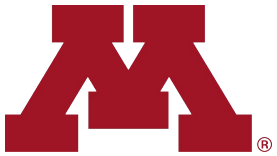
How can Physics Inform Deep Learning Methods in Scientific Problems:

Recent Progress and Future Prospects



Anuj Karpatne

Post-Doctoral Associate,

University of Minnesota

karpa009@umn.edu

http://www.cs.umn.edu/~anuj

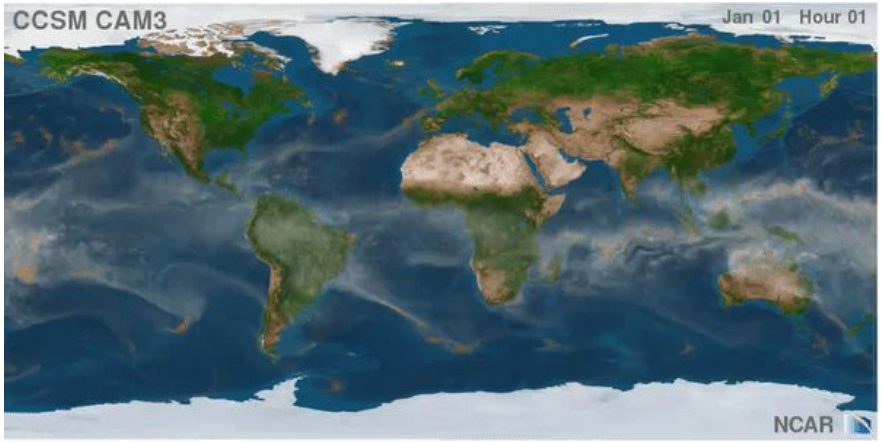
Outline

- Why Deep Learning Needs Physics?
- “Theory-guided Data Science”
- Recent Progress
- Future Prospects

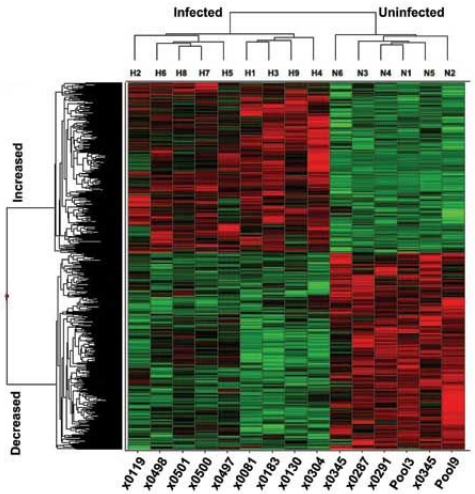
Big Data in Physical and Life Sciences

Earth Science

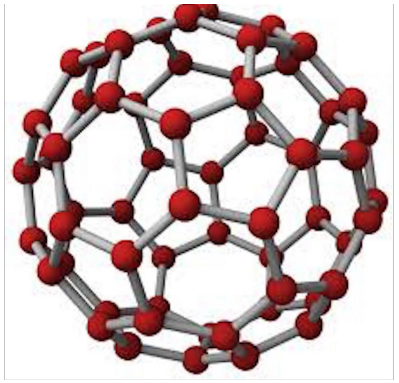
- Satellite Data
- In-situ Sensors
- Model Simulations
- Experimental Data
- Survey Reports



