Interpretability and Explainability: Framework and Methods

Finale Doshi-Velez
Harvard University

Why Interpretability?

Project in our lab: What antidepressants work best for which patients?

patient data → prozac works

Why Interpretability?

Project in our lab: What antidepressants work best for which patients?

patient data → prozac works

pregancy, UTIs → prozac works

Why Interpretability?

Project in our lab: What antidepressants work best for which patients?

patient data → prozac works

pregancy, UTIs → prozac works

female biology → prozac works

Why Interpretability?

Project in our lab: What antidepressants work best for which patients?

patient data → prozac works

pregancy, UTIs → prozac works

female biology → prozac works

OBs/GYNs → prozac prescribed, happens to work

Why Interpretability?

Project in our lab: What antidepressants work best for which patients?

patient data → prozac works

pregancy, UTIs → prozac works

female biology → prozac works

OBs/GYNs → prozac prescribed, happens to work

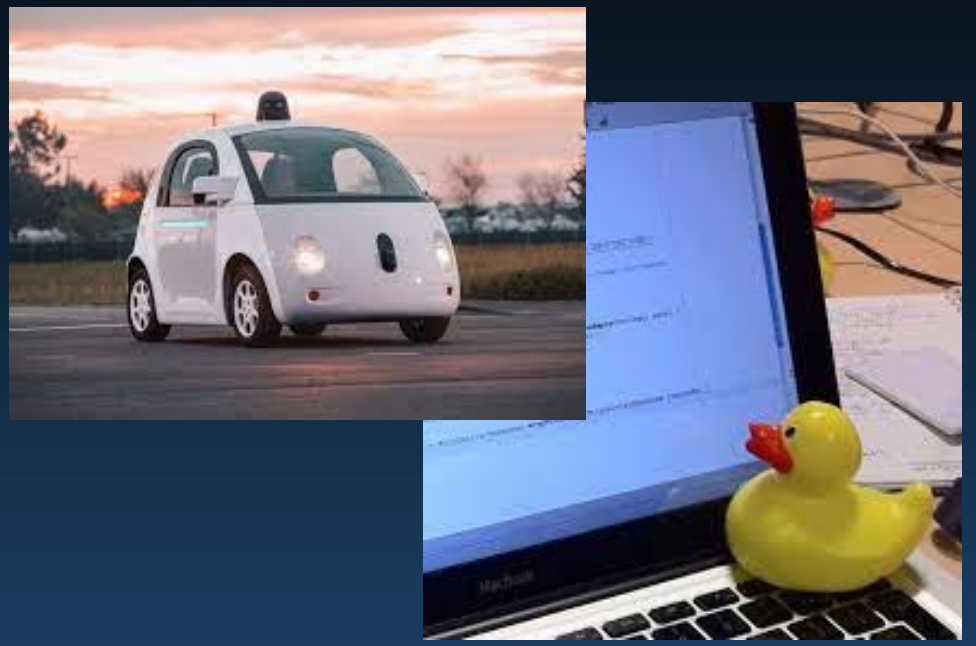
Interpretation was needed to draw the correct insight

What are other reasons that we may desire explanations?

Insight

Choosing what meds fit your needs; advancing scientific understanding

Oversight



Improving safety, debugging systems

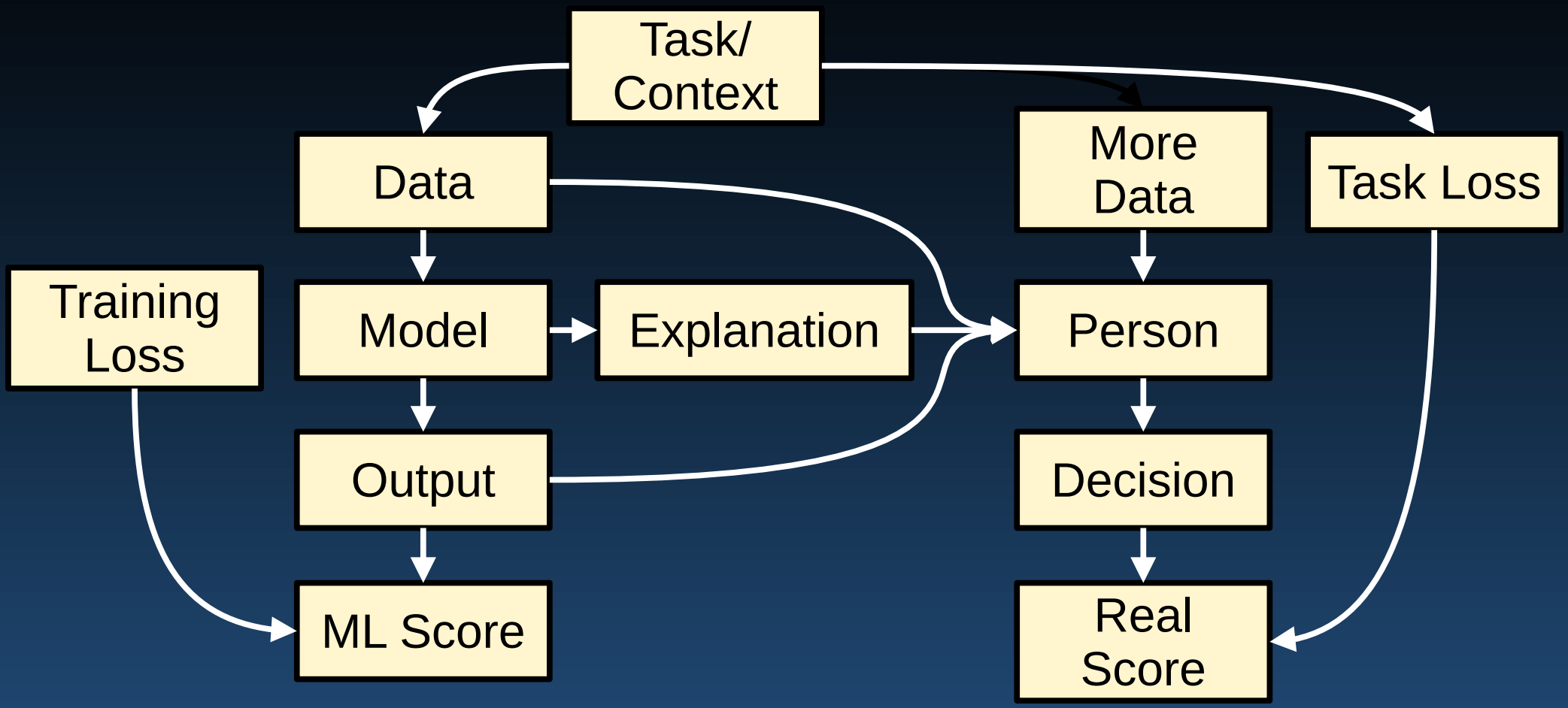
Common idea: the problem cannot be specified and solved computationally

Think/Chat Exercise: Explanation Needed?

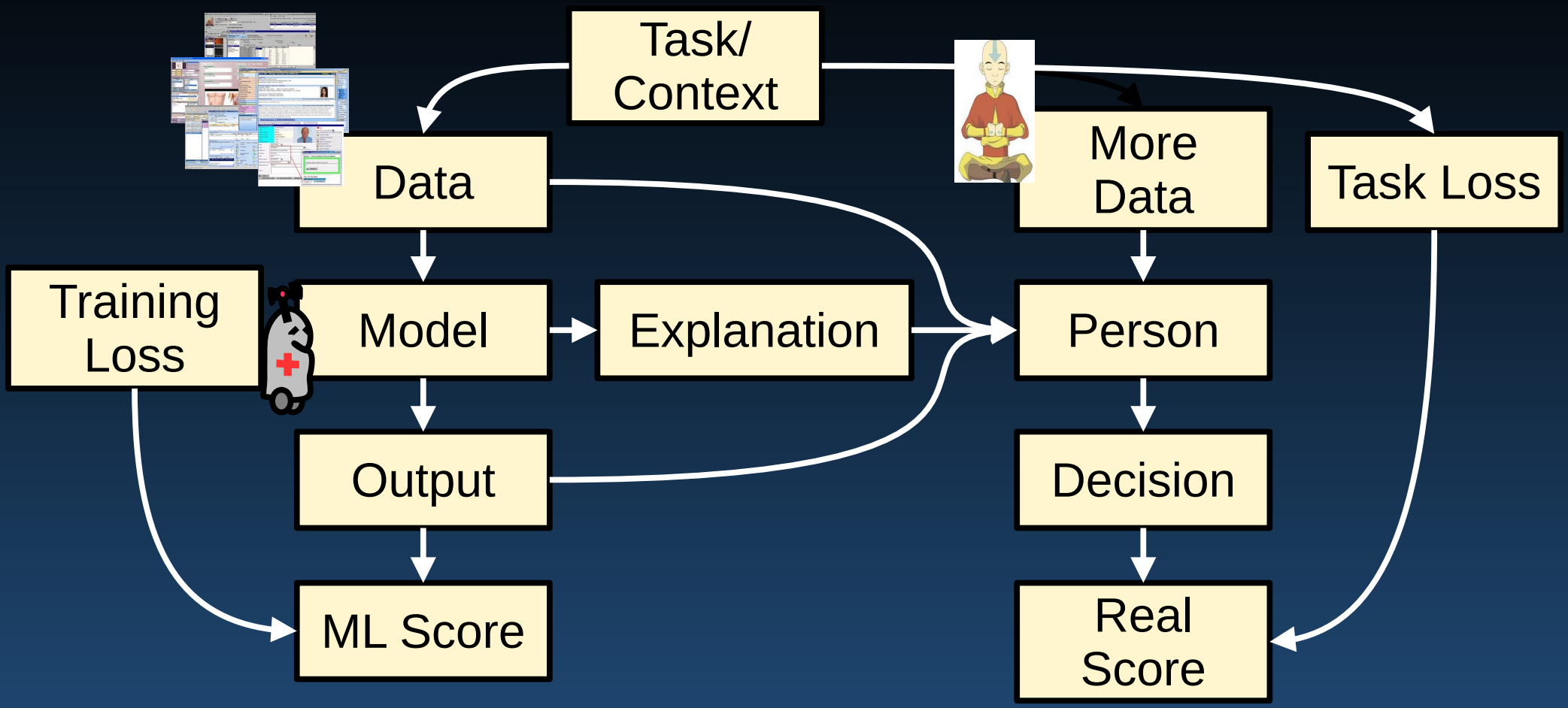
For the following cases – would an explanation be helpful? If so, for whom and for what?

1. A social media content ranking system deployed across millions of users.
2. A smartwatch-based atrial fibrillation algorithm.
3. A negligence-detection algorithm to determine whether a child is at-risk.
4. A smart grid system to minimize energy usage and prevent brown-outs on hot days.

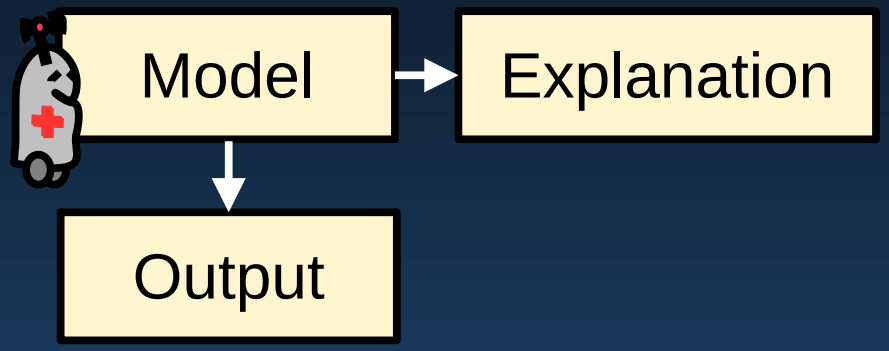
The Interpretable ML Ecosystem



The Interpretable ML Ecosystem



Rationales, Reasonings, Explanations



Rationales, Reasonings, Explanations

solve 2x = 6, explain why

To solve the equation:

$$2x = 6$$

✅ **Step-by-step solution:**

We want to **isolate** x. Currently, it's being **multiplied** by 2.
To undo multiplication, we use **division** (the inverse operation).

$$\frac{2x}{2} = \frac{6}{2}$$

$$x = 3$$

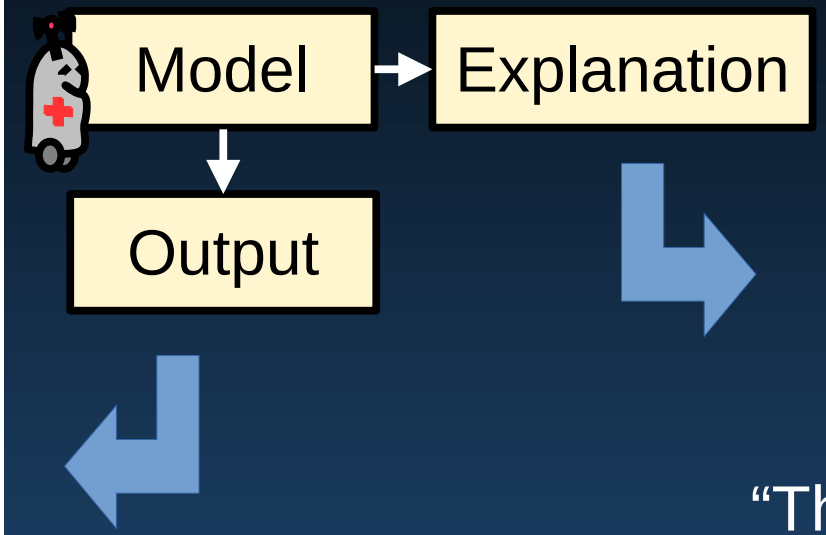
🧠 **Why this works:**

An equation is like a balance scale — what you do to one side, you must do to the other to keep it balanced.
Since $2x = 6$ means "2 times something equals 6", we divide both sides by 2 to find out what that "something" is.

✔ **Final Answer:**

$$x = 3$$

In this case, lets us check, but doesn't reveal anything about the model internals.



"The model divided 6 by 2 to get the answer."

Tells us exactly how the model performed the computation.

Rationales, Reasonings, Explanations

solve 2x = 6, explain why

To solve the equation:

$$2x = 6$$

✅ **Step-by-step solution:**

We want to **isolate** x. Currently, it's being **multiplied** by 2.
To undo multiplication, we use **division** (the inverse operation).

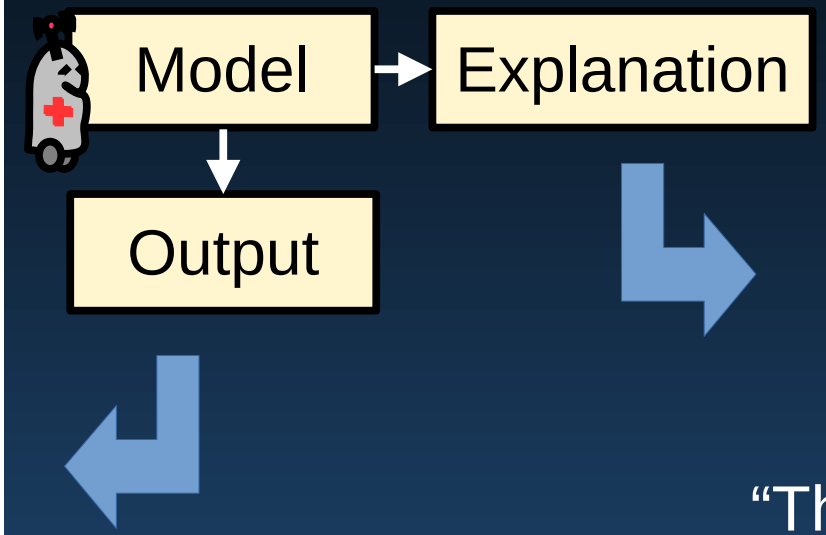
$$\frac{2x}{2} = \frac{6}{2}$$

$$x = 3$$

🧠 **Why this works:**

An equation is like a balance scale — what you do to one side, you must do to the other to keep it balanced.
Since $2x = 6$ means "2 times something equals 6", we divide both sides by 2 to find out what that "something" is.

✔ **Final Answer:**



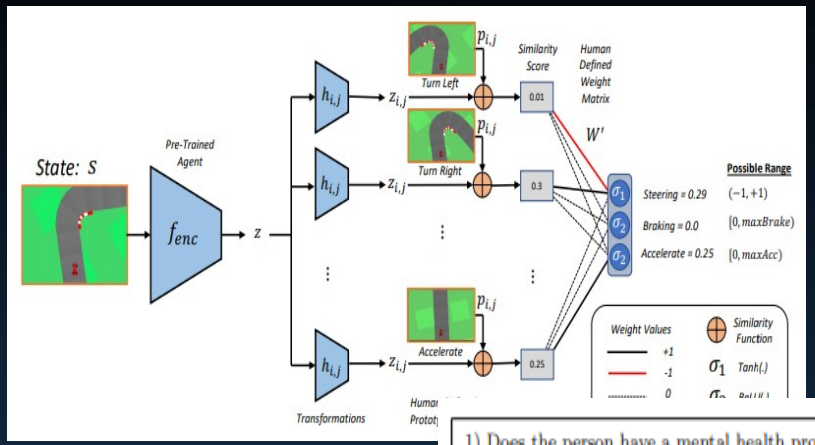
"The model divided 6 by 2 to get the answer."

For this collection of modules, we're going to focus on explanations that reveal model internals.

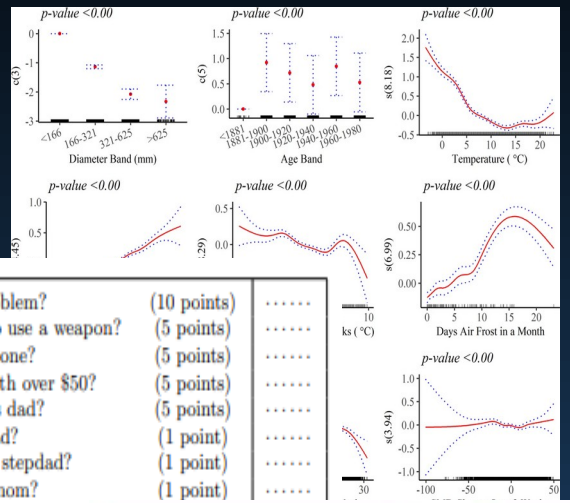
So what can explanations look like?

So what can explanations look like? LOTS of choices

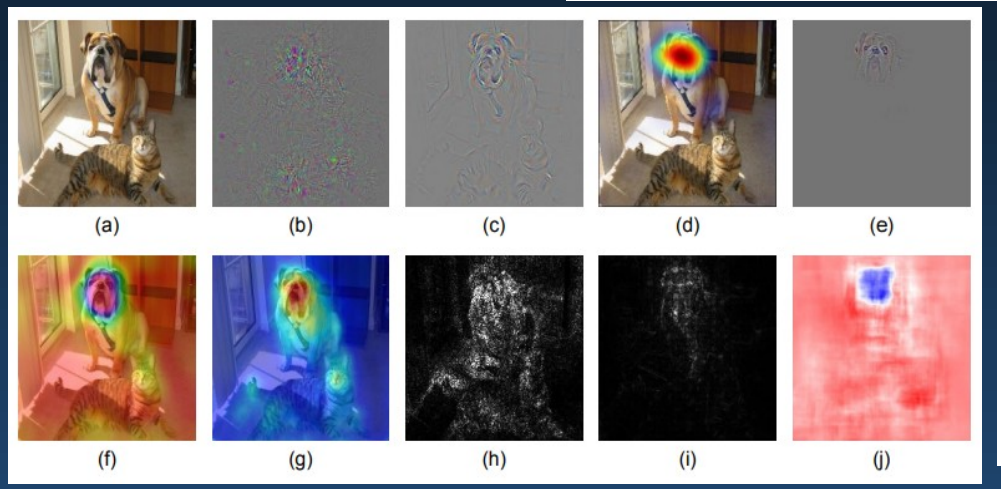
PWNet
Kenny
et al. 2023



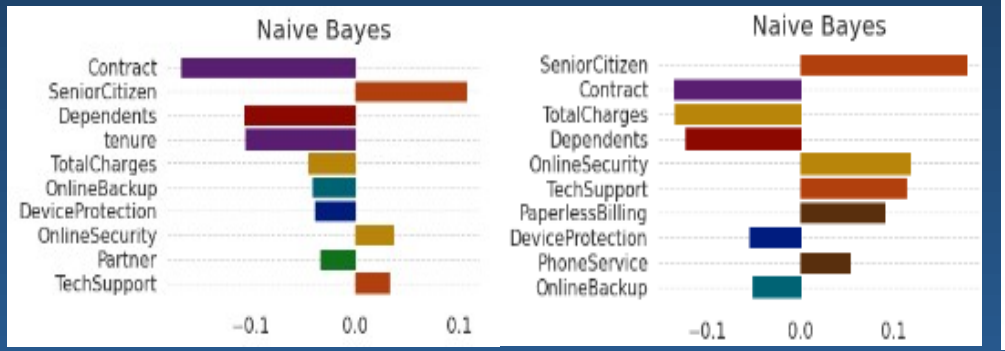
GAMs: Barton 2022

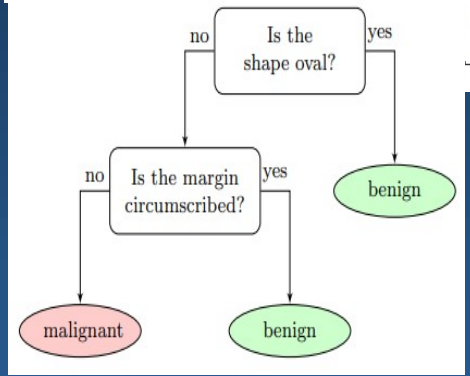


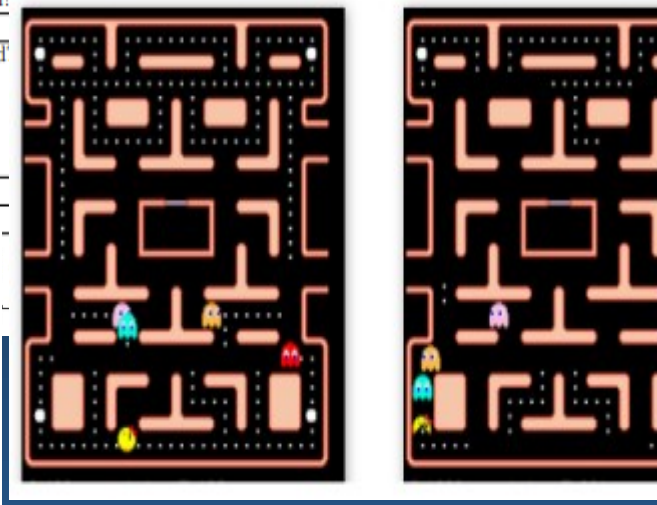
Survey: Abhishek and Kamath, 2022



1) Does the person have a mental health problem?	(10 points)
2) Has the person ever used or threatened to use a weapon?	(5 points)
3) Has the person ever shot or stabbed someone?	(5 points)
4) Has the person ever stolen something worth over $50?	(5 points)
5) Is the person male and distanced from his dad?	(5 points)
6) Does the person not have a dad or stepdad?	(1 point)
7) Is the person male and not have a dad or stepdad?	(1 point)
8) Does the person not have a mom or stepmom?	(1 point)
9) Is the person male and not have a mom or stepmom?	
Sum points from 1 to 9	
10) Is the person female and not have a dad or stepdad	
11) Does the person have college plans?	
12) Is the person employed?	
13) Is the person in school and employed?	
14) Likelihood to use child welfare system.	
Sum points from 10 to 14	
Subtract Total B from Total A	







Highlights: Amir and Amir 2018

Cloudera Blog, 2020

SLIM: Ustun et al. 2014

Guiding Principle: The right form of interpretability depends on context.

Patient Details:
Susan is a 31 year old woman who is single and works part time. She has a history of diabetes, arrhythmia and hypertensive heart disease. She presents with 14 months of depressed mood. Current medications include amoxicillin, and prior treatment with Paroxetine was ineffective.

System.13 Recommendation: **DULOXETINE**

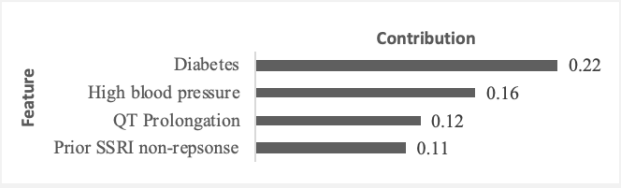
Top 5 therapies with highest probability for stability:

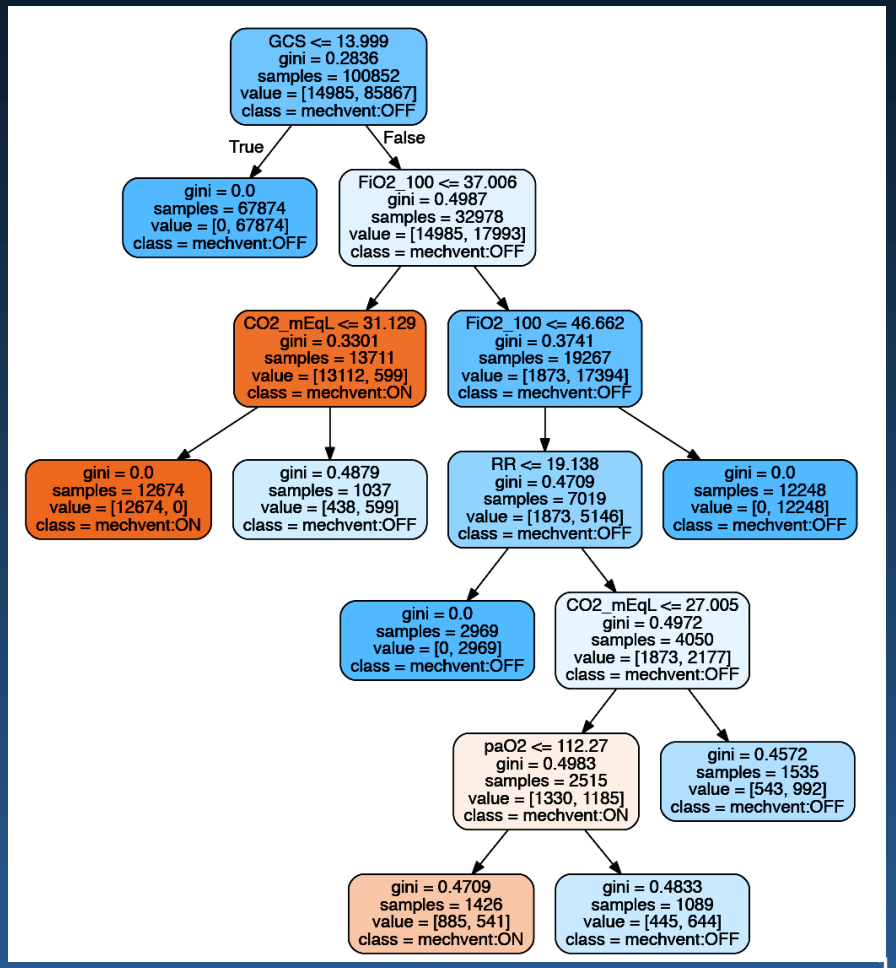
Therapy	Predicted Stability*	Predicted Dropout Risk**
Duloxetine	.80	.06
Fluoxetine	.68	.12
Citalopram	.67	.13
Escitalopram	.59	.16
Bupropion	.57	.20

**Stability: continued use of the same medication for at least 3 months*
***Dropout: early treatment discontinuation following prescription*

Why are these therapies being recommended?

The following **patient features** had the highest contributions to system.13's predictions:





Plan for these modules:

Module 1: Techniques

- Inherently interpretable models
- Partial views for more complex models

Module 2: Computational Evaluations

Module 3: Human Factors

Plan for these modules:

Module 1: Techniques

- Inherently interpretable models
- Partial views for more complex models

Module 2: Computational Evaluations

Module 3: Human Factors

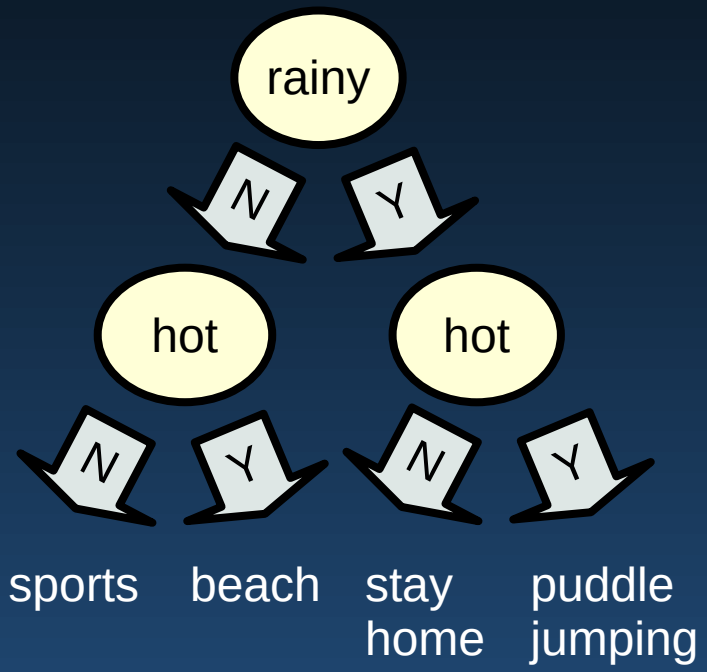
Inherently Interpretable Model

An inherently interpretable model is a model that you can literally “look at the whole thing and understand it”

Let’s start with some simple examples…

Decision Trees, Sets, Tables, Lists

Many different formats for logical expressions:




```
IF (rainy and hot):
    puddle-jump
ELSEIF (rainy):
    stay home
ELSEIF (hot):
    beach
ELSE
    sports
```

```
IF (rainy and hot):
    puddle-jump
IF (rainy and ¬hot):
    stay home
IF(¬rainy and hot):
    beach
IF(¬rainy and ¬hot):
    sports
```

Wetness	Temperature	Decision
rain	hot	puddle-jump
rain	cold	stay home
dry	hot	beach
dry	cold	sports

Super Sparse Linear Models

Have the form where features contribute linearly to a score threshold with integer weights

PREDICT PATIENT HAS OBSTRUCTIVE SLEEP APNEA IF SCORE > 1

1.	*age* $\geq$ 60	4 points		.....
2.	*hypertension*	4 points	+	.....
3.	*body mass index* $\geq$ 30	2 points	+	.....
4.	*body mass index* $\geq$ 40	2 points	+	.....
5.	*female*	-6 points	+	.....
	ADD POINTS FROM ROWS 1 – 5	**SCORE**	=	.....

Ustun and Rudin, 2009

Super Sparse Linear Models

Have the form where features contribute linearly to a score threshold with integer weights

PREDICT PATIENT HAS OBSTRUCTIVE SLEEP APNEA IF SCORE > 1

1.	*age* $\geq$ 60	4 points		.....
2.	*hypertension*	4 points	+	.....
3.	*body mass index* $\geq$ 30	2 points	+	.....
4.	*body mass index* $\geq$ 40	2 points	+	.....
5.	*female*	-6 points	+	.....
	ADD POINTS FROM ROWS 1 – 5	**SCORE**	=	.....

Note: These can look simple, but with the right *task-focused* training, can be highly predictive!

Ustun and Rudin, 2009

Generalized Additive Models

Have the form

$$g(E[Y]) = \beta_0 + f_1(x_1) + f_2(x_2) + \ldots + f_k(x_k)$$

Hastie and Tishbirani, 1990

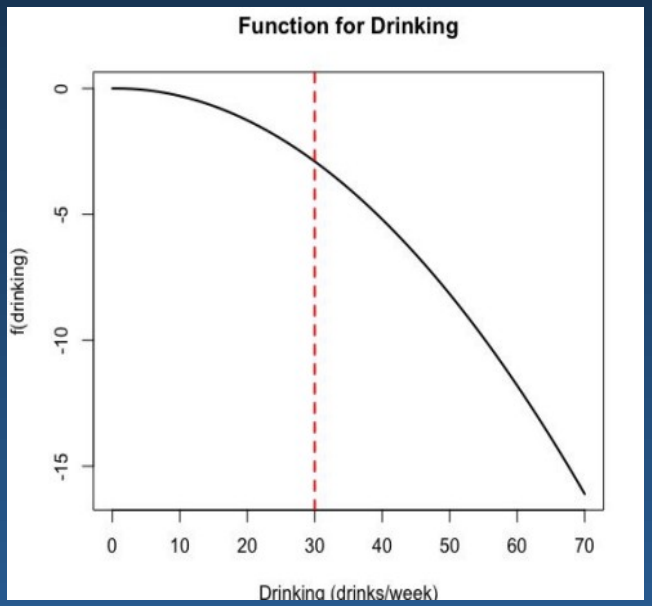
Generalized Additive Models

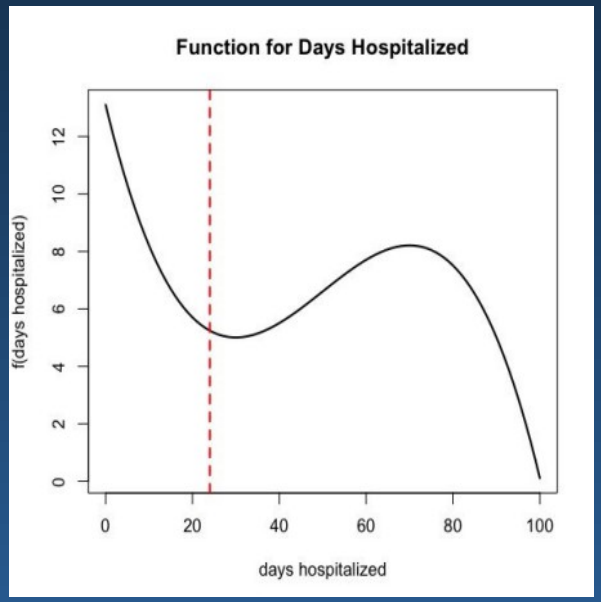
Have the form

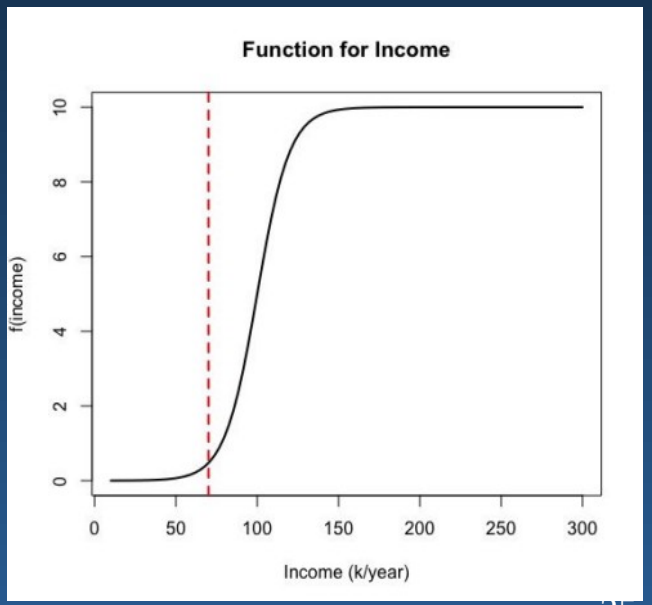
$$g(E[Y]) = \beta_0 + f_1(x_1) + f_2(x_2) + \ldots + f_k(x_k)$$

making them somewhat easy to visualize.

ExpectedYearsToLive = 25 - .3 * age + f1(drinks) + f2(days) + f3(income)







Credit: Adapted from the Harvard Embedded EthiCS program.

