Trustworthy AI Systems

-- Explainability of AI

Instructor: Guangjing Wang

guangjingwang@usf.edu

Last Lecture

- Uncertainty and Robustness
- Source of Uncertainty
- Measure the Quality of Uncertainty
- Reduce Uncertainty and Enhance Robustness

This Lecture

- Motivation for Explainable AI
- Overview of Explainable AI Techniques
- Case Studies

Explanation - From a Business Perspective (1)

- Data is not necessarily as massive
- Human is usually behind to interpret and take final decisions
- Those humans need support and tools for understanding patterns, models, prediction, decisions

From the inside of a submarine, attempting to remove WW-II mines using signals such as sonar images.

Explanation - From a Business Perspective (2)

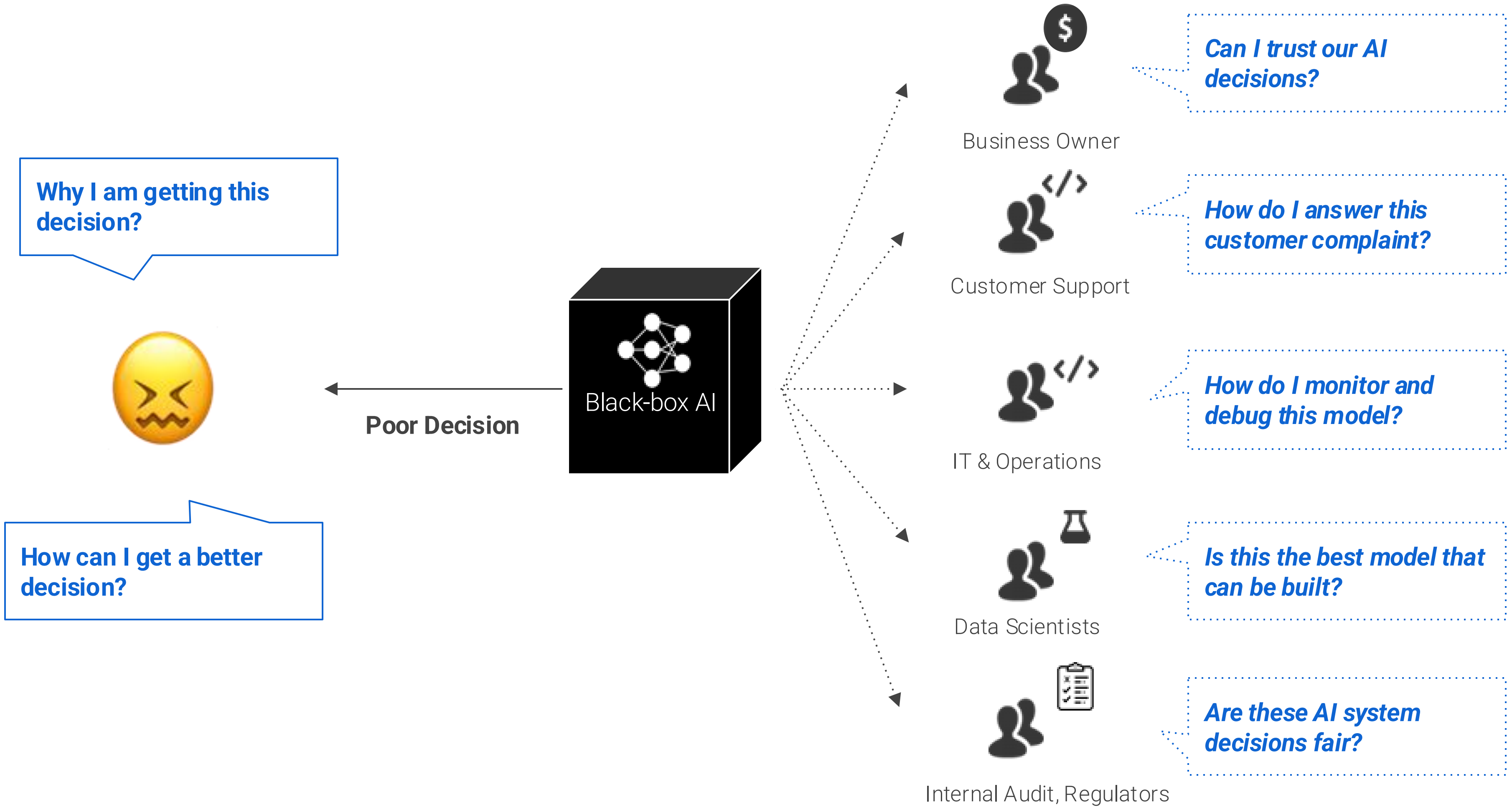
If something (bad) is happening, we need to trace back the cause and even have the explanation in real-time to limit any bad consequences

Explanation - From a Business Perspective (3)

Finance:

- Credit scoring, loan approval
- Insurance quotes
- FICO challenge in finance to understand why mortgage could be not approved - that is in front line with human, who ask for more transparency and understanding.

