Microsoft Research

KDD 2022 Tutorial

Have Your Cake And Eat It Too:

Glass-Box ML, Missing Values, Differential Privacy, and Editability with Boosted Trees & Neural Nets

Rich Caruana & Harsha Nori

Microsoft

Microsoft Research

KDD 2022 Tutorial

Have Your Cake And Eat It Too:

Glass-Box ML, Missing Values, Differential Privacy, and Editability with Boosted Trees & Neural Nets

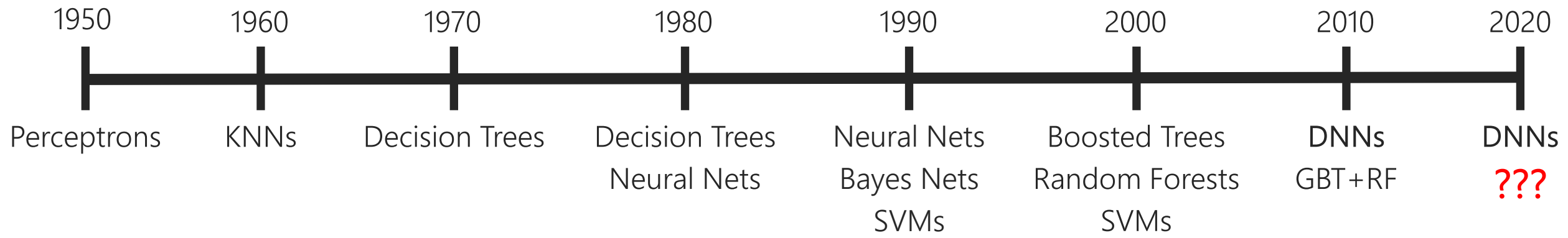
Rich Caruana & Harsha Nori
Yin Lou, Sarah Tan, Xuezhou Zhang, Ben Lengerich, Kingsley Chang, Jay Wang, Zhi Chen
Paul Koch, Harsha Nori, Sam Jenkins, Jessica Wolk, Luis França, Levi Melnick, Urszula Chajewska
Greg Cooper MD/PhD, Mike Fine MD, Vivienne Souter MD, Yin Aphinyanaphongs MD/PhD,
Giles Hooker, Johannes Gehrke, Tom Mitchell, Marc Sturm, Niloo Steele, Noemie Elhadad, Jacob Bien,
Noah Snavely, Eric Horvitz MD/PhD, Nick Craswell, Jenn Wortmann Vaughan, Mihaela Vorvoreanu

Microsoft

Tutorial Outline:

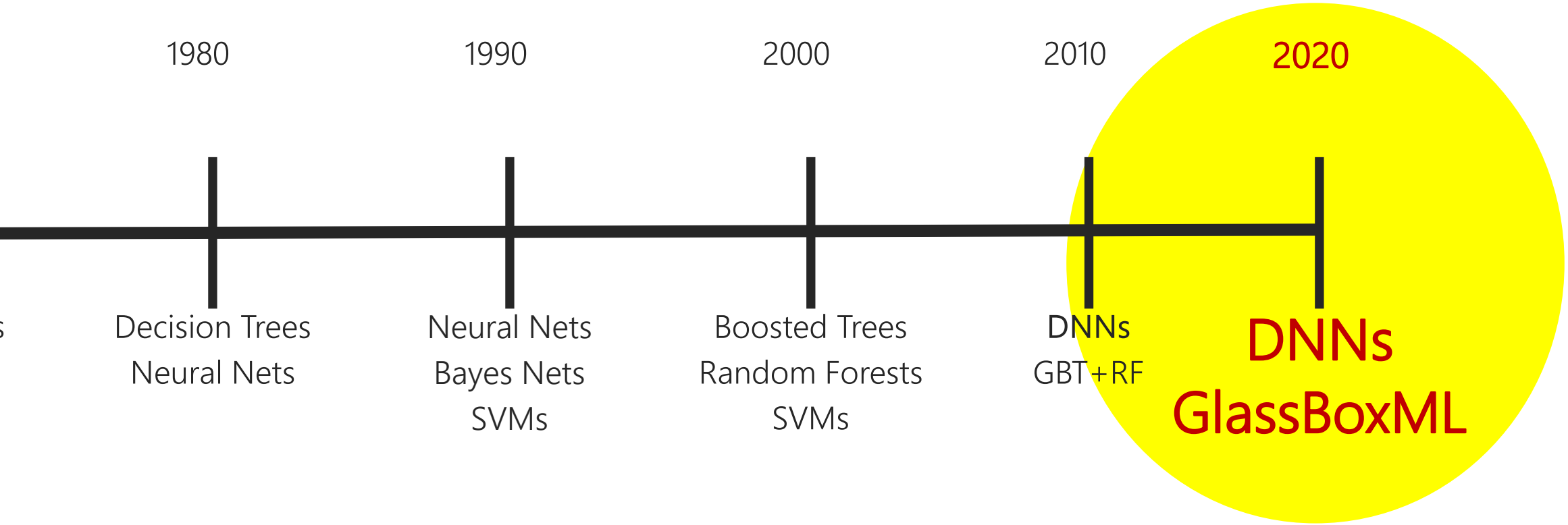
- Glass-box machine learning with EBMs: Explainable Boosting Machines
  - Glass-box Case Studies
  - How do we train EBMs?
- Glass-box learning with NAMs: Neural Additive Models
  - How do we train NAMs?
- Glass-box models vs. Missing Values
- Hands-on demos:
  - InterpretML and EBMs
  - DP-EBMs: Differentially Private EBMs
  - Editing glass-box models with GAM-Changer

A Brief History of Machine Learning

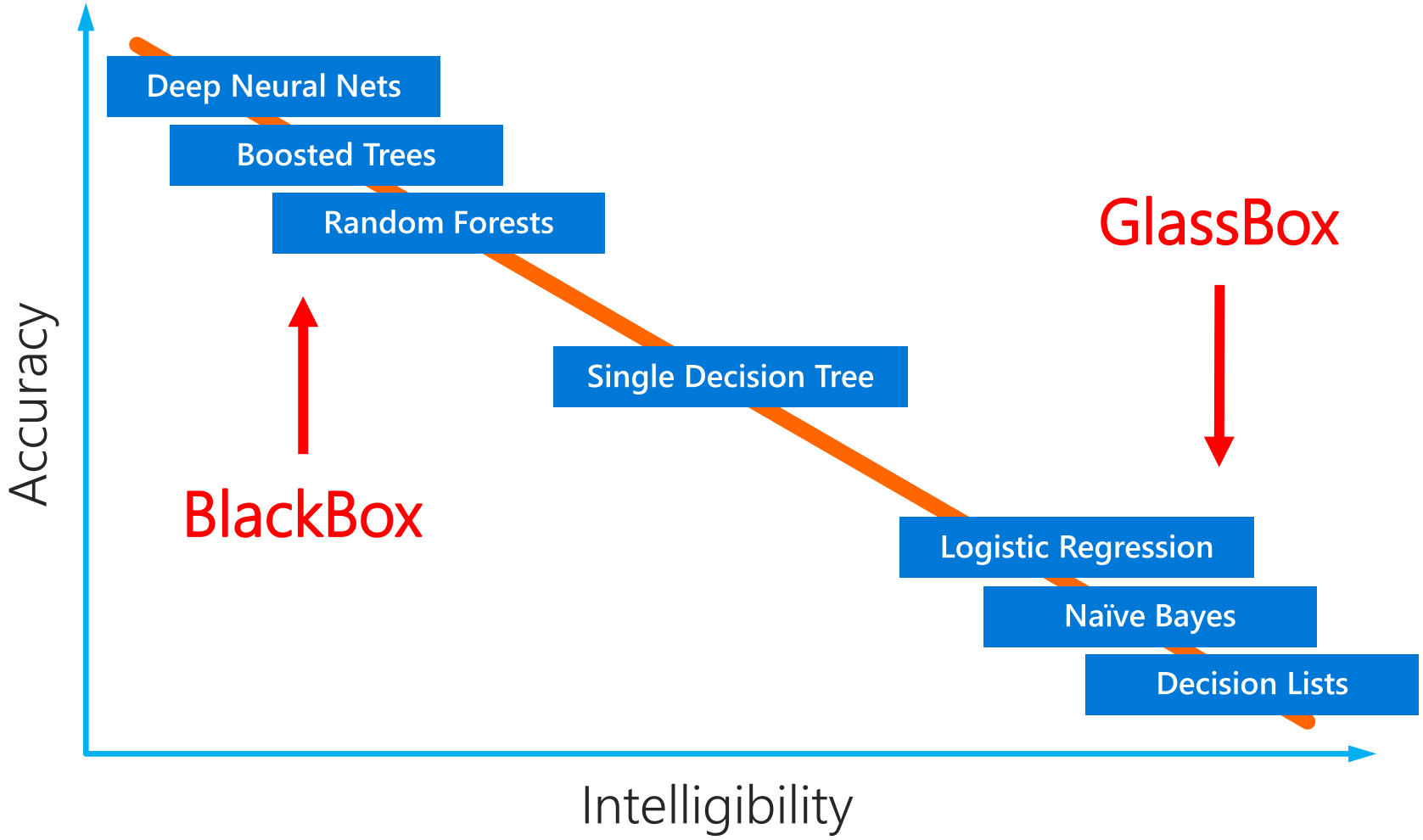


The Future of Machine Learning

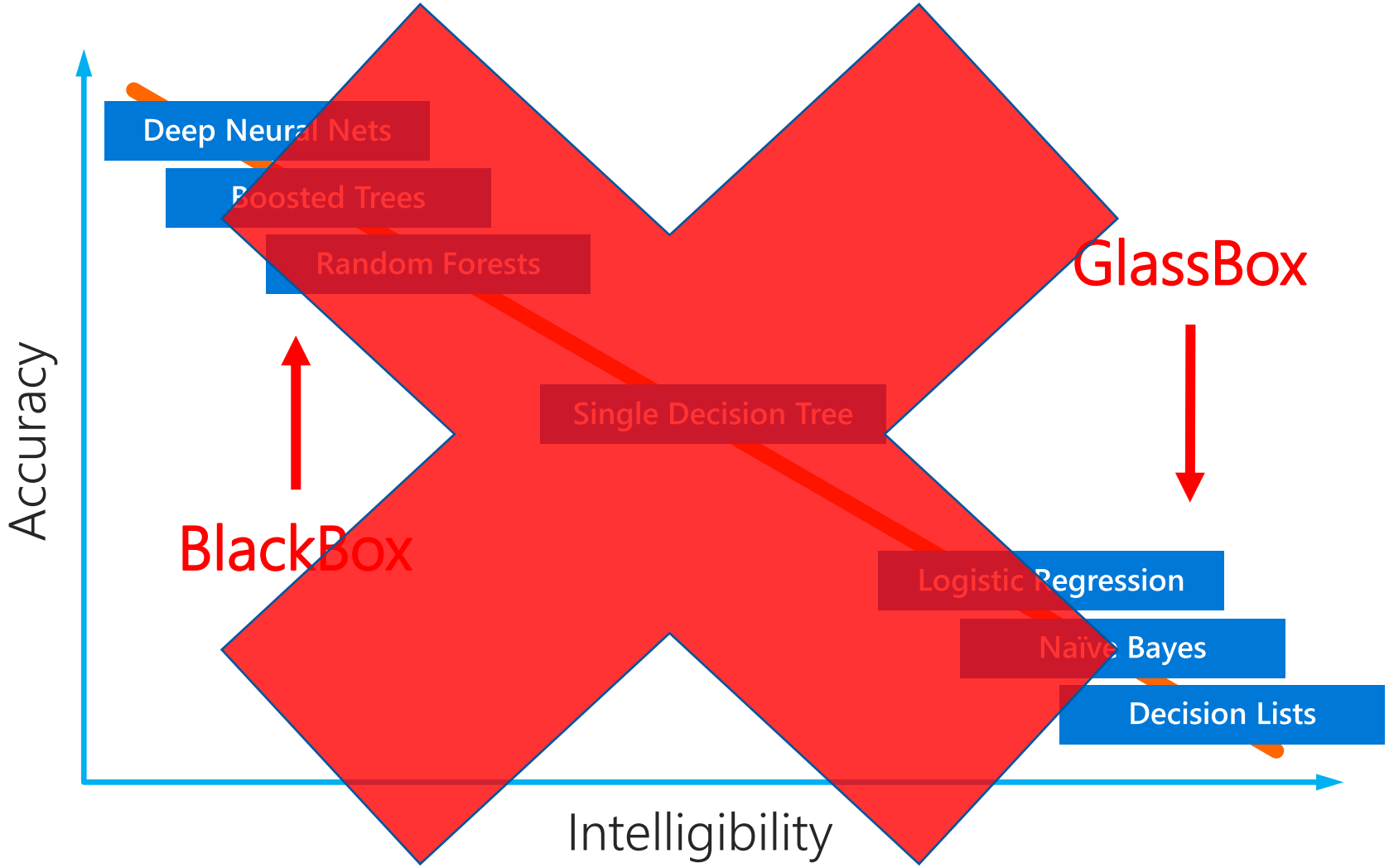
Data scientists now want/need to use ML methods that are interpretable



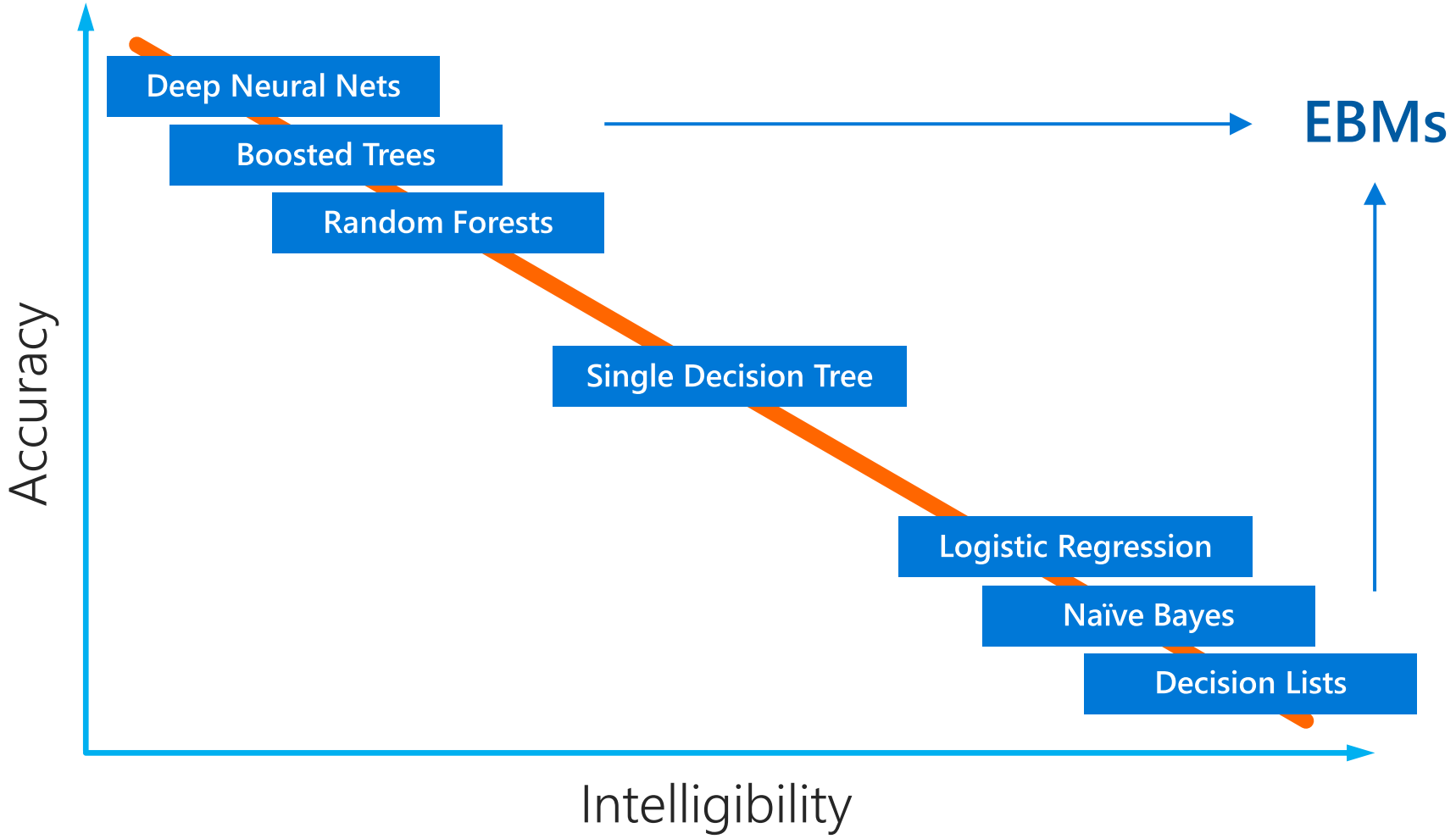
Accuracy vs. Intelligibility Tradeoff ???