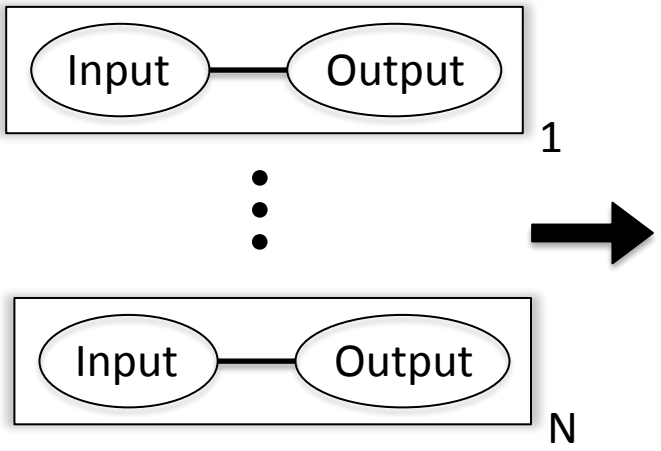
Genomics



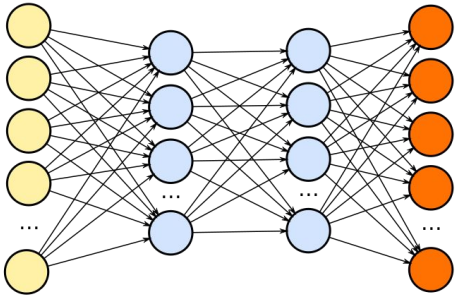
Material Science



Age of Data Science

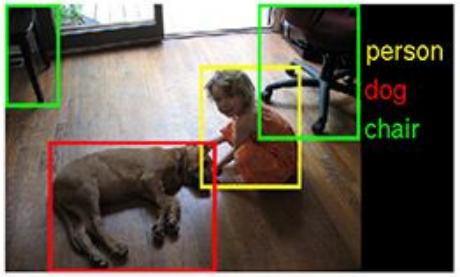


Deep Learning



"**Black-box**" models learn patterns and models solely from data without relying on scientific knowledge

- Hugely successful in commercial applications:



DeepMind

Google Ads

facebook

NETFLIX

Promise of Data Science in Transforming Scientific Discovery

"Unlike earlier attempts … [AI systems] can see patterns and spot anomalies in data sets far larger and messier than human beings can cope with."

July 7 2017 Issue

Promise of Data Science in Transforming Scientific Discovery

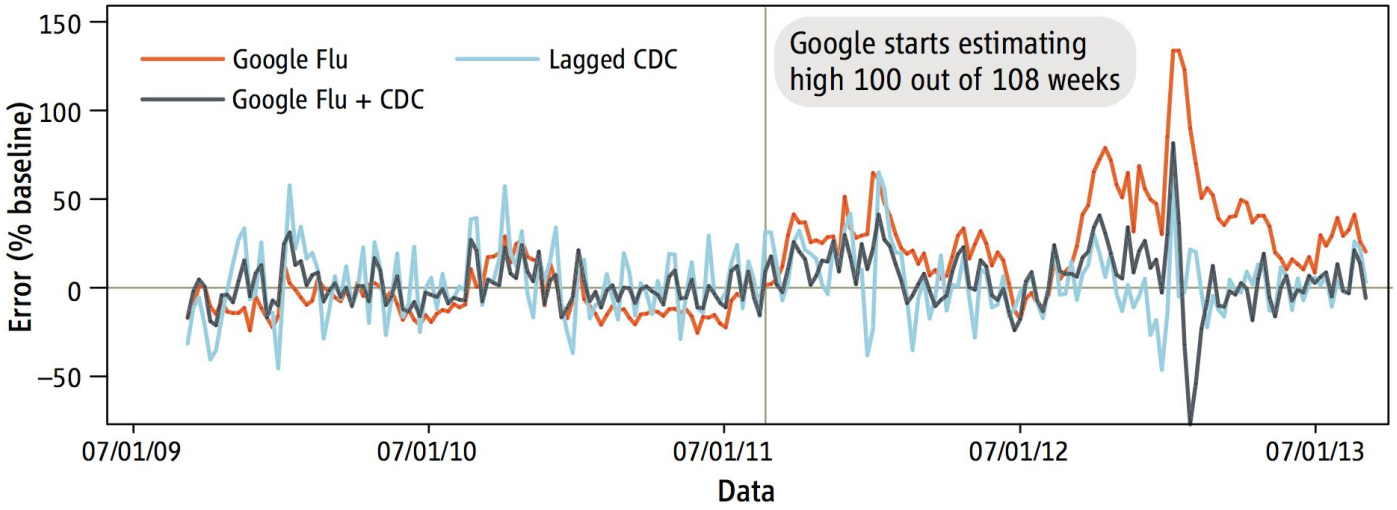
Will the rapidly growing area of "black-box" data science models make existing theory-based models obsolete?

Limits of “Black-box” Data Science Methods

Science

The Parable of Google Flu: Traps in Big Data Analysis

- Predicted flu using Google search queries
- Overestimated by twice in later years



***Climate Science*:**

Geophysical Research Letters

Statistical significance of climate sensitivity predictors obtained by data mining

The New York Times

The Opinion Pages | OP-ED CONTRIBUTORS

Eight (No, Nine!) Problems With Big Data

By GARY MARCUS and ERNEST DAVIS APRIL 6, 2014

" . . . you will always need to start with an analysis that relies on an understanding of physics and biochemistry."

