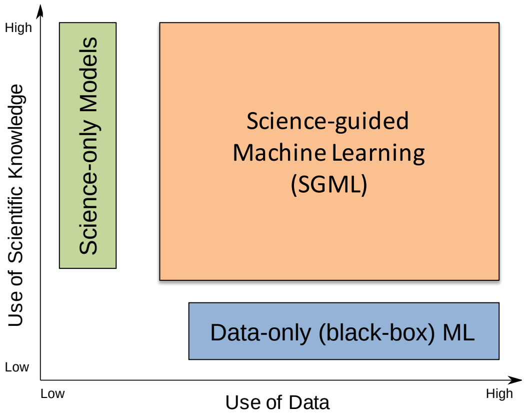
CS 6804: Science-guided Machine Learning (3 credits, CRN: 82313)

Department of Computer Science, Virginia Tech

Instructor: **Anuj Karpatne** (http://people.cs.vt.edu/karpatne)
Class Type: Online Course
Class Timings: MW: 4:00 pm - 5:15 pm Eastern; Zoom URL: https://virginiatech.zoom.us/j/99539879818
Instructor Office Hours: MW: 5:30 pm - 6:30 pm; Zoom URL: https://virginiatech.zoom.us/j/95075800579
Course Website: http://people.cs.vt.edu/karpatne/teaching/6804-f20/



Course Overview: While the impact of machine learning (ML) in commercial disciplines involving vision, speech, and text related problems is well understood, the promise of ML is yet to be fully realized for accelerating discoveries in scientific and engineering disciplines. This is because mainstream "black-box" ML models, that only rely on data, are susceptible to learning spurious relationships that do not generalize well outside the data they are trained upon. Moreover, black-box ML models do not provide any mechanistic insights about the scientific processes being studied, thus making them unfit to be used as building blocks in scientific discovery. What is fundamentally lacking in black-box ML is their inability to ingest the rich background of scientific knowledge driving real-world phenomena along with the information contained in data. To address this, there is a growing research trend to deeply integrate scientific knowledge in the ML process, referred to as the paradigm of Science-guided ML (SGML). This course will introduce the foundations of SGML and provide a coherent perspective of research themes in SGML. These research themes will be illustrated using recent examples of cutting-edge research from diverse scientific disciplines. The course will also impart hands-on experience in conducting SGML research through a semester-long project. All course activities will be conducted online.

Course Topics: This course will **tentatively** cover the following list of research themes in SGML:

1. **Science-guided Learning:** Techniques for modifying the learning algorithms of ML, e.g., with the help of priors, constraints, and loss functions to ensure that the learned ML solutions are scientifically consistent.
2. **Science-guided Design:** Techniques for hard-coding (or "baking in") scientific knowledge in the design of ML models, e.g., using neural network architectures that capture the physics of the scientific process.
3. **Science-guided Refinement:** Techniques for refining the outputs of ML models using scientific knowledge, e.g., by pruning or post-processing.
4. **Discovery of Scientific Laws from Data:** Techniques for automatically discovering the governing equations of a scientific problem from simulated or real-world data using ML.
5. **Inferring parameters in science-based models:** Techniques for inferring parameters or state-variables in science-based forward models using ML-based inversion techniques.
6. **Hybrid-science-ML Modeling:** Techniques for integrating ML models with science-based models to augment systematic biases or replace sub-components of science-based models that are currently lacking.

Background Required: This advanced-topics course does not require any formal pre-requisite courses and is broadly open to students with the interest and ability to learn topics in SGML. Specifically, this course is meant for two categories of graduate students: (a) students familiar in ML who are eager and willing to learn about scientific problems and pursue SGML research, and (b) students from scientific disciplines with little familiarity in ML who are eager to learn and apply SGML in an area they are familiar with. Students can assess their preparedness for the course by discussing with the instructor and attending the first class.

Learning Aims: 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

Learning Activities: This is a project-based course that will use a mix of learning activities. These activities are designed to provide an overview of the foundations and recent trends in SGML research, as well as to inculcate skills necessary to pursue research in SGML. We will specifically make use of the following learning activities:

- **Lectures:**
  We will have introductory lectures in the first few weeks by the instructor that will cover the basics of ML, foundations of SGML, and research themes in SGML. We will also have occasional guest lectures during the course of the semester by leading researchers in SGML covering special topics of interest.
- **Paper Discussions and Reviews:**
  Every student will get to lead a paper discussion from a reading list of relevant literature in SGML. Along with presenting the technical content of the paper, students will be encouraged to turn the paper discussion into an interactive event by posing questions to the class, presenting their perspective on the strengths and limitations of prior work and their applicability to other application domains, and identifying promising ideas for future research. Presentation files should be emailed to the instructor by 3 pm on the day of paper discussion, which will be uploaded on Canvas. Students will also submit peer-evaluations of the presentation and their individual reviews of the paper after every discussion session (*due before 1 pm on the day of the next class*). Sample paper reviews from the previous class will be discussed in the beginning of every class.
- **Course Project:**
  A major component of the course will be a semester-long project where students will get to work on a research problem in SGML of their interest at the intersection of ML and science from scratch to finish. Students are *highly encouraged* to choose a problem from an application domain they are most familiar with where they can leverage their unique perspective on the scientific background of the problem to be integrated with ML, although a list of sample projects will be provided by the instructor along with regular project pitches by students. Students will get to work in groups to identify and formulate a research problem, apply, explore, and design SGML algorithms to solve the problem, and demonstrate the real-world effectiveness of SGML procedures using rigorous evaluation setups, potentially leading to publications. Project deliverables include project proposals, midterm presentation, final presentation, and