Accuracy vs. Intelligibility Tradeoff – No Longer True for Tabular Data



Accuracy vs. Intelligibility Tradeoff – No Longer True for Tabular Data



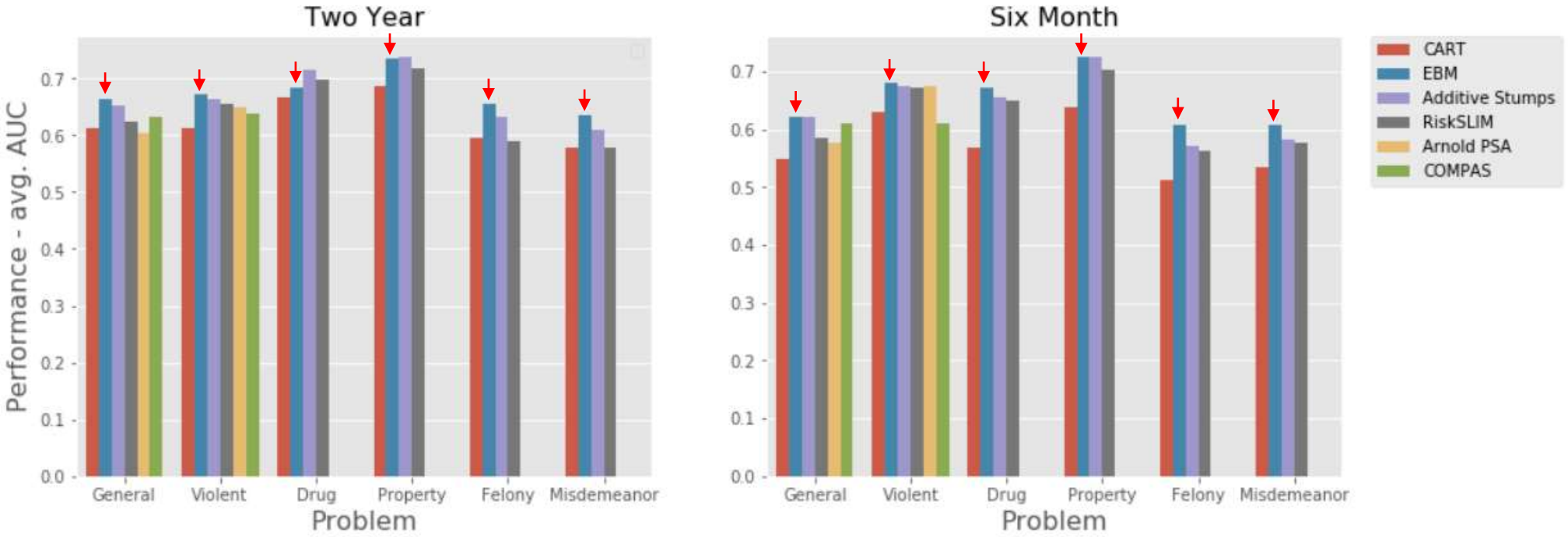
BlackBox

GlassBox

Table 1: Test set AUCs across 10 datasets. Best number in each row in **bold**.

	GAM									Full Complexity	
	EBM	EBM-BF	XGB	XGB-L2	FLAM	Spline	iLR	LR	mLR	RF	XGB-d3
Adult	**0.930**	0.928	0.928	0.917	0.925	0.920	0.927	0.909	0.925	0.912	**0.930**
Breast	0.997	0.995	0.997	0.997	**0.998**	0.989	0.981	0.997	0.985	0.993	0.993
Churn	**0.844**	0.840	0.843	0.843	0.842	**0.844**	0.834	0.843	0.827	0.821	0.843
Compas	0.743	**0.745**	**0.745**	0.743	0.742	0.743	0.735	0.727	0.722	0.674	**0.745**
Credit	0.980	0.973	0.980	0.981	0.969	**0.982**	0.956	0.964	0.940	0.962	0.973
Heart	0.855	0.838	0.853	0.858	0.856	0.867	0.859	**0.869**	0.744	0.854	0.843
MIMIC-II	0.834	0.833	0.835	0.834	0.834	0.828	0.811	0.793	0.816	**0.860**	0.847
MIMIC-III	0.812	0.807	**0.815**	**0.815**	0.812	0.814	0.774	0.785	0.776	0.807	0.820
Pneumonia	**0.853**	0.847	0.850	0.850	**0.853**	0.852	0.843	0.837	0.845	0.845	0.848
Support2	0.813	0.812	0.814	0.812	0.812	0.812	0.800	0.803	0.772	**0.824**	0.820
Average	**0.866**	0.862	**0.866**	0.865	0.864	0.865	0.852	0.853	0.835	0.855	**0.866**
Rank	3.70	6.70	**3.40**	4.90	5.05	4.60	8.70	7.75	9.70	7.40	4.10
Score	**0.893**	0.781	0.873	0.818	0.836	0.810	0.474	0.507	0.285	0.543	0.865

Chang, C.H., Tan, S., Lengerich, B., Goldenberg, A. and Caruana, R. "How Interpretable and Trustworthy are GAMs?" *KDD2021*



"We observed that the best interpretable models can perform approximately as well as the best black-box models(XGBoost)"

Wang, C., Han, B., Patel, B., Mohideen, F. and Rudin, C., 2020. In Pursuit of Interpretable, Fair and Accurate Machine Learning for Criminal Recidivism Prediction. *arXiv preprint arXiv:2005.04176*.

Table 1: AUC on the classification datasets for different learning methods. Each cell contains the mean AUC ± one standard deviation obtained via 5-fold cross validation. Higher AUCs are better.

Model	COMPAS	MIMIC-II	Credit Fraud
Logistic Regression	0.730± 0.014	0.791 ± 0.007	0.975 ± 0.010
Decision Trees	0.723 ± 0.010	0.768 ± 0.008	0.956 ± 0.004
NAMs	0.741 ± 0.009	0.830 ± 0.008	0.980 ± 0.002
EBMs	0.740 ± 0.012	0.835 ± 0.007	0.976 ± 0.009
XGBoost	0.742 ± 0.009	0.844 ± 0.006	0.981 ± 0.008
DNNs	0.735 ± 0.006	0.832 ± 0.009	0.978 ± 0.003

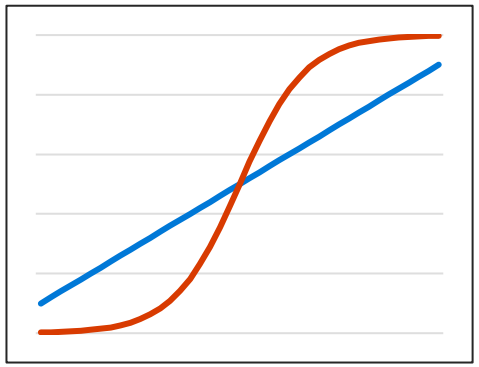
Table 2: RMSE on regression datasets for different learning methods. Each cell contains the mean RMSE ± one standard deviation obtained via 5-fold cross validation. Lower RMSE is better.

Model	California Housing	FICO Score
Linear Regression	0.728 ± 0.015	4.344 ± 0.056
Decision Trees	0.720 ± 0.006	4.900 ± 0.113
NAMs	0.562 ± 0.007	3.490 ± 0.081
EBMs	0.557 ± 0.009	3.512 ± 0.095
XGBoost	0.532 ± 0.014	3.345 ± 0.071
DNNs	0.492 ± 0.009	3.324 ± 0.092

Agarwal, R., Melnick, L., Lengerich, B. ,Frosst, N., Zhang, X., Caruana, R. & Hinton, G.E., *Neural Additive Models: Interpretable Machine Learning with Neural Nets,* NeurIPS 2021.

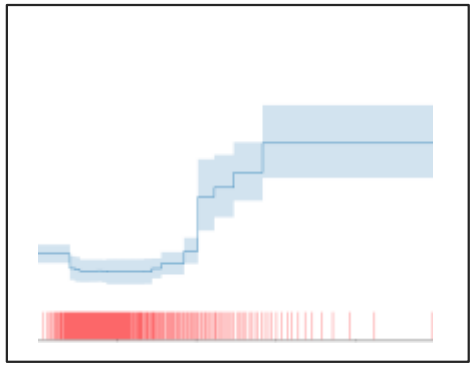
EBMs: Generalized Additive Models (GAMs)

Linear/Logistic Regression



- Interpretable
- Not very accurate
- Can't model nonlinearities
- Sometimes gets sign wrong!

GAMs/EBMs



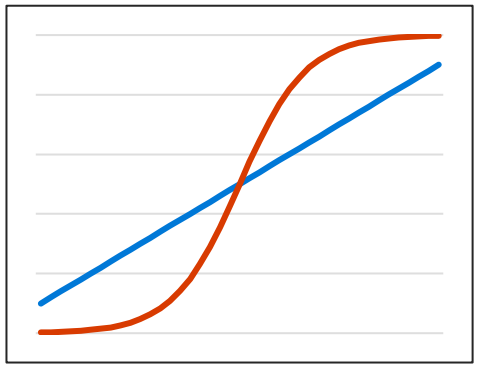
- More interpretable than linear/logistic
- Can be very accurate
- Can model nonlinearities
- More likely to show important effects
- Invented by Hastie & Tibshirani 1980's

BlackBox Machine Learning

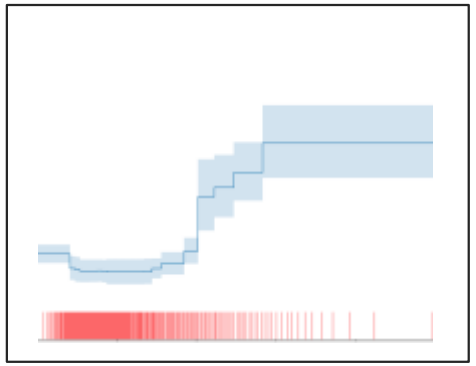
- Not interpretable (blackbox)
- Can be very accurate
- Can model nonlinearities
- More likely to learn spurious effects

EBMs: Generalized Additive Models (GAMs)

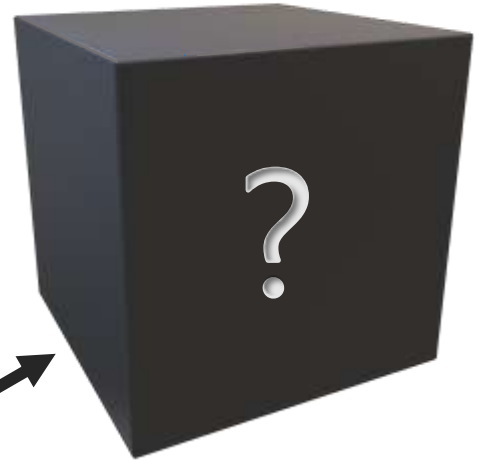
Linear/Logistic Regression



GAMs/EBMs



BlackBox Machine Learning



- Linear Model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$
- Generalized Additive Model: $y = f_1(x_1) + f_2(x_2) + \ldots + f_n(x_n)$
- Additive Model with Pairwise Interactions: $y = \Sigma_i f_i(x_i) + \Sigma_{ij} f_{ij}(x_i, x_j) + \Sigma_{ijk} f_{ijk}(x_i, x_j, x_k)$
- Full Complexity Models: $y = f(x_1, \ldots, x_n)$

Example 1: Pneumonia Mortality

Pneumonia Dataset (collected 1989): 46 Features

Patient-history findings
Age (years)
Gender
A re-admission to the hospital
Admitted from a nursing home
Admitted through the ER
Has a chronic lung disease
Has asthma
Has diabetes mellitus
Has congestive heart failure
Has ischemic heart disease
Has cerebrovascular disease
Has chronic liver disease
Has chronic renal failure
Has history of seizures
Has cancer
Number of above disease conditions
Pleuritic of chest pain

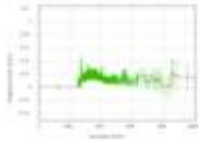
Physical examination findings
Respiration rate (resps/min)
Heart rate (beats/min)
Systolic blood pressure (mmHg)
Temperature (°C)
Altered mental status (disorientation, lethargy, or coma)
Wheezing
Stridor
Heart murmur
Gastrointestinal bleeding

Laboratory findings
Sodium level (mEq/l)
Potassium level (mEq/l)
Creatinine level (mg/dl)
Glucose level (mg/dl)
BUN level (mg/dl)
Liver function tests (coded only as normal* or abnormal)
Albumin level (gm/dl)
Hematocrit
White blood cell count (1000 cells/μl)
Percentage bands
Blood pH
Blood pO_2 (mmHg)
Blood pCO_2 (mmHg)

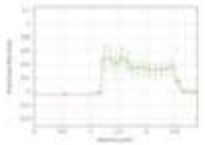
Chest X-ray findings
Positive chest X-ray
Lung infiltrate
Pleural effusion
Pneumothorax
Cavitation/empyema
Lobe or lung collapse
Chest mass

Age => -0.23

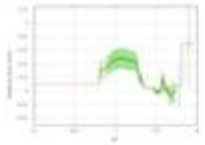
Asthma => -0.15



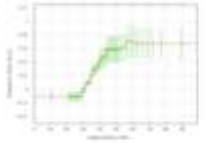
Glucose => +0.18



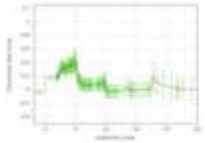
Albumin => +0.01



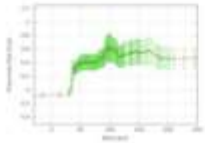
Blood pH => +0.38



Respiration => +0.21



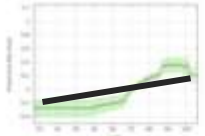
Creatinine => -0.01



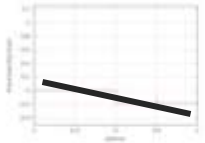
BUN => -0.21

$$Score = baseline + \sum_{i=0}^{n} f_i(variable_i)$$

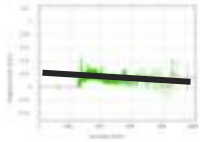
$$\text{POD} = \frac{1}{1+e^{-Score}}$$



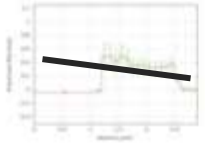
Age => -0.23



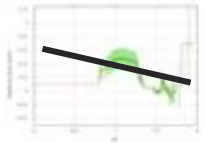
Asthma => -0.15



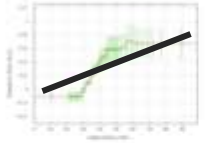
Glucose => +0.18



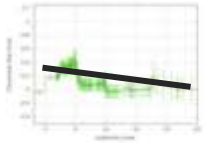
Albumin => +0.01



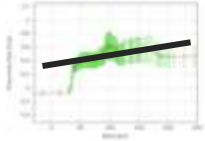
Blood pH => +0.38



Respiration => +0.21



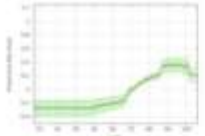
Creatinine => -0.01



BUN => -0.21

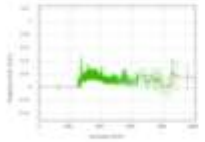
$$Score = baseline + \sum_{i=0}^{n} f_i(variable_i)$$

$$POD = \frac{1}{1+e^{-\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n}}$$

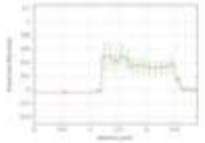


Age => -0.23

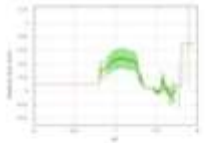
Asthma => -0.15



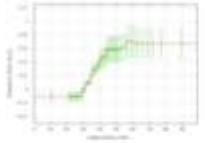
Glucose => +0.18



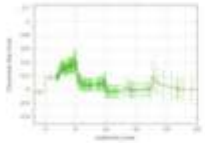
Albumin => +0.01



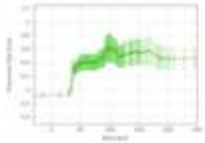
Blood pH => +0.38



Respiration => +0.21



Creatinine => -0.01



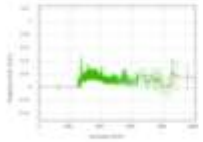
BUN => -0.21

$$Score = baseline + \sum_{i=0}^{n} f_i(variable_i)$$

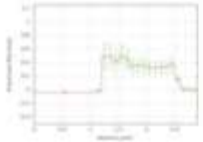
$$POD = \frac{1}{1+e^{-}} + \dots +$$

Age => -0.23

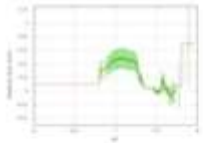
Asthma => -0.15



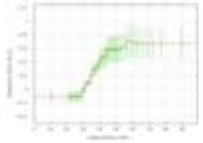
Glucose => +0.18



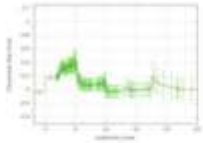
Albumin => +0.01



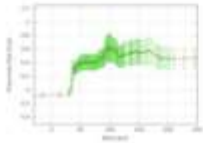
Blood pH => +0.38



Respiration => +0.21



Creatinine => -0.01



BUN => -0.21

$$Score = baseline + \sum_{i=0}^{n} f_i(variable_i)$$

$$\text{POD} = \frac{1}{1+e^{-Score}}$$

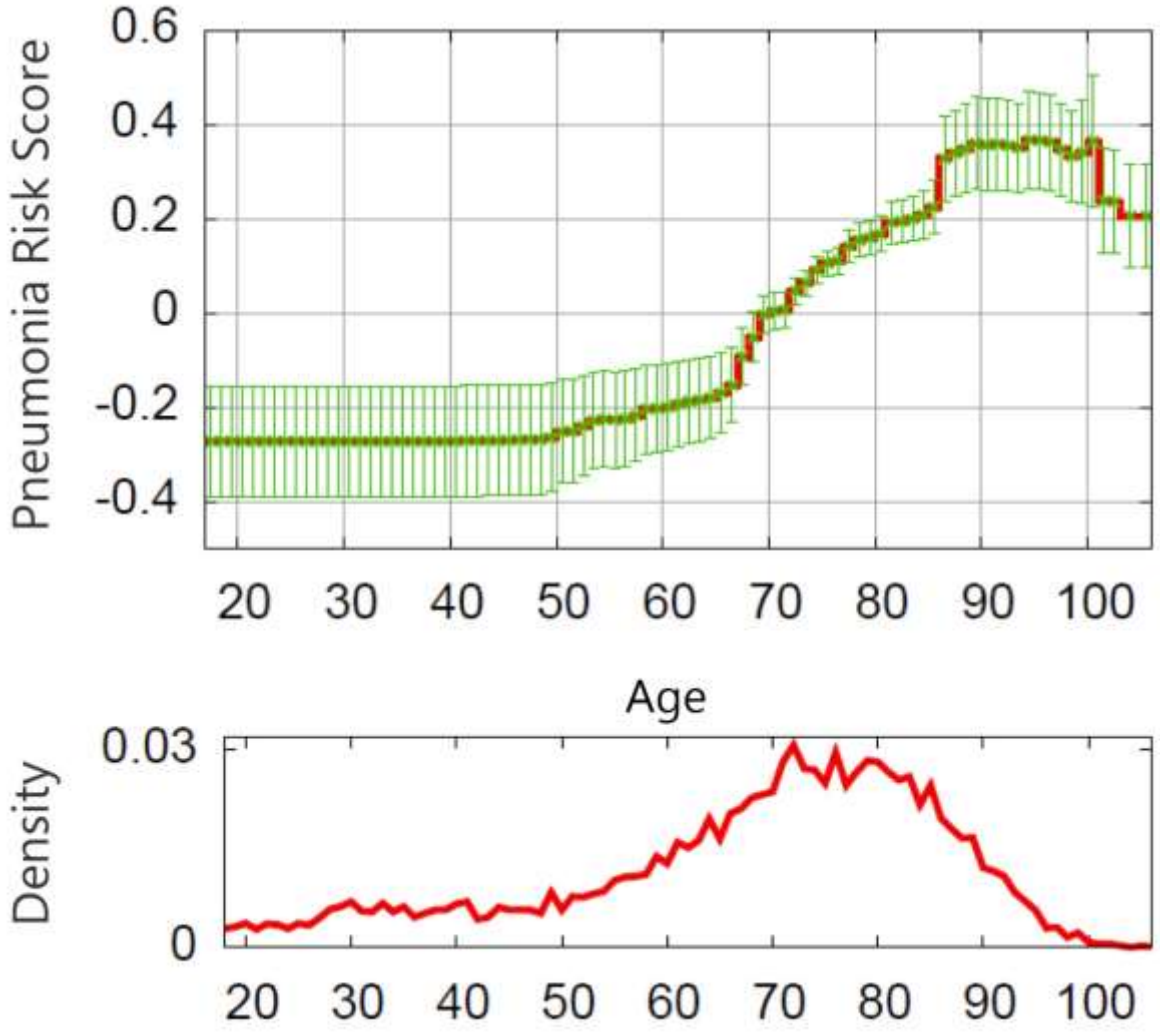
$$Score = -2.11 - 0.23 - 0.15 + 0.18 + 0.01 + 0.38 + \ \ldots$$

