BMI 702: Biomedical Artificial Intelligence

Foundations of Biomedical Informatics II, Spring 2024

Lecture 4: Interpretability and explainability in biomedical AI

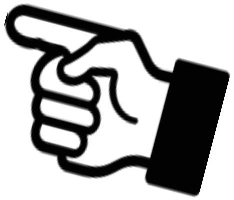
HARVARD MEDICAL SCHOOL

Marinka Zitnik

marinka@hms.harvard.edu

Outline for today's class

1. What is trustworthy ML and why should I care?



2. Interpretability vs. explainability

3. Explaining ML predictions

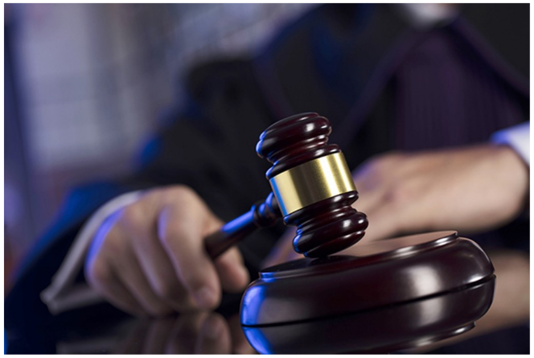
4. Case studies
   - Drug repurposing
   - Treatment recommendation

Trustworthy ML

- ML models are increasingly being deployed in real-world applications
  - It is critical to ensure that these models are behaving responsibly and are trustworthy
- There has been growing interest to develop and deploy ML models and algorithms that are:
  - Not only accurate
  - But also explainable, fair, privacy-preserving, causal, and robust
- This broad area of research is commonly referred to as trustworthy ML

Motivation

Model understanding is absolutely critical in several domains - particularly those involving **high stakes decisions**



Why model understaning?

Input

This model is
relying on incorrect
features to make
iction!! Let
model

Model understanding facilitates debugging

Predictive Model

Prediction = Siberian Husky

Hima Lakkaraju

Why model understanding?

This prediction is

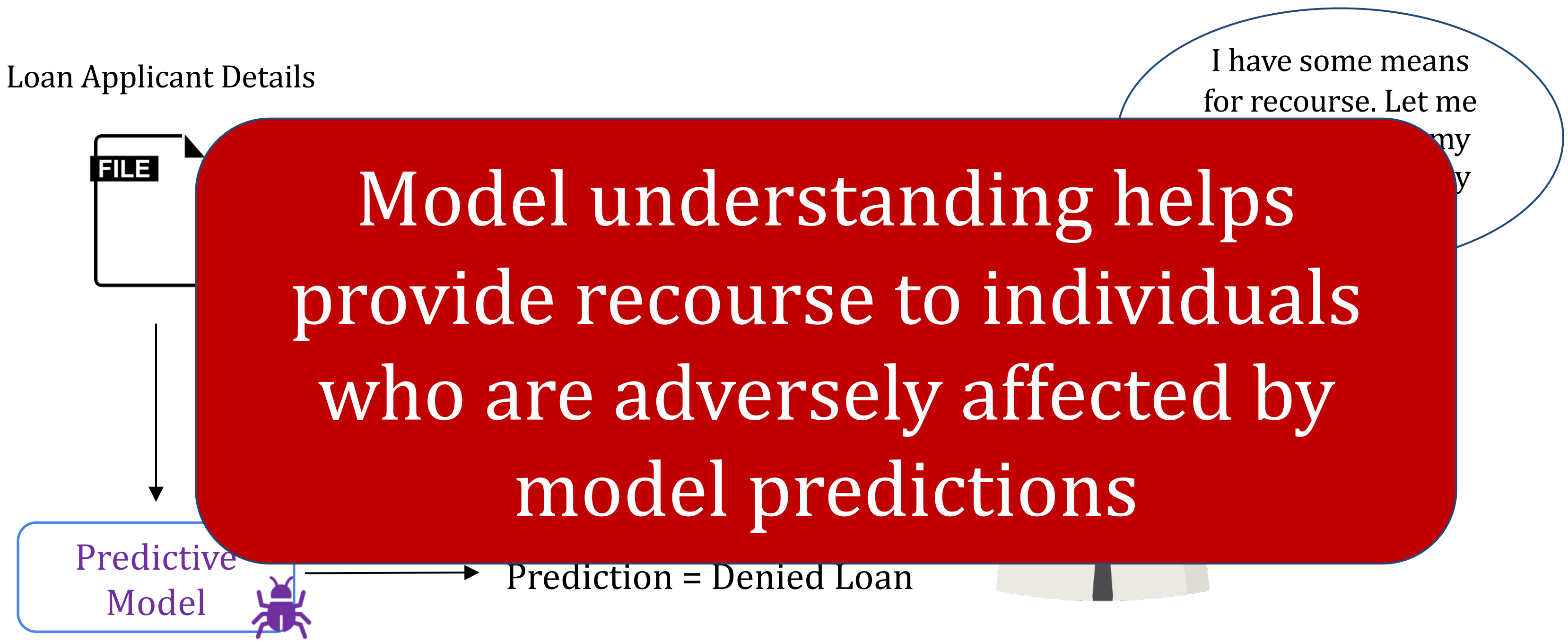
Defendant Detail

Model understanding facilitates bias detection

Predictive Model

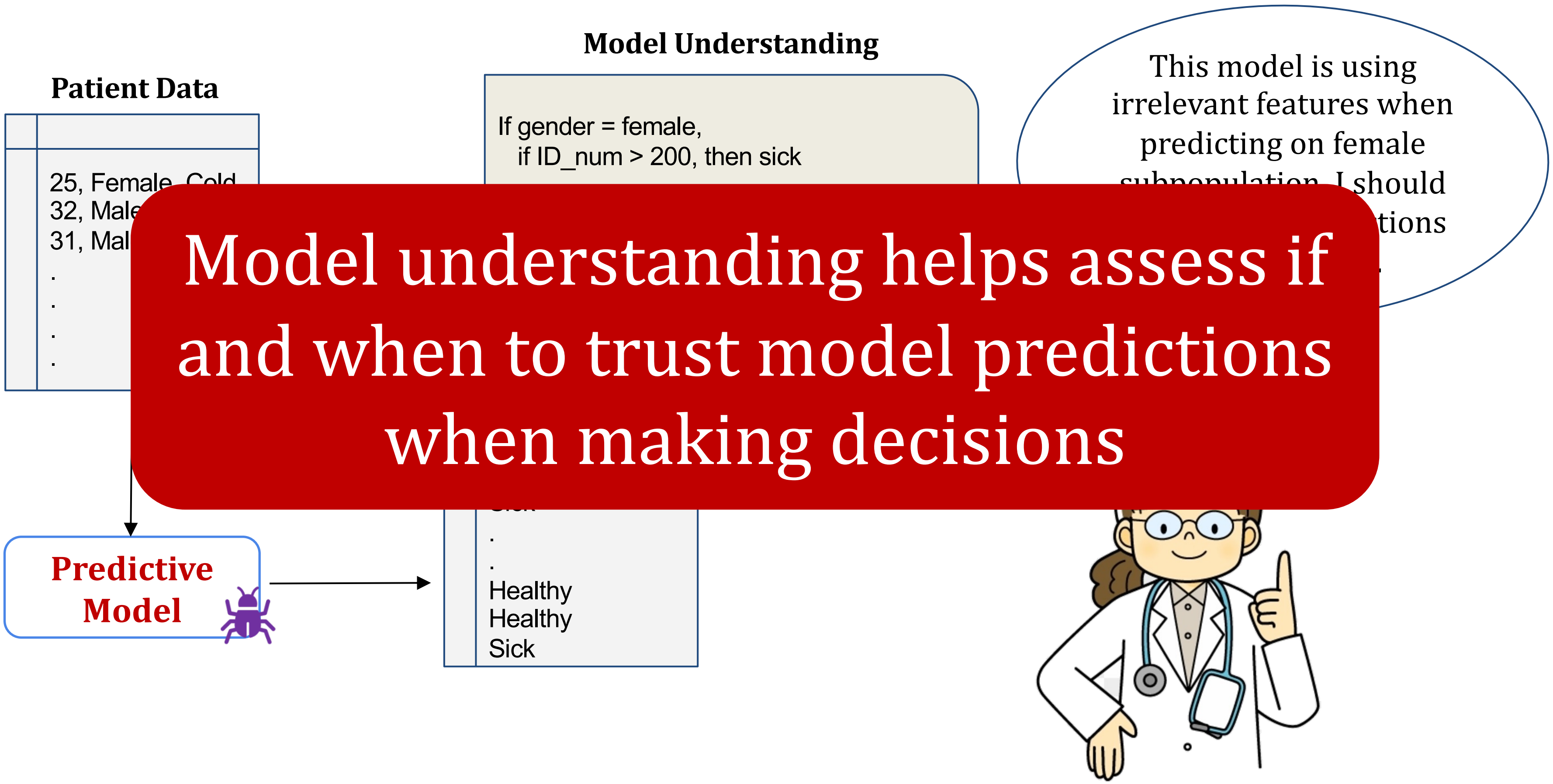
Prediction = Do not release on bail

Why model understanding?



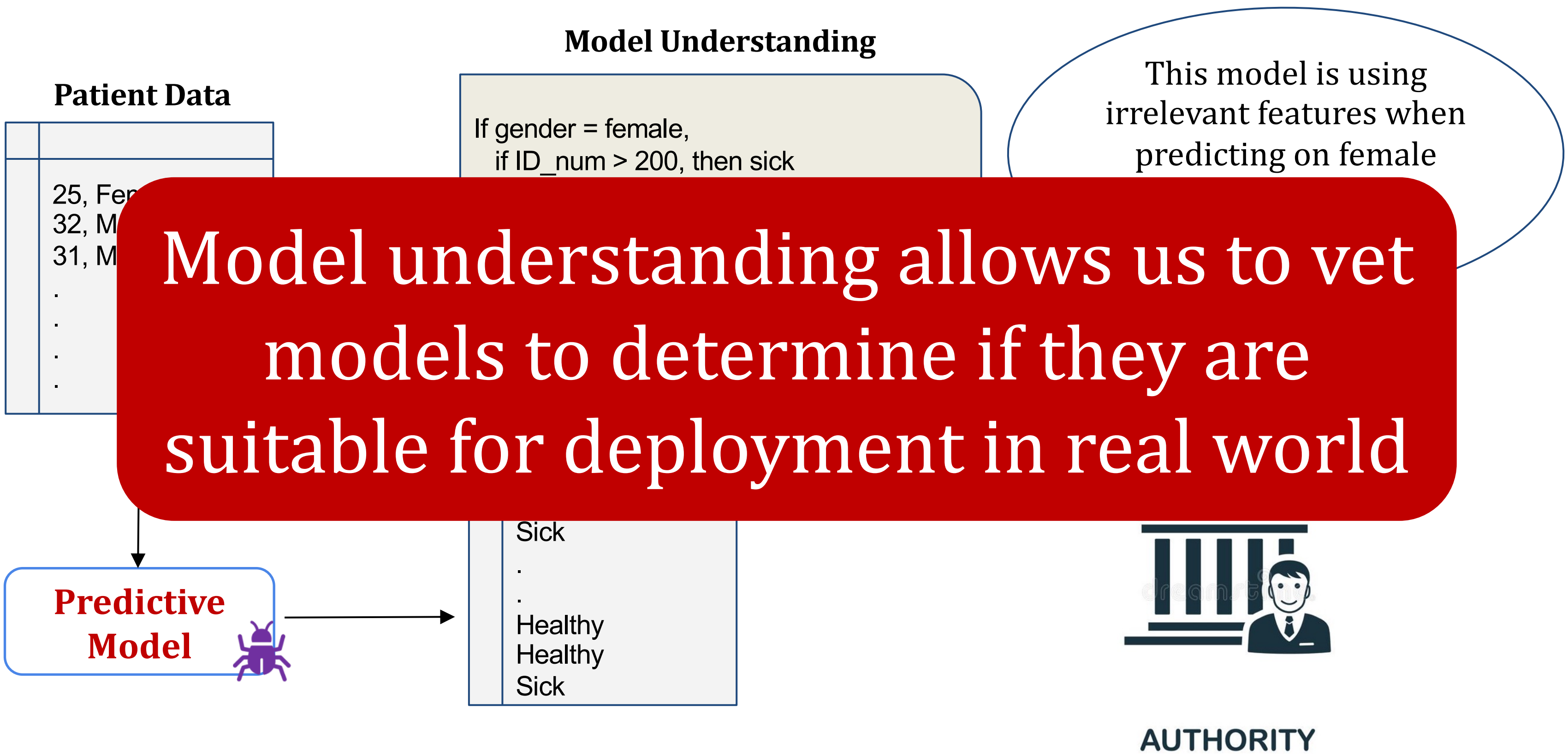
Model understanding helps provide recourse to individuals who are adversely affected by model predictions

Motivation: Why model understanding?



Model understanding helps assess if and when to trust model predictions when making decisions

Motivation: Why model understanding?



Model understanding allows us to vet models to determine if they are suitable for deployment in real world

Why should I care about understanding ML models?