Inherently Interpretable Model

An inherently interpretable model is a model that you can literally “look at the whole thing and understand it”

Let’s start with some simple examples…

Inherently Interpretable Model

An inherently interpretable model is a model that you can literally “look at the whole thing and understand it”

Let’s start with some simple examples…

… and more complex examples, from our work

Case 1: Interpretable Representations

What an Agent Sees (EHR Codes)

30921	ICD-9-CM	Diagnosis	Separation anxiety disorder
29384	ICD-9-CM	Diagnosis	Organic anxiety syndrome
3000	ICD-9-CM	Diagnosis	Anxiety states
30002	ICD-9-CM	Diagnosis	
32709	ICD-9-CM	Diagnosis	Other organic insomnia
32702	ICD-9-CM	Diagnosis	Insomnia due to mental disorder
C96115	CPT-4	Procedure	Neurobehavioral status exam (clinical assessment of thinking, reasoning and judgment, eg, acquired knowledge, attention, memory, visual spatial abilities, language functions, planning) with interpretation and report, per hour
78902	ICD-9-CM	Diagnosis	Abdominal pain, left upper quadrant
C78900	CPT-4	Procedure	Abdominal pain, unspecified site
CG8440	CPT-4	Procedure	Documentation of pain assessment (including location, intensity and description) prior to initiation of treatment or documentation of the absence of pain as a result of assessment through discussion with the patient including the use of a standardized tool and a follow-up plan is documented
d04313	Multum	Medication	fentanyl topical
d04852	Multum	Medication	bupivacaine-fentanyl
d03432	Multum	Medication	aspirin-oxycodone

What a Clinician Sees (Patient, Notes)

Patient has anxiety and is being treated for abdominal pain

Case 1: Interpretable Representations

What an Agent Sees (EHR Codes)

30921 ICD-9-CM Diagnosis Separation anxiety disorder
29384 ICD-9-CM Diagnosis Organic anxiety syndrome
3000 ICD-9-CM Diagnosis Anxiety states
30002 ICD-9-CM Diagnosis
32709 ICD-9-CM Diagnosis Other organic insomnia
32702 ICD-9-CM Diagnosis Insomnia due to mental disorder
C96115 CPT-4 Procedure Neurobehavioral status exam (clinical assessment of thinking, reasoning and judgment, eg, acquired knowledge, attention, memory, visual spatial abilities, language functions, planning) with interpretation and report, per hour
78902 ICD-9-CM Diagnosis Abdominal pain, left upper quadrant
C78900 CPT-4 Procedure Abdominal pain, unspecified site
CG8440 CPT-4 Procedure Documentation of pain assessment (including location, intensity and description) prior to initiation of treatment or documentation of the absence of pain as a result of assessment through discussion with the patient including the use of a standardized tool and a follow-up plan is docum
d04313 Multum
d04852 Multum
d03432 Multum

What a Clinician Sees (Patient, Notes)

Patient has anxiety and is being treated for abdominal pain

We need representations that align the human and AI views

Example: Predicting Prescription of Antipsychotics in Psychiatry

Consider a predictor of the following form:

c2y: $\hat{y} = \sigma($ (0.704 x **Insomnia**) + (0.589 x **Anxiety**) + (-0.231 x **Overweight**) + bias)

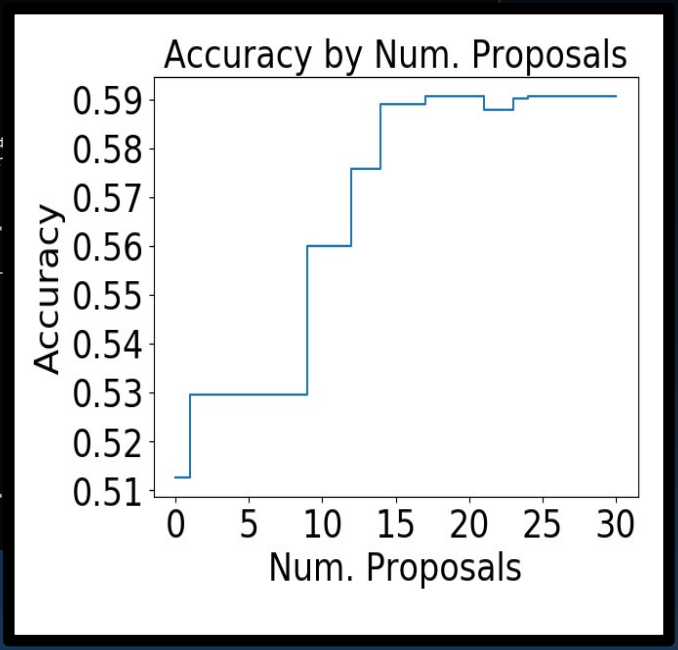
f2c:

If sum(**Features**) > 1 → **Insomnia**
Features:
Other insomnia - 78052
Trazodone - rxnorm:10737

If sum(**Features**) > 1 → **Anxiety**
Features:
Generalized anxiety - 30002
Anxiety, unspecified - 30000
Lorazepam - rxnorm:6470
Clonazepam - rxnorm:2598
Alprazolam - rxnorm:596

If sum(**Features**) > 1 → **Overweight**
Features:
Obesity, unspecified - 27800
Other hyperlipidemia - 2724
Glucose - c82962
Type II diabetes - 25002
Type II diabetes - 25000
Glyburide - 4815

```
Adding features to concept: 0 - seed feature: ***Other insomnia_78052***

Feature to concept rules
        Concept 0 - Count: 1
                        Other insomnia_78052
        Concept 1 - Count: 1
                        Generalized anxiety disorder_30002
        Concept 2 - Count: 1
                        Obesity, unspecified_27800




***COMPUTING NEXT PROPOSAL***

[       Do you want to add: ***Anxiety state, unspecified_30000*** to concept seeded with: ***Other insomnia_78052***?  Enter 'y' for yes, 'n' for no, or 'o' to ad
[               You have chosen not to add ***Anxiety state, unspecified_30000*** to the concept seeded with ***Other insomnia_78052***.  Please type 'n' to confir

***COMPUTING NEXT PROPOSAL***

[       Do you want to add: ***Displacement of cervical intervertebral disc without myelopathy_7220*** to concept seeded with: ***Other insomnia_78052***?  Enter '
 n
[               You have chosen not to add ***Displacement of cervical intervertebral disc without myelopathy_7220*** to the concept seeded with ***Other insomnia_
ge your mind: n

***COMPUTING NEXT PROPOSAL***

[       Do you want to add: ***Trazodone_rxnorm:10737*** to concept seeded with: ***Other insomnia_78052***?  Enter 'y' for yes, 'n' for no, or 'o' to add it to a
                You have added ***Trazodone_rxnorm:10737*** to this concept.  The full list of terms in the concept is now:
                        ***Other insomnia_78052***
                        ***Trazodone_rxnorm:10737***
[               To confirm this is correct, type 'y' for yes.  To remove this term from the concept, type 'n' for no: y

***COMPUTING NEXT PROPOSAL***

        Do you want to add: ***Group psychotherapy (other than of a multiple-family group)_c90853*** to concept seeded with: ***Other insomnia_78052***?  Enter 'y'
```



This predictor came from a pilot study with an expert psychiatrist and real EHR data

c2y: $\hat{y} = \sigma(\,(\,0.704 \times \textbf{Insomnia}\,) + (\,0.589 \times \textbf{Anxiety}\,) + (\,-0.231 \times \textbf{Overweight}\,) + \text{bias}\,)$

f2c:

If sum(**Features**) > 1 → **Insomnia**
Features:
Other insomnia - 78052
Trazodone - rxnorm:10737

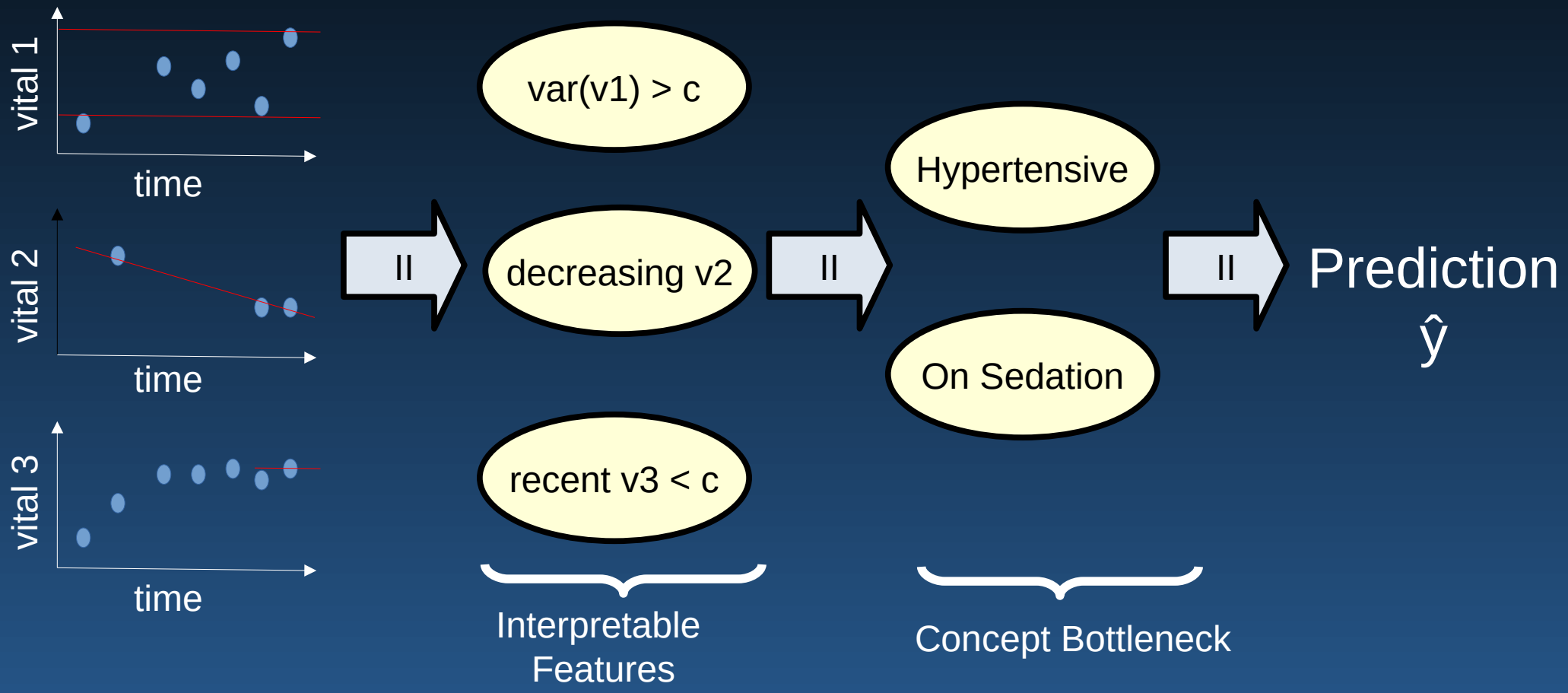
If sum(**Features**) > 1 → **Anxiety**
Features:
Generalized anxiety - 30002
Anxiety, unspecified - 30000
Lorazepam - rxnorm:6470
Clonazepam - rxnorm:2598
Alprazolam - rxnorm:596

If sum(**Features**) > 1 → **Overweight**
Features:
Obesity, unspecified - 27800
Other hyperlipidemia - 2724
Glucose - c82962
Type II diabetes - 25002
Type II diabetes - 25000
Glyburide - 4815

Lage and Doshi-Velez, ICML WS 2020

Example: Predicting Vasopressor use from timeseries

We began with how *clinicians* describe the data.



Wu et al. MLHC 2022

Example: Predicting Vasopressor use from timeseries

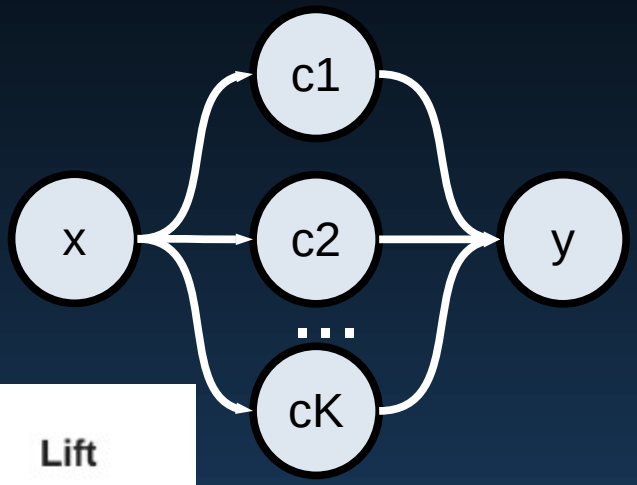
And our colleagues could completely inspect the model!

	Concept Num	Concept Weight	Feat Name	Feat Summary	Feat Weight
0	0	0.527	urine	var of indicators	0.0002
1	0	0.527	map	last time measured	0.0263
2	0	0.527	bun_ind	N/A	-0.0002
3	1	0.420	po2	last time measured	-0.0005
4	1	0.420	lactate	ever measured	0.0001
5	2	0.162	pco2_ind	N/A	-0.0283
6	3	-2.102	po2	hours below threshold	-0.0744
7	3	-2.102	GCS	mean of indicators	0.0708
8	3	-2.102	inr	hours above threshold	-0.2857
9	3	-2.102	hr_ind	N/A	0.2521
10	3	-2.102	spontaneousrr	last time measured	0.0716
11	3	-2.102	alt	slope std err	-0.1284
12	3	-2.102	hr	last time measured	0.1028
13	3	-2.102	spo2	last time measured	0.0739

Wu et al. MLHC 2022

Detour: Concept Bottlenecks Can Leak

- “Soft” concept layers (e.g. sigmoids) allow for efficient optimization.
- But, these can leak info: e.g. if p(c1) is in .9-1, both c1 and c2 are present, if p(c1) is in .7-.9, only c1 is present, etc.

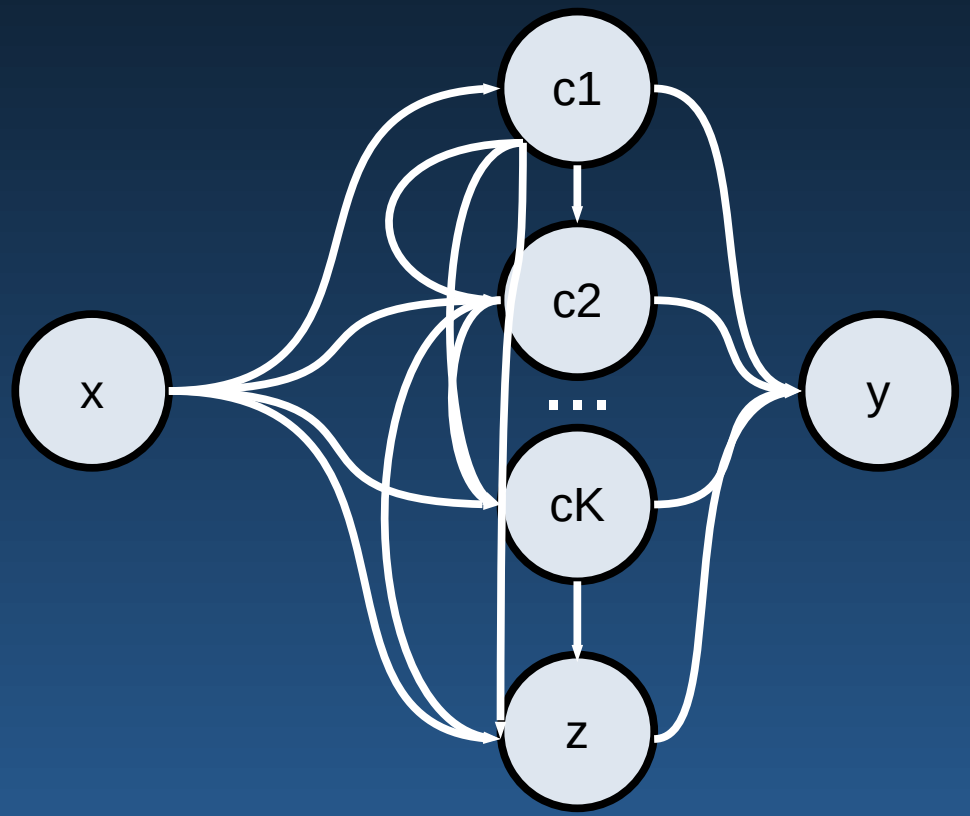


No. Concepts	Mean "Hard" Accuracy	Mean "Soft" Accuracy	Lift
1	55.76%	59.16%	3.40%
2	65.31%	71.57%	6.26%
5	72.19%	80.62%	8.43%
10	73.29%	82.69%	9.39%
50	78.18%	89.68%	11.50%
100	78.27%	92.39%	14.12%

Mahinpei and Clark et al. ICML WS 2021

Leakage can be fixed!

We can get hard CBMs (explicitly consider concepts as 0 or 1, and marginalize) to have the same predictive power as soft ones.



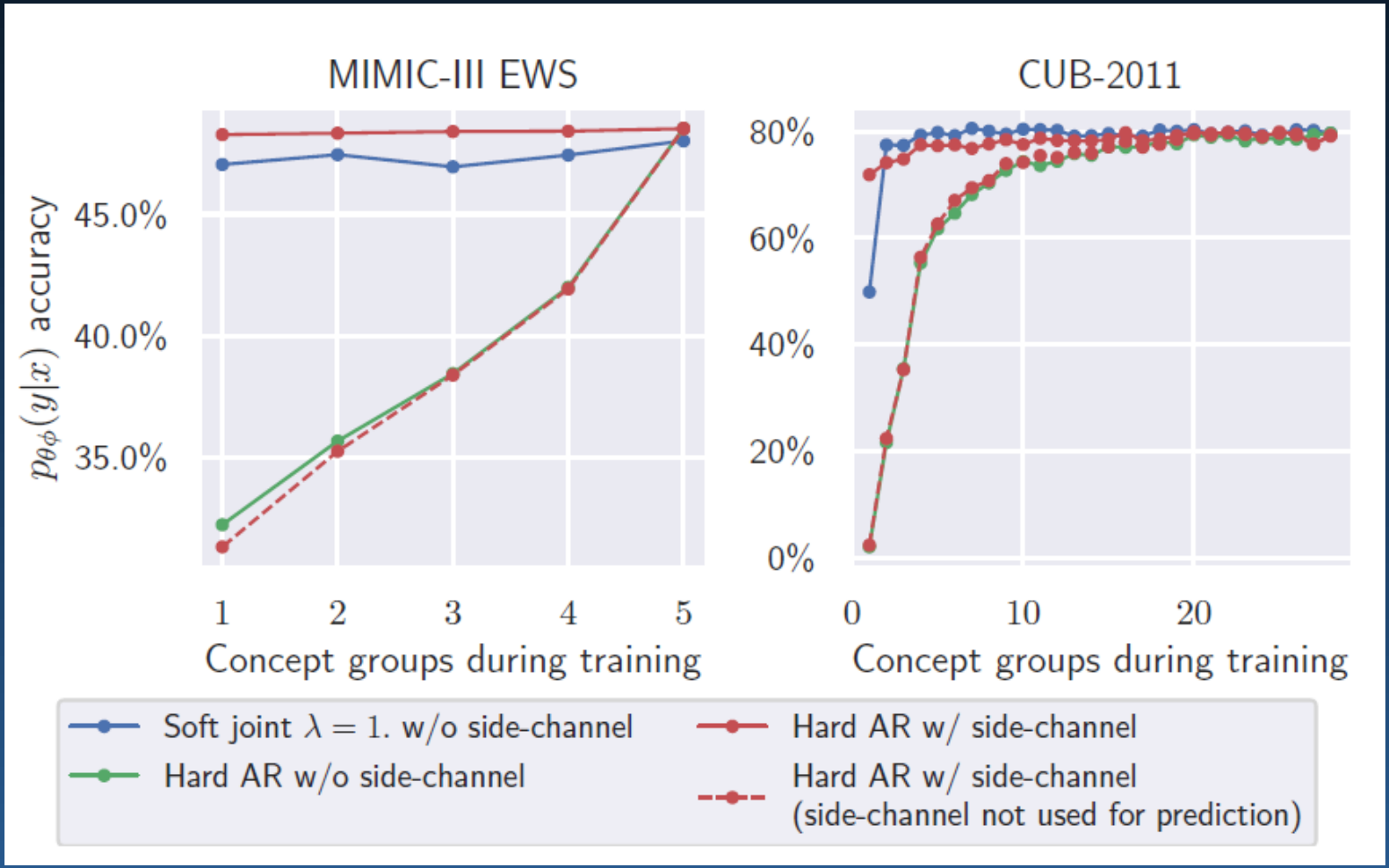
Autoregressive
Concept Modeling

Capturing
Additional Information

Havasi et al, NeurIPS 2022

Leakage can be fixed!

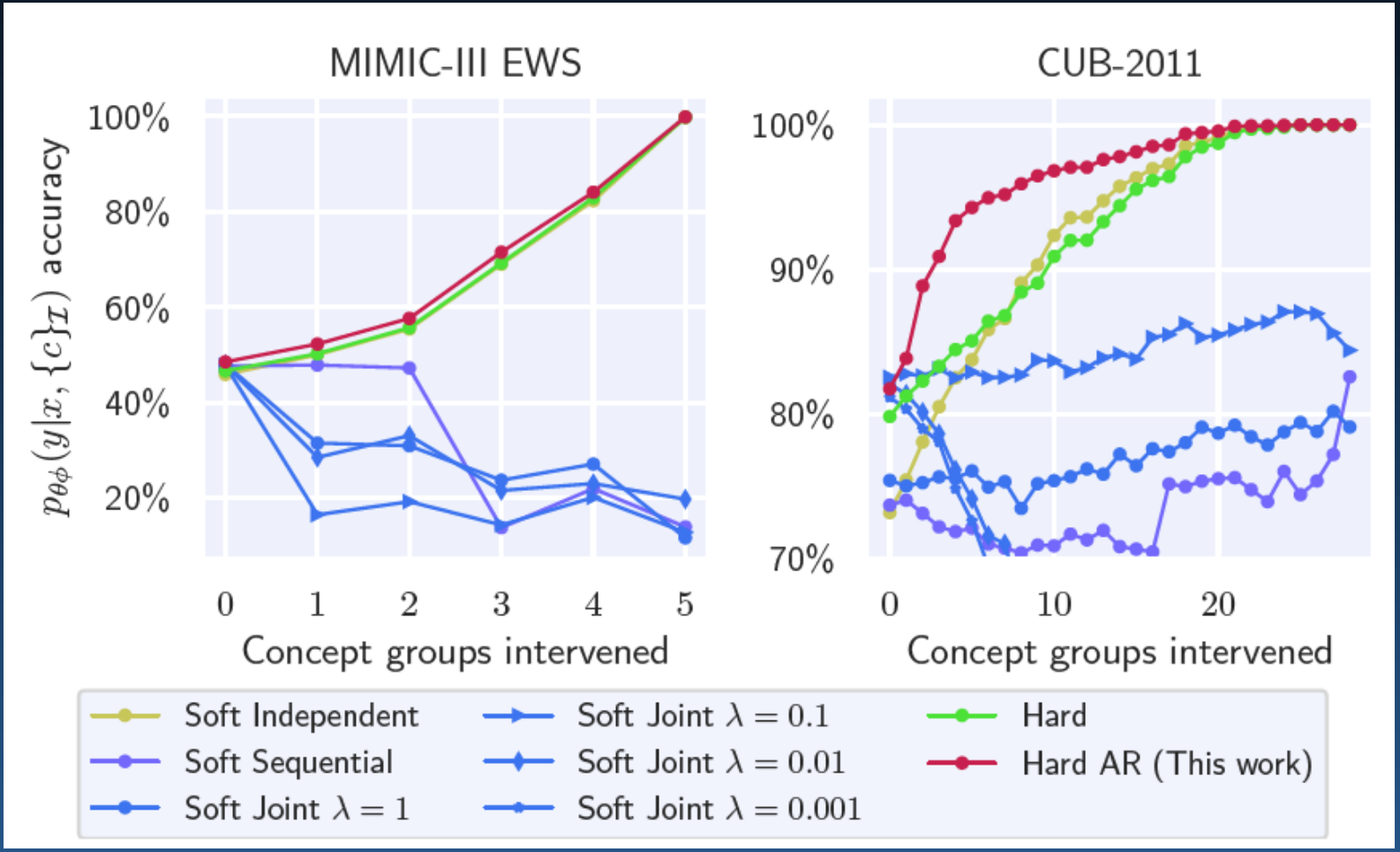
Result: Get the same predictive performance as leaky models that cheat…



Havasi et al, NeurIPS 2022

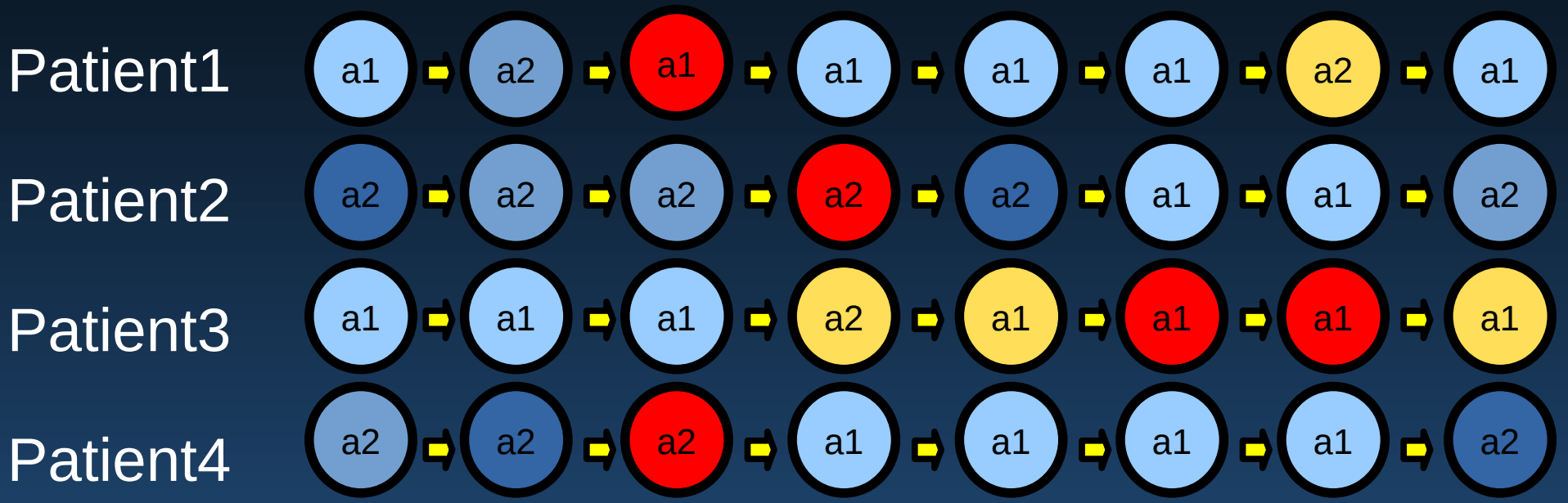
Leakage can be fixed!

…without the issues that come with cheating!



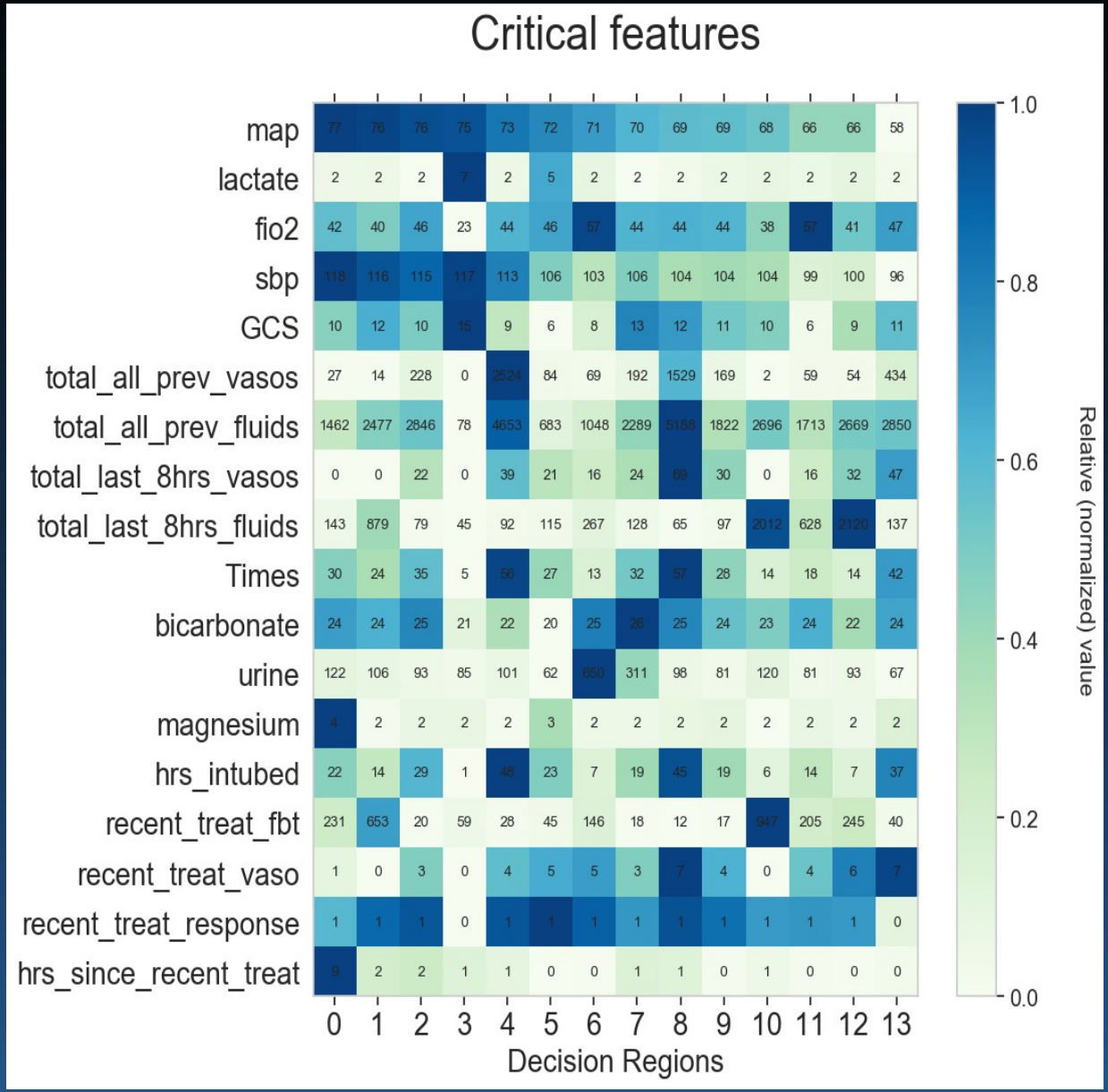
Havasi et al, NeurIPS 2022

Case 2: Batch RL: Seeking out areas of clinician disagreement



Just two states to optimize!

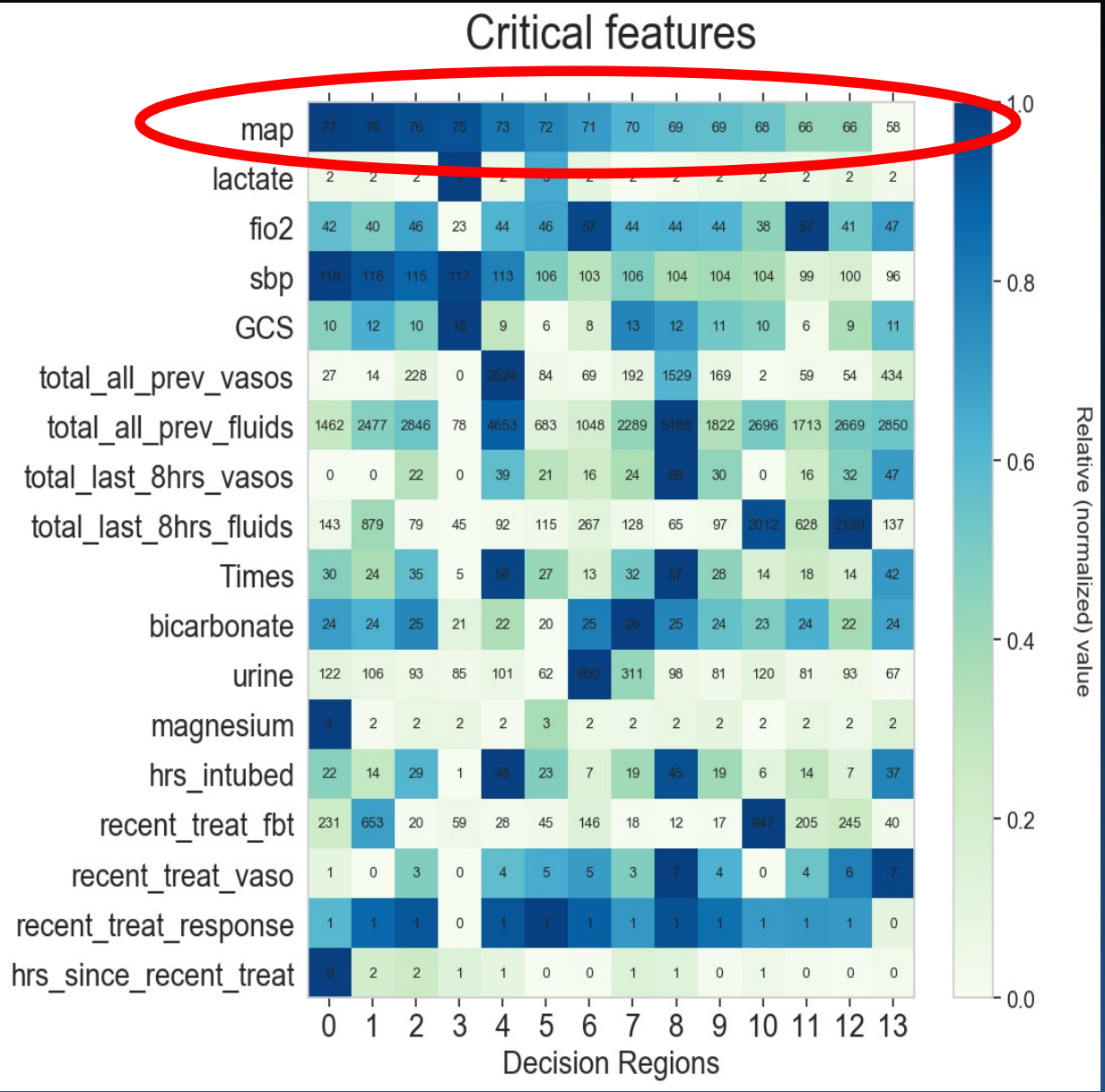
Decision States for Hypotension Management in the ICU



Zhang et al. NDM 2022

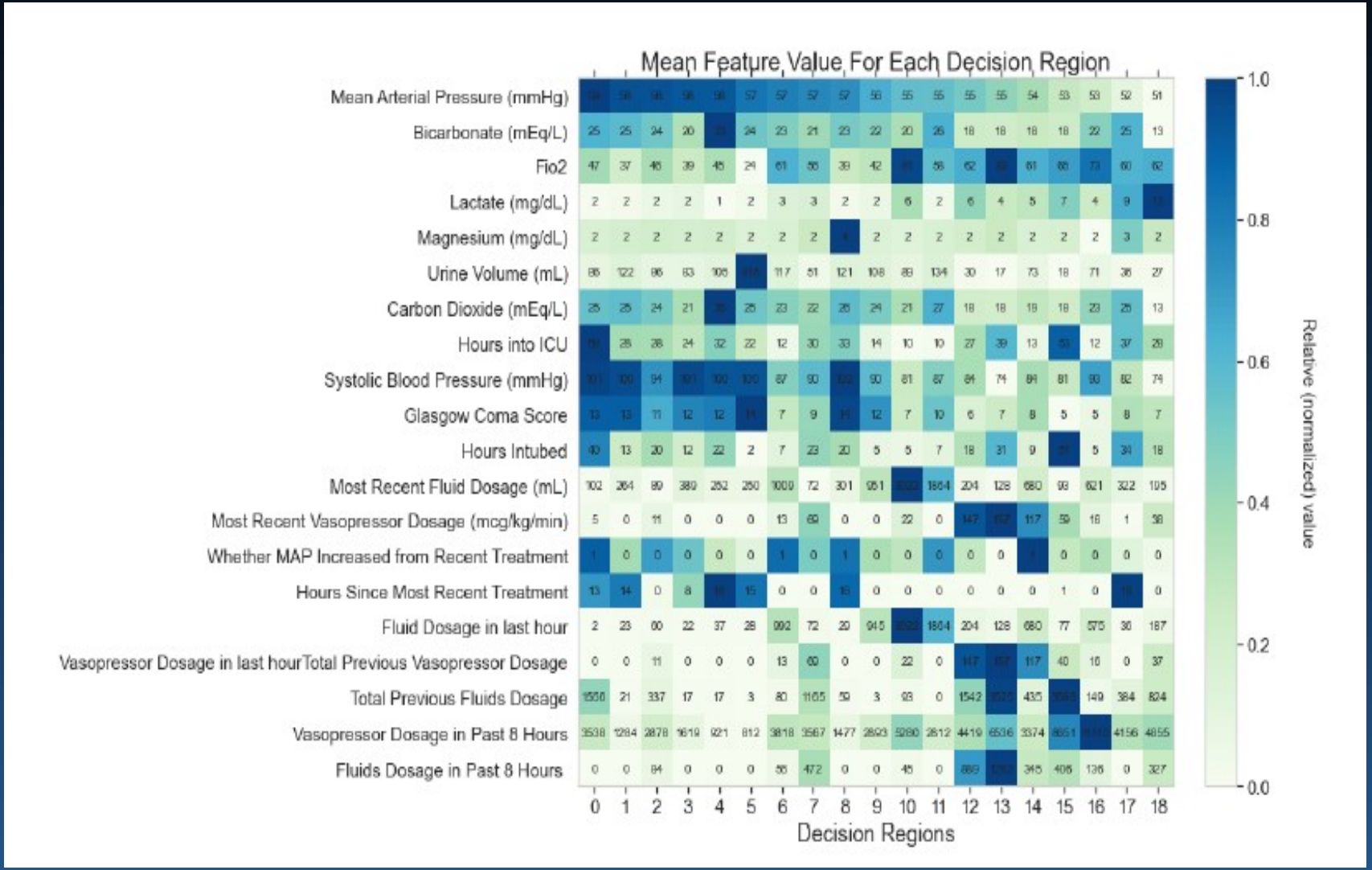
Decision States for Hypotension Management in the ICU

Clinicians: Why are the MAPs so high? Are recent dosages considered properly?



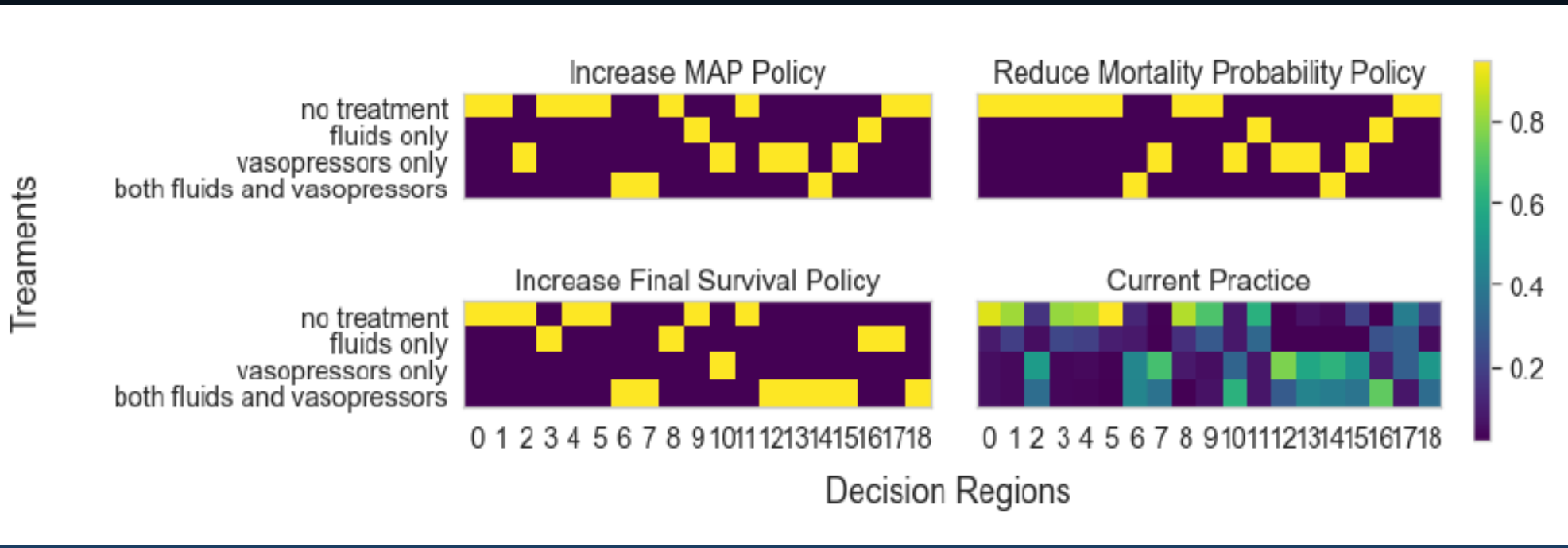
Zhang et al. NDM 2022

UPDATED Decision States



Zhang et al. NDM 2022

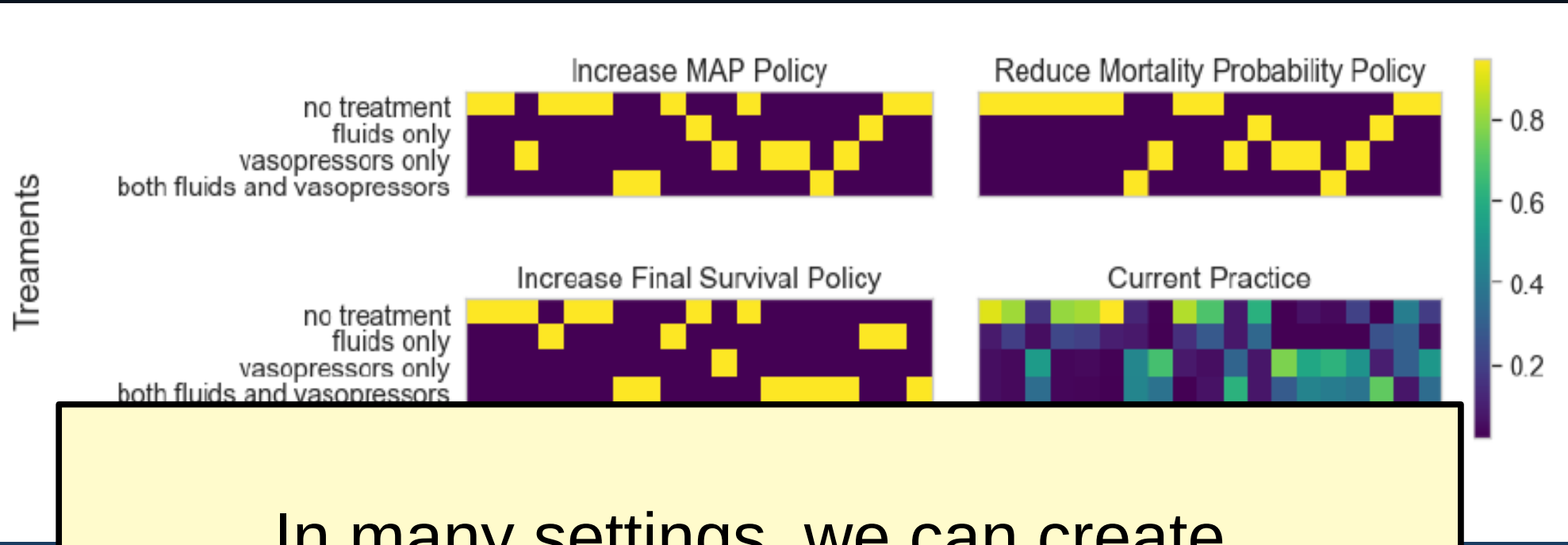
Associated Policies



Effective Sample Size (Statistical Support): ~1000; our prior works in this space ranged from ~10 to ~100.