$$Score = -0.78$$

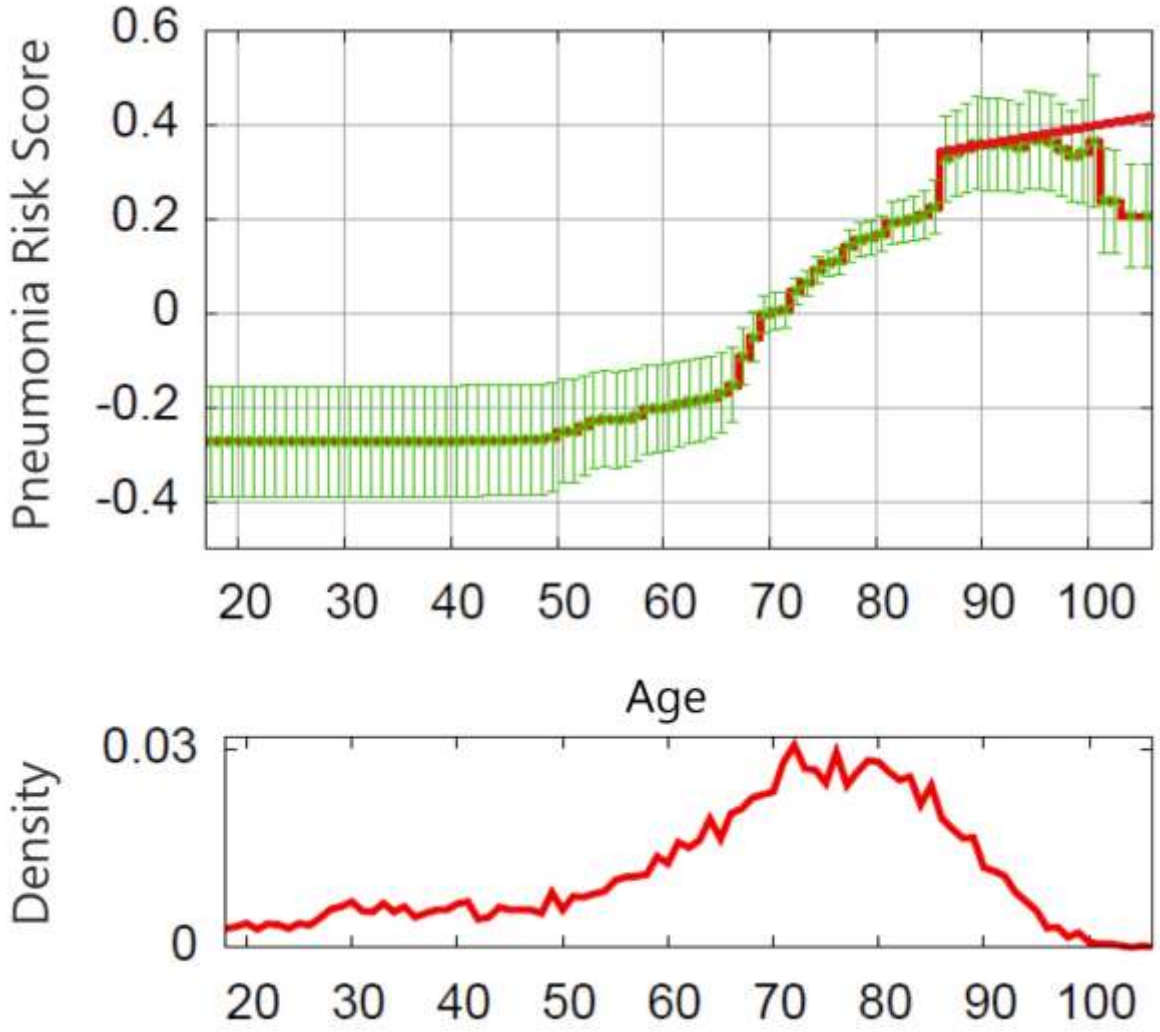
$$\text{POD} = \frac{1}{1+e^{-(-0.78)}}$$

$$\text{POD} = 0.3143$$

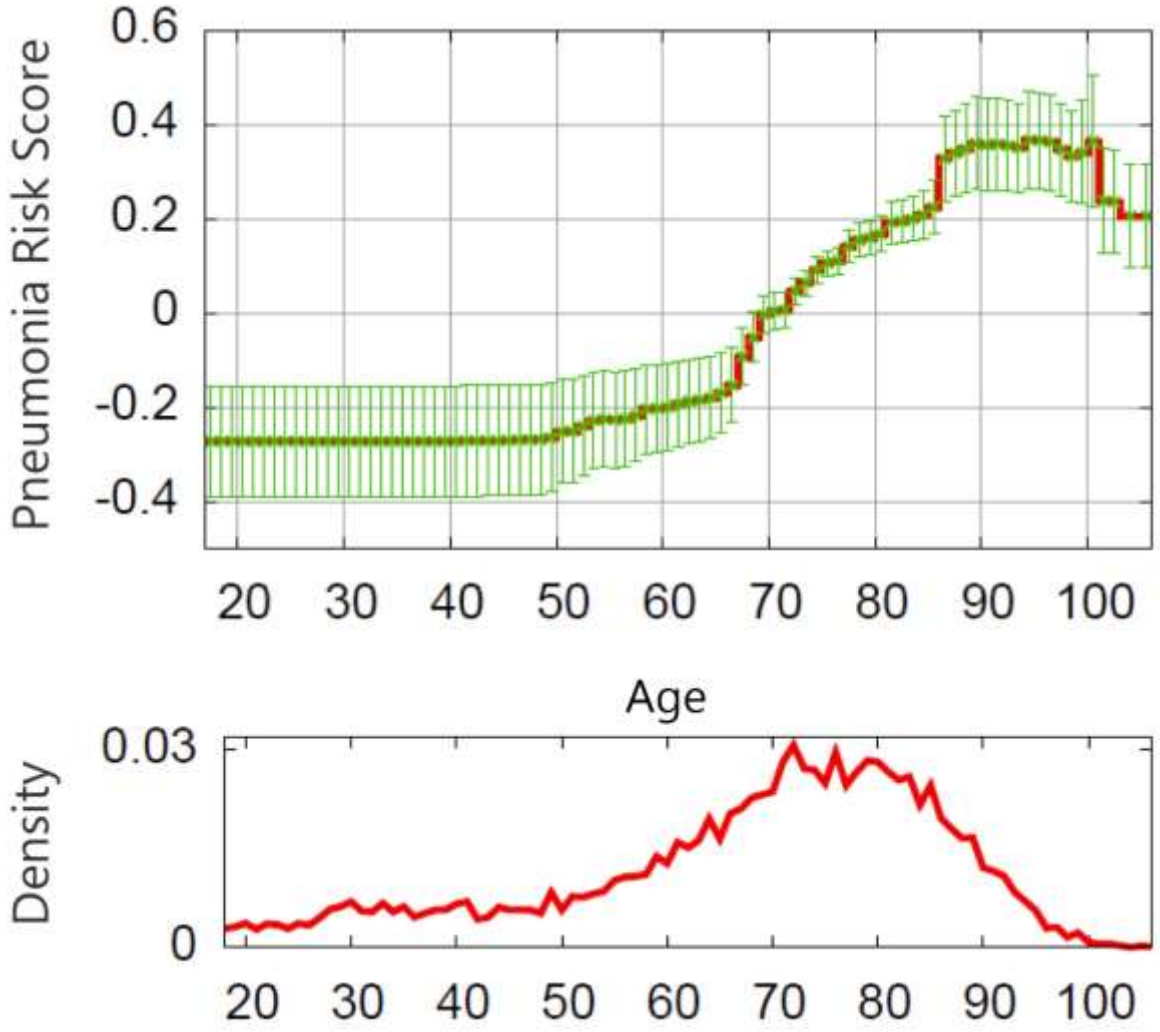
What EBMs Learn about Pneumonia Risk vs. Age



Fix Age > 100 Problem (Enforce Monotonicity)

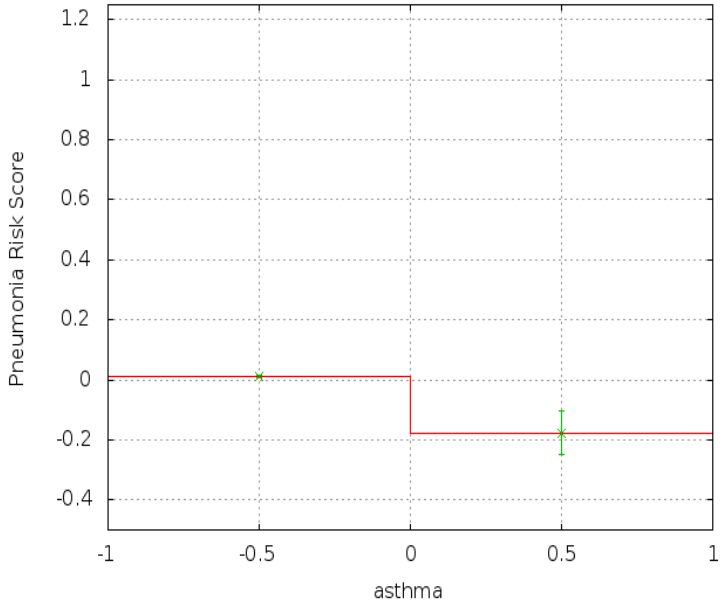


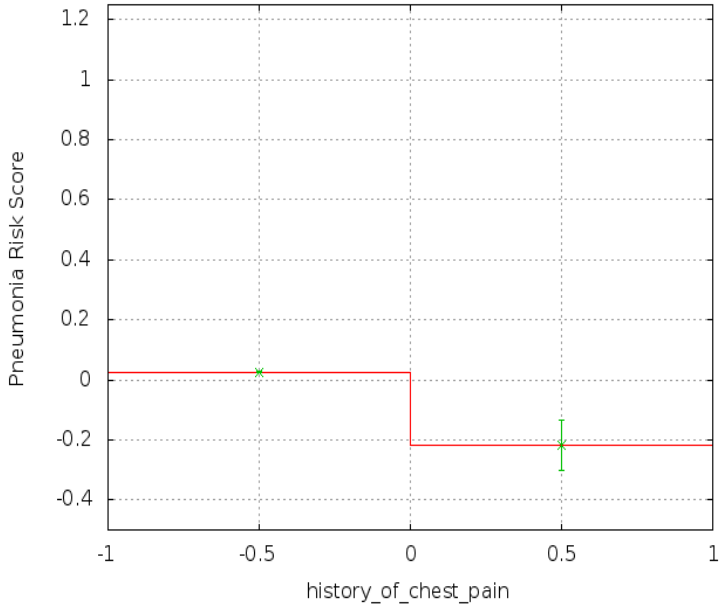
Original Model is Correct for Actuarial Use

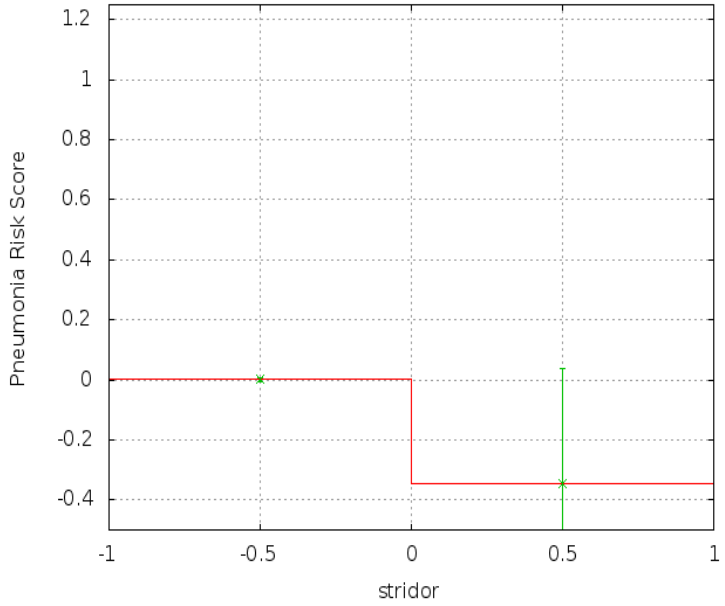


- Model correctness (and accuracy) depends on how model will be used
  - This is a good model for health insurance provides
  - But needs to be repaired before using for patient treatment decision

- A few things intelligible model learned:
  - Beware of jumps at round numbers --- almost always due to human/social/policy effects
  - Asthma => lower risk
  - History of chest pain => lower risk
  - History of heart disease => lower risk
  - Obstructed airway => lower risk

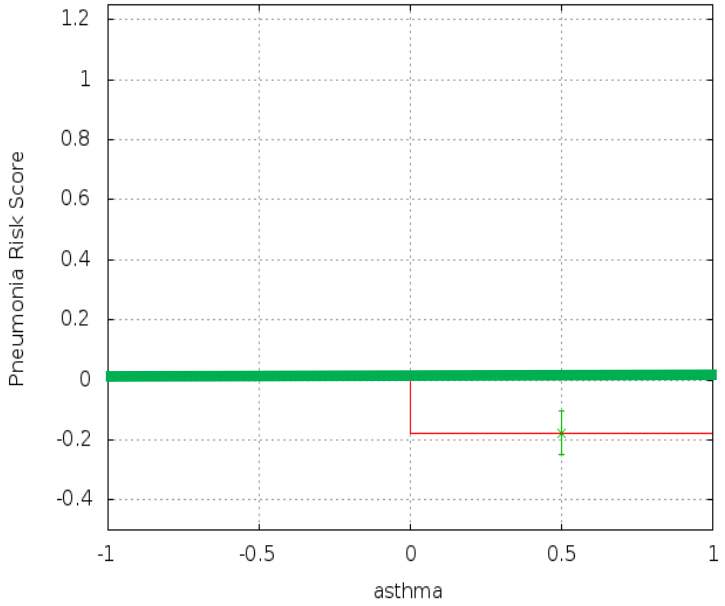


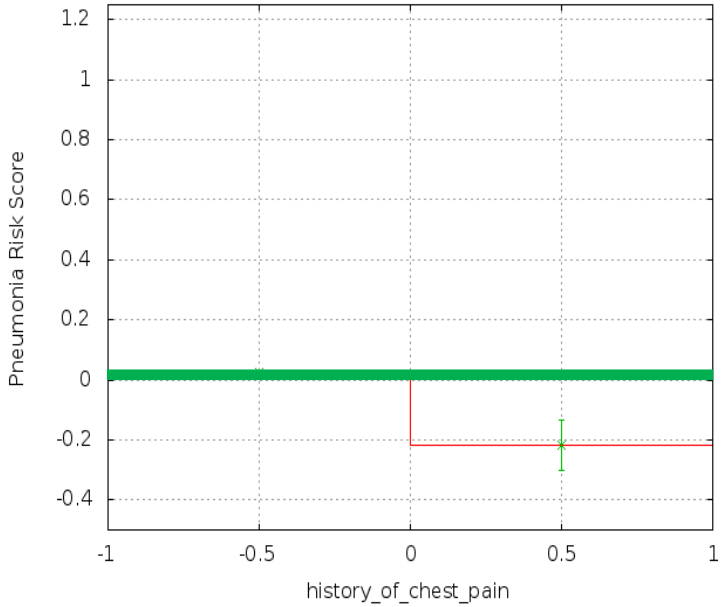


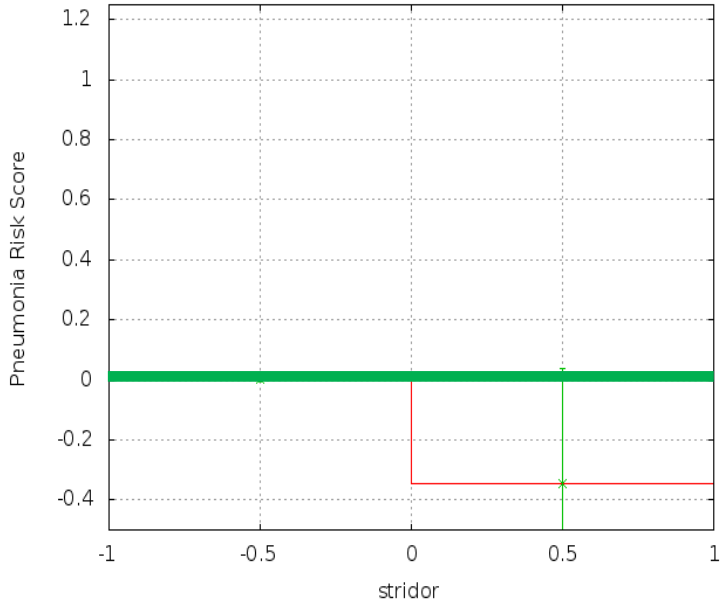


- Model correctness (and accuracy) depends on how model will be used
  - This is a good model for health insurance provides
  - But needs to be repaired before using for patient treatment decision

