

# Explanations

Anupam Datta

John Mitchell

Stanford CS 329T, Spring 2021

# Course update

- This is Week Two!
- This is a new course. We are developing it as we proceed.
  - You have our complete attention for the quarter!
  - But we might not get everything right the first time.
- Planned weekly schedule
  - Sunday evening: announce new videos for the week
  - Monday - Tuesday: watch videos before Tuesday class meeting
  - Tuesday: Fireside chat, general discussion of topics for the week
  - Wednesday – Friday: meet with your lab section, work on homework
  - Friday 5pm: Homework due

# Course update

- This is Week Two!
- This is a new course. We are developing it as we proceed.
  - You have our complete attention for the quarter!
  - But we might not get everything right the first time.
- Planned weekly schedule
  - Sunday evening: announce new videos for the week
  - Monday - Tuesday: watch videos before Tuesday class meeting
  - Tuesday: Fireside chat, general discussion of topics for the week
  - Wednesday – Friday: meet with your lab section, work on homework
  - Friday 5pm: Homework due

Class poll  
Use the Q&A  
feature of Zoom

Is the course organization  
clear? If not, why?

# Recall course objective

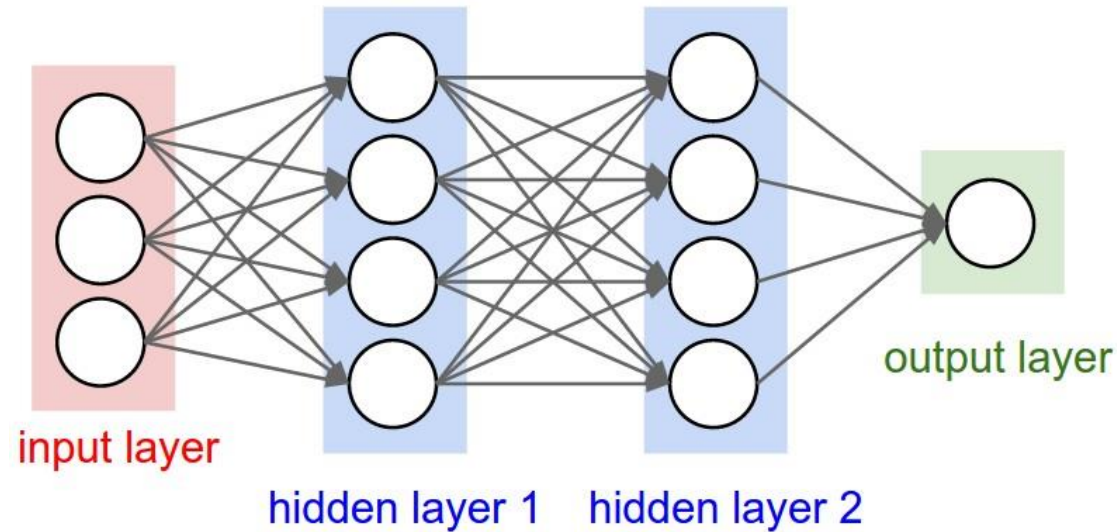
Understand how to assess and improve  
trustworthiness of ML Models

# Recall course objective

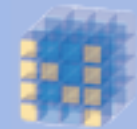
Understand how to assess and improve trustworthiness of ML Models

- Module 1
  - Background – recorded video, discussion last week
  - Homework 1 – due end of this week
- Module 2
  - Explanations – recorded video, discussion this week and next
  - Homework 2 – posted end of this week, due in *two weeks*

# Course module 1: Machine Learning Review



python



NumPy



Keras

# Homework 1

Practice using basic tools:

- Scikit-learn
- Numpy
- Keras

By implementing digit classifiers using:

- Logistic regression
- Deep neural network models

Sample application exercises with logistic regression:

- Explanations
- Adversarial examples
- Model stealing





Class poll  
Use the Q&A  
feature of Zoom

Are there questions  
about the review?

# Course module 2: Explainability

Explanations enable understanding and debugging of ML models

- How can we explain individual predictions from ML models?
- How can we explain an ML model's behavior at a global level?
- How can explanations enable testing, debugging, and improvement of ML models?

# Videos announced this week

Trustworthy Machine Learning

## Local Interpretable Model-agnostic Explanations (LIME)

Anupam Datta  
John Mitchell

Stanford CS 329T, Spring 2021



Trustworthy Machine Learning



## Quantitative Input Influence

ML Explanations with Causality & Game Theory

Anupam Datta  
John Mitchell

Stanford CS 329T, Spring 2021

# Fireside chat

- The fireside chats were a series of evening radio addresses given by Franklin D. Roosevelt, President of the United States, between 1933 and 1944. ... On radio, he was able to quell rumors, counter conservative-dominated newspapers and explain his policies directly to the American people.