Utility

Debugging

Bias Detection

Recourse

If and when to trust model predictions

Vet models to assess suitability for deployment

Stakeholders

End users (e.g., loan applicants)

Decision makers (e.g., doctors, judges)

Regulatory agencies (e.g., FDA, European commission)

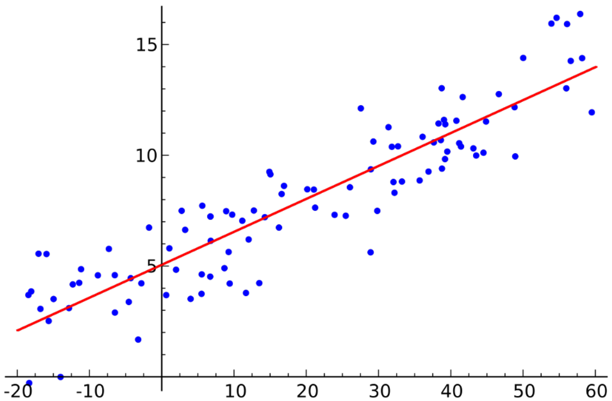
Researchers and engineers

Outline for today's class

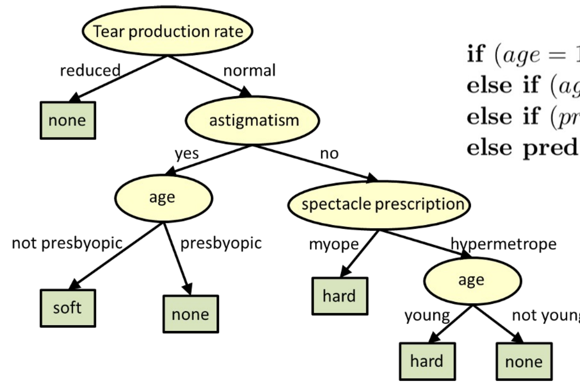
1. What is trustworthy ML and why should I care?
2. Interpretability vs. explainability
3. Explaining ML predictions
4. Case studies
   - Drug repurposing
   - Treatment recommendation

Achieving model understanding

Goal: Build <u>inherently interpretable</u> predictive models



Linear regression



Decision trees

if $(age = 18-20)$ **and** $(sex = male)$ **then predict** yes
else if $(age = 21-23)$ **and** $(priors = 2-3)$ **then predict** yes
else if $(priors > 3)$ **then predict** yes
else predict no

Decision rules

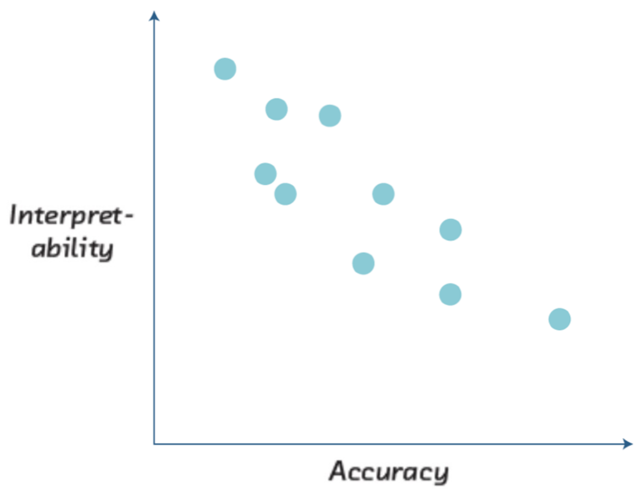
Saliency map of a black box (deep learning) model does not explain anything except where the model is looking: We have no idea why this image is labeled as either a dog or a musical instrument when considering only saliency. The explanations look essentially the same for both classes

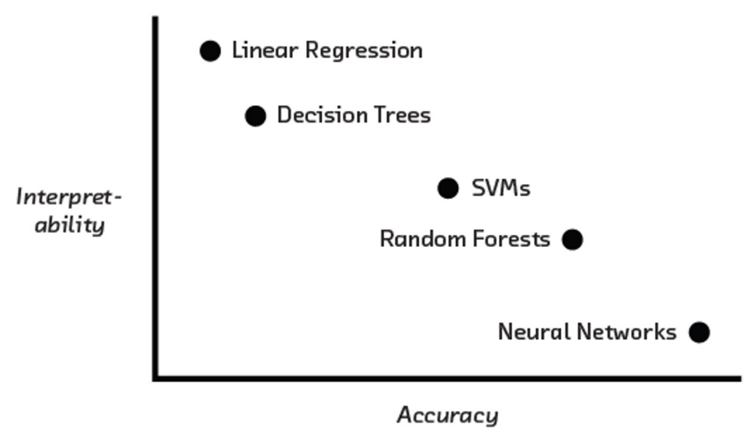
Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, *Nature Machine Intelligence* 2019

Inherently interpretable models vs. post hoc explanations

Accuracy-interpretability trade offs may exist in certain settings

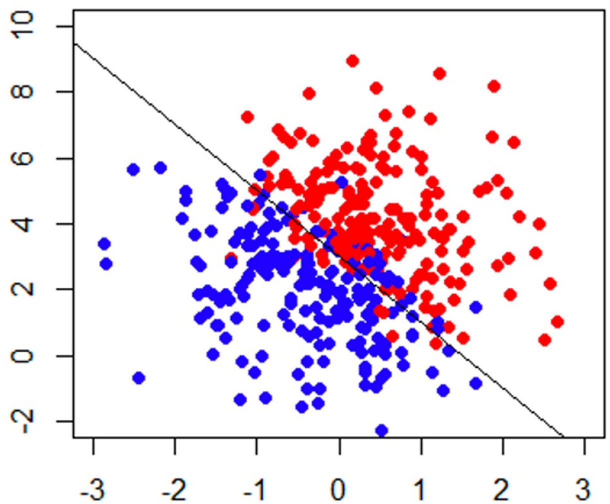
Example



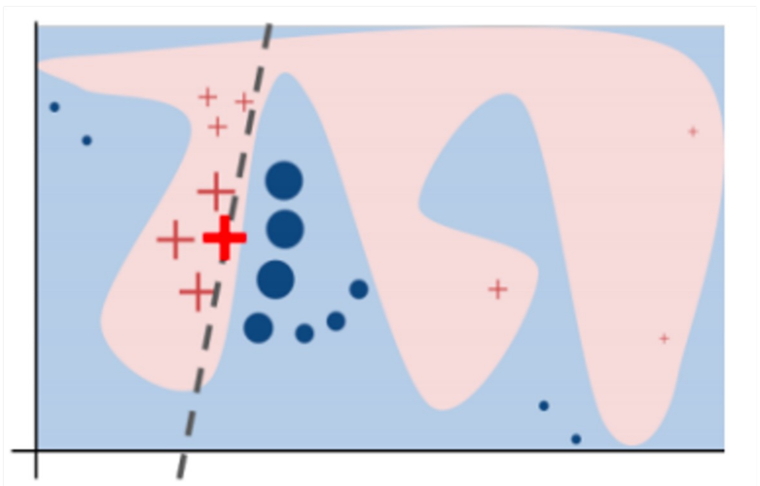


[Cireșan et. al. 2012, Caruana et. al. 2006, Frosst et. al. 2017, Stewart 2020]

Inherently interpretable models vs. post hoc explanations



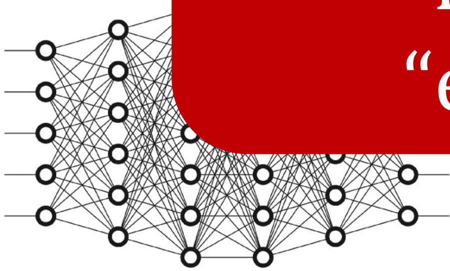
Build interpretable and accurate models

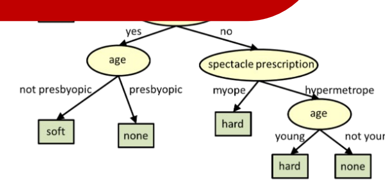


Complex models might achieve higher accuracy

Achieving model understanding

Explain pre-built models *in a post-hoc manner*





Interpretability/accuracy tradeoffs and proliferation of black box models force us to rely on post hoc "explanations" of ML models

[Ribeiro et. al. 2016, 2018; Lakkaraju et. al. 2019]

Inherently interpretable models vs. post hoc explanations

- If you can build an interpretable model which is also adequately accurate for your setting, DO IT!

- Sometimes, you don't have enough data to build your model from scratch
- And, all you have is a (proprietary) black box!
- Post hoc explanations come to the rescue!

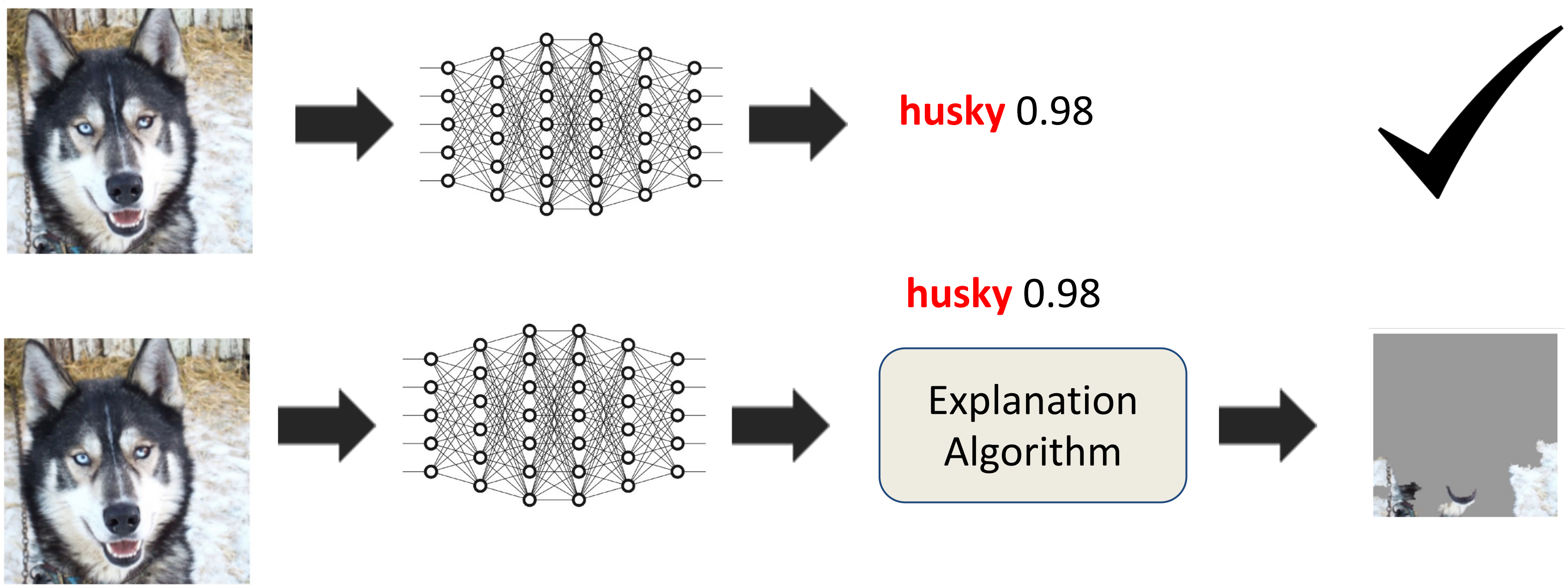
Next: Overview of post hoc explanations methods

Outline for today's class

1. What is trustworthy ML and why should I care?
2. Interpretability vs. explainability
3. Explaining ML predictions
4. Case studies
   - Drug repurposing
   - Treatment recommendation

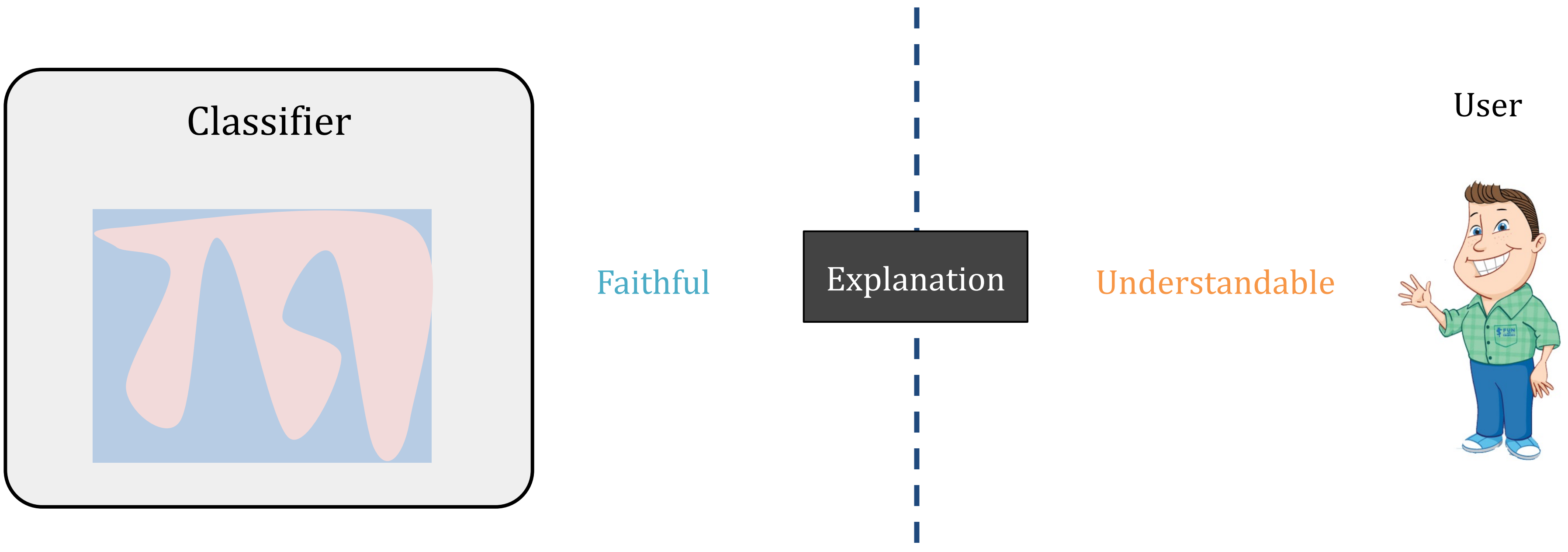
Explainable AI

"Explainable AI refers to the set of approaches that provide an interpretable description of the behavior of a given (complex) model to end users."



What is an explanation?

