CS 6804: Science-guided Machine Learning (SGML)

An Emerging Field of Research
Combining Scientific Knowledge with Machine Learning

Course Webpage: http://people.cs.vt.edu/karpatne/teaching/6804-f20/index.html

Anuj Karpatne

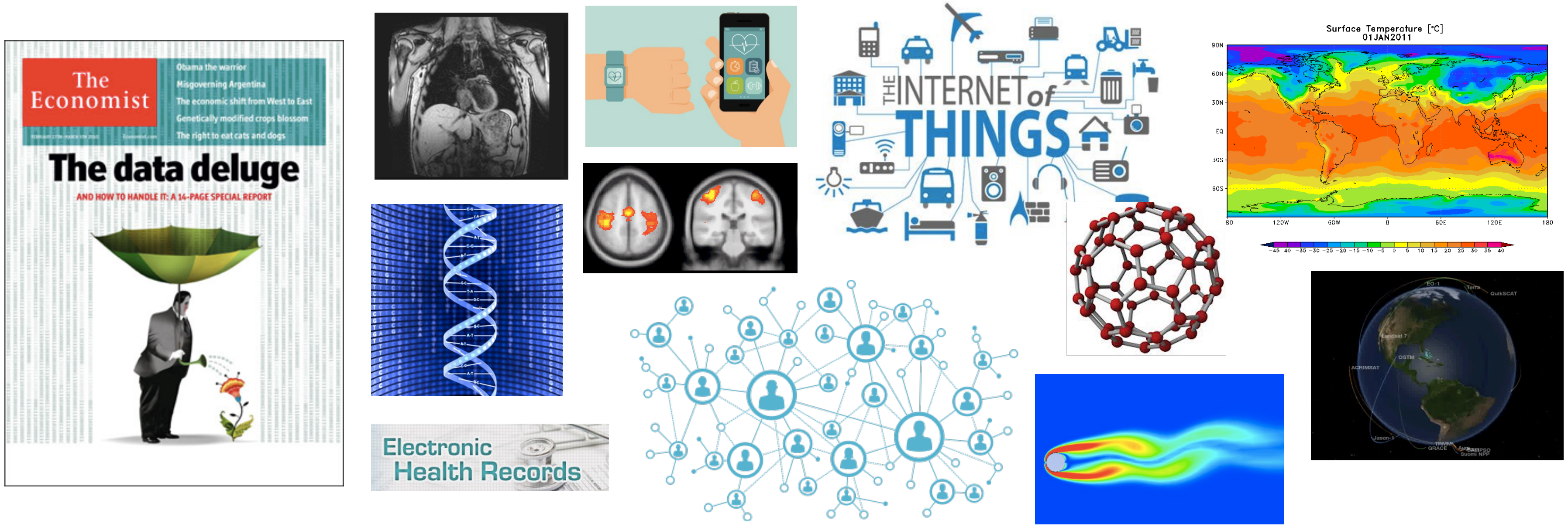
Assistant Professor, Computer Science

Virginia Tech

karpatne@vt.edu

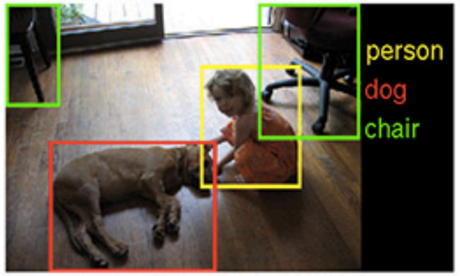
https://people.cs.vt.edu/karpatne/

Golden Age of Machine Learning / Artificial Intelligence



- Hugely successful in commercial applications:

IMAGENET



DeepMind

Google Ads

facebook

NETFLIX

Google AI algorithm masters ancient game of Go

Golden Age of Machine Learning / Artificial Intelligence

- Promise of Machine Learning (ML) in Accelerating Scientific Discovery

Will the rapidly growing area of **“black-box”** ML models make existing theory-based models obsolete?

