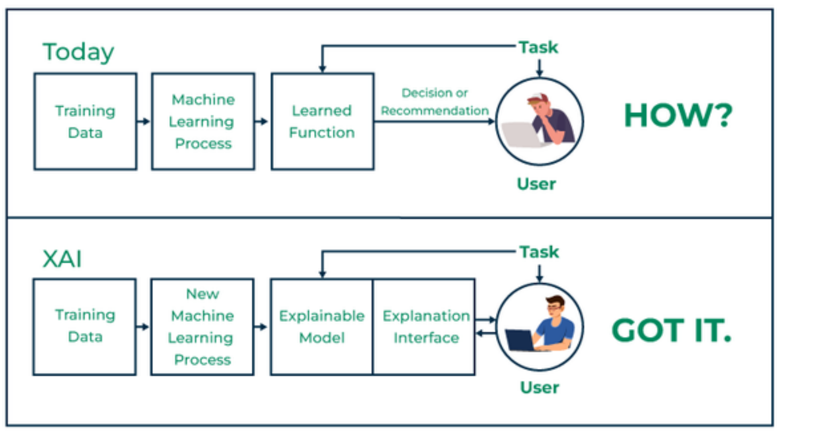
Explainability of AI models

Hanchen Zhou

SOFIA UNIVERSITY

Making Black-Box AI Decisions Transparent

- AI models make critical decisions, but user can't understand why
- Explanations brings trust and accountability for predictions



SOFIA UNIVERSITY

TL;DR

- Motivation
  - different XAI methods can disagree, so we measured their agreement.
- Main Idea
  - Built a Credit Explainability Comparison on LendingClub loan applications using SHAP (global/local), LIME (local), and counterfactuals; analyzed method agreement.
- Results
  - LIME–SHAP Top-3 feature agreement: 72.22%
  - Feature-importance correlation: 0.672
  - Method disagreements signal uncertainty to surface in decisions

SOFIA UNIVERSITY

What is AI Explainability

- Ability to understand and interpret AI model decisions in human terms
- Local: "Why did the model make THIS specific decision?"
- Global: "How does the model generally behave?"

Why Does It Matter?

- Trust
- Compliance
- Debugging
- Fairness

Foundation Methods

- LIME(2016): Explains individual predictions by learning local surrogate models
  - Intuitive but Unstable
- SHAP(2017): Uses game theory (Shapley values) to assign importance scores
  - Consistent but Expensive
- Counterfactual Explanations (2017): Find minimal changes needed to flip the decision
- Anchors(2018): "IF-THEN" rules that locally govern predictions

SOFIA UNIVERSITY

Explanation Enhancement

LLM-Based Explanations: a complementary approach that enhances traditional explanation methods

Traditional Methods	**LLM-Based Explanations**
Feature importance scores	Natural language narratives
Technical visualizations	Human-readable explanations
Expert interpretation needed	Accessible to all users

SOFIA UNIVERSITY

Demo

- Data: 10,000 Loan application
- Model: Random Forest Classifier
- **Explainability methods**: SHAP and LIME and Counterfactual

- Link: https://www.youtube.com/watch?v=jAFenYrTGrM&t=1s

SOFIA UNIVERSITY

Conclusion & Future Work

- **Explainability builds trust** in AI by showing *why* predictions are made.
- **Different methods disagree** (72% LIME–SHAP overlap), highlighting uncertainty that should be surfaced.
- **Future Work:** Combine fairness + explainability and use **LLMs** for human-centered narrative explanations.

SOFIA UNIVERSITY

Reference

- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *"Why Should I Trust You?": Explaining the Predictions of Any Classifier.* Retrieved from https://arxiv.org/pdf/1602.04938
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30.* Retrieved from https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf
- Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology, 31*(2), 841–887. Retrieved from https://arxiv.org/abs/1711.00399
- Molnar, C. (2022). *Interpretable machine learning: A guide for making black box models explainable* (2nd ed.). Retrieved from https://christophm.github.io/interpretable-ml-book/
- Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys, 51*(5), 93. Retrieved from https://arxiv.org/abs/1802.01933
- Microsoft Research. (2020). InterpretML: A unified framework for machine learning interpretability. Retrieved from https://github.com/interpretml/interpret

SOFIA UNIVERSITY