Lecture 1A: Syllabus & Logistics

ECE 696B

Trustworthy Machine Learning

Spring 2025

Instructor: Dr Ravi Tandon

Department of ECE

THE UNIVERSITY OF ARIZONA

Today’s Agenda

- Introduction
- Logistics, Syllabus, Course Overview
- Pre-requisites on Probability and Machine Learning
- Introduction to LLMs and Plan for the next 3 lectures.

ECE 696B: Logistics-I

- **Class Timings**: Tues, Thurs, 11:00-12:15 pm
- **Location**: Modern Languages, Room 504
- **Class Meeting Zoom Link**: https://arizona.zoom.us/j/87997122418
- **E-mail**: tandonr@arizona.edu (quickest way to reach me)
- **Office Hours:** By appointment
- **D2L:** we will be using D2L for lecture slides, videos, assignments and projects.

THE UNIVERSITY OF ARIZONA

ECE 696B: Books & Reference Material

- **Textbook:** None

- Reference books (also see the syllabus)

  Christopher Bishop, Pattern Recognition and Machine Learning, Springer, 2006

  Shai Shalev-Shwartz and Shai Ben-David, Understanding Machine Learning from Theory to Algorithms, Cambridge University Press, 2014.

  S. Bubeck, Convex Optimization: Algorithms and Complexity, NOW Publishers, 2015.

  T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, 2nd Edition, 2017

  I. Goodfellow, Y. Bengio and A. Courville, Deep Learning, Cambridge Uni. Press

- HWs, lecture notes + any other additional material will be posted on D2L. Please check D2L regularly.

THE UNIVERSITY OF ARIZONA

ECE 696B: Programming/Software

- PyTorch (preferred) or Tensorflow
- Proficiency in Python (encourage posting coding related questions on Discussion Forum)
- Ability to implement & validate ML algorithms on commonly encountered datasets
- **Pytorch Tutorial:**

https://www.youtube.com/playlist?list=PLqnslRFeH2UrcDBWF5mfPGpqQDSta6VK4

ECE 696B: Grading

- In-class Presentations (60%)- each student will present a total of 4-5 papers
- Final Project Presentation (15%)
- Final Project Report (15%)
- Attendance & Class Participation (10%)

THE UNIVERSITY OF ARIZONA®

Probability Pre-requisites

- Random variables (PDF, PMF, CDF)
- Basic distributions (Binomial, Normal, Laplace, Poisson, Chi-squared)
- Conditional Probability, Bayes' rule, Total Probability Theorem
- Independence
- Expectation, Variance, Moment Generating Functions (MGF)
- Markov, Chebychev's inequalities
- Weak Law of Large Numbers (WLLN)
- Central Limit Theorem (CLT)
- KL Divergence, Total-variation Distance, Cross-Entropy

Machine Learning Pre-requisites

- Exposure to basic concepts in Machine Learning
- Distinction between various "types" of learning
  - Supervised (Linear reg, Logistic regression, SVM, MLP, CNNs)
  - Unsupervised (KNN, Decision Trees, PCA)
- Exposure to optimization methods for training ML models (e.g., gradient descent, SGD, ADAM etc)
- Prior exposure/ability to implement & validate basic learning algorithms

Big Data & ML

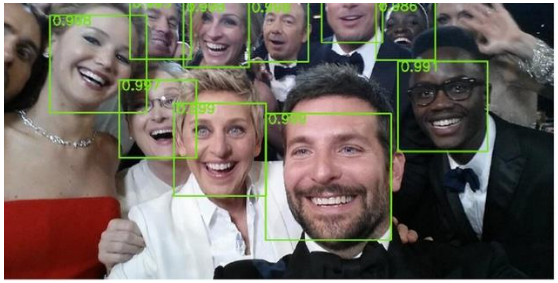
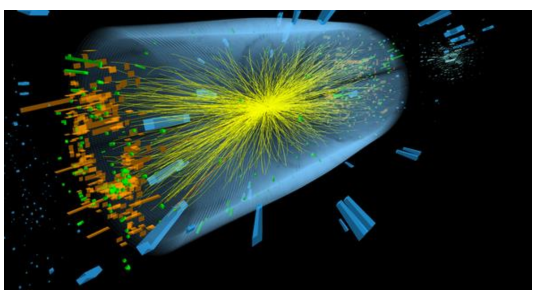