Black-box AI Creates Business Risk for Industry



Why Explainability: Debug (Mis-)Predictions

Top label: **"clog"**

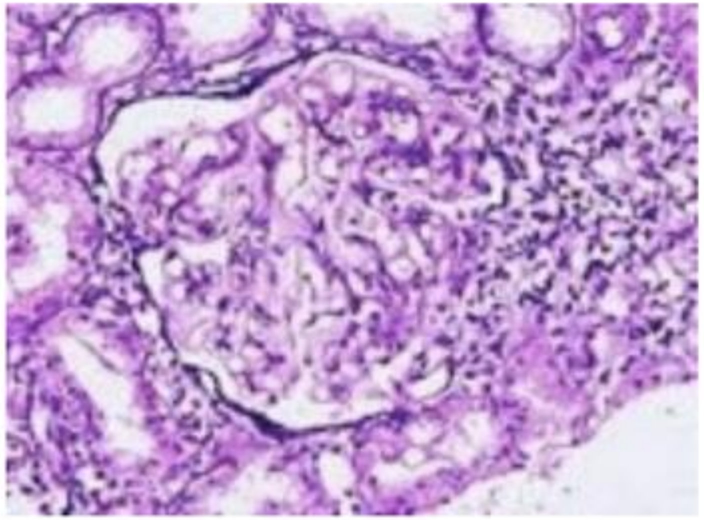
Why did the network label this image as **"clog"**?

Why Explainability: Verify the ML Model / System

Wrong decisions can be costly and dangerous

"Autonomous car crashes, because it wrongly recognizes …"

"AI medical diagnosis system misclassifies patient's disease …"



Why Explainability: Laws against Discrimination

Citizenship

Immigration Reform and Control Act

Sex

Equal Pay Act of 1963;
Civil Rights Act of 1964

Age

Age Discrimination in Employment Act of 1967

Race

Civil Rights Act of 1964

Disability status

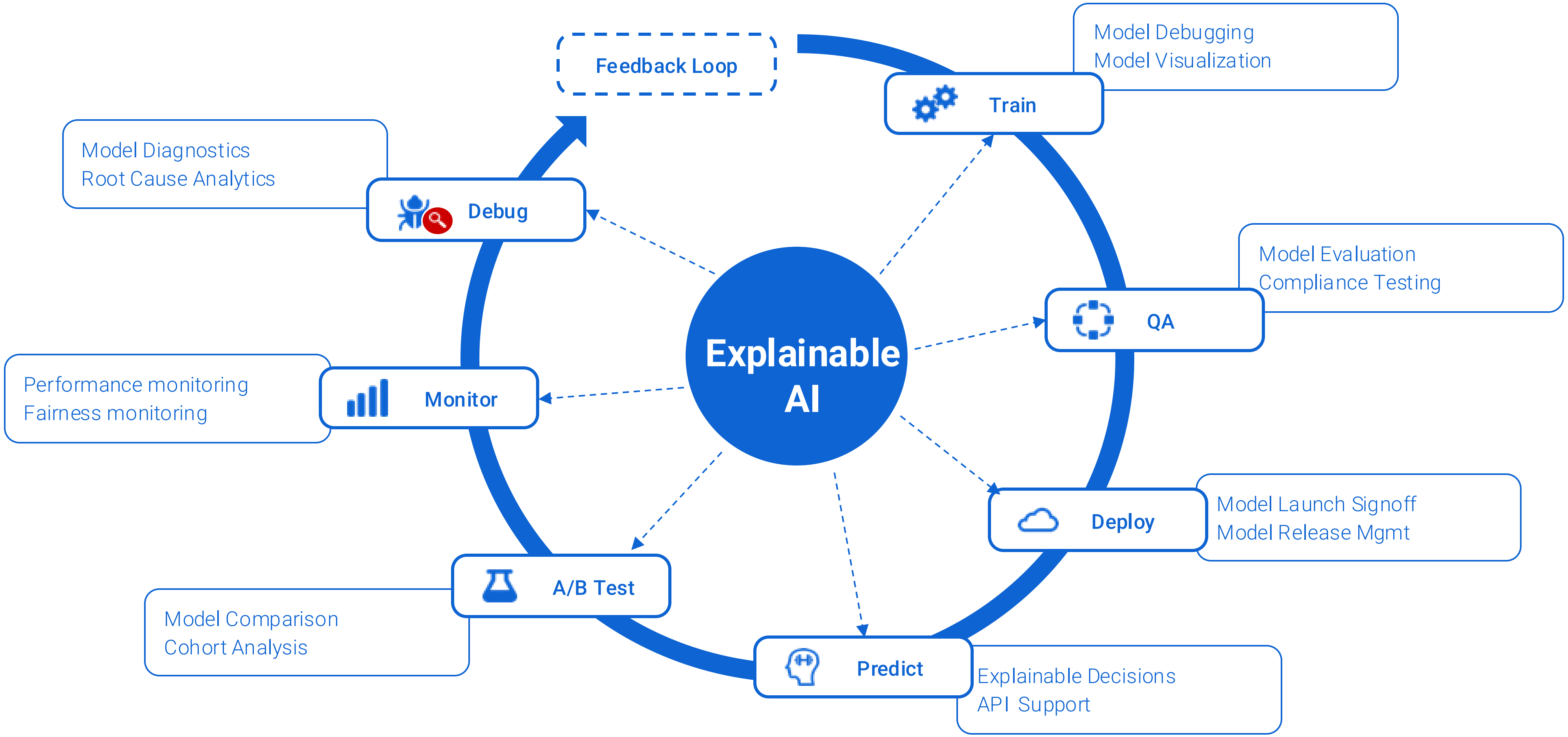
Rehabilitation Act of 1973;
Americans with Disabilities Act of 1990

And more...