Zhang et al. NDM 2022

Associated Policies



In many settings, we can create *inherently inspectable* models that *both* predictive and inspectable!

Zhang et al. NDM 2022

Plan for these modules:

Module 1: Techniques

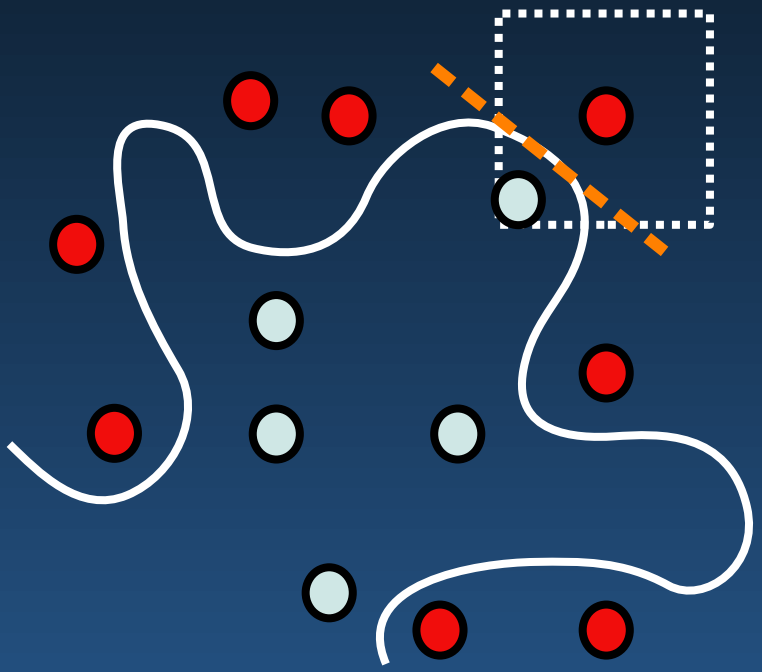
- Inherently interpretable models
- Partial views for more complex models

Module 2: Computational Evaluations

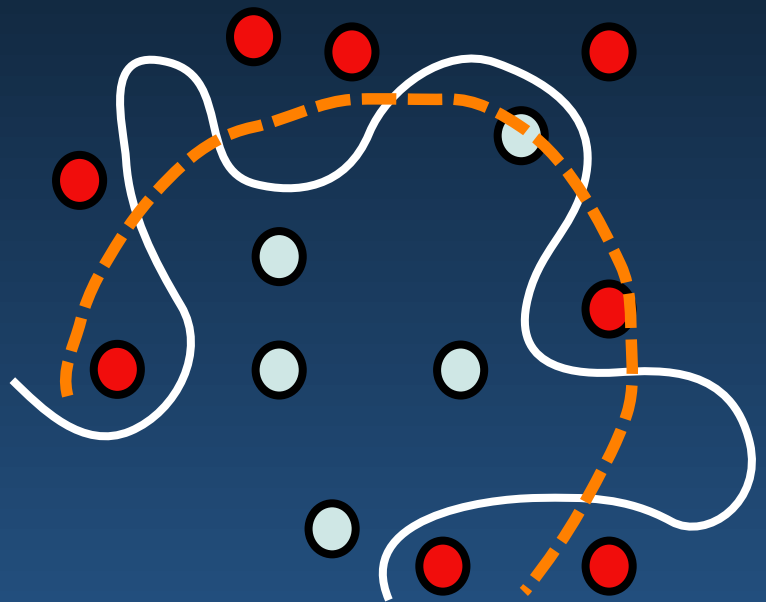
Module 3: Human Factors

Partial views: When the model is too complex for inherent interpretability

A partial view explanation is one that tells us how some – not all – of a model works.



Sensitivity around a point:
Local function, or features.



A simpler model that captures the overall shape.

Partial views: When the model is too complex for inherent interpretability

Jacob et al.

Patient Details:
Susan is a 31 year old woman who is single ar
history of diabetes, arrhythmia and hypertensiv
with 14 months of depressed mood. Current m
and prior treatment with Paroxetine was ineffe

System.13 Recommendation: **DULOXETINE**

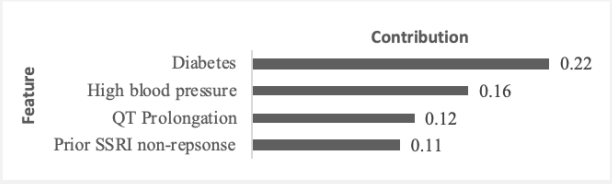
Top 5 therapies with highest probability for stability:

Therapy	Predicted Stability*
Duloxetine	.80
Fluoxetine	.68
Citalopram	.67
Escitalopram	.59
Bupropion	.57

*Stability: continued use of the same medication for at least 3 months
**Dropout: early treatment discontinuation following prescription

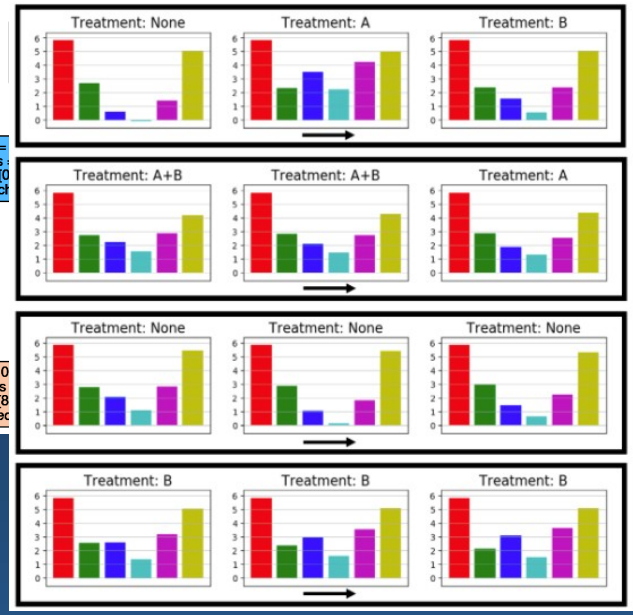
Why are these therapies being recommended?

The following **patient features** had the highest contributions to system.13's predictions:



Simonyan et al.

Wu et al., Lage et al.

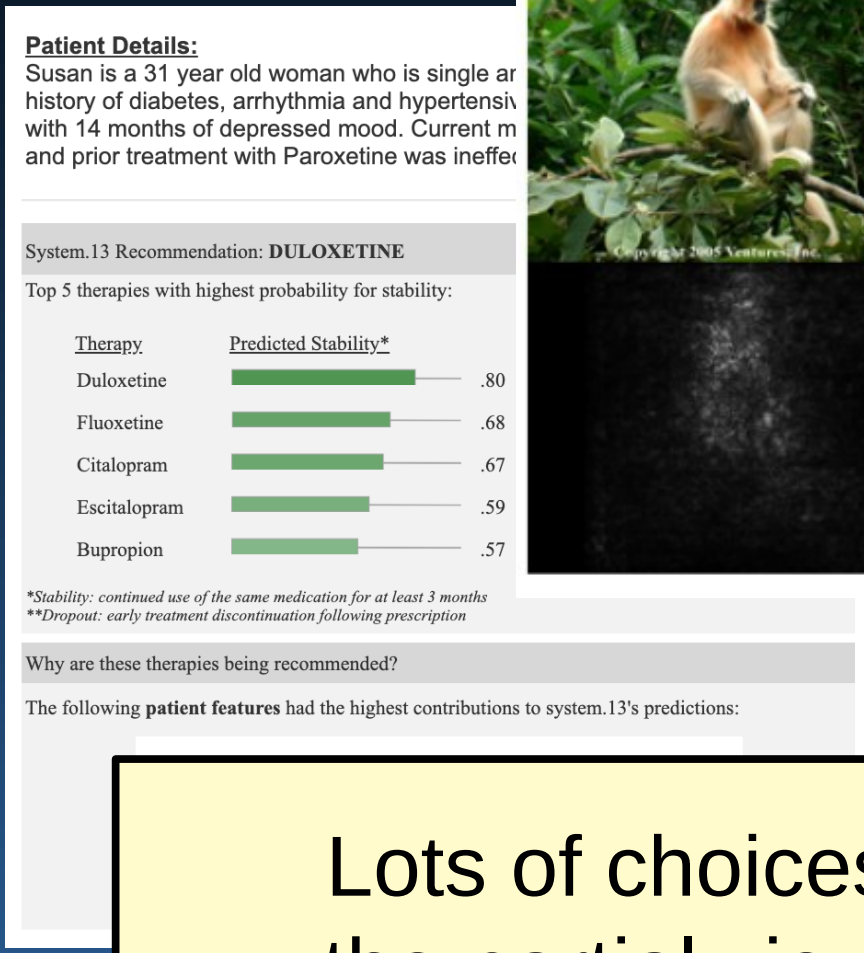


Sensitivity around a point: Local function, or features.

A simpler model that captures the overall shape.

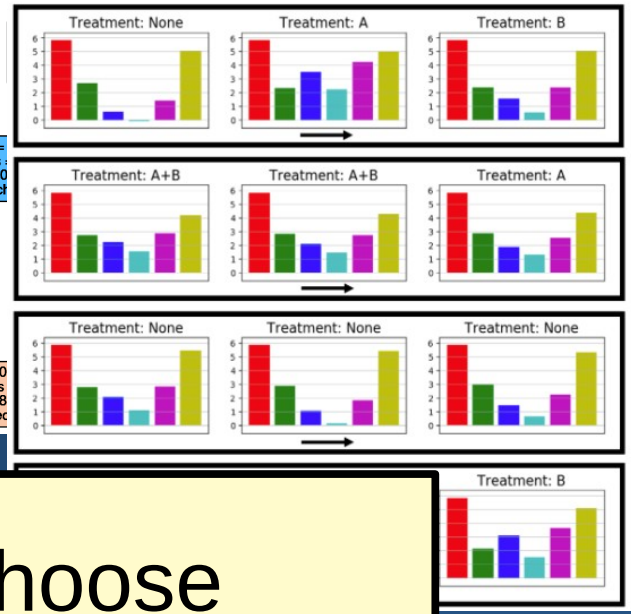
Partial views: When the model is too complex for inherent interpretability

Jacob et al.



Simonyan et al.

Wu et al., Lage et al.



Lots of choices! We have to choose the partial view needed for our task.

S
Local function, or features.
captures the overall shape.

Example: What features matter?

Predict “LikesIceCream” if

[.9 (LikesCandy) + .9 (LikesLemonade) + .2 (FriendsLikeIceCream) > 1]

Given a person who LikesCandy and FriendsLikeIceCream, what features matter?

Example: What features matter?

Predict “LikesIceCream” if

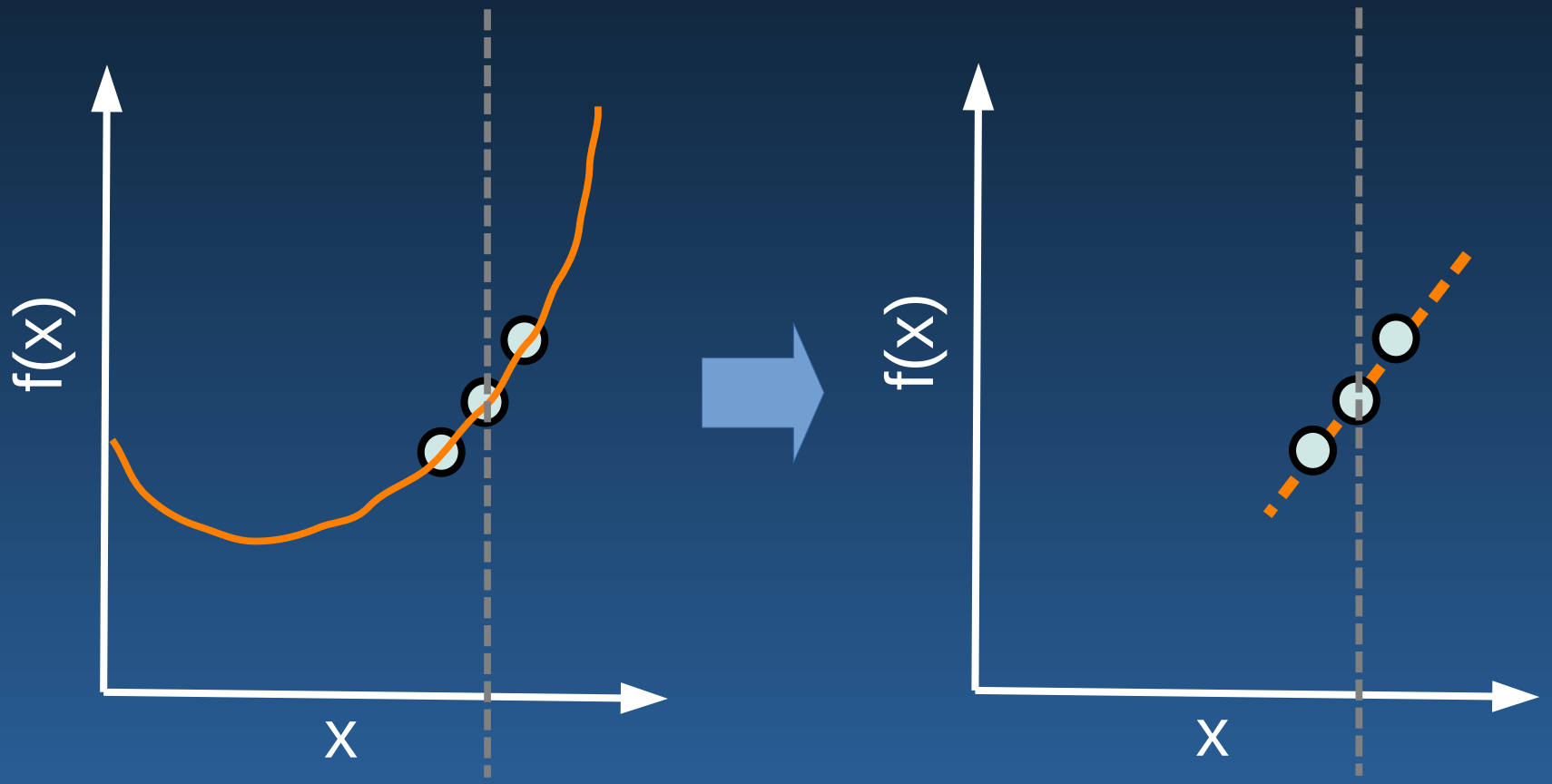
[.9 (LikesCandy) + .9 (LikesLemonade) + .2 (FriendsLikeIceCream) > 1]

Given a person who LikesCandy and FriendsLikeIceCream, what features matter?

Let’s take a look at some of the choices...

Feature Attributions: LIME

Given point of interest x , we sample $x_i \sim N(x,\sigma^2)$. Then we fit a linear function to the dataset of $\{(x_i,f(x_i))\}$. The weights are the explanation.



Feature Attributions: SHAP

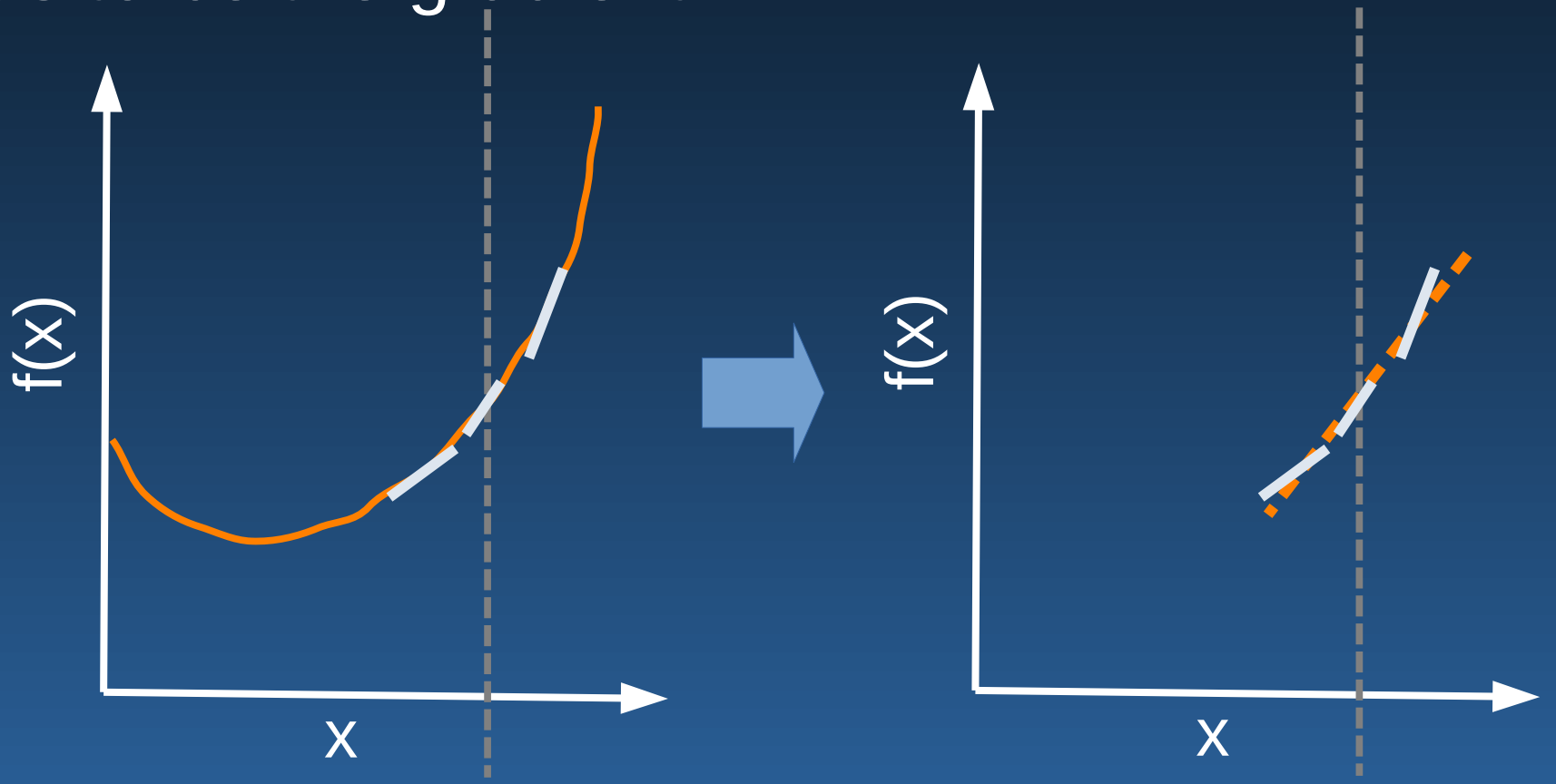
1. Sample many coalitions in which features are in and which are out m_i (“masks”).
2. Let $x_i = m_i \odot x + (1-m_i) \odot x_0$. Create $\{(x_i, f(x_i))\}$
3. Fit a weighted linear regression with weights

$$\rho(x_i) = \frac{(D-1)}{\binom{D}{|m_i|}|m_i|(D-|m_i|)}$$

Idea: “coalitions” where only one dimension changes are more important for assigning importance vs. if many dimensions change.

Feature Attributions: SmoothGrad

Given point of interest x, we sample $x_i \sim N(x,\sigma^2)$. Then we fit the gradient of linear function to the dataset of $\{(w, \nabla_x f(x_i))\}$. which, since it is linear, w has to be the gradient.



Feature Attributions: Integrated Gradients

Integrated Gradients: for each dimension, averages gradients between the input x and reference x_0:

$$(x_d - x_{0,d}) \times \int_{\alpha=0}^{1} \frac{\partial F(x_0 + \alpha(x - x_0))}{\partial x_d} d\alpha$$

by calculus:

$$\sum_d (x_d - x_{0,d}) \times \int_{\alpha=0}^{1} \frac{\partial F(x_0 + \alpha(x - x_0))}{\partial x_d} d\alpha = F(x) - F(x_0)$$

and has other nice properties e.g. if changing only dim d changes the prediction, then d will get a nonzero weight.

It is also the same as sample many points between x and some baseline x_0 and apply the same gradient matching loss as Smoothgrad with some weighting by dimension.

Feature Attribution Explanations

Many common feature attribution explanations can be cast as optimizing a linear function given points near the input:

$$E_x = w, \hat{y} = w^T x, min_{x \in Z} E[l]$$

Explanation Method	Local Neighborhood $\mathcal{Z}$ around $\mathbf{x}_0$	Loss Function ℓ
C-LIME	$\mathbf{x}_0 + \xi;\ \xi(\in \mathbb{R}^d) \sim \text{Normal}(0, \sigma^2)$	Squared Error
SmoothGrad	$\mathbf{x}_0 + \xi;\ \xi(\in \mathbb{R}^d) \sim \text{Normal}(0, \sigma^2)$	Gradient Matching
Vanilla Gradients	$\mathbf{x}_0 + \xi;\ \xi(\in \mathbb{R}^d) \sim \text{Normal}(0, \sigma^2), \sigma \to 0$	Gradient Matching
Integrated Gradients	$\xi\mathbf{x}_0;\ \xi(\in \mathbb{R}) \sim \text{Uniform}(0, 1)$	Gradient Matching
Gradients $\times$ Input	$\xi\mathbf{x}_0;\ \xi(\in \mathbb{R}) \sim \text{Uniform}(a, 1), a \to 1$	Gradient Matching
LIME	$\mathbf{x}_0 \odot \xi;\ \xi(\in \{0,1\}^d) \sim$ Exponential kernel	Squared Error
KernelSHAP	$\mathbf{x}_0 \odot \xi;\ \xi(\in \{0,1\}^d) \sim$ Shapley kernel	Squared Error
Occlusion	$\mathbf{x}_0 \odot \xi;\ \xi(\in \{0,1\}^d) \sim$ Random one-hot vectors	Squared Error

Han et al., 2022

Feature Attribution Explanations

Many common feature attribution explanations can be cast as optimizing a linear function given points near the input:

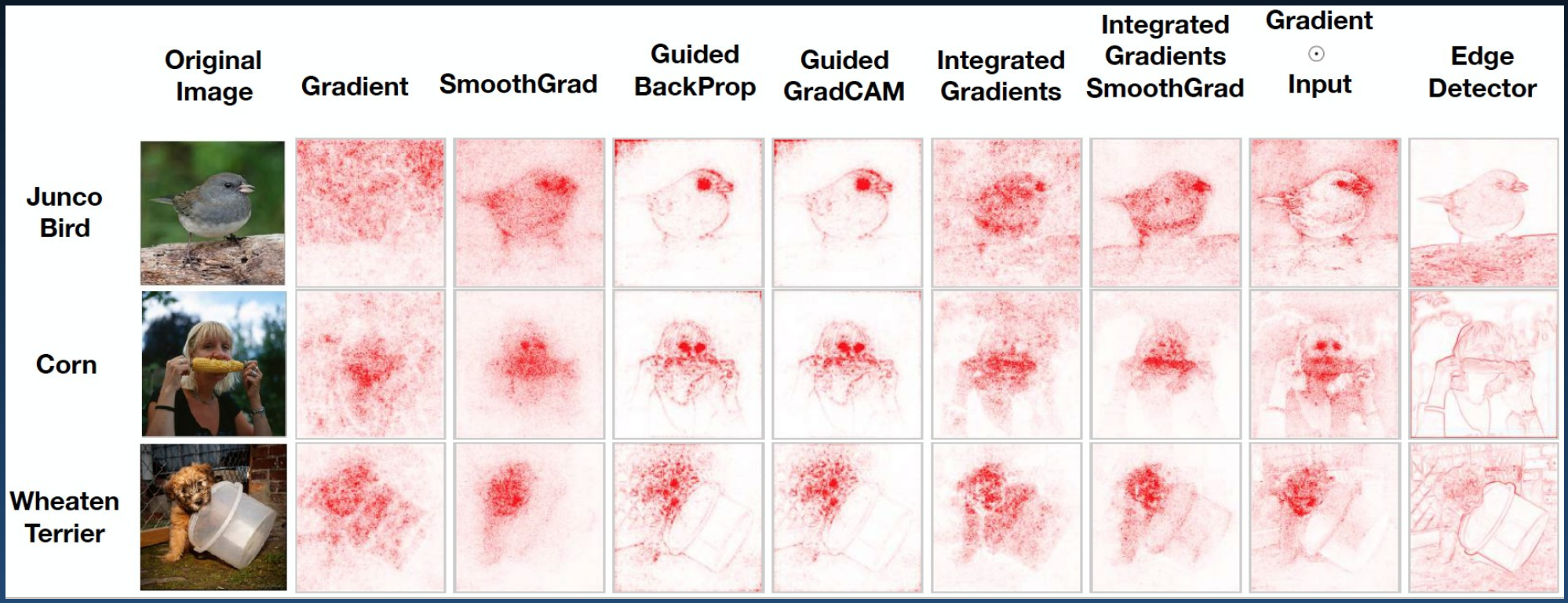
$$E_x = w, \hat{y} = w^T x, min_{x \in Z} E[l]$$

Explanation Method	Local Neighborhood $\mathcal{Z}$ around $\mathbf{x}_0$	Loss Function ℓ
C-LIME	$\mathbf{x}_0 + \xi;\ \xi(\in \mathbb{R}^d) \sim \text{Normal}(0, \sigma^2)$	Squared Error
SmoothGrad	$\mathbf{x}_0 + \xi;\ \xi(\in \mathbb{R}^d) \sim \text{Normal}(0, \sigma^2)$	Gradient Matching
Vanilla Gradients	$\mathbf{x}_0 + \xi;\ \xi(\in \mathbb{R}^d) \sim \text{Normal}(0, \sigma^2), \sigma \to 0$	Gradient Matching
Integrated Gradients	$\xi\mathbf{x}_0;\ \xi(\in \mathbb{R}) \sim \text{Uniform}(0, 1)$	Gradient Matching
Gradients × Input	$\xi\mathbf{x}_0;\ \xi(\in \mathbb{R}) \sim \text{Uniform}(a, 1), a \to 1$	Gradient Matching
		rror
		rror
		rror

Most other partial view approaches can also be viewed as function approximation

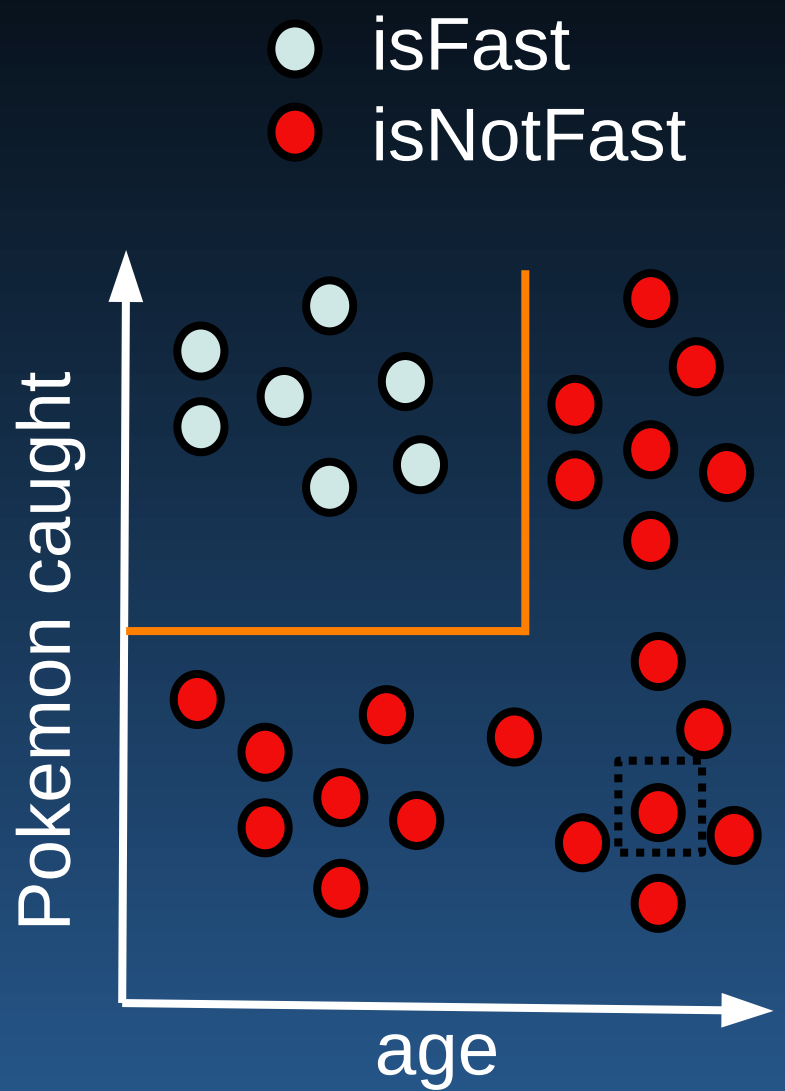
Han et al., 2022

Different choices produce different explanations!

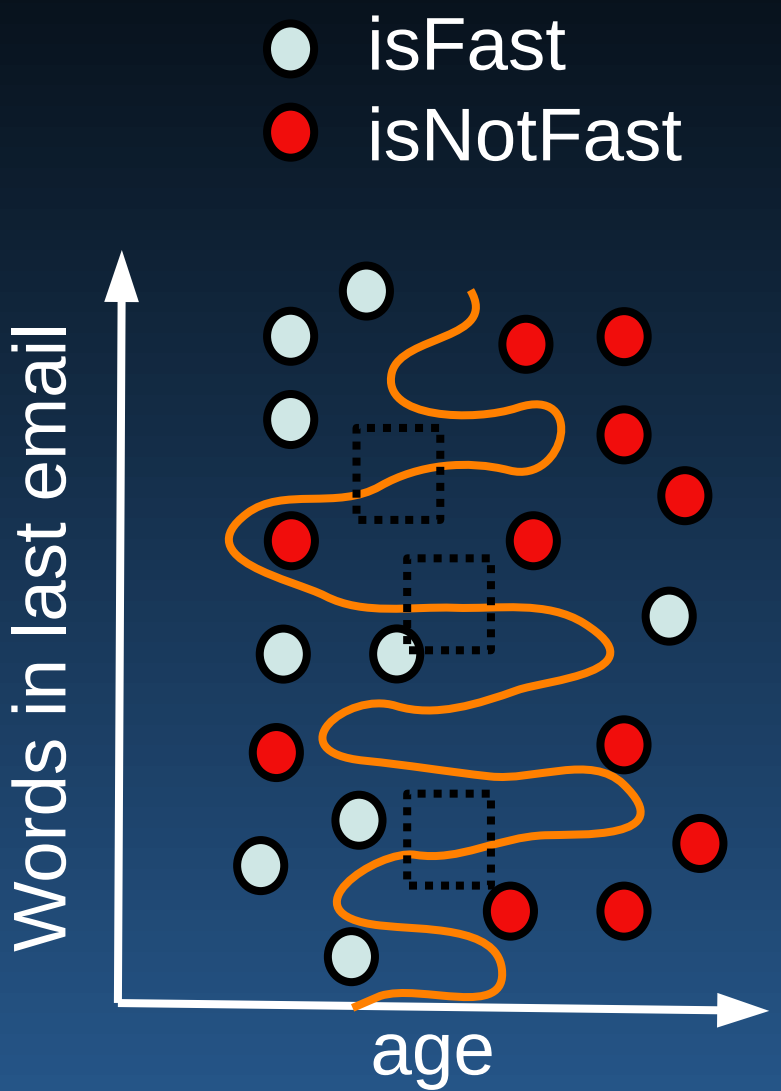


Sanity checks for Saliency Maps, Adebayo et al.

Subtleties of Partial Views



Seems like nothing matters



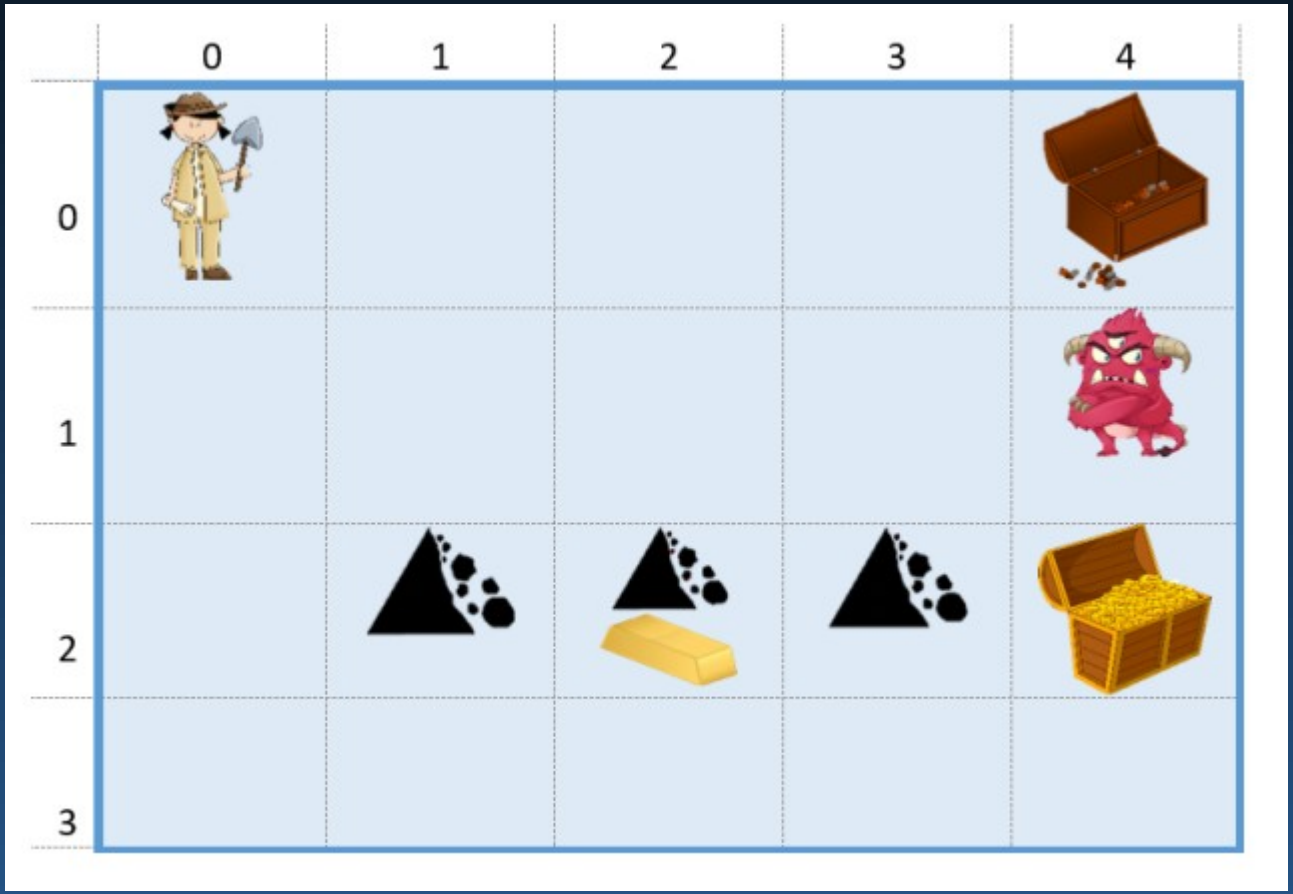
Seems like age doesn't matter

Explaining Reinforcement Learning

- We could just explain the policy $\pi(s,a)$ just like any other function – have it be inherently interpretable, or provide a partial view.
- But: A key difference in RL is that actions are justified based on what they enable in *future*.

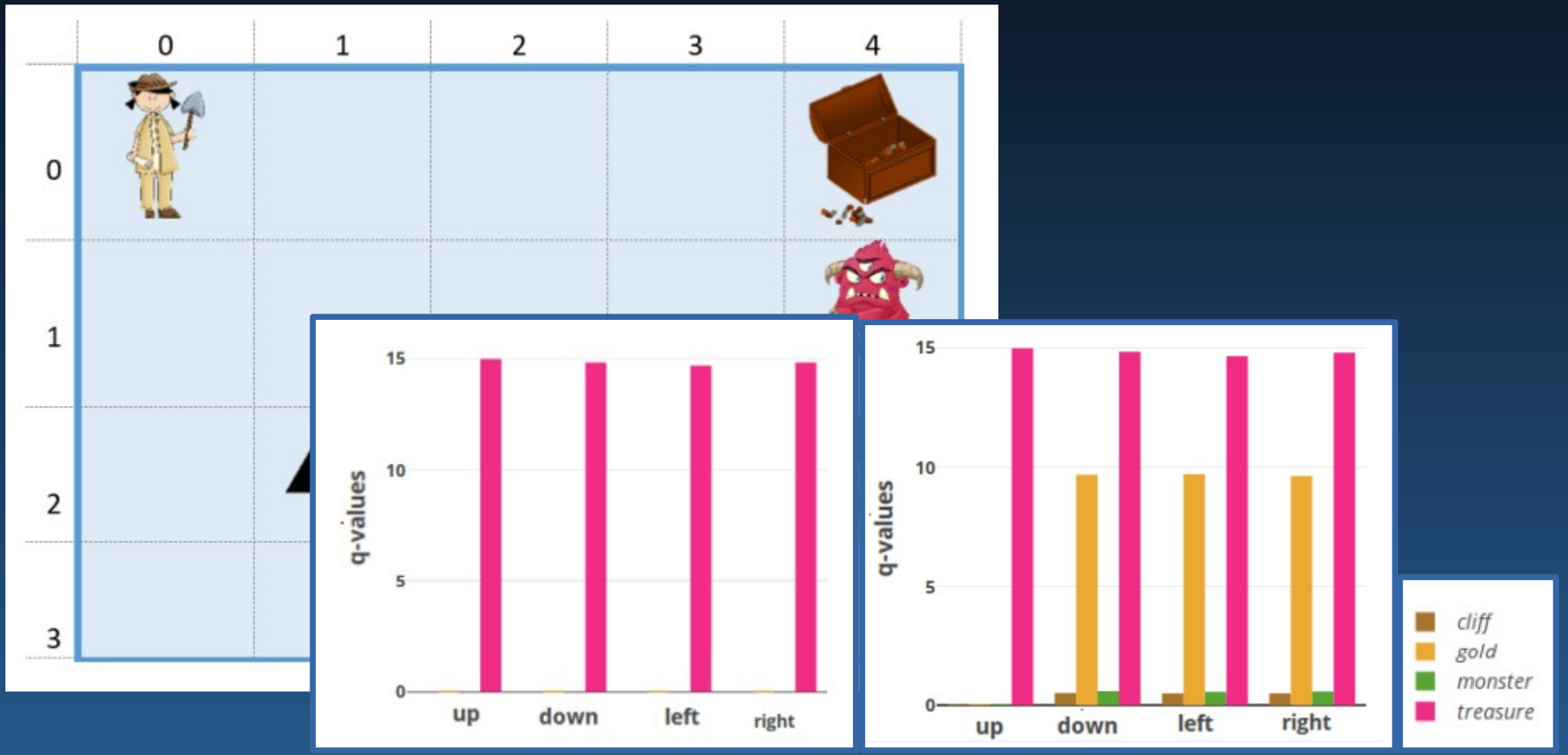
XRL that quantifies futures

Explain by what rewards/outcomes to expect.