- But disappointing results in scientific domains!
  - Require lots of labeled data
  - Unable to provide valuable physical insights

Science **The Parable of Google Flu: Traps in Big Data Analysis**

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

Science-based vs. Data-based Models

- Scientific Rules and Equations

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

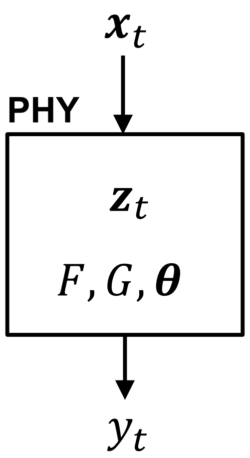
$$\frac{\partial \rho\mathbf{u}}{\partial t} = -\nabla.\left(\frac{1}{\rho}(\rho\mathbf{u})\otimes(\rho\mathbf{u}) + p\mathbf{I}\right) + \rho\mathbf{g}$$

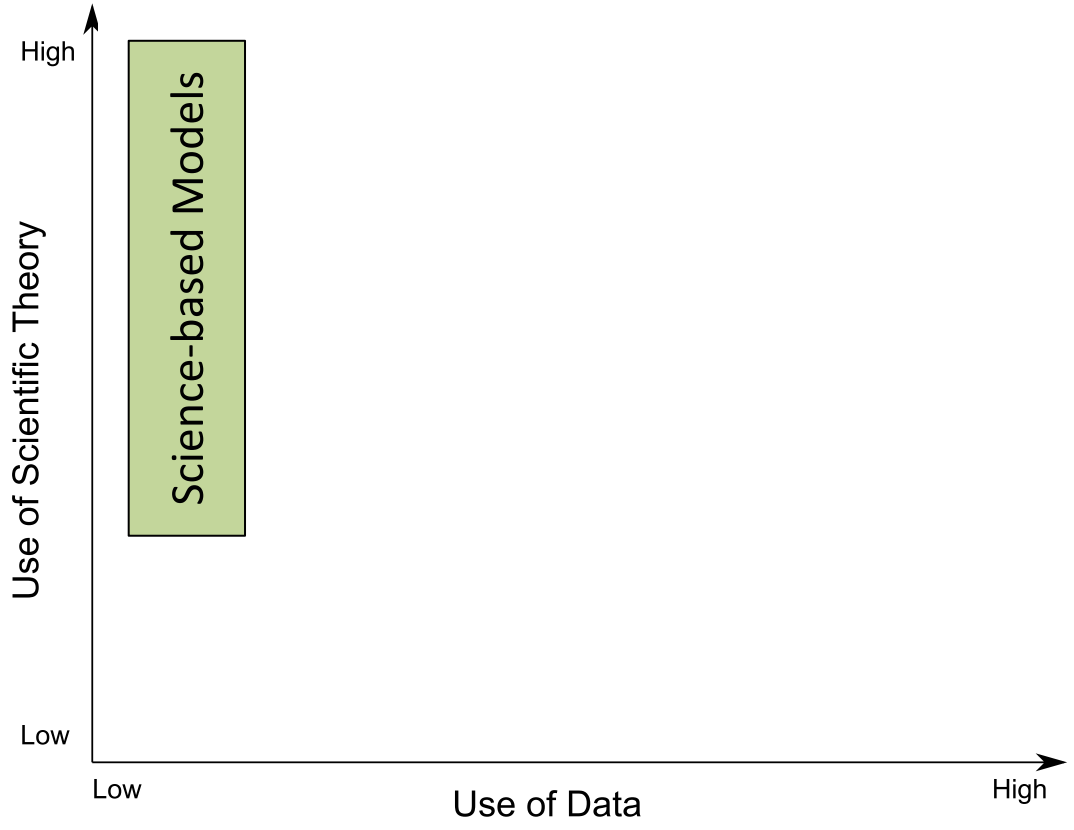
$$\frac{\partial E}{\partial t} = -\nabla.\left(\frac{1}{\rho}(E+p)(\rho\mathbf{u})\right) + \mathbf{u}.\rho\mathbf{g}$$

