EEP 509 Privacy-Preserving Machine Learning

Spring 2026

Instructor: Tamara Bonaci (t.bonaci@)

Office hours:

- Tuesdays after class
- By appointment

TA: Arnav Das (arnavmd2@uw.edu)

Office hours:

- TBD

Course website: https://canvas.uw.edu/courses/1883704

Course assignment submission: https://canvas.uw.edu/courses/1883704/assignments

Course gradebook: https://canvas.uw.edu/courses/1883704/gradebook

Course Piazza: https://piazza.com/class/mngjj08o2mwb5/

Course Overview:

Driven by recent progress in data science, machine learning and AI, our data is being collected at an ever-increasing pace. Collectors, aggregators, and processors of such data are various private for-profit and non-profit entities, as well as public organizations. While machine learning models, trained on our data, can be beneficial to us personally, as well as to societies at large, they can also lead to a slew of undesirable, negative, and occasionally catastrophic privacy incidents. We therefore must balance out two conflicting objectives: the need to maximize accuracy, utility, and efficiency of the used machine learning models, while at the same time protecting the privacy of the data used for and by those models, as well as the privacy of the models themselves.

Privacy Preserving Machine Learning (PPML) is an important and very active research area, that focuses on this question exactly – how to benefit from machine learning techniques while preserving the privacy of training data and learned models.

This course provides a rigorous, technically grounded treatment of the theory, mechanisms, and practice of privacy-preserving machine learning (PPML), updated to reflect the rapid rise of foundation models and generative AI.

Through this course, we will fluency with the central formal frameworks — differential privacy, federated learning, and cryptographic approaches — and will apply them to a full range of ML settings: traditional supervised learning, deep neural networks, large language models, and diffusion-based generative models. The course balances theoretical depth with systems-level thinking, and addresses the distinct privacy challenges posed by LLMs, including verbatim memorization, training data extraction, and the tension between generative capability and privacy.

Course Prerequisites:

This course assumes a basic familiarity with stochastic mathematics (notion of a random variable and a random process, expectation, variance, independence, Markov process).

Course Goals:

By taking this course, the students will:

- Gain familiarity with statistical and information-theoretic notion of privacy.
- Gain familiarity with privacy attacks against machine learning models.
- Gain familiarity with privacy engineering techniques, including secure multi-party computation (MPC), and differential privacy.
- Gain knowledge of privacy preserving approaches, including federated learning and split learning.
- Gain familiarity with federated learning in adversarial attacks.
- Gain knowledge of state-of-the-art Python libraries and tools used for PPML.

Course Outcomes:

Upon taking this course, students will be able to:

- Formally define and reason about privacy guarantees using differential privacy and related frameworks, across classical and deep learning settings.
- Apply DP-SGD, PATE, and related mechanisms to privately train models ranging from logistic regression to large transformers.
- Design and evaluate federated learning systems with appropriate privacy, efficiency, and fairness properties.
- Apply cryptographic techniques — MPC, secret sharing, and homomorphic encryption — to ML inference workloads.
- Identify and execute major classes of privacy attacks on ML models, including training data extraction from LLMs.
- Analyze the unique privacy risks of large language models and generative AI, including memorization, synthetic data risks, and copyright tensions.
- Critically read and present primary research literature in PPML.

Course Logistics

Lectures: lecture slides, relevant reading and additional material will be made available in the following way:

- Lecture notes and relevant reading material will be posted on Canvas, under the appropriate weekly module.
- Pre-recorded video material (if available for the current module) will be posted on Canvas, under the appropriate weekly module.

- Every **Tuesday from 6:30-9:50pm PT,** we will meet ***in person in Electrical Engineering Building, classroom 045, and/or on Zoom, using our recurring Zoom link.*** Our lectures will focus on the outline weekly topics. Additionally, we will use lecture time to discuss weekly readings. We will do our best to record lectures, and make it available through Canvas + Panopto.

Office hours: on Tuesdays after class, or by appointment using UW Zoom meeting.

Office hours will not be recorded, but may be attended by multiple students at the same time. If you would like to talk to me in private, please send me a note, and we can schedule a different time to meet.

Classroom recordings: This course, or parts of this course, may be recorded for educational purposes. These recordings will be made available only to students enrolled in the course, instructor of record, and any teaching assistants assigned to the course.

Course Progression:

The following is a preliminary class progression covering the 10 weeks of the course. It is subject to changes.

- **Week 1:**
  - **Lecture 1**: Course overview. Why Machine Learning Needs Privacy-Preserving Manner?
- **Week 2:**
  - **Lecture 2:** Crash Course to Machine Learning: traditional ML, LLMs and Generative AI.
- **Week 3:**
  - **Lecture 3:** Mathematical and cryptographic preliminaries (probability and information theory, L1/L2 sensitivity, noise distribution, basic cryptographic primitives, threat modeling).
- **Week 4:**
  - **Lecture 4:** Privacy Attacks on ML Models (membership inference, model inversion, attribute inference, gradient leakage, training data extraction).
- **Week 5:**
  - **Lecture 5:** Introduction to differential privacy.
- **Week 6:**
  - **Lecture 6:** Differential privacy – practice (DP-SGD, PATE framework, accounting and auditing).
- **Week 7:**
  - **Lecture 7:** Introduction to Decentralized Privacy-Preserving Machine Learning Algorithms. Federated Learning.
- **Week 8:**
  - **Lecture 8:** Cryptographic approaches to privacy-preserving machine learning. Introduction to homomorphic encryption and secure multi-party computation.
- **Week 9:**
  - **Lecture 9:** Privacy in LLMs and Generative AI.