Why Do Black-box Methods Fail? (1/2)

- Scientific problems are often under-constrained
  - Complex, dynamic, and non-stationary relationships
  - Large number of variables, small number of samples

- Standard methods for evaluating ML models (e.g., cross-validation) break down
  - Easy to learn *spurious relationships* that look deceptively good on training and test sets
  - But lead to poor generalization outside the available data

Huge number of samples is critical to success of methods such as deep learning

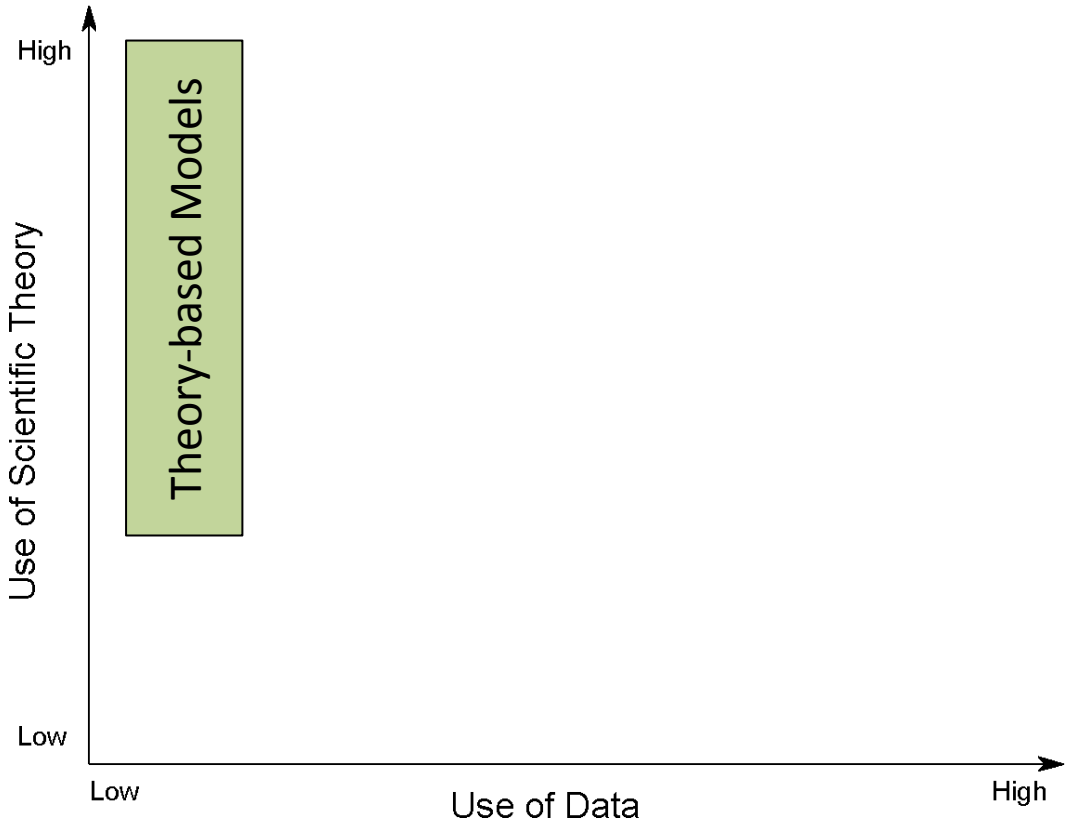
Why Do Black-box Methods Fail? (2/2)

- Interpretability is an important end-goal (esp. in scientific problems)

- Need to explain or discover mechanisms of underlying processes to …
  - Form a basis for scientific advancements
  - Safeguard against the learning of non-generalizable patterns

Theory-based vs. Data Science Models

Contain knowledge gaps in describing certain processes (turbulence, groundwater flow)



Gravitational Law

$$F = G\frac{m_1 m_2}{r^2}$$

Conservation of Mass, Momentum, Energy

$$\frac{\partial \rho}{\partial t} = -\nabla.(\rho \mathbf{u})$$

$$\frac{\partial \rho \mathbf{u}}{\partial t} = -\nabla.\left(\frac{1}{\rho}(\rho \mathbf{u})\otimes(\rho \mathbf{u}) + p\mathbf{I}\right) + \rho \mathbf{g}$$

$$\frac{\partial E}{\partial t} = -\nabla.\left(\frac{1}{\rho}(E+p)(\rho \mathbf{u})\right) + \mathbf{u}.\rho \mathbf{g}$$