XRL that quantifies futures

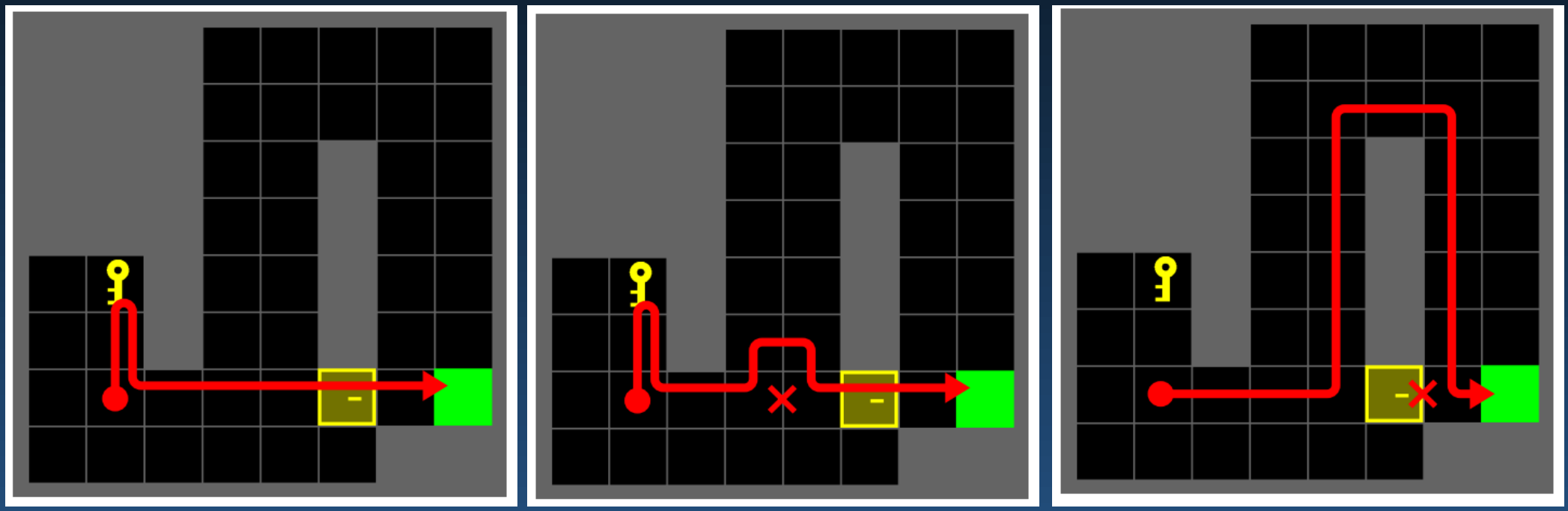
Explain by what rewards/outcomes to expect.



Juozapaitis et al., IJCAI 2019

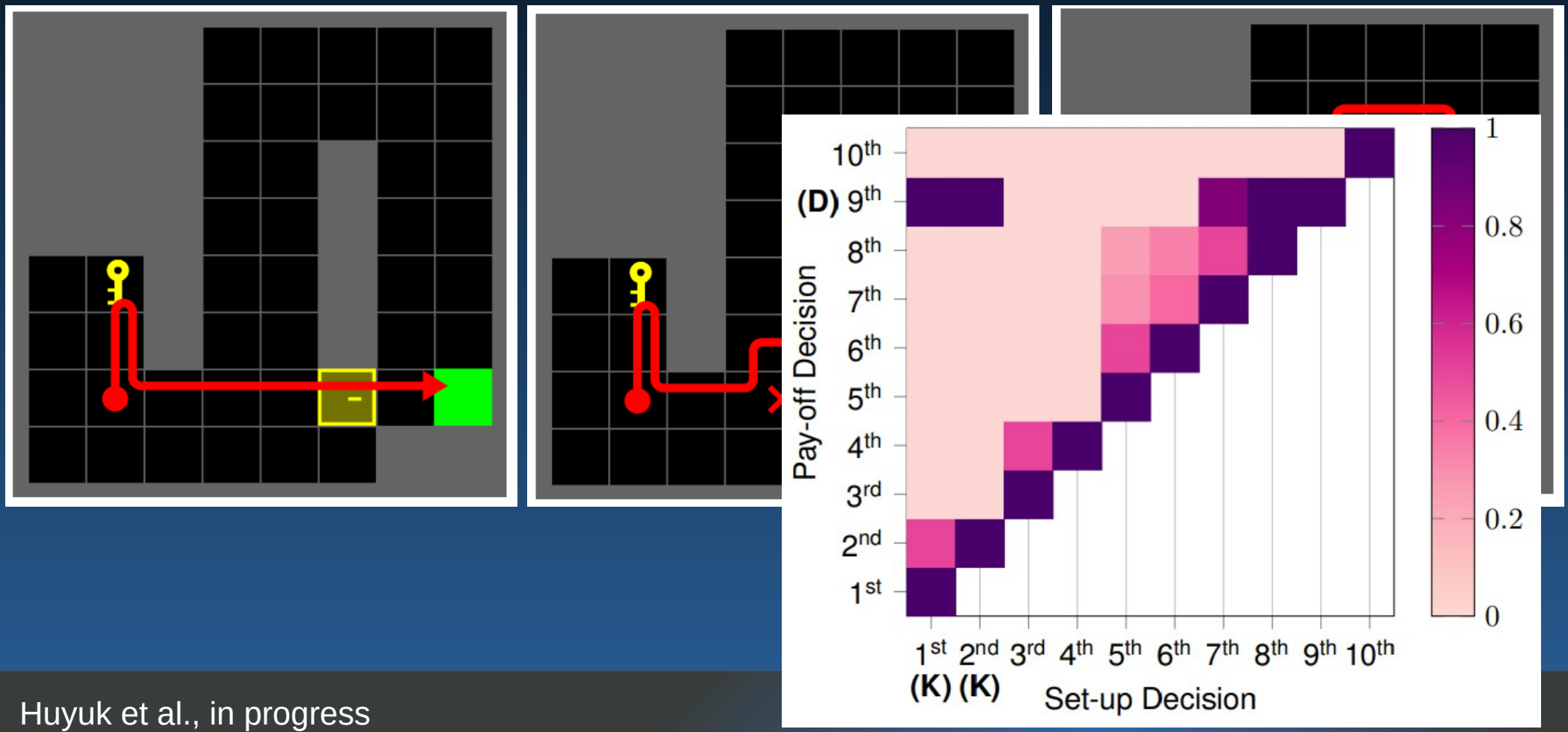
XRL that exposes connected actions

For some tasks, we need to know that taking an action now is to enable another action in future:



XRL that exposes connected actions

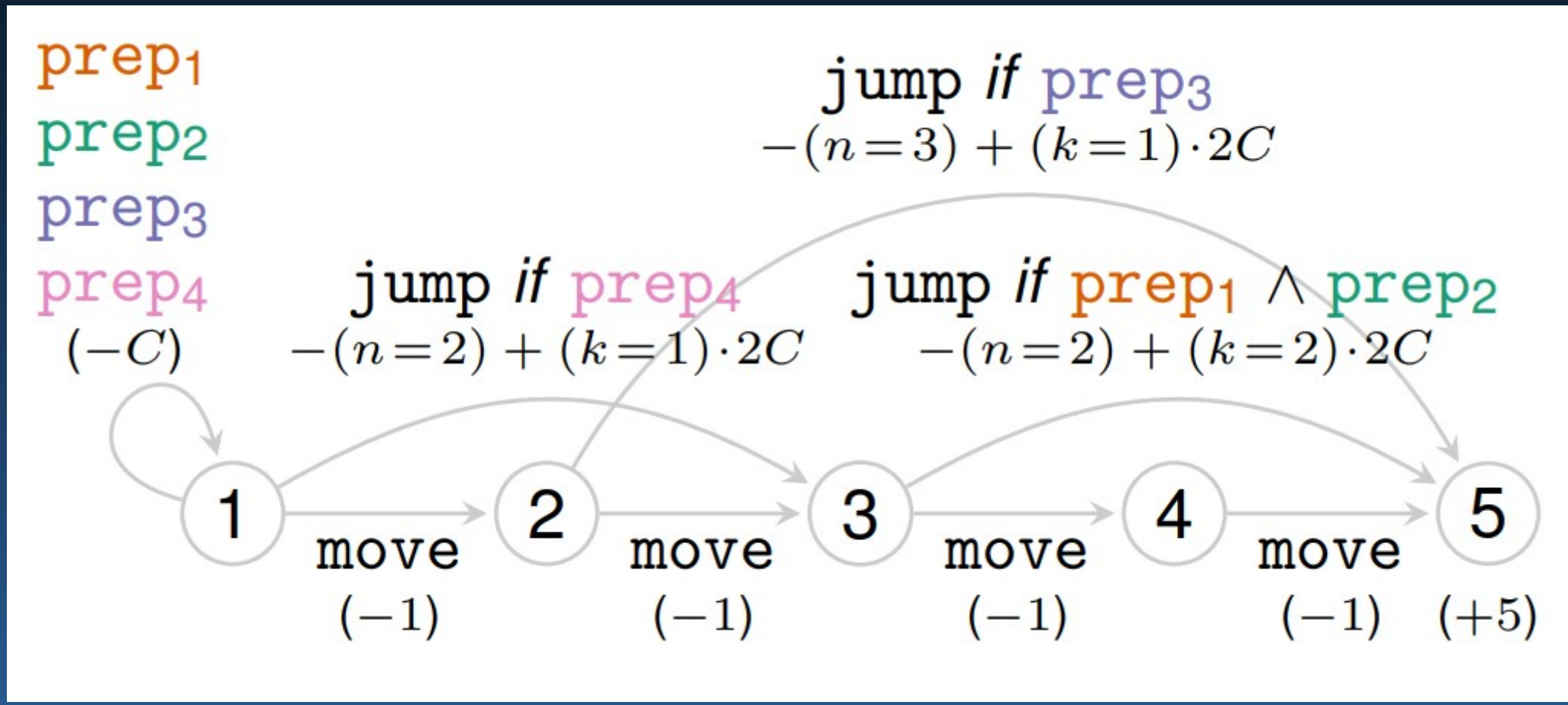
For some tasks, we need to know that taking an action now is to enable another action in future:



Huyuk et al., in progress

XRL that exposes connected actions

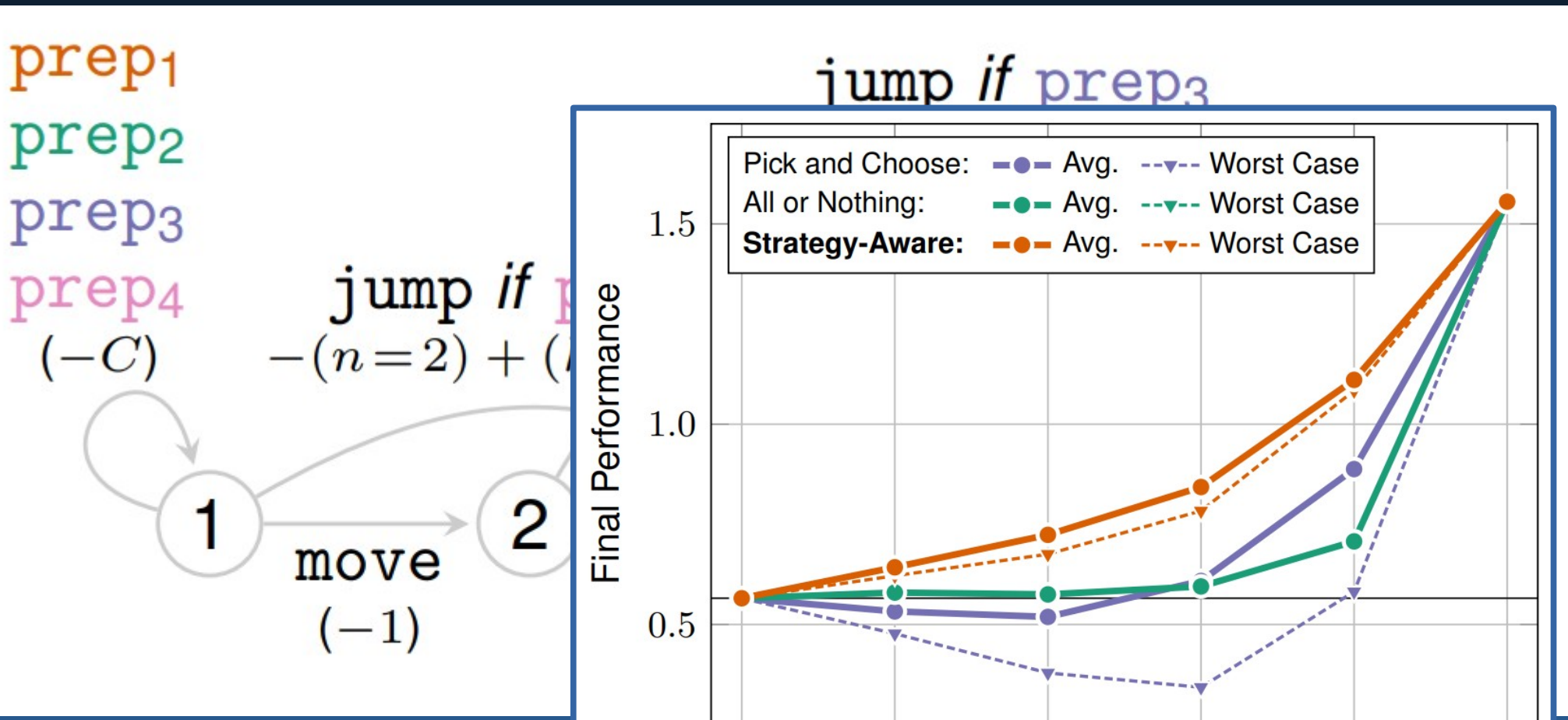
For some tasks, we need to know that taking an action now is to enable another action in future:



Huyuk et al., in progress

XRL that exposes connected actions

For some tasks, we need to know that taking an action now is to enable another action in future:



Huyuk et al., in progress

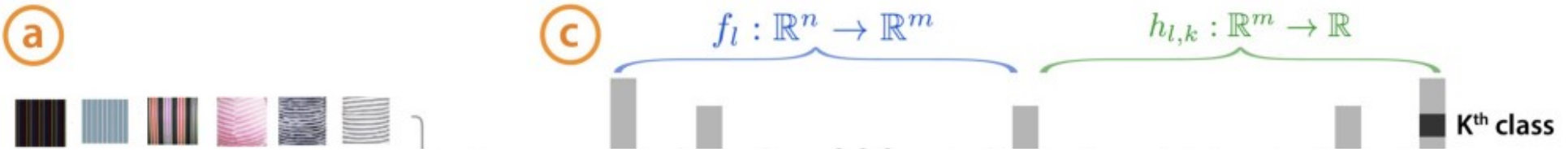


Fig
exa
sen
a c
the

Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)

Been Kim Martin Wattenberg Justin Gilmer Carrie Cai James Wexler
Fernanda Viegas Rory Sayres

Summary for Explanation Methods

- Inherently interpretable models are ones that you can just look at and fully understand.
- Partial views let us peek into more complex models.
  - We need to know what we're looking for!
  - Note: training a model and its explanations together can produce explanations with better quality explanations – and sometimes, better models too!

Evaluation

Plan for these modules:

Module 1: Techniques

Module 2: Computational Evaluations

- Conceptualizing evaluation
- Computational properties

Module 3: Human Factors

Plan for these modules:

Module 1: Techniques

Module 2: Computational Evaluations

- Conceptualizing evaluation
- Computational properties

Module 3: Human Factors

Evaluations: Where were we? (pre-2016)

"It's like porn, you know it when you see it."

Generalized additive models (GAMs) are the gold standard for intelligibility when low-dimensional terms are considered [4, 5, 6]. Standard GAMs have the form

$$g(E[y]) = \beta_0 + \sum f_j(x_j),$$

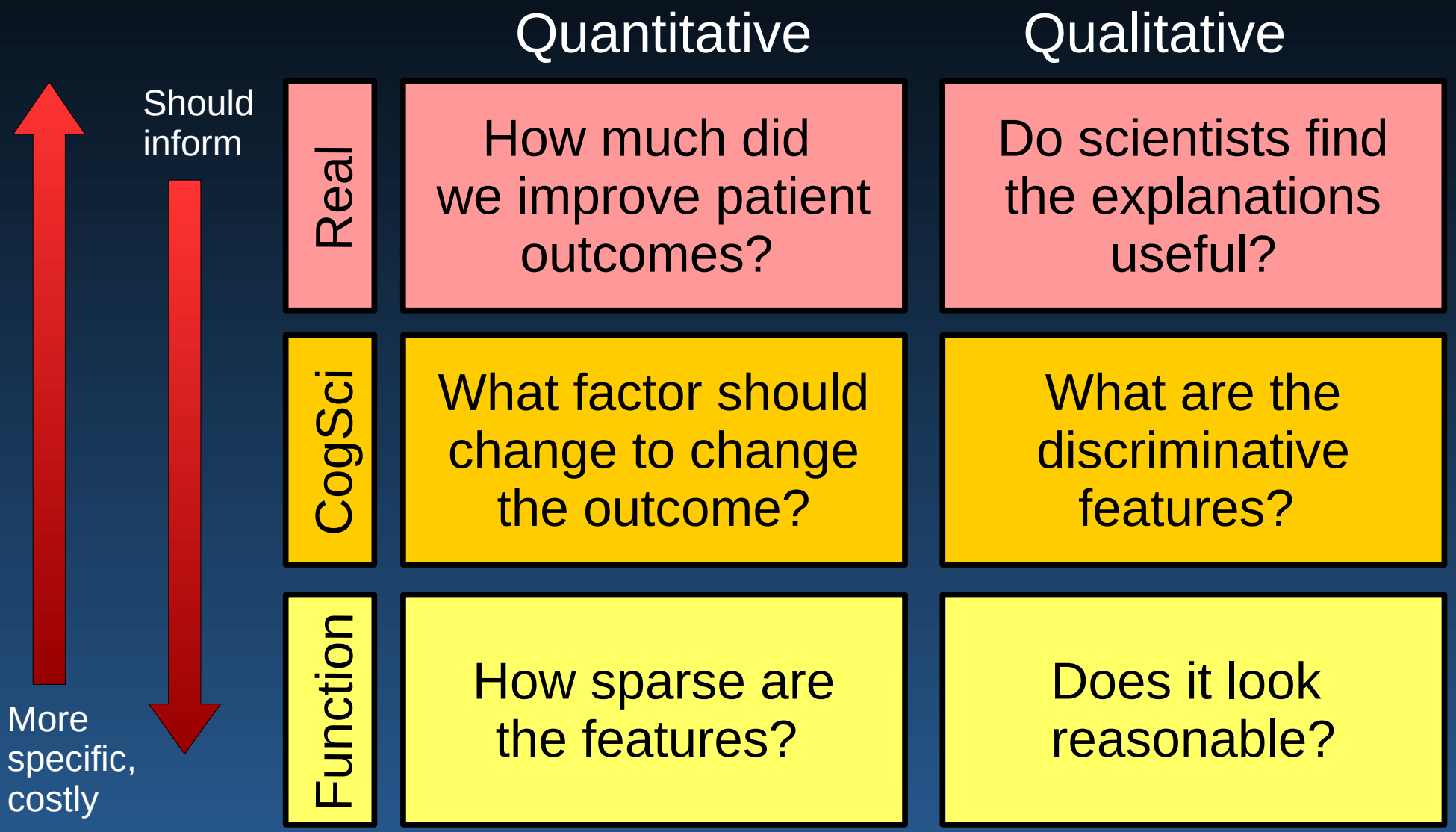
of the overall learning problem. We limit ourselves to extractive (as opposed to abstractive) rationales. From this perspective, our rationales are simply subsets of the words from the input text that satisfy two key properties. First, the selected words represent short and coherent pieces of text (e.g., phrases) and, second, the selected words must alone suffice for prediction as a substitute of the original text. More

accurate, yet are highly interpretable. These predictive models will be in the form of sparse *decision lists*, which consist of a series of *if*... *then*... statements where the *if* statements define a partition of a set of features and the *then* statements correspond to the predicted outcome of interest. Because of this form, a decision list model naturally provides a reason for

n this case, explanations are a small list of symptoms with relative weights - symptoms that either ontribute towards the prediction (in green) or are evidence against it (in red). In this, and other examples

ing robust prediction while mimicking the performance of deep learning models. Unlike the standard mimic learning [1], our interpretable mimic learning framework uses Gradient Boosting Trees (GBT) to learn interpretable features from deep learning models. We use GBT as our mimicking model since they not only provide interpretable decision rules and tree structures, but also successfully maintain the performance of original complex models such as deep networks. Our main contributions in this paper include:

A spectrum of evaluation



Doshi-Velez and Kim, 2017

Good news: More user studies!

Riveiro and Thill, “That’s (not) the output I expected! ...”. AI Journal 2021.

In this paper, we report an empirical study with 181 participants who were shown outputs from a text classifier system along with an explanation of why the system chose a particular class for each text. Explanations were both *factual*, explaining why produced a certain output or *counterfactual*, explaini system produced one output instead of another. Ou

5.3 User Studies

Finally, to investigate whether our **CET** is easy for human-users to understand, we conducted user studies with 35 participants. Each participant work in research and development departments related to artifi-

Kanamori et al. Counterfactual Explanations Trees: Transparent and Consistent Actionable Recourse with Decision Trees. AISTATS 2022.

Chromik et al. “I think I get your point AI!” The illusion of explanatory depth in explainable AI IUI 2021.

ABSTRACT

Unintended consequences of deployed AI systems fueled the call for more interpretability in AI systems. Often explainable AI (XAI) systems provide users with simplifying local explanations for individual predictions but leave it up to them to construct a global understanding of the model behavior. In this work, we examine if non-technical users of XAI fall for an illusion of explanatory depth when interpreting additive local explanations. We applied a mixed methods approach consisting of a moderated study with 40 participants and an unmoderated study with 107 crowd workers using a spreadsheet-like explanation interface based on the SHAP framework. We observed what non-technical users do to form their mental models of global AI model behavior from local explanations and how their perception of understanding decreases when it is examined.

Paleja et al. The Utility of Explainable AI in Ad Hoc Human-

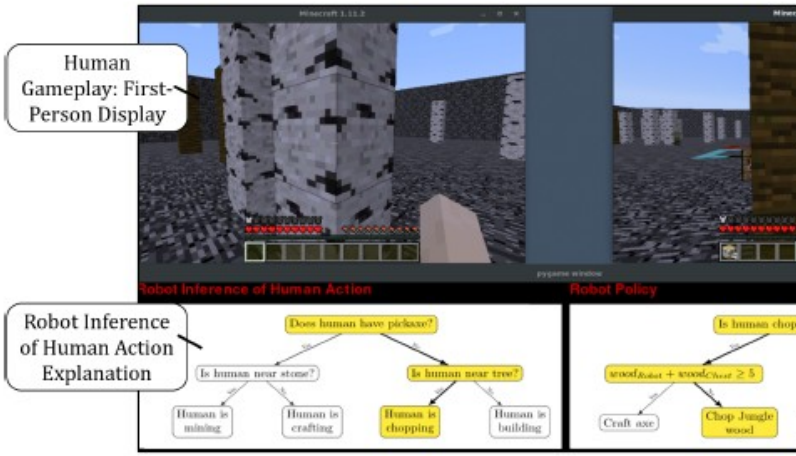


Figure 2: This figure displays a sample gameplay image where th decision-tree explanation. Note this shows IV1:SA1-2-3 condition an of Human Policy and Cobot Policy in Section 5.

trained ...e, which impedes a regret analysis in this setting. For the sake of completeness, the empirical policy values can be found in Appendix K.

User study. In a user study, we asked 10 practitioners to rate the interpretability of our policy and the baseline policies on a scale between 0 (black box) and 10 (fully transparent) and whether they would use our policies in practice (see Appendix L). All but one practitioner perceived our policies as interpretable (i. e., a rating larger than 5, which denotes neutral). Furthermore, 9 out of 10 would consider using our treatment decisions in practice. This demonstrates that our policies fulfill interpretability demands from clinical practice. A detailed comparison to our baselines can be found in Appendix L.

Tschernutter et al. Interpretable Off-Policy Learning via Hyperbox Search. ICML 2022.

Good news: More user studies!

Riveiro and Thill, "That's (not) the output I expected! ...". AI Journal 2021.

In this paper, we report an empirical study with 181 participants who were shown outputs from a text classifier system along with an explanation of why the system chose a particular class for each text. Explanations were both *factual*, explaining why

5.2 Real-Wor

produced a certain output or *counterfactual*, explaini

Paleja et al. The Utility of Explainable AI in Ad Hoc Human-

Figure 2: This figure displays a sample gameplay image where th decision-tree explanation. Note this shows IV1:SA1-2-3 condition an of Human Policy and Cobot Policy in Section 5.

But: How do we gain generalizable knowledge from a user study?

Actionable Recourse with Decision Trees. AISTATS 2022.

search and development departments related to artifi-

ES,
For M
real

trained

Chromik et al. "I think I get your point AI!" The illusion of explanatory depth in explainable AI IUI 2021.

ABSTRACT

Unintended consequences of deployed AI systems fueled the call for more interpretability in AI systems. Often explainable AI (XAI) systems provide users with simplifying local explanations for individual predictions but leave it up to them to construct a global understanding of the model behavior. In this work, we examine if non-technical users of XAI fall for an illusion of explanatory depth when interpreting additive local explanations. We applied a mixed methods approach consisting of a moderated study with 40 participants and an unmoderated study with 107 crowd workers using a spreadsheet-like explanation interface based on the SHAP framework. We observed what non-technical users do to form their mental models of global AI model behavior from local explanations and how their perception of understanding decreases when it is examined.

ıe, which impedes a regret analysis in this setting. For the sake of completeness, the empirical policy values can be found in Appendix K.

User study. In a user study, we asked 10 practitioners to rate the interpretability of our policy and the baseline policies on a scale between 0 (black box) and 10 (fully transparent) and whether they would use our policies in practice (see Appendix L). All but one practitioner perceived our policies as interpretable (i. e., a rating larger than 5, which denotes neutral). Furthermore, 9 out of 10 would consider using our treatment decisions in practice. This demonstrates that our policies fulfill interpretability demands from clinical practice. A detailed comparison to our baselines can be found in Appendix L.

Tschernutter et al. Interpretable Off-Policy Learning via Hyperbox Search. ICML 2022.

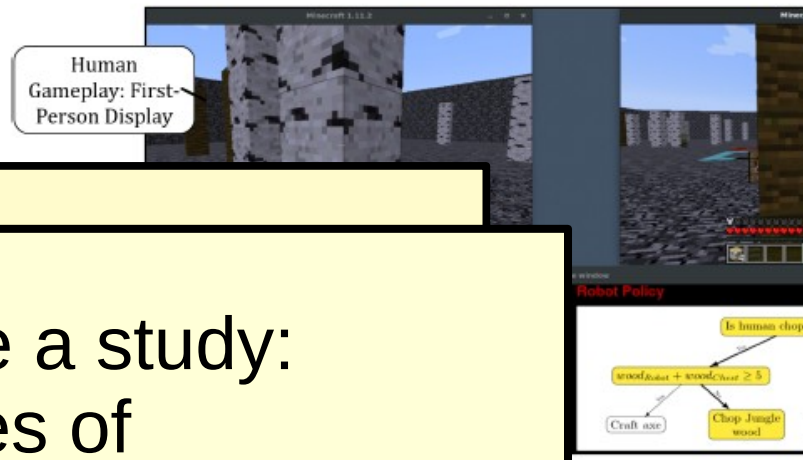
Good news: More user studies!

Riveiro and Thill, “That’s (not) the output I expected! ...”. AI Journal 2021.

In this paper, we report an empirical study with 181 participants who were shown outputs from a text classifier system along with an explanation of why the system chose a particular class for each text. Explanations were both *factual*, explaining why

5.2 Real-Wor

Paleja et al. The Utility of Explainable AI in Ad Hoc Human-



Specificity issue: e.g. imagine a study:
“Do fewer but longer lines of explanation result in better decisions?”
Needs abstractions to generalize:
1) “Does an explanation that can be read quickly result in better decisions?”
2) For different explanation forms (lists, etc.) what makes them readable?

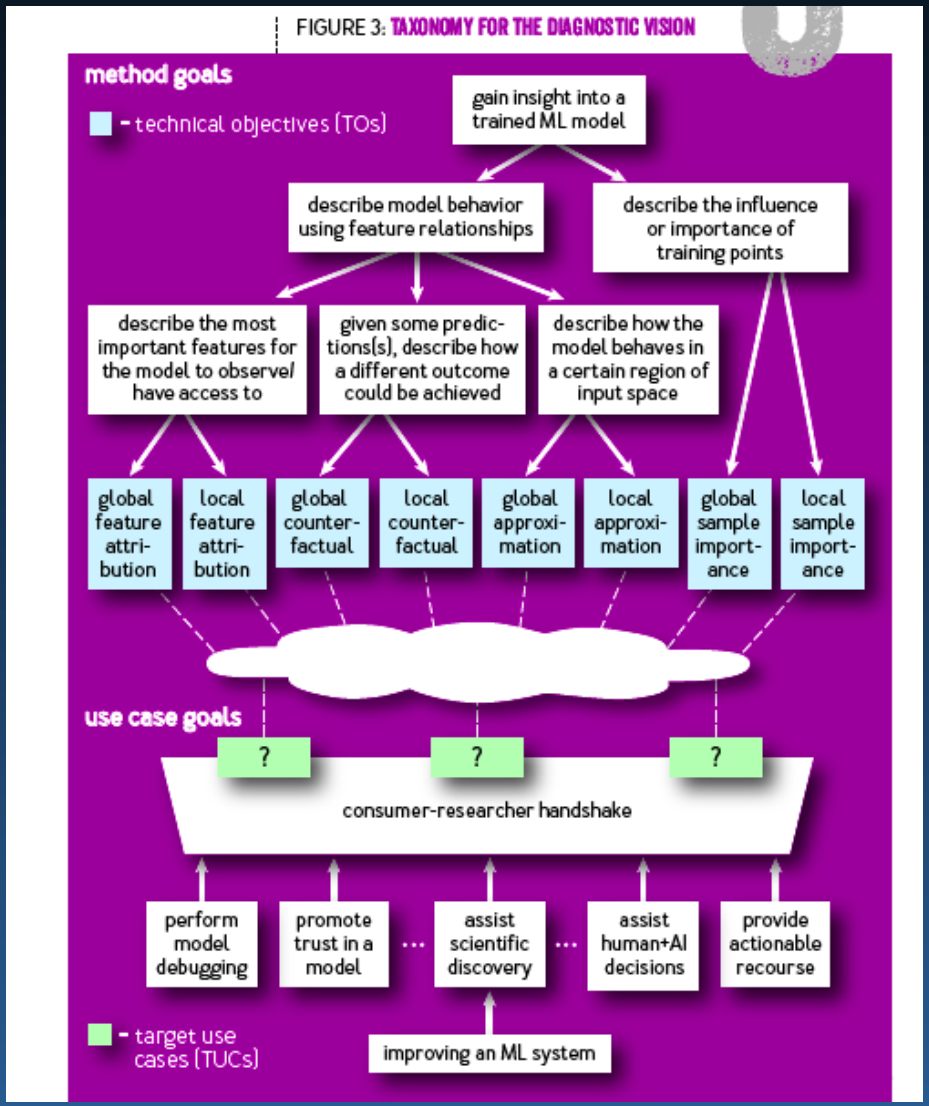
lay image where th
A1-2-3 condition an

chernutter al. terpretable ff-Policy earning via yperbox earch. ML 2022.

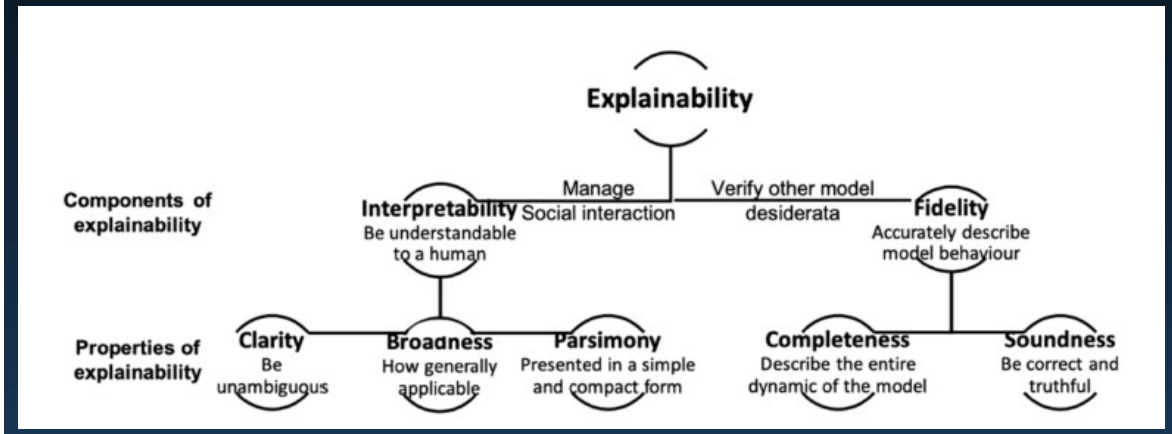
Appendix L.

Work on taxonomies and synthesis

Interpretable ML: Moving from Mythos to Diagnostics
Chen et al. ACMQueue 2021.