- **Week 10:**
  - **Lecture 10:** Emerging topics: machine unlearning and the right to erasure. Fairness and privacy tensions, agentic AI.
- **Final exam week:**
  - Project presentations.

About the Course:

The course will consist of ***weekly readings, quizzes, labs and a project.***

Weekly readings:

Weekly readings of assigned papers are an important part of this course. The goal of this exercise is to work together on understanding broader implications of the foundational privacy-preserving machine learning research. Additionally, we will also try to keep up with the state-of-the-art PPML research.

Every week, a single research paper will be assigned, and you will be expected to post to the class discussion board about it. Your post should contain something original beyond what others have already posted. You may consider posting:

- A summary of the paper,
- Evaluation of the paper's strengths and weaknesses,
- Open research question related to the topic of the paper, or
- Question you would like to discuss in class.

All posts are due by **11:59pm PT on a Saturday,** and they will be graded on the scale of 0-2, where:

- 0 means missed or irrelevant post,
- 1 means a relevant post submitted, and
- 2 means a good and interesting post.

Post submitted after the deadline, and those generated by an AI tool, and blatantly reposted without an acknowledgement will receive no credit. There will be a total of 9 reading assignments, and we will take 7 best scores when determining your grade.

Quizzes:

Several times during the quarter, we will review the relevant course material using short online quizzes.

You will have a week to choose a convenient time, and work on several short questions at your own convenience. Once you start the quiz, you will have **an hour** to finish it. You will have several attempts at every quiz. **Additionally, we will drop your lowest quiz score.**

Expected quizzes topics:

- Introduction to machine learning

- Mathematical and cryptographic preliminaries
- ML and differential privacy
- Federated Learning
- Privacy attacks against ML, LLM and generative AI

Labs:

We will have up to five Python-based lab assignments, intended to give you experience with some simple privacy-preserving techniques.

The labs will utilize publicly available data sets. Additionally, we will us the some of the following tools:

- **Python (PyTorch / HuggingFace Transformers):** primary framework for our labs, including LLM fine-tuning
- **Opacus (Meta):** differentially private training for PyTorch; supports DP-SGD for both small models and transformer fine-tuning
- **TensorFlow Federated / Flower:** federated learning simulation and deployment
- **PySyft / OpenMined:** MPC and privacy-preserving computation
- **Microsoft SEAL / HElib:** homomorphic encryption libraries
- **IBM diffprivlib:** general-purpose differential privacy for classical ML (sklearn-compatible)
- **Memorization evaluation tools:** Carlini et al. extraction scripts and the 'privacy meter' library for auditing

Please note that no prior experience with any of these tools is expected. We will introduce the tools as we need them.

Overview of labs topics:

- Lab 1 :Machine learning with Python
- Lab 2: Mathematical and cryptographic primitives.
- Lab 3: Machine learning and differential privacy.
- Lab 4: Federated learning.
- Lab 5: Privacy attacks on ML, LLMs and generative AI.

Project:

The final component of this course is a project, and its goal is to give you a deeper understanding of how to think about, and how to solve a real-life problem related to privacy-preserving machine learning.

For the project, you will choose a topic related to privacy-preserving machine learning. You can work on the project either individually, or in groups of two. When working in a group, your end-result should reflect the fact that it is a multi-person effort.

Your work on the project will consists of several milestones:

- Project proposal – due in Week 4,
- Project update – due in Week 7

- Draft project submission and peer review - due in Weeks 9 and 10
- Project presentation – due in finals week
- Project report – due in finals week

Grading:

Your grade in this course will be based on readings and discussion, homework assignments, quizzes and project. The expected grade breakdown is:

- Quizzes - 10%
- Readings and discussion – 20%
- Labs: 25%
- Project – 40%
- Participation: 5%

Course Material:

There is no required textbook for this course, but some recommended books that you might want to consider include:

- Jin Li, Ping Li, Zheli Liu, Xiaofeng Chen, Tong Li – Privacy Preserving Machine Learning, Springer Briefs on Cyber Security Systems and Networks, Springer 2022
- Muhammad Habib ur Rehman, Mohamed Mdhat Gaber – Federated Learning Systems, Towards Next-Generation AI, Springer, 2021.

Assignment Turn-In and Late Submission Policy:

Collaboration: In this course, we want you to learn from each other. Therefore, you are allowed (and encouraged!) to talk to your classmates and other students about all course assignments. You may also consult outside reference materials, and the instructor. ***However, all material that you decide to turn in should reflect your own understanding of the subject matter at the time of writing.*** This means that you should write your own posts about the assigned papers, and code up your own lab assignments, and prepare your own course project. If you work with someone else on any assignment, please include their names on the material that you turn in.

Assignment Turn-in: Posts about the assigned papers should be submitted using the Canvas discussion board. All other material (lab and project assignments) should be submitted using course website on Canvas. Please, ***do not use*** email for assignment submissions.

Late Assignment Turn-in: All assignments are due **by 11:59pm PT on the assigned date,** but we understand that you may have to sometimes turn them in late. The grading penalty is 5% of the grade that you would otherwise receive for each day, or part of the day, that you are late. No submissions will be accepted after two weeks.

Checking grades: Grades will be posted to the course gradebook.