Navier-Stokes Equation

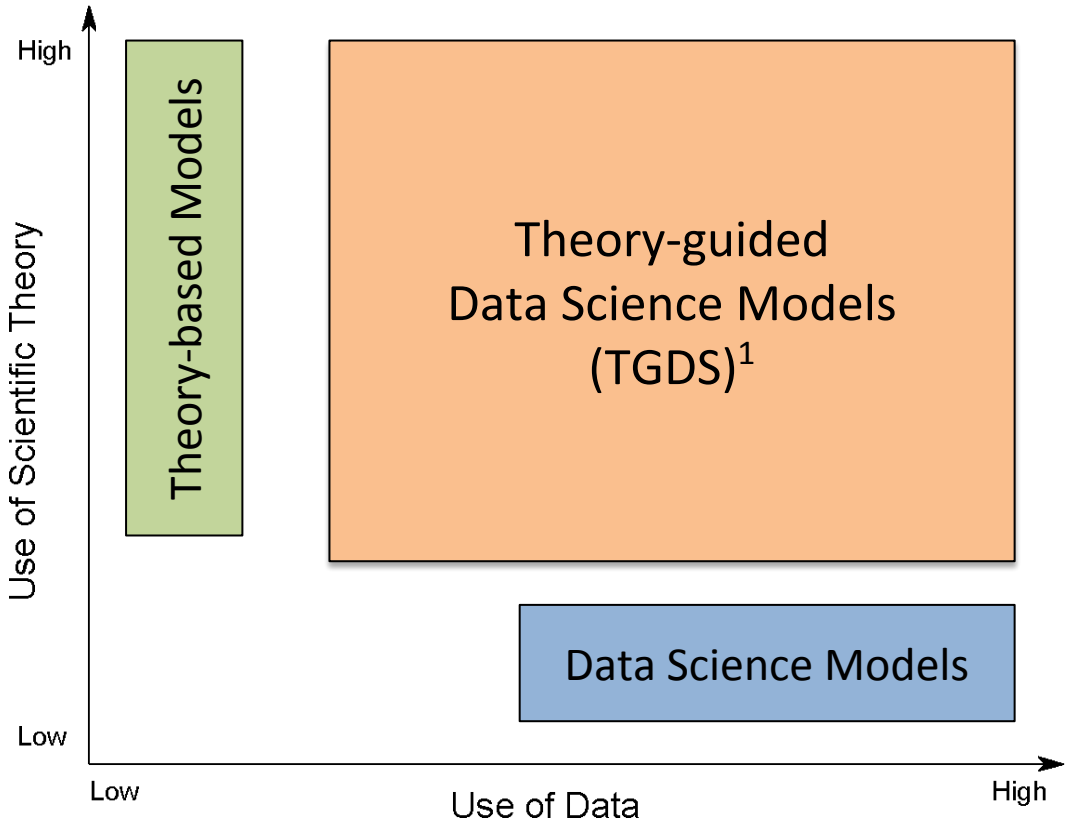
$$\rho\left(\frac{\partial v}{\partial t} + v \cdot \nabla v\right) = -\nabla p + \nabla \cdot T + f$$

Schrodinger's Equation

$$\mathbf{H}\Psi = E\Psi$$

Theory-based vs. Data Science Models

Contain knowledge gaps in describing certain processes (turbulence, groundwater flow)



Take full advantage of data science methods without ignoring the treasure of accumulated knowledge in scientific "theories"

Require large number of representative samples

[1] Karpatne et al. "**Theory-guided data science: A new paradigm for scientific discovery**," TKDE 2017

Theory-guided Data Science:
Emerging Applications

- Earth Science:
  - Karpatne et al., "Physics-guided Neural Networks: Application in Lake Temperature Modeling," SDM 2018 (in review).
  - Faghmous et al., "Theory-guided data science for climate change," IEEE Computer, 2014.
  - Faghmous and Kumar, "A big data guide to understanding climate change: The case for theory-guided data science," Big data, 2014.
- Fluid Dynamics:
  - Singh et al., "Machine learning- augmented predictive modeling of turbulent separated flows over airfoils," arXiv, 2016.
- Material Science:
  - Curtarolo et al., "The high-throughput highway to computational materials design," Nature Materials, 2013.
- Computational Chemistry:
  - Li et al., "Understanding machine-learned density functionals," International Journal of Quantum Chemistry, 2015.
- Neuroscience, Biomedicine, Particle Physics, …

PHYSICS INFORMED MACHINE LEARNING

Symposium by Los Alamos National Laboratory, 2016, 2018

Workshop on Deep Learning for Physical Sciences 2017

FORESIGHT INSTITUTE

AI for Scientific Progress, 2016

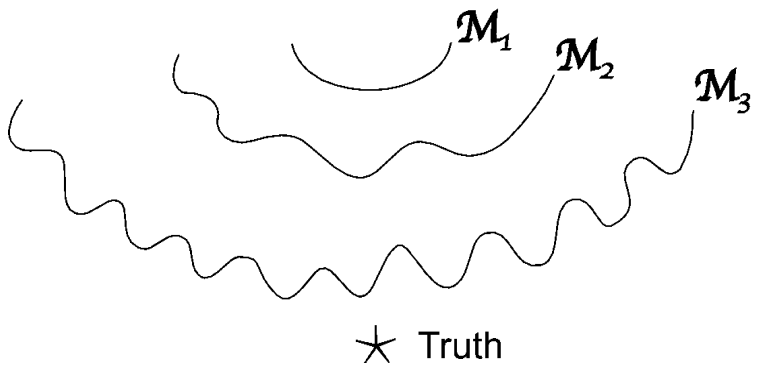
IBM Research

"Physical Analytics" Research Division

An Overarching Objective of TGDS

Learning Physically Consistent Models

- Traditionally, “simpler” models are preferred for generalizability
  - Basis of several statistical principles such as bias-variance trade-off