First address March 12, 1933







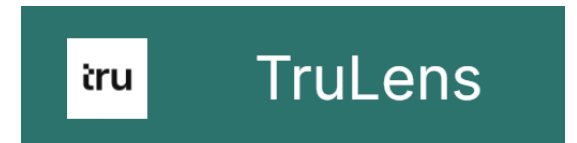
# Homework 2

## Explanations for traditional models

- Model-agnostic: Shapley Values, local approximations (QII, LIME, SHAP)
- Model-specific optimizations: TreeSHAP

## Explanations for deep models

- Saliency Maps
- Integrated Gradients
- Influence-directed Explanations



## What will I be doing?

- Understand, reason about, and extend the theoretical principles backing explanation frameworks like QII, LIME, SHAP, Integrated Gradients, etc.
- Implement a subset of these methods yourself (LIME, some deep explanations)
- Become familiar with open-source packages that enable explanations (SHAP, Trulens)

Class poll  
Use the Q&A  
feature of Zoom

Are there questions  
about the organization of  
Module 2, Explanations?

# Fireside Chat 2 – Explanations

With some overview of the videos on LIME and QII because this is the first week and some of you may not have planned time to see them before this class meeting.








# Explanations



# What's an *explanation?*

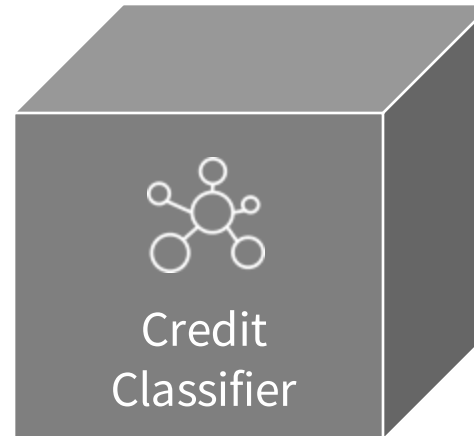
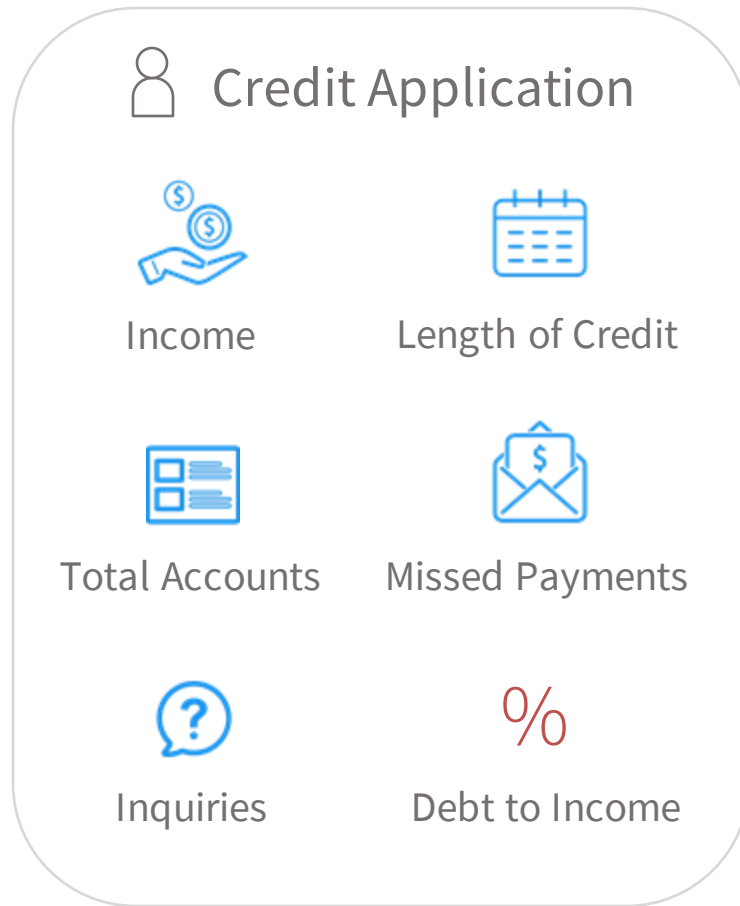
- “Textual or visual artifacts that provide qualitative understanding of the relationship between the instance’s components (e.g, words in text, patches in image) and the model’s prediction.” [Ribeiro et al]
- Example: A list of components and weights indicating their significance in making a prediction.



# Why do we want one?

- Determine trust in prediction
  - The model suggests a medical diagnosis for a serious health condition. Should the doctor act on this diagnosis?
- Determine trust in the model
  - Several possible models could be deployed
    - Which do we trust most?
    - Why?
  - And are there improvements?