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

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

- Computational Models of Dynamical Systems





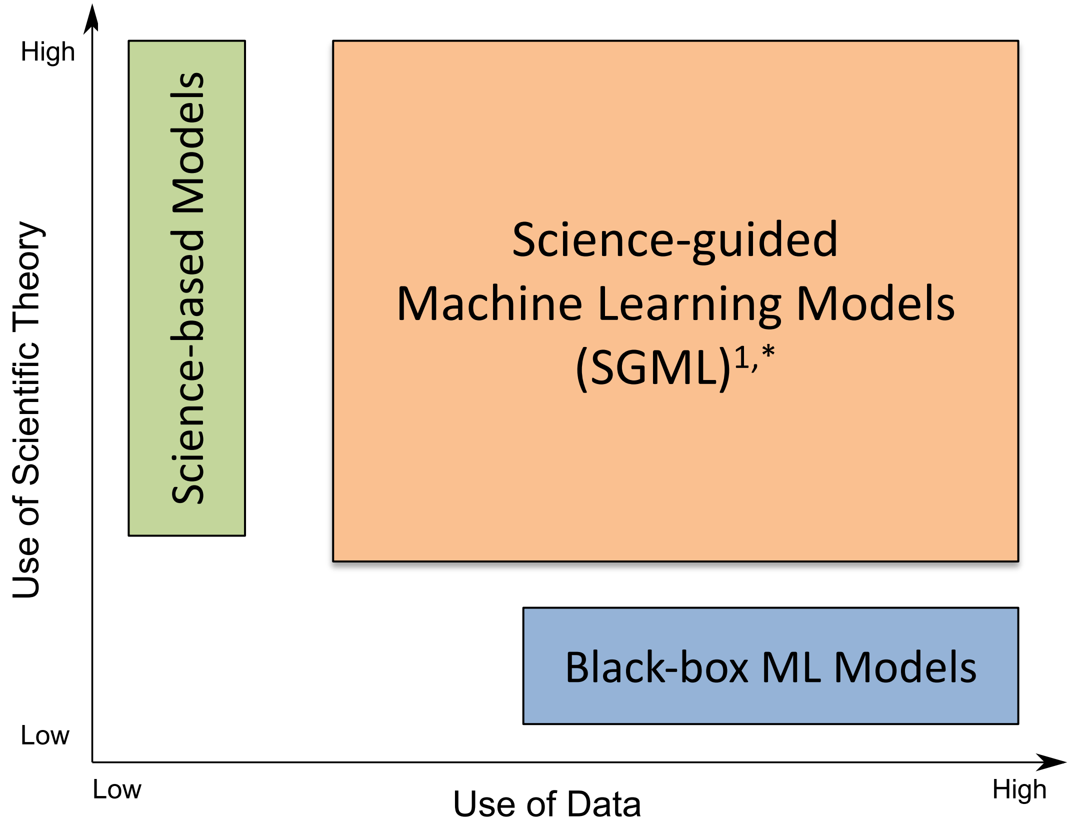
Limitations of Science-based Models

- Large number of parameters/states
- Incomplete or missing physics / process knowledge

Science-based vs. Data-based 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"



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

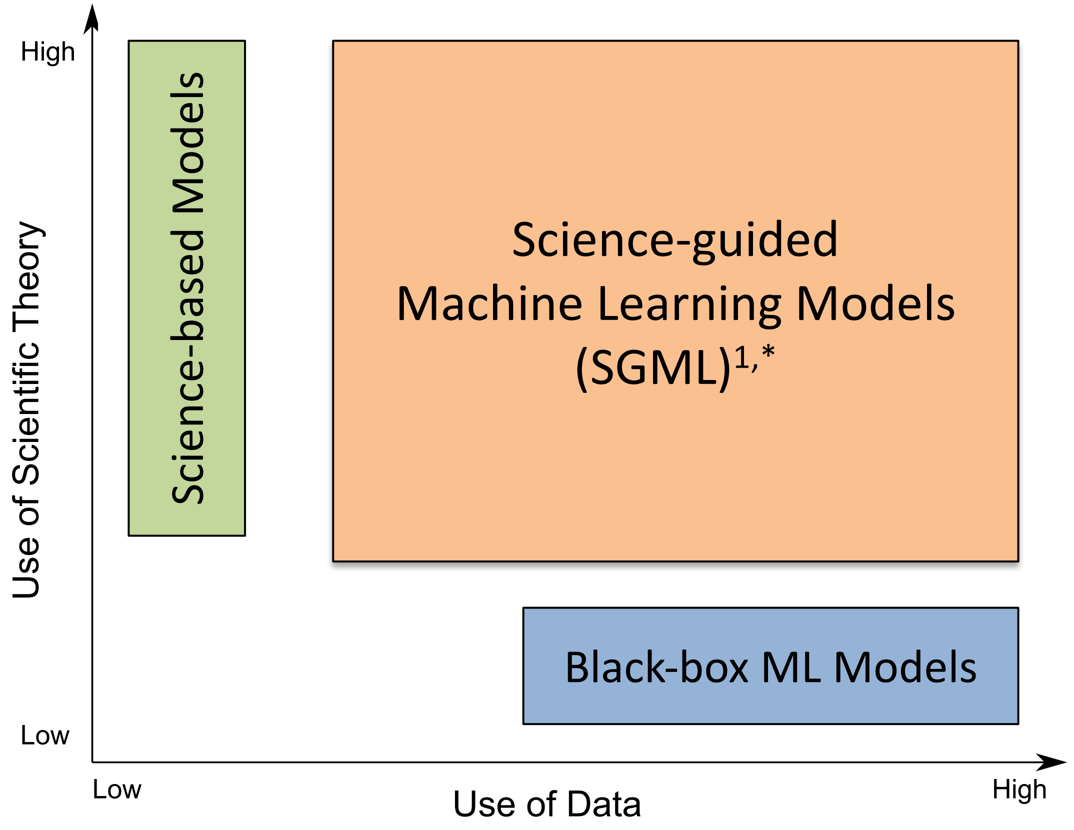
Require large number of representative samples