M1 (less complex model):
High bias—Low variance

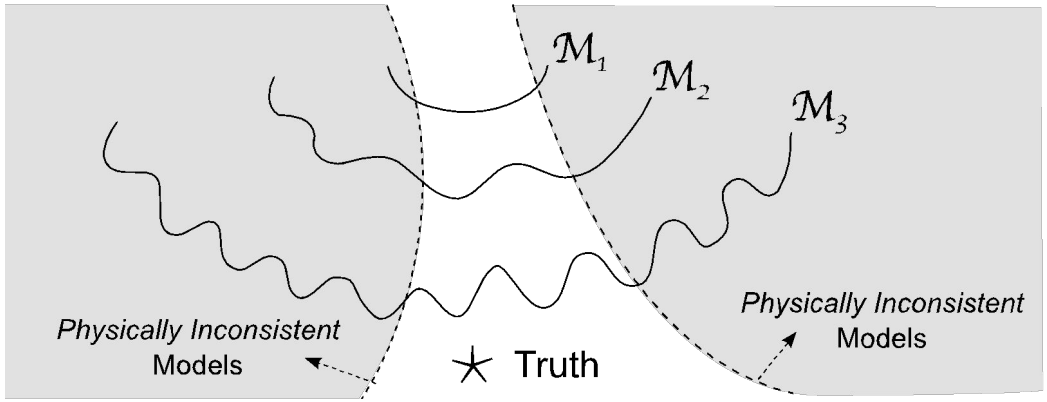
M3 (more complex model):
Low bias—High variance

Generalization Performance $\propto$ Accuracy + Simplicity

An Overarching Objective of TGDS

Learning Physically Consistent Models

- Traditionally, "simpler" models are preferred for generalizability
  - Basis of several statistical principles such as bias-variance trade-off



M1 (less complex model):
High bias—Low variance

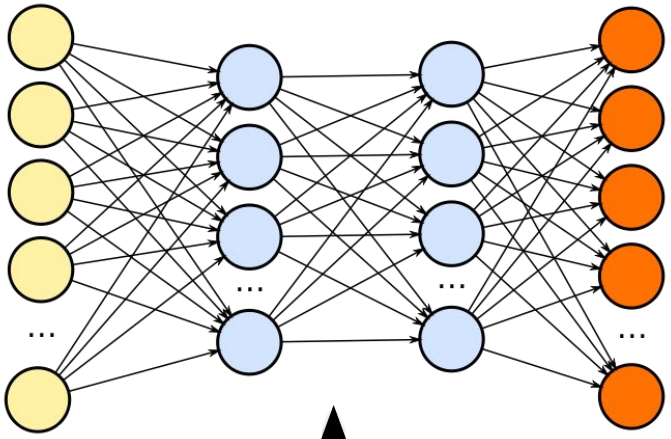
M3 (more complex model):
Low bias—High variance

- In scientific problems, **"physical consistency"** can be used as another measure of generalizability
  - Can help in pruning large spaces of inconsistent solutions
  - Result in generalizable *and* physically meaningful results

Generalization Performance $\propto$ Accuracy + Simplicity + ***Consistency***

Physics-Guided Neural Networks (PGNN)

A Framework for Learning **Physically Consistent** Deep Learning Models



Scientific Knowledge (Physics)

Used to guide selection of model architecture, activation functions, loss functions, …

Karpatne et al., "Physics-guided neural networks (PGNN): Application in Lake Temperature Modeling," SDM 2018 (in review; **arXiv: 1710.11431**).