# Explanations are Necessary



**DENIED**

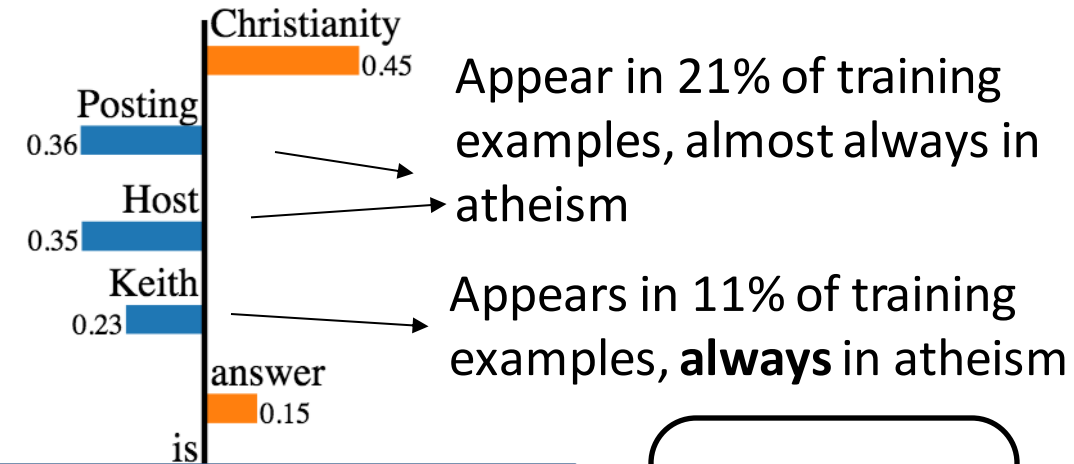
# Explanations based on components and weights

From: Keith Richards  
Subject: Christianity is the answer  
NTTP-Posting-Host: x.x.com

I think Christianity is the one true religion.  
If you'd like to know more, send me a note

atheism

christian



→ Will not generalize  
→ Don't trust this model!

Why did this happen? How do I fix it?

# Feature Importance



Class poll  
Use the Q&A  
feature of Zoom

State one difference  
between the LIME  
explanation (words with  
weights) and the QII  
explanation (features and  
significance)