- **Definition:** Interpretable description of the model behavior



Overview of explanation methods

Local explanations

Explain individual predictions

Help unearth biases in the *local neighborhood* of a given instance

Help vet if individual predictions are being made for the right reasons

Global explanations

Explain complete behavior of the model

Sheds light on *big picture biases* affecting larger subgroups

Help vet if the model, at a high level, is suitable for deployment

Hima Lakkaraju

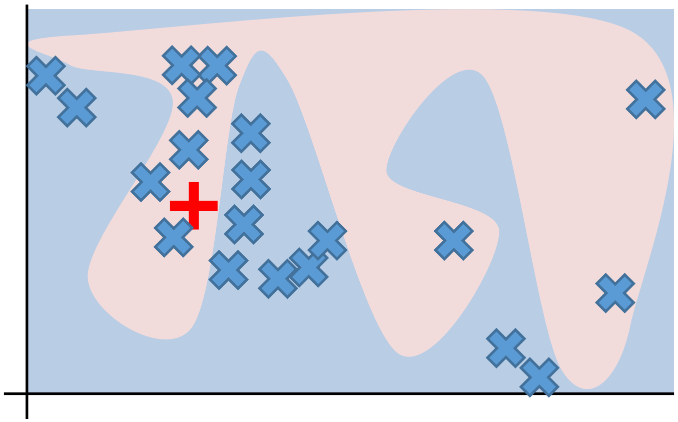
Overview of explanation methods

- Local explanation methods:
  - <u>Feature importance scoring</u>
  - Integrated gradients
  - Prototype explanations
  - Counterfactuals

- Global explanation methods:
  - Collection of local explanations
  - Representation-based explanations
  - Model distillation

LIME: Local interpretable model-agnostic explanations

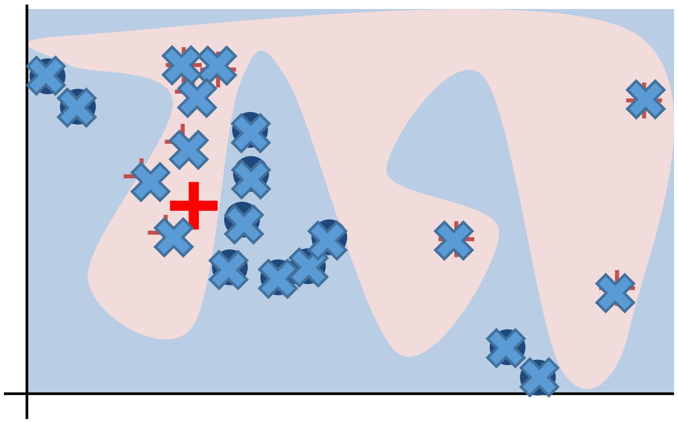
1. Sample points around x_i



[Ribeiro et al. 2016]

LIME: Local interpretable model-agnostic explanations

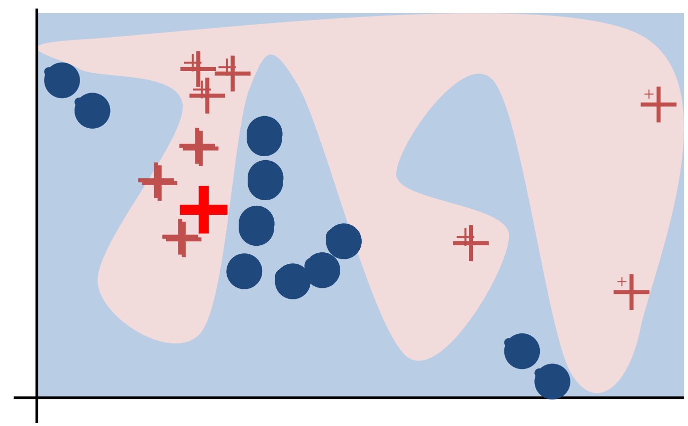
1. Sample points around x_i
2. Use model to predict labels for each sample



[Ribeiro et al. 2016]

LIME: Local interpretable model-agnostic explanations

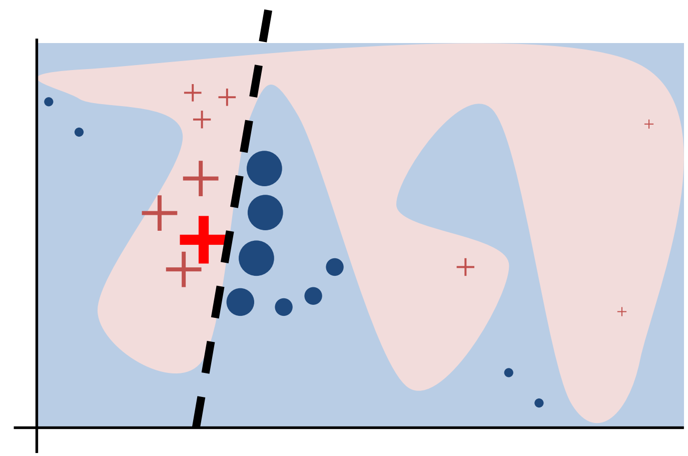
1. Sample points around x_i
2. Use model to predict labels for each sample
3. Weigh samples according to distance to x_i



[Ribeiro et al. 2016]

LIME: Local interpretable model-agnostic explanations

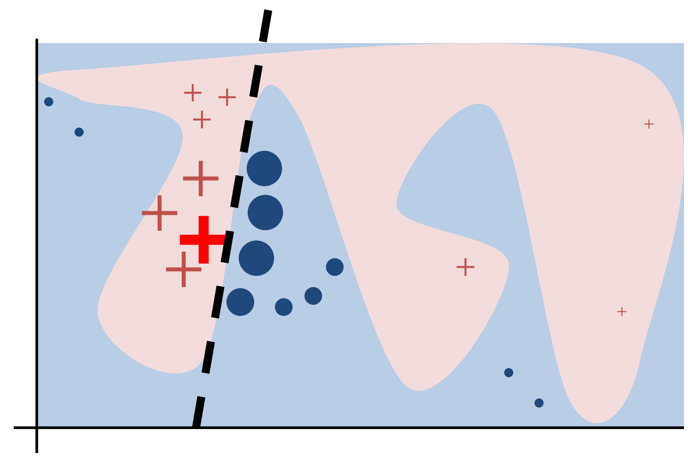
1. Sample points around x_i
2. Use model to predict labels for each sample
3. Weigh samples according to distance to x_i
4. Learn simple linear model on weighted samples



[Ribeiro et al. 2016]

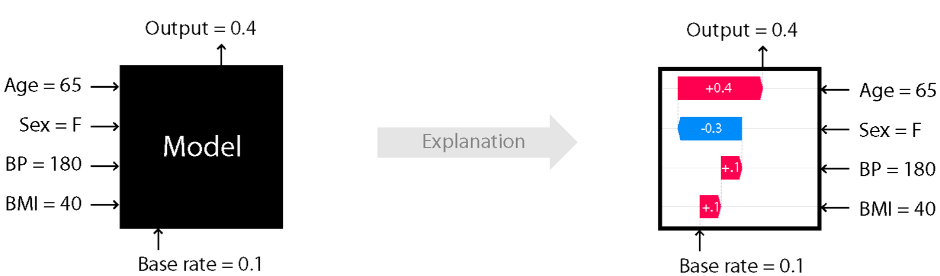
LIME: Local interpretable model-agnostic explanations

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



Another popular method which outputs feature importance scores: SHAP

SHAP values are based on game theory and assign an importance value to each feature in a model. Features with positive SHAP values positively impact the prediction, while those with negative values have a negative impact. The magnitude is a measure of how strong the effect is



[Ribeiro et al. 2016; Lundberg & Su-In Lee 2017]

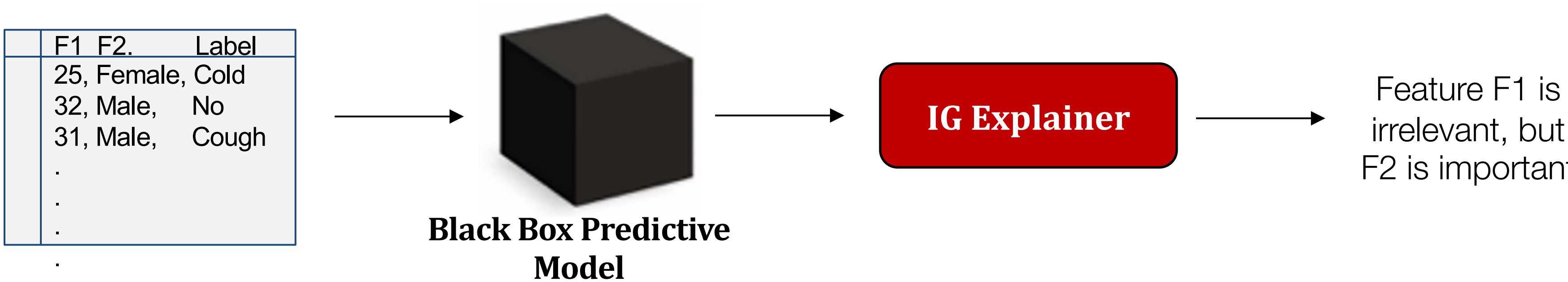
Overview of explanation methods

- Local explanation methods:
  - Feature importance scoring
  - Integrated gradients
  - Prototype explanations
  - Counterfactuals

- Global explanation methods:
  - Collection of local explanations
  - Representation-based explanations
  - Model distillation

Integrated Gradients (IG)

- Integrated Gradients (IG) is an explanation method for deep neural networks
- It identifies important features that contribute most to the model's prediction



- Appealing properties of integrated gradients:
  - It can be applied to any differentiable model like models for images, text, or structured data
  - It requires no modification to the original ML model

Axiomatic Attribution for Deep Networks, *ICML 2017*

https://distill.pub/2020/attribution-baselines

How does IG work?

- IG computes gradients of the model's prediction w.r.t. input features
- IG is built on two axioms which need to be satisfied:
  - Sensitivity and
  - Implementation invariance
- Sensitivity:
  - We establish a baseline instance as a starting point
  - We then build a sequence of instances which we interpolate from a baseline instance to the actual instance to calculate
- Implementation invariance:
  - Implementation invariance is satisfied when two functionally equivalent models have identical attributions for the same input image and the baseline image.
  - Two models are functionally equivalent when their outputs are equal for all inputs despite having very different implementations

Axiomatic Attribution for Deep Networks, *ICML 2017*

Calculating and visualizing IG

- Setup:
  - Let's consider an ML model for image classification
  - We aim to use IG to explain the predicted image label

- Step 1:
  - Start from a baseline where the baseline can be a black image whose pixel values are all zero or an all-white image, or a random image
  - Baseline input is one where the prediction is neutral and is central to any explanation method and visualizing pixel feature importance scores

Axiomatic Attribution for Deep Networks, *ICML 2017*

Calculating and visualizing IG

- Step 2:
  - Generate a linear interpolation between the baseline and the original image
  - Interpolated images are small steps(α) in the feature space between your baseline and input image and consistently increase with each interpolated image's intensity

Axiomatic Attribution for Deep Networks, *ICML 2017*

Calculating and visualizing IG

- Step 3: Calculate gradients to measure the relationship between changes to a feature and changes in the model's predictions
- The gradient informs which pixel has the strongest effect on the models predicted class probabilities
  - Varying variable changes the output, and the variable will receive some attribution to help calculate the feature importances for the input image
  - Variable that does not affect the output gets no attribution
- Step 4: Compute the numerical approximation through averaging gradients (that's why the method's name is integrated gradients)

Axiomatic Attribution for Deep Networks, *ICML 2017*

Calculating and visualizing IG

- Step 5:
  - Scale IG to the input image to ensure that the attribution values are accumulated across multiple interpolated images are all in the same units
  - Represent the IG on the input image with the pixel importances

IG helps us explain what an ML model looks at to make a prediction by highlighting the feature importances.
It does this by computing the gradient of the model's prediction output to its input features.

Attribution mask   Overlay IG on Input image

Overview of explanation methods

- Local explanation methods:
  - Feature importance scoring
  - Integrated gradients
  - <u>Prototype explanations</u>
  - Counterfactuals

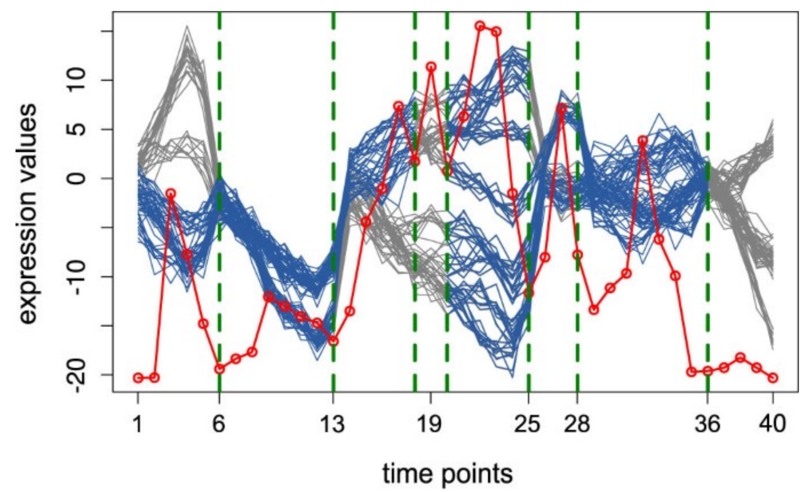
- Global explanation methods:
  - Collection of local explanations
  - Representation-based explanations
  - Model distillation

Prototype-based explanations

- Use examples (synthetic or natural) to explain individual predictions
- Influence Functions (Koh & Liang 2017)
  - Identify instances in the training set that are responsible for the prediction of a given test instance
- Activation Maximization (Erhan et al. 2009)
  - Identify examples (synthetic or natural) that strongly activate a function (neuron) of interest

Prototypes for explaining time series models

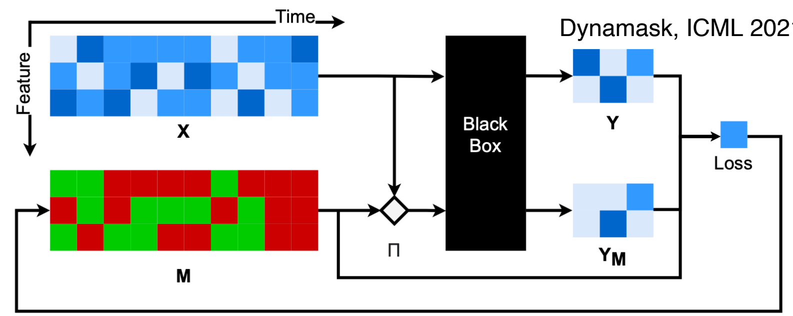
- Time series are not easily visually interpretable
  - Noisy samples
  - Dense informative features, unlike imaging and text modalities
- **Temporal patterns**
  - Only show up when looking at time segments and long-term behaviors
- Perturbations matter
  - Setting a value to zero does not ignore that time point
  - Temporal dependencies cannot be ignored



Omranian et al., 2015

Existing time series explainers are inadequate

1. **Perturbations are continuous**
   - Can deform shape of samples
2. **Give only instance-based explanations**
   - Cannot relate patterns across samples
3. **Fail to match performance of generic explainers**
   - Post-hoc methods suffer from a lack of faithfulness and stability