Case Study: Lake Temperature Modeling

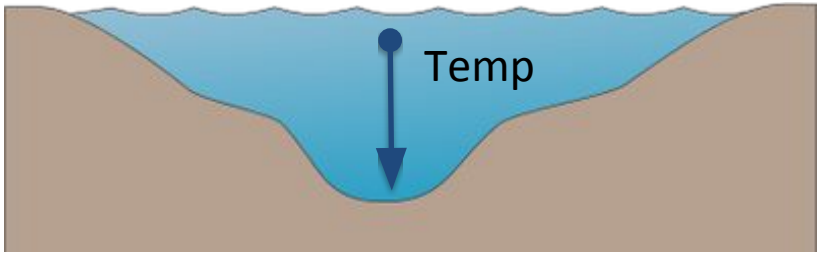
Center for Limnology
University of Wisconsin–Madison

Input Drivers:

Short-wave Radiation,
Long-wave Radiation,
Air Temperature,
Relative Humidity,
Wind Speed,
Rain, …

Target Output:

Temp. of water at every depth



Physics-based Approach: General Lake Model (GLM)[1]

- Captures physical processes responsible for energy balance
- Requires *lake-specific calibration* using large amounts of data and computational resources

RMSE of Uncalibrated Model: 2.57
RMSE of Calibrated Model: 1.26

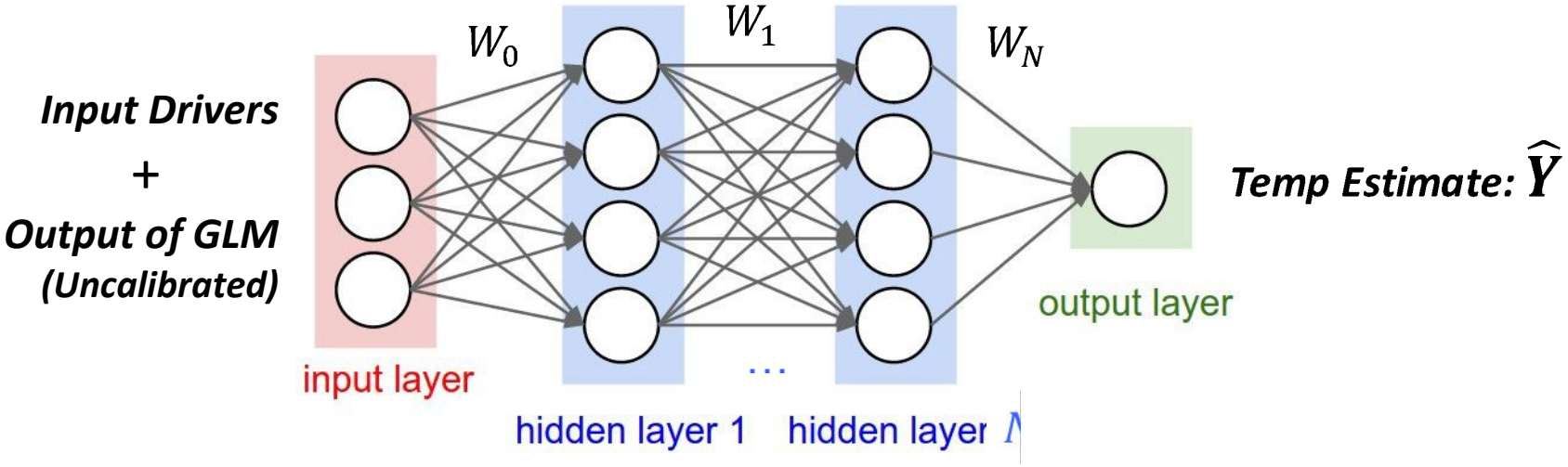
(for Lake Mille Lacs in Minnesota)

[1]Hipsey et al., 2014

PGNN 1:

Use GLM Output as Input in Neural Network

- Deep Learning can augment physics-based models by modeling their *errors*
- Part of a broader research theme on creating hybrid-physics-data models

