DS-GA 3001.003: Learning from small data
Syllabus - Fall 2025
Please note that this syllabus is not final and there may be further adjustments.

Semester: Fall 2025
Lecture Location: 31 Washington Place (Silver Center) Room 411
Lecture Meeting Time: Tuesday, 12:30 pm - 2:30 pm (Starting Sept 2)
Lab Location: 60 Fifth Ave. Room 150 - Room 401
Lab Meeting Time: Tuesday, 4:55 pm - 5:45 pm (Starting Sept 2)
Communication Platform: Slack (request to be added if you aren't already after the first week)

Instructor(s): Ilia Sucholutsky
Email: is3060@nyu.edu
Office Hours Time: Wednesday, 12:30 pm - 1:30 pm
Office Hours Location: Zoom - https://nyu.zoom.us/j/92703344081

Section Leader: Ying Wang
Email: yw3076@nyu.edu
Office Hours Time: Thursday, 1:30-3:30 pm
Office Hours Location: 60 Fifth Ave. Room 763

Section Leader: Vivek Agarwal
Email: vka244@nyu.edu
Office Hours and Location: Monday, 12-2 pm
Office Hours Location: 60 Fifth Ave. Room 244

Section Leader: Jiayi Cheng
Email: jc10077@nyu.edu
Office Hours and Location: Friday, 11:30-1:30 pm
Office Hours Location: 60 Fifth Ave. Room 242

Prerequisites:
This will be a project-based course; some experience with either machine learning or cognitive science is recommended. Coding experience is highly recommended.

Suggested courses (while optional, having taken several of these or similar courses will help you get the most out of this course):

- DS-GA 1002 Probability and Statistics for Data Science
- CSCI-GA.2565 / DS-GA 1003 Machine Learning
- CSCI-GA.2569 / DS-GA 1005 Inference and Representation
- CSCI-GA 2572 / DS-GA 1008 Deep Learning
- PSYCH-GA 3405 / DS-GA 1016 Computational Cognitive Modeling
- DS-GA 1011 Natural Language Processing with Representation Learning

Course Description:
Deep learning has resulted in substantial progress on a variety of problems, but learning from small amounts of data remains a challenge. The current big data paradigm of training increasingly large models on increasingly large datasets is not feasible for many users working in settings with limited or expensive data (e.g., rare diseases, niche modalities, low-resource languages, etc.) or computational resources. By contrast, humans have an incredible ability to generalize and draw accurate inferences from very little data. In this course, we will aim to understand, harness, and improve that ability by drawing on insights from both cognitive science and computer science to explore how we can adapt machine learning systems to learn from less data. We will challenge the assumption that deep learning requires big data and show that we can develop methods for humans to teach AI systems in ways that better support efficient generalization and personalization. Throughout the course we will have a diverse set of guest speakers with expertise in very different research areas discuss the small data problems they face in their research.

Schedule of Classes (Weekly Schedule)
Note: the schedule below is only a rough guide, and the exact content will be determined as the course progresses. We will also aim to have some guest talks from other researchers .

- Tue, Sept 2: Introduction; Overview of course
- Tue, Sept 9: Probing the limits of learning with small data - Part 1 Dataset distillation
    - Additional resources (optional):
        - Few-shot, one-shot, and less-than-one-shot learning
            - One-shot learning of object categories
            - LibFewShot: A Comprehensive Library for Few-Shot Learning
            - 'Less Than One'-Shot Learning: Learning N Classes From M<N Samples
            - Can humans do less-than-one-shot learning?
            - Using Compositionality to Learn Many Categories from Few Examples
        - Coresets, prototypes, and dataset distillation
            - Prototype Selection for Nearest Neighbor Classification: Taxonomy and Empirical Study
            - Dataset distillation
- Tue, Sept 16: Probing the limits of learning with small data - Part 2 Less-than-one-shot learning
    - Additional resources (optional):
        - High dimensional, tabular deep learning with an auxiliary knowledge graph
- Tue, Sept 23: Probing hidden representations; Representational alignment
    - Additional resources (optional):
        - Getting aligned on representational alignment
- Tue, Sept 30: Guest lecture (Computation & Cognition) - Prof. Todd Gureckis (NYU)
- Tue, Oct 7: Guest lecture (Decision-making & Cognition) - Prof. Mark Ho (NYU)
- Tue, Oct 14: NO LECTURE (Monday schedule)
- Tue, Oct 21: Guest lecture (Human-in-the-loop algorithms) - Prof. Nori Jacoby (Cornell)

