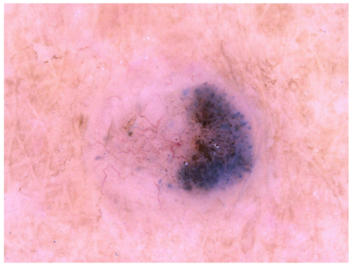
Pigmented Bowen's



Pigmented Benign Keratoses



Basal Cell Carcinoma

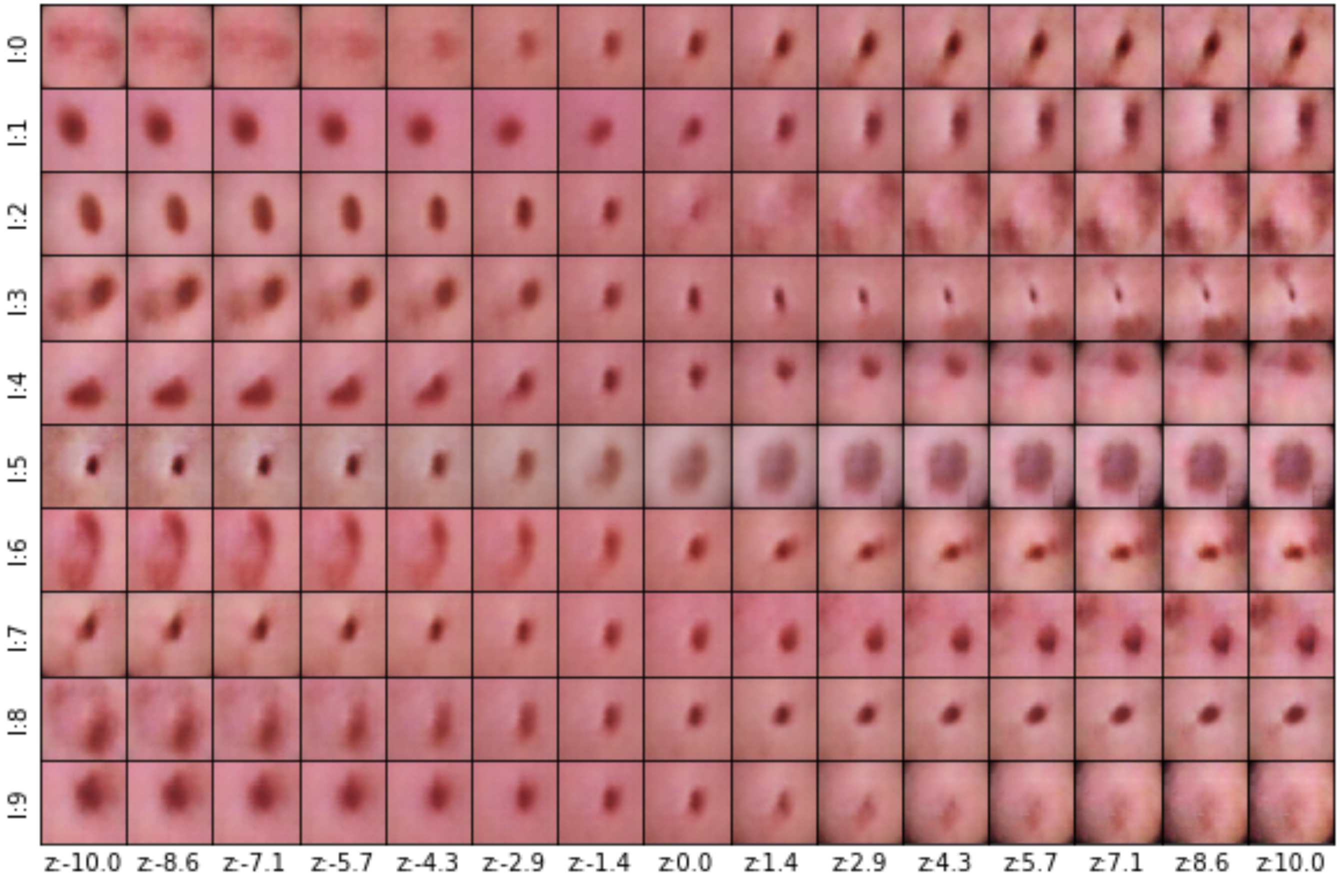


Explore Learned Representations using Scientific Priors

Example 2: Dermoscopy images from subjects diagnosed with different types of skin lesions

Latent space traversal from a disentangled VAE

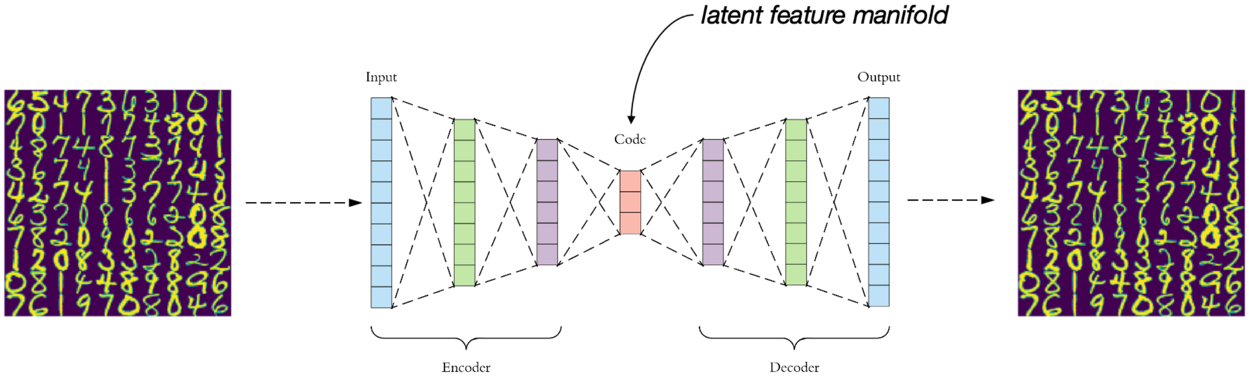
A: Asymmetry
B: Border
C: Color
D: Diameter



Distribution shifts are features, not bugs!

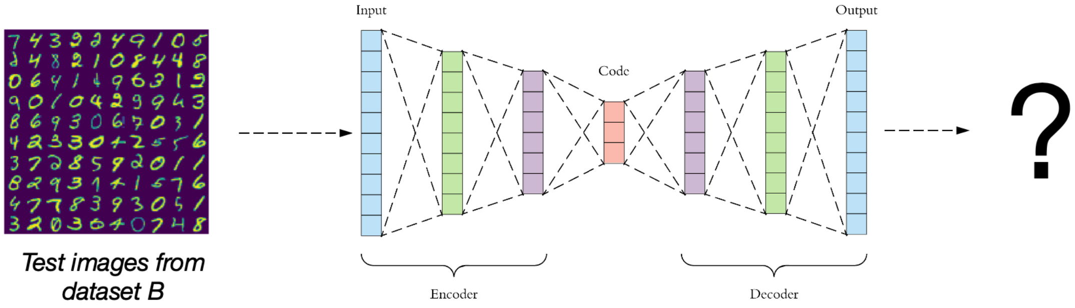
Understanding Inductive Biases of ML Models is Critical to Characterizing their Behavior

