🤌 Explainable Artificial Intelligence

Understanding the Foundations and Motivations

FAMAF - UNC

September 5, 2025

Section 1

Course Overview

Course Goals & Logistics

- **Course Goals**: Learn and improve upon state-of-the-art ML interpretability
- **Format**: Lectures, guest speakers, student presentations
- **Components**: Research project (60%), Paper presentations (30%), Participation (10%)
- **Emerging Field**: Opportunity to make real contributions

Course Staff

- **Instructor**: Hima Lakkaraju
- **TAs**: Jiaqi Ma, Suraj Srinivas
- **Office Hours**:
  - Hima: Monday 1:30-2:30pm
  - TAs: Thursday 1:00-2:00pm
- **Location**: Longeron Meeting Room, SEC 6th floor + Zoom
- **Webpage**: https://canvas.harvard.edu/courses/117650

Course Structure (14 Weeks)

1. **Week 1**: Introduction & overview
2. **Week 2**: Evaluating interpretability
3. **Weeks 3-4**: Learning inherently interpretable models
4. **Weeks 5-9**: Post-hoc explanations and vulnerabilities
5. **Weeks 10-11**: Theory + connections with robustness, fairness, DP
6. **Weeks 12-14**: Understanding LLMs and Foundation Models

Section 2

Why Model Understanding?

Machine Learning is Everywhere

- Healthcare diagnostics
- Criminal justice systems
- Financial lending decisions
- Autonomous vehicles
- Social media recommendations
- And many more high-stakes applications...

Use Case 1: Debugging

Model predicts "Siberian Husky" but relies on snow background

- **Problem**: Model using irrelevant features
- **Solution**: Understanding reveals the issue
- **Action**: Fix the model to focus on correct features

Use Case 2: Bias Detection

Criminal justice prediction system

- **Input**: Defendant details
- **Prediction**: “Risky to Release”
- **Issue**: Model using race and gender inappropriately
- **Insight**: “This prediction is biased!”

Use Case 3: Providing Recourse

Loan application denied

- **Explanation**: “Increase salary by \$50K + pay credit card bills on time for 3 months”
- **Result**: Individual has actionable steps to improve their situation
- **Benefit**: Provides path forward for applicants

Use Case 4: Trust Assessment

Medical diagnosis system

- **Finding**: Model uses irrelevant features for female patients
- **Decision**: "I should not trust predictions for that group"
- **Importance**: Knowing when NOT to trust the model

Use Case 5: Regulatory Approval

Model approval process

- **Authority concern**: "This model uses irrelevant features"
- **Decision**: "This cannot be approved!"
- **Requirement**: Models must be vetted before deployment

Section 3

Approaches to Model Understanding

Two Main Approaches

Take 1: Inherently Interpretable

- Linear regression
- Decision trees
- Rule-based models
- Built-in transparency

Take 2: Post-hoc Explanations

- LIME, SHAP
- Attention mechanisms
- Gradient-based methods
- External explanation tools

Accuracy vs. Interpretability Trade-offs

- **Sometimes** accuracy-interpretability trade-offs exist
- **Linear models**: High interpretability, potentially lower accuracy
- **Neural networks**: High accuracy, lower interpretability
- **Context matters**: Not all applications require the same balance

When to Use Each Approach

Decision Framework

If you can build an interpretable model that is adequately accurate for your setting: **DO IT!**
Otherwise, post-hoc explanations come to the rescue!

Additional considerations: - Limited data availability - Proprietary black-box systems - Legacy system constraints

Section 4

Defining Interpretability

What is Interpretability?

Definition: Ability to explain or present in understandable terms to a human

Challenges: - No clear consensus in psychology about explanations - What makes some explanations better than others? - When are explanations sought?

When Do We Need Interpretability?