- A few things intelligible model learned:
  - Beware of jumps at round numbers --- almost always due to human/social/policy effects
  - Asthma => lower risk
  - History of chest pain => lower risk
  - History of heart disease => lower risk
  - Obstructed airway => lower risk

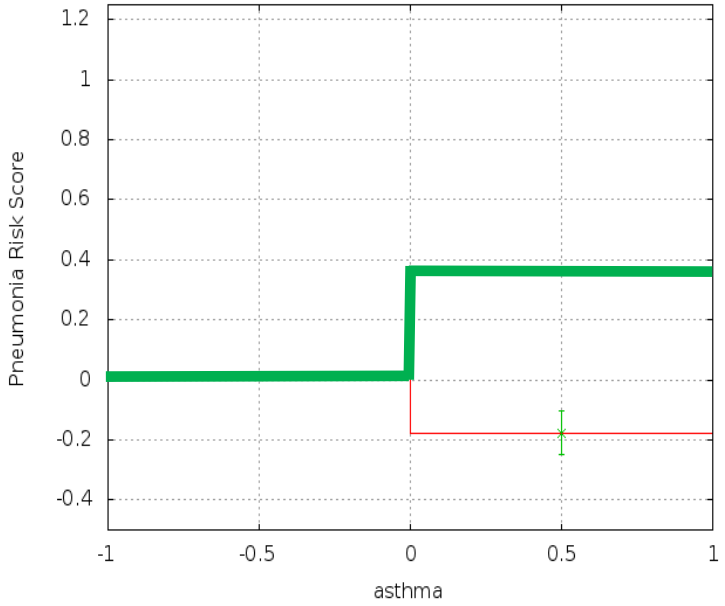


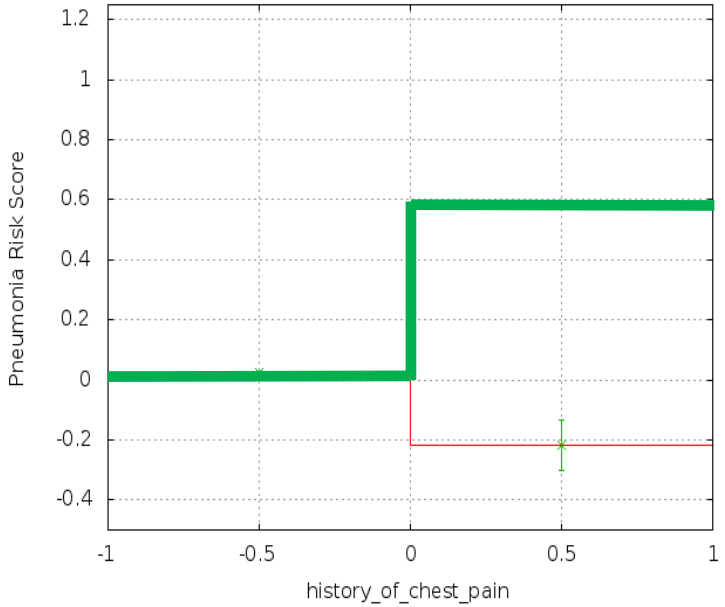


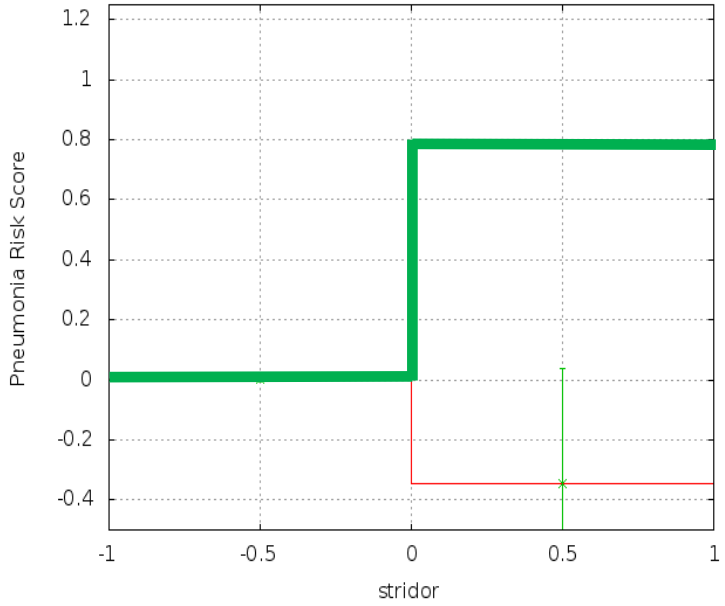


- Model correctness (and accuracy) depends on how model will be used
  - This is a good model for health insurance provides
  - But needs to be repaired before using for patient treatment decision

- A few things intelligible model learned:
  - Beware of jumps at round numbers --- almost always due to human/social/policy effects
  - Asthma => lower risk
  - History of chest pain => lower risk
  - History of heart disease => lower risk
  - Obstructed airway => lower risk

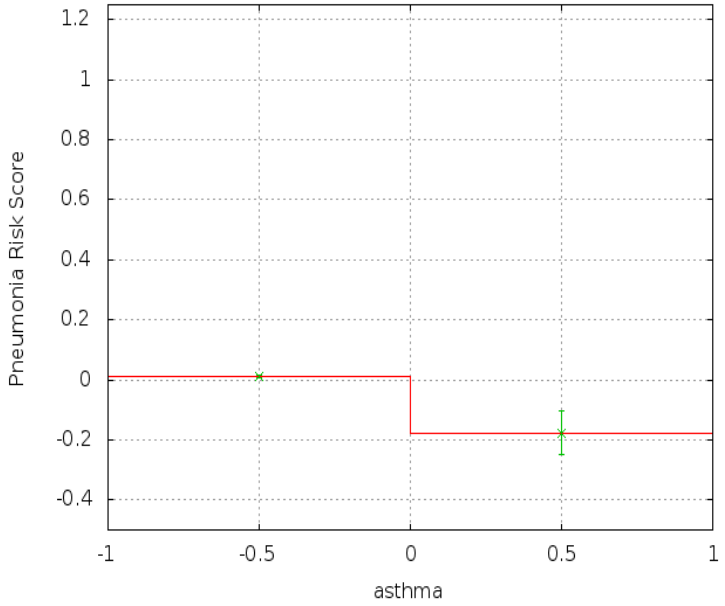


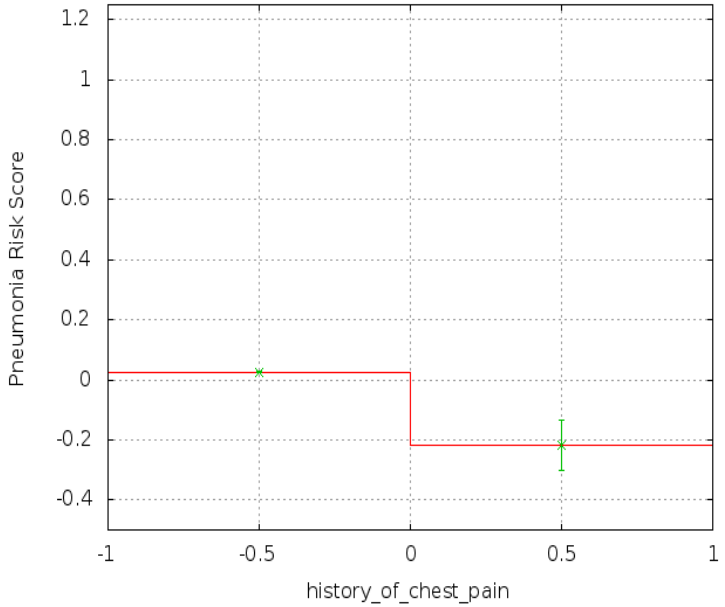


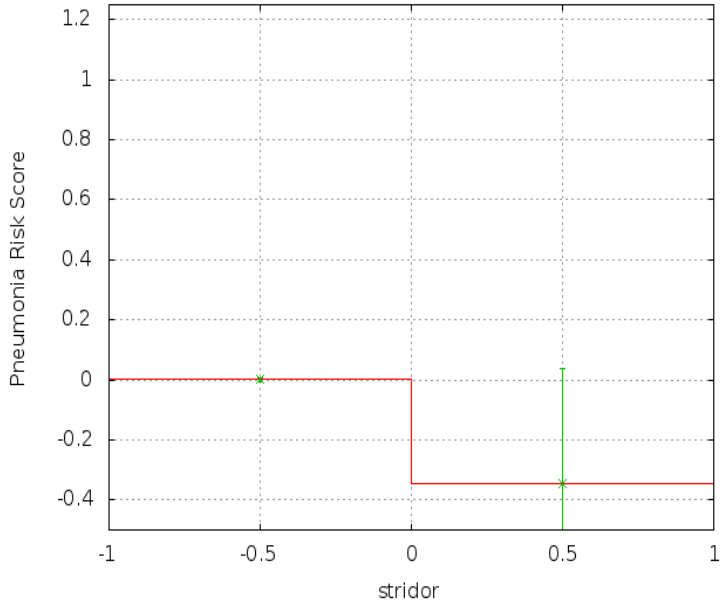


- Model correctness (and accuracy) depends on how model will be used
  - This is a good model for health insurance provides
  - But needs to be repaired before using for patient treatment decision

- A few things intelligible model learned:
  - Beware of jumps at round numbers --- almost always due to human/social/policy effects
  - Asthma => lower risk
  - History of chest pain => lower risk
  - History of heart disease => lower risk
  - Obstructed airway => lower risk





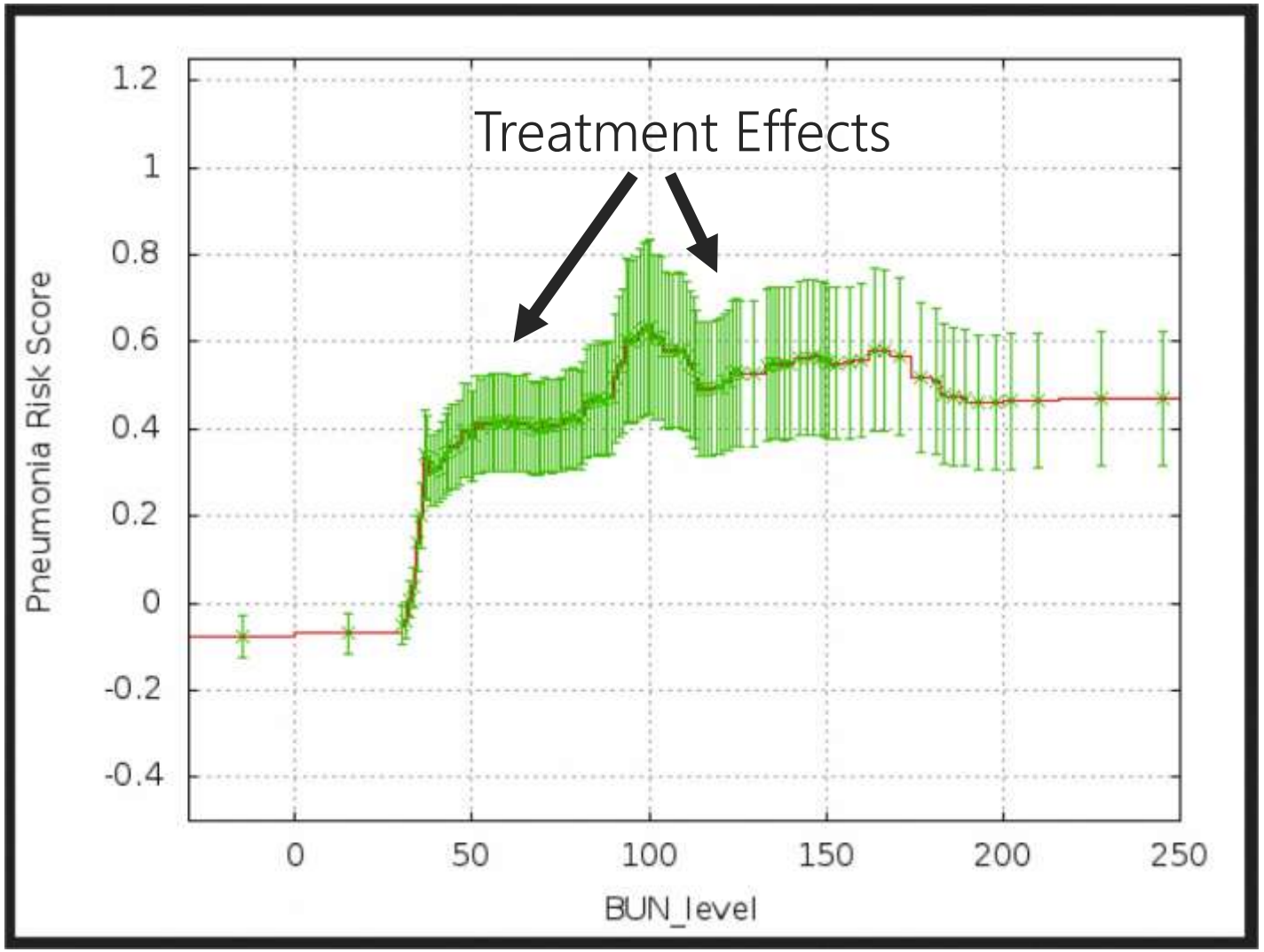


- Model correctness (and accuracy) depends on how model will be used
  - This is a good model for health insurance provides
  - But needs to be repaired before using for patient treatment decision

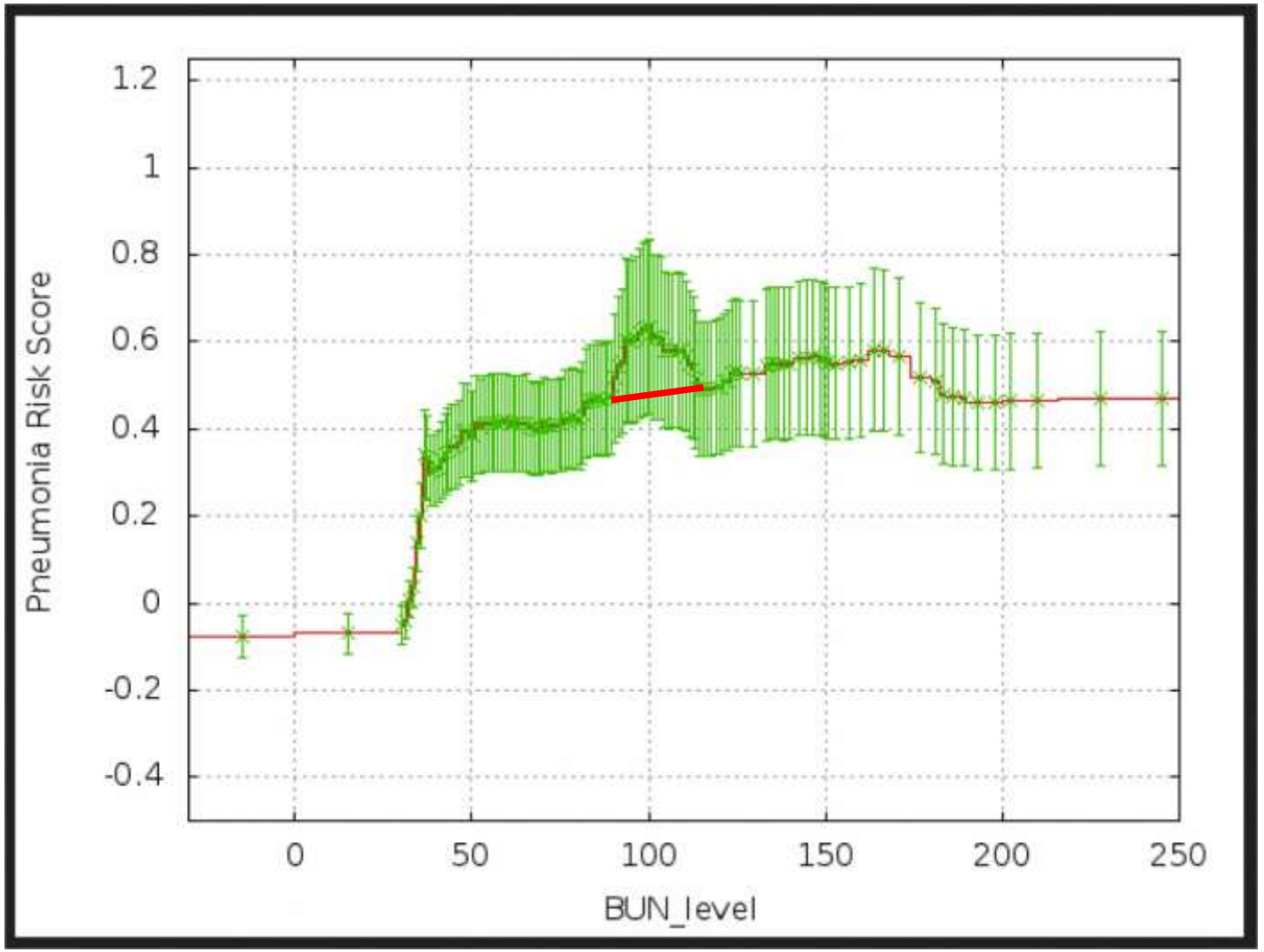
- A few things intelligible model learned:
  - Beware of jumps at round numbers --- almost always due to human/social/policy effects
  - Asthma => lower risk
  - History of chest pain => lower risk
  - History of heart disease => lower risk
  - Obstructed airway => lower risk
  - …
  - Model is rewarded with high accuracy on test set for predicting these things!