Zhou et al. "Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics." Electronics. 2021



Statistics Surveys
Vol. 16 (2022) 1–85
ISSN: 1935-7516
https://doi.org/10.1214/21-SS133

Interpretable machine learning: Fundamental principles and 10 grand challenges*

Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, Chudi Zhong†

e-mail: cynthia@cs.duke.edu; chaofan.chen@maine.edu; zhi.chen1@duke.edu haiyang.huang@duke.edu; lesia.semenova@duke.edu; chudi.zhong@duke.edu

Abstract: Interpretability in machine learning (ML) is crucial for high stakes decisions and troubleshooting. In this work, we provide fundamental principles for interpretable ML, and dispel common misunderstandings that dilute the importance of this crucial topic. We also identify 10 technical challenge areas in interpretable machine learning and provide history and

Plan for these modules:

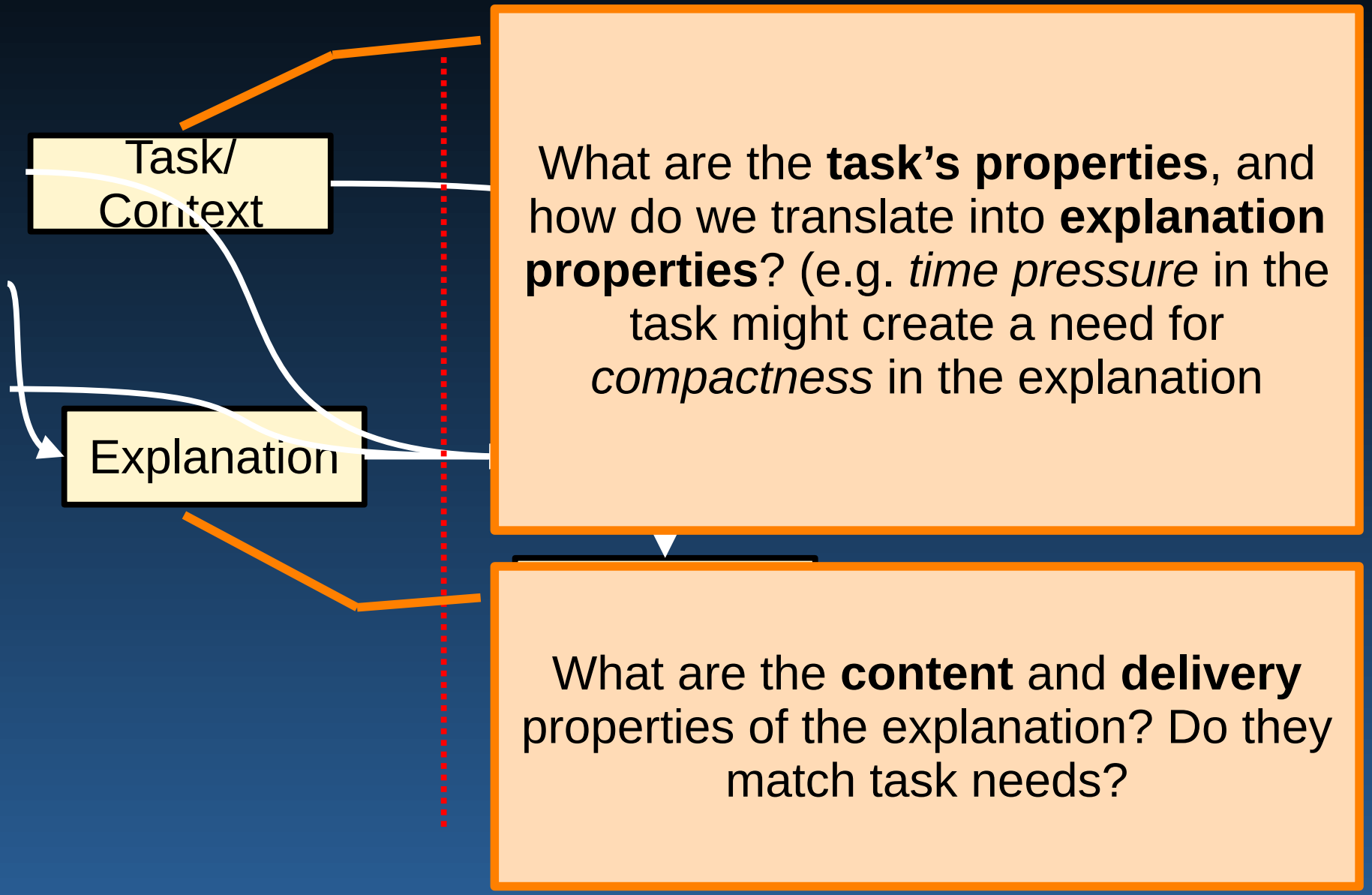
Module 1: Techniques

Module 2: Computational Evaluations

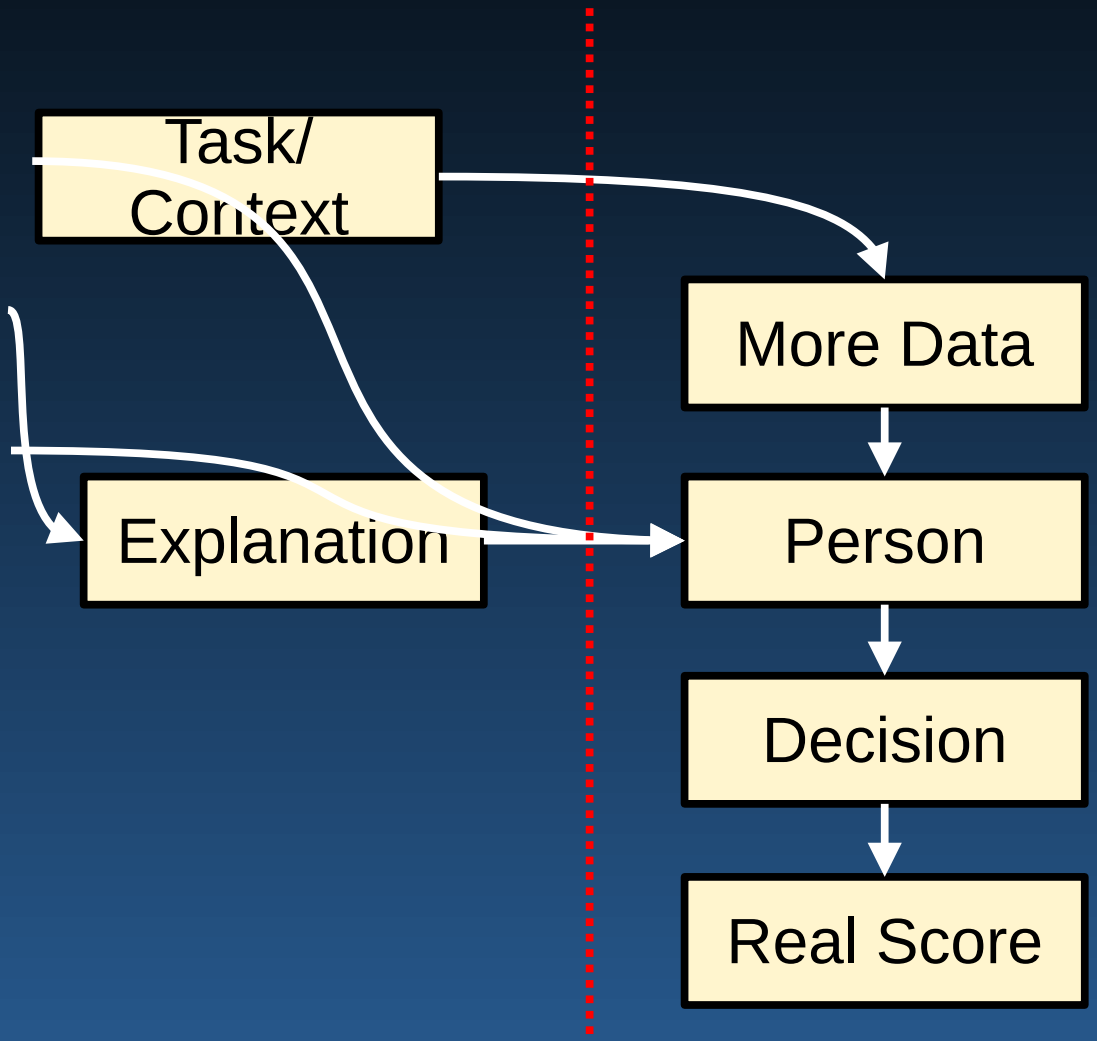
- Conceptualizing evaluation
- Computational properties

Module 3: Human Factors

Conjecture:
Properties as Abstractions for Generalization

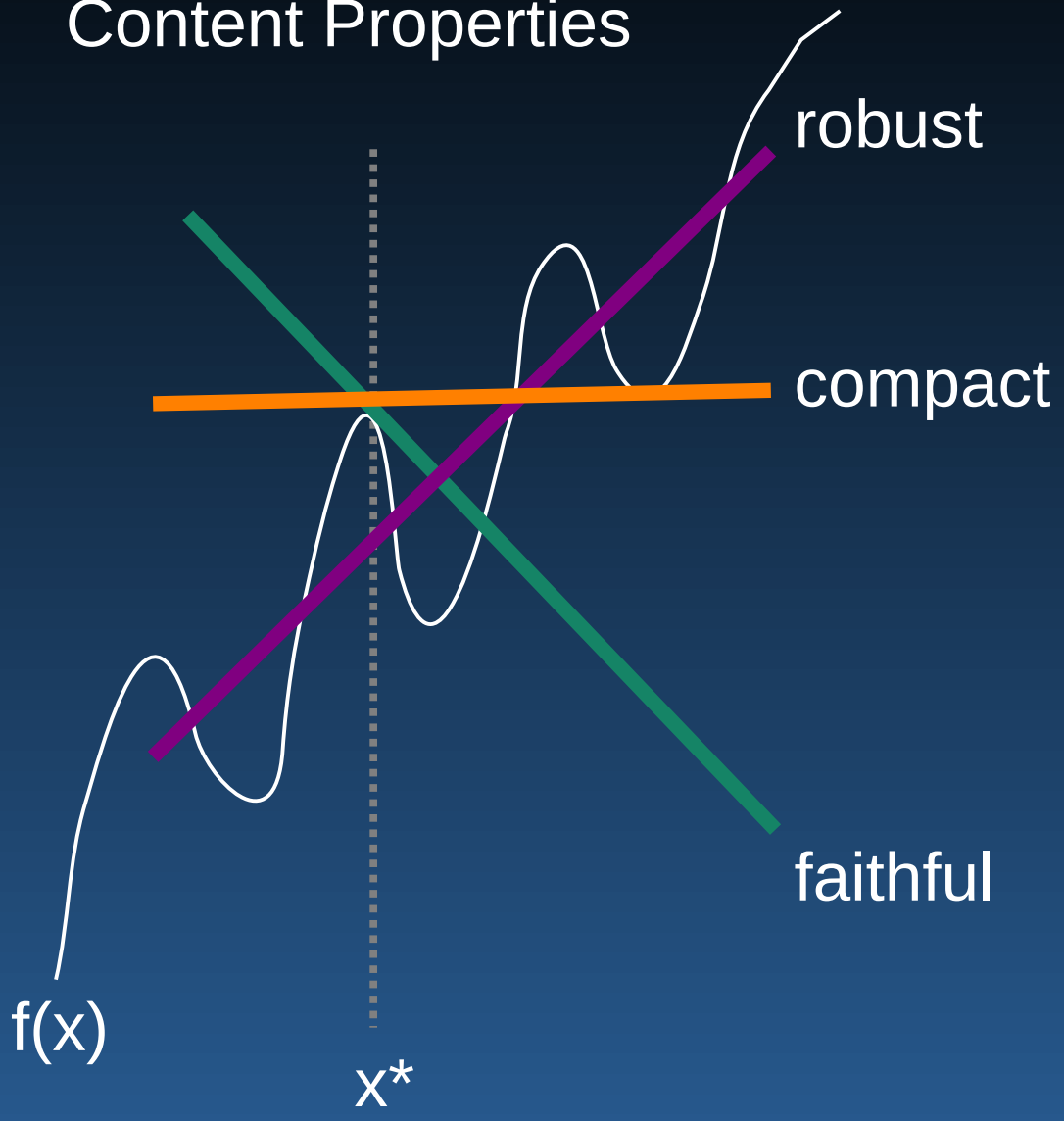


Conjecture:
Properties as Abstractions for Generalization



Examples of Explanation Properties

Content Properties

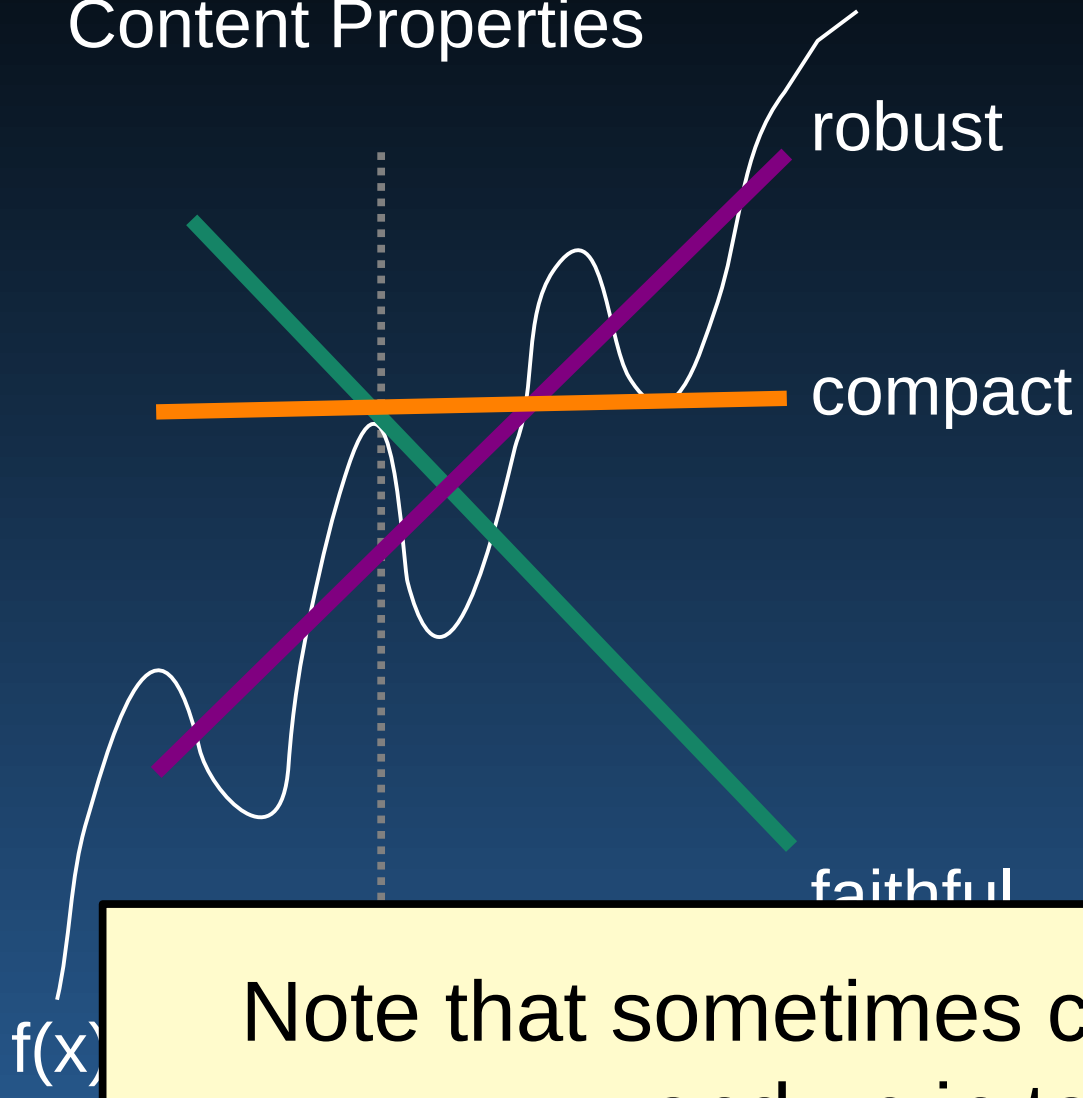


Delivery Properties

- When to provide the explanation (e.g. let the user think first).
- Whether to provide a recommendation, or explanation only.
- Interactivity to adjust and dig deeper.

Examples of Explanation Properties

Content Properties



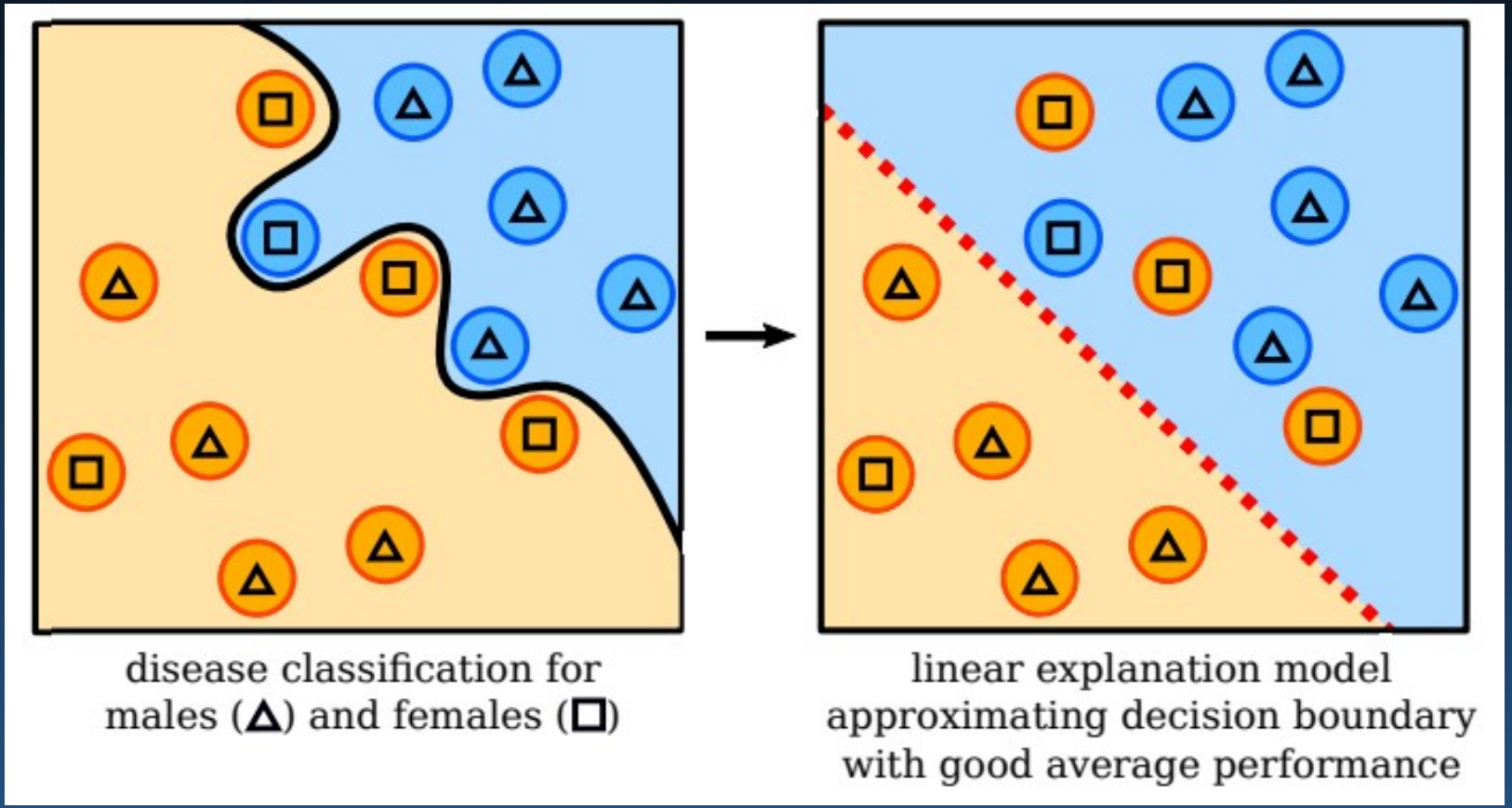
Delivery Properties

- When to provide the explanation (e.g. let the user think first).
- Whether to provide a recommendation, or explanation only.
- Interactivity to adjust and dig deeper.

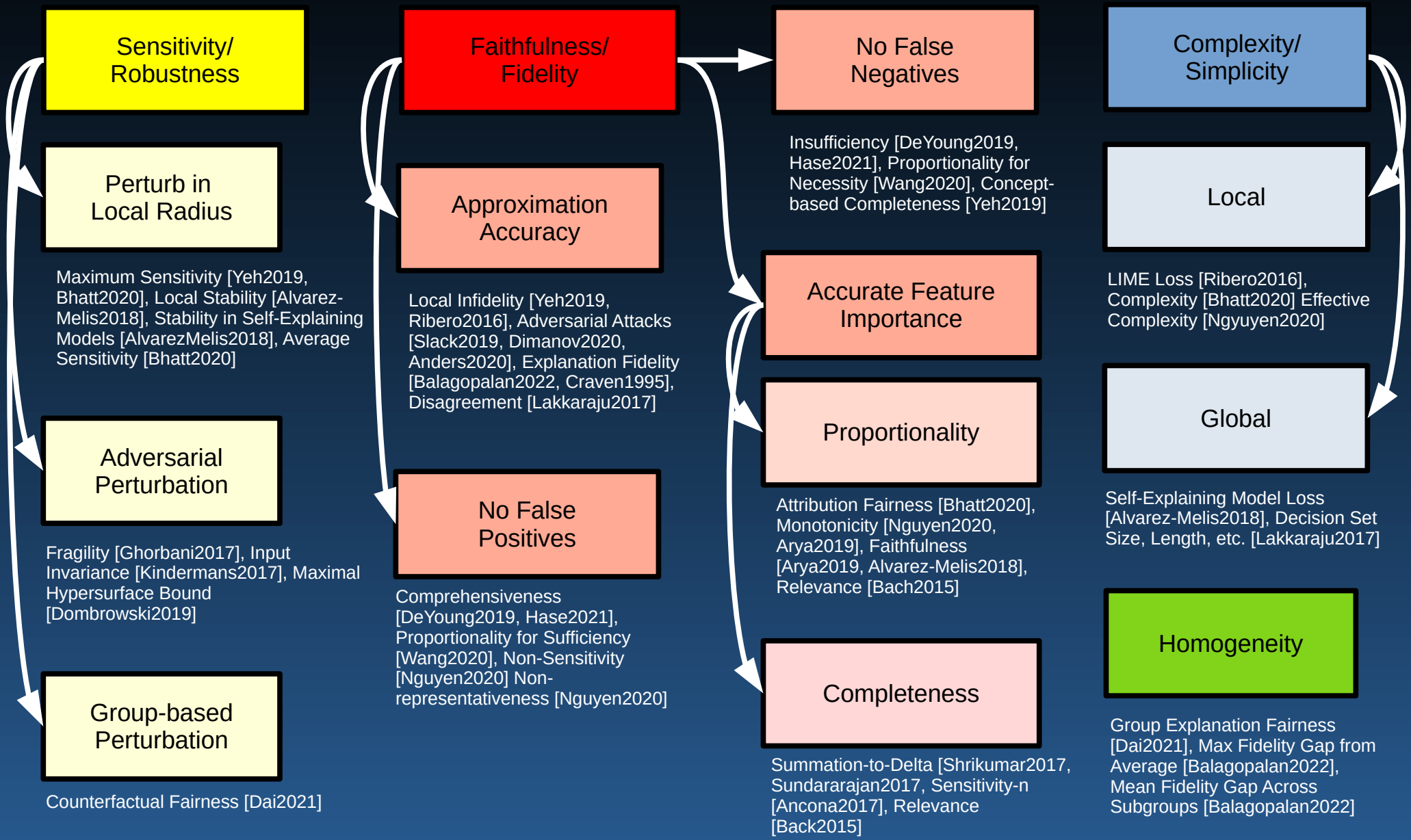
Note that sometimes content properties end up in tension!

Another example of properties in tension:

What is the tension here?



Also, so many properties!



Subhash et al. 2024

Property-Optimized Explanations

Current explanation methods tend to be defined as some forward computation: e.g. given an input, take the derivatives around nearby sampled points and average (smoothgrad).

Even when there's a connection to a loss e.g. different explanations minimize a particular infidelity loss under different perturbations (Yeh 2019), it may not be the desired formalization nor can we manage trade-offs.

What if explanations were **defined** by what properties they were optimized for?

Example: Fidelity, Robustness, Smoothness

We start by writing forms for the transductive setting in which we know all N points for which we desire explanations:

$$\mathcal{L}^{\text{transductive}}_{\text{Faithful}}(\mathbf{W}_{\text{E}}) = \sum_{n=1}^{N} \|W_{\text{E}_n} - \nabla f(\mathbf{x}_n)\|_2^2$$

$$\mathcal{L}^{\text{transductive}}_{\text{Robust}}(\mathbf{W}_{\text{E}}) = \sum_{n=1}^{N} \sum_{n'=1}^{N} \|W_{\text{E}_n} - W_{\text{E}_{n'}}\|_2^2 \, \text{Sim}_{n,n'}$$

$$\mathcal{L}^{\text{transductive}}_{\text{Smooth}}(\mathbf{W}_{E}) = \sum_{N=1}^{N} \sum_{i=1}^{D} \sum_{j=1}^{D} \|(W_{E_n})_i - (W_{E_n})_j\|_2^2 \text{Sim}'_{ij}$$

Example: Fidelity, Robustness, Smoothness

We start by writing forms for the transductive setting in which we know all N points for which we desire explanations:

$$\mathcal{L}^{\text{transductive}}_{\text{Faithful}}(\mathbf{W}_{\text{E}}) = \sum_{n=1}^{N} \|W_{\text{E}_n} - \nabla f(\mathbf{x}_n)\|_2^2$$

$$\mathcal{L}^{\text{transductive}}_{\text{Robust}}(\mathbf{W}_{\text{E}}) = \sum_{n=1}^{N} \sum_{n'=1}^{N} \|W_{\text{E}_n} - W_{\text{E}_{n'}}\|_2^2 \, \text{Sim}_{n,n'}$$

$$\mathcal{L}^{\text{transductive}}_{\text{Smooth}}(\mathbf{W}_{E}) = \sum_{N=1}^{N} \; D \; D$$

Note: Can also efficiently optimize other property forms with QCQPs

Tadesse and Huyuk et al. UAI 2025

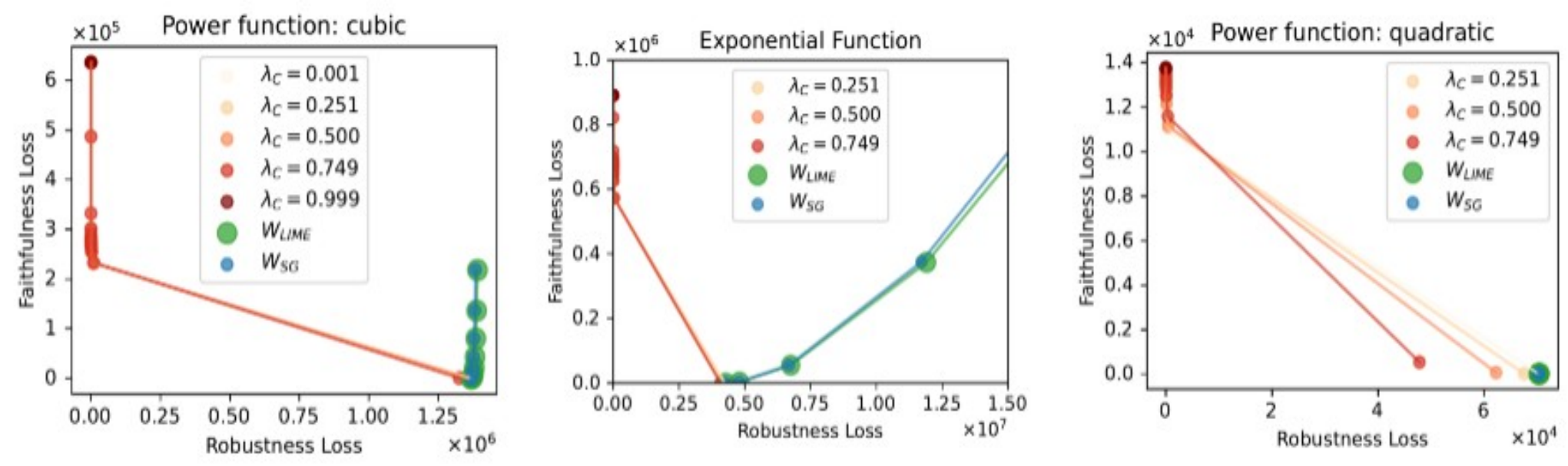
Optimization in the Transductive Setting

For any trade-off between these three objectives, optimization simply involves solving a linear system.

$$
\begin{aligned}
\mathbf{W}^*_{\text{opt}} =& \underset{\mathbf{W}_\text{E}}{\text{argmin}}\ \lambda_F\ \text{Faithful\_loss}(\mathbf{W}_\text{E}) + \lambda_R\ \text{Robust\_loss}_k(\mathbf{W}_\text{E}) \\
&+ \lambda_S\ \text{Smoothness\_loss}_k(\mathbf{W}_\text{E}) + \lambda_C\ \text{Complexity\_loss}(\mathbf{W}_\text{E}) \\
=& \underset{\mathbf{W}_\text{E}}{\text{argmin}}\ \lambda_F \sum_n \|\mathbf{W}_{\text{E}_n} - \nabla f(\mathbf{x}_n)\|_2^2 + \lambda_R \sum_n \sum_{n'} \|\mathbf{W}_{\text{E}_n} - \mathbf{W}_{\text{E}_{n'}}\|_2^2\ \text{Sim}_{n,n'} \\
&+ \lambda_S \sum_n \sum_{i=1}^{D} \sum_{j=1}^{D} \|(W_{E_n})_i - (W_{E_n})_j\|_2^2\ \text{Sim}'_{i,j} + \lambda_C \sum_n \|\mathbf{W}_{\text{E}_n}\|_1
\end{aligned}
$$

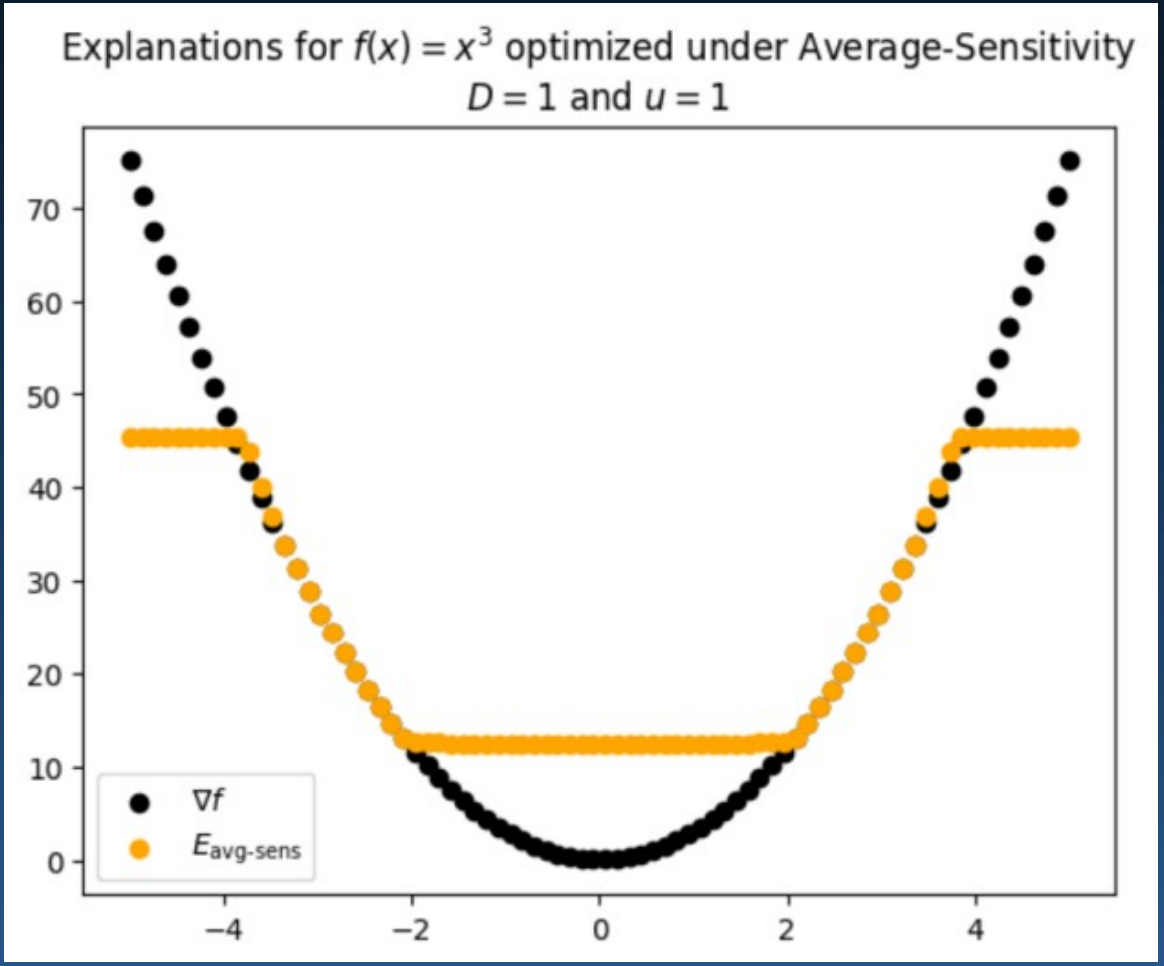
Optimizing results in better properties

For any trade-off between these three objectives, optimization simply involves solving a linear system.



However, be careful as you optimize!

Actual solution (yellow) for a quadratic, optimized for both fidelity and robustness: explanations pick and choose which!



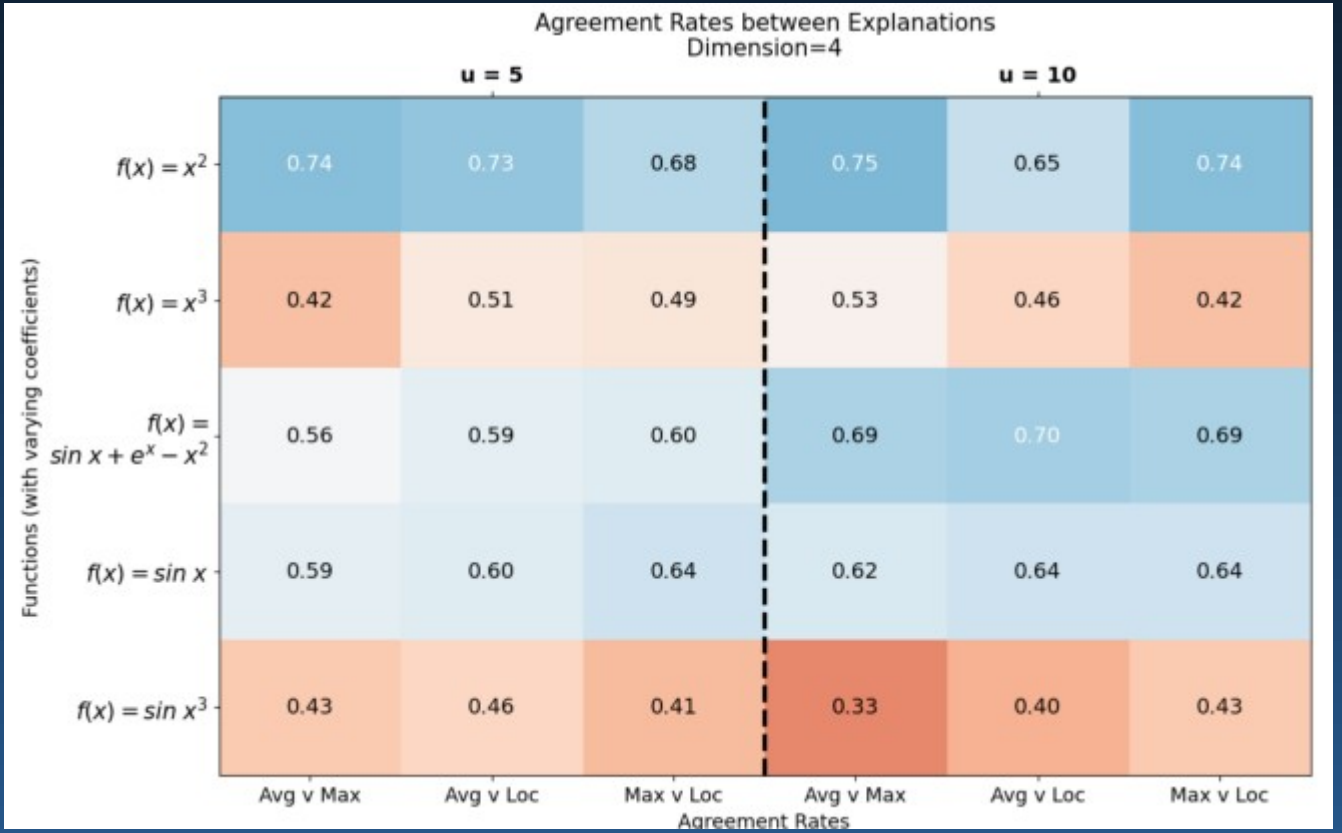
Different formalizations of properties yield different solutions

Explored three different definitions of robustness alongside the gradient definition of fidelity.

Metric	Formalization
Faithfulness	$\mathcal{L}_{\text{faithful}}(\mathbf{E}) = \sum_{n=1}^{N} \|\nabla \mathbf{f_n} - \mathbf{E_n}\|_2^2$
Max-Sensitivity (Yeh et al., 2022)	$\mathcal{L}_{\text{max\_sensitivity}}(\mathbf{E}) = \sum_{n=1}^{N} \max_{\substack{n' \in N s.t. \\ \|X_n - X_{n'}\| \leq u}} \|\mathbf{E_n} - \mathbf{E_{n'}}\|$
Local-Stability (Alvarez-Melis & Jaakkola, 2018a)	$\mathcal{L}_{\text{local\_stability}}(\mathbf{E}) = \sum_{n=1}^{N} \max_{\substack{n' \in N s.t. \\ \|X_n - X_{n'}\| \leq u}} \frac{\|\mathbf{E_n} - \mathbf{E_{n'}}\|}{\|X_n - X_{n'}\|}$
Average-Sensitivity (Bhatt et al., 2020a)	$\mathcal{L}_{\text{average\_sensitivity}}(\mathbf{E}) = \sum_{n=1}^{N} \sum_{n'=1}^{N} \|\mathbf{E_n} - \mathbf{E_{n'}}\|^2 \cdot p_n(n')$

Different formalizations of properties yield different solutions

Explored three different definitions of robustness alongside the gradient definition of fidelity.



… mapping properties to tasks

Scenario: InvestAssistant is a new AI system that assists personal finance by recommending investment plans (e.g. stocks to sell and buy). The AI system monitors the financial market and considers a user's financial profile and personal preferences such as risk tolerance in order to make recommendations. It learns from historical investment records.

The AI system can generate explanations, e.g. how it judges the risk and return of investment plans, and why it recommends a particular investment plan.

Task 1 of 5

You are an investor and were just introduced to the AI system. While testing the system, you request explanations to evaluate whether the AI's capability is good enough for you to use; e.g., whether the way it judges the risk and returns of investment plans is reasonable, reliable or a aligns with your investment knowledge.

For the task described above, rate how important the model explanations should satisfy each criterion. If you feel a criterion could be compromised or is not/less applicable to the task, you can rate it as less important.