Data you expect vs. Data you get



Digits Autoencoder

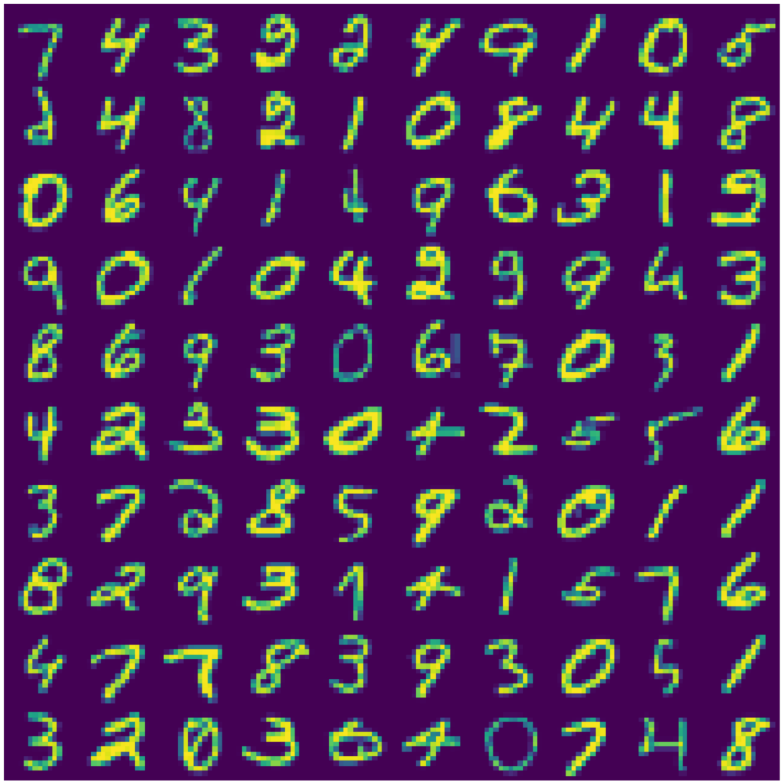
(trained on dataset A)



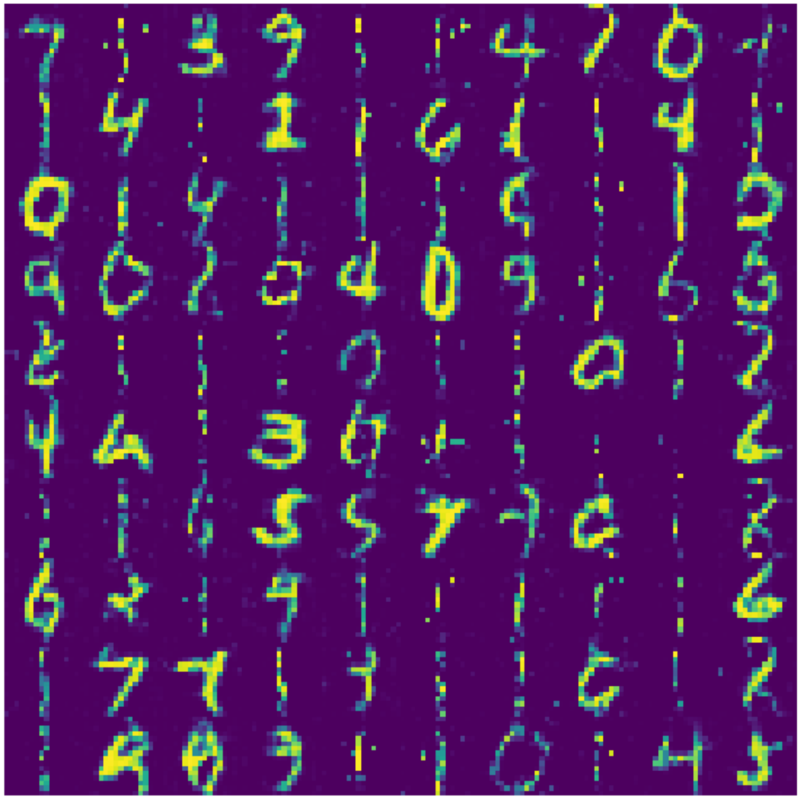
Digits Autoencoder

Understanding Inductive Biases of ML Models is Critical to Characterizing their Behavior

What we put in

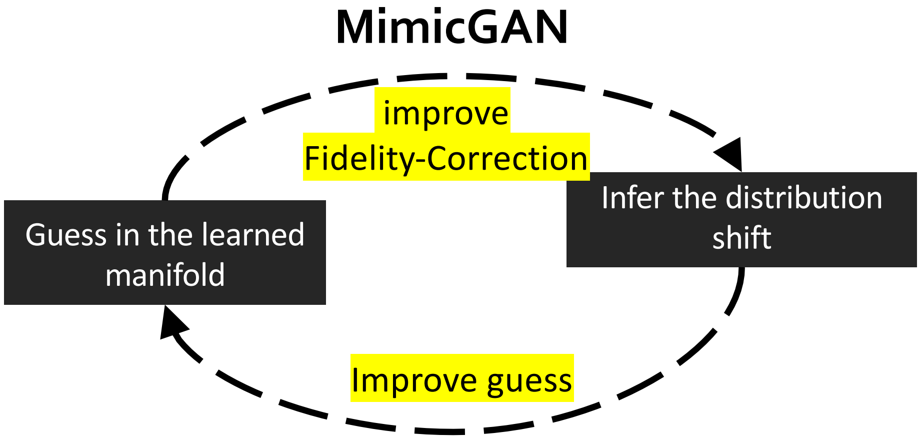


What we got out



Inferring the Unknown” Distribution Shifts is Essential to Effectively Utilize the Representations in Practice

How should we transform “unseen” observations to look like they were produced by the “known” data generation process?