Image detection

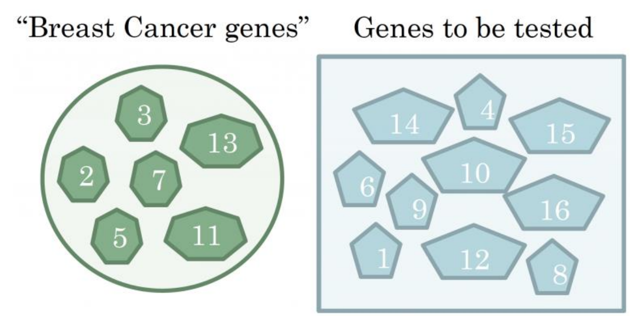
Medical Imaging

Behavioral Analysis



Scientific Applications

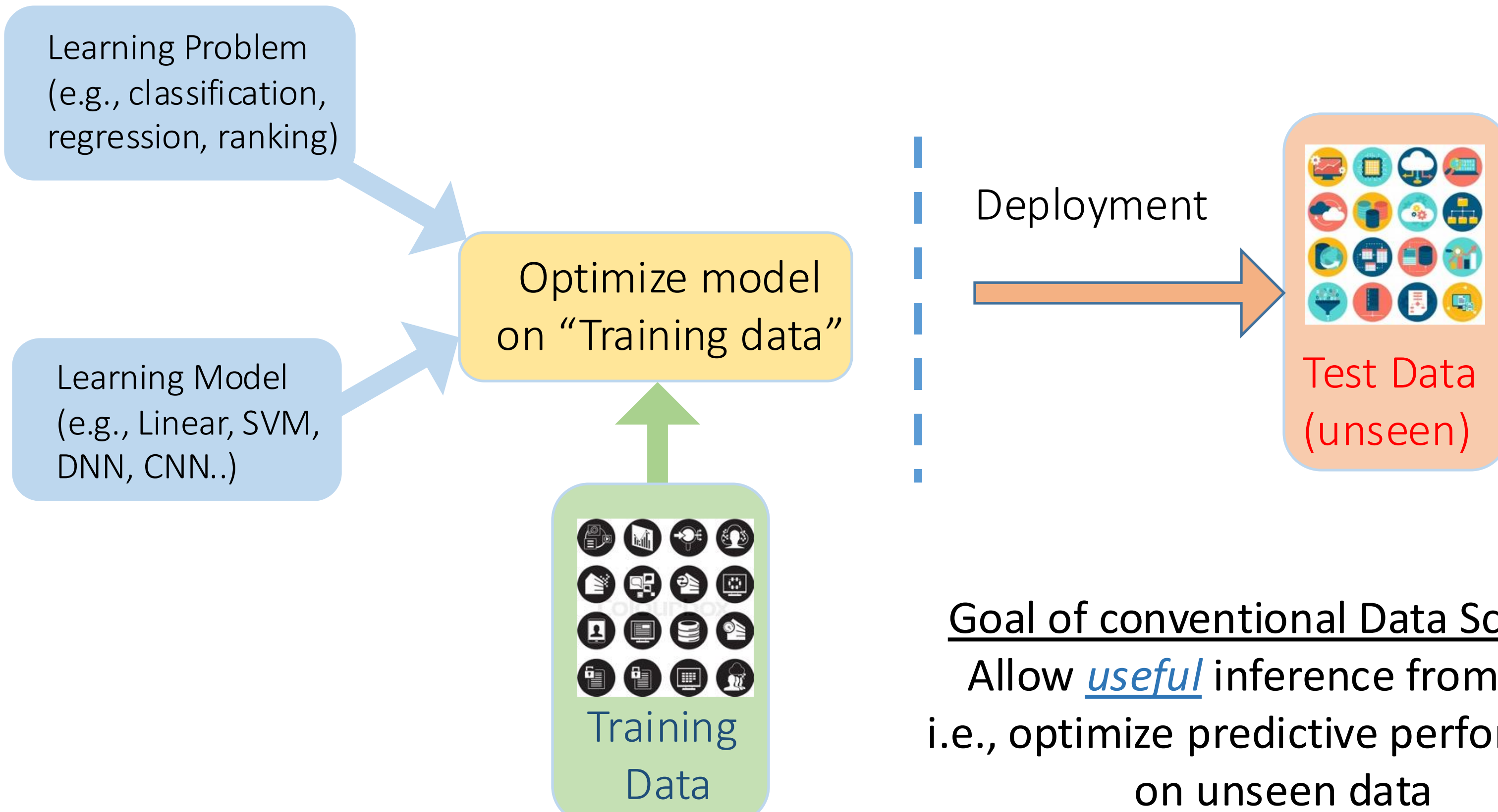
Cybersecurity



Bioinformatics

THE UNIVERSITY OF ARIZONA

The *Classical* Learning "Pipeline"



<u>Goal of conventional Data Science:</u>
Allow *useful* inference from Data
i.e., optimize predictive performance on unseen data

THE UNIVERSITY OF ARIZONA

What is this course about ?
Modern Trustworthy ML

Architectural & algorithmic features

- Transformers
- Multi-modality
- In-context
- Test-time Compute

"Modern" Deep Learning

Desirable Features & Constraints

- Privacy
- Robustness
- Fairness
- Explainability

ECE 696B— Tentative Topics & Timeline

**Module 1 (Introduction to LLMs) (3 weeks)

- Overview of Attention, Transformers
- Foundational papers on LLMs

Module 2 (Privacy Preserving Learning) (2 weeks)

- Basics of Differential Privacy
- Privacy Preserving ML

Module 3 (Robust & Adversarial ML) (2 weeks)

- Adversarial Machine Learning
- Techniques for Attacks & Defenses

Module 4 (Fairness in Machine Learning) (2 weeks)

- Notions of Fairness
- Algorithmic Techniques for learning fair classifiers