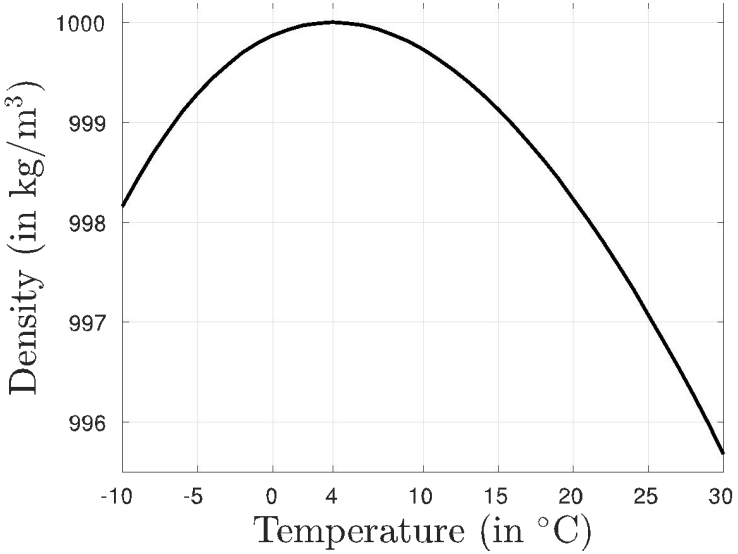


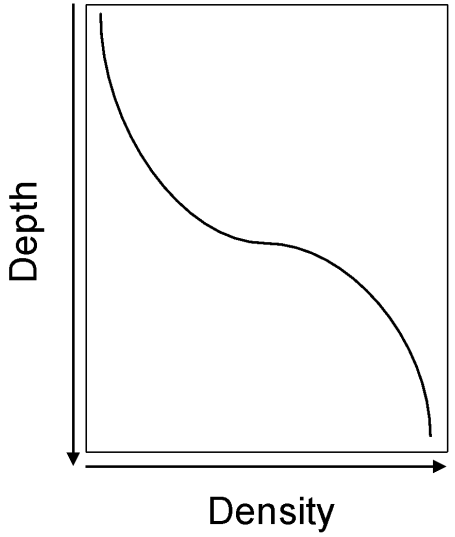
$$\text{Loss Function} = \text{Training Loss}\left(Y, \hat{Y}\right) + \lambda_1 \|W\|_1 + \lambda_2 \|W\|_2$$

PGNN 2:

Use **Physics-based Loss Functions**

- Temp estimates need to be consistent with physical relationships b/w temp, density, and depth





Physical Constraint:
Denser water is at higher depth

$$\Delta_i = \hat{\rho}_i - \hat{\rho}_{i+1} \geq 0$$

If depth $d_i >$ depth d_{i+1}

Convert $\hat{Y}$ to density estimate $\hat{\rho}$ **Physics-based Loss** = Sum of Physical Violations