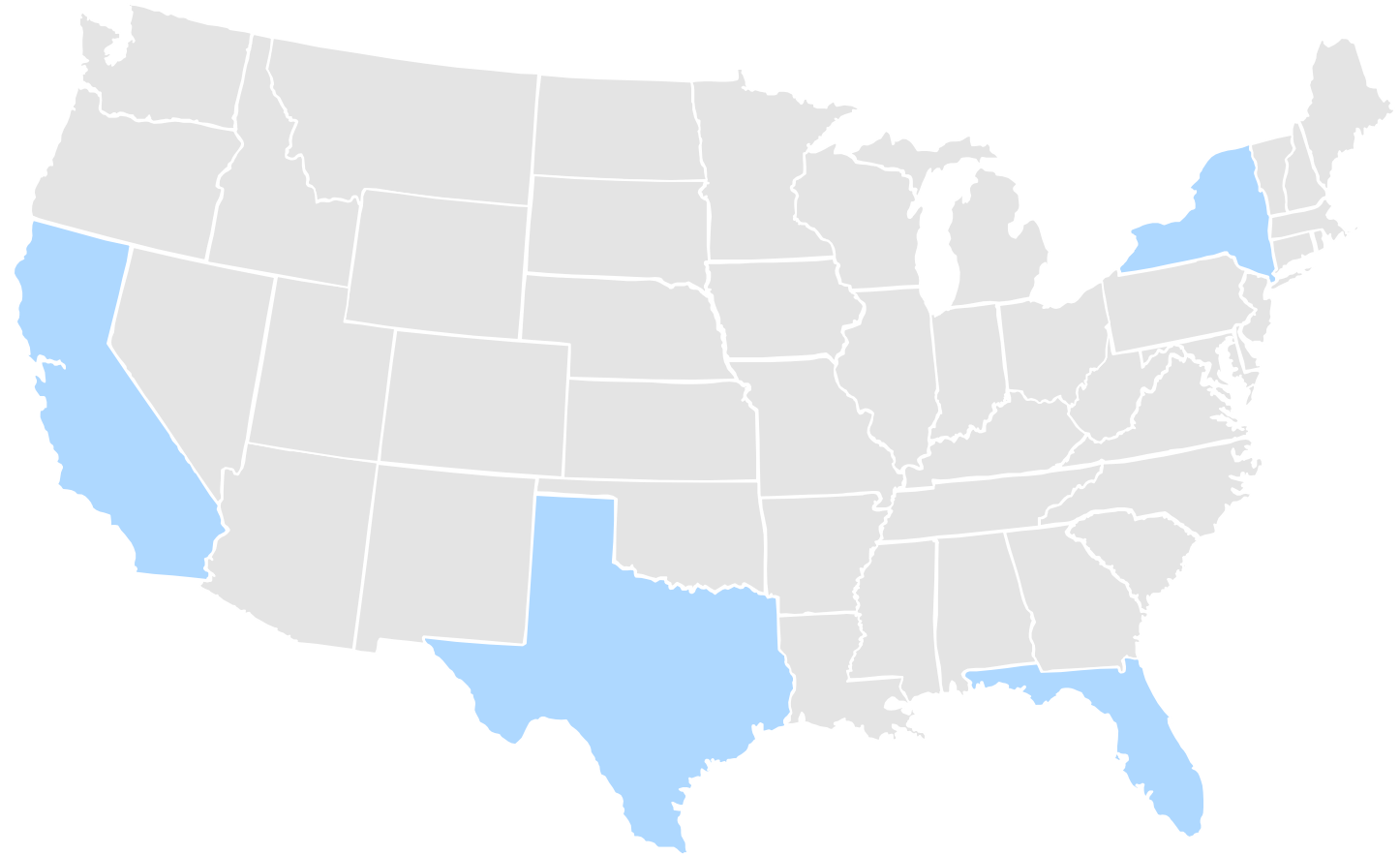


# Query Definition



# Explaining more than one question is critical

Which states contribute the most electoral votes?



# Explaining more than one question is critical

# Which states contribute the most electoral votes?

# Which states decide the winner?

# Query Definition by Method

	QII	PI	LIME	KernelSHAP	TreeSHAP
Individual Predictions (all output types)	✓	✓	✓	✓	x (log- odds)
Aggregate Model Properties (Disparate Impact Ratio)	✓	x	x	x	x

# Breakout discussion

Talk with your small group:

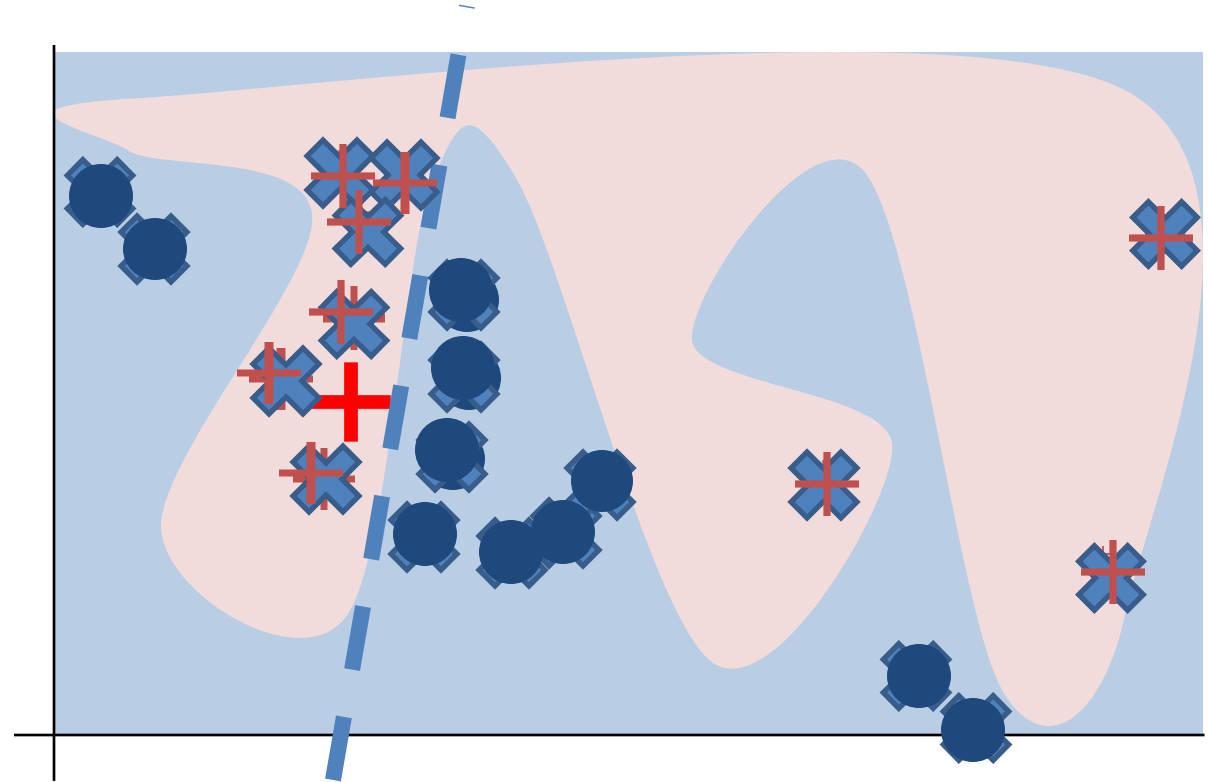
- How would *you* explain which states determine the presidential election?
- What *statistics* would you use?