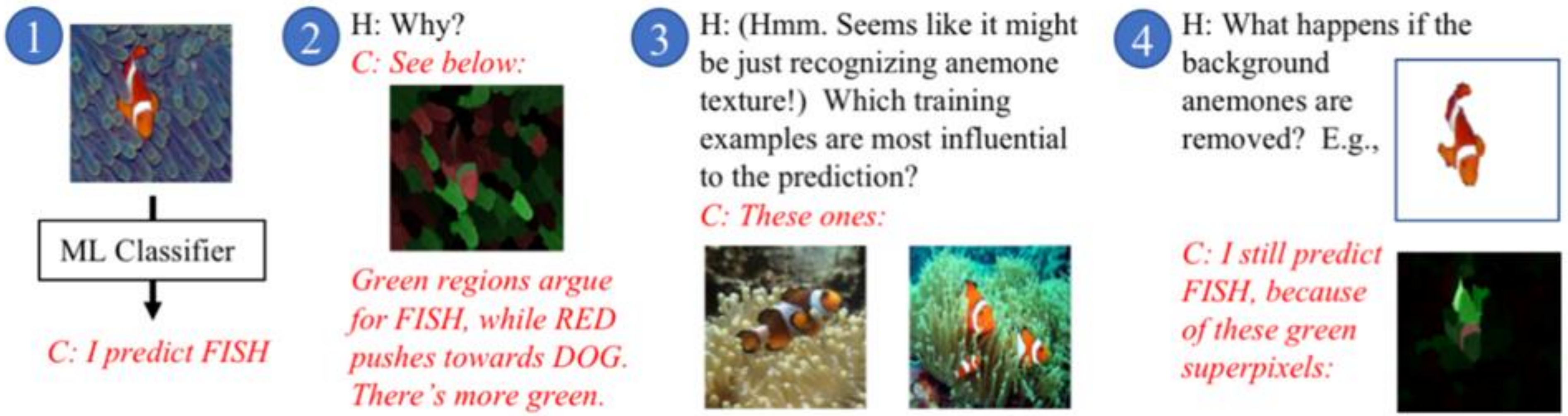
This Lecture

- Motivation for Explainable AI
- Overview of Explainable AI Techniques
- Case Studies

"Explainability by Design" for AI products



Example of an End-to-End XAI System



- Get a prediction
- Asking why and getting saliency map like explanations
- Keep iterating by asking more examples
- User is asking to remove / add some information to the results
- We could even imagine the user to add content, to add context, to ask for counterfactual...

Achieving Explainable AI

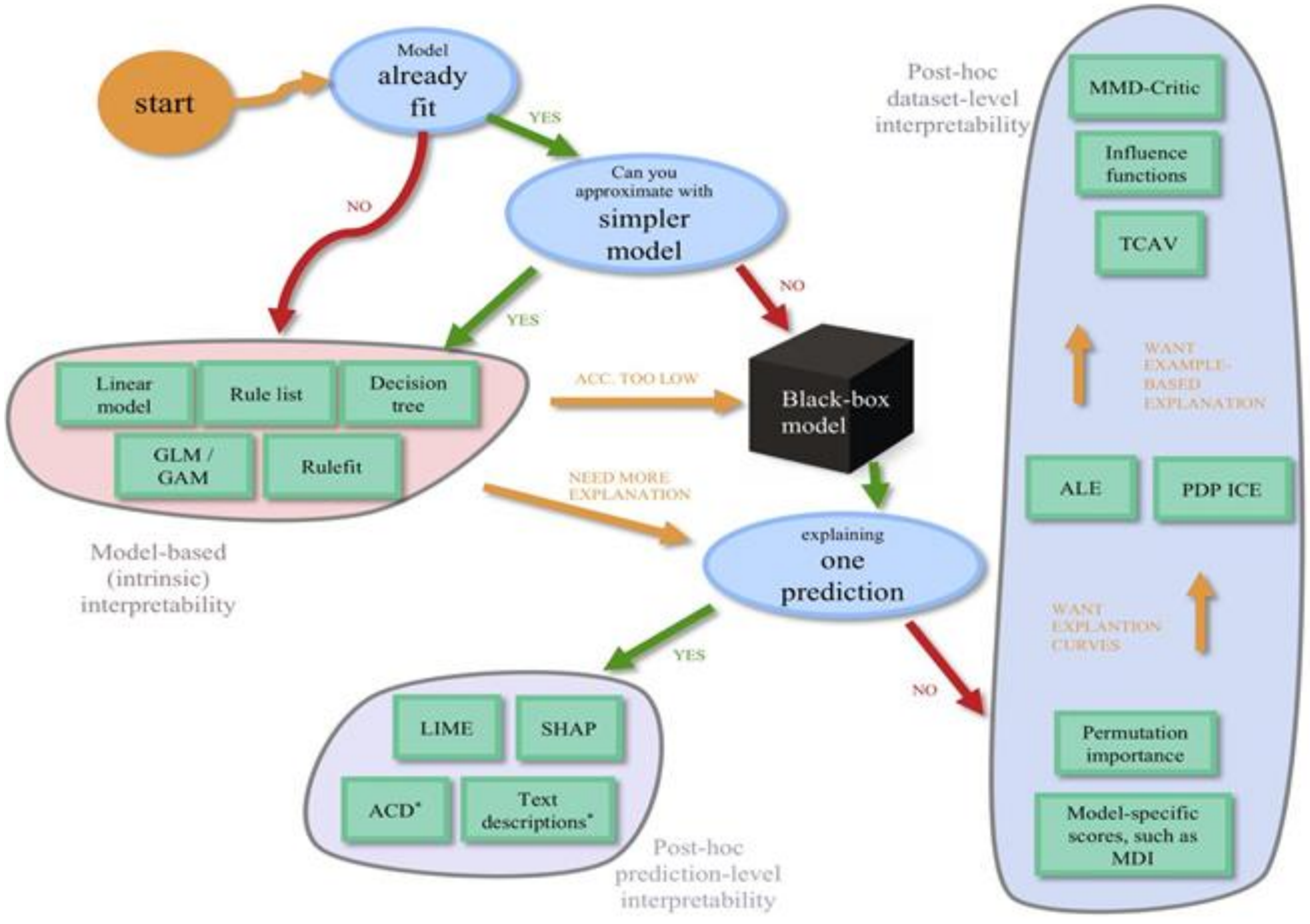
Approach 1: **Post-hoc explain a given AI model**