	Not important	Slightly important	Moderately Important	Important	Very important
Translucence: the explanation is transparent about its limitations, for examples the conditions for it to hold	○	○	○	○	○

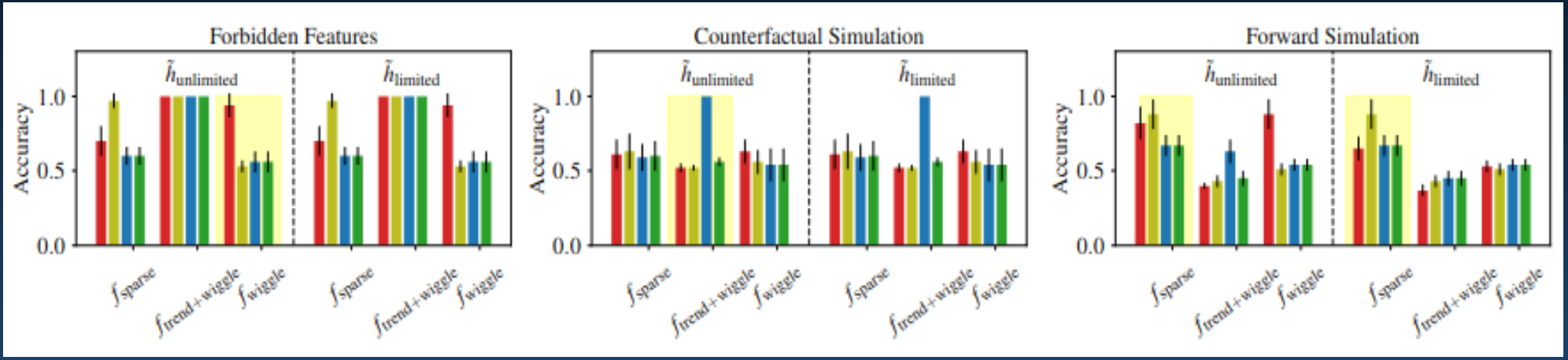
Liao et al. 2022, https://arxiv.org/abs/2206.10847

… mapping properties to tasks: surveys

	Model improvement		Capability assessment		Decision support		Adapting control		Domain learning		Model auditing		Overall	
						Expert Survey								
1	**Faithfulness**	4.61	**Translucence**	4.40	**Interactivity**	4.29	**(Un)Certainty**	4.31	**Translucence**	4.23	**Faithfulness**	4.54	**Faithfulness**	4.30
2	**Stability**	4.27	**(Un)Certainty**	4.29	**(Un)Certainty**	4.26	**Interactivity**	4.31	**Faithfulness**	4.20	**Translucence**	4.43	**Translucence**	4.25
3	(Un)Certainty	4.24	**Comprehen.**	4.20	**Faithfulness**	4.14	**Translucence**	4.29	**Stability**	4.17	**Completeness**	4.37	(Un)Certainty	4.15
4	Translucence	4.06	**Faithfulness**	4.09	**Translucence**	4.11	**Faithfulness**	4.26	**Comprehensibility**	4.14	(Un)Certainty	4.00	Stability	4.01
5	Comprehen.	3.94	**Stability**	4.00	**Comprehen.**	3.97	**Stability**	4.00	**Completeness**	4.00	Interactivity	3.89	Interactivity	3.92
6	Completeness	3.67	**Compactness**	4.00	**Actionability**	3.91	**Actionability**	3.86	**Interactivity**	3.86	Stability	3.83	Comprehen.	3.89
7	Actionability	3.48	Interactivity	3.80	Stability	3.83	Completeness	3.80	(Un)Certainty	3.83	Comprehen.	3.49	Completeness	3.70
8	Compactness	3.42	Actionability	3.69	**Personalization**	3.43	Comprehen.	3.63	Compactness	3.60	Actionability	3.06	Actionability	3.55
9	Interactivity	3.33	Personalization	3.63	Compactness	3.31	Coherence	3.46	Personalization	3.57	Compactness	3.03	Compactness	3.45
10	Coherence	3.18	Coherence	3.54	Novelty	3.20	Compactness	3.31	Actionability	3.29	Coherence	2.97	Coherence	3.23
11	Novelty	2.88	Completeness	3.23	Coherence	3.17	Personalization	2.94	Novelty	3.17	Personalization	2.57	Personalization	3.14
12	Personalization	2.67	Novelty	2.51	Completeness	3.11	Novelty	2.91	Coherence	3.03	Novelty	2.37	Novelty	2.84
						End-User Survey								
1			**Comprehen.**	4.28	**(Un)Certainty**	4.38	**Translucence**	4.44	**Stability**	4.38	**Faithfulness**	4.78	**Faithfulness**	4.38
2			**Personalization**	4.28	**Translucence**	4.38	**Faithfulness**	4.38	**Faithfulness**	4.28	Translucence	4.28	**(Un)Certainty**	4.26
3			**(Un)Certainty**	4.19	**Faithfulness**	4.31	**(Un)Certainty**	4.38	**(Un)Certainty**	4.25	**Completeness**	4.19	**Translucence**	4.25
4			**Faithfulness**	4.12	**Comprehen.**	4.19	**Stability**	4.19	**Comprehen.**	4.25	(Un)Certainty	4.12	Stability	4.15
5			**Stability**	4.12	**Actionability**	4.03	**Personalization**	4.09	**Translucence**	4.09	Stability	4.12	Comprehen.	4.05
6			**Translucence**	4.06	Stability	3.94	Actionability	4.03	**Completeness**	4.09	Interactivity	3.97	Interactivity	3.91
7			Actionability	3.88	Interactivity	3.81	Interactivity	3.97	**Personalization**	4.06	Comprehen.	3.72	Actionability	3.89
8			Coherence	3.84	Personalization	3.75	Comprehen.	3.81	**Interactivity**	4.03	Actionability	3.66	Personalization	3.88
9			Interactivity	3.75	Coherence	3.72	Completeness	3.72	Actionability	3.84	Coherence	3.59	Completeness	3.86
10			Completeness	3.69	Completeness	3.59	Coherence	3.66	Coherence	3.72	Personalization	3.19	Coherence	3.71
11			Compactness	3.12	Novelty	3.16	Novelty	2.88	Compactness	3.59	Novelty	3.12	Novelty	3.04
12			Novelty	3.00	Compactness	2.84	Compactness	2.78	Novelty	3.06	Compactness	2.62	Compactness	2.98
						Combined Results								
1			**Comprehen.**	4.24	**(Un)Certainty**	4.31	**Translucence**	4.46	**Stability**	4.27	**Faithfulness**	4.66	**Faithfulness**	4.31
2			**(Un)Certainty**	4.24	**Translucence**	4.24	**(Un)Certainty**	4.34	**Faithfulness**	4.24	Translucence	4.36	**Translucence**	4.27
3			**Translucence**	4.24	**Faithfulness**	4.22	**Faithfulness**	4.31	**Comprehen.**	4.19	**Completeness**	4.28	(Un)Certainty	4.20
4			**Faithfulness**	4.10	**Comprehen.**	4.07	**Interactivity**	4.15	**Translucence**	4.16	(Un)Certainty	4.06	Stability	4.05
5			**Stability**	4.06	**Interactivity**	4.06	**Stability**	4.09	**Completeness**	4.04	Stability	3.97	Interactivity	3.97
6			**Personalization**	3.94	**Actionability**	3.97	Actionability	3.94	**(Un)Certainty**	4.03	Interactivity	3.93	Comprehen.	3.96
7			Actionability	3.78	Stability	3.88	Completeness	3.76	**Interactivity**	3.94	Comprehen.	3.60	Completeness	3.78
8			Interactivity	3.78	Personalization	3.58	Comprehen.	3.72	Personalization	3.81	Actionability	3.34	Actionability	3.72
9			Coherence	3.69	Coherence	3.43	Coherence	3.55	Actionability	3.55	Coherence	3.27	Personalization	3.54
10			Compactness	3.58	Completeness	3.34	Personalization	3.49	Compactness	3.55	Personalization	2.87	Coherence	3.46
11			Completeness	3.45	Novelty	3.18	Compactness	3.06	Coherence	3.36	Compactness	2.84	Compactness	3.22

… mapping properties to tasks: simulations

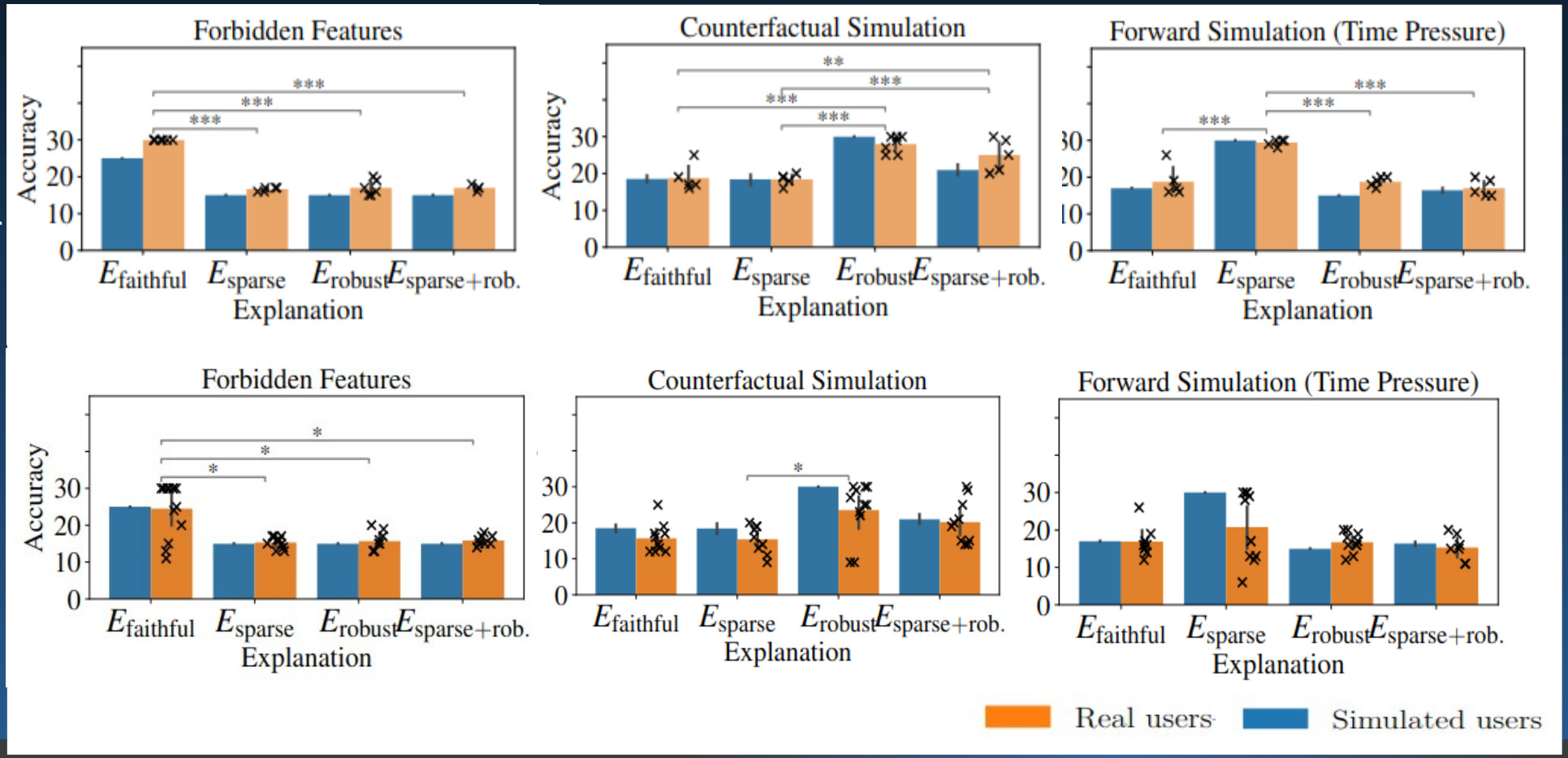
Testing with proxy humans helps us identify when optimizing for properties will matter, for user studies.



Nofshin and Brown et al., in submission

… mapping properties to tasks: simulations to reality

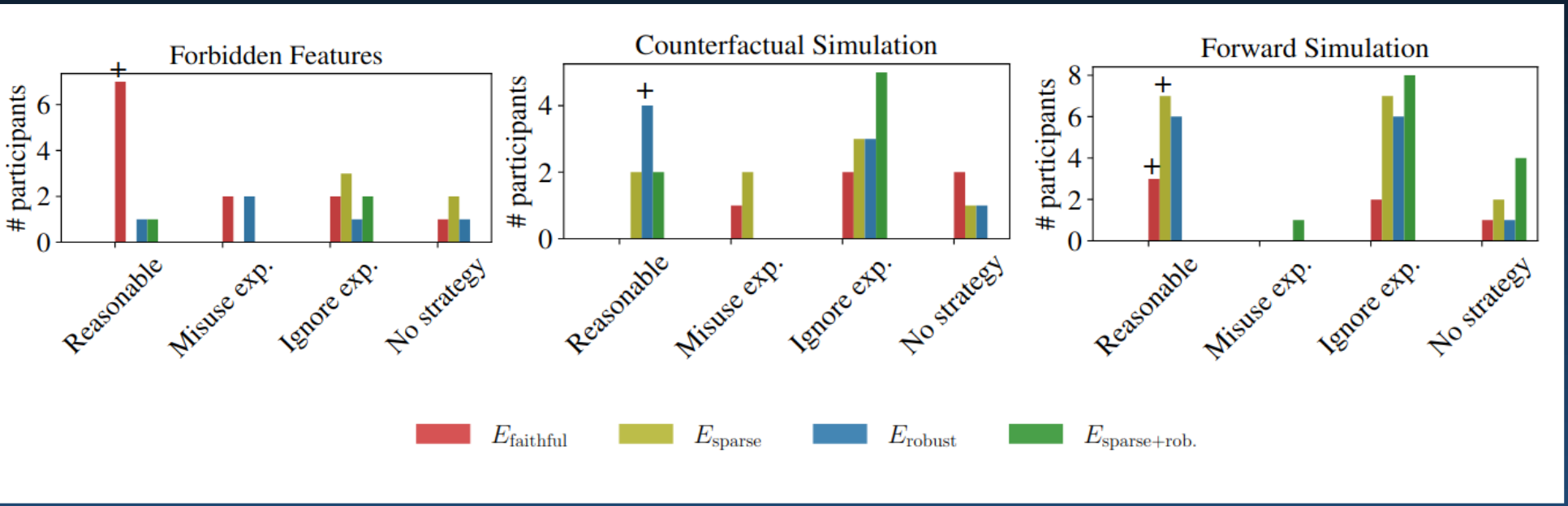
Testing with proxy humans helps us identify when optimizing for properties will matter, for user studies – and then validate!



Nofshin and Brown et al., in submission

… mapping properties to tasks: simulations to reality

Testing with proxy humans helps us identify when optimizing for properties will matter, for user studies – and then validate!



Nofshin and Brown et al., in submission

… mapping properties to tasks: simulations to reality

Measurements for alien #GP2U2

- Core temperature (originally 4.04) changed by **-15**
- Pulse rate (originally 0.13) changed by **0**
- Antenna length (originally 1.94) changed by **0**
- Glow (originally 2.22) changed by **0**
- Hearing score (originally 2.53) changed by **0**
- Skin moisture (originally 1.28) changed by **0**
- Eye reflex (originally 1) changed by **0**

Helpful information from the alien researcher

The alien researcher thinks the the measurements affect the risk score in this way:

- Core temperature: **20**
- Pulse rate: **0**
- Antenna length: **-20**
- Glow: **0**
- Hearing score: **0**
- Skin moisture: **0**
- Eye reflex: **0**

*Note: the closer to 0, the smaller the influence on the risk score.

Old risk score

63

Now that the measurements have changed, will the risk score go up or down?

- ○ Up
- ○ Down

Nofshin and Brown et al., in submission

Human Factors

Plan for these modules:

Module 1: Techniques

Module 2: Computational Evaluations

Module 3: Human Factors

- Ways in which people's behavior may be unexpected
- How to improve human+AI performance

Plan for these modules:

Module 1: Techniques

Module 2: Computational Evaluations

Module 3: Human Factors

- Ways in which people's behavior may be unexpected
- How to improve human+AI performance

Key points for today:

People don’t always interact with explanations in a fully engaged, analytical fashion.

Indeed, there’s research showing that just the presence of an explanation can make a system appear more authoratative!

In addition, people are not the same across individuals and across time.

Sheepdog or mop?

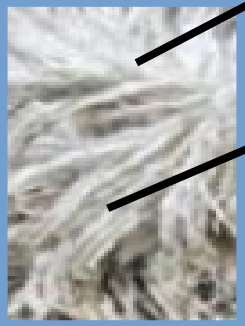
https://www.beano.com/random/fun/sheepdog-or-mop

Sheepdog or mop?

https://www.beano.com/random/fun/sheepdog-or-mop

Sheepdog or mop?

AI: Dog

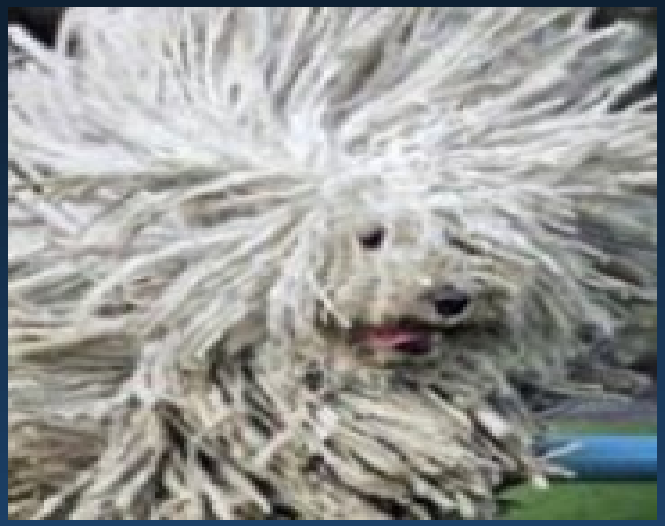


XAI: Mop

https://www.beano.com/random/fun/sheepdog-or-mop

Sheepdog or mop?

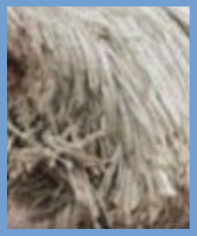
AI: Dog



XAI: Mop

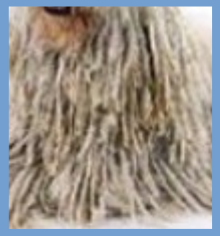
https://www.beano.com/random/fun/sheepdog-or-mop

AI:Dog

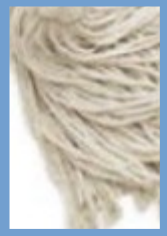


AI:Mop

AI:Dog



AI:Mop

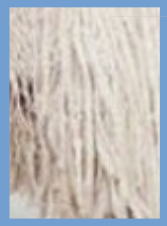


AI:Mop

AI:Dog

AI:Mop

AI:Dog



AI:Mop

AI:Dog

AI:Dog

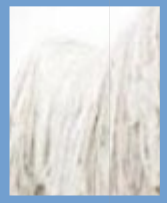
AI:Dog

AI:Mop

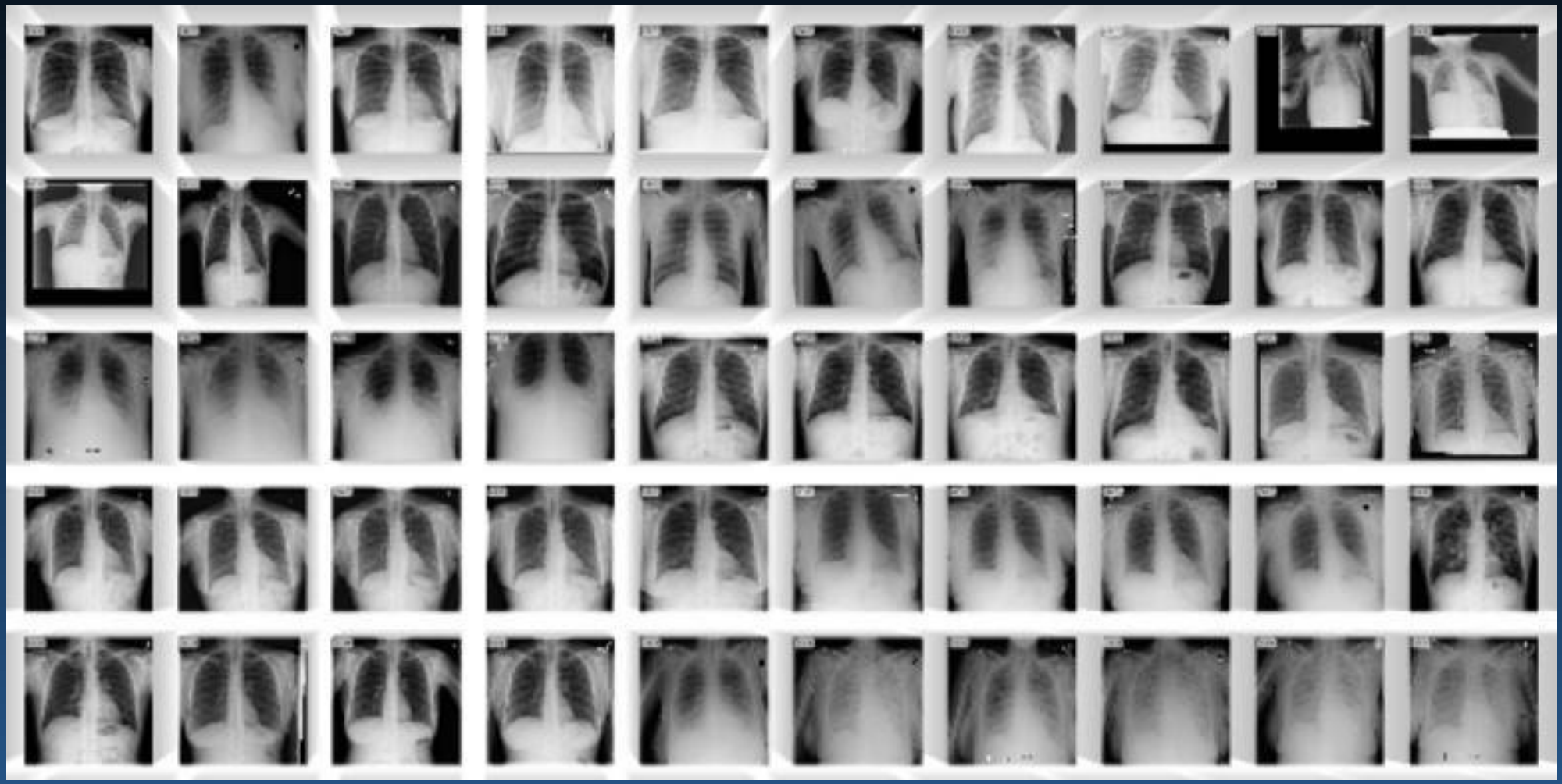
AI:Dog

AI:Mop

AI:Mop



A slightly more important problem…



https://datasets.activeloop.ai/docs/ml/datasets/nih-chest-x-ray-dataset/

Can we do better than human or AI alone?

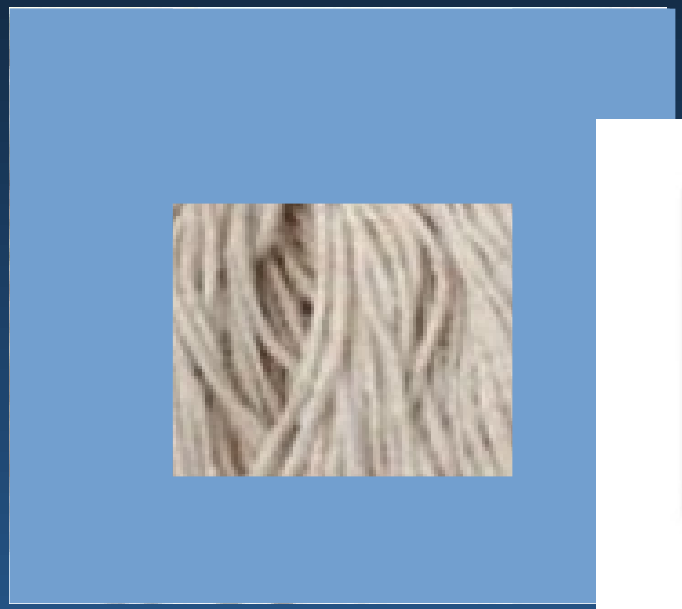
Human+AI teams may help if strengths are complementary.
Humans know goals, extra facts, can notice data issues --
    but: may be inattentive and cannot process lots of info.
AIs can process lots of data and manage probabilities --
    but: data may be incomplete and task not fully specified.

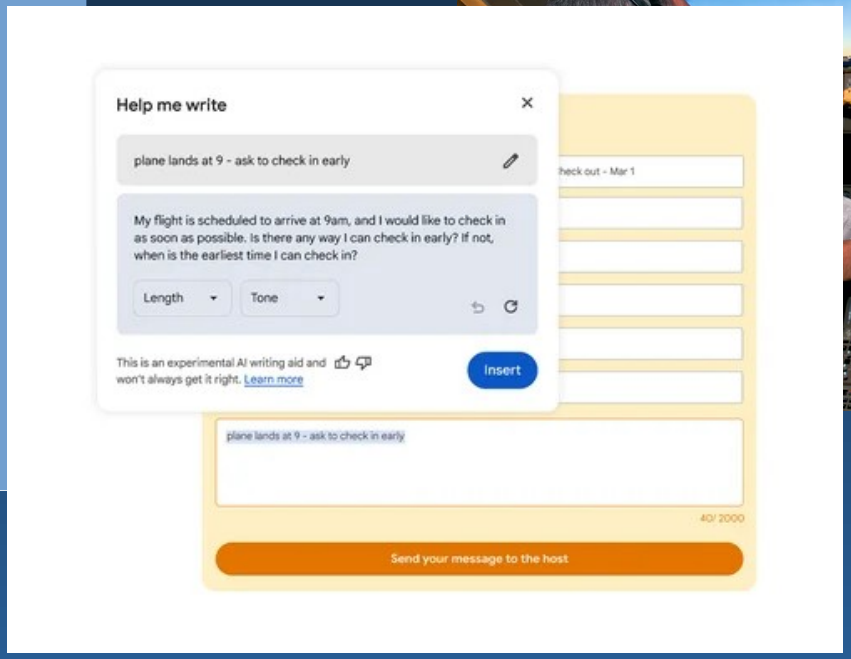
Will we do better than human or AI alone?

When a tool mostly seems to work – often quite well! – and we are busy, we may stop paying attention. AIs also often appear objective and authoritative.

These factors lead to overreliance on AIs.

(Or, less often, people ignoring them)



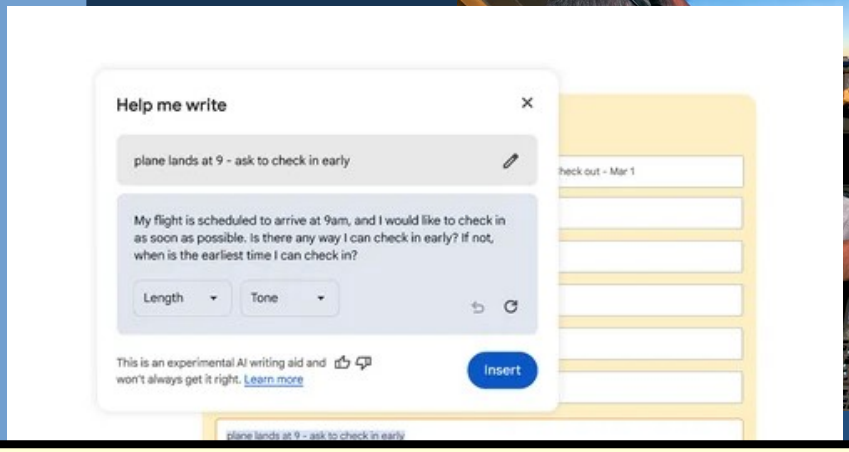


Will we do better than human or AI alone?

When a tool mostly seems to work – often quite well! – and we are busy, we may stop paying attention. AIs also often appear objective and authoritative.

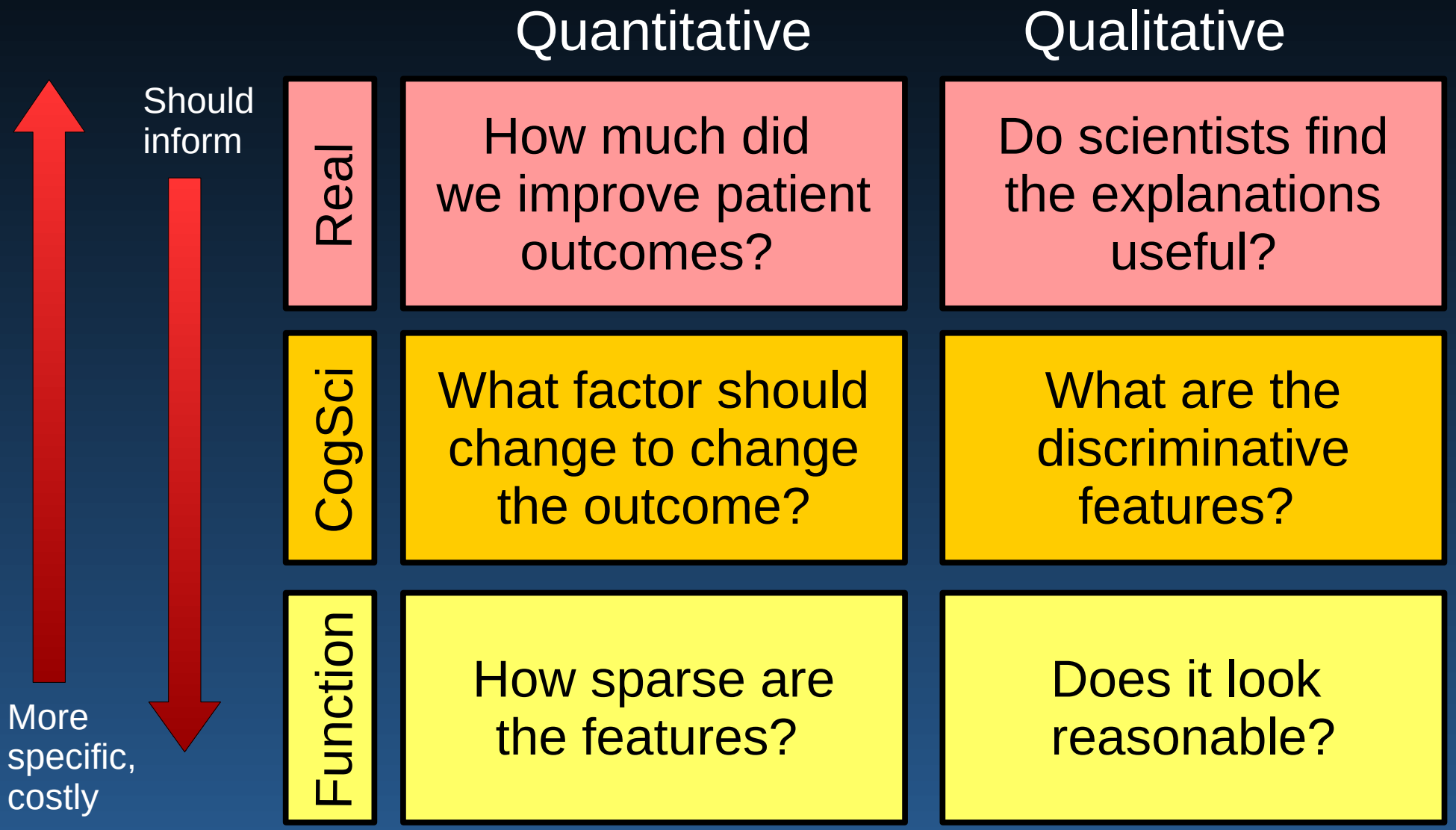
These factors lead to overreliance on AIs.

(Or, less often, people ignoring them)



Short answer: it’s not simple!

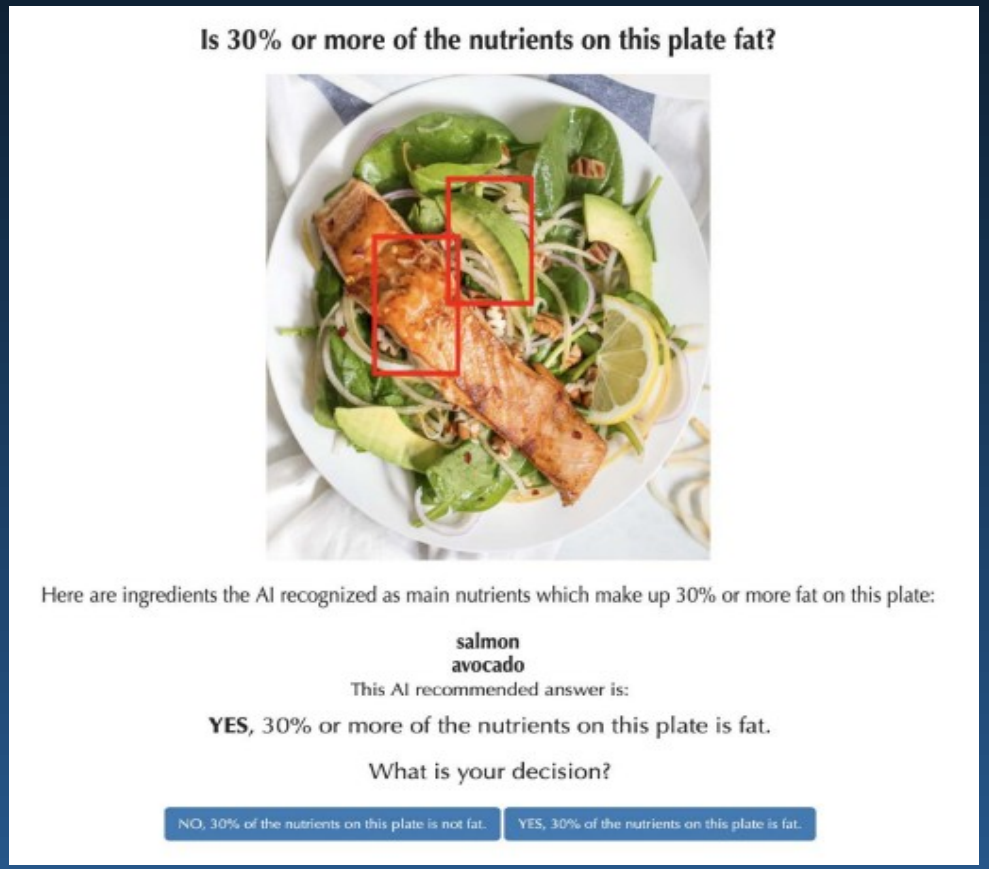
A spectrum of evaluation



Doshi-Velez and Kim, 2017

Different explanations help with different tasks

Imagine two explanation types, one that provides examples and one that provides features:

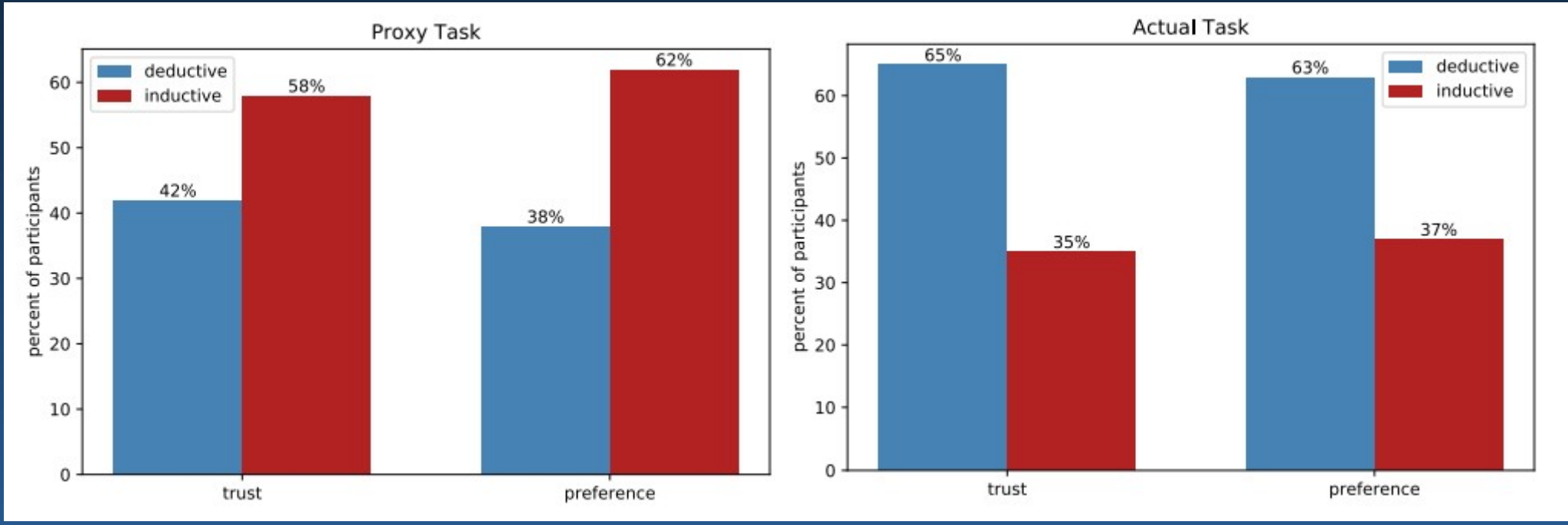


Bucinca, Lin, et al. CHI 2018

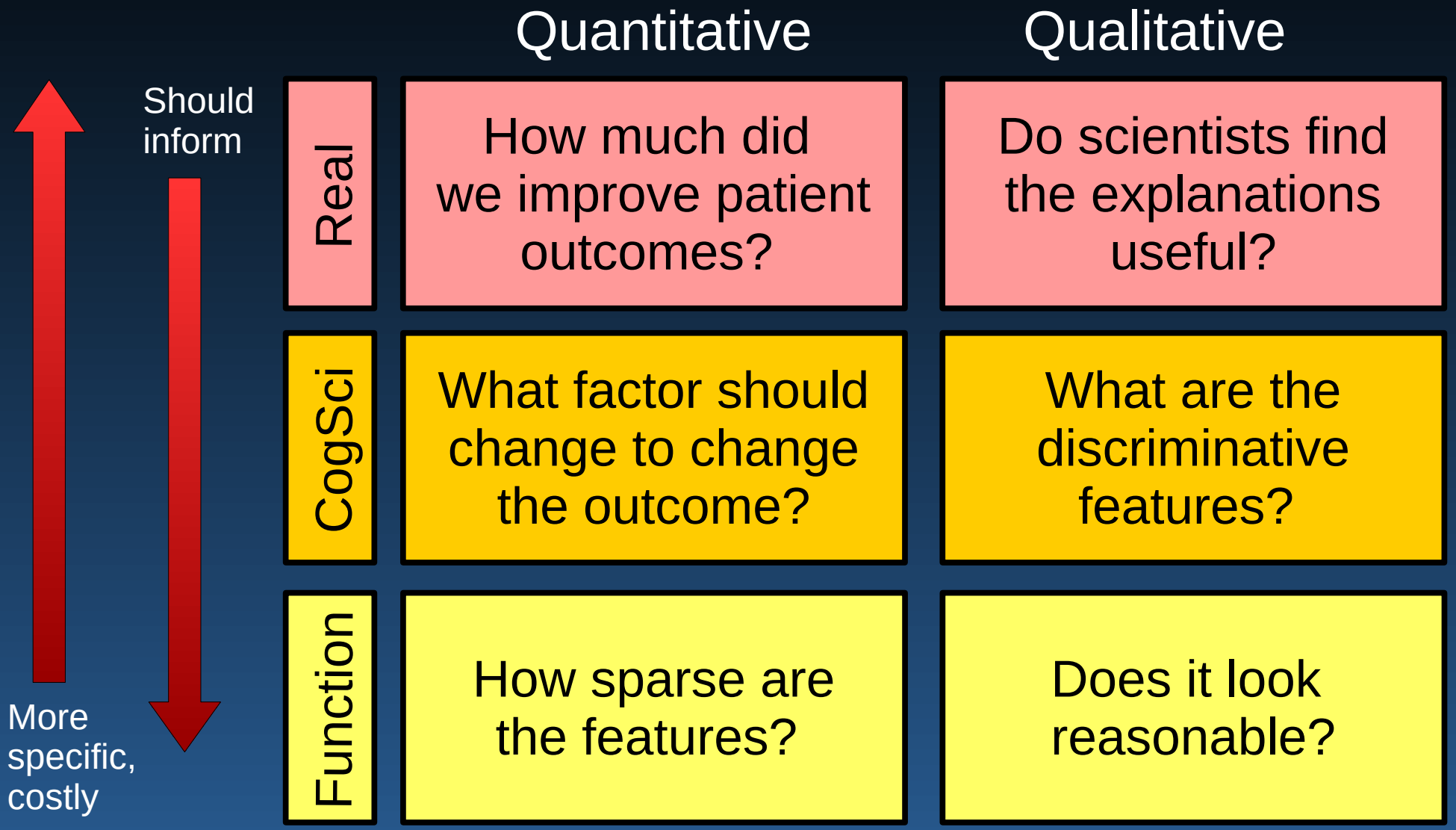
Different explanations help with different tasks

Imagine two explanation types, one that provides examples and one that provides features.

Depending on the task – predicting the AI output, or making the right choice – different explanations help.



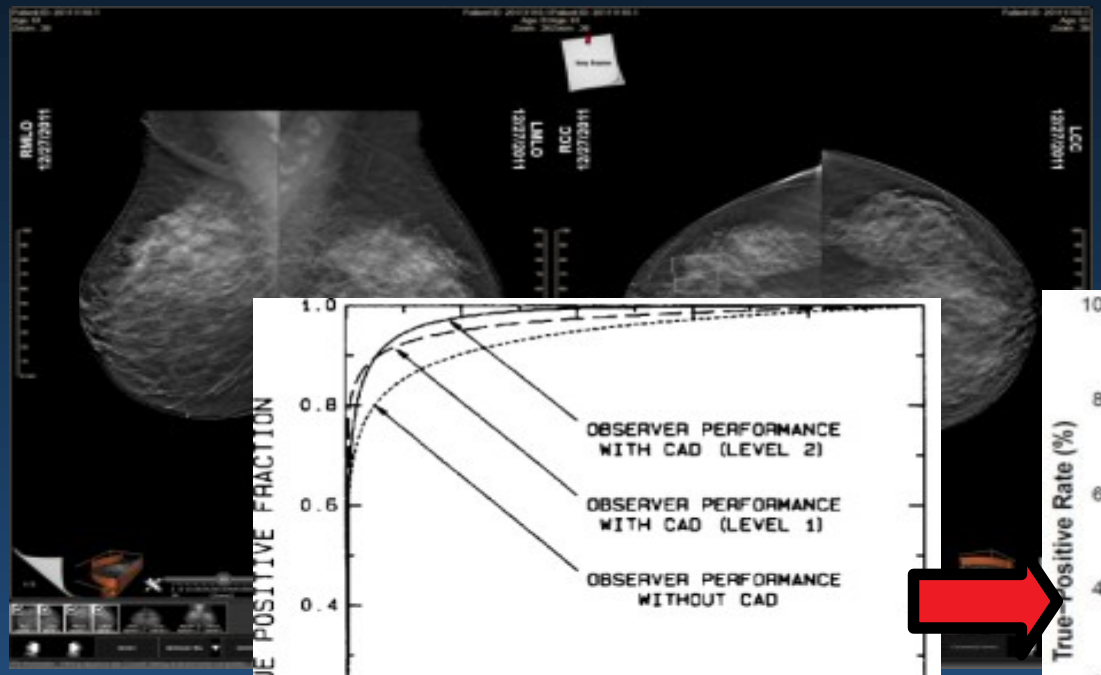
A spectrum of evaluation

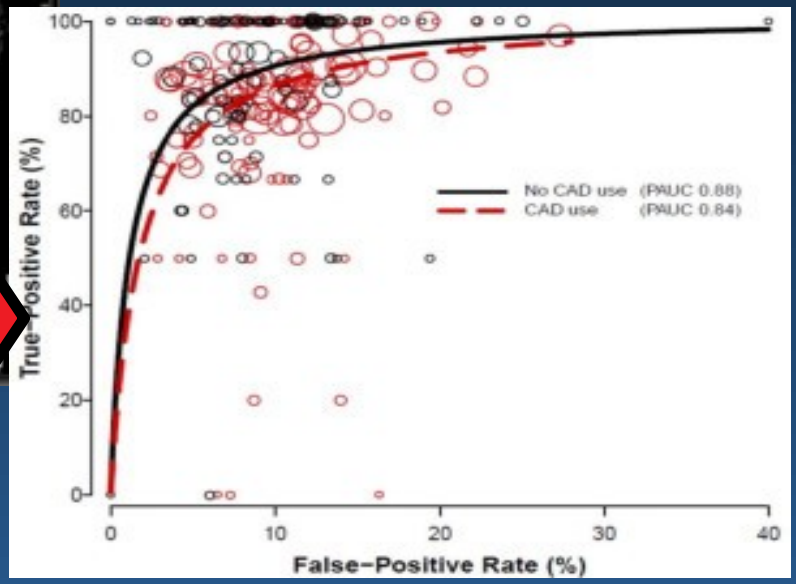


Doshi-Velez and Kim, 2017

This happens in real life!