The icon consists of a red outline of a person's head and shoulders on the left, followed by a blue bracket '[' and a blue arrow '→' pointing to the right.

# Output Comparisons

# LIME: explain complex model's prediction for input $x_i$

1. Sample points around  $x_i$
2. Use complex model to predict labels for each sample
3. Weigh samples according to distance to  $x_i$
4. Learn new simple model on weighted samples
5. Use simple model to explain



# Power of a State (Feature)

# Which states decide the winner?

# Causal Influence of Pennsylvania is high

# Power Depends on Marginal Influence

Goal: Capture feature interactions

What is the effect of PA after results from IN, GA, MD are in?





# Output Comparisons by Methods

	QII	PI	LIME	TreeSHAP	KernelShap
Feature Interactions	✓	x	✓	✓	✓
Flexible Comparison Groups	✓	x	x	✓	x
Causal influence	✓	✓	x	o	o
Model Agnostic	✓	✓	✓	x	✓



Further discussion

# Part II – Explanations | Week 1

## Fireside Chat

- How to perturb/define a neighborhood of points in a region
  - A common refrain: “What happens if you go out of distribution?”
    - LIME: What happens if the local region you are exploring/perturbing is not in the support of the training data?
    - QII: Generating near clones in a reasonable way
  - Mechanisms for perturbation: how do we do this?
    - Can be tough to even define a distance metric for certain types of input?
    - Why should the weighting method matter? Do different weightings need different interpretations?

# Part II – Explanations | Week 1

## Fireside Chat

- Causality
  - How do we capture causality within explanations? Is this critical?
- Feature interaction
  - If we have a sense of which features are interacting, how do we measure the importance of interactions within the QII framework?
  - How do we do this if we don't know which features are interacting?

# Breakout discussion

## Causality and feature interaction:

- What is the difference between causality and correlation?
- Is causality important for explanations?
- How do feature interactions and correlations change our understanding of causality? Of explanation?

# Part II – Explanations | Week 1

## Fireside Chat

- What constitutes a good explanation framework?
  - Flexibility in outcome, in comparison group, internal consistency, scaling
- Schools of thought:
  - inherently interpretable models vs. standalone explanation methods (and the continuum in between)

# Part II – Explanations | Week 1

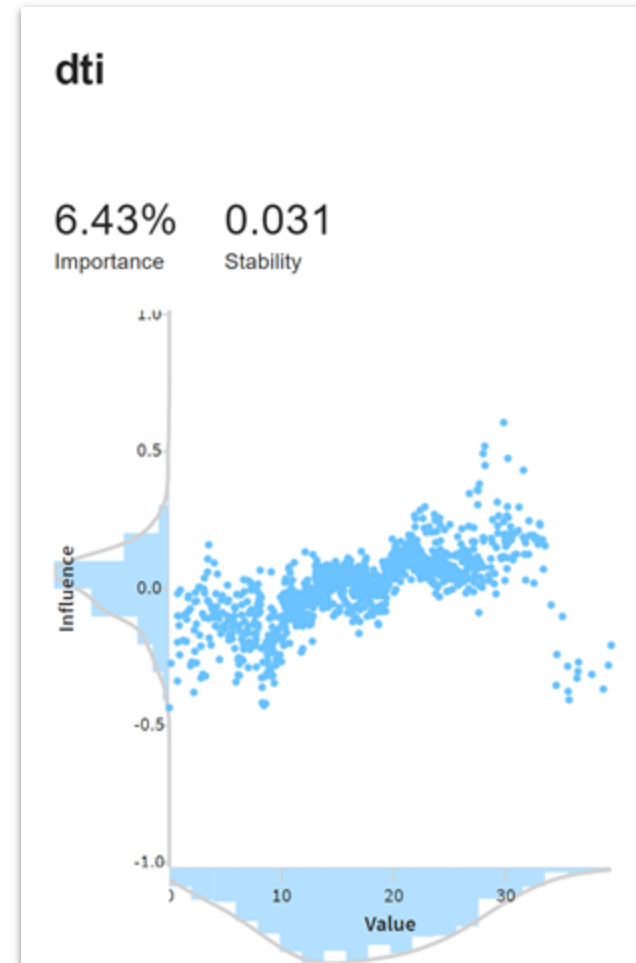
## Fireside Chat

- Global explanations
  - Sensitivity plots
- Counterfactual Explanations
  - Reasons vs actionable recourse (counterfactual explanations)
- Explanations as a building block to assess model quality
  - Fairness, Privacy, Accuracy, Stability,...(later in the course)

# Global Explanations

Is the model (in)consistent with my domain knowledge?