final report. All project activities starting from idea generation to report preparation will be facilitated through online peer discussions coordinated by the instructor.

Workload and Grade Breakdown:

Paper presentation	15%
Project Proposal *A short report (max. 2 pages) summarizing the problem that will be studied, the goals and intended outcomes of the project, and research directions that will be explored.*	10%
Mid-term Project Review *In-class presentation of the problem being studied, progress made so far, challenges faced, and potential directions to explore next.*	15%
Final Project Presentation and Report *A final report (max. 6 pages) and in-class presentation of the methods and results.*	15 + 15 = 30%
Paper Reviews and Course Participation *Includes paper reviews and regular feedback on in-class participation and participation on course forum.*	20 + 10 = 30%

Tentative Outline of Course Activities (will be modified based on class strength and other factors):

Aug 24 – Sep 9	Introductory lectures on SGML
Sep 14 – Dec 9	Paper Discussions
Sep 14	Project Proposal Due
Oct 12 – Oct 19	Midterm Presentations
Dec 2 – Dec 9	Final Presentations
Dec 11	Project Report Due

Grading Scheme:

Grade	Aggregate Score Range
A	92 – 100
A-	87 – 91
B+	80 – 86
B	75 – 79
B-	70 – 74
C+	65 – 69
C	60 – 64
C-	55 – 59
D+	50 – 54
D	45 – 49
D-	40 – 44
F	< 40

Zoom Best Practices: We will be using Zoom for conducting all class activities and office hour discussions. The Zoom URLs are provided at the top of this document and can also be accessed from the Zoom tab on the left panel of the Canvas page of the class. Please familiarize yourself with Zoom and student tips for remote learning (see: https://tutorials.tlos.vt.edu/index/zoom.html and https://teaching.vt.edu/OurServices/StudentTips.html). You should keep your video turned on during the class to remain attentive and compensate for the lack of physical interactions in an online environment, unless restricted by low internet bandwidth. You may keep your audio muted unless you have a question to ask or need to respond to an on-going discussion to avoid interference and feedback. You can also type your question in Zoom's chat interface or provide nonverbal feedback and express opinions by clicking on icons in the Participants panel at any point during the lecture. All Zoom recordings of the class will be posted on Canvas within 6 hours after every class. We will be using Zoom waiting rooms during office hours to facilitate one-on-one or group interactions of the instructor with students. When you enter the waiting room, please fill in your name in the following Google Spreadsheet to keep track of your order of entry: https://tinyurl.com/cs6804-f20-office-hours.

Communications and Feedback: We will be using Canvas Announcements as our preferred mode of communication to notify any changes to the class schedule and activities, so please ensure that your Canvas Notification Preferences are set to notify you (typically via email) when an Announcement has been posted. Regular feedback will be provided to students on all submissions and class participation. At any time during the course, if you are facing any difficulties to meet the course deliverables or would like to discuss any concerns, you are welcome to talk to the instructor during office hours, over email, or using the following link for submitting anonymous feedback: https://tinyurl.com/cs6804-f20-feedback.

Academic Integrity: The tenets of the Virginia Tech's Honor Codes will be strictly enforced in this course, and all assignments shall be subject to the stipulations of the Undergraduate and Graduate Honor Codes. For more information on the Graduate Honor Code, please refer to the GHS Constitution at http://ghs.graduateschool.vt.edu. All paper reviews, project reports, and other submissions must represent your own individual effort. Students are encouraged to consult with one another about project design and evaluation issues, whether performed individually or in groups, as long as the individual submissions represent their individual efforts. Be particularly careful to avoid plagiarism, which essentially means using materials (ideas, code, designs, text, etc.) that you did not create without giving appropriate credit to the creator (using quotation marks, citations, comments in the code, link to URL, etc.). We will also adhere to Virginia Tech's Principles of Community for all in-class discussions and activities, to maintain a safe, welcoming, and respectful environment for every student in the class. For more information, see: https://www.inclusive.vt.edu/Initiatives/vtpoc0.html.

Accommodations for Students with Special Needs: Students with special needs will be provided additional resources and materials to aid in their learning. Mode of communication during the class will be adjusted in lieu of the respective needs of the student. Please discuss your requirements with the instructor so that we can work together to make a comfortable environment for everyone. Please see: https://www.ssd.vt.edu/ for more information. If you have an emergency medical information, please let me know privately as soon as possible.