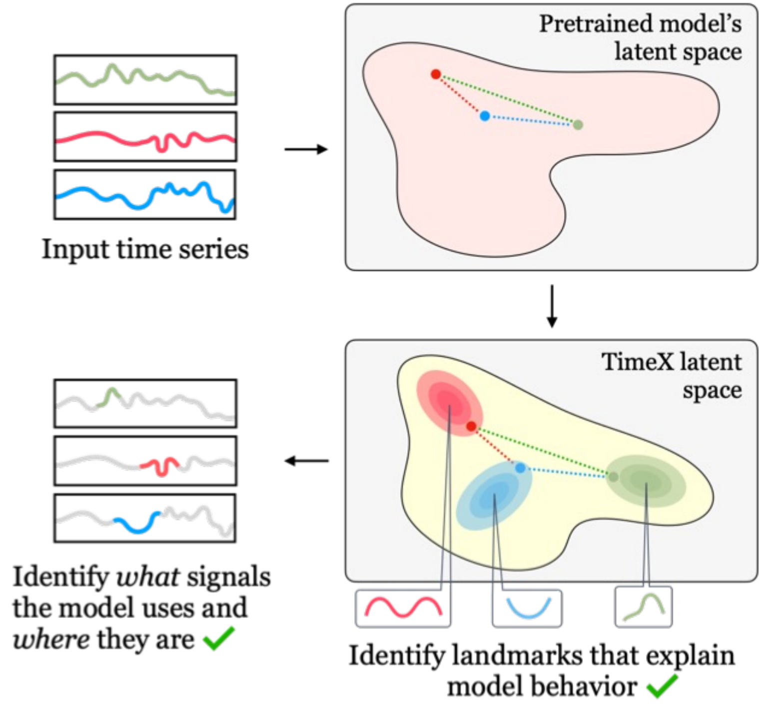
Dynamask, ICML 2021

Desiderata for time series explanations

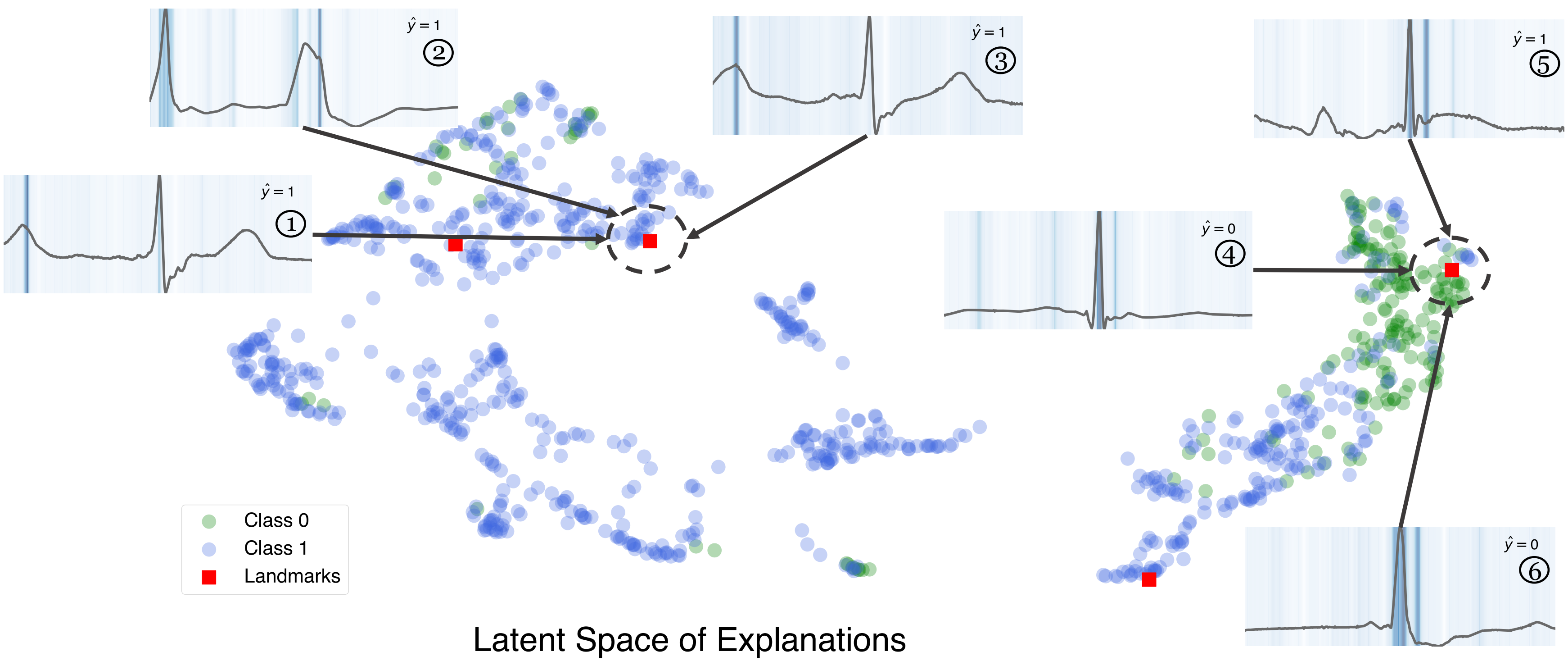
- Temporally connected and visually digestible
- Identify the location of predictive time series signals and underlying interpretable patterns
- Connect explanations across samples

TimeX is a time-series consistency explainer

- Surrogate model to **mimic the behavior of a pretrained time series model**
- TimeX makes inferences on masked samples
- **Model behavior consistency**
  - Enforces faithfulness at the level of the latent space
  - Learns a flexible latent space of explanations



Learned landmarks represent important patterns in physiological time series



Landmarks partition the latent space of explanations into interpretable temporal patterns

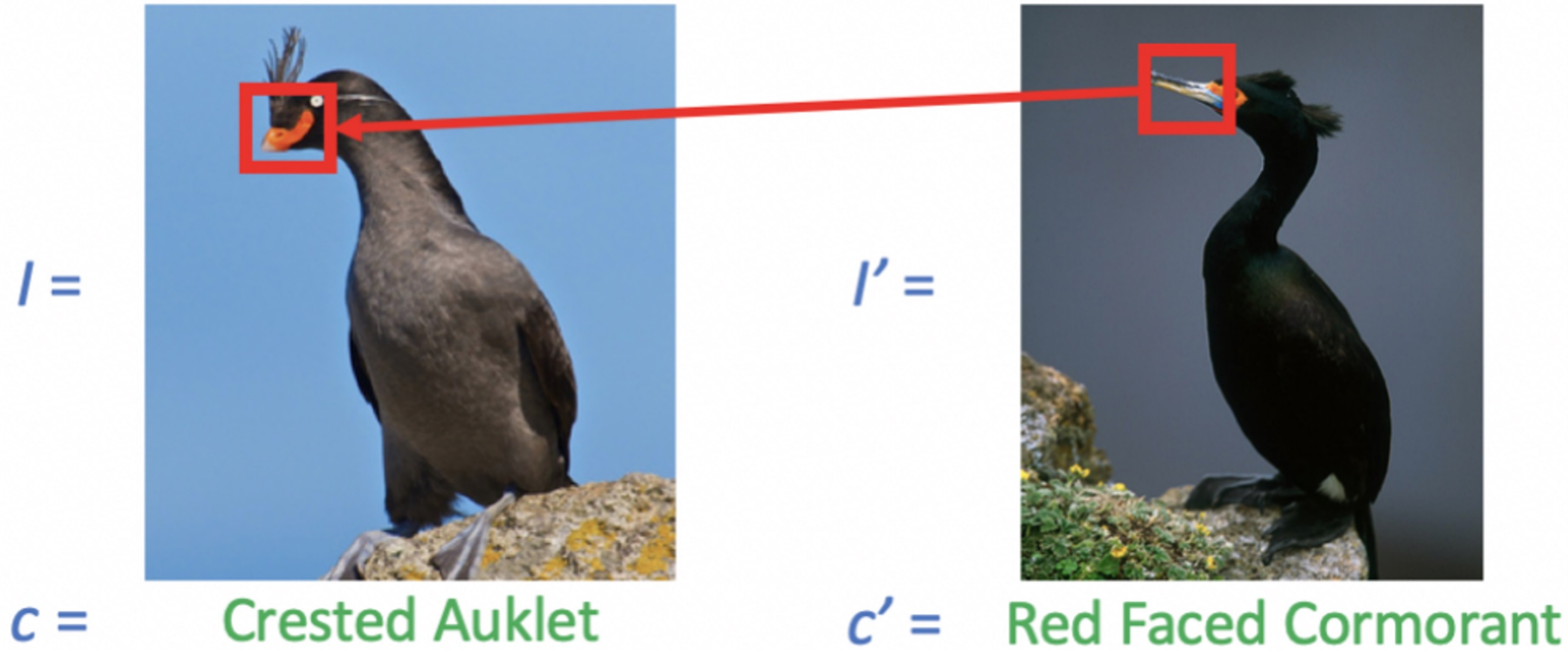
Overview of explanation methods

- Local explanation methods:
  - Feature importance scoring
  - Integrated gradients
  - Prototype explanations
  - <u>Counterfactuals</u>

- Global explanation methods:
  - Collection of local explanations
  - Representation-based explanations
  - Model distillation

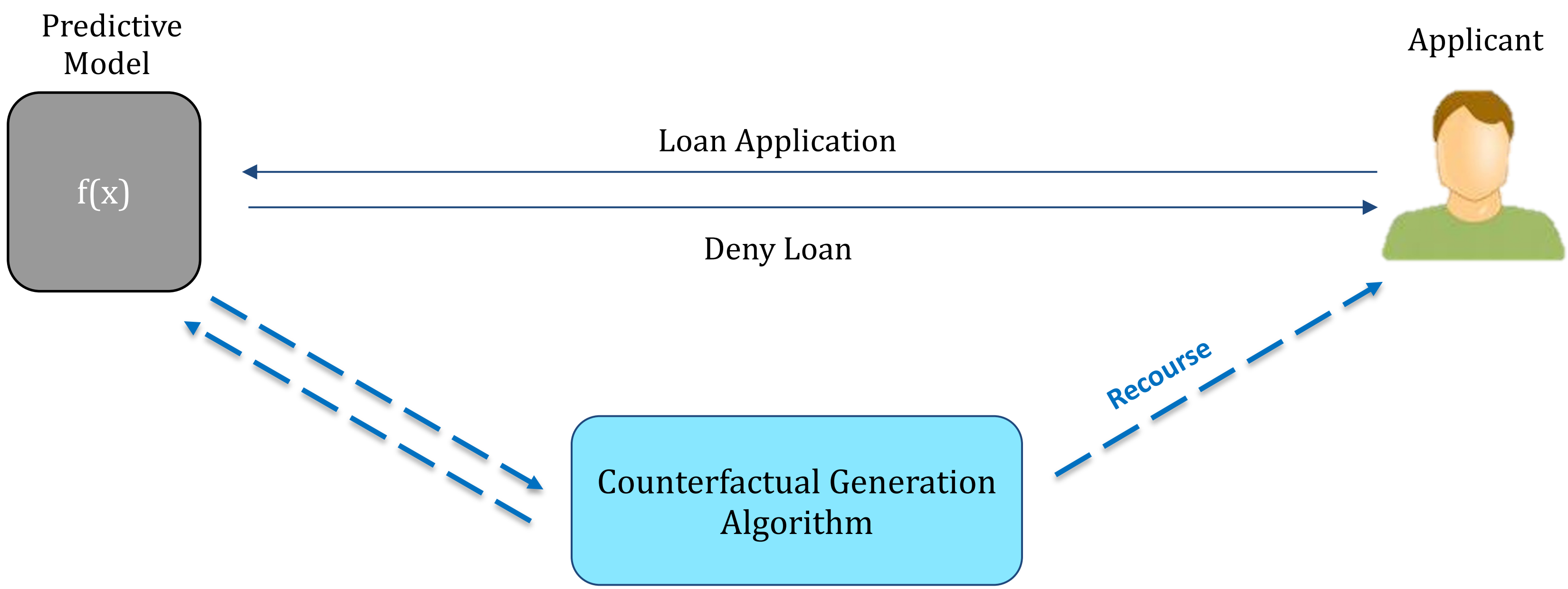
Counterfactual explanations

What features need to be changed and by how much to flip a model's prediction?



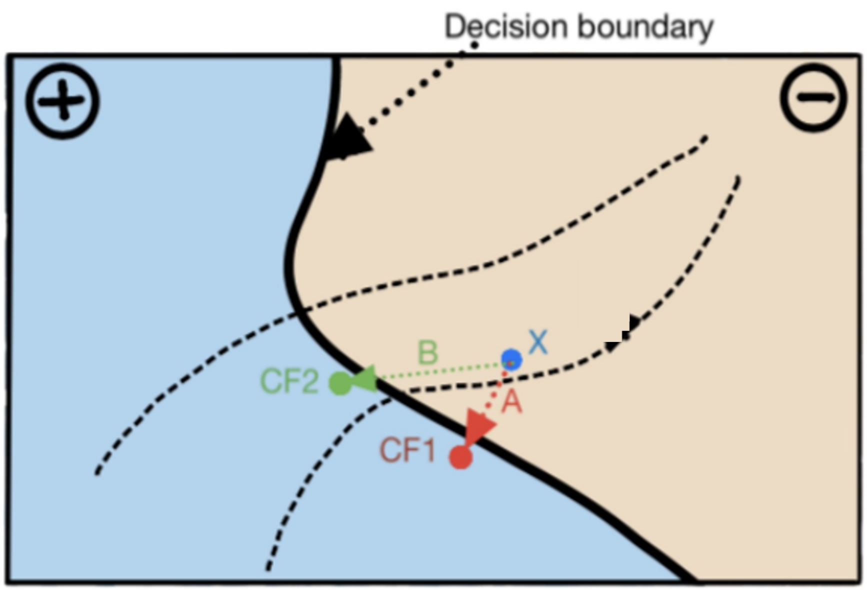
[Goyal et. al., 2019]

Counterfactual explanations



Recourse: Increase your salary by 50K & pay your credit card bills on time for next 3 months

Generating counterfactual explanations: Intuition



Proposed solutions differ on:

1. How to choose among candidate counterfactuals?

1. How much access is needed to the underlying predictive model?

[Verma et. al., 2020]

Quick Check

https://forms.gle/An2ZzQHbc568XAhe9

BMI 702: Biomedical Artificial Intelligence

Foundations of Biomedical Informatics II, Spring 2024

Quick check quiz for lecture 4: Interpretability and explainability in biomedical AI

Course website and slides: https://zitniklab.hms.harvard.edu/BMI702

Not shared

* Indicates required question

First and last name *

Your answer

Harvard email address *

Your answer

Describe a scenario in which a predictive model is created using a healthcare or biomedical dataset and the LIME explainability method is used to analyze its behavior. What can be expected from the LIME explanations? *

Your answer

Describe a scenario in which a predictive model is created using a healthcare or biomedical dataset and the Integrated Gradients explainability method is used to analyze its behavior. What can be expected from the Integrated Gradients explanations? *

Your answer

Submit

Clear form

Overview of explanation methods

- **Local explanation methods:**
  - Feature importance scoring
  - Integrated gradients
  - Prototype explanations
  - Counterfactuals

- **Global explanation methods:**
  - Collection of local explanations
  - Representation-based explanations
  - Model distillation

Global explanations from local feature importances: SP-LIME

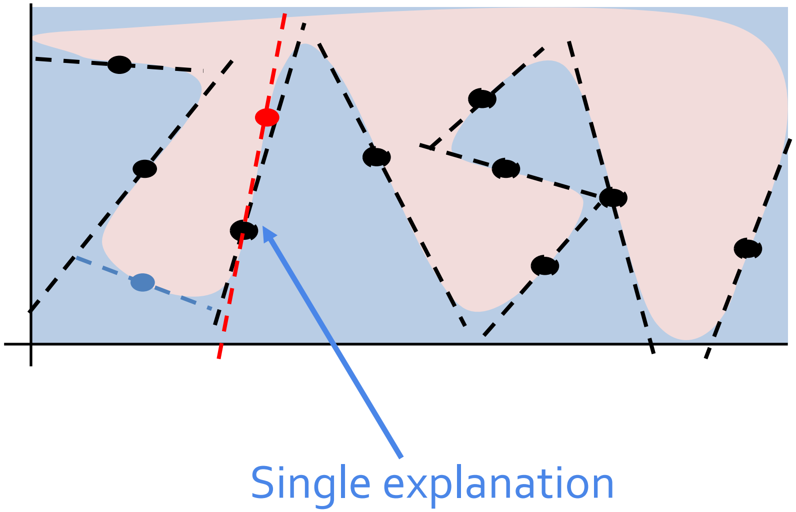
LIME explains a single prediction
local behavior for a single instance