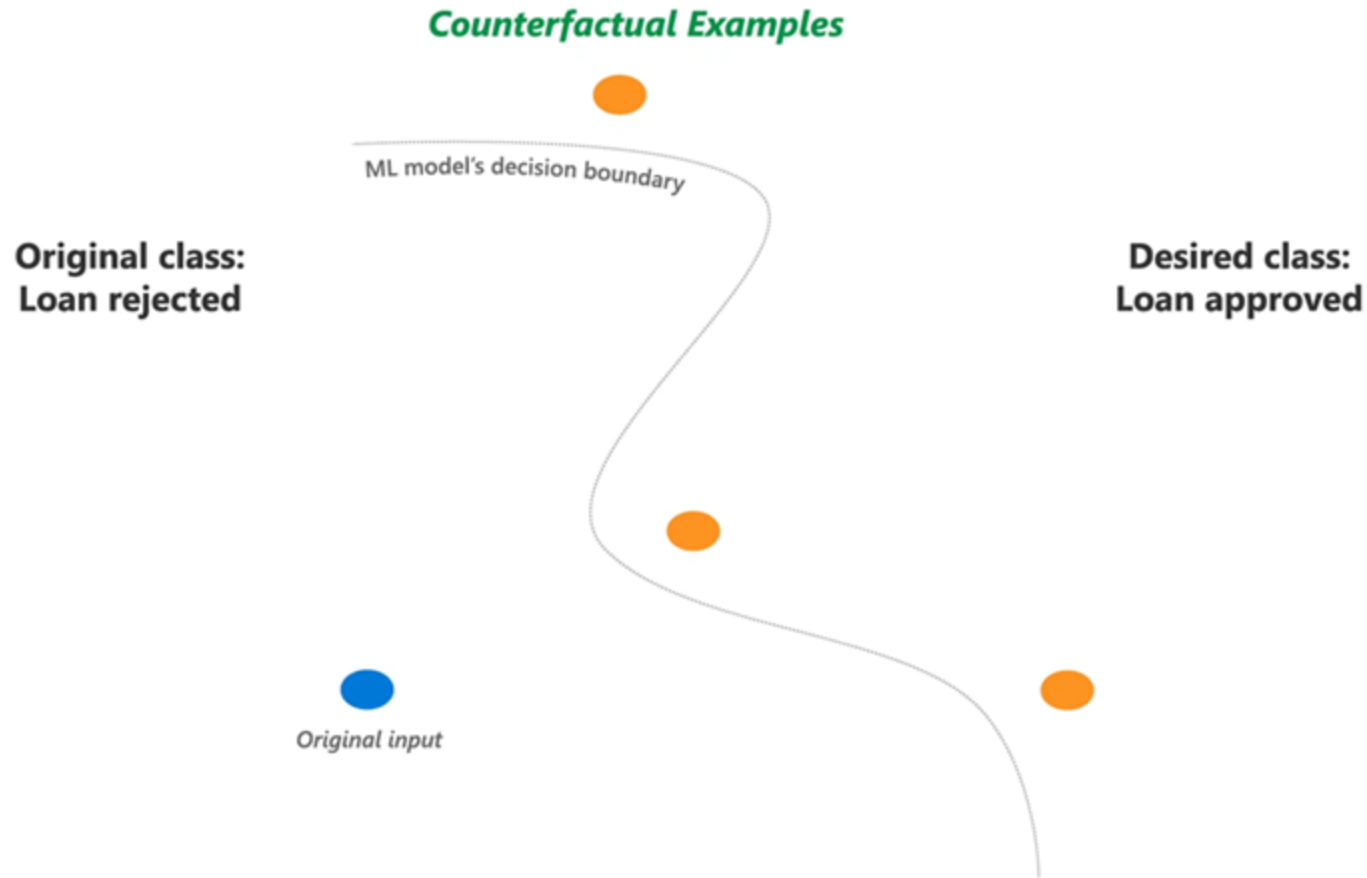
Influence sensitivity plots show how the model's output depends on a feature (debt to income ratio) while accounting for feature interactions





# Actionable Recourse (Action Codes)

If Jane's loan application is rejected, how can she improve her outcomes for the future?  
Need to provide **action codes** for what she can change.



Building on existing literature:

- [Poyiadzi et al](#) (AAAI '20)
- [Utsun et al](#) (FAT '19)
- [Wachter et al](#) (Harvard Journal of Law & Technology)

# Action Codes for Adverse Action Notices

Recourse frameworks must satisfy a set of principles:



## Validity

Will this suggestion change the outcome?  
→ Increase your salary by \$100!



## Actionability

Can the user follow up on the suggestion?  
→ Reduce your age by ten years!



## Proximity

Is this the smallest change possible?  
→ Increase your salary by \$1,000,000!



## Sparsity

Is the user forced to change many things?  
→ Earn more, decrease debt, relocate.