- **Individual prediction explanations** in terms of input features, influential examples, concepts, local decision rules
- **Global prediction explanations** in terms of entire model in terms of partial dependence plots, global feature importance, global decision rules

Approach 2: **Build an interpretable model**

- Logistic regression, Decision trees, Decision lists and sets, Generalized Additive Models (GAMs)

Achieving Explainable AI



* Denotes that a method only works on certain models (e.g. only neural networks)

interpretability cheat-sheet

View on github

Based on this interpretability review and the sklearn cheat-sheet. More in this book + these slides.

Summaries and links to code

RuleFit – automatically add features extracted from a small tree to a linear model

LIME – linearly approximate a model at a point

SHAP – find relative contributions of features to a prediction

ACD – hierarchical feature importances for a DNN prediction

Text – DNN generates text to explain a DNN's prediction (sometimes not faithful)

Permutation importance – permute a feature and see how it affects the model

ALE – perturb feature value of nearby points and see how outputs change

PDP ICE – vary feature value of all points and see how outputs change

TCAV – see if representations of certain points learned by DNNs are linearly separable

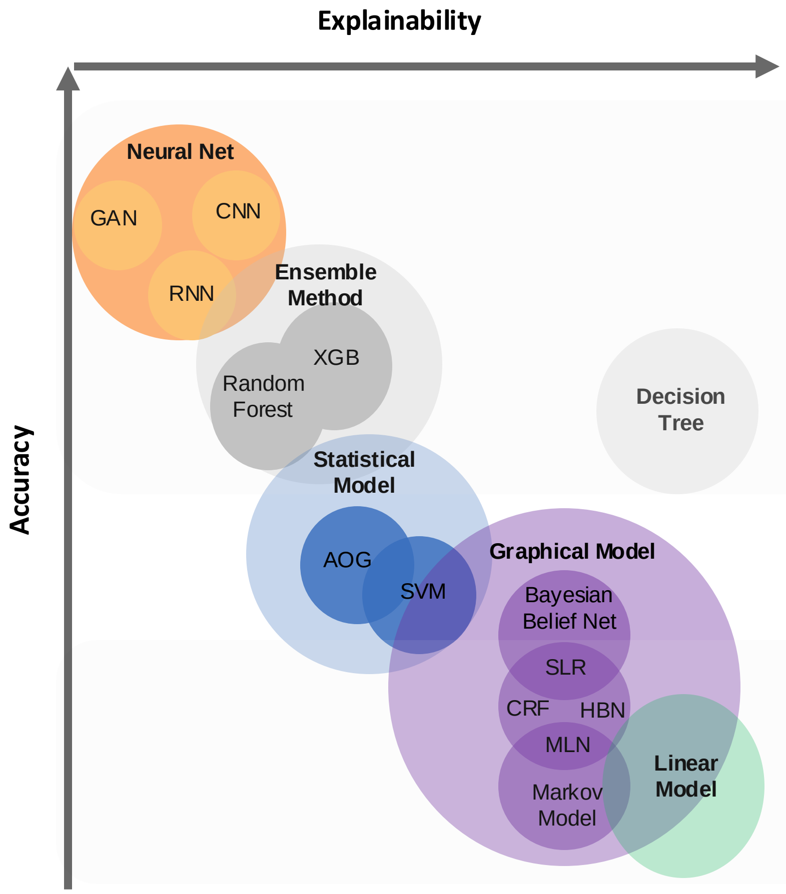
Influence functions – find points which highly influence a learned model

MMD-CRITIC – find a few points which summarize classes

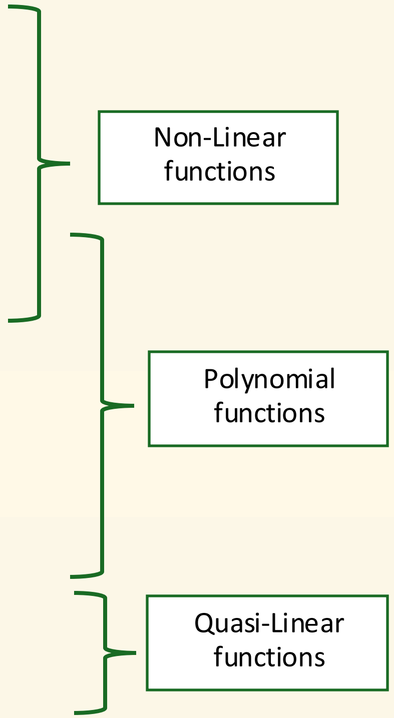
How to Explain? Accuracy vs. Explainability

Learning

- Challenges:
  - Supervised
  - Unsupervised learning
- Approach:
  - Representation Learning
  - Stochastic selection
- Output:
  - **Correlation**
  - **No causation**



Interpretability



This Lecture

- Motivation for Explainable AI
- Overview of Explainable AI Techniques
  - Individual Prediction Explanations
- Case Studies

Example: Individual Example

Top label: “**fireboat**”

Why did the network label this image as “**fireboat**”?