Module 4 (Student Project Presentations) (3 weeks)

THE UNIVERSITY OF ARIZONA

ECE 696B: Spring 2025
Trustworthy Machine Learning

Lecture 1B: *Introduction to Large Language Models (LLMs)*

Instructor: Dr Ravi Tandon

Department of ECE

Lecture Outline

- What are Large Language Models?
- Evolution of Natural Language Processing (NLP)
- Key Components of LLMs
- Applications of LLMs
- Challenges and Limitations
- Plan for the next few lectures

THE UNIVERSITY OF ARIZONA

What are LLMs ?

• Definition: Neural network models designed to understand, generate, and manipulate human language.

• Key Characteristics:

- Large number of parameters (e.g., GPT-3 with 175 billion parameters).
- Trained on massive datasets spanning diverse topics.
- Contextual understanding of text inputs.

• Examples:

- **GPT (OpenAI)**: Autoregressive model excelling in text generation.
- **BERT (Google)**: Bidirectional encoder for understanding sentence context.
- **T5 (Google)**: Unified text-to-text framework for multiple NLP tasks.
- **LLaMA (Meta)**: Lightweight language model for efficient inference.

THE UNIVERSITY OF ARIZONA

How LLMs work ?

Prompt
What is the capital of France?

→ ChatGPT →

Response
The capital of France is Paris.