## Privacy

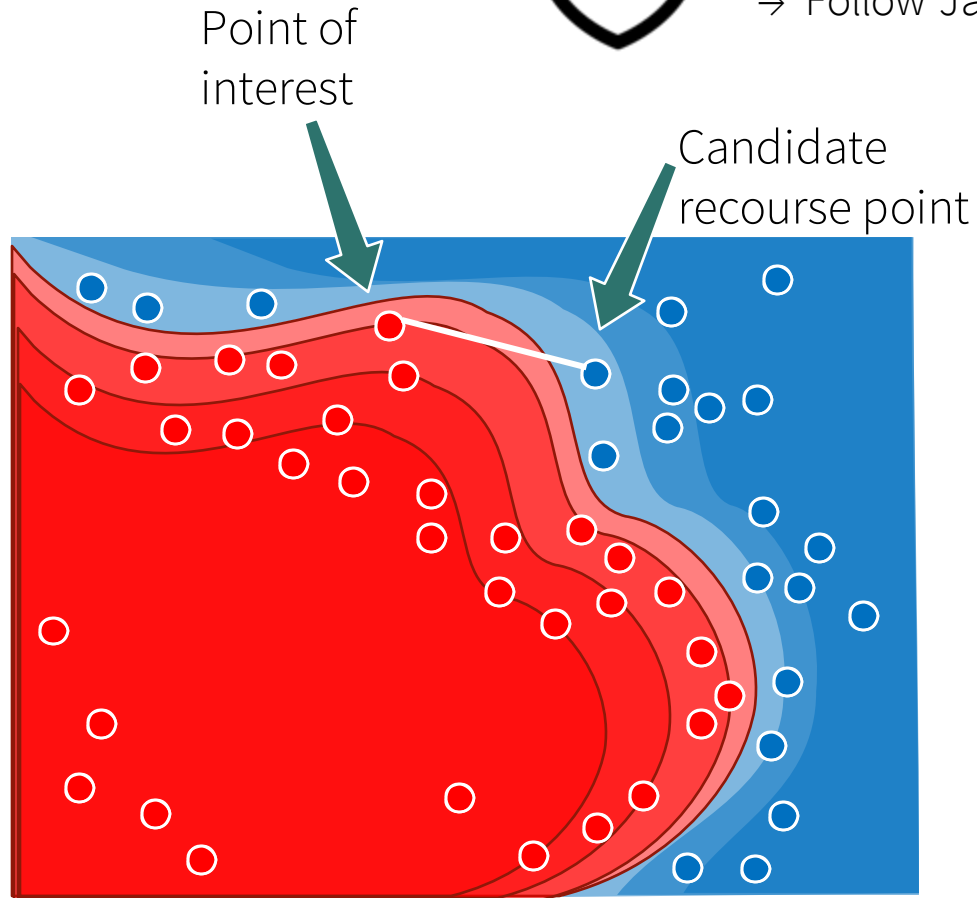
Will this suggestion leak customer data?  
→ Follow Jane's financial profile.

# Action Codes for Adverse Action Notices



## Privacy

Will this suggestion leak customer data?  
→ Follow Jane's financial profile.



Suggesting a specific point as a recourse strategy violates privacy for the individual we are trying to emulate.

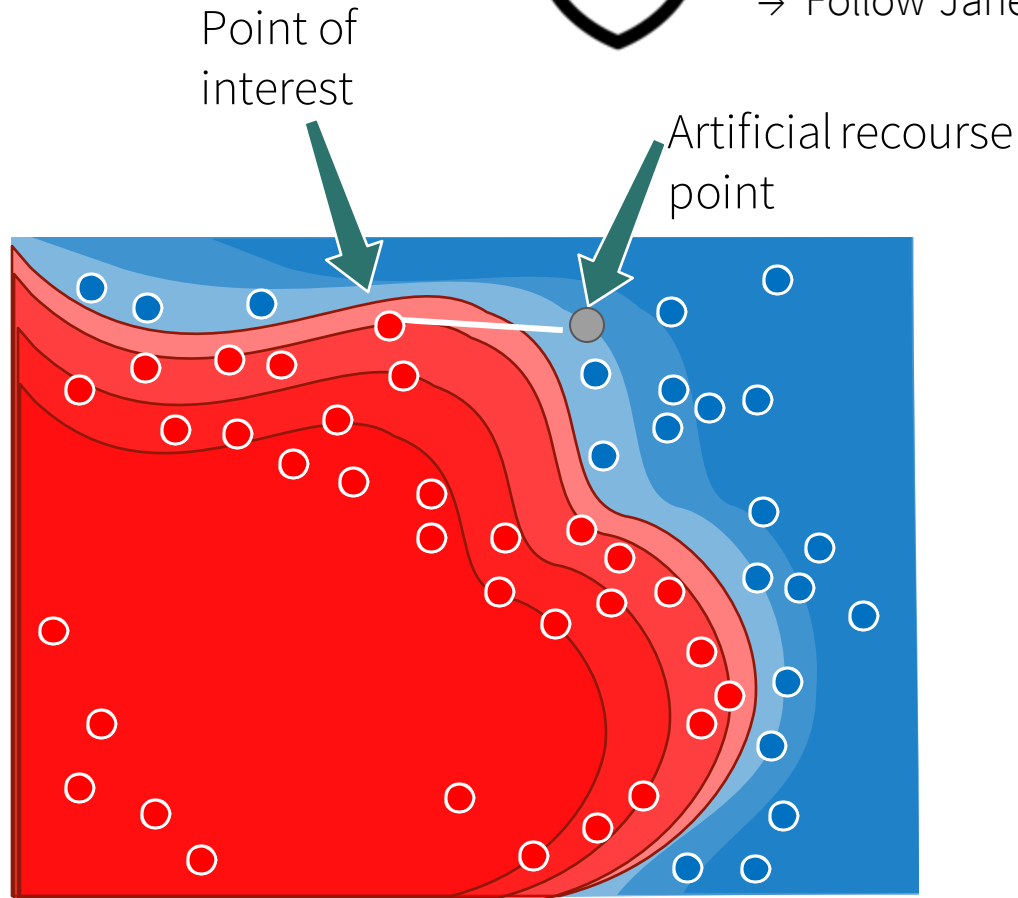
This is a gap in the literature.