- Important: Must keep potentially offending features in model!
  - Let model become as biased as it can be
  - Then delete or edit terms after seeing what model learned

Intelligibility Can Create New Medical Science

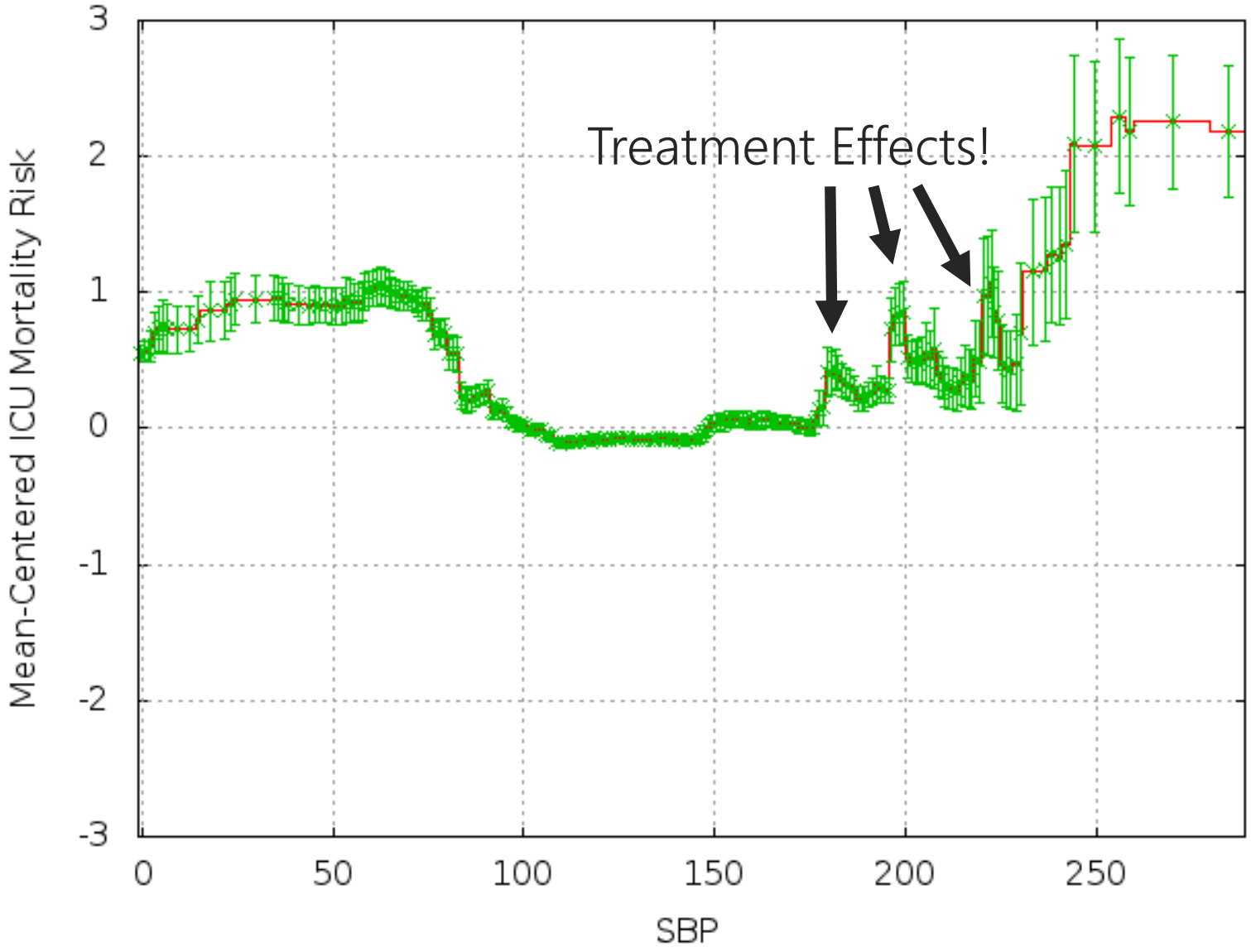


Intelligibility Can Create New Medical Science



Can save 2500 lives per year in U.S. alone

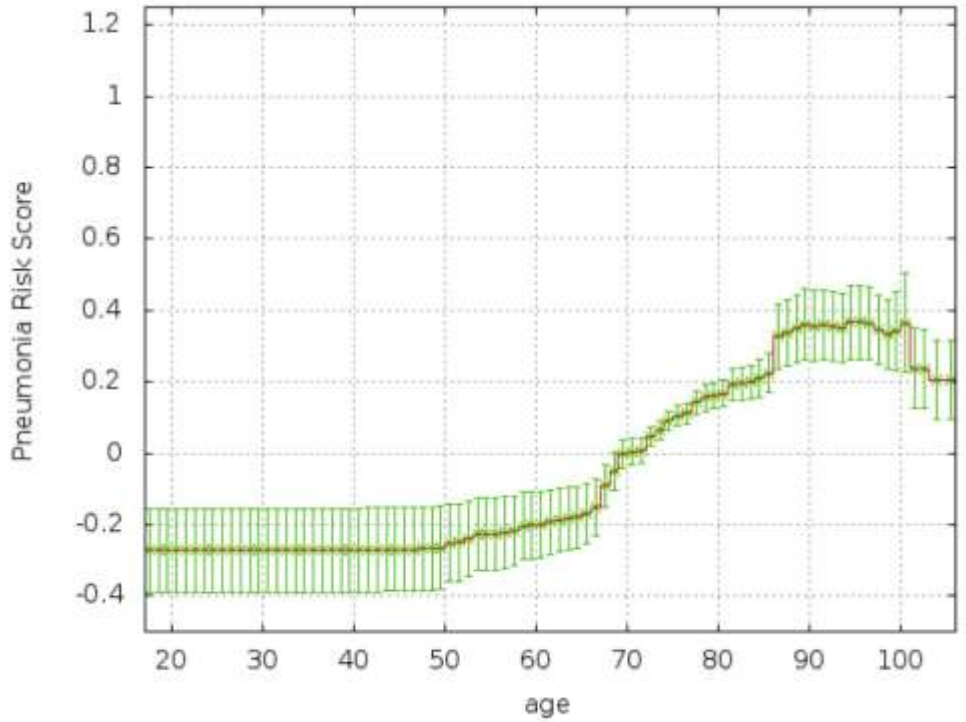
Treatment Effects Ubiquitous in All Medical Data

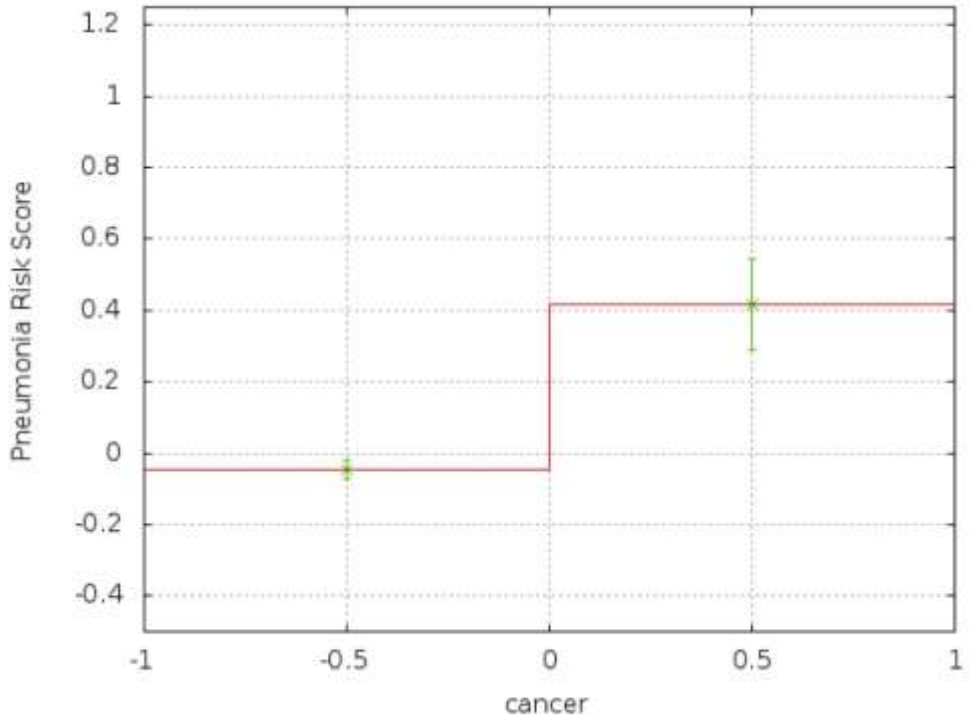


Pairwise Interactions?

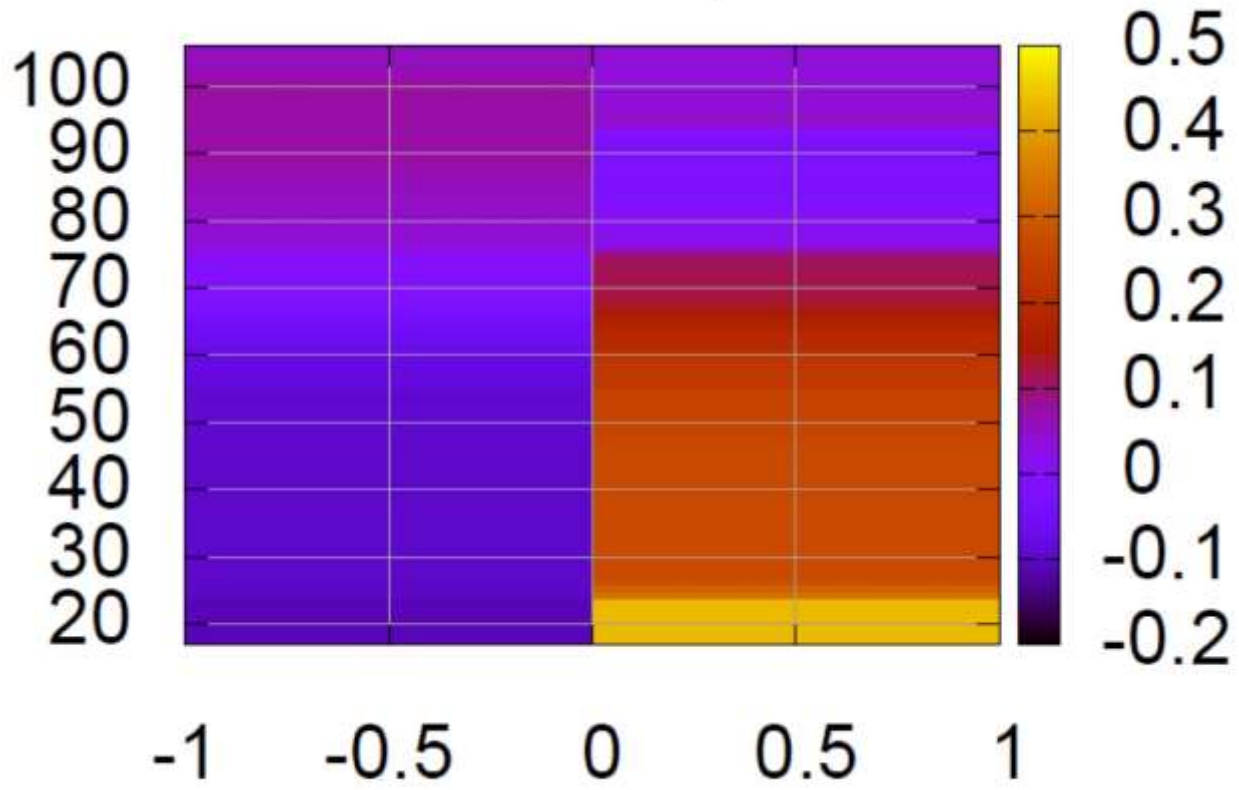
Like XOR (parity), interactions can't be modeled as a sum of independent effects:

$$f(b_1) + f(b_2) \neq f(b_1, b_2)$$





Pairwise Interaction: Age x Cancer (Pneumonia-95)



Age vs. Cancer

Example 2: Housing Price Data

Housing Pricing Data

Housing Pricing Data

In [74]:
```
df_filt[df_filt['YearBuilt'] == 1989].sort_values('SoldPrice', ascending=False)
```

executed in 83ms, finished 00:52:17 2020-08-14

Out[74]:

	SoldPrice	NEW House Type	NEW Zipcode	Bedrooms	Bathrooms	HouseSizeSqm	LotSizeSqm	YearBuilt	New City
58799	8094000	Condo/Coop/Timeshare	98136	1	1.00	50.91	1375.93	1989	Seattle
58798	8094000	Condo/Coop/Timeshare	98136	1	1.00	50.91	1375.93	1989	Seattle
58797	8094000	Condo/Coop/Timeshare	98136	1	1.00	48.31	1375.93	1989	Seattle
58789	8094000	Condo/Coop/Timeshare	98136	2	2.00	70.98	1393.17	1989	Seattle
58788	8094000	Condo/Coop/Timeshare	98136	2	2.00	70.61	1393.17	1989	Seattle
58787	8094000	Condo/Coop/Timeshare	98136	2	2.00	66.89	1393.17	1989	Seattle
58786	8094000	Condo/Coop/Timeshare	98136	1	1.00	47.94	1375.93	1989	Seattle
53120	1940000	Single Family	98102	4	3.00	318.66	340.68	1989	Seattle

Example 3: Wikipedia Malicious Edits

Wikipedia/Wikimedia

- 160,000 edits per day, 10-15% of which are flagged as damaging (e.g., malicious)
- Current ML tools are not intelligible, do not give help or explanations to editors

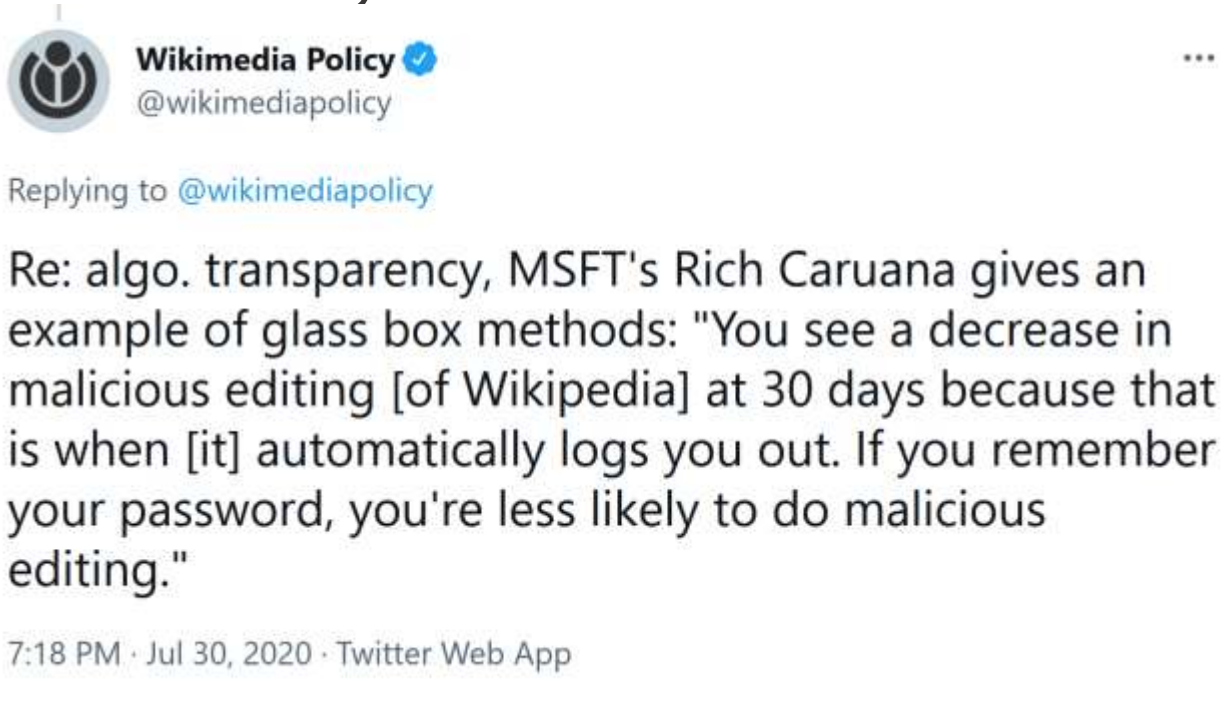
Wikipedia/Wikimedia

- 160,000 edits per day, 10-15% of which are flagged as damaging (e.g., malicious)
- Current ML tools are not intelligible, do not give help or explanations to editors

Wikipedia/Wikimedia

- 160,000 edits per day, 10-15% of which are flagged as damaging (e.g., malicious)
- Current ML tools are not intelligible, do not give help or explanations to editors

Impact on Real Policy



WIKIMEDIA FOUNDATION

- Overall, Wikipedia protected about **2,000 election-related pages**. Restrictions were put in place so that many of the most important election-related pages, such as the main page about the U.S. 2020 Presidential Election, could be edited only by the most trusted and experienced Wikipedia editors.