Many studies on human+AI interaction on over-trust and other factors that impede performance.





https://lukeoakdenrayner.wordpress.com/2019/01/21/medical-ai-safety-doing-it-wrong/

More Unexpected Uses: Explanations for Pre-Trial Risk Assessments

Some jurisdictions use pretrial risk assessments to help denoise bail decisions. Our question: can explanation help judges still apply discretion when appropriate?

Case/Charge Information:

Class/Type:	Simple Misdemeanor	Count(s):	1
Statute:	123.46	Booking:	910217
Description:	Public consumption or intoxication		

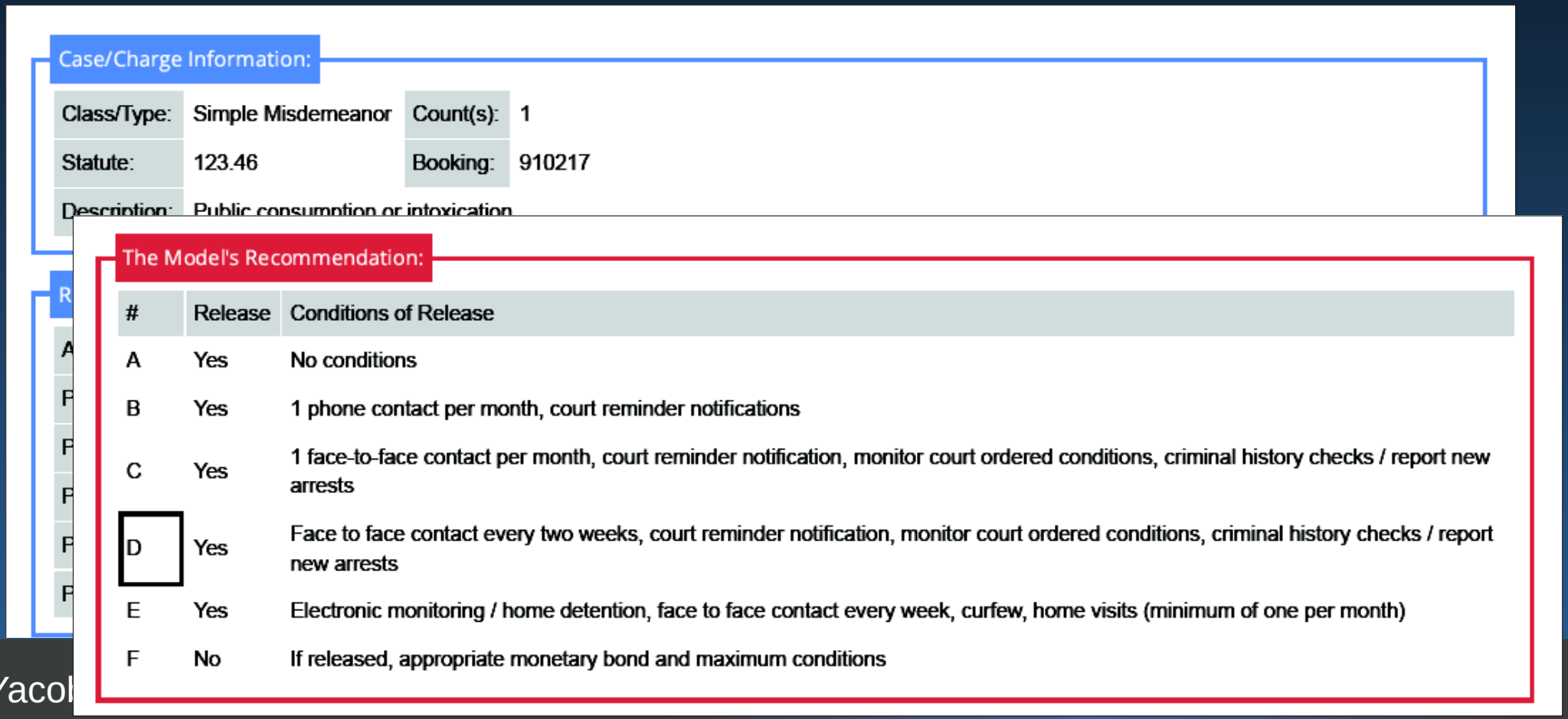
Risk Factors:

Age at current arrest:	23 or Older	Prior felony convictions:	No
Pending charge at the time of the offense:	No	Current violent offense:	No
Prior sentence to incarceration:	Yes	Prior convictions:	Yes
Prior failures to appear pretrial in past 2 years:	Yes, 2 or more	Prior violent convictions:	No
Prior failures to appear pretrial older than 2 years:	Yes	Current violent offense & 20 years old or younger:	No
Prior misdemeanor convictions:	Yes		

Yacoby et al. https://arxiv.org/abs/2205.05424

More Unexpected Uses: Explanations for Pre-Trial Risk Assessments

Some jurisdictions use pretrial risk assessments to help denoise bail decisions. Our question: can explanation help judges still apply discretion when appropriate?



Case/Charge Information:

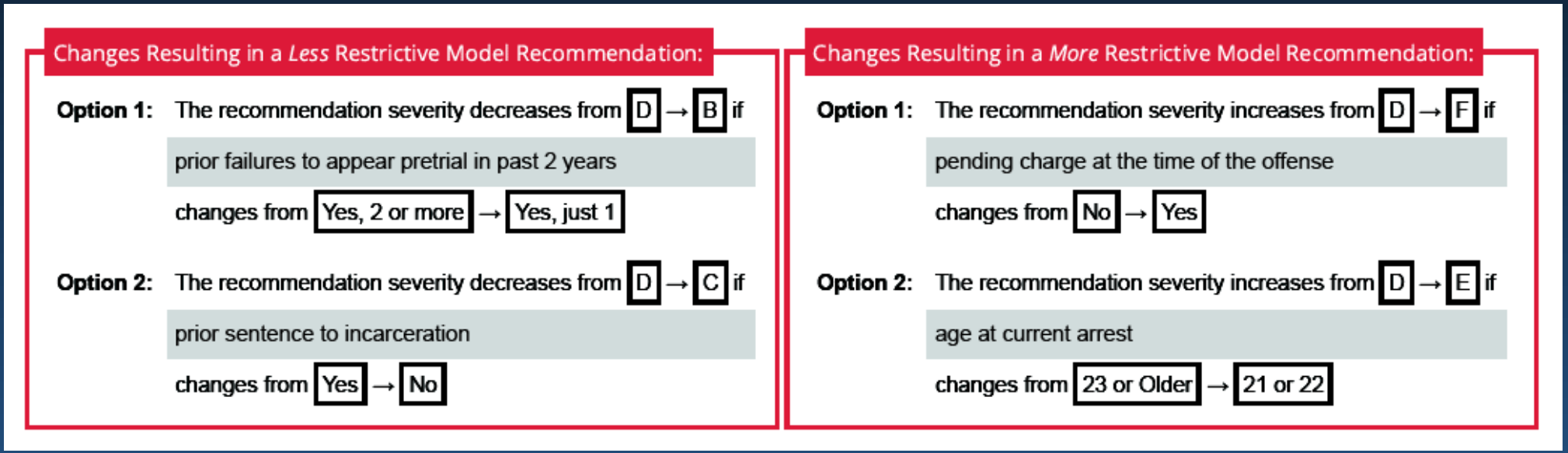
Class/Type:	Simple Misdemeanor	Count(s):	1
Statute:	123.46	Booking:	910217
Description:	Public consumption or intoxication		

The Model's Recommendation:

#	Release	Conditions of Release
A	Yes	No conditions
B	Yes	1 phone contact per month, court reminder notifications
C	Yes	1 face-to-face contact per month, court reminder notification, monitor court ordered conditions, criminal history checks / report new arrests
D	Yes	Face to face contact every two weeks, court reminder notification, monitor court ordered conditions, criminal history checks / report new arrests
E	Yes	Electronic monitoring / home detention, face to face contact every week, curfew, home visits (minimum of one per month)
F	No	If released, appropriate monetary bond and maximum conditions

More Unexpected Uses: Explanations for Pre-Trial Risk Assessments

Some jurisdictions use pretrial risk assessments to help denoise bail decisions. Our question: can explanation help judges still apply discretion when appropriate?



Yacoby et al. https://arxiv.org/abs/2205.05424

More Unexpected Uses: Explanations for Pre-Trial Risk Assessments

Main Result: Judges had a difficult time using counterfactuals to assess the model.

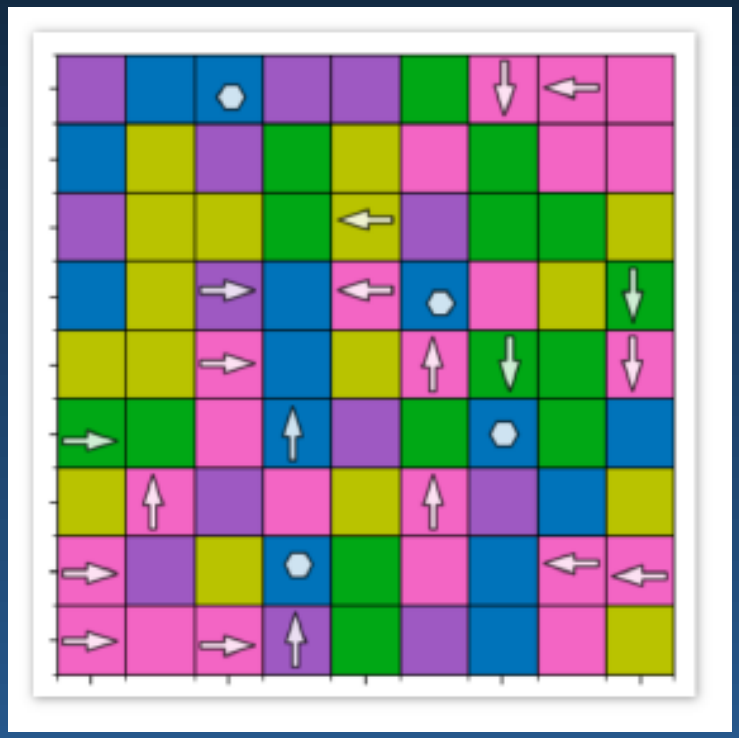
"No [I would not change my decision], because nothing has changed. It's a little awkward thought process… What you are saying is: these didn't happen. If they didn't happen, which they didn't happen… why would I change my decision?"

Yacoby et al. https://arxiv.org/abs/2205.05424

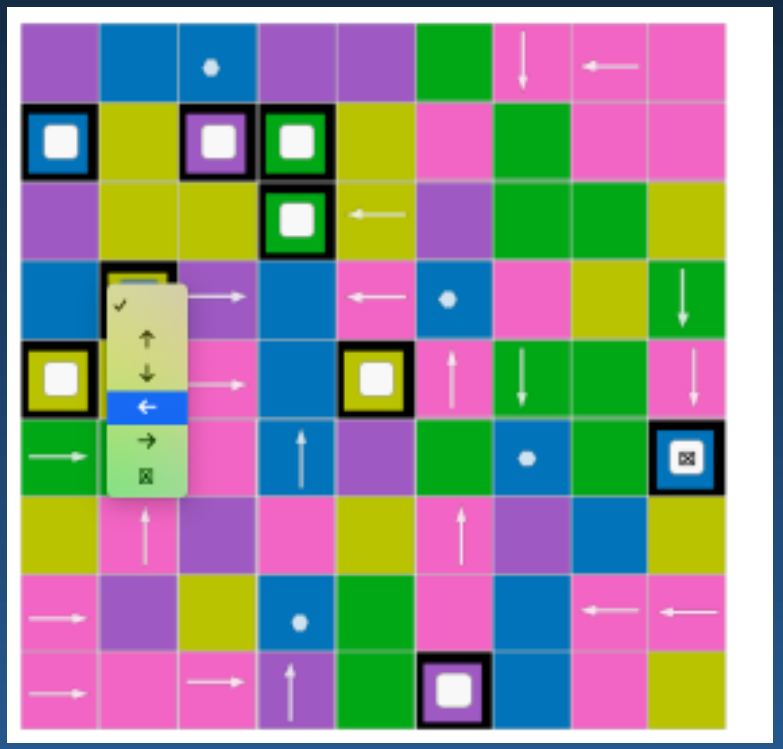
Unexpected Variation: Explaining Policies with Examples

Example: List some gridworld actions

Given:



Test: What happens here?

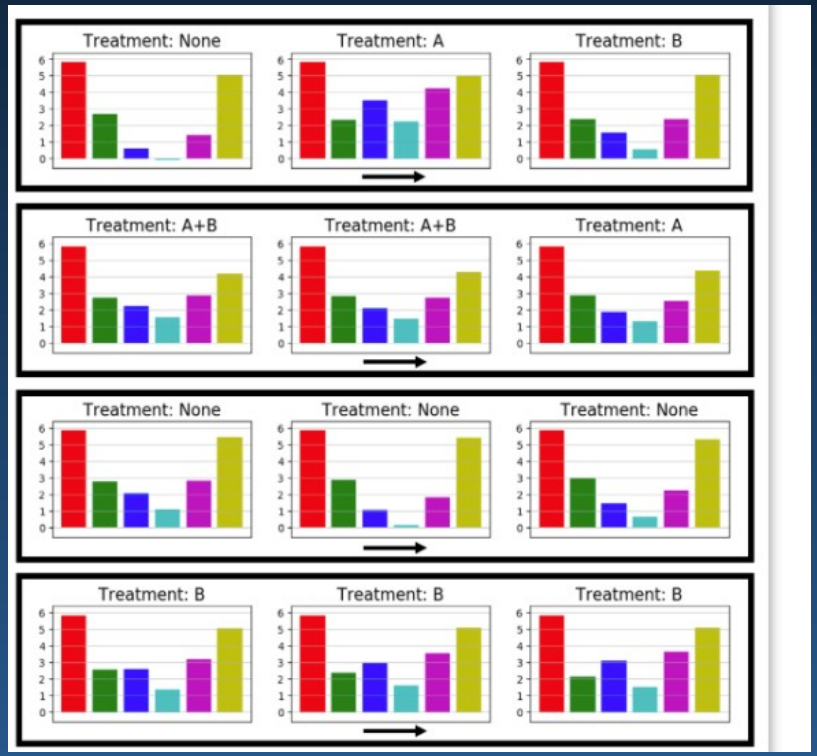


Lage et al, IJCAI 2019

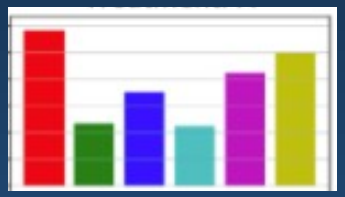
Unexpected Variation: Explaining Policies with Examples

Example: List some HIV actions

Given:



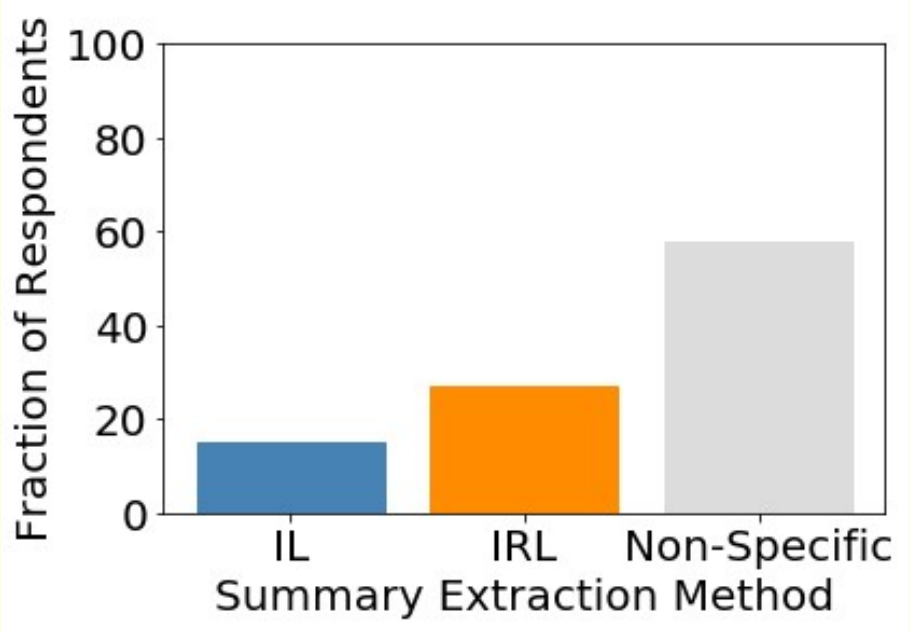
Test: What happens here?



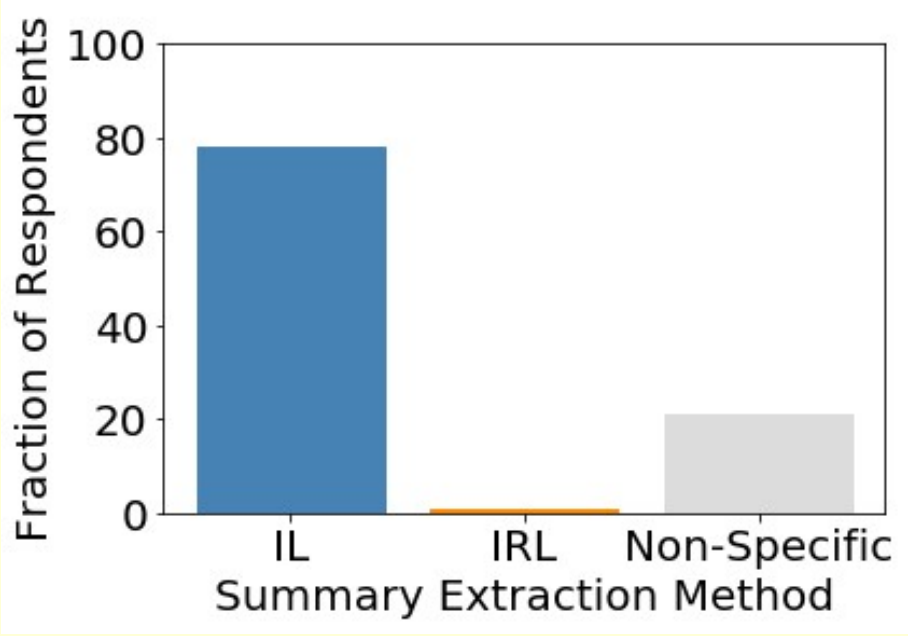
Lage et al, IJCAI 2019

Unexpected Variation: Explaining Policies with Examples

Gridworld



HIV



Lage et al, IJCAI 2019

Jacob et al. TransPsych 2021

A Study with 200+ Psychiatrists on Antidepressant Recommendations

Patient Details:
Patricia is a 31 year old woman who is married and works full time. She has a history of seizure disorder and lack of appetite, and presents with 11 months of depressed mood. Current medications include Omeprazole and Celecoxib. Prior treatment with Citalopram did not cause a reduction of depression symptoms.

Given a description of the patient, what should be the medication recommendation? We tried several different approaches.

Jacob et al. TransPsych 2021

A Study with 200+ Psychiatrists on Antidepressant Recommendations

Patient Details:
Jennifer is a 40 year old woman who is married and works from home. She is diabetic and has a history of hypertensive heart disease and arrhythmia. She presents with 10 months of depressed mood. Current medications include amoxicillin, and prior treatment with Paroxetine had no effect on depressed mood.

System.07 Recommendation: **DULOXETINE**

Top 5 therapies with highest probability for stability:

Therapy	Predicted Stability*	Predicted Dropout Risk**
Duloxetine	.79	.03
Fluoxetine	.65	.03
Citalopram	.60	.06
Escitalopram	.59	.07
Bupropion	.55	.12

**Stability: continued use of the same medication for at least 3 months*
***Dropout: early treatment discontinuation following prescription*

Jacob et al. TransPsych 2021

A Study with 200+ Psychiatrists on Antidepressant Recommendations

Patient Details:
Jennifer is a 40 year old woman who is married and works from home. She is diabetic and has a history of hypertensiv
presents with 10 months of depressed m
amoxicillin, and prior treatment with Paro
mood.

System.07 Recommendation: **DULOXETINE**

Top 5 therapies with highest probability for stability:

Therapy	Predicted Stability*
Duloxetine	
Fluoxetine	
Citalopram	
Escitalopram	
Bupropion	

**Stability: continued use of the same medication for at least 3 m*
***Dropout: early treatment discontinuation following prescriptio*

Patient Details:
David is a 43 year old man who is widowed and works full time. He has a history of diabetes, arrhythmia and hypertensive heart disease. He presents with 14 months of depressed mood. Current medications include amoxicillin, and prior treatment with Paroxetine was ineffective.

System.10 Recommendation: **DULOXETINE**

Top 5 therapies with highest probability for stability:

Therapy	Predicted Stability*	Predicted Dropout Risk**
Duloxetine	.78	.06
Fluoxetine	.65	.08
Citalopram	.54	.10
Escitalopram	.52	.14
Bupropion	.52	.14

**Stability: continued use of the same medication for at least 3 months*
***Dropout: early treatment discontinuation following prescription*

Why are these therapies being recommended?

System.10's predictions are **based on the patient's ICD-9 codes.**

Jacob et al. TransPsych 2021

A Study with 200+ Psychiatrists on Antidepressant Recommendations

Patient Details:
Susan is a 31 year old woman who is single and works part time. She has a history of diabetes, arrhythmia and hypertensive heart disease. She presents with 14 months of depressed mood. Current medications include amoxicillin, and prior treatment with Paroxetine was ineffective.

System.13 Recommendation: **DULOXETINE**

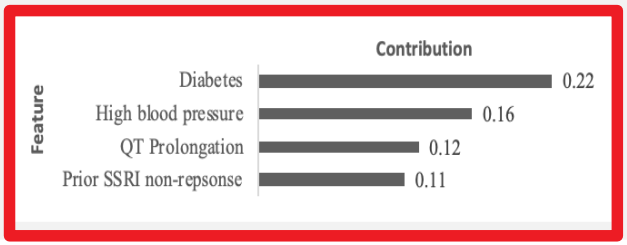
Top 5 therapies with highest probability for stability:

Therapy	Predicted Stability*	Predicted Dropout Risk**
Duloxetine	.80	.06
Fluoxetine	.68	.12
Citalopram	.67	.13
Escitalopram	.59	.16
Bupropion	.57	.20

**Stability: continued use of the same medication for at least 3 months*
***Dropout: early treatment discontinuation following prescription*

Why are these therapies being recommended?

The following **patient features** had the highest contributions to system.13's predictions:



Patient Details:
Thomas is a 38 year old man who is single and works full time. He has a history of diabetes, hypertensive heart disease, and arrhythmia. He presents with 10 months of depressed mood. Current medications include amoxicillin, and prior treatment with Paroxetine was ineffective.

System.16 Recommendation: **DULOXETINE**

Top 5 therapies with highest probability for stability:

Therapy	Predicted Stability*	Predicted Dropout Risk**
Duloxetine	.77	.04
Fluoxetine	.69	.07
Citalopram	.65	.07
Escitalopram	.65	.07
Bupropion	.51	.18

**Stability: continued use of the same medication for at least 3 months*
***Dropout: early treatment discontinuation following prescription*

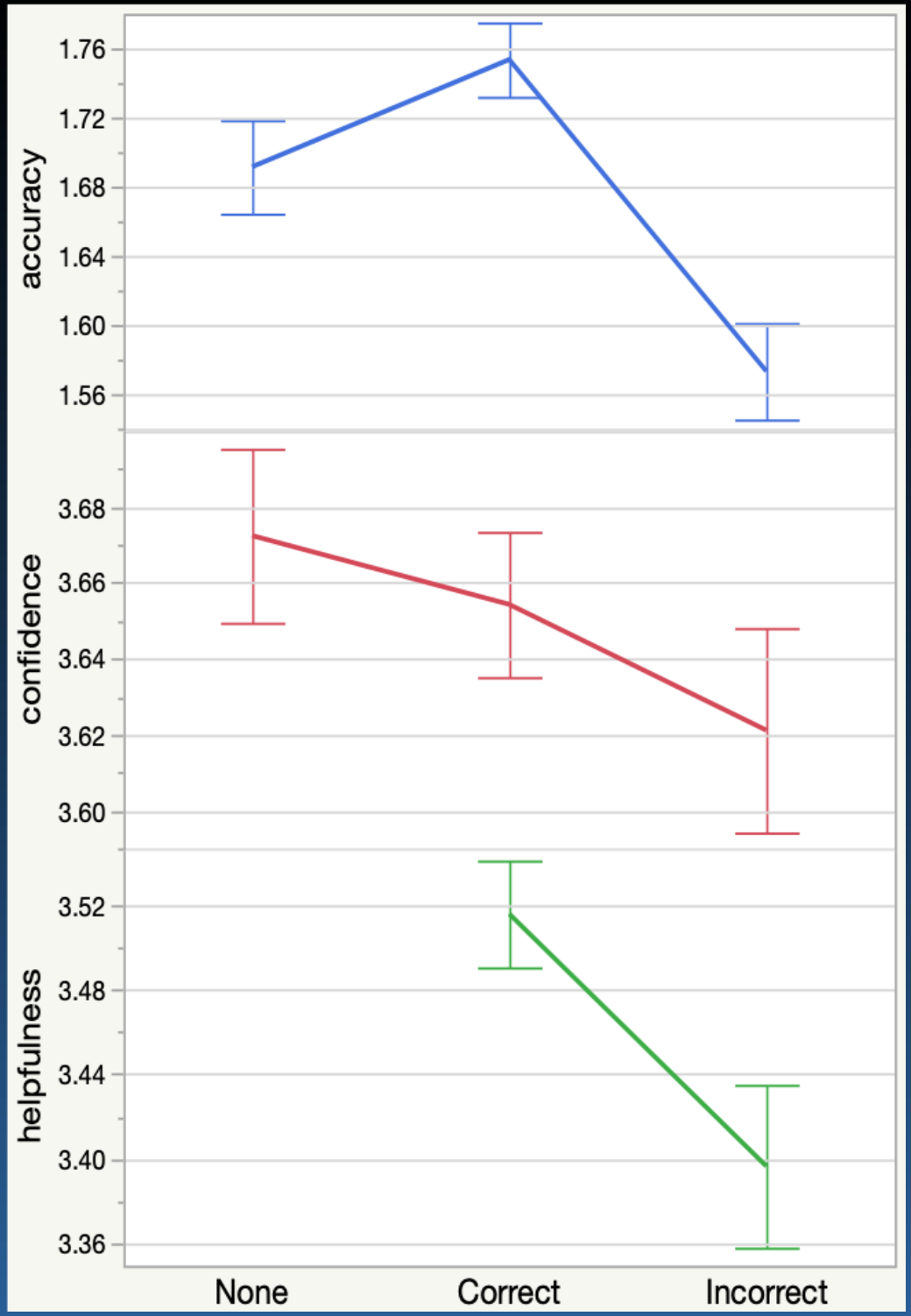
Why are these therapies being recommended?

The following **rules** had the highest contributions to system.16's predictions:

1. *If concern for QT prolongation, favor Sertraline, avoid Citalopram*
2. *If avoiding weight gain, favor weight loss, favor Bupropion, avoid Mirtazapine*
3. *If concern for increased blood pressure, avoid SNRI's*
4. *If lack of response to Paroxetine, avoid SSRI's*

Results

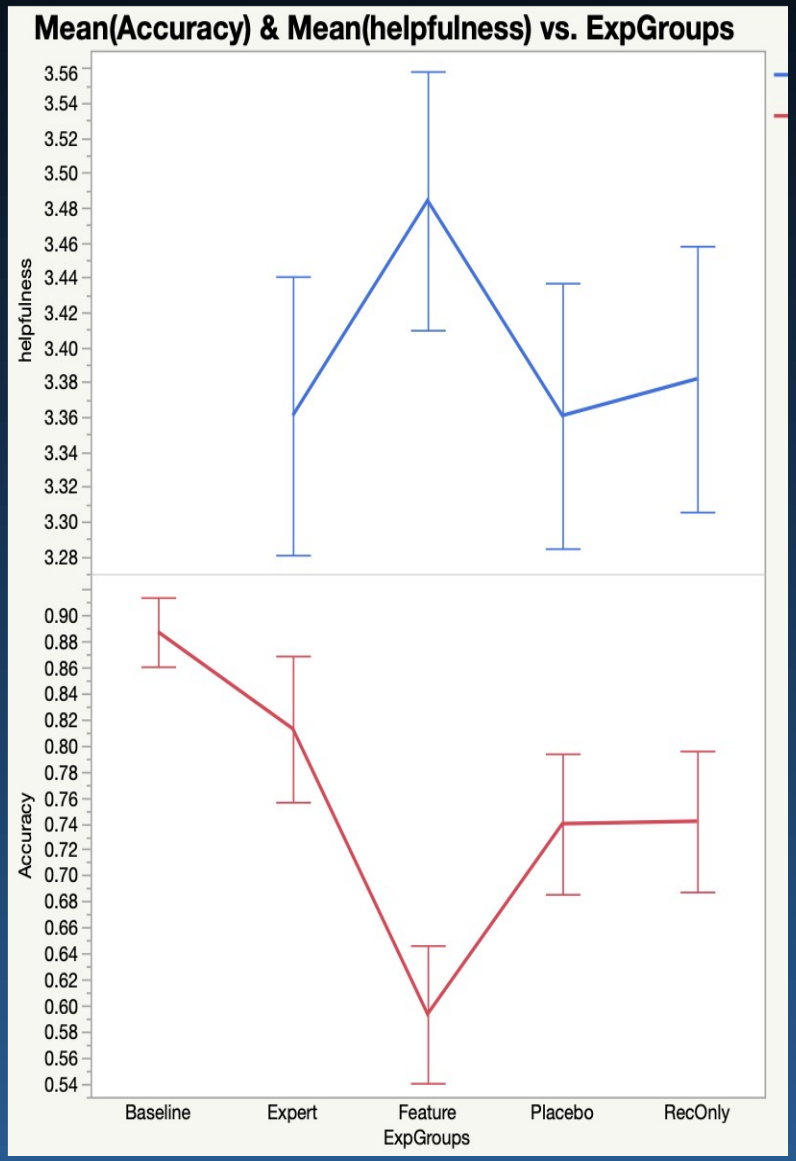
- Accuracy increases if the recommendation is correct, decreases if not correct.
- Some awareness of helpfulness w.r.t. the correctness of the explanation.



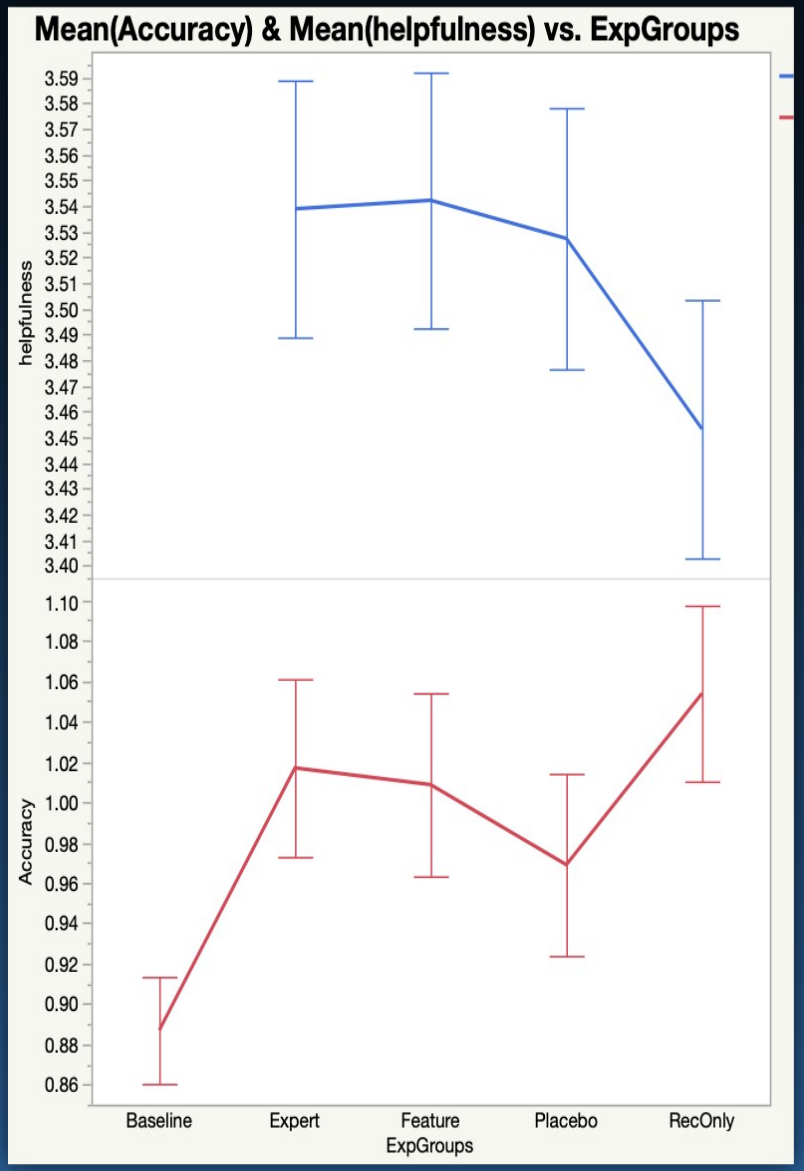
Jacob et al. TransPsych 2021

Perceived Helpful ≠ Accurate

Rec incorrect



Rec correct



Plan for these modules:

Module 1: Techniques

Module 2: Computational Evaluations

Module 3: Human Factors

- Ways in which people's behavior may be unexpected
- How to improve human+AI performance

Being Smarter about Seeking Engagement

Information about the alien

Instructions

The alien's treatment plan:

(**brain fog**) and (**slurred speech** or **confusion** or **jaundice** or **blisters**) → **fatigued**
(**brain fog**) and (**chills** or **blurry vision**) and (**migraine**) → **low blood pressure**
(**blisters** or **confusion** or **rash** or **puffy eyes**) → **muscle weakness**
(**jaundice** or **puffy eyes** or **fatigued**) and (**bloating** or **chills**) → **stimulants**
(**jaundice** or **muscle weakness** or **rash** or **low blood pressure**) → **laxatives**
(**slurred speech** or **fatigued** or **chills** or **sleepy**) → **vitamins**
(**jaundice** or **sleepy** or **puffy eyes**) and (**slurred speech**) → **antibiotics**
(**blisters** or **rash** or **sleepy** or **puffy eyes**) → **tranquilizers**

Observed symptoms: **brain fog, blurry vision, shortness of breath, migraine, bloating**

AI input
The AI recommends prescribing **laxatives**,
because the alien has the symptom(s): **low blood pressure**.

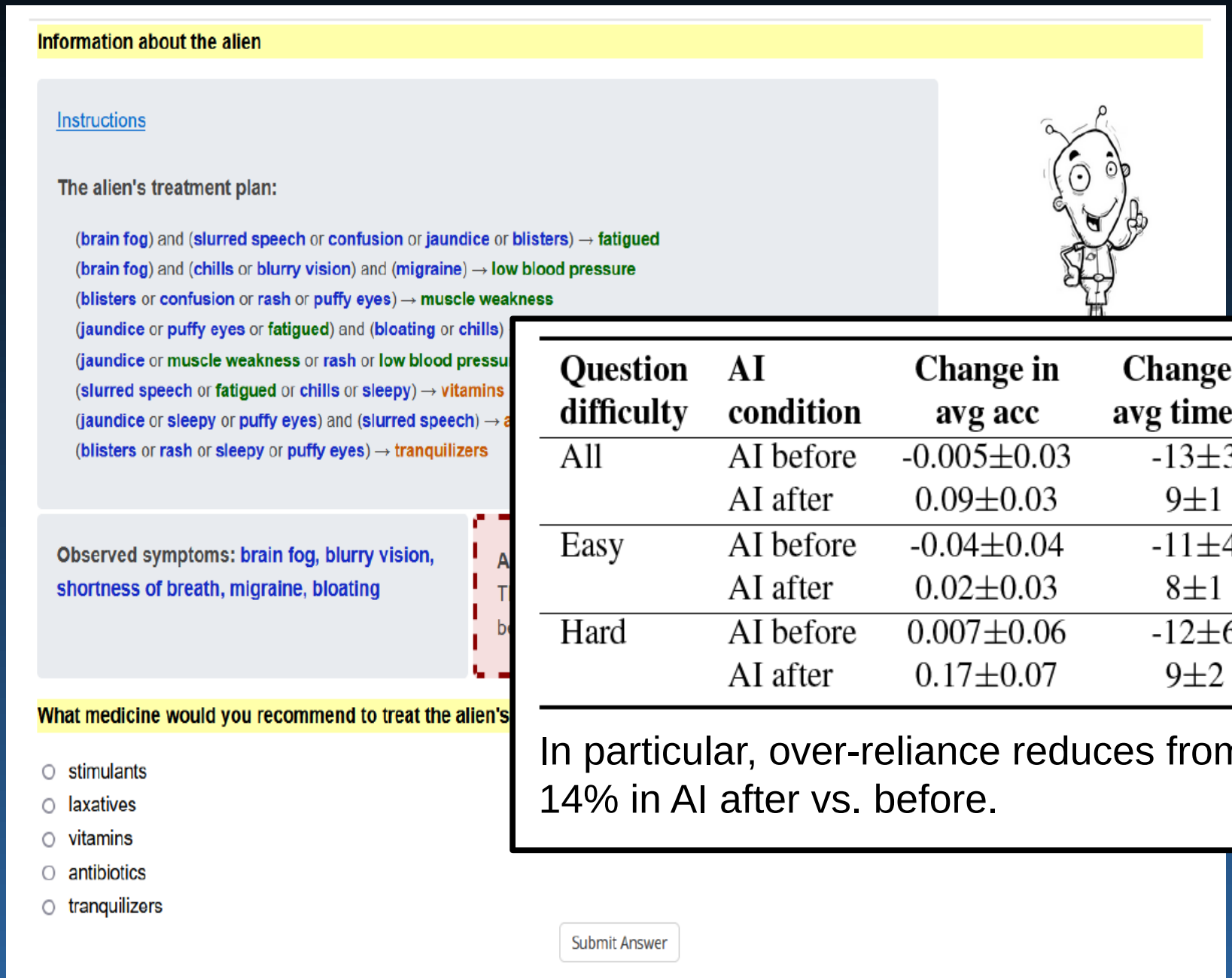
What medicine would you recommend to treat the alien's observed symptoms?