The Attribution Problem

Attribute a model’s prediction on an input to **features of the input**

Examples:

- Attribute an object recognition network’s prediction to its pixels
- Attribute a text sentiment network’s prediction to individual words
- Attribute a lending model’s prediction to its features

A reductive formulation of “why this prediction” but surprisingly useful

Attribution: Ablation-based Method

Drop each feature and attribute the change in prediction to that feature

Pros:

- Simple and intuitive to interpret

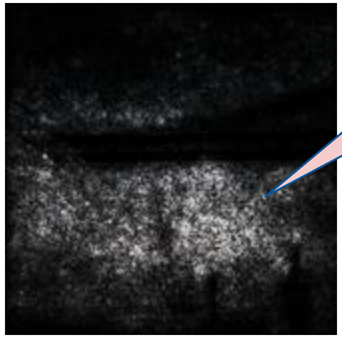
Cons:

- Unrealistic inputs
- Improper accounting of interactive features
- Can be computationally expensive

Attribution: Gradient-based method

Attribution to a feature is feature value times gradient, i.e., $x_i * \partial y/\partial x_i$

- Gradient captures sensitivity of output w.r.t. feature
- Equivalent to Feature*Coefficient for linear models
  - **First-order Taylor approximation** of non-linear models
- Popularized by SaliencyMaps [NIPS 2013], Baehrens et al. [JMLR 2010]



Gradients in the vicinity of the input seem like noise?

Attribution: Game Theory-based Method

Shapley Value: Classic result in game theory on distributing gain in a **coalition game**

- Coalition Games
  - Players collaborating to generate some **gain** (think: revenue)
  - Set function **v(S)** determining the gain for **any subset S** of players

- Shapley Values are a fair way to attribute the total gain to the players based on their contributions
  - Concept: **Marginal contribution** of a player to a subset of other players (v(S U {i}) - v(S))
  - Shapley value for a player is a **specific weighted aggregation of its marginal** over all possible subsets of other players

$$\textbf{Shapley Value for player } i = \sum_{S \subseteq N} w(S) * (v(S \cup \{i\}) - v(S))$$

(where $w(S) = N! / |S|! \, (N - |S| - 1)!$)

Shapley Value Justification

Shapley values are unique under four simple axioms

- **Dummy:** If a player never contributes to the game then it must receive zero attribution
- **Efficiency:** Attributions must add to the total gain
- **Symmetry:** Symmetric players must receive equal attribution
- **Linearity:** Attribution for the (weighted) sum of two games must be the same as the (weighted) sum of the attributions for each of the games

Shapley Values for Explaining ML models

- Define a coalition game for each model input X
  - **Players are the features in the input**
  - **Gain is the model prediction** (output), i.e., gain = F(X)
- Feature attributions are the Shapley values of this game

Challenge: Shapley values require the gain to be defined for all subsets of players

- What is the prediction when **some players (features) are absent**?

  i.e., what is **F(x_1, <absent>, x_3, ..., <absent>)**?

Modeling Feature Absence

Key Idea: Take the expected prediction when the (absent) feature is sampled from a certain distribution.

Different approaches choose different distributions

- [SHAP, NIPS 2018] Use conditional distribution w.r.t. the present features
- [QII, S&P 2016] Use marginal distribution
- [Strumbelj et al., JMLR 2009] Use uniform distribution

Computing Shapley Values

Exact Shapley value computation is exponential in the number of features

- Shapley values can be expressed as an expectation of marginals

$$\phi(\mathbf{i}) = \mathbf{E}_{S \sim \mathcal{D}}\,[\mathbf{marginal}(S, i)]$$

  - Sampling-based methods can be used to approximate the expectation
  - See: "Computational Aspects of Cooperative Game Theory", Chalkiadakis et al. 2011

- The method is still computationally infeasible for models with hundreds of features, e.g., image models

Attributions don’t explain everything

Some things that are missing:

- Feature interactions (ignored or averaged out)
- What training examples influenced the prediction (training agnostic)
- Global properties of the model (prediction-specific)

An instance where attributions are useless:

- A model that predicts TRUE when there are **even number** of black pixels and FALSE otherwise