The Economist

"For America's recent presidential election, editing articles was restricted to accounts more than 30 days old, and with at least 500 edits …" – The Economist, Jan 7th 2021

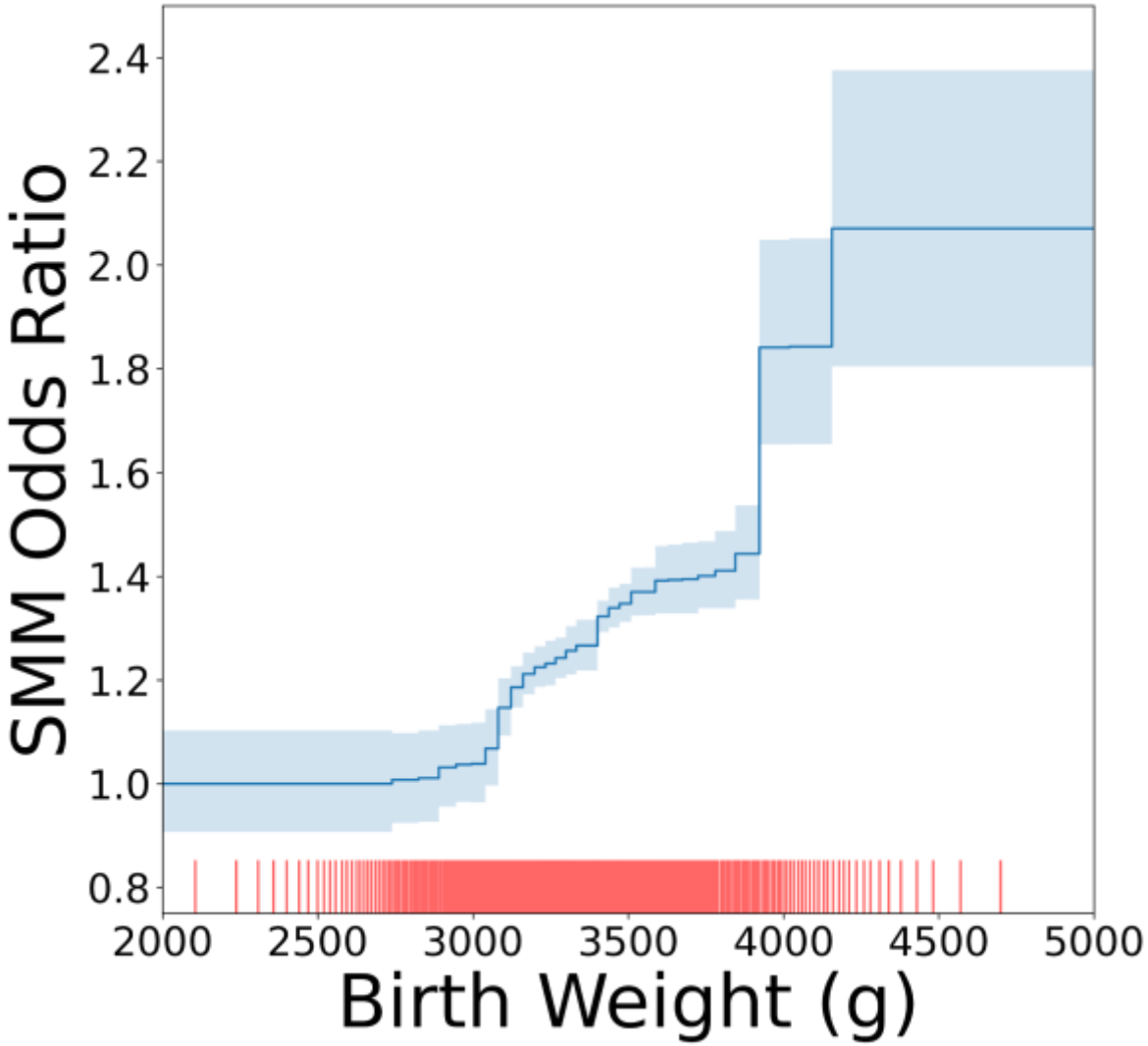
Example 4: Severe Maternal Morbidity

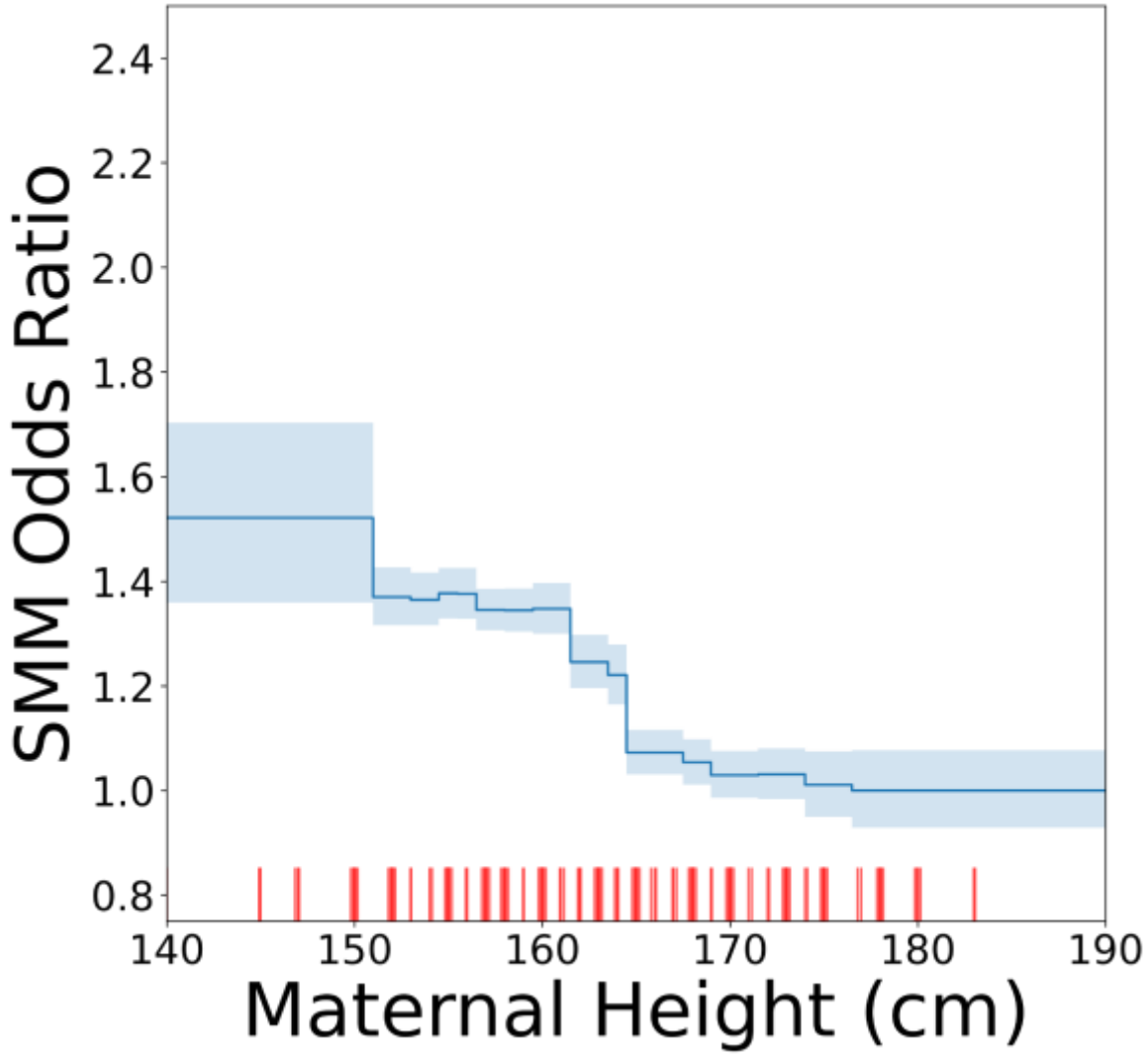
Pregnancy & Severe Maternal Morbidity (SMM)

- SMM: predicting maternal risk during labor in NTSV population:
  - Hemorrhage or need for blood transfusion
  - Thromboembolism
  - Hysterectomy
  - Eclampsia
  - …
- Before our work, the main risk factors for severe maternal morbidity (SMM) were:
  - Maternal hypertension (pre-eclampsia)
  - Maternal diabetes
  - Maternal obesity
  - …

Rich Caruana, Ben Lengerich (CMU), Vivienne Suiter M.D. (FHCQ)

Intelligible ML Says Most Important Factors Are…

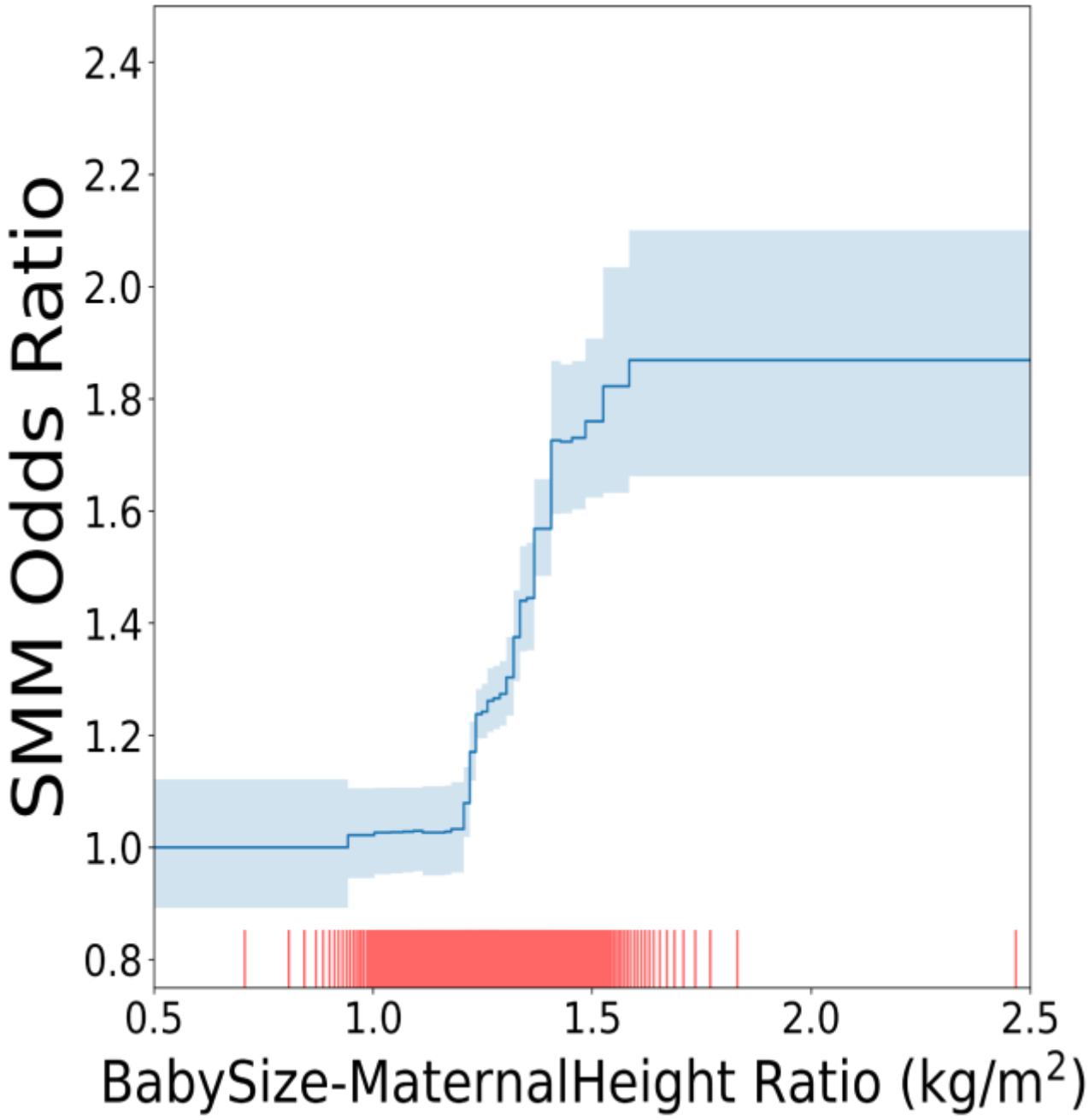




Joint work with Ben Lengerich CMU/MIT), Vivienne Souter (FHCQ)

“BMI” for Pregnancy

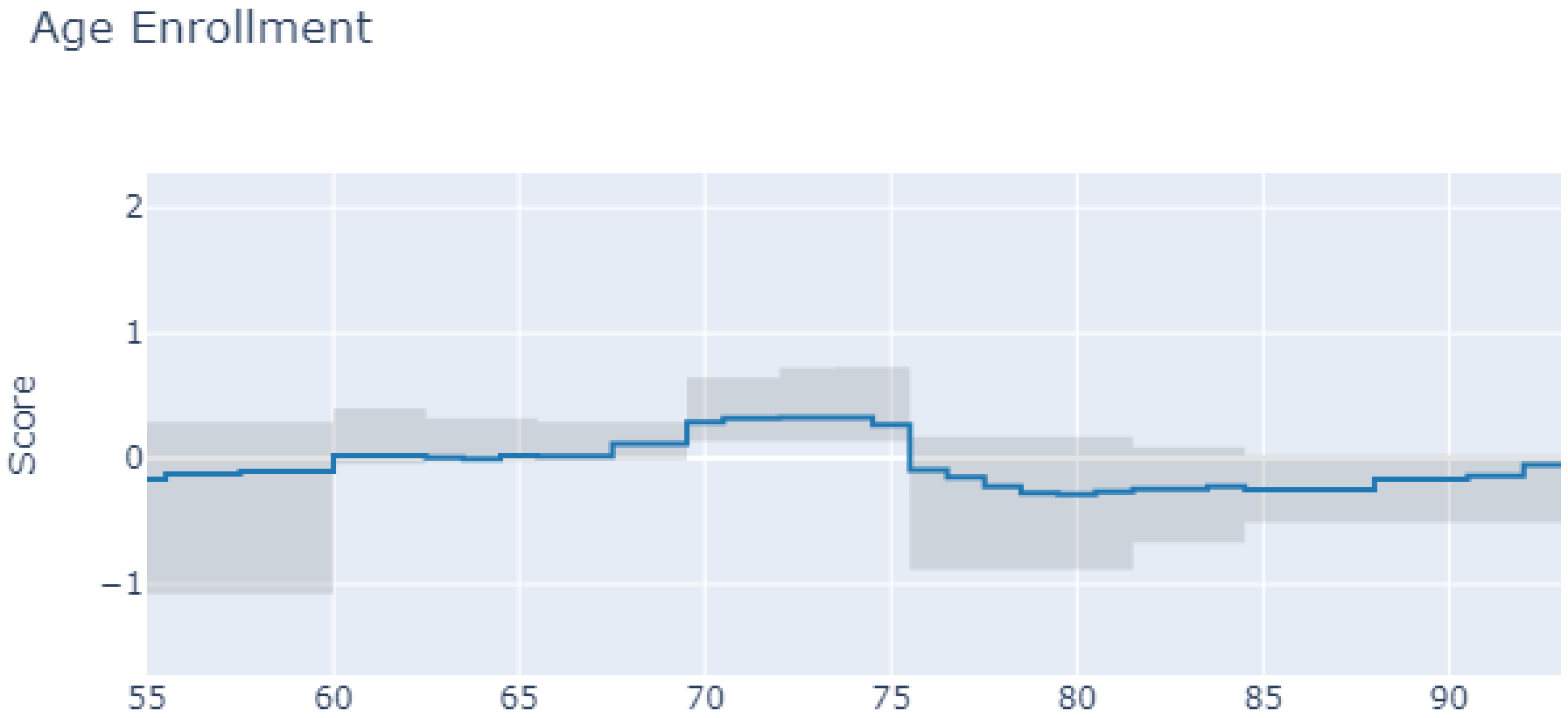
$$\frac{BabyBirthWeight}{MaternalHeight^2}$$



Joint work with Ben Lengerich CMU/MIT), Vivienne Souter (FHCQ)