- Tue, Oct 28: Guest lecture (Robotics) - Prof. Andreea Bobu (MIT)
- Tue, Nov 4: Guest lecture (Language & Compositionality): Prof. Bonan Zhao (Edinburgh)
- Tue, Nov 11: Guest lecture: Prof. Weiji Ma (NYU)
- Tue, Nov 18: Project presentations
- Tue, Nov 25: Project presentations
- Tue, Dec 2: Project presentations
- Tue, Dec 9: CANCELLED

Schedule of Labs (Weekly Schedule)
Labs will be led by the section leaders. These will primarily be used as dedicated time for you to meet with your groupmates, check in with the section leaders, ask questions, get help, etc. Labs will be mandatory to make sure that all group members are contributing (see grading details below).

- Tue, Sept 2: Introduction; brainstorming
- Tue, Sept 9: Brainstorming
- Tue, Sept 16: Pitch project ideas, split into groups
- Tue, Sept 23: Work on proposal
- Tue, Sept 30: Setting up compute
- Tue, Oct 7: Work on project
- Tue, Oct 14: NO LAB, Monday schedule
- Tue, Oct 21: Work on project
- Tue, Oct 28: Work on project
- Tue, Nov 4: Work on project
- Tue, Nov 11: Work on project
- Tue, Nov 18: Work on project
- Tue, Nov 25: Work on project
- Tue, Dec 2: Work on project
- Tue, Dec 9: CANCELLED

Course Project
Projects will be done in groups of 3-4 students (PhD students can work in smaller groups though you are encouraged to work on a new project or direction; otherwise permission required to work individually or in larger groups). Project topics are flexible including applications to interesting small datasets, extending an existing method, developing a novel algorithm, studying data-efficiency of humans or machines (e.g., LLMs) in interesting domains, and reproducing and reviewing a recent result. For interested groups, we can explore extending projects that have promising novel contributions into conference papers (author list will be discussed on a case-by-case basis but may include all members of your group, one or more section leaders who mentored your project, the instructor, and any other collaborators who end up contributing to the project). There may also be some other research or internship opportunities available after the end of the course; I'll try to connect you with researchers, guest lecturers, or industry folks who might have shared interests with you. Here are some papers we've previously written that might provide some ideas for your project: [1] [2] [3] [4] [5] [6] [7] [8] [9] [10]

Use of AI tools
The course project aims to give you the opportunity to conduct independent research. As a result, we will apply the same standards as for scholarly publications. While these standards may differ between different fields and different journals/conferences, we will adopt the (fairly relaxed) standards from NeurIPS 2024 (copied from https://neurips.cc/Conferences/2024/CallForPapers):

"*Use of Large Language Models (LLMs)*: We welcome authors to use any tool that is suitable for preparing high-quality papers and research. However, we ask authors to keep in mind two important criteria. First, we expect papers to fully describe their methodology, and any tool that is important to that methodology, including the use of LLMs, should be described also. For example, authors should mention tools (including LLMs) that were used for data processing or filtering, visualization, facilitating or running experiments, and proving theorems. It may also be advisable to describe the use of LLMs in implementing the method (if this corresponds to an important, original, or non-standard component of the approach). Second, authors are responsible for the entire content of the paper, including all text and figures, so while authors are welcome to use any tool they wish for writing the paper, they must ensure that all text is correct and original."

You must acknowledge any AI-generated material that informed your work (this includes in-text citations and/or use of quotations, and in your reference list). Using a generative AI tool to generate content without proper attribution qualifies as academic dishonesty. To indicate the use of a generative AI resource, you should include the following statement in your project: "*The author(s) acknowledge the usage of [generative AI tool Name], a [type of model] model developed by [generative AI tool Provider], in the preparation of this assignment. The [generative AI tool Name] was employed in the following manner(s) within this assignment [e.g., brainstorming, grammatical correction, specific section of the assignment].*"

Some suggested tools:
- Cursor - AI code editor; I like it better than copilot
- Github Copilot - Code completion tool
- Perplexity - AI-powered search engine; better than just asking ChatGPT or some other LLM, actually cites real sources, might have free Pro access via NYU email
- Grammarly - AI-powered grammar and spell-checking tool
- JupyterLab - Interactive development environment; Not an AI tool, but a great environment for quickly iterating on research code
- Otter - AI notetaker for meetings