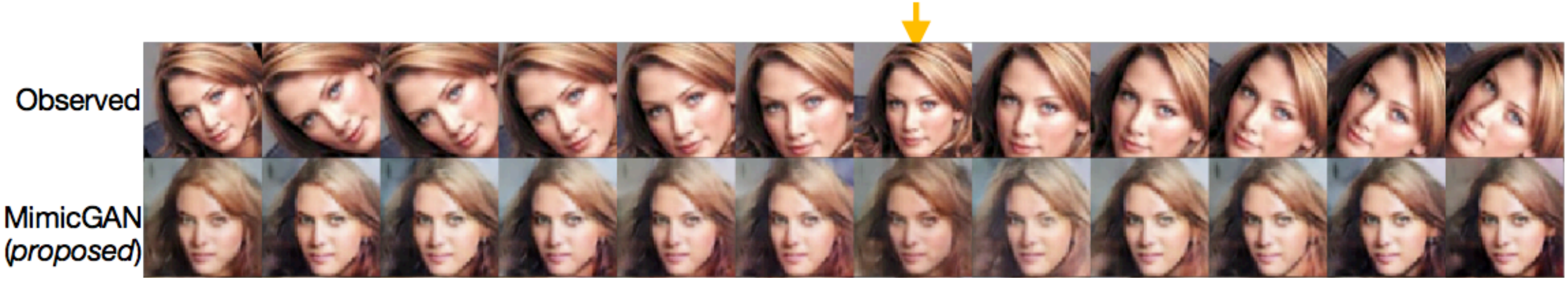
MimicGAN

Guess in the learned manifold

improve Fidelity-Correction

Infer the distribution shift

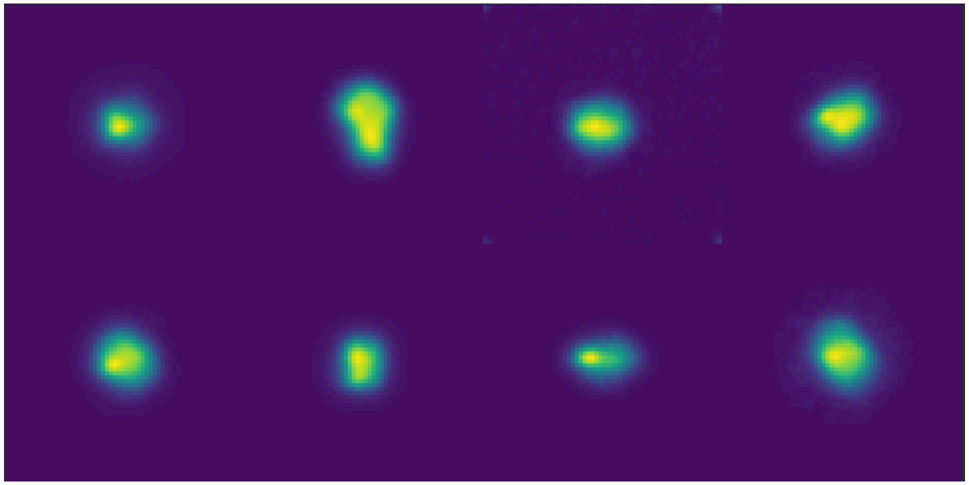
Improve guess



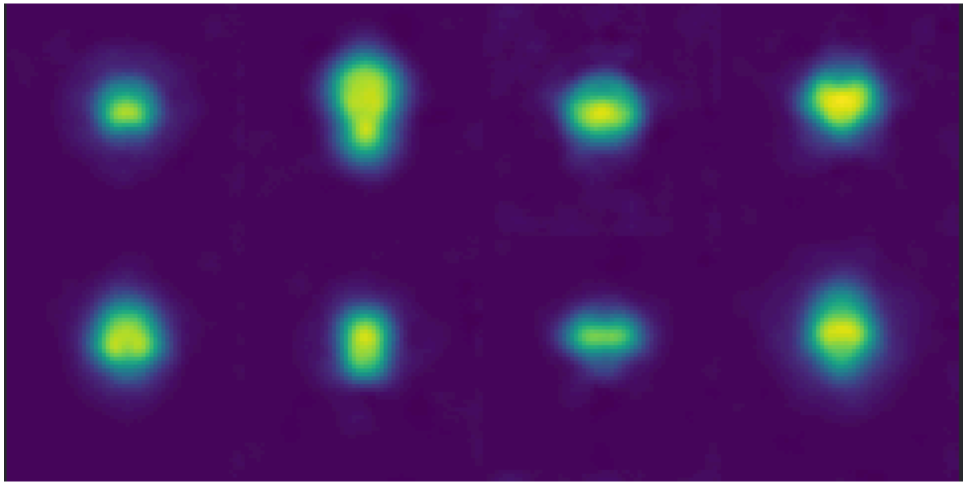
Understanding Inductive Biases of ML Models is Critical to Characterizing their Behavior

Example: Matching real-experiments (X-ray images) to simulations from an inertial confinement fusion simulator (hydra).

Experiments – Asymmetric



Equivalent Simulations – Symmetric



Your model is as good as your loss function!

Machine Learning Eventually Boils Down to Optimizing Parameters Based on a Fidelity Metric

Does patient A have cancer? → Does the patient’s scan look different enough from the database of healthy subjects? → Minimize the cross-entropy loss

In many predictive modeling problems in sciences, we deal with continuous-valued targets and we use mean-squared loss as the objective.

Choice of loss function places a prior on the distribution of residuals

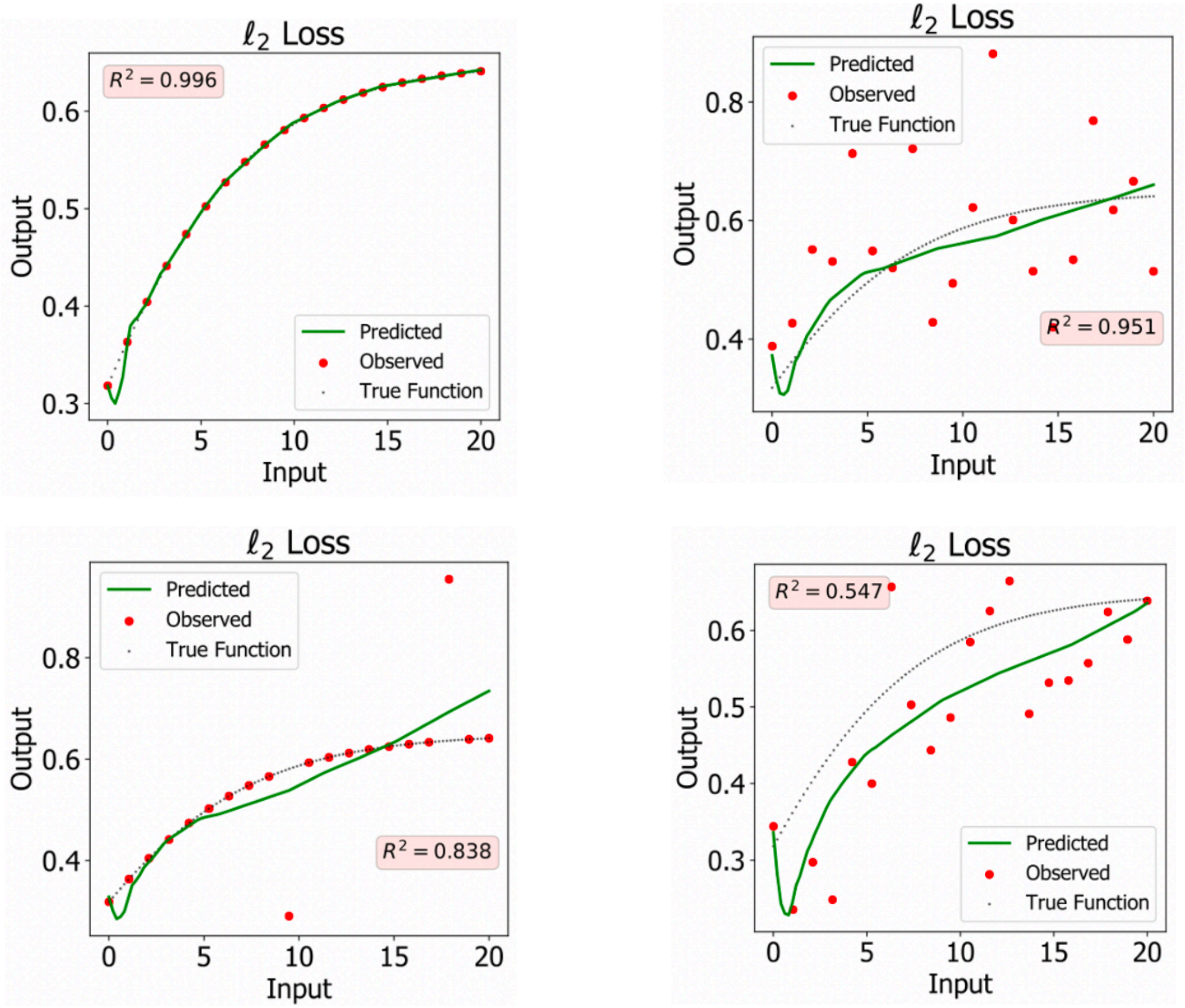
$$\mathbf{y} - f(\mathbf{x})$$

L2 error optimal when the distribution is symmetric

Not robust when there are outliers

Machine Learning Eventually Boils Down to Optimizing Parameters Based on a Fidelity Metric