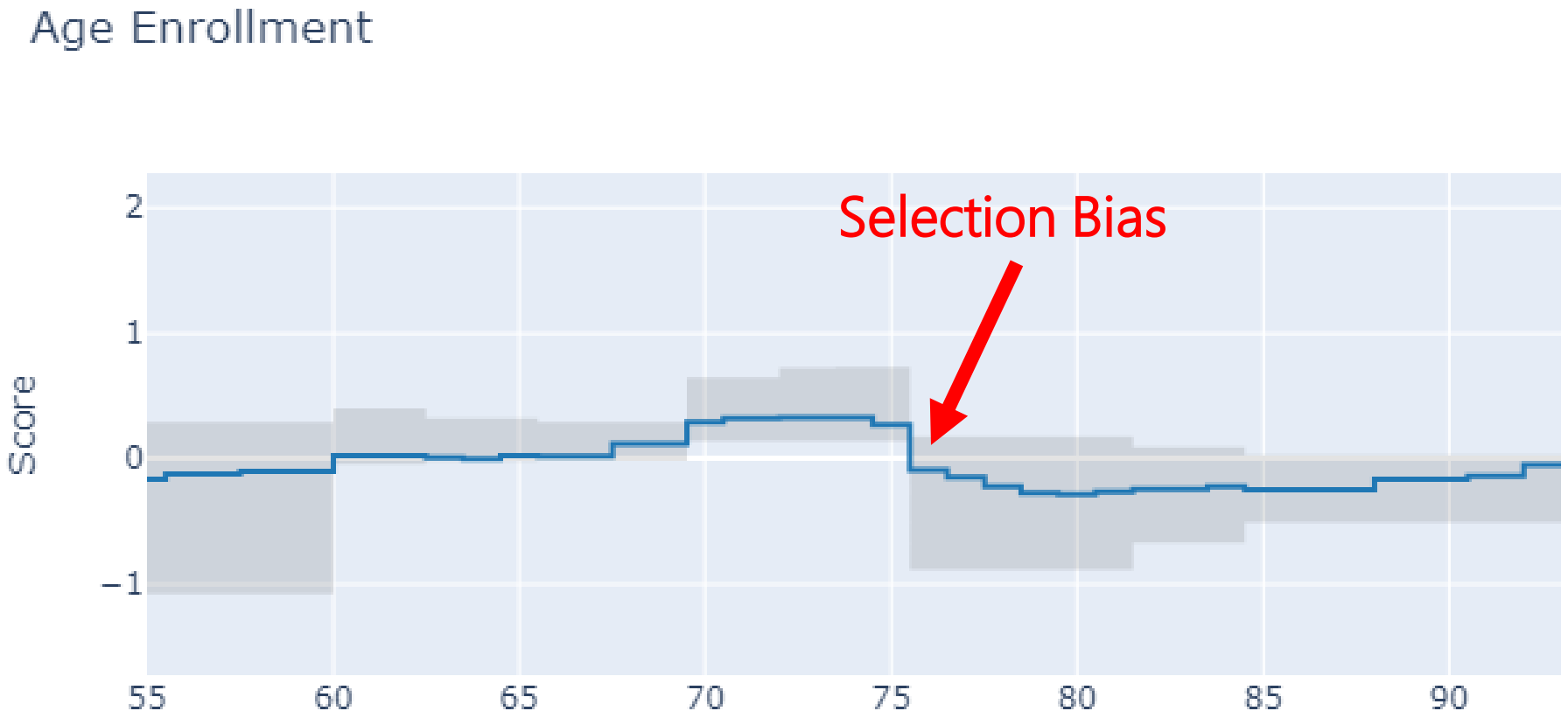
Example 5: Cancer Clinical Trial

Cancer Mortality Risk vs. Age



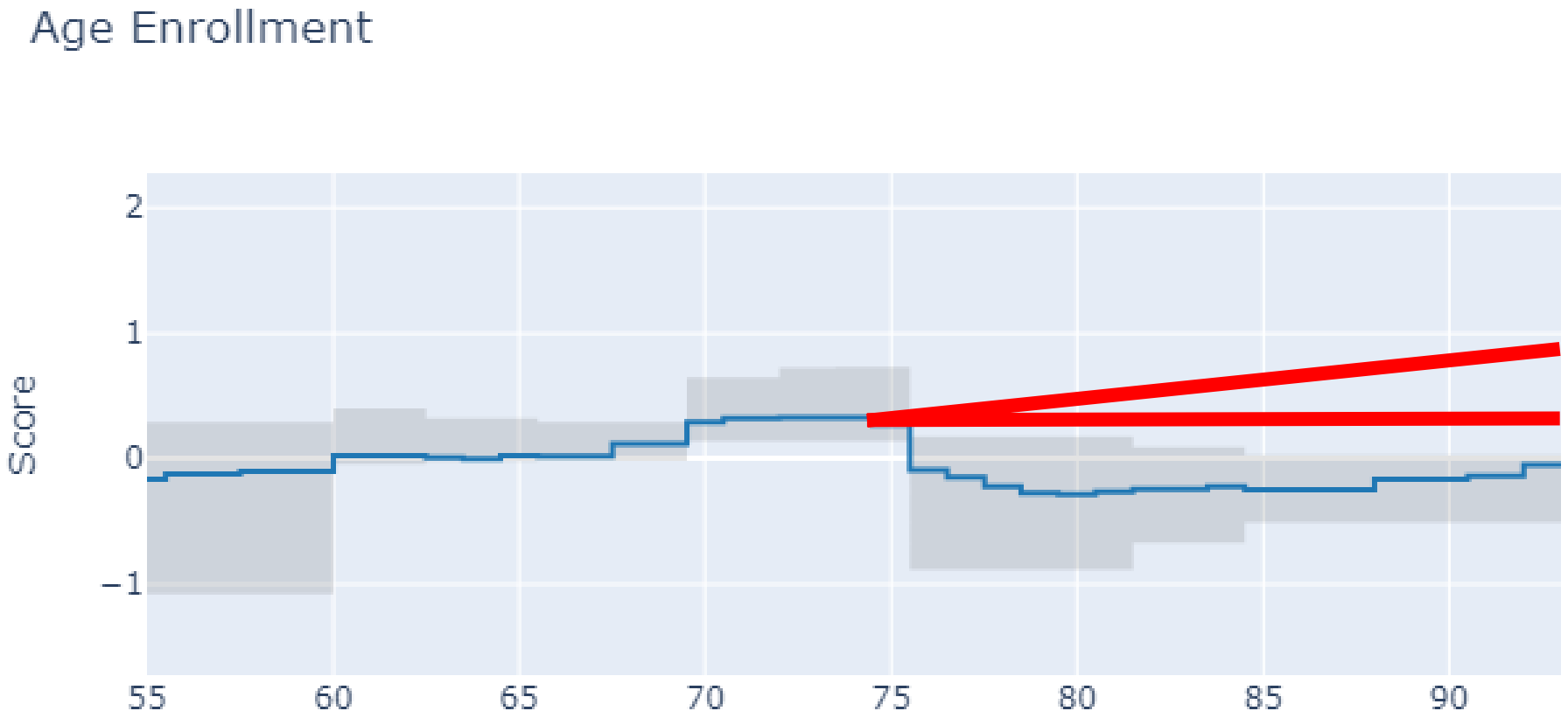
Zheng Zhang, Ying Xiao M.D., Sang Ho Lee (University of Pennsylvania), Rich Caruana (Microsoft)

Cancer Mortality Risk vs. Age



Zheng Zhang, Ying Xiao M.D., Sang Ho Lee (University of Pennsylvania), Rich Caruana (Microsoft)

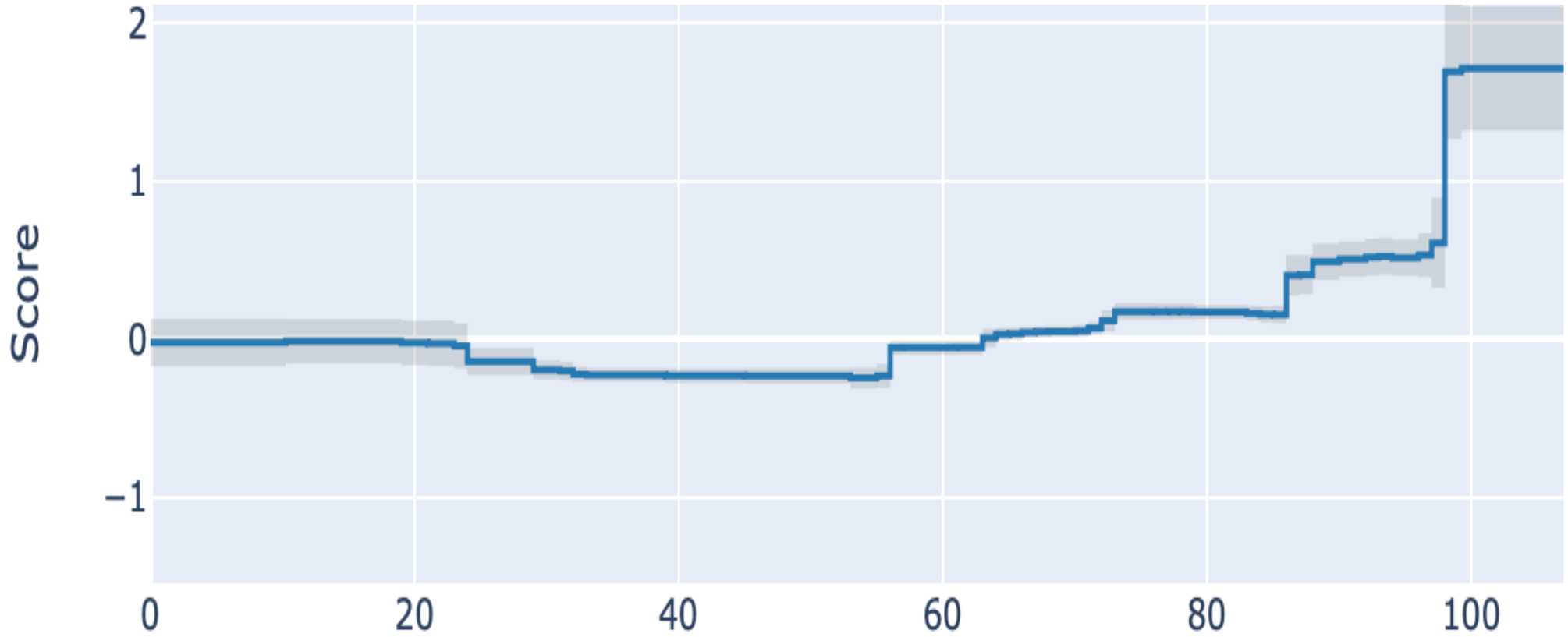
Cancer Mortality Risk vs. Age



Zheng Zhang, Ying Xiao M.D., Sang Ho Lee (University of Pennsylvania), Rich Caruana (Microsoft)

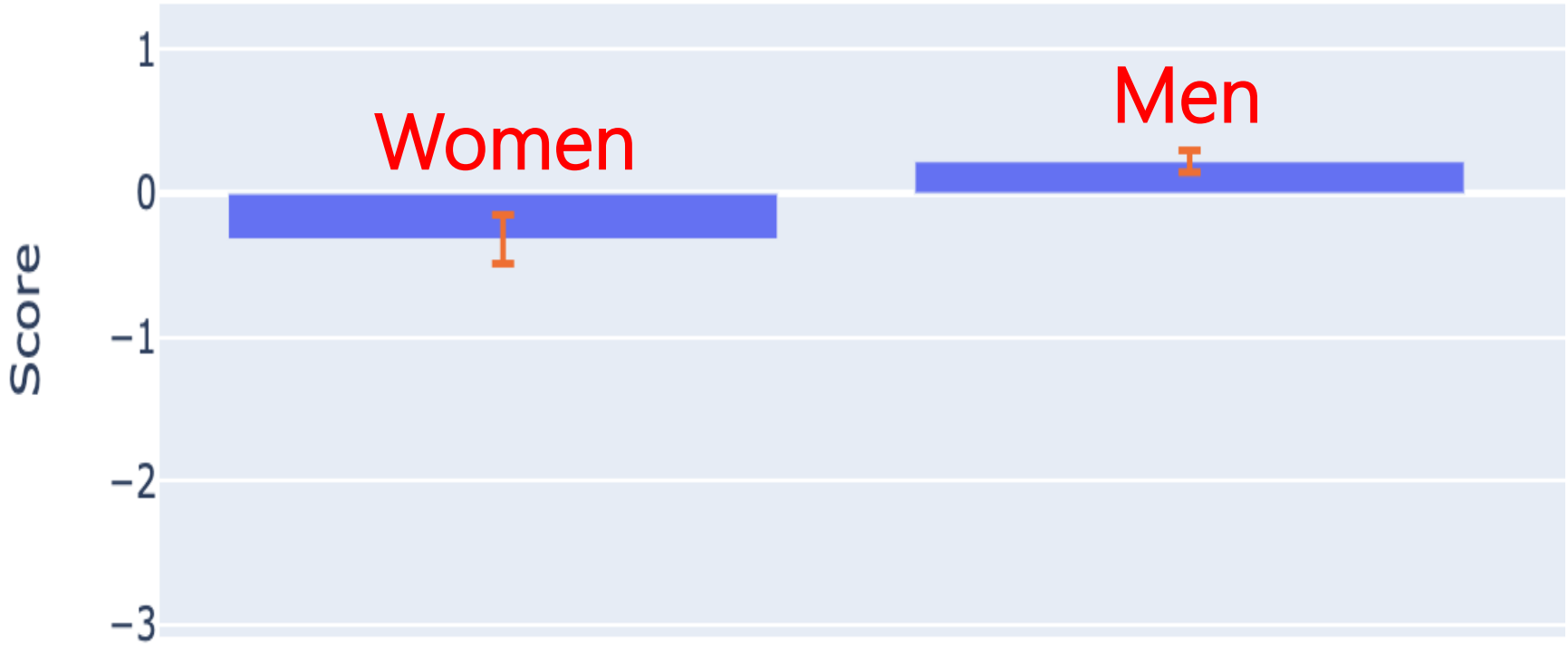
Example 6: COVID-19 Mortality

COVID-19 Mortality Risk vs. Age



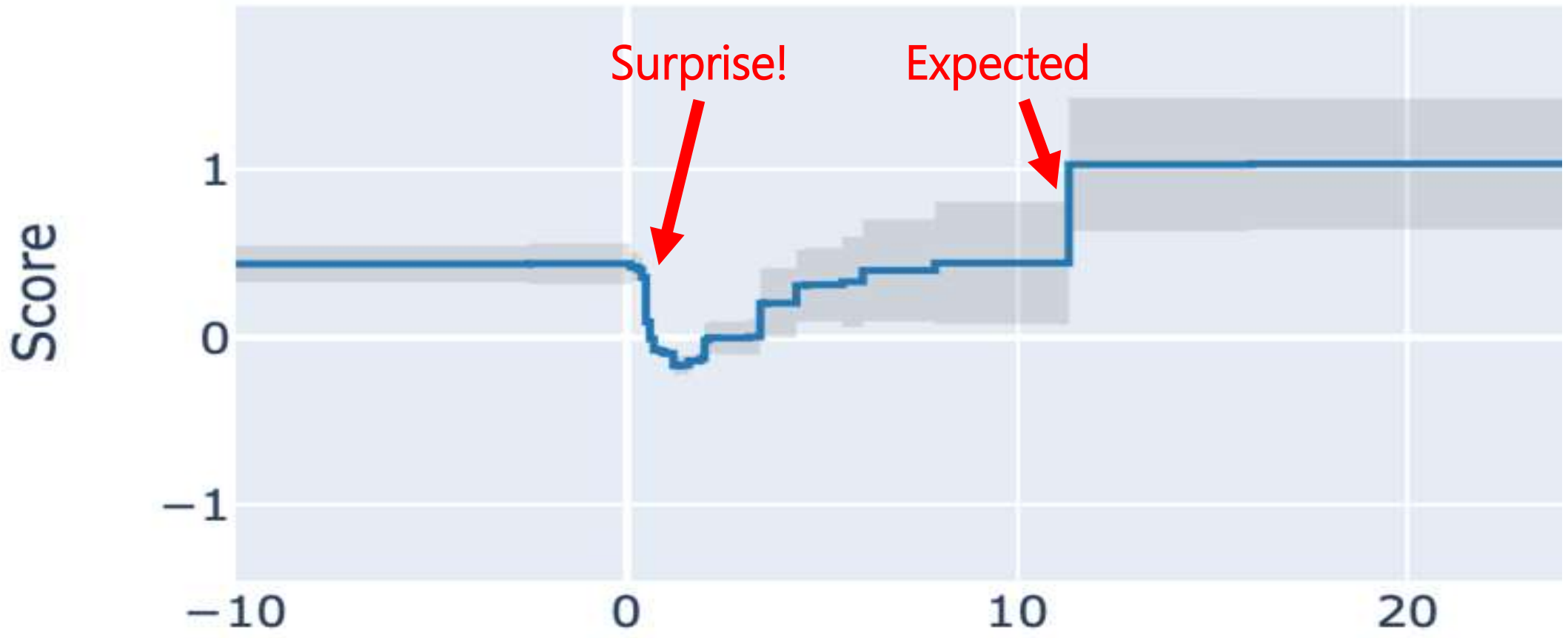
Ben Lengerich (CMU/MIT), Rich Caruana (Microsoft), Aphinyanaphongs Yindalon (NYU)

COVID-19 Mortality Risk vs Gender



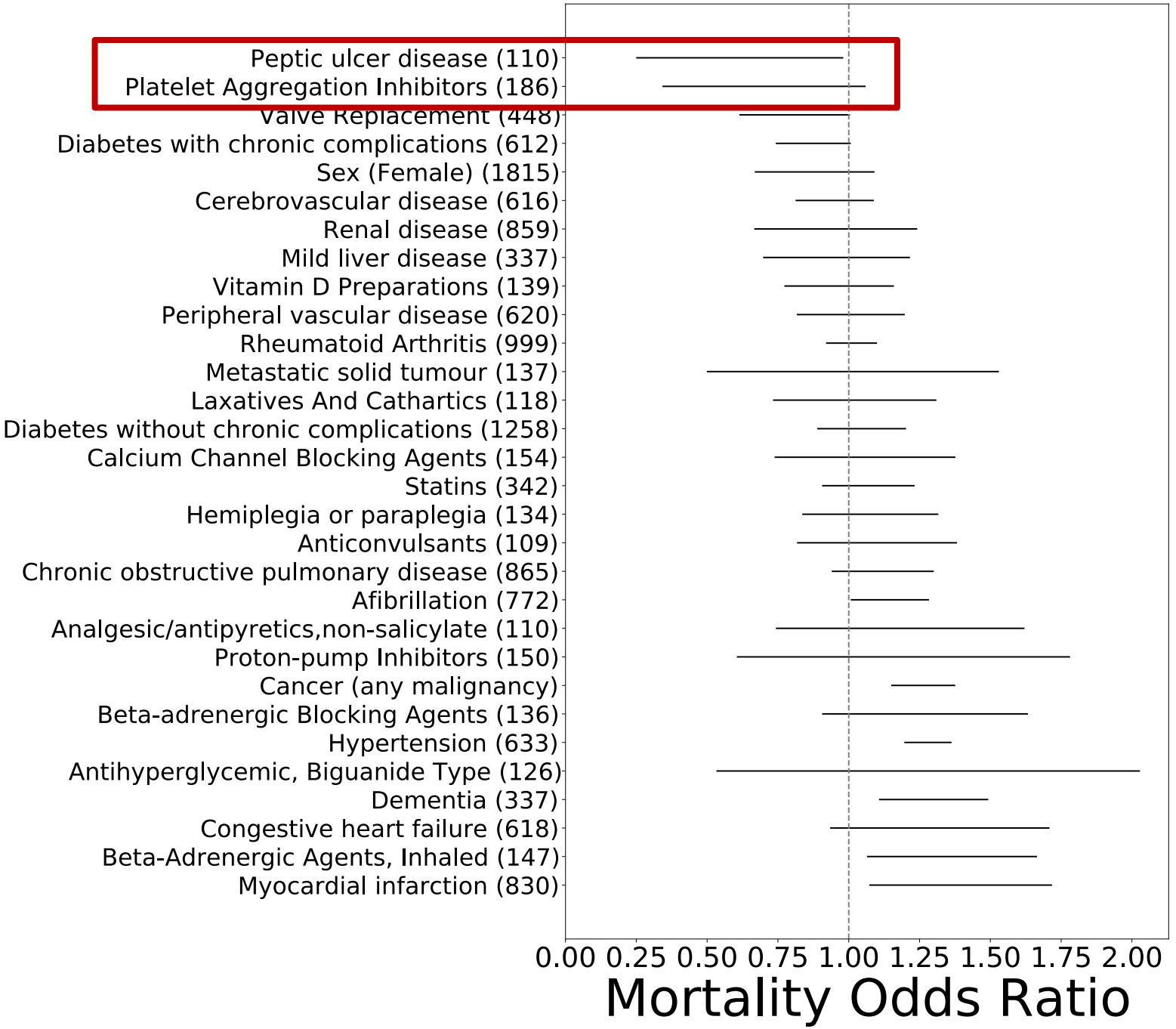
Ben Lengerich (CMU/MIT), Rich Caruana (Microsoft), Aphinyanaphongs Yindalon (NYU)