Can't examine all explanations
Instead pick *k* explanations to show to the user

Representative
Should summarize the model's global behavior

Diverse
Should not be redundant in their descriptions

SP-LIME uses submodular optimization and *greedily* picks k explanations



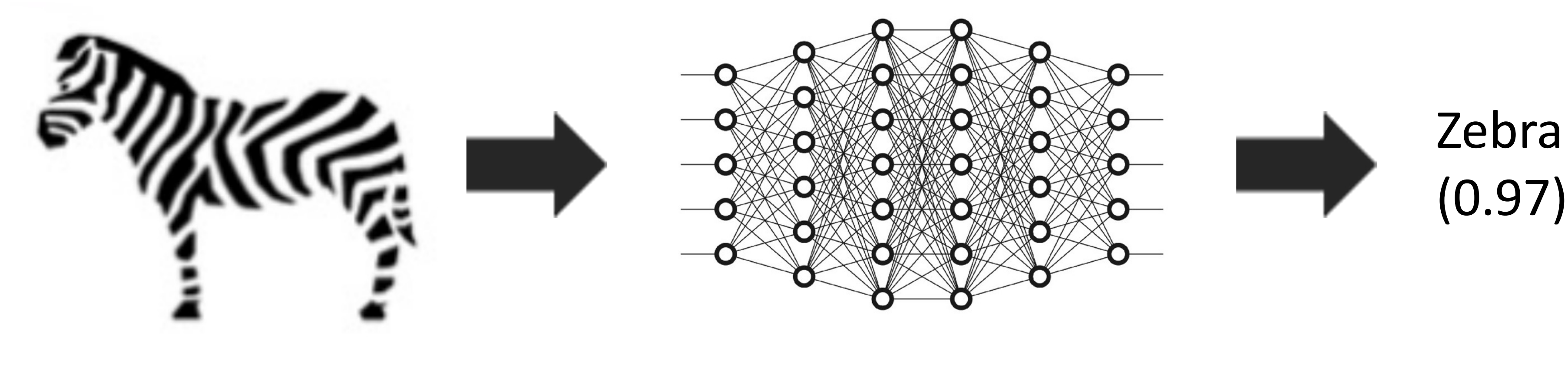
[Ribeiro et. al., 2016]

Overview of explanation methods

- Local explanation methods:
  - Feature importance scoring
  - Integrated gradients
  - Prototype explanations
  - Counterfactuals

- Global explanation methods:
  - Collection of local explanations
  - Representation-based explanations
  - Model distillation

Representation-based explanations



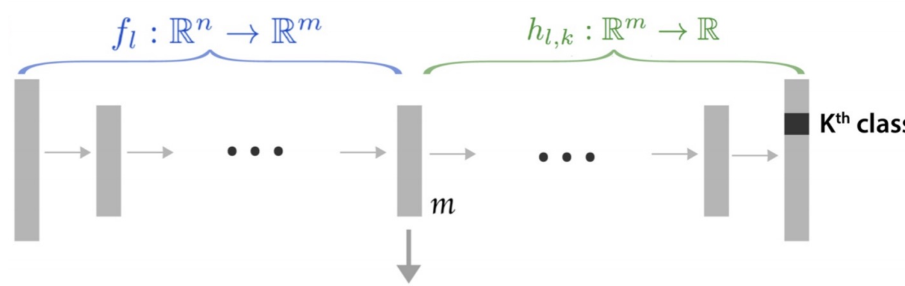
How important is the notion of “stripes” for this prediction?

[Kim et. al., 2018]

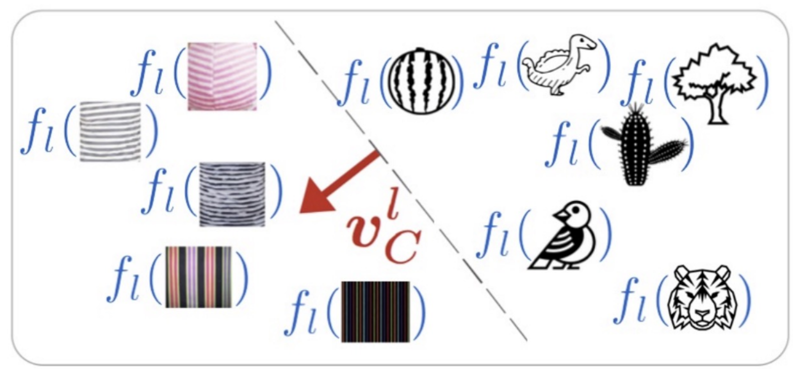
Representation-based explanations: TCAV approach

Examples of the concept "stripes"

Random examples



Train a linear classifier to separate activations



The vector orthogonal to the decision boundary denotes the concept "stripes"

Compute gradient w.r.t. this vector to determine how important is the notion of stripes for a prediction

[Kim et. al., 2018]

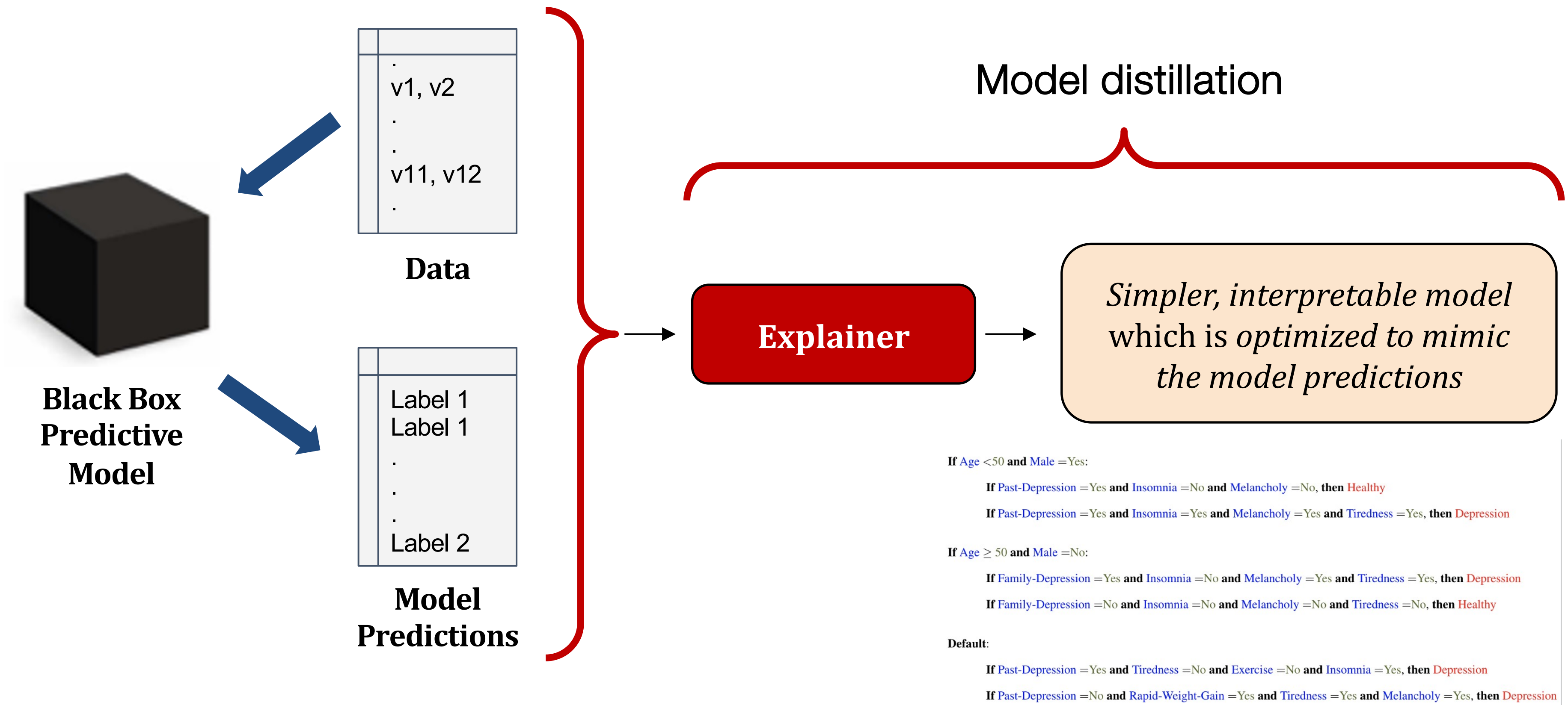
TCAV = testing with concept activation vectors

Overview of explanation methods

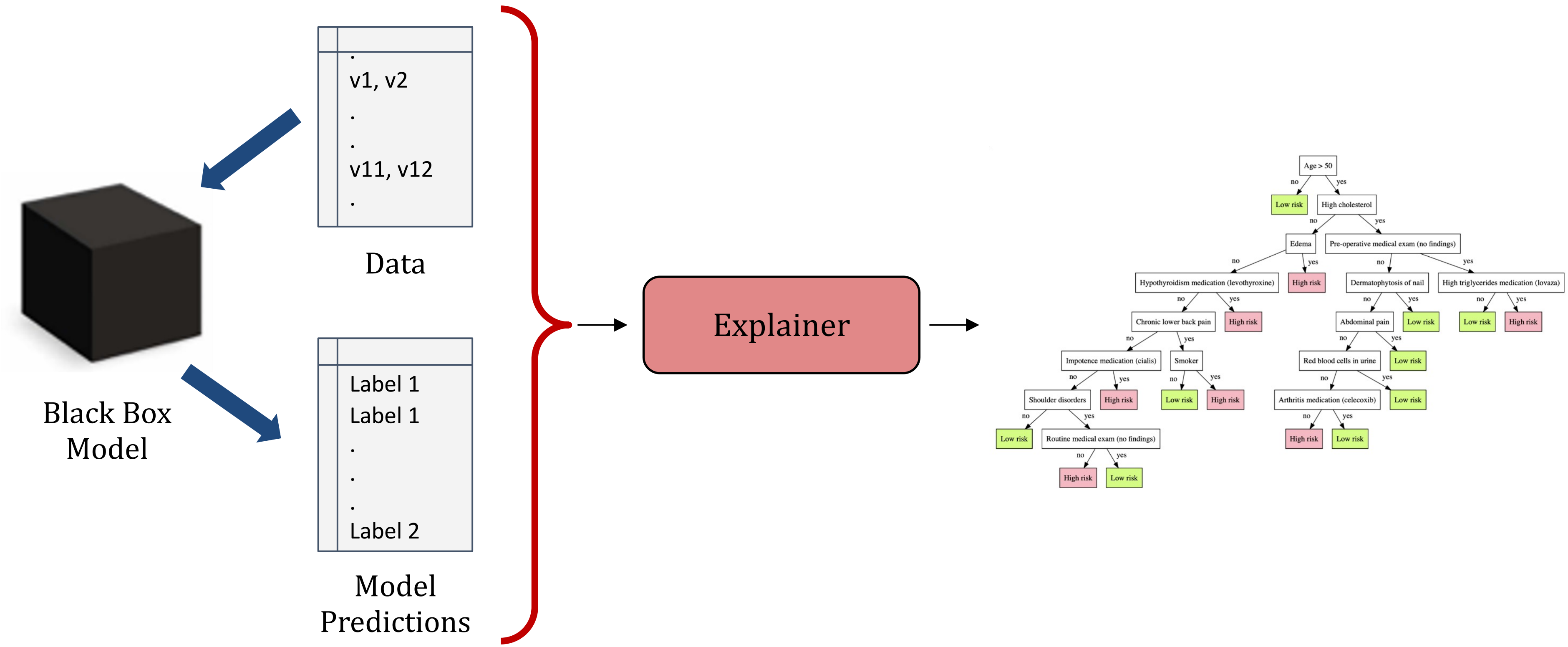
- Local explanation methods:
  - Feature importance scoring
  - Integrated gradients
  - Prototypes/Example-based explanations
  - Counterfactuals

- Global explanation methods:
  - Collection of local explanations
  - Representation-based explanations
  - <u>Model distillation</u>

Model distillation

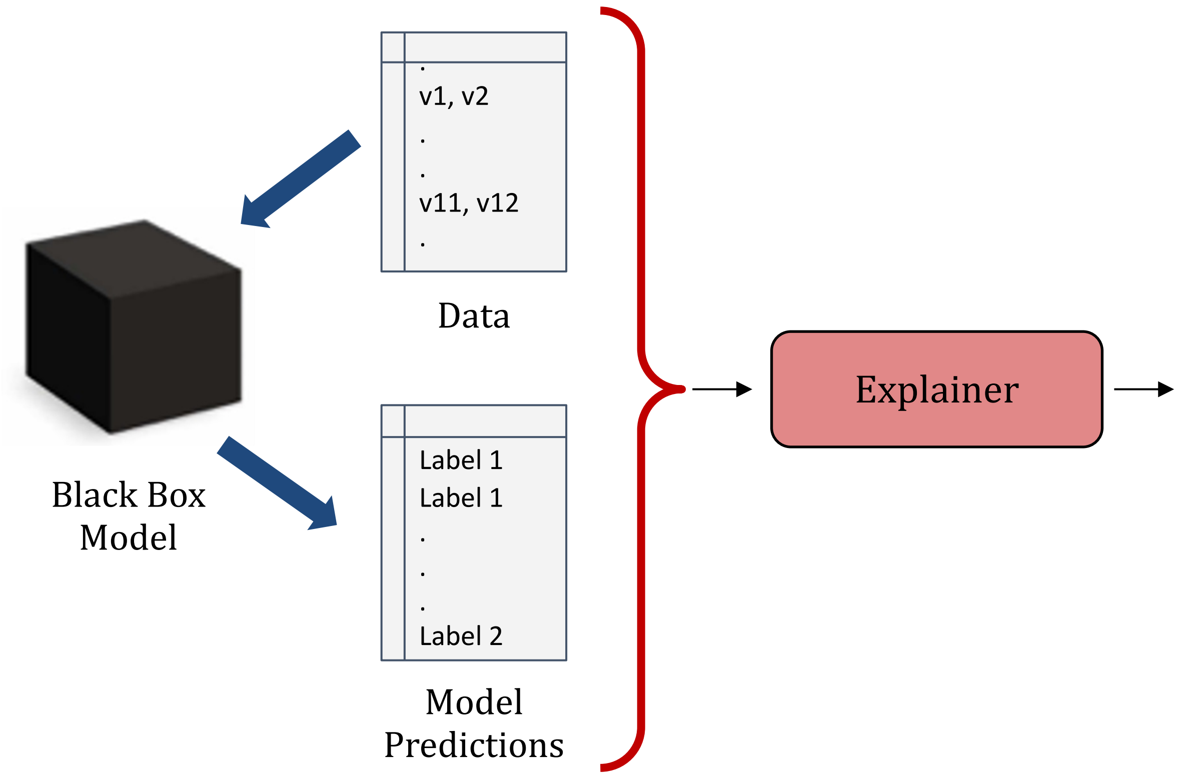


Model distillation using decision trees



[Bastani et. al., 2019]