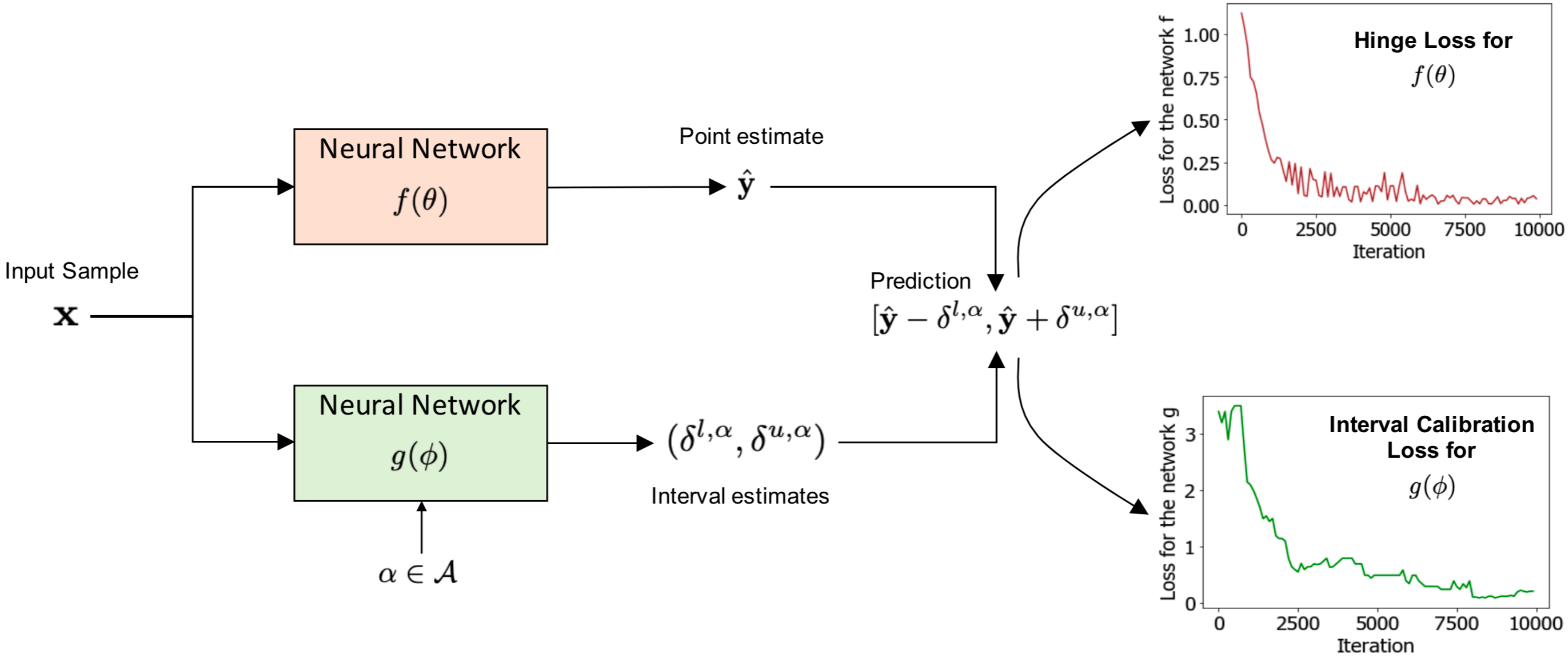
Example: Approximating a synthetic function using a 1-layer MLP



Learn-by-Calibrating: Prior-free Loss Function Design Based on Interval Calibration

If a model produces a prediction interval $[\hat{\mathbf{y}} - \delta^l, \hat{\mathbf{y}} + \delta^u]$

$$\boldsymbol{p}\big((\hat{\mathbf{y}} - \boldsymbol{\delta}^l) \leq \mathbf{y} \leq (\hat{\mathbf{y}} + \boldsymbol{\delta}^u)\big) = \boldsymbol{\alpha} \longrightarrow \text{Confidence level}$$


Input Sample

$\mathbf{x}$

Neural Network $f(\theta)$

Neural Network $g(\phi)$

$\alpha \in \mathcal{A}$

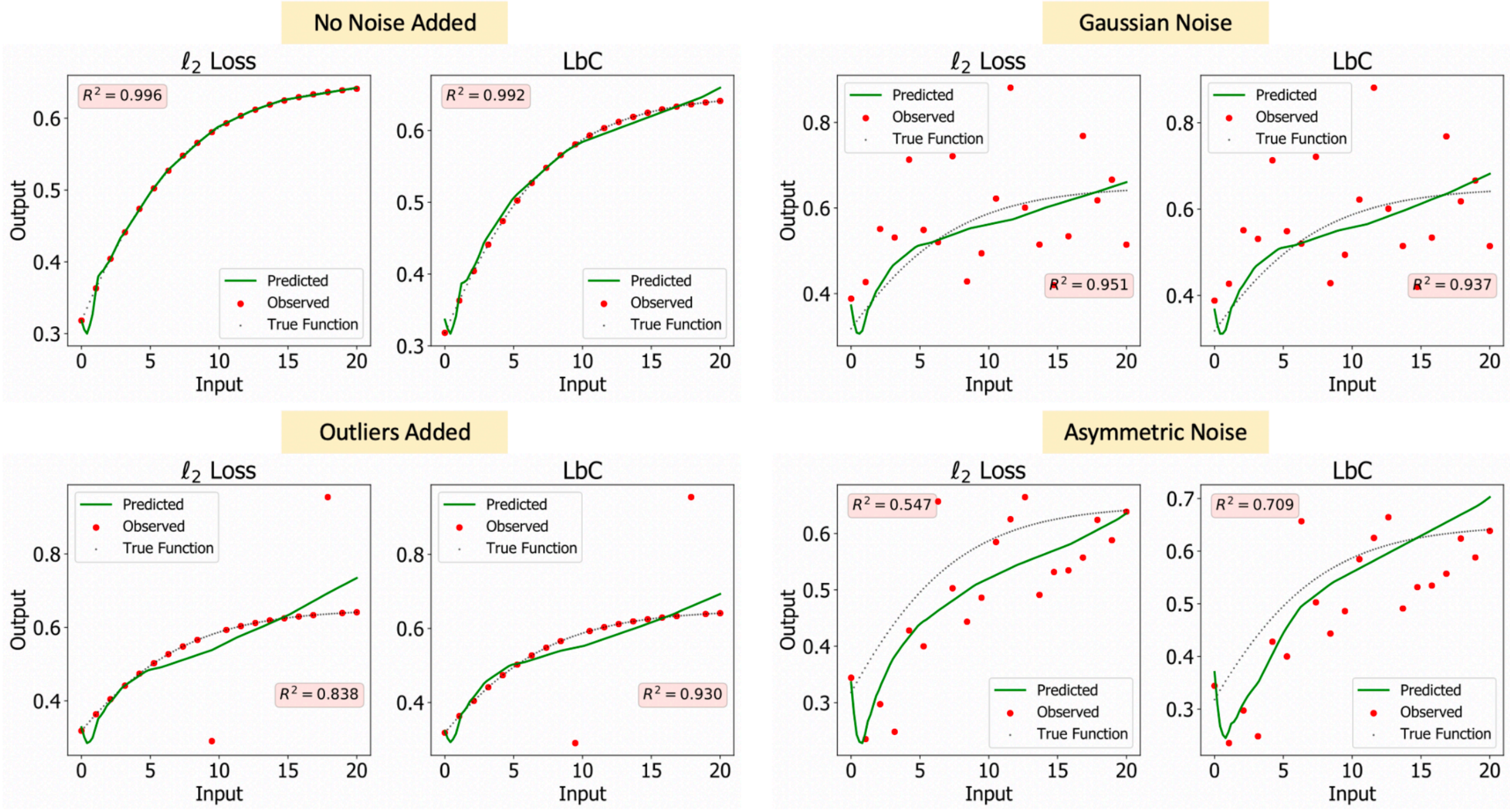
Point estimate $\hat{\mathbf{y}}$

$(\delta^{l,\alpha}, \delta^{u,\alpha})$ Interval estimates

Prediction $[\hat{\mathbf{y}} - \delta^{l,\alpha}, \hat{\mathbf{y}} + \delta^{u,\alpha}]$

Machine Learning Eventually Boils Down to Optimizing Parameters Based on a Fidelity Metric