- ○ stimulants
- ○ laxatives
- ○ vitamins
- ○ antibiotics
- ○ tranquilizers

Submit Answer

Swaroop et al., ICML WS 2023

Being Smarter about Seeking Engagement



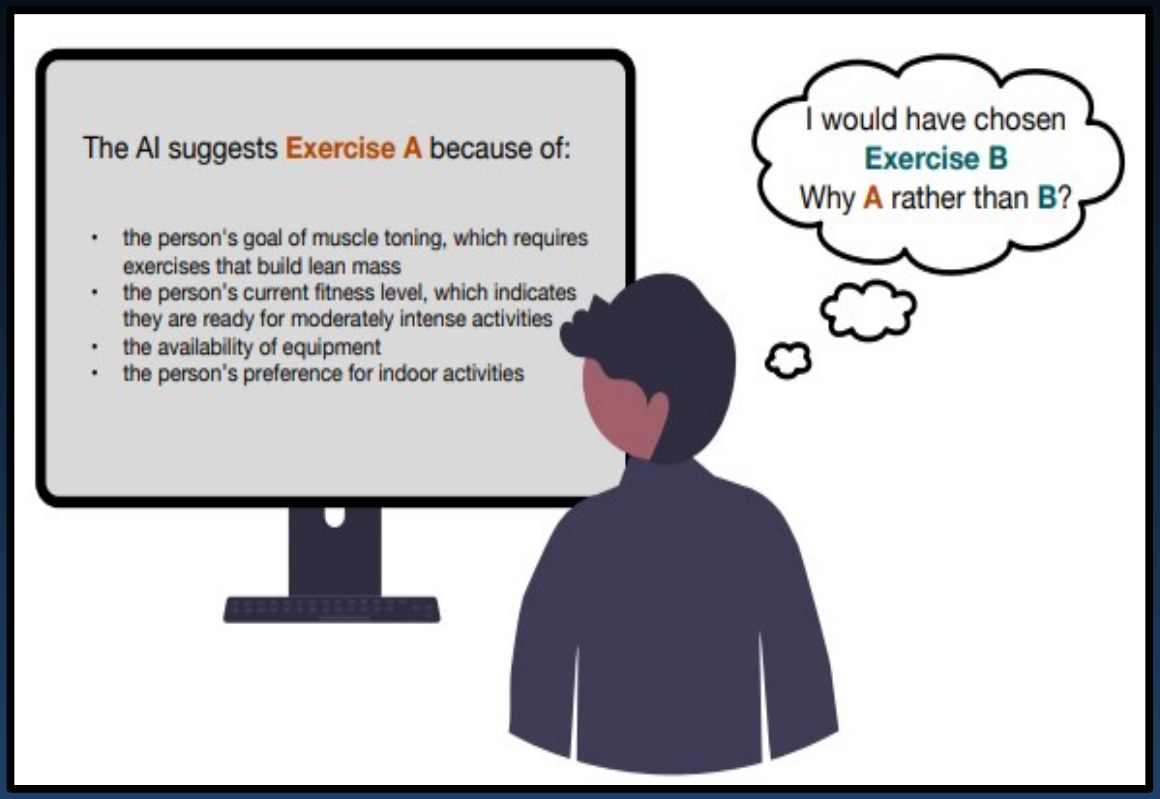
Question difficulty	AI condition	Change in avg acc	Change in avg time (s)
All	AI before	-0.005±0.03	-13±3
	AI after	0.09±0.03	9±1
Easy	AI before	-0.04±0.04	-11±4
	AI after	0.02±0.03	8±1
Hard	AI before	0.007±0.06	-12±6
	AI after	0.17±0.07	9±2

In particular, over-reliance reduces from 49% to 14% in AI after vs. before.

The timing of the AI advice is just one interaction choice; there are many options to explore!

Bucina et al., CHI 2025

Providing contrasts to *anticipated* errors



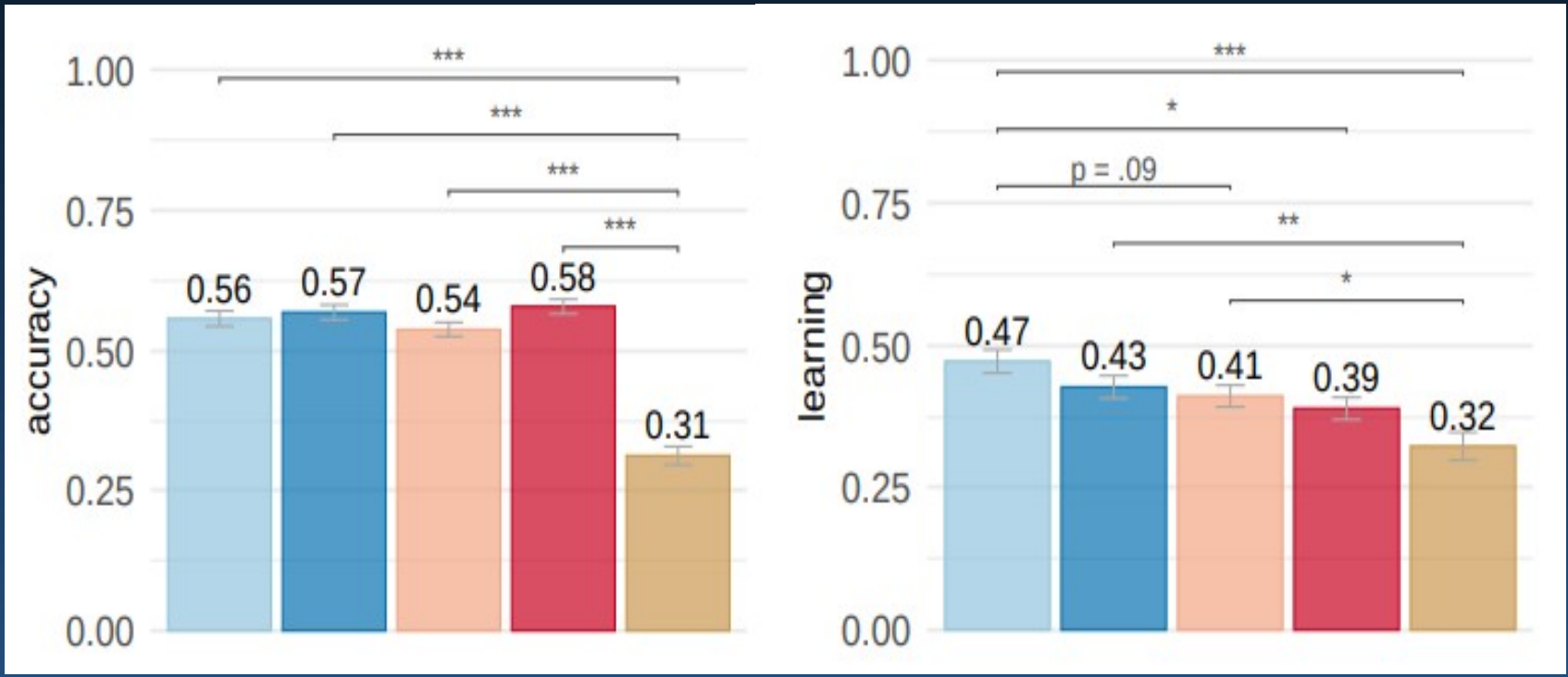
Bucina et al., CHI 2025

Providing contrasts to *anticipated* errors

Bucina et al., CHI 2025

Providing contrasts to *anticipated* errors

Approach increased learning, had strong performance, and avoided the wait-then-suggest friction.



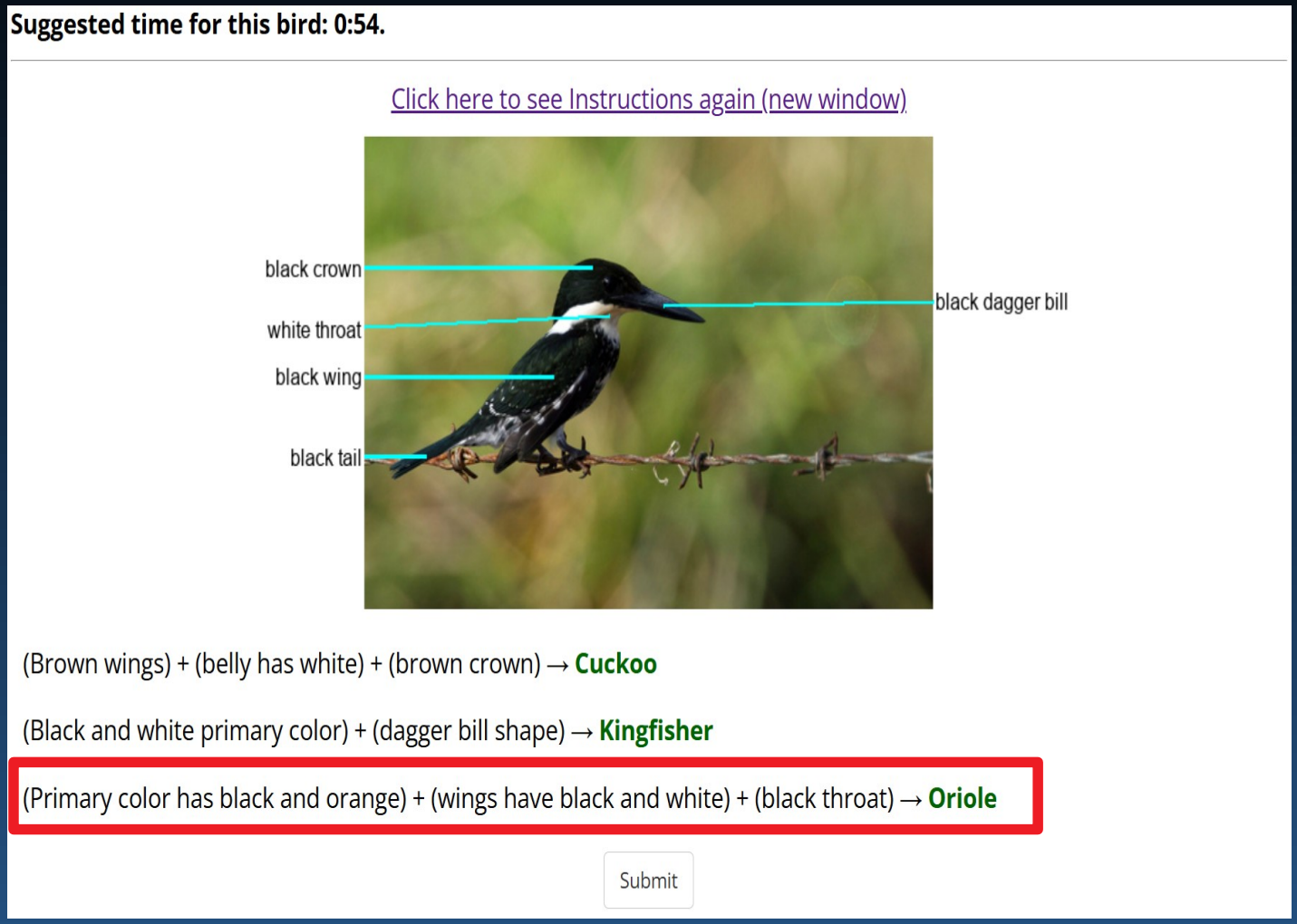
Swaroop et al; So et al. ongoing

Different people make different errors

Novices tend to collect facts but not synthesize them.

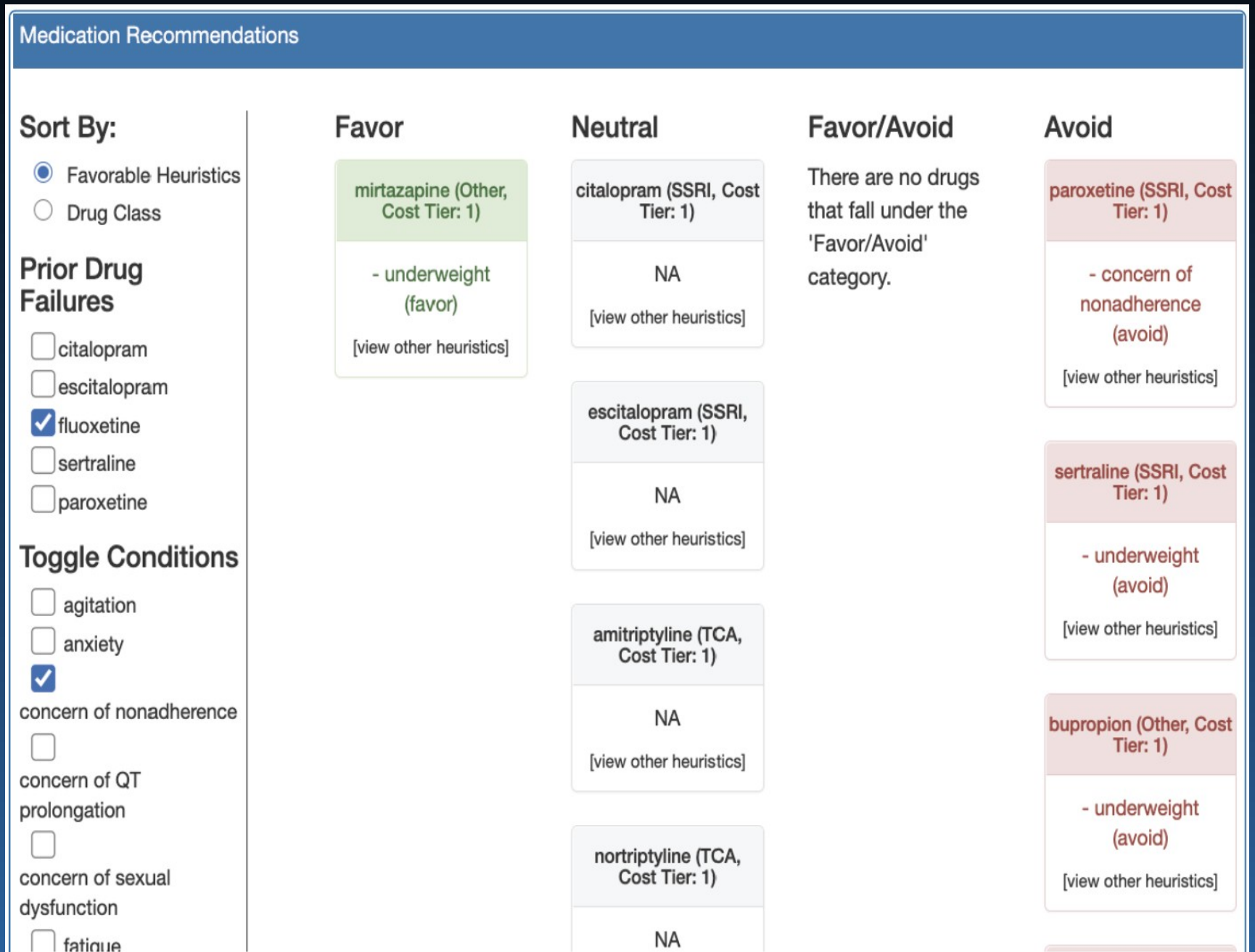
Experts tend to work backwards from hypotheses but miss key facts.

People vary in their visual and logical reasoning skills.



Jacob et al, CHI 2021

Interactivity also helps with engagement



Batool et al, ongoing

Sometimes the AI needs to support process

Some problems have a right answer – e.g. whether a tumor is cancerous – and we need the AI to get to the correct outcome.

Other problems involve values – e.g. whether one should opt for surgery or radiation – and we need the AI to make sure we considered all key factors, that is, to support correct process.

Batool et al, ongoing

Sometimes the AI needs to support process
Example: Negligence Determinations

Penny was using the pallet stacker to move a heavy load when a light drizzle turned to downpour. The stacker jerked wildly. Penny fell and suffered a serious back injury.

Penny: Penny claims that the incident was caused by the Metro's failure to maintain the stacker. She states she reported minor issues with the stacker's stability a month before the incident, and no action was taken. Historical records report slippage-related but not stability-related issues with stackers in poor weather. A
prevent her from supportin

Metro: Metro points to thei
one of the best places to w
Penny, they claim, has alw
color of company hats. The
ago. The machine is only 1
completed at purchase and
received Penny's complain
unexpected weather condi

SUMMARY: Metro had a duty of care towards Penny.

Did Metro fail their duty of care?

- Thorough inspection of stacker at purchase passed all checks.
- Stability issue reported one month prior.
- Internal email claims issue was checked and stacker functional.
- Routine checks passed two weeks prior.

Did that alleged failure cause the harm?

- A stability issue caused Penny to fall.
- The fall resulted in the injuries.
- It was unexpected rainy.
- Rain has not caused stability issues in the past.

With all these options, one thing we can do is personalize to the user and/or context.

Being Smarter about Seeking Engagement

Information about the alien

Instructions

The alien's treatment plan:

(**brain fog**) and (**slurred speech** or **confusion** or **jaundice** or **blisters**) → **fatigued**
(**brain fog**) and (**chills** or **blurry vision**) and (**migraine**) → **low blood pressure**
(**blisters** or **confusion** or **rash** or **puffy eyes**) → **muscle weakness**
(**jaundice** or **puffy eyes** or **fatigued**) and (**bloating** or **chills**) → **stimulants**
(**jaundice** or **muscle weakness** or **rash** or **low blood pressure**) → **laxatives**
(**slurred speech** or **fatigued** or **chills** or **sleepy**) → **vitamins**
(**jaundice** or **sleepy** or **puffy eyes**) and (**slurred speech**) → **antibiotics**
(**blisters** or **rash** or **sleepy** or **puffy eyes**) → **tranquilizers**

Observed symptoms: **brain fog, blurry vision, shortness of breath, migraine, bloating**

AI input
The AI recommends prescribing **laxatives**,
because the alien has the symptom(s): **low blood pressure**.

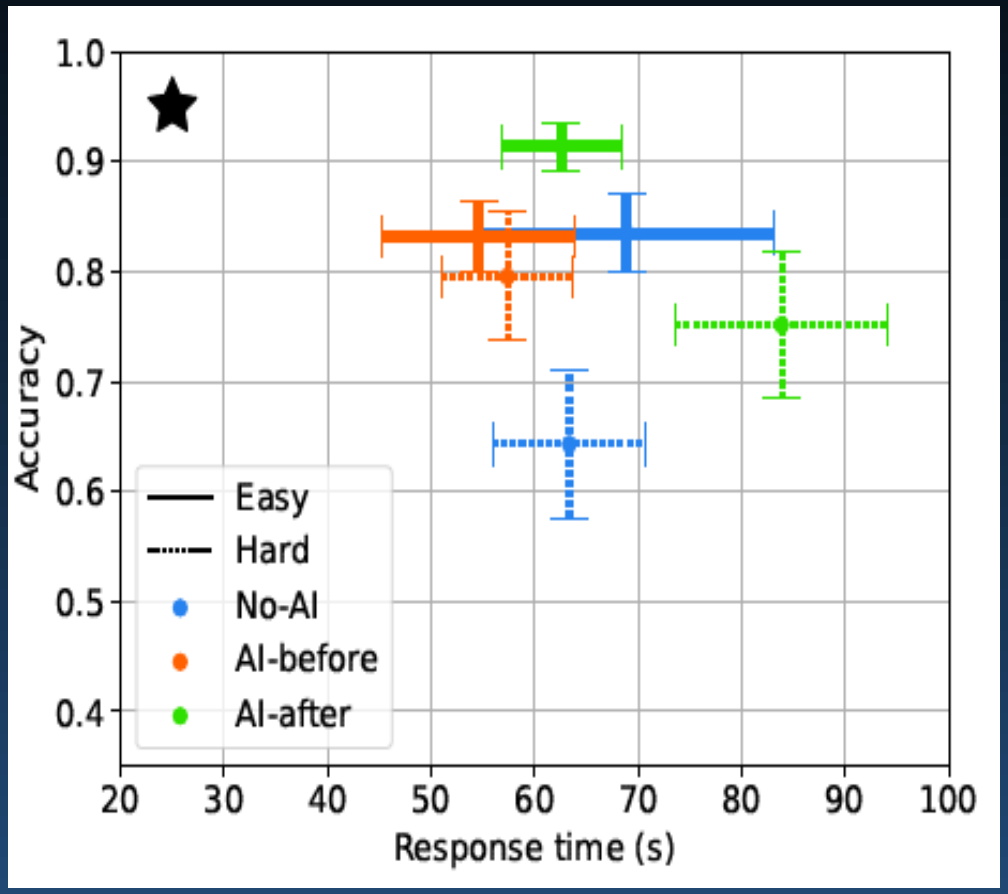
What medicine would you recommend to treat the alien's observed symptoms?

- ○ stimulants
- ○ laxatives
- ○ vitamins
- ○ antibiotics
- ○ tranquilizers

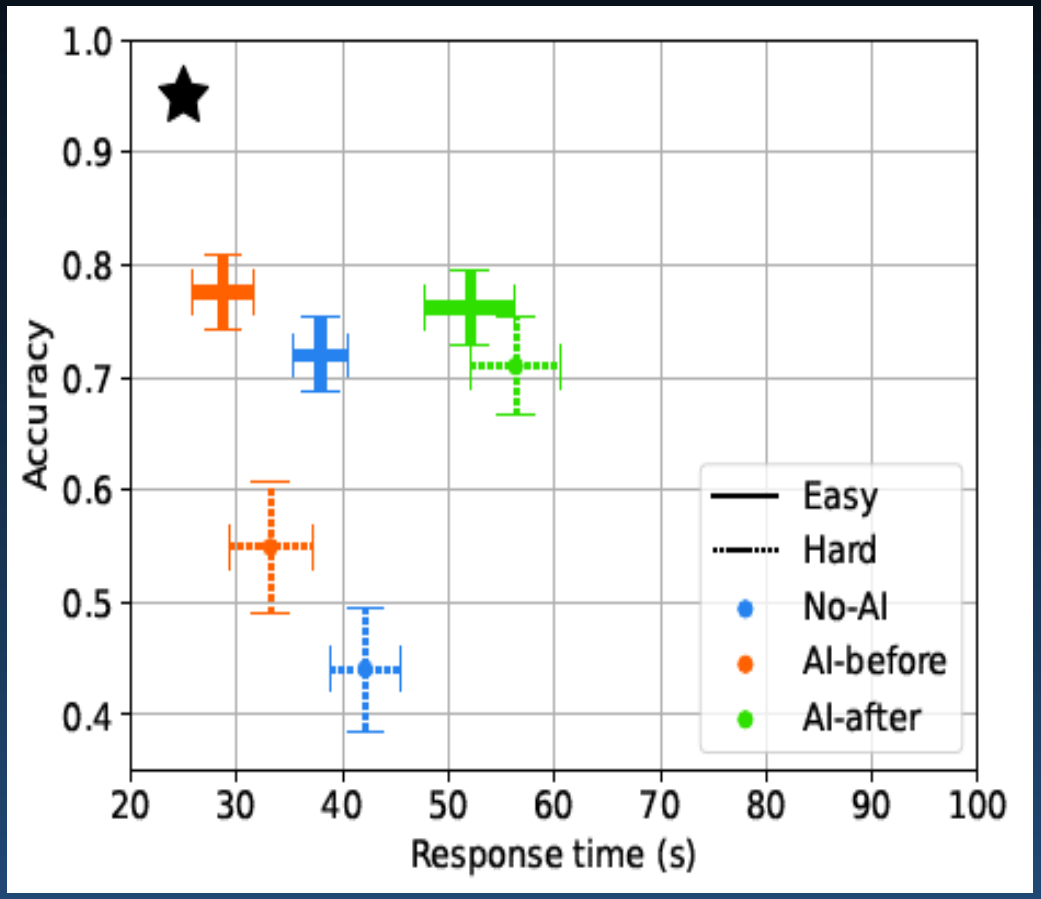
Submit Answer

Swaroop et al., IUI 2024

Not all users, questions are the same



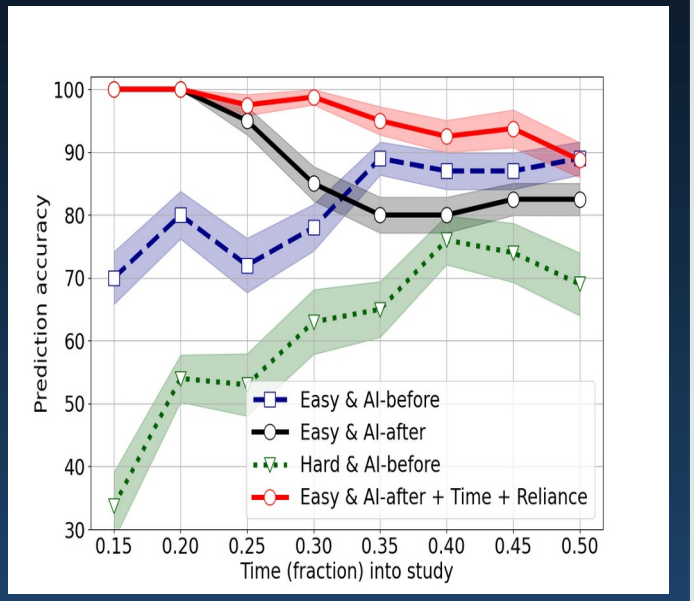
Not Overreliers



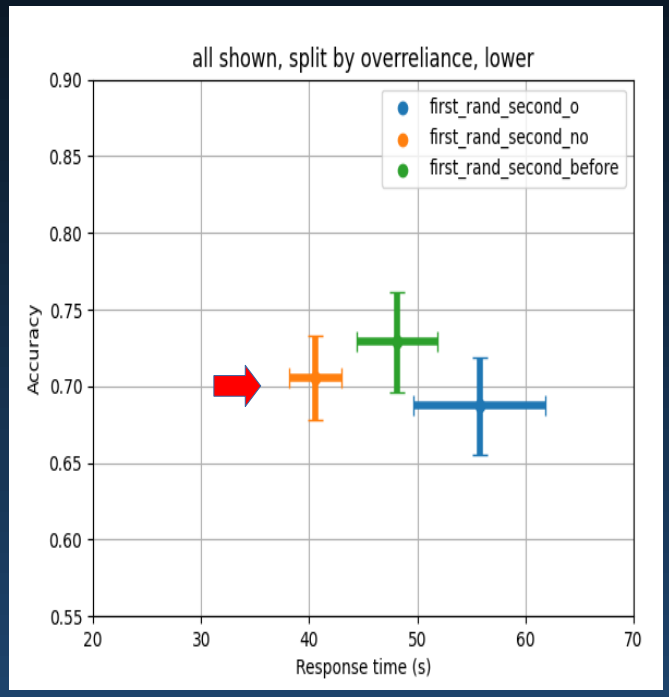
Overreliers

Swaroop et al., IUI 2025

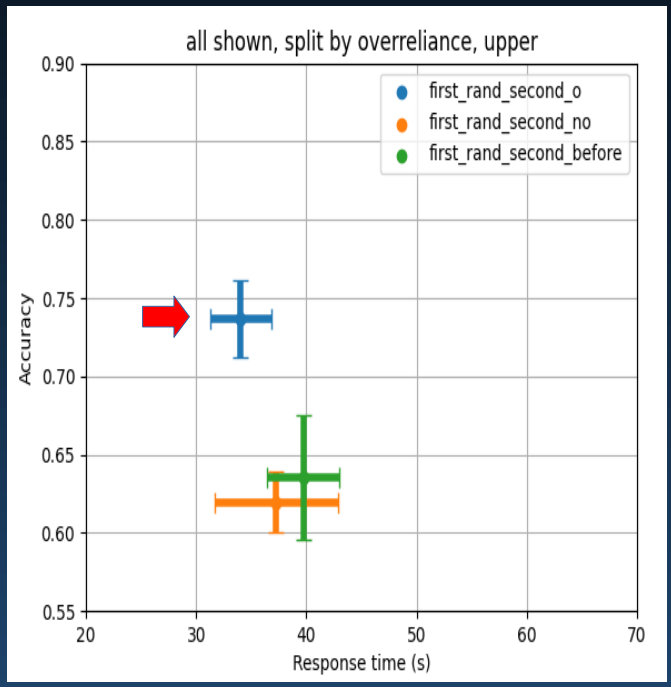
Which is a chance to personalize!



Identifying overreliers
from their interactions



Not Overreliers



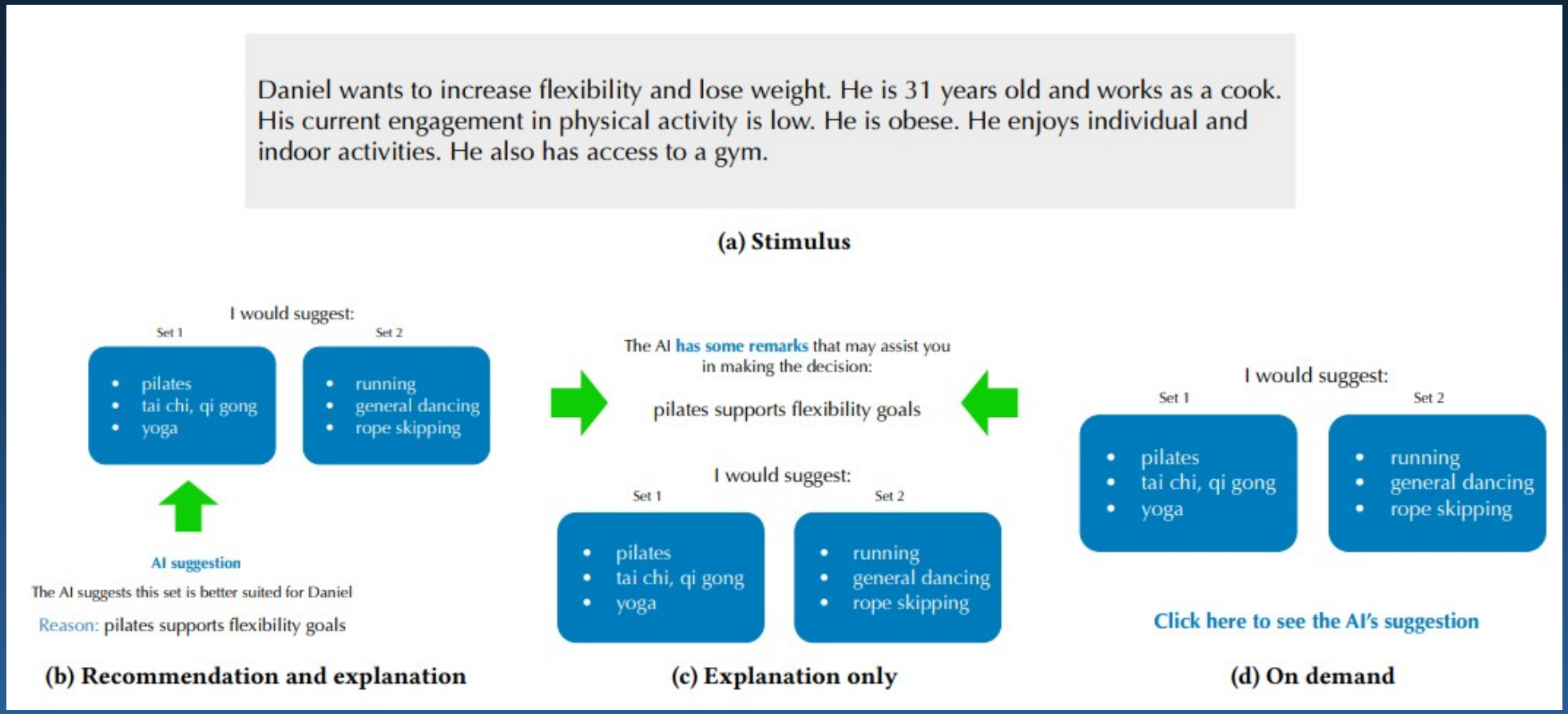
Overreliers

State defined by need for cognition (traits), task, person's expertise

Bucina et al., CHI 2024

Another example with multiple objectives

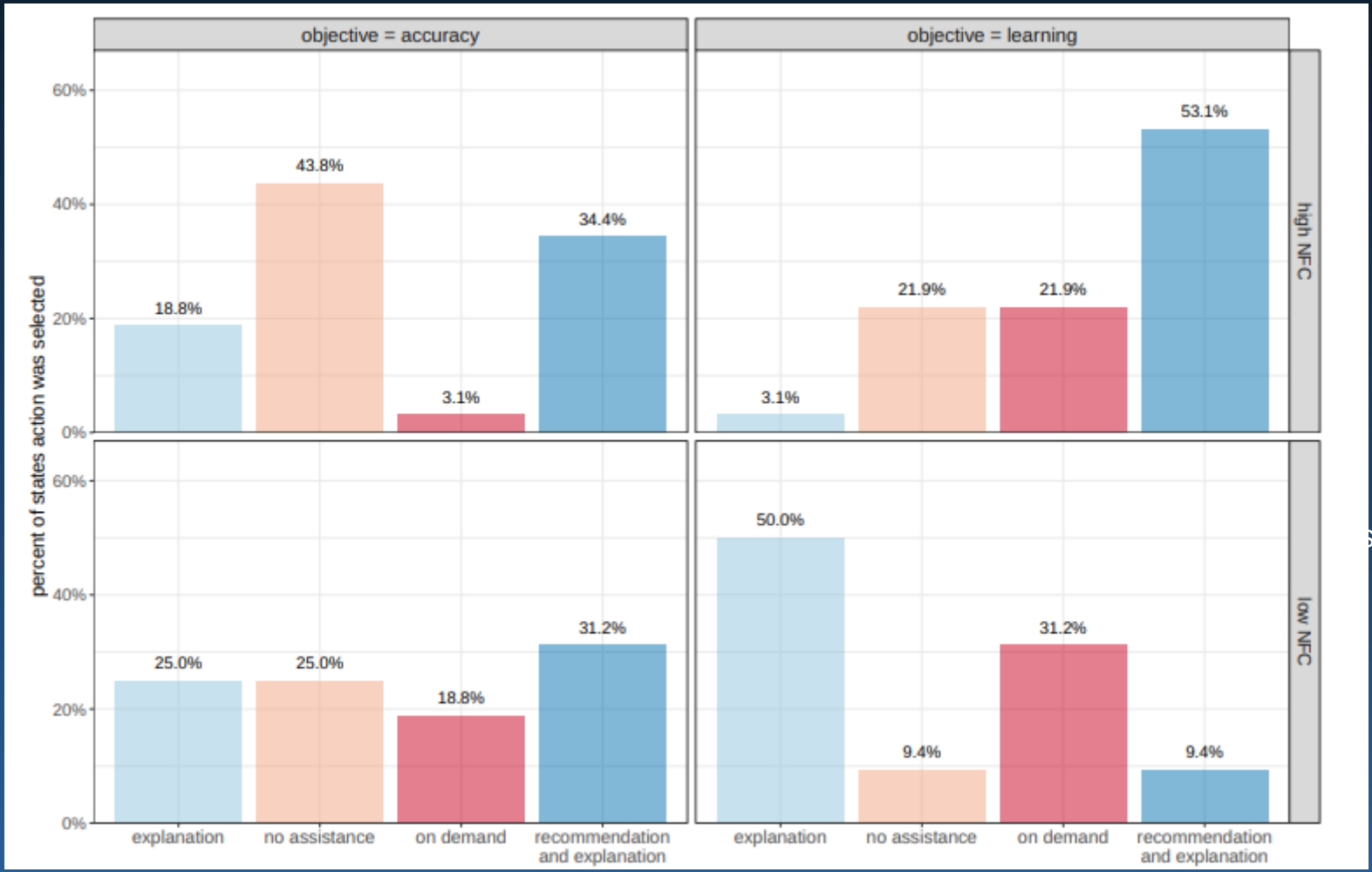
Here the state space was larger: 64 conditions based on the user's traits (need for cognition), needs of the task, and the person's expertise (from initial questions). The action space had four different support options.



Bucina et al., CHI 2024

Another example with multiple objectives

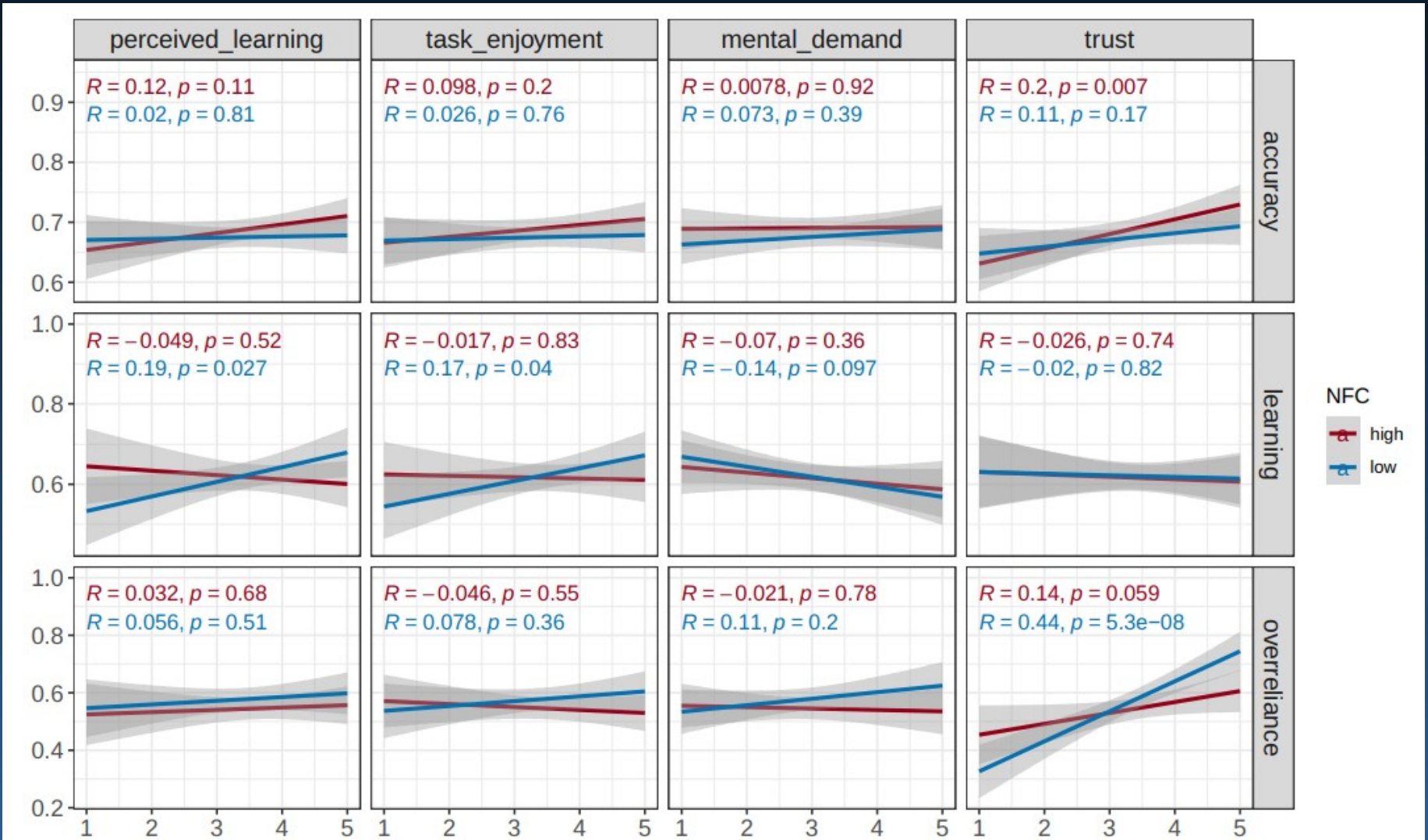
Authors found they could personalize for accuracy and learning-based objectives from offline data (more OPE next week!)



Bucina et al., CHI 2024

Another example with multiple objectives

Authors found they could personalize for accuracy and learning-based objectives from offline data (more OPE next week!)



Summary

Interpretability (and explainability) is important when both human and model are needed to do a task. The right form of interpretability will depend on the task.

(1) With clever design and optimization, we can often make models inherently interpretable.
(2) There are many metrics for evaluating partial view and inherently interpretable explanations – and the right one will depend on the ultimate task.
(3) Successful interpretable ML requires also engaging with the human factors of how people use explanations.

Simulation/game this afternoon!!

Different Explanations have Different Properties

A gradient-based measure of importance gives different answers for the two functions below, while one based on larger perturbations will not.