Local Interpretable Model-agnostic Explanations

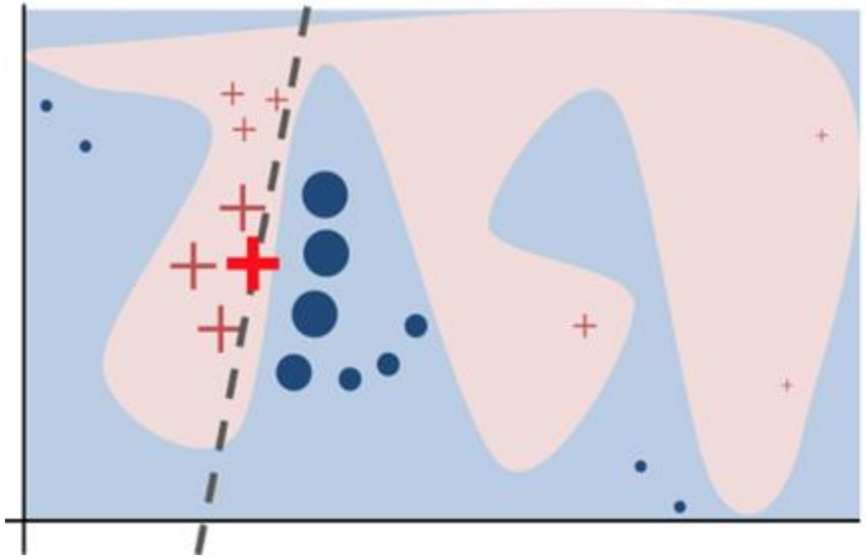
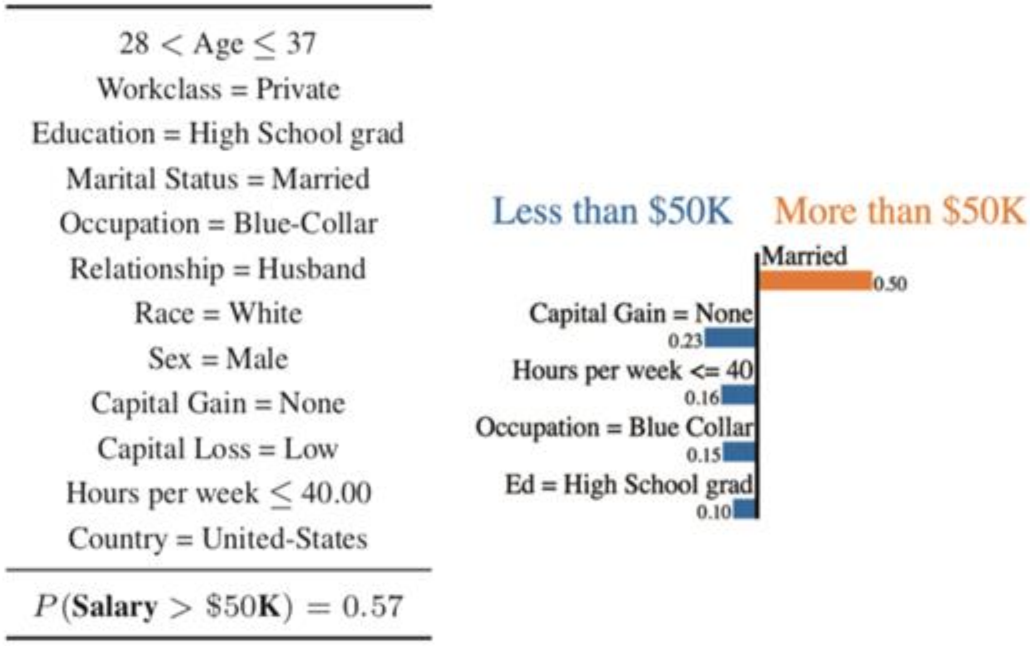


Figure credit: Ribeiro et al. KDD 2016



(a) Instance and prediction (b) LIME explanation

Figure credit: Anchors: High-Precision Model-Agnostic Explanations. Ribeiro et al. AAAI 2018