Example: Approximating a synthetic function using a 1-layer MLP


No Noise Added

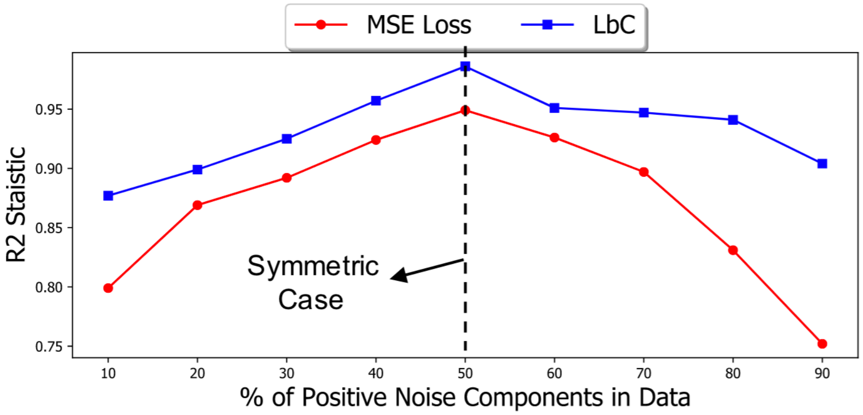
Gaussian Noise

Outliers Added

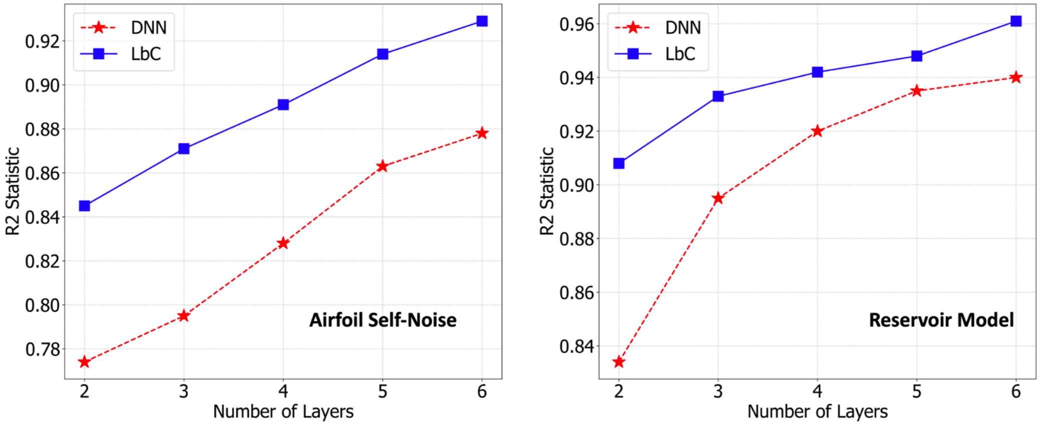
Asymmetric Noise

Machine Learning Eventually Boils Down to Optimizing Parameters Based on a Fidelity Metric

LbC consistently outperforms existing symmetric losses when the residual distribution is skewed



LbC requires lesser number of model parameters to match the performance of a standard model



Sometimes, we need to build our own architectures…

In Deep Learning, Model Architectures Act as a Prior in Defining the Hypothesis Space

Convolutional networks that enforce local smoothness act as a strong prior for images – Even untrained networks can provide useful feature representations.

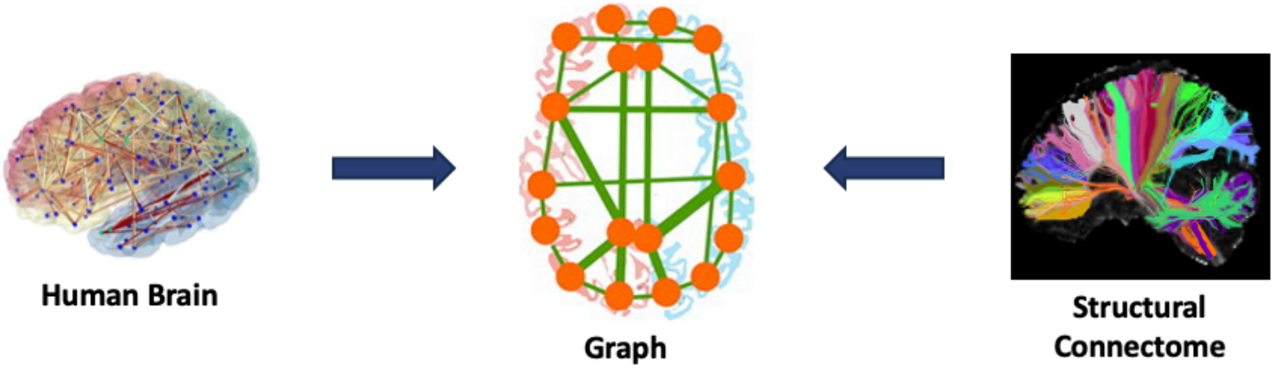
In scientific problems, sometimes we need to design custom architectures to provide strong inductive biases.

These solutions rarely generalize, however, can provide significant performance gains for the problem at hand.

Designing Custom Model Architectures to Provide Strong Inductive Biases

Example: Modeling human brain connectomes for predictive modeling

Connectome is a comprehensive map of structural and functional connectivity of neural pathways in the brain.

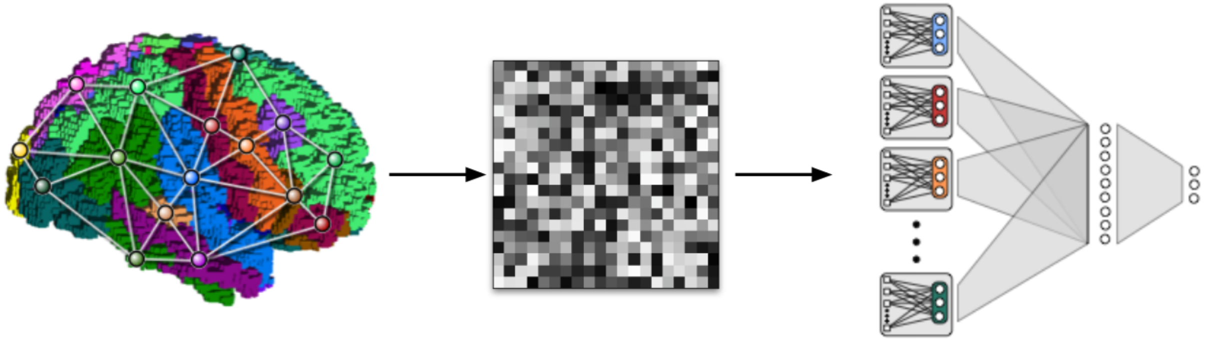


Connectome Graph generation from T1 images and probabilistic Tractography

Designing Custom Model Architectures to Provide Strong Inductive Biases

Example: Modeling human brain connectomes for predictive modeling

Since connectomes can be viewed as a graph, one might be tempted to use graph neural networks (e. g. molecular chemistry) – Information diffusion property does not hold.