Science-based vs. Data-based Models

*Work on this topic has been referred to by various names such as:

- Knowledge-guided ML
- Science-guided ML
- Physics-guided ML
- Physics-informed ML / Physics-informed NN
- Physics-aware AI
- Theory-guided Data Science

In these works, “**physics**” or "**physics-guided**" should be more generally interpreted as “**science”** or “**scientific knowledge”.**



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

Require large number of representative samples

Recent Developments in SGML

Defense Advanced Research Projects Agency › Program Information

Physics of Artificial Intelligence (PAI)

DARPA

The Physics of Artificial Intelligence (PAI) program is part of a broad DAPRA initiative t
and adversarial spoofing, and that incorporate domain-relevant knowledge through gen

It is anticipated that AI will play an ever larger role in future Department of Defense (Do
processing, to control and coordination of composable systems. However, despite rapi
subfield of machine learning – AI's successful integration into numerous DoD applicati
development of causal, predictive models and dealing with incomplete, sparse, and noi

To facilitate better incorporation of AI into DoD systems, the PAI program is exploring n
physics, mathematics, and prior knowledge relevant to DoD application domains. PAI a
will help to overcome the challenges of sparse data and will facilitate the development

CCC Computing Community Consortium Catalyst

Catalyzing the computing research community and enabling the pursuit of innovative, high-impact research.

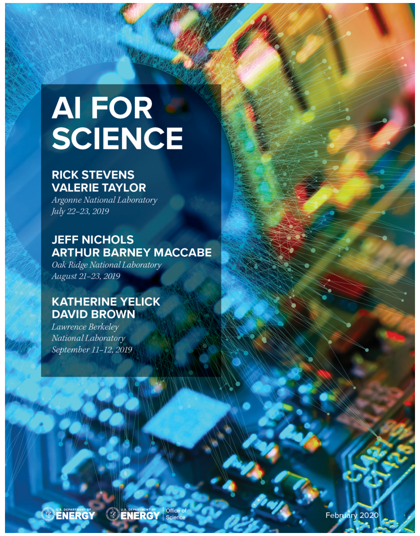
ABOUT | VISIONING | LEADERSHIP DEVELOPMENT | TASK FORCES | RESOURCES | EVENTS | BLOG | CCC BY

Visioning Activity

Artificial Intelligence Roadmap

In fall 2018, the Computing Community Consortium (CCC) initiated an effort to create a 20-Year Roadmap for Artificial Intelligence, led by Yolanda Gil (University of Southern California and President of AAAI) and Bart Selman (Cornell University and President-Elect of AAAI). The goal of the initiative was to identify challenges, opportunities, and pitfalls in the AI landscape, and to create a compelling report to inform future decisions, policies, and investments in this area.

The Roadmap was based on broad community input gathered via a number of forums and communication channels: three topical workshops during the fall and winter of 2018/2019, a Town Hall at the annual meeting of the AAAI, and feedback from other groups of stakeholders in industry, government, academia,



Report on DOE Town halls on "AI for Science"

Many conferences/workshops

- 2020 AAAI Fall Symposium on Physics-guided AI
- 2020 and 2021 AAAI Spring Symposium on ML in Physical Sciences