Model distillation using decision sets



If Age <50 **and** Male =Yes:

If Past-Depression =Yes **and** Insomnia =No **and** Melancholy =No, **then** Healthy

If Past-Depression =Yes **and** Insomnia =Yes **and** Melancholy =Yes **and** Tiredness =Yes, **then** Depression

If Age ≥ 50 **and** Male =No:

If Family-Depression =Yes **and** Insomnia =No **and** Melancholy =Yes **and** Tiredness =Yes, **then** Depression

If Family-Depression =No **and** Insomnia =No **and** Melancholy =No **and** Tiredness =No, **then** Healthy

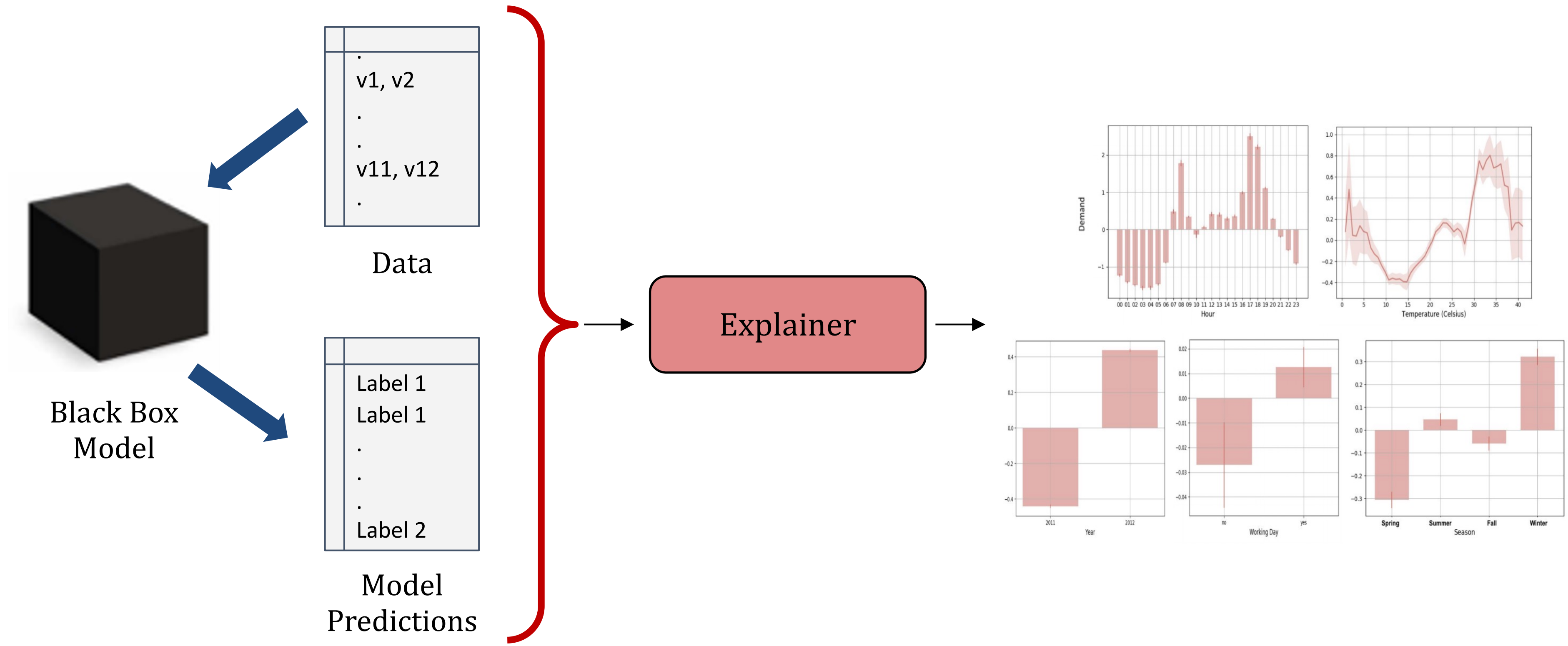
Default:

If Past-Depression =Yes **and** Tiredness =No **and** Exercise =No **and** Insomnia =Yes, **then** Depression

If Past-Depression =No **and** Rapid-Weight-Gain =Yes **and** Tiredness =Yes **and** Melancholy =Yes, **then** Depression

[Lakkaraju et. al., 2019]

Model distillation using generalized additive models



[Tan et. al., 2019]

Overview of explanation methods

- **Local explanation methods:**
  - Feature importance scoring
  - Integrated gradients
  - Prototype explanations
  - Counterfactuals

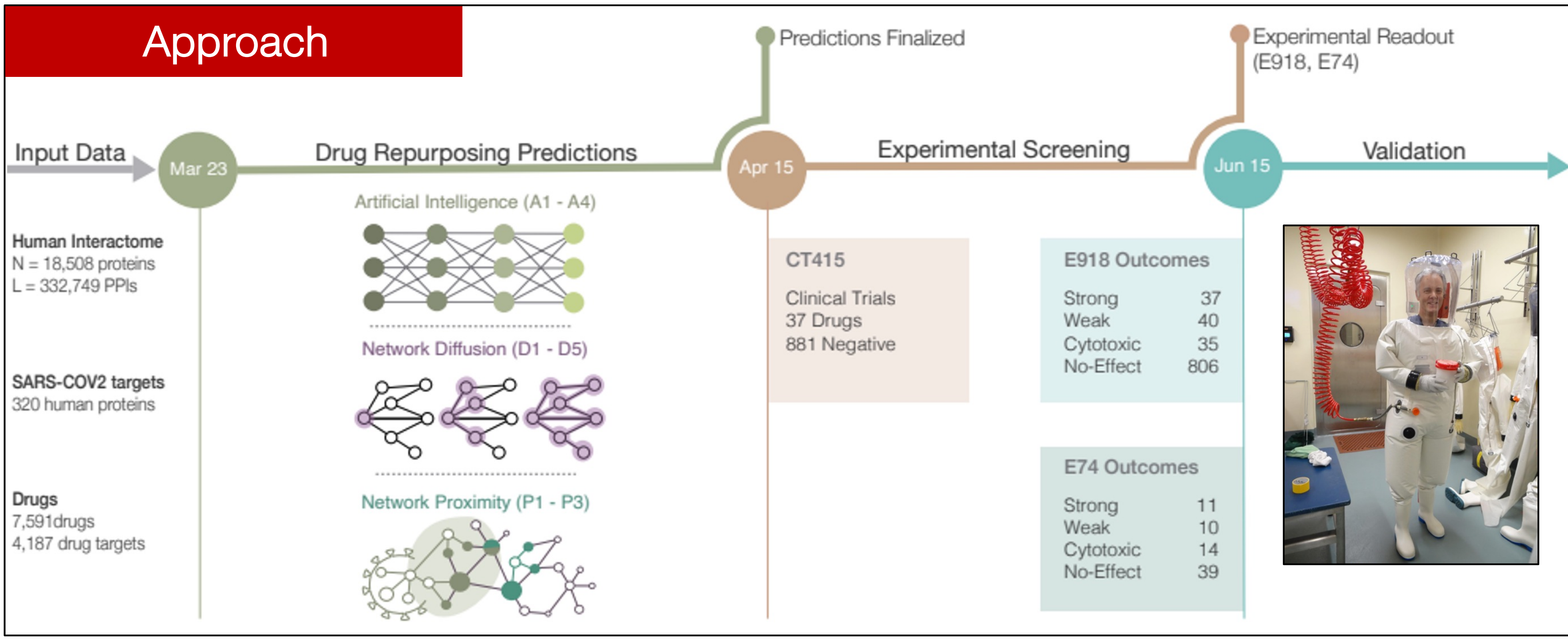
- **Global explanation methods:**
  - Collection of local explanations
  - Representation-based explanations
  - Model distillation

Outline for today's class

1. What is trustworthy AI/ML and why should I care?
2. Interpretability vs. explainability
3. Explaining AI/ML predictions
4. Case studies
   - Drug repurposing
   - Treatment recommendation

Rapid therapeutic innovation

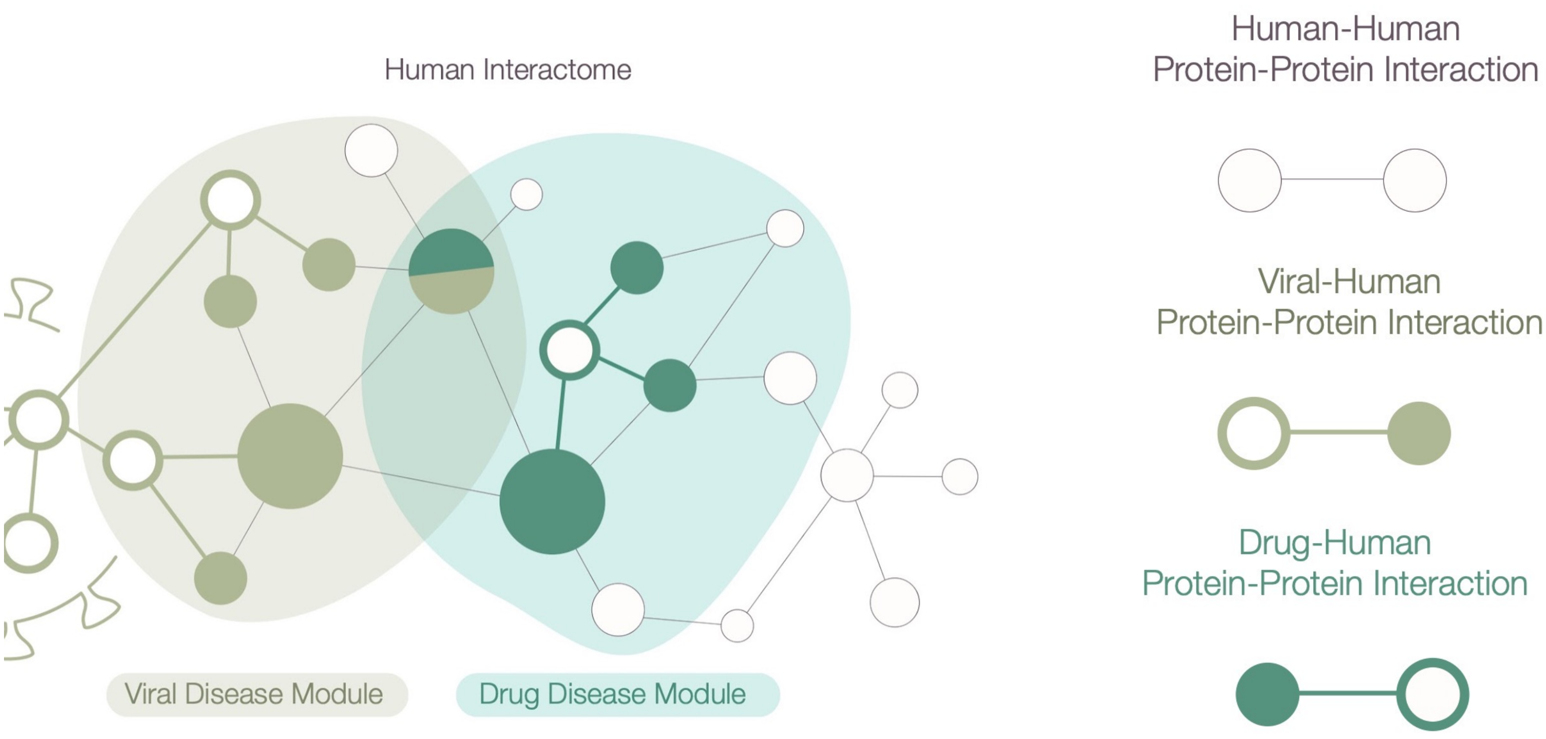
- Pandemics demand safe and effective therapies developed at an unprecedented speed
- Traditional, iterative development, experimental and clinical testing, and approval of new drugs not feasible
- **Challenge:** How to compress years of work into months or even weeks through AI, automation, and new data resources?



Network Medicine Framework for Identifying Drug Repurposing Opportunities for Covid-19, *PNAS*, 2021

Design therapies to target biological networks

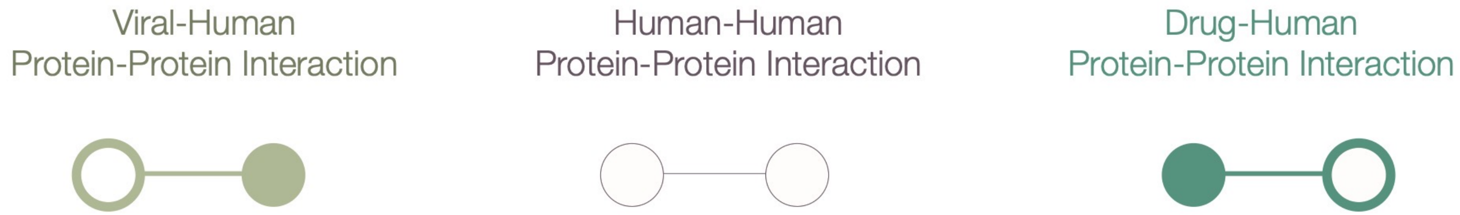
Disease disrupts the normal behavior of genes. Drugs intervene against the disease by restoring the function of disrupted genes. **Goal:** What chemical compounds can intervene against disease?