Parcellation of Brain to 84 different regions

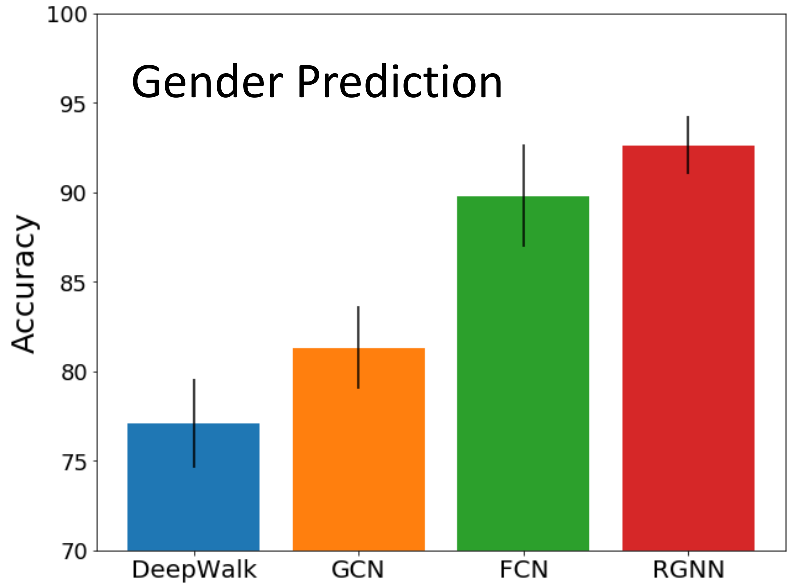
Weighted connectivities between regions indicating number of fibres

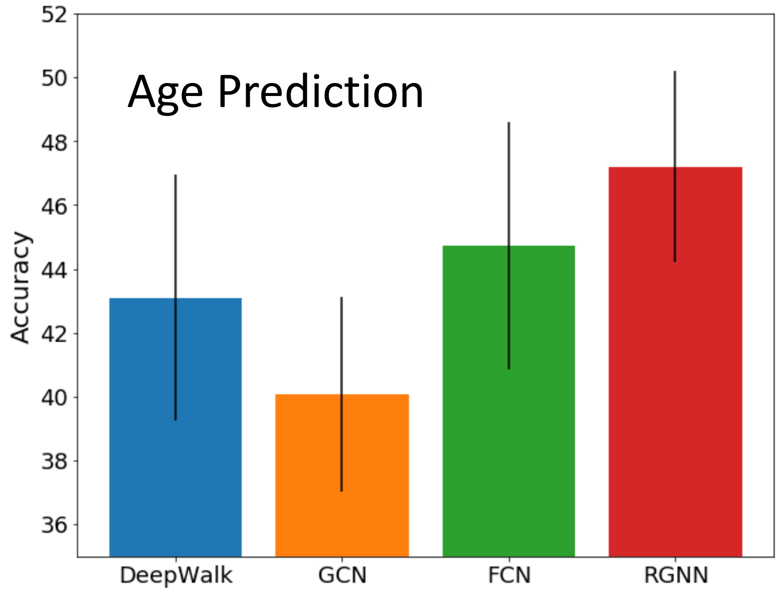
RGNN mirrors the hierarchical structure in connectome

We proposed a novel message passing framework that is aptly suited to process relational structure without the need for diffusion.

Designing Custom Model Architectures to Provide Strong Inductive Biases

Example: Modeling human brain connectomes for predictive modeling



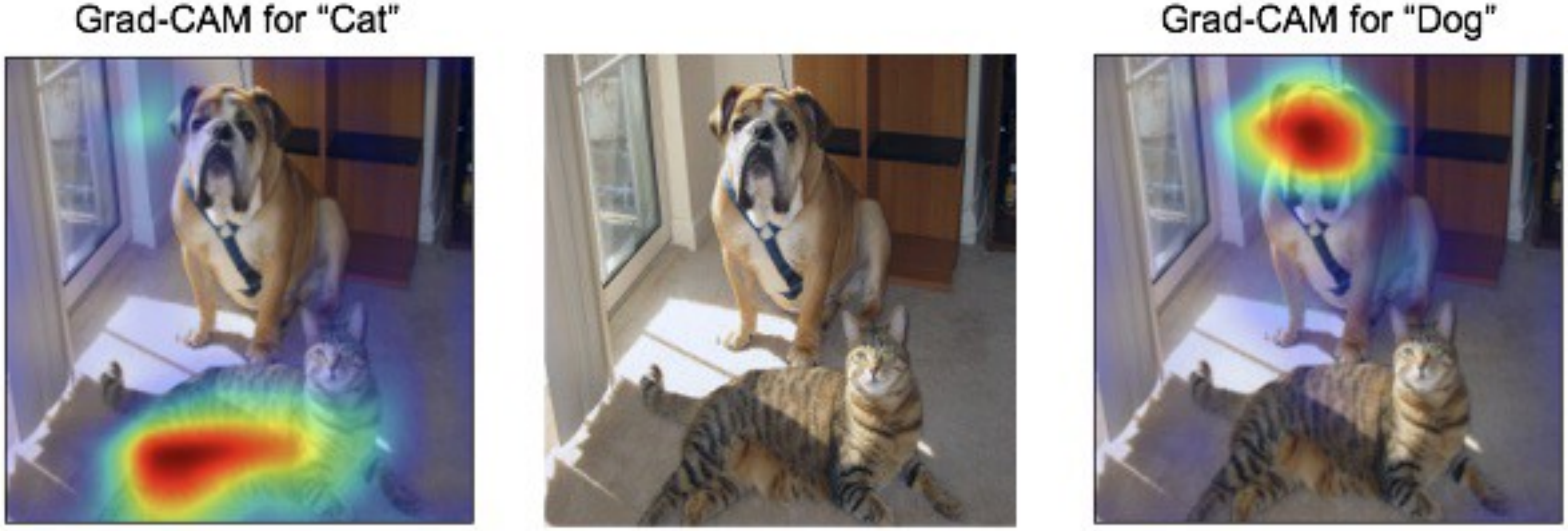


Do not just predict, but also introspect!

Using Machine Learning Models in Critical Applications Requires a Rigorous Characterization of its Behavior

Gaining scientific insights from learned models needs introspection driven by *user-specified hypotheses*.

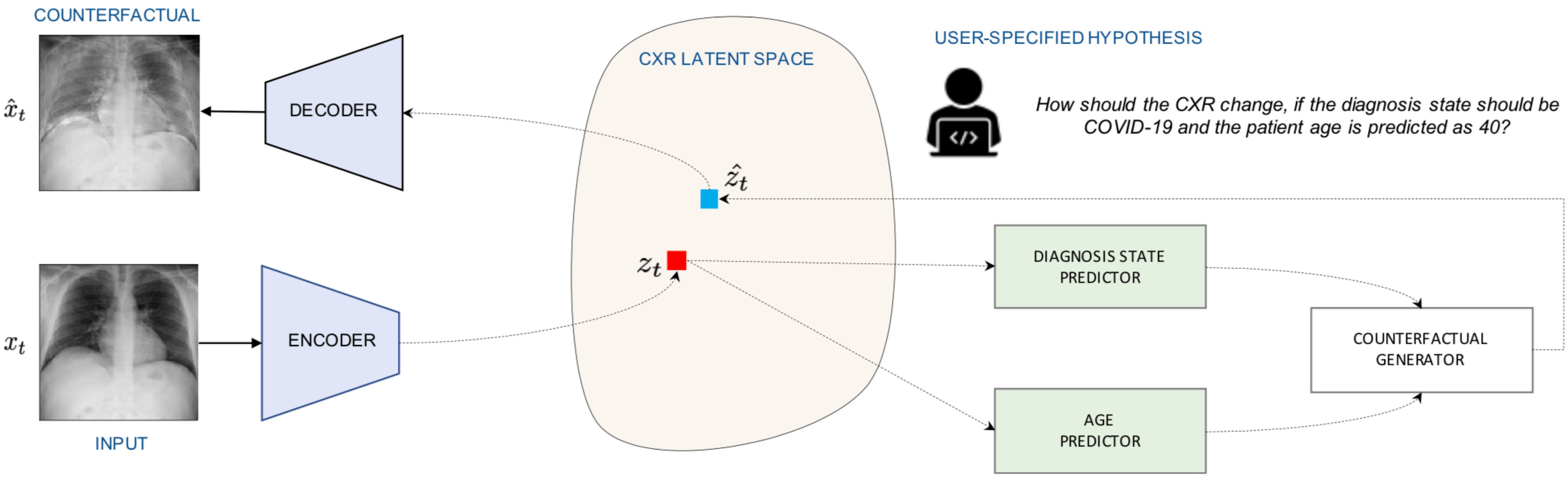
A broad class of interpretability techniques exist to uncover the “most plausible” explanations for a decision.