Physics-Informed Learning Machines for Multiscale and Multiphysics Problems

Pacific Northwest NATIONAL LABORATORY

Machine Learning for Physics and the Physics of Learning

ipam

PHYSICS INFORMED MACHINE LEARNING

Workshop by Los Alamos National Laboratory, 2016, 2018, 2020

Integrating Physics-Based Modeling With Machine Learning: A Survey

JARED WILLARD* and XIAOWEI JIA*, University of Minnesota
SHAOMING XU, University of Minnesota
MICHAEL STEINBACH, University of Minnesota
VIPIN KUMAR, University of Minnesota

Surveys more than 300 papers

https://arxiv.org/pdf/2003.04919.pdf

Guiding Principles of SGML

- **How can Science help ML?**

- **How can ML advance Science?**

Guiding Principles of SGML

- **How can Science help ML?**
  - Guide the learning of AI models to *scientifically consistent* solutions
  - Ensure *generalizability* even when training data is limited

Generalization Error $\propto$ Training Error + Complexity + **Scientific Inconsistency**

- **How can ML advance Science?**

Guiding Principles of SGML

- **How can Science help ML?**
  - Guide the learning of AI models to *scientifically consistent* solutions
  - Ensure *generalizability* even when training data is limited

$$\text{Generalization Error} \propto \text{Training Error} + \text{Complexity} + \textbf{Scientific Inconsistency}$$

- **How can ML advance Science?**
  - Discover new scientific laws from data
  - Augment or replace components of science-based models

Research Themes in TGDS

Science helps AI

ML advances Science

1. Science-guided Design
   - Choice of Response Function
   - Design of Model Architecture
   - …
2. Science-guided Learning
   - Using Loss Functions, Constraints, Priors, Training Labels
   - …
3. Science-guided Refinement
   - Post-processing
   - Pruning
   - …
4. Discovery of Scientific Laws from Data
   - Symbolic Regression, Autoencoders, …
5. Inferring Parameters in Science-based Models
   - Model Calibration, Inverse Modeling, Data Assimilation, …
6. Hybrid-Science-ML Modeling
   - Residual Modeling, Augmenting system components using ML, Pretraining, …

What to Expect from this Course?

- By the end of the course, students will…
  - Be well-versed with the **foundations and theme areas of SGML**, as well as recent developments in every theme area
  - Be able to compare and contrast different SGML research themes and identify their strengths, limitations, and opportunities for future research
  - Be equipped to cross-pollinate SGML ideas from one application domain to another
  - **Develop essential research skills** including reading, discussing, and critiquing research papers, identifying research gaps and brainstorming solutions, and communicating research ideas through technical writing and oral presentations
  - Gain practical experience in pursuing SGML research through a course project

Who Should Take this Course?

- No pre-requisites except <u>interest</u> and <u>ability</u> to learn and apply SGML topics
- Students familiar in ML:
  - Who are eager and willing to learn about scientific problems and pursue SGML research
  - Ready to cross-disciplinary boundaries and work on inter-disciplinary projects
- Students from scientific disciplines:
  - Who have little familiarity in ML but are eager to learn and apply SGML in an application area they are familiar with

Who Should Not Take this Course?

- Students looking for a course on “Introduction to Machine Learning”
  - There are alternate courses for this purpose, including CS 5824: Advanced Machine Learning, CS 5525: Data Analytics
- Students who want to explore “black-box” applications of ML on conventional benchmark data sets (e.g., ImageNet or UCI data sets)
- Students looking for a regular lecture-based course with homework assignments and exams
- Students not looking forward to reading papers, writing reviews, and doing collaborative research

Let Us Get to Know Each Other!

- Quick Round of Introductions:
  - Name
  - Department
  - Program (BS-MS / MS / PhD / …)
  - What brings you to this course?

Next Steps and What is Coming Up Next

- Background Survey (Assignment 0) due next class:
  - https://virginiatech.qualtrics.com/jfe/form/SV_8x2NMN6FemRKh6t (also available on course webpage)
- Next Class:
  - Basic Introduction to ML
- Suggested Readings for Next Week:
  - A. Karpatne, G. Atluri, J. Faghmous, M. Steinbach, A. Banerjee, A. Ganguly, S. Shekhar, N. Samatova, and V. Kumar, "Theory-guided Data Science: A New Paradigm for Scientific Discovery from Data**,"** *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 29(10), 2318–2331, 2017.
  - Willard, Jared, Xiaowei Jia, Shaoming Xu, Michael Steinbach, and Vipin Kumar. "Integrating physics-based modeling with machine learning: A survey." arXiv preprint arXiv:2003.04919 (2020).
- Full Reading List to be posted on Canvas by Sep 31