Network Medicine Framework for Identifying Drug Repurposing Opportunities for Covid-19, *PNAS*, 2021
Zero-shot prediction of therapeutic use with geometric deep learning and clinician centered design, medRxiv 2023

Dataset and experimental setup

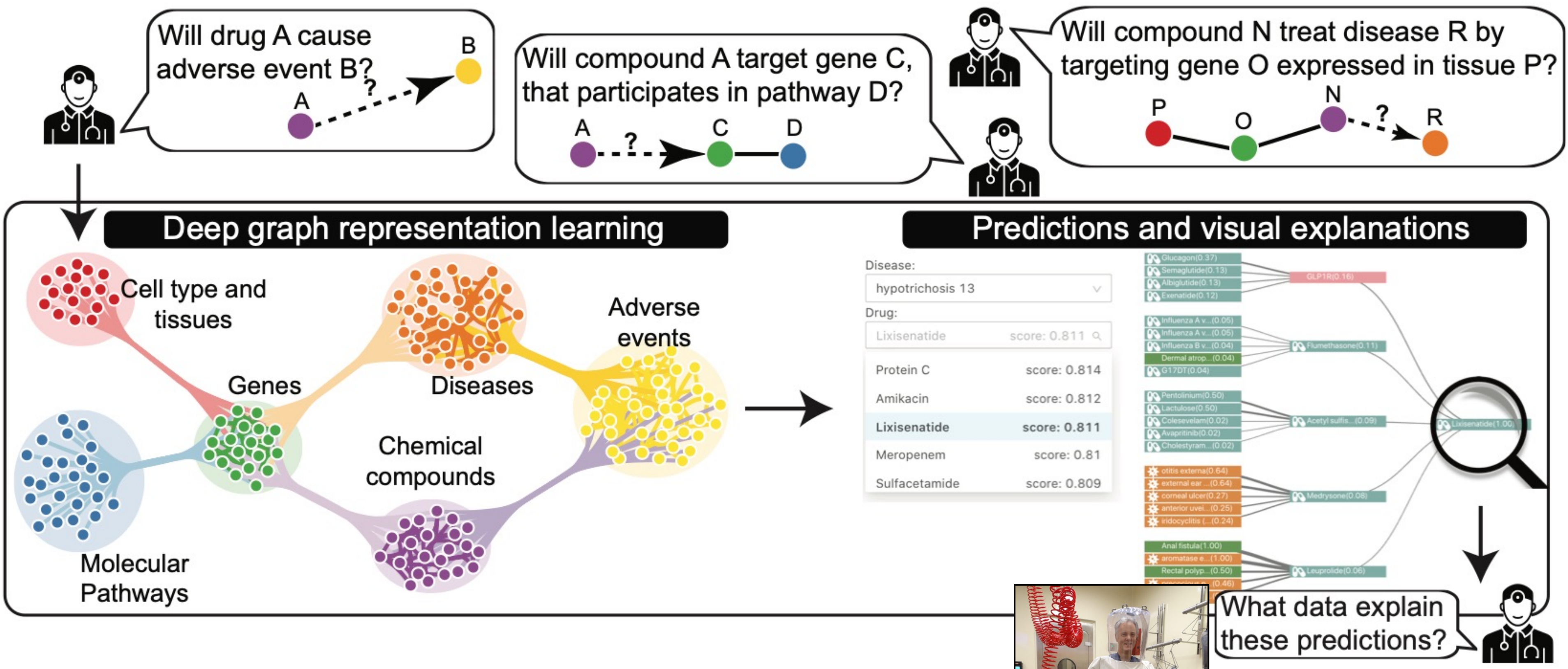
- **COVID-19 repurposing knowledge graph:**
  - Human protein-protein interaction graph
  - Approved drugs and proteins that each drug targets
  - Diseases and proteins perturbed in each disease
  - Approved drug-disease treatments



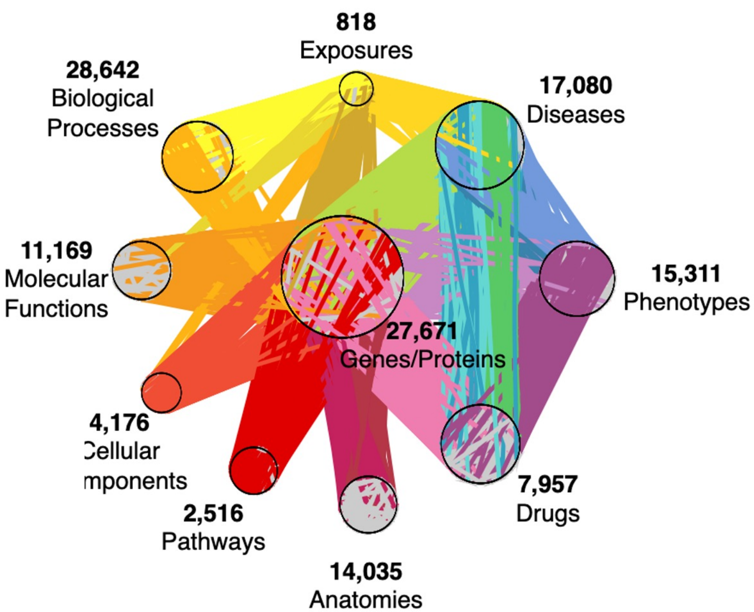
- **ML task**: Given approved drug-disease treatments, identify candidate treatments for COVID-19

Network Medicine Framework for Identifying Drug Repurposing Opportunities for Covid-19, *PNAS*, 2021

Approach: Graph ML model

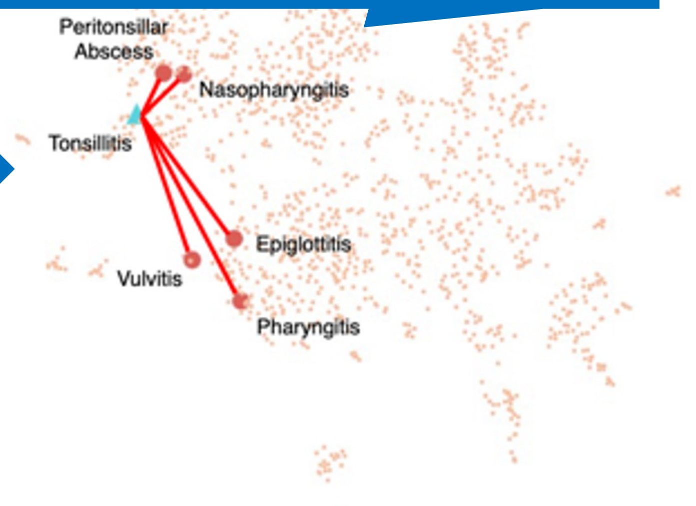


Approach: Graph ML model

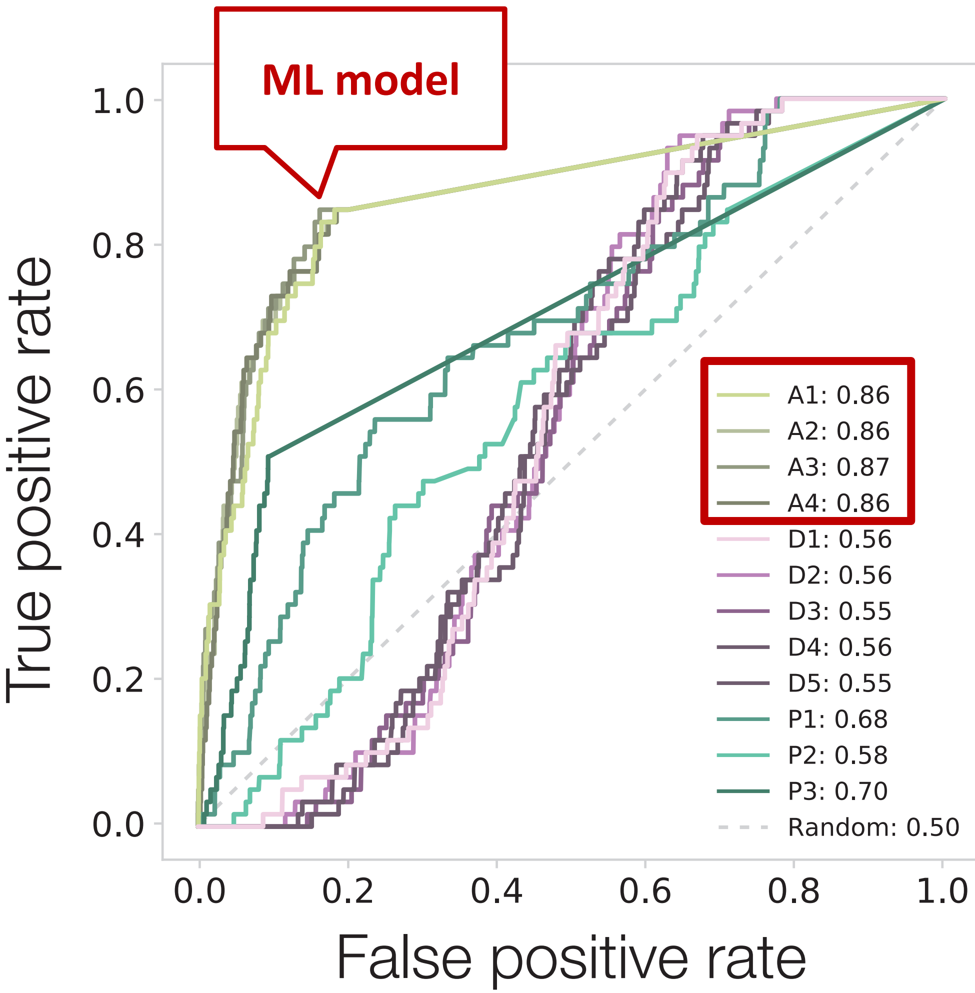


Model

Goal: What compounds/drugs can intervene against (treat) disease?



Results: COVID-19 repurposing



We test each method's ability to recover drugs currently in clinical trials for COVID-19 (67 drugs from ClinicalTrials.gov)

The best individual ROC curves are obtained by the GNN methods

The second-best performance is provided by the proximity P3. Close behind is P1 with AUC = 0.68 and AUC = 0.58

Diffusion methods offer ROC between 0.55-0.56

Network Medicine Framework for Identifying Drug Repurposing Opportunities for Covid-19, *PNAS*, 2021

Results: Experimental screening

National Emerging Infectious Diseases Laboratories (NEIDL)

CRank	Drug Name
1	Ritonavir
2	Isoniazid
3	Troleandomycin
4	Cilostazol
5	Chloroquine
6	Rifabutin
7	Flutamide
8	Dexamethasone
9	Rifaximin
10	Azelastine
11	Crizotinib

17	Celecoxib
18	Betamethasone
19	Prednisolone
20	Mifepristone
21	Budesonide
22	Prednisone
23	Oxiconazole
24	Megestrol acetate
25	Idelalisib
26	Econazole

Predicted lists of drugs

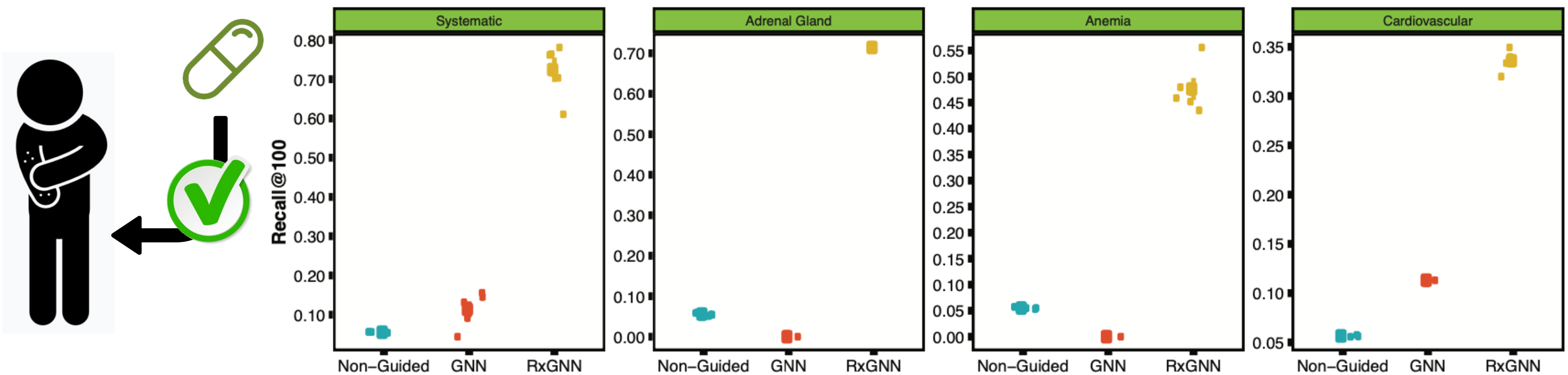
New algorithms:
Prioritizing Network Communities, *Nature Communications* 2018
Subgraph Neural Networks, *NeurIPS* 2020
Graph Meta Learning via Local Subgraphs, *NeurIPS* 2020

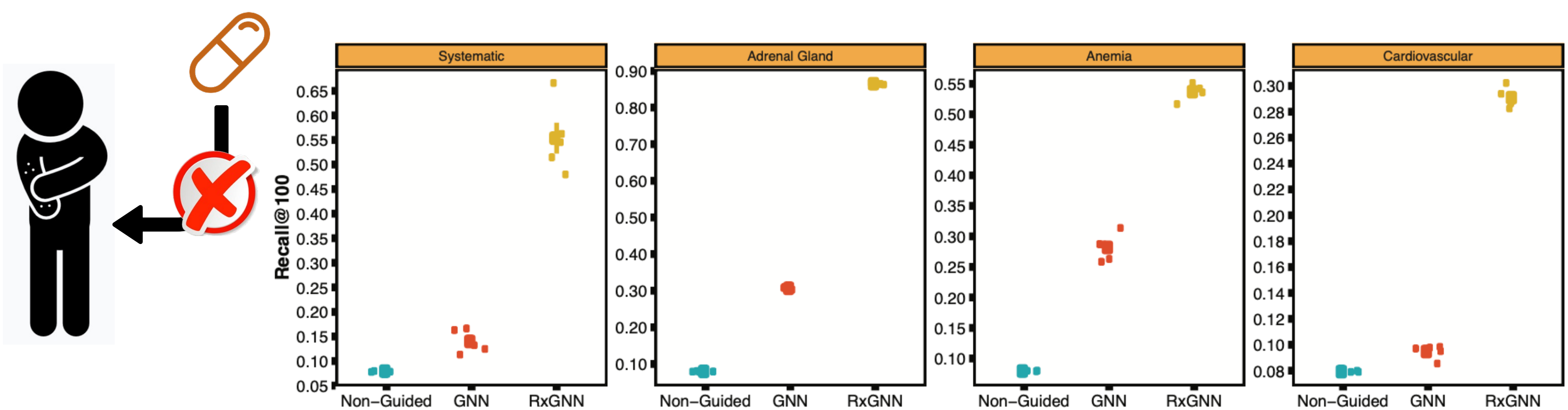
Results: 918 compounds screened for their efficacy against SARS-CoV-2 in VeroE6 & human cells:

- We screened in human cells the top-ranked drugs, obtaining a 62% success rate, in contrast to the 0.8% hit rate of nonguided screenings
- This is an **order of magnitude higher hit rate** among top 100 drugs than **alternative approach**

Network Medicine Framework for Identifying Drug Repurposing Opportunities for Covid-19, *PNAS*, 2021