Influence functions

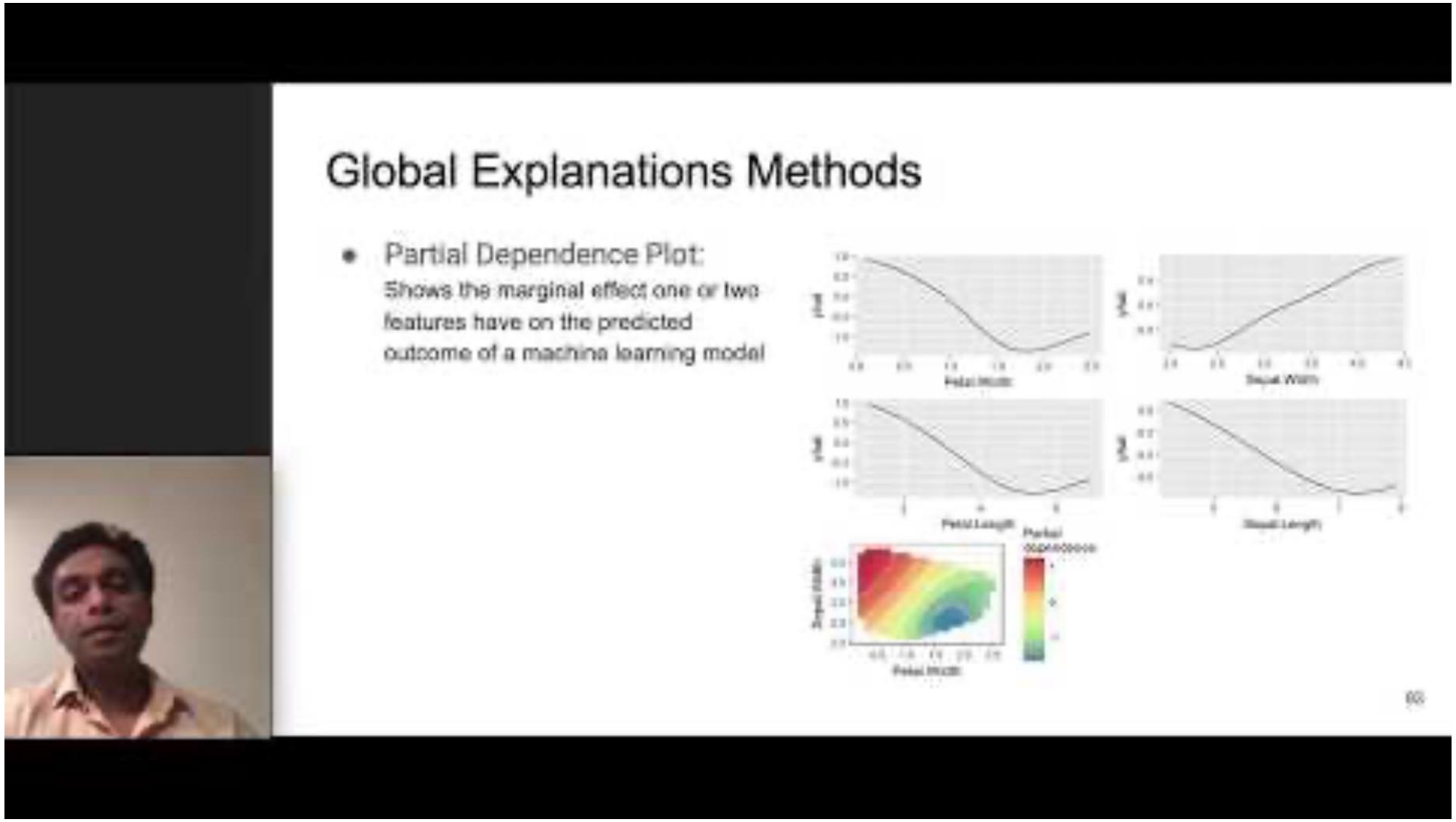
- Trace a model's prediction through the learning algorithm and back to its training data
- Training points "responsible" for a given prediction

Figure credit: Understanding Black-box Predictions via Influence Functions. Koh and Liang. ICML 2017

This Lecture

- Motivation for Explainable AI
- Overview of Explainable AI Techniques
  - Global Explanations
- Case Studies