# Action Codes for Adverse Action Notices



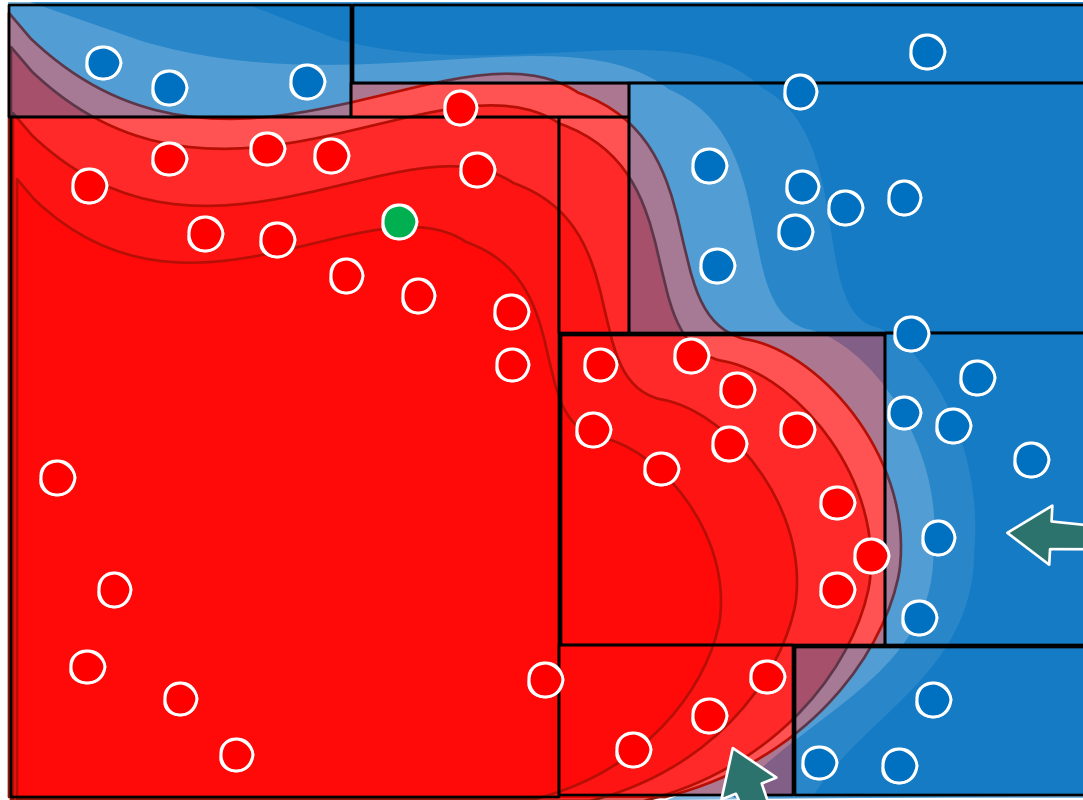
## Privacy

Will this suggestion leak customer data?  
→ Follow Jane's financial profile.



Even if we suggest emulating a point that isn't a real customer, this leaks model data about decision boundaries, sensitivities to input perturbation, etc.  
Users can game the system.

# Action Codes for Adverse Action Notices



Our approach:

Working paper Datta et al (2020):



Segment the input space

Create a sequence of rules partitioning the input space

Regions are dense, preserving (differential) **privacy**

Segments are interpretable and can be defined by a **sparse** sequence of rules