Grading
Proposal (**due 11:59 p.m. EST on Wed, Sept 24**): 10%
- Submit here: https://forms.gle/pZwWCQTKhkTh4co57
- No more than 2 pages + references.
- Should cover motivation/summary, description of how idea fits into related work, key contribution(s) you want to make (e.g., hypothesis you want to test, algorithm you propose to develop), and your plan for making and testing this contribution (i.e., experiments you will run, models/datasets/performance metrics you will use, etc.) including a timeline (which will help us determine feasibility)
- Some tips on reading research papers:
  - https://osf.io/preprints/psyarxiv/p37zj
- Grading - you will be graded on three criteria:
  - Completeness - are all the components listed above reflected in the proposal?
  - Clarity - does the story make sense? is it clear what the exact problem & plan are?

    - Feasibility - does the plan make sense? is it feasible to complete it this semester?

Final presentation (Nov 18 - Dec 9): 25%

- Presented in class in last 4 weeks, the project does not need to be finished by this point, you can just present your progress.
- Max 15 minutes (aim for 12 minutes)
- Should cover motivation/summary and how it's situated within related literature, key contributions you want to make, your progress so far (models, experiments, results, etc), what you plan to do next.
- Grading - you will be graded on the three criteria outlined below along with guiding questions explaining what graders will be looking for:
    - Completeness - are all the components listed above reflected in the presentation?
    - Clarity - does the story make sense? is it clear what the exact contributions (e.g. problem & solution or question & answer) are?
    - Correctness - Are all claims supported by evidence (either from existing literature or from own experiments)?

Final report (**due 11:59 pm Mon, Dec 15**): 40%

- Submit here: https://forms.gle/qJrbzk6Yd6sicppE9
- Compose this as you would a research paper. While the exact structure is up to you, a typical paper will have an abstract, introduction, related works, methods, experiments, results, discussion, and conclusion. Aim for about 6 pages + references. Here's a CogSci template and ICLR template you can use. Here are a few examples of our papers using this format: [1] [2] [3] [4]
- Some tips on writing ML research papers:
    - https://grigorisg9gr.github.io/machine%20learning/research%20paper/how-to-write-a-research-paper-in-machine-learning/
    - https://icml.cc/Conferences/2002/craft.html
    - https://www.approximatelycorrect.com/2018/01/29/heuristics-technical-scientific-writing-machine-learning-perspective/
    - https://www.alignmentforum.org/posts/eJGptPbbFPZGLpjsp/highly-opinionated-advice-on-how-to-write-ml-papers
- Grading - you will be graded on the three criteria outlined below along with guiding questions explaining what graders will be looking for:
    - Completeness - are all the components listed above reflected in the report?
    - Clarity - does the story make sense? is it clear what the exact contributions (e.g. problem & solution or question & answer) are?
    - Correctness - Are all claims supported by evidence (either from existing literature or from own experiments)?

Participation: 25%

- 1% per each lecture and lab attended (there are 28 in total so up to 3% bonus marks)

Student Responsibilities

Students are expected to read/view assigned material of the class for which they are scheduled, attend class, participate in class, complete assignments, complete projects, and ask for help early

if they are having trouble. Students are expected to be respectful of their fellow students and the instructor, including interactions outside of the lectures and labs, such as physical and virtual group meetings, Slack messages, etc. This class is a collaborative learning experience.

Instructor Responsibilities
The instructor is expected to read/view the material before the class for which they are scheduled, prepare and deliver high-quality introductions to the material, and provide comments on assignments and projects intended to help students develop their abilities to work on the course material.

Academic Honesty
NYU students and faculty are expected to maintain the highest standards of academic honesty. Students can find information on the core principles and standards in the University's policy on academic integrity, which is accessible at http://www.nyu.edu/about/policies-guidelines-compliance/policies-and-guidelines/academic-integrity-html.

Student Accessibility
Academic accommodations are available to any student with a chronic, psychological, visual, mobility, learning disability, or who is deaf or hard of hearing. Students should please register with the Moses Center for Student Accessibility at 212-998-4980.

The Moses Center for Student Accessibility
726 Broadway, 2nd Floor
New York, NY 10003-6675
Telephone: 212-998-4980
Email: mosescsa@nyu.edu
Website:
https://www.nyu.edu/students/communities-and-groups/student-accessibility.html