Using Machine Learning Models in Critical Applications Requires a Rigorous Characterization of its Behavior

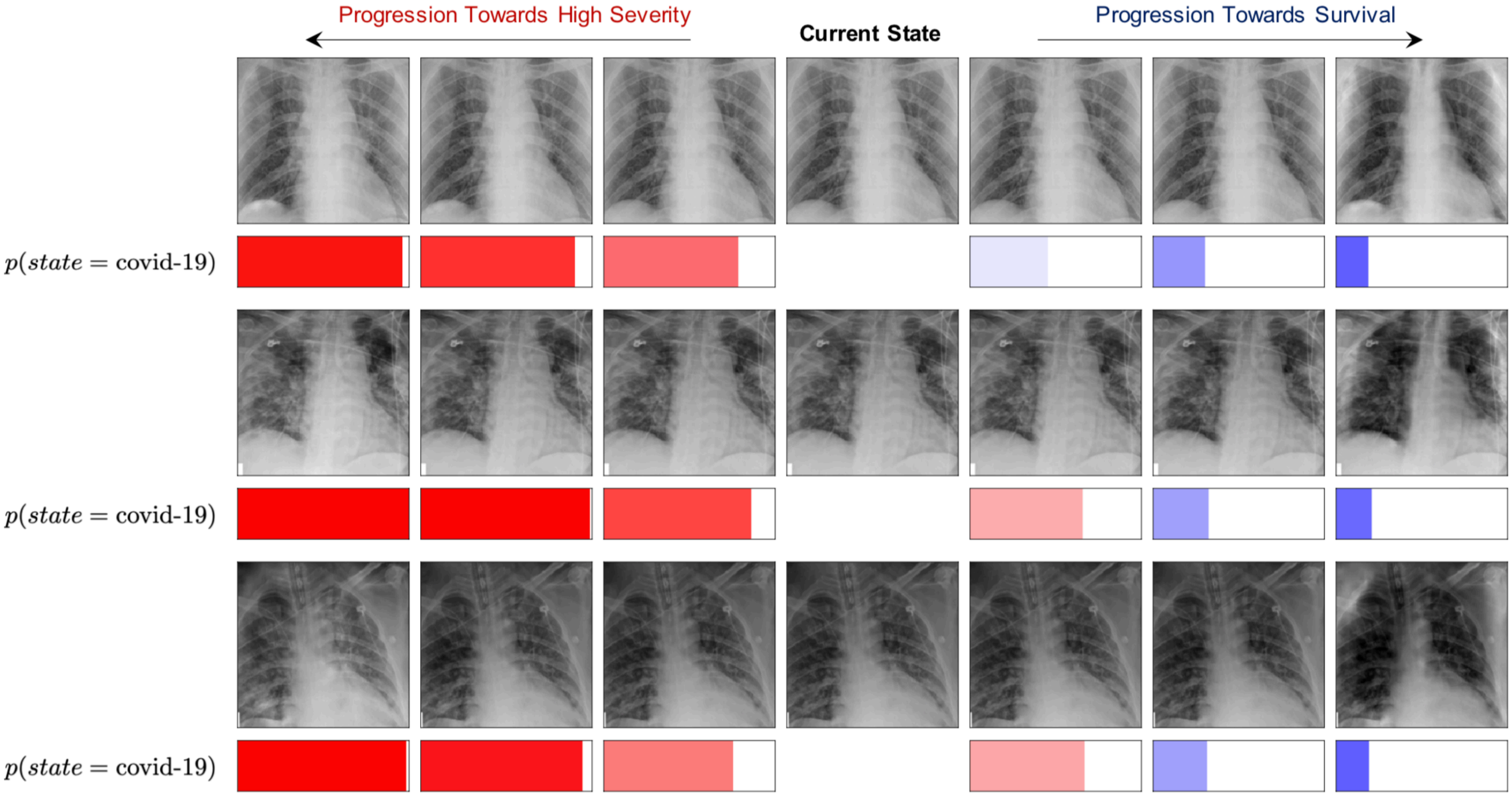
However, in science problems, we need to deal with more complex hypotheses.

Example: Analysis of COVID-19 infections from CXR data using counterfactual reasoning


COUNTERFACTUAL
$\hat{x}_t$
DECODER
CXR LATENT SPACE
$\hat{z}_t$
z_t
USER-SPECIFIED HYPOTHESIS
How should the CXR change, if the diagnosis state should be COVID-19 and the patient age is predicted as 40?
DIAGNOSIS STATE PREDICTOR
COUNTERFACTUAL GENERATOR
AGE PREDICTOR
x_t
ENCODER
INPUT

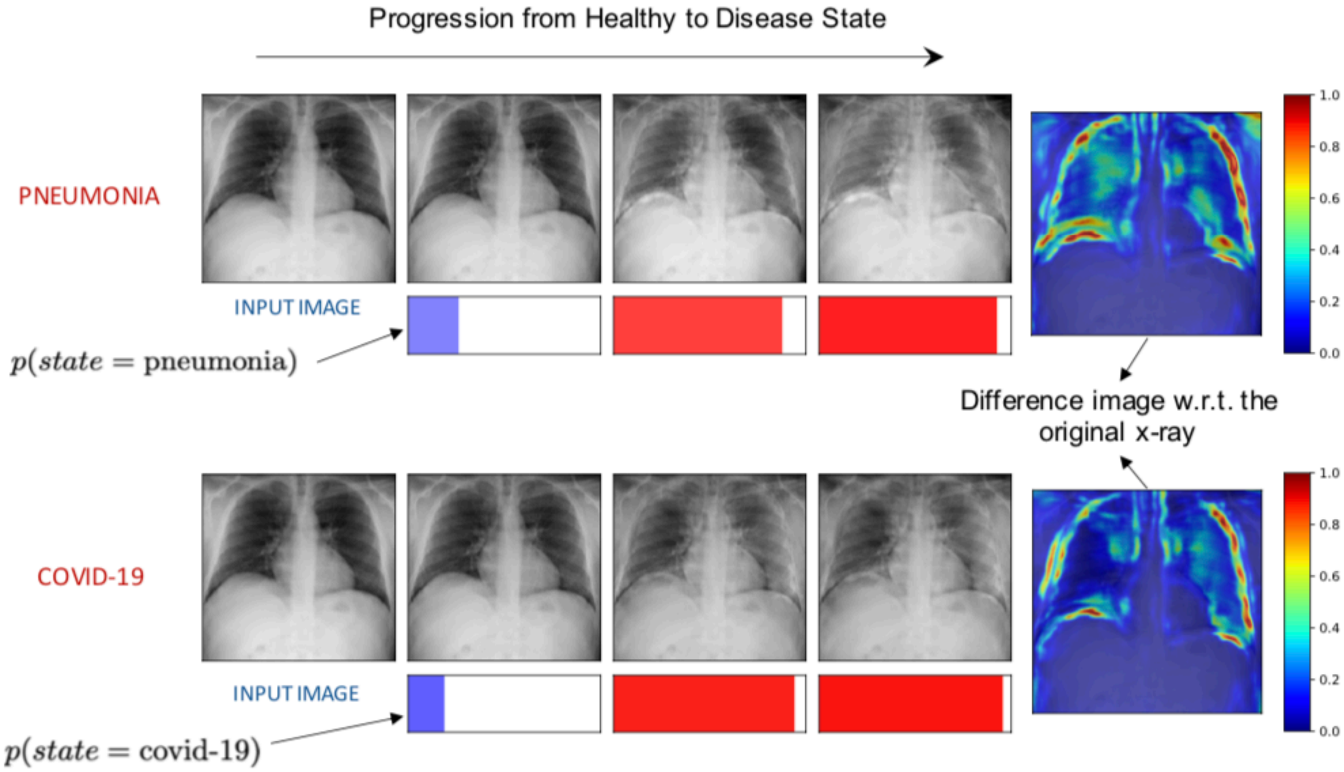
Using Machine Learning Models in Critical Applications Requires a Rigorous Characterization of its Behavior