Global Explanations

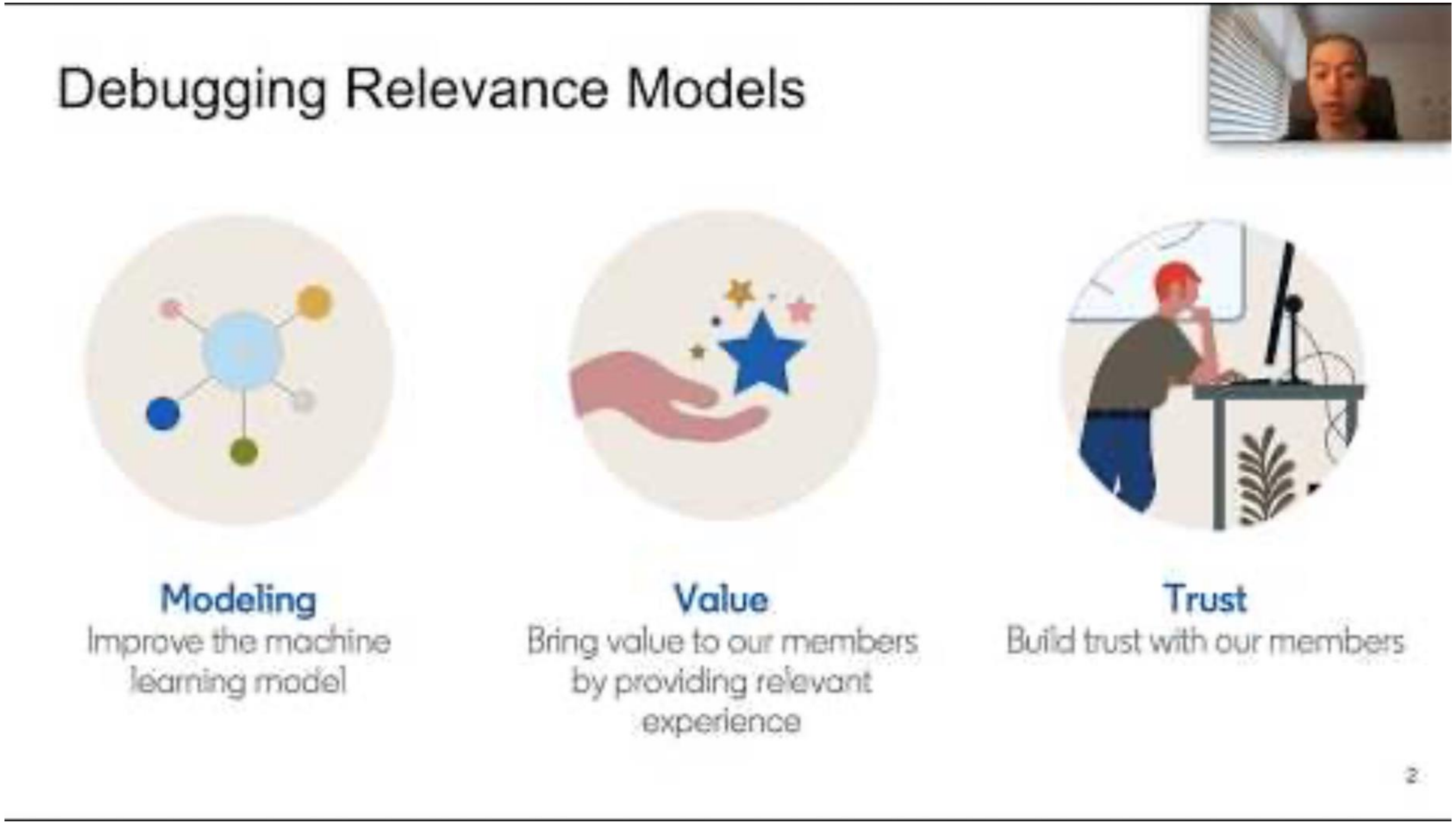


https://www.youtube.com/watch?v=Do_ito-X5KY

This Lecture

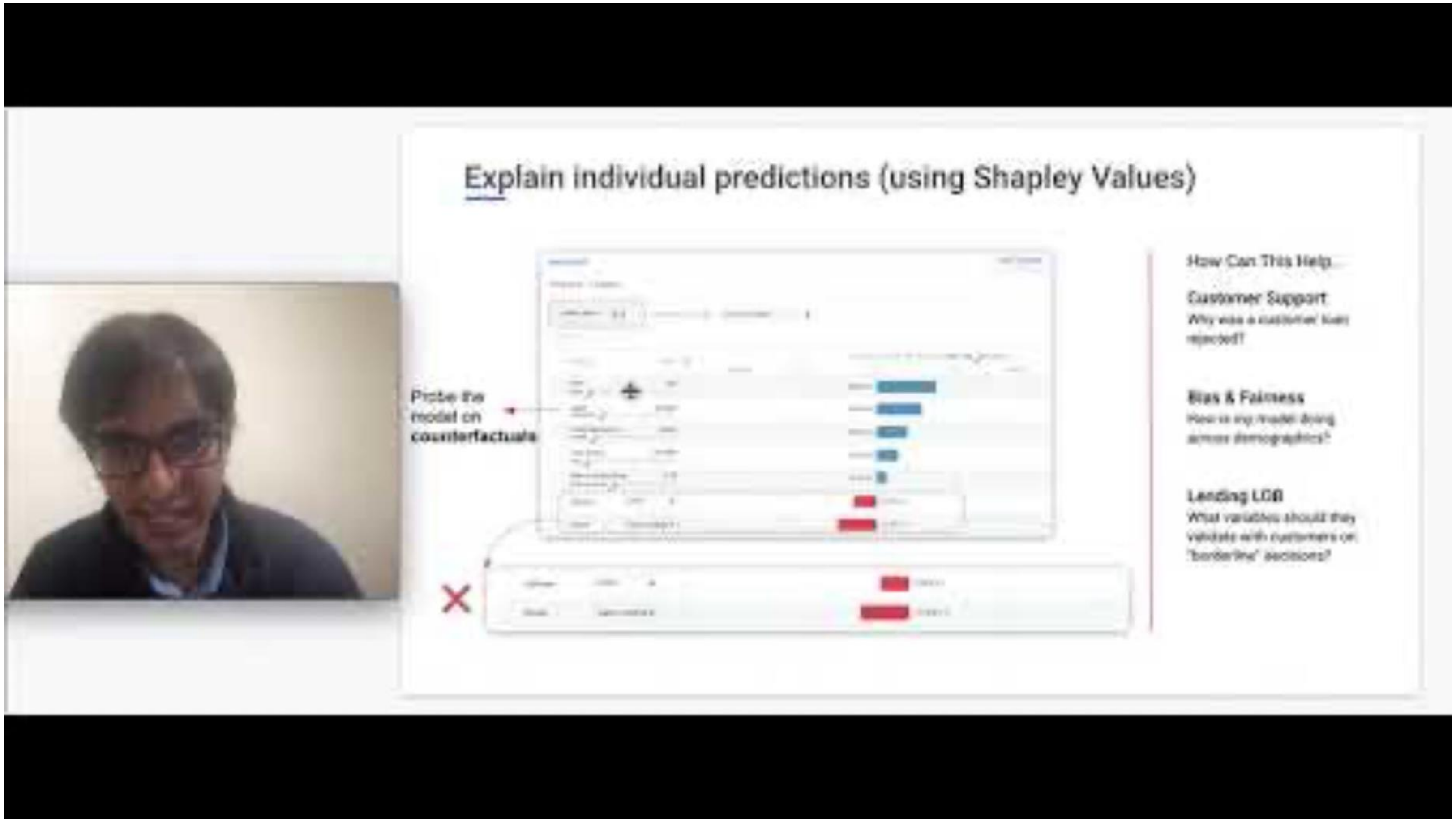
- Motivation for Explainable AI
- Overview of Explainable AI Techniques
  - Global Explanations
- Case Studies

LinkedIn Relevance Debugging & Explaining



https://www.youtube.com/watch?v=WsOrjE4Muio

Diabetic Retinopathy & Fiddler Case Studies



https://www.youtube.com/watch?v=iMHrl1hAr6U

References

- https://sites.google.com/view/explainable-ai-tutorial
- https://www.slideshare.net/slideshow/explainable-ai-in-industry-www-2020-tutorial/231998856
- https://christophm.github.io/interpretable-ml-book/