How LLMs Work (key steps):

1. Input Text: Users provide input text (e.g., a prompt or query).

2. Tokenization: Text is split into smaller units (tokens) for processing.

3. Neural Network Processing: Tokens are fed into a Transformer-based architecture.

4. Output Generation: The model predicts the next tokens or generates the desired output.

Evolution of NLP Models & Architectures

- Pre-Neural Era:
  - Rule-Based Systems: Early systems relied on handcrafted rules for language processing.
  - Statistical Models: n-Grams and Hidden Markov Models (HMMs)
- Neural Networks Era:
  - Recurrent Neural Networks (RNNs):
    - Process sequences of data with memory of previous inputs.
    - Struggled with long-term dependencies due to vanishing gradients.
  - Long Short-Term Memory Networks (LSTMs):
    - Improvement over RNNs with gates to handle long-term dependencies.
    - Widely used for machine translation and text generation.
- Transformer Era:
  - Introduction of Transformers:
    - Key paper: Vaswani et al. (2017) "Attention Is All You Need".
    - Self-attention mechanism replaced recurrent structures, enabling parallel processing.
  - BERT and GPT Families etc.
- Key Innovations Across Eras:
  - Shift from handcrafted rules to data-driven models.
  - Increased computational power and dataset sizes.

Key Components of LLMs

- **Neural Networks:**
  1. Backbone of LLMs.
  2. Layers of interconnected nodes to process data.

- **Attention Mechanism:**
  1. Focuses on important parts of the input.
  2. Key innovation in the Transformer architecture.

- **Tokenization:**
  1. Splitting text into smaller units (tokens) for processing.
  2. Common methods: Byte Pair Encoding (BPE), WordPiece.

- **Positional Encoding:**
  1. Provides information about the order of tokens in the input sequence.
  2. Uses mathematical functions (e.g., sinusoidal) to encode position information.
  3. Essential for Transformers to capture sequence structure without recurrence.

Applications of LLMs

- **Text Generation:** creative writing and content generation.
- **Summarization:** Condensing information into concise formats.
- **Machine Translation**: Translating text between different languages accurately.
- **Question Answering:** Assisting in retrieving precise answers from large datasets.
- **Sentiment Analysis:** Understanding public opinion from text.
- **Reasoning and Planning:**
  - Solving complex problems by simulating logical reasoning.
  - Applications in step-by-step problem-solving (e.g., mathematics, programming).
  - Strategic decision-making in games or simulations.
- **Coding Assistance**: AI coding tools (e.g., GitHub Copilot) to enhance productivity.

and many more…

Challenges & Limitations

- Computational Costs:
  - High energy consumption during training & increasingly inference/deployment.
  - Large infrastructure requirements.
- Bias in Outputs:
  - Reflects biases in training data.
- Ethical Concerns:
  - Potential misuse (misinformation, deepfakes).
- Context Limitations:
  - Struggles with very long documents or nuanced reasoning.
- Privacy Issues:
  - Risks of exposing sensitive information during training or inference.
  - Challenges in ensuring secure handling of user data.
- Hallucinations:
  - Generation of false or fabricated information that appears plausible.
  - Issues in reliability for critical applications like healthcare or law.

Next 4 lectures: LLM Foundational Papers

1. "Attention paper"
   - Attention is all you need (2017) – *Presented by Ravi*

2. "GPT-1 paper"
   - Improving Language understanding by Generative Pre-Traning (2018) – *Presented by Ravi*

3. "BERT paper"
   - BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (2019) – *Presented by ?*

4. "GPT-2 paper"
   - Language Models are Unsupervised Multitask Learners (2019) – *Presented by ?*

Next few lectures: LLM Foundational Papers (1)

5. "Scaling Laws paper"
   - Scaling Laws for Neural Language Models (2020) – *Presented by ?*

6. "GPT-3 paper"
   - Language Models are Few Shot Learners (2020) – *Presented by ?*

7. "Instruction Tuning paper"
   - Fine Tuned Language Models are Zero Shot Learners (2022) – *Presented by ?*

8. "Codex paper"
   - Evaluating Large Language Models Trained on Code (2021) – *Presented by ?*

9. "Chain of Thought (COT) Paper"
   - Chain of thought prompting elicits reasoning in Large Language Models (2022) – *Presented by ?*

THE UNIVERSITY OF ARIZONA