Results: Predicting therapeutic use

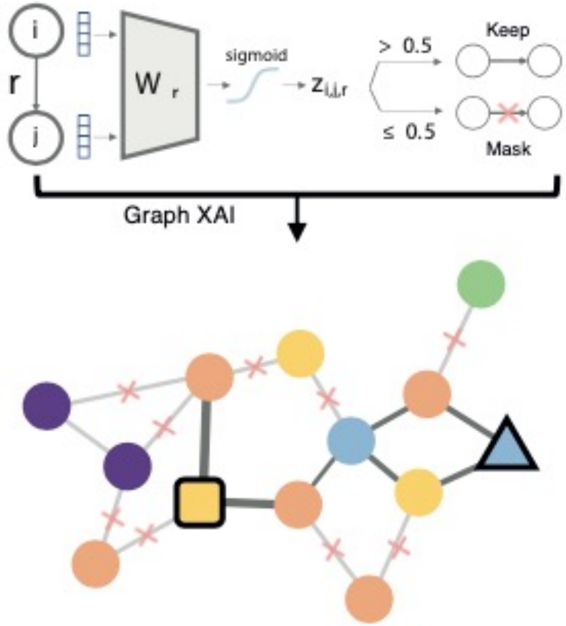


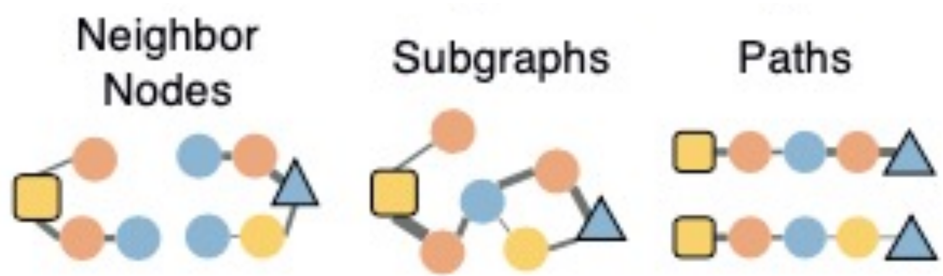


Explaining model predictions

Key idea:

- Summarize where in the data the model "looks" for evidence for its prediction
- Find a small subgraph most influential for the prediction

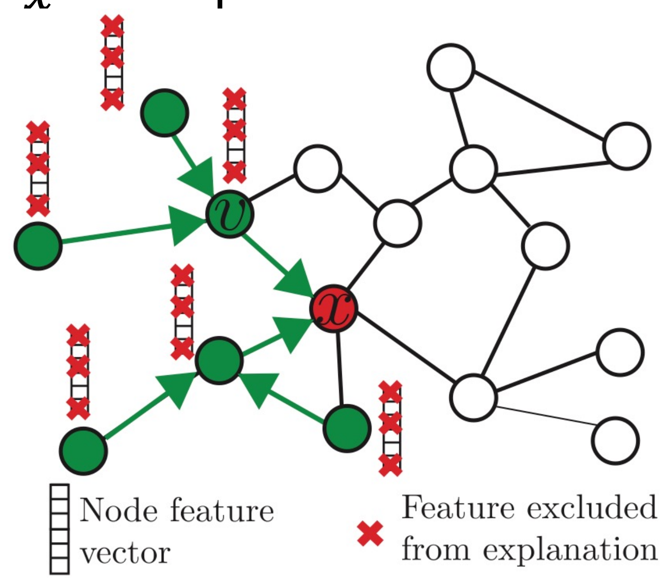




Interpreting Graph Neural Networks for NLP With Differentiable Edge Masking, *ICLR* 2021
GNNExplainer: Generating Explanations for Graph Neural Networks, *NeurIPS* 2019

GNNExplainer: Key idea

- Input: Given prediction $f(x)$ for node/link x
- Output: Explanation, a small subgraph M_x together with a small subset of node features:
  - M_x is most influential for prediction $f(x)$
- Approach: Optimize mask M_x in a post-hoc manner
  - Intuition: If removing v from the graph strongly decreases the probability of prediction $\Rightarrow$ v is a good counterfactual explanation for the prediction



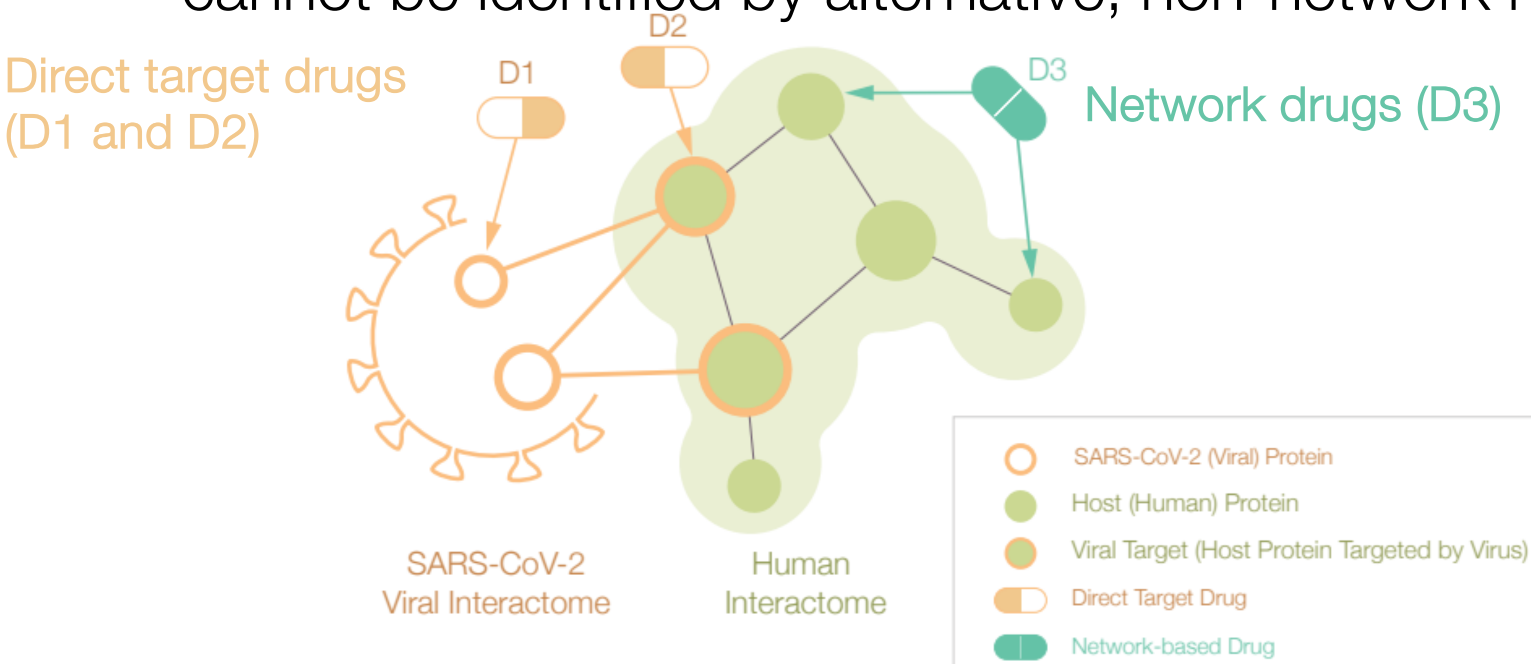
GNN Explainer: Generating Explanations for Graph Neural Networks, *NeurIPS* 2019

Explanations: Network drugs

"What is the disease treatment mechanism for drugs with **positive experimental outcomes**?"

76/77 predicted drugs with positive experimental outcomes <u>do not directly bind</u> to SARS-CoV-2 targets:

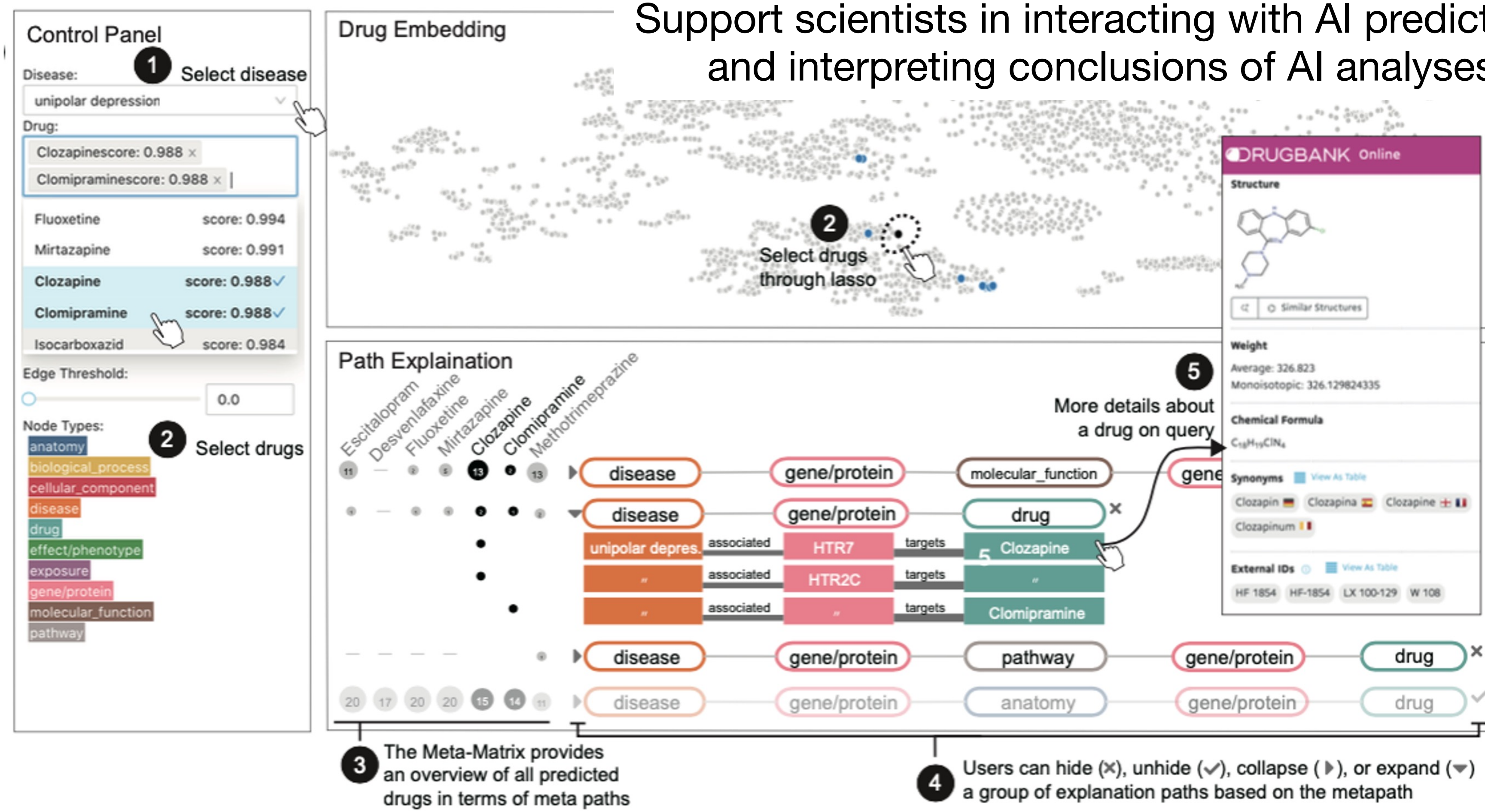
- Instead, the drugs rely on **network-based actions** and cannot be identified by alternative, non-network methods



AI-clinician collaboration

”Will clozapine treat unipolar depression? What is the disease treatment mechanism?”

Support scientists in interacting with AI predictions and interpreting conclusions of AI analyses



Probing GNN Explainers: A Rigorous Theoretical and Empirical Analysis of GNN Explanation Methods, *AISTATS* 2022

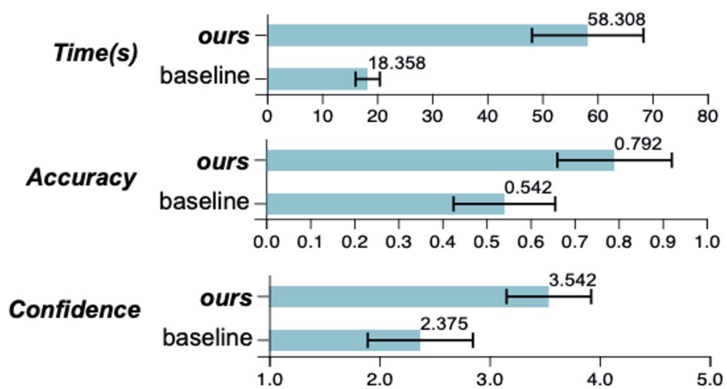
Extending the Nested Model for User-Centric XAI: A Design Study on GNN-based Drug Repurposing, *IEEE VIS* 2022 (Best Paper Award)

Identification of Disease Treatment Mechanisms through the Multiscale Interactome, Nature Communications 2021

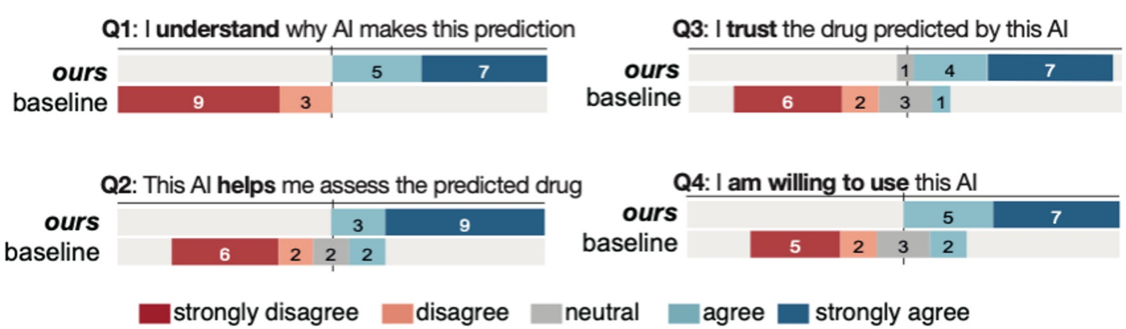
Clinician-centric study

Compared to a no-explanation baseline in terms of **user answer accuracy, exploration time, user confidence**, and **user agreement** across a spectrum of usability questions

Mount Sinai



Error bars indicate the 95% confidence intervals



Agree scores are placed to the right, disagree to the left

Extending the Nested Model for User-Centric XAI: A Design Study on GNN-based Drug Repurposing, *IEEE VIS* 2022 (**Best Paper Award**)

Practical and ethical challenges

Q: Are decision-makers benefitting from explanations?
A: (Mixed) evidence of real-world benefit

Q: How are explanations calibrating trust in AI?
A: Explanations can be used to manipulate & miscalibrate trust

Q: How are explanations calibrating perceptions of fairness?
A: Explanations can be used to change fairness perceptions

Q: Can adversaries fool explanation algorithms & hence users?
A: Adversaries can easily obfuscate true model behavior

Outline for today's class

1. What is trustworthy AI/ML and why should I care?
2. Interpretability vs. explainability
3. Explaining AI/ML predictions
4. Case studies
   - Drug repurposing
   - Treatment recommendation