Not always needed: - Ad servers - Postal code sorting - Well-validated systems with no serious consequence

Required when there is incompleteness in: - Problem formalization - Safety requirements - Ethical considerations

Incompleteness vs. Uncertainty

Incompleteness $\neq$ Uncertainty

- **Uncertainty**: Can be quantified (e.g., small dataset)
- **Incompleteness**: Abstract goals, unmeasurable criteria

Examples of incompleteness: - Scientific knowledge discovery - Safety (impossible to test all scenarios) - Ethics (abstract discrimination concepts)

Section 5

Evaluation Framework

Taxonomy of Interpretability Evaluation

Evaluation Type	**Humans**	**Tasks**
Application-grounded	Real Humans	Real Tasks
Human-grounded	Real Humans	Simple Tasks
Functionally-grounded	No Real Humans	Proxy Tasks

Important

Claim of the research should match the type of evaluation!

Application-grounded Evaluation

Characteristics: - Real humans (domain experts) - Real tasks or simplified versions - Most specific and costly - Gold standard for validation

Benefits: - Highest validity - Direct applicability

Challenges: - Expensive - Time-consuming - Limited subject pool

Human-grounded Evaluation

Characteristics: - Real humans (can be lay people) - Simplified tasks - Larger pool, less expensive

Typical experiments: - Pairwise comparisons - Model output simulation - Counterfactual reasoning tasks

Functionally-grounded Evaluation

When appropriate: - Model class already validated (e.g., decision trees) - Method not yet mature - Human experiments would be unethical

Proxy measures: - Model complexity - Number of rules/features - Computational metrics

Section 6

Taxonomies for Analysis

Application-based Taxonomy

Global vs. Local: - High-level patterns vs. specific decisions

Degree of Incompleteness: - What part is incomplete? - How incomplete is it?

Time Constraints: - How much time for understanding?

User Expertise: - Domain expert vs. lay user - Affects information processing capacity

Method-based Taxonomy

Basic Units of Explanation: - Raw features (pixel values) - Semantic features (objects) - Prototypes

Number of Units: - How many explanatory elements? - How do different types interact?

Compositionality: - Structured organization - Hierarchical relationships

Interactions: - Linear vs. non-linear combinations - Understandability of combinations

Section 7

Course Structure & Requirements

Course Components

Research Project (60%): - 3 checkpoints (10% each): Proposal, Baseline, Progress - Final Report (20%) - Final Presentation (10%) - Teams of 2-3 students

Paper Presentations (30%): - Teams of 2-3 students - Each team presents two papers

Class Participation (10%): - Active discussion participation - Regular attendance

Project Milestones

Proposal (10%): - 2-page project overview - Problem definition and motivation - Proposed solution approach - Success metrics

Baseline Implementation (10%): - Implement existing method - Reproduce published results - Critical analysis and improvement ideas

Project Milestones (cont.)

Midterm Progress (10%): - 2-3 page update - Formal problem statement - Detailed solution description - Preliminary results

Final Report (20%): - 5-6 page comprehensive writeup - Complete methodology - Thorough empirical evaluation - Findings and conclusions

Prerequisites

Required Background: - Linear algebra - Probability theory - Algorithms - Machine learning (CS181 or equivalent) - Python programming - NumPy, scikit-learn

Helpful Experience: - Statistics - Optimization theory

Section 8

Research Opportunities

Course Research Impact

Previous Success: - 11 research papers from previous course iterations - Publications at top venues: NeurIPS, ICML, AIES

Research Focus: - Not just surface-level applications - Goal: Push boundaries and make new contributions - Question existing work critically

Relevant Conferences

Core ML Venues: - ICML, NeurIPS, ICLR - UAI, AISTATS - KDD, AAAI

Interdisciplinary Venues: - FAccT (Fairness, Accountability, Transparency) - AIES (AI, Ethics, and Society) - CHI, CSCW, HCOMP (Human-Computer Interaction)

Section 9

Discussion Questions

Breakout Session Topics

Getting Acquainted: - Introduce yourselves - What topics excite you most in this course?

Philosophical Questions: - Are you convinced interpretability is important? - Can we really interpret/explain models correctly?

Technical Preferences: - Inherently interpretable models vs. post-hoc explanations? - Which approach do you favor and why?

Thank You!

Next Steps: - Review course materials on Canvas - Start thinking about research interests - Form initial project teams - Prepare for next week's readings

Questions? Contact the teaching team through Canvas or office hours.