Hypothesis: The severity of the disease infection increases or decreases


Progression Towards High Severity
Current State
Progression Towards Survival
$p(state = \text{covid-19})$
$p(state = \text{covid-19})$
$p(state = \text{covid-19})$

Using Machine Learning Models in Critical Applications Requires a Rigorous Characterization of its Behavior

Hypothesis: Manifestation of COVID-19 is different from other types of pneumonia known before


Progression from Healthy to Disease State
PNEUMONIA
INPUT IMAGE
$p(state = \text{pneumonia})$
1.0
0.8
0.6
0.4
0.2
0.0
Difference image w.r.t. the original x-ray
COVID-19
INPUT IMAGE
$p(state = \text{covid-19})$
1.0
0.8
0.6
0.4
0.2
0.0

Summary

Scientific data provide unique challenges and opportunities towards advancing machine learning

🤝

Using suitable machine learning methods will help accelerate scientific analysis and discovery

Team

Rushil Anirudh

Bhavya Kailkhura

Timo Bremer

Vivek N

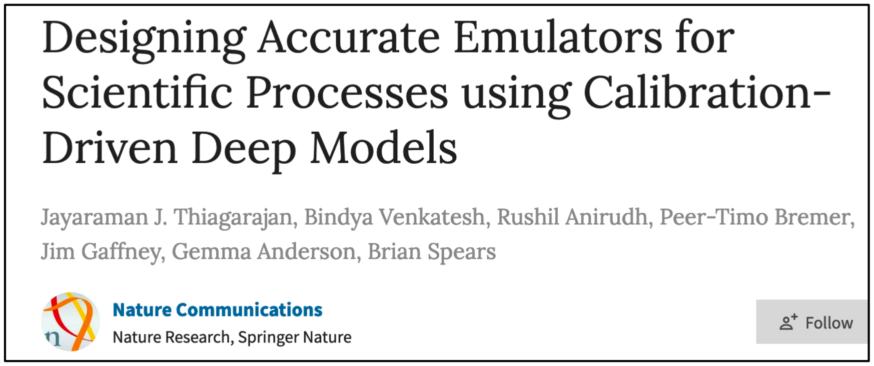
Bindya Venkatesh

Uday Shankar S

Relevant Papers

RESEARCH ARTICLE

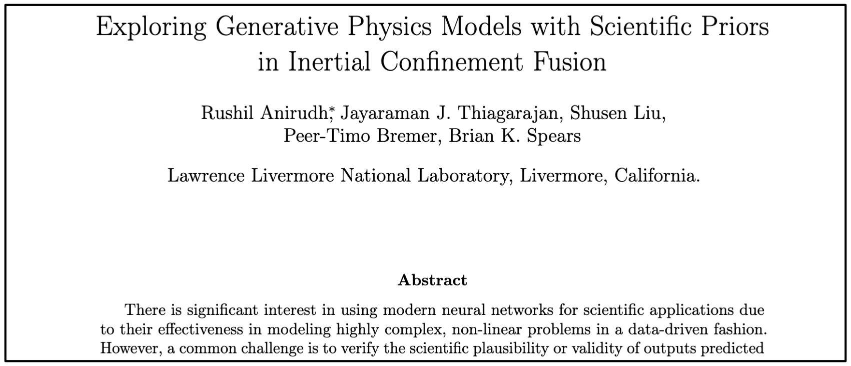
Improved surrogates in inertial confinement fusion with manifold and cycle consistencies



Designing Accurate Emulators for Scientific Processes using Calibration-Driven Deep Models

Jayaraman J. Thiagarajan, Bindya Venkatesh, Rushil Anirudh, Peer-Timo Bremer, Jim Gaffney, Gemma Anderson, Brian Spears

Nature Communications
Nature Research, Springer Nature



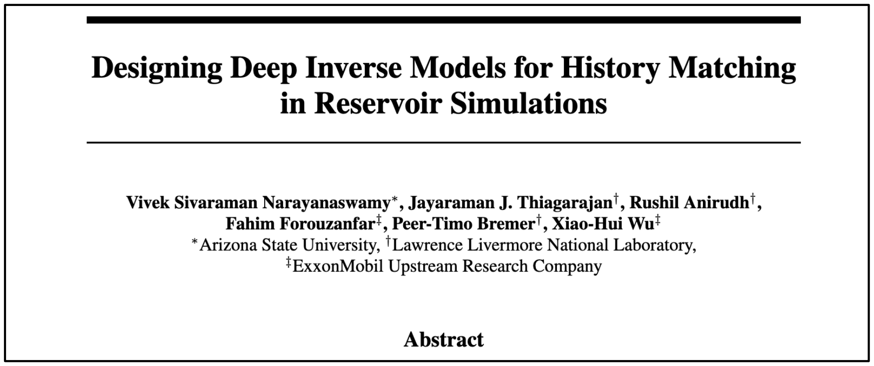
Exploring Generative Physics Models with Scientific Priors in Inertial Confinement Fusion

Rushil Anirudh*, Jayaraman J. Thiagarajan, Shusen Liu, Peer-Timo Bremer, Brian K. Spears

Lawrence Livermore National Laboratory, Livermore, California.

Abstract

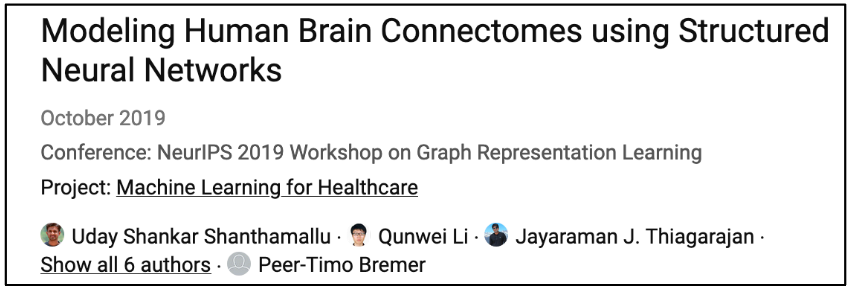
There is significant interest in using modern neural networks for scientific applications due to their effectiveness in modeling highly complex, non-linear problems in a data-driven fashion. However, a common challenge is to verify the scientific plausibility or validity of outputs predicted



Designing Deep Inverse Models for History Matching in Reservoir Simulations

Vivek Sivaraman Narayanaswamy*, Jayaraman J. Thiagarajan†, Rushil Anirudh†, Fahim Forouzanfar‡, Peer-Timo Bremer†, Xiao-Hui Wu‡
*Arizona State University, †Lawrence Livermore National Laboratory, ‡ExxonMobil Upstream Research Company

Abstract



Modeling Human Brain Connectomes using Structured Neural Networks

October 2019

Conference: NeurIPS 2019 Workshop on Graph Representation Learning

Project: Machine Learning for Healthcare

Uday Shankar Shanthamallu · Qunwei Li · Jayaraman J. Thiagarajan · Show all 6 authors · Peer-Timo Bremer

Published: 05 March 2020

MimicGAN: Robust Projection onto Image Manifolds with Corruption Mimicking

Rushil Anirudh, Jayaraman J. Thiagarajan, Bhavya Kailkhura & Peer-Timo Bremer

Questions?