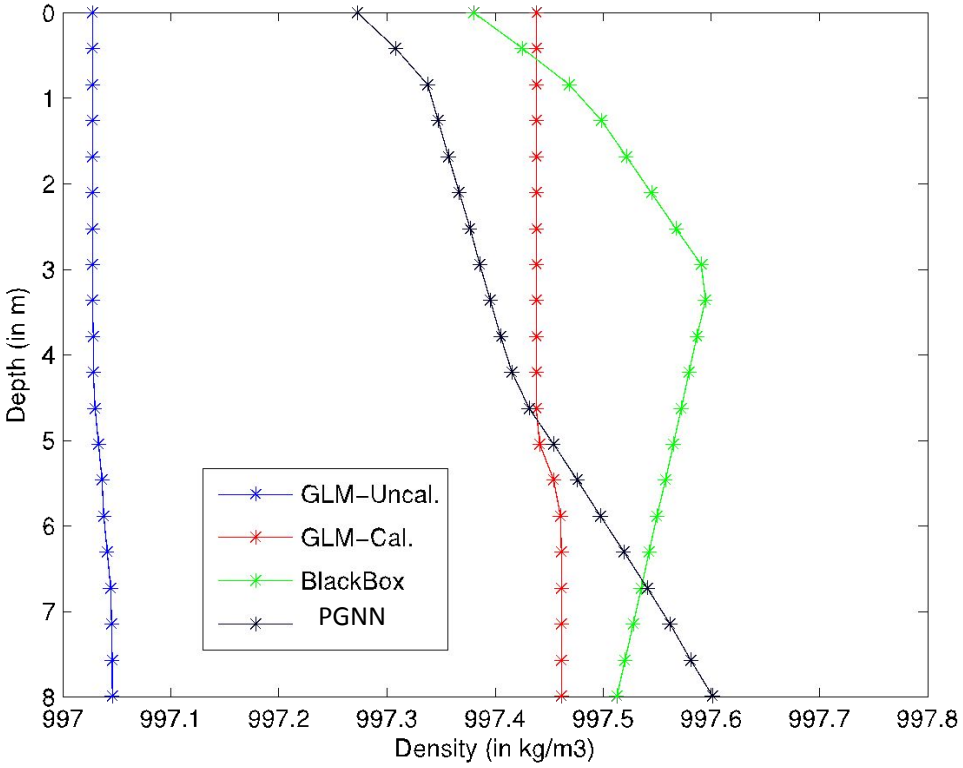
$$= \sum_i \mathrm{ReLU}(\Delta_i)$$

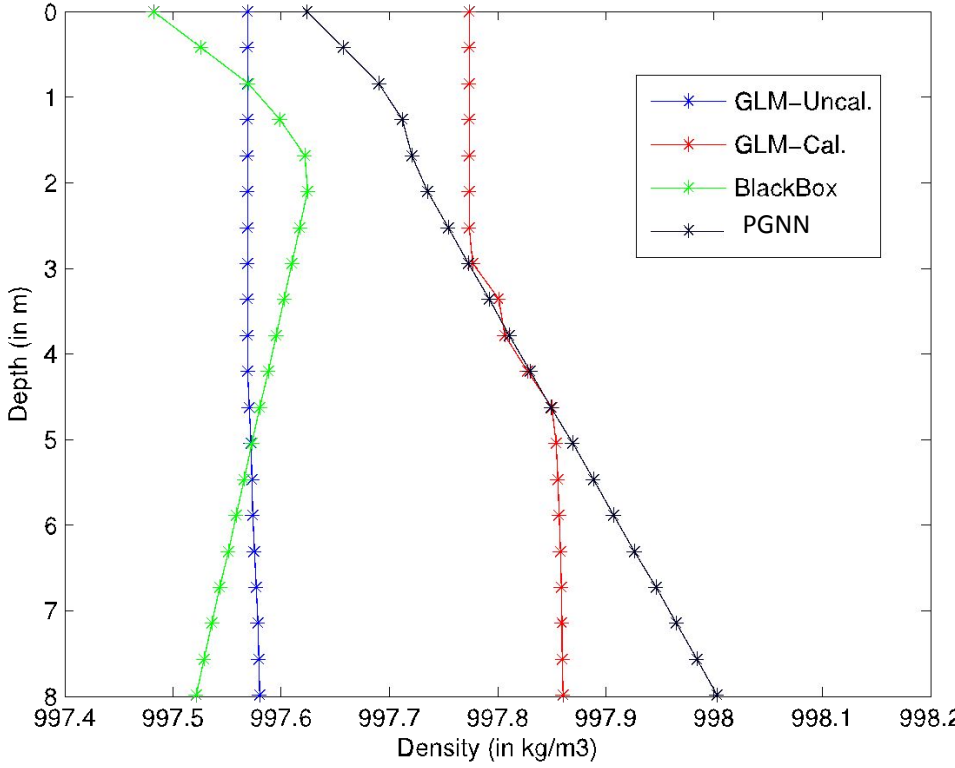
$$\text{Loss Function} = \text{Tr. Loss}\left(Y, \hat{Y}\right) + \lambda_1 \|W\|_1 + \lambda_2 \|W\|_2 + \lambda_{\text{PHY}} \textbf{ Physics-based Loss}$$

Does not require labels!

Physical Consistency Ensures Generalizability

	GLM (Uncalibrated)	Black-box Neural Network	PGNN	GLM (Calibrated)
RMSE (in °C)	2.57	1.77	**1.16**	1.26





Future Prospects: Theory-guided Data Science

1. Theory-guided Learning

- Choice of Loss Function
- Constrained Optimization Methods
- Probabilistic Models

[Limnology, Chemistry, Biomedicine, Climate, Genomics]

2. Theory-guided Design

- Choice of Response/Loss Functions
- Design of Model Architecture

[Turbulence Modeling, Neuroscience]

3. Theory-guided Refinement

- Post-processing
- Pruning

[Remote Sensing, Material Science]

4. Creating Hybrid Models of Theory and Data Science

- Residual Modeling
- Predicting Intermediate Quantities

[Hydrology, Turbulence Modeling]

5. Augmenting Theory-based Models using Data

- Calibrating Model Parameters
- Data Assimilation

[Hydrology, Climate Science, Fluid Dynamics]

Concluding Remarks

- "Black-box" deep learning methods not sufficient for knowledge discovery in scientific domains
- Physics can be combined with deep learning in a variety of ways under the paradigm of **"theory-guided data science"**
- Use of physical knowledge ensures physical consistency as well as generalizability
- Theory-guided data science is already starting to gain attention in several disciplines:
  - Climate science and hydrology
  - Turbulence modeling
  - Bio-medical science
  - Bio-marker discovery
  - Material discovery
  - Computational chemistry, …

Thank You!

- Karpatne, A., Atluri, G., Faghmous, J.H., Steinbach, M., Banerjee, A., Ganguly, A., Shekhar, S., Samatova, N. and Kumar, V., *"Theory-Guided Data Science: A New Paradigm for Scientific Discovery from Data"*. IEEE Transactions on Knowledge and Data Engineering, 29(10), pp.2318-2331, 2017.

- Karpatne, A., Watkins W., Read, J., and Kumar, V., *"Physics-guided Neural Networks (PGNN): An Application in Lake Temperature Modeling"*. SIAM International Conference on Data Mining 2018 (in review; arXiv: 1710.11431).

- Contact: karpa009@umn.edu