Surprising Discovery: Lymphocytes_Absolute_a

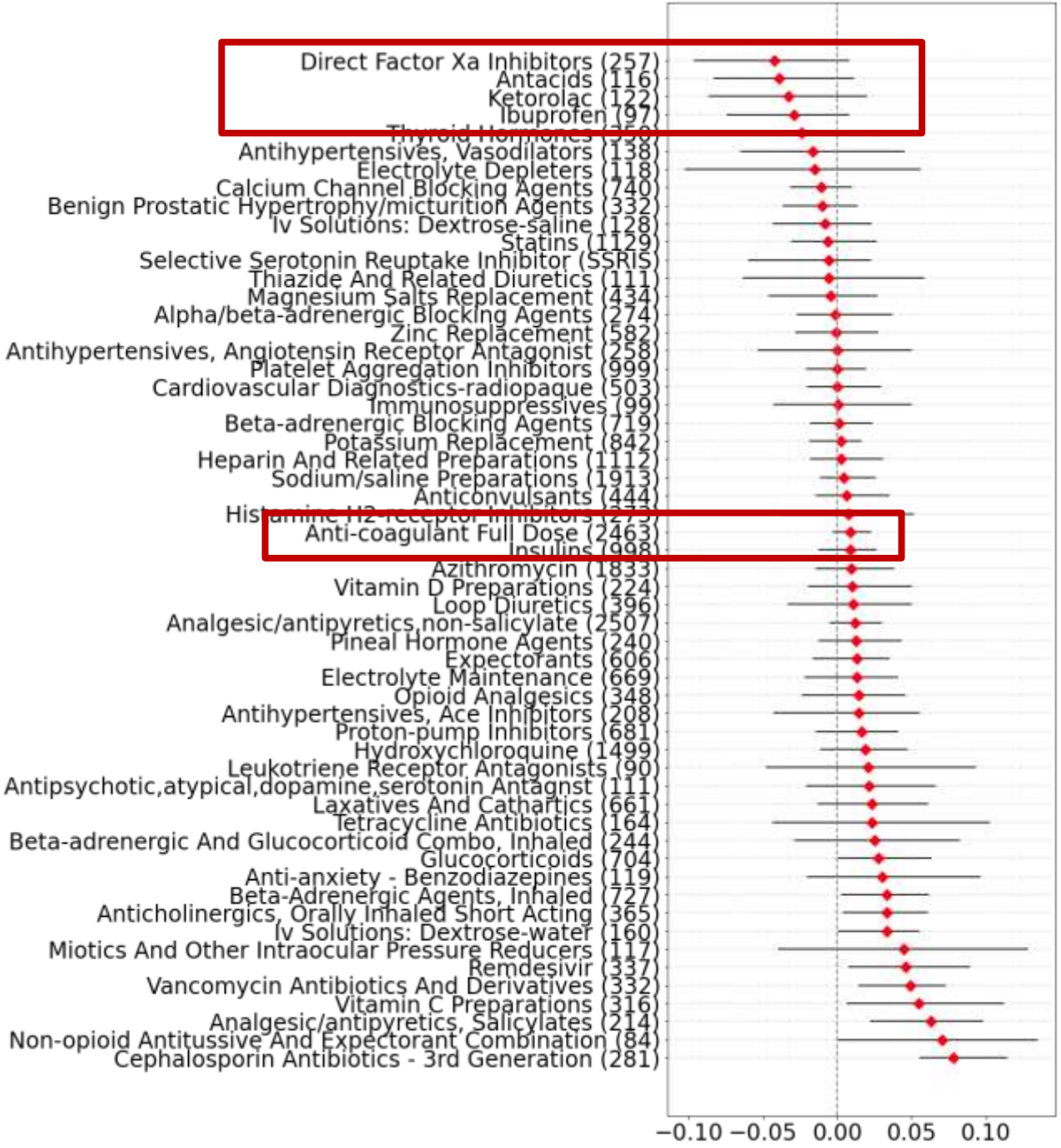


Joint work with Aphinyanaphongs Yindalon (NYU)

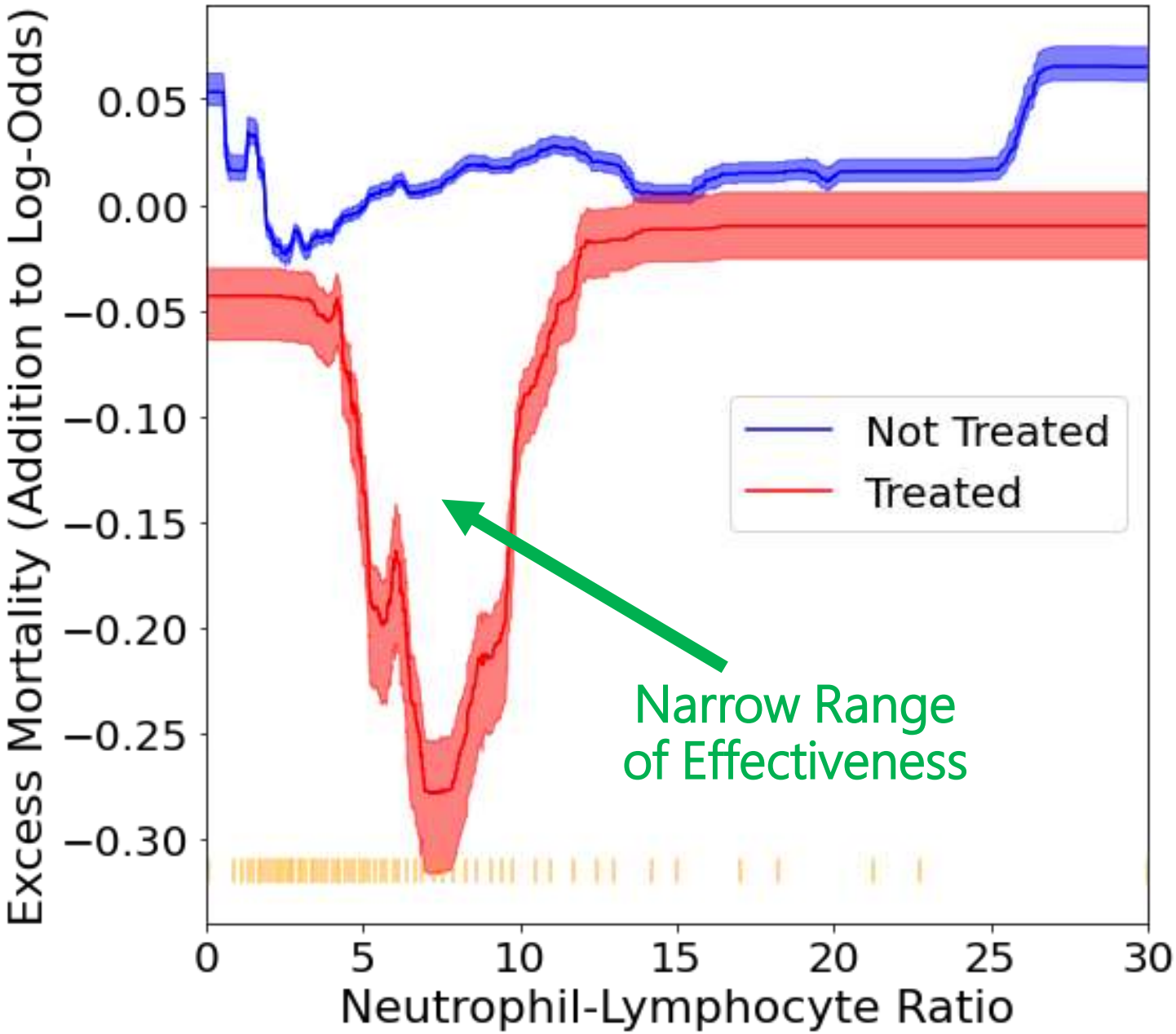
Mortality Risk from Comorbidities and Out-patient Meds



Mortality Risk from In-patient Meds



Glucocorticoid Steroids



Joint work with Ben Lengerich (CMU) & Aphinyanaphongs Yindalon (NYU)

Example 7: Hospital 30-Day Readmission

Ways of Leveraging Transparent EBM Models

- Understand **GLOBAL MODEL**, find and fix problems before deployment
- **EXPLAIN PREDICTIONS** by sorting features that contribute most to prediction
- **“OPEN-UP"** other black-box models to see what's inside them

Ways of Leveraging Transparent EBM Models

- Understand **GLOBAL MODEL**, find and fix problems before deployment
- **EXPLAIN PREDICTIONS** by sorting features that contribute most to prediction
- **“OPEN-UP"** other black-box models to see what's inside them

Readmission Demo

Example 8: Recidivism Prediction

FAT*/ML: ProPublica COMPAS Recidivism Data

- COMPAS is a black-box model used to predict future criminal behavior
  - Model is black-box because it is protected by IP, not because it is a deep net
  - Criminal justice officials use risk prediction to inform bail, sentencing and parole decisions

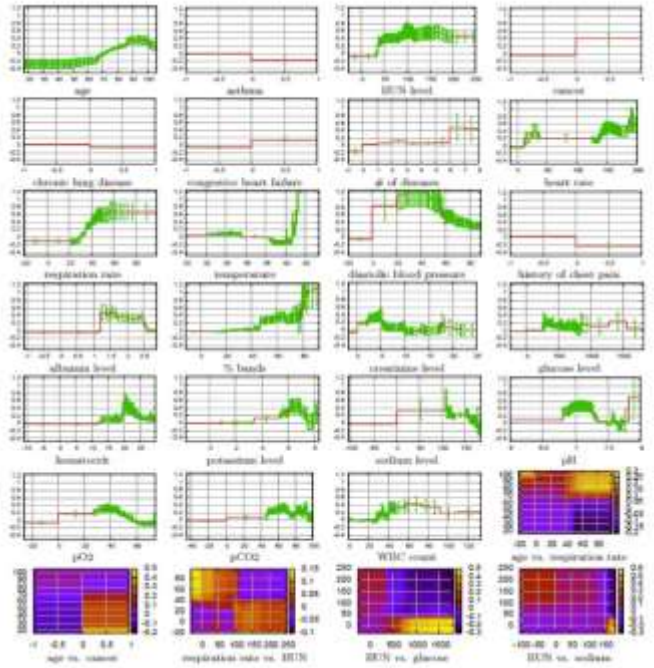
- Is COMPAS model biased?
- Is COMPAS model accurate?
- Is COMPAS model complex?

Distillation Trick to Open Up Black-Box Models

Soft labels from Teacher

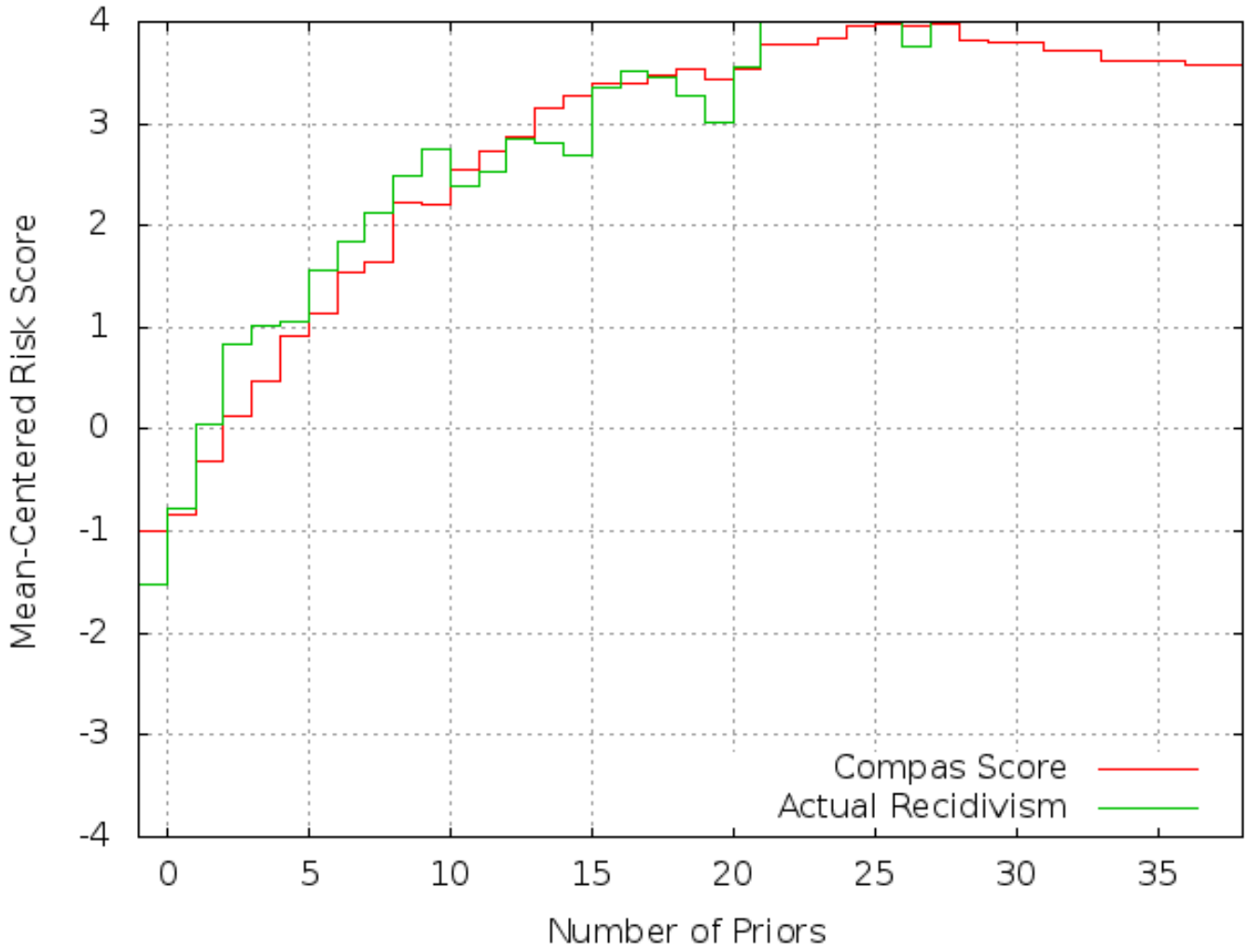
Black-Box Compas Model

Black-Box Teacher Model

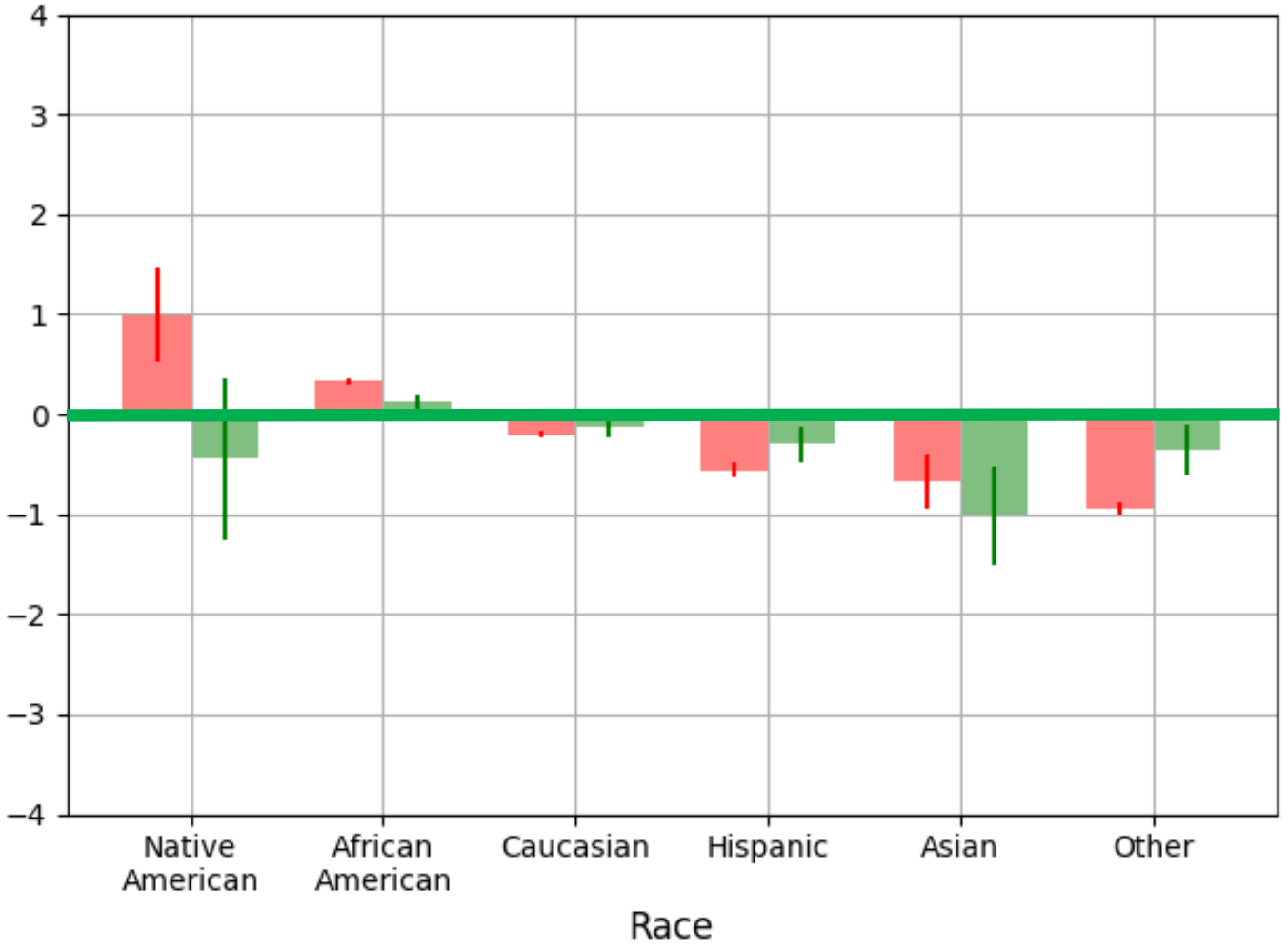


Glass-Box EBM Student Model

Recidivism Risk vs. Number of Prior Convictions



Recidivism Risk vs. Race



InterpretML

Open-Source EcoSystem for Intelligibility

github.com/interpretml/interpret 

Algorithm Sketch for EBMs (Explainable Boosting Trees)

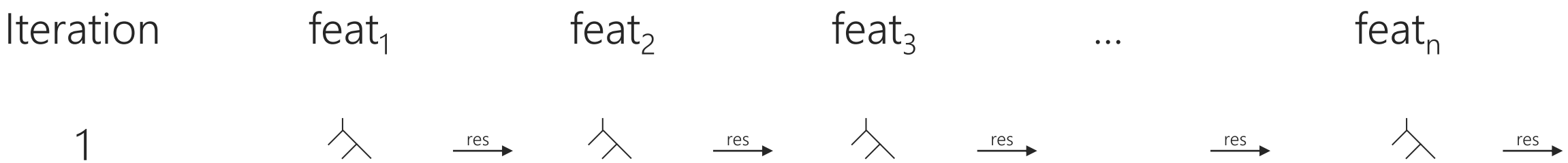
Iteration	$feat_1$	$feat_2$	$feat_3$	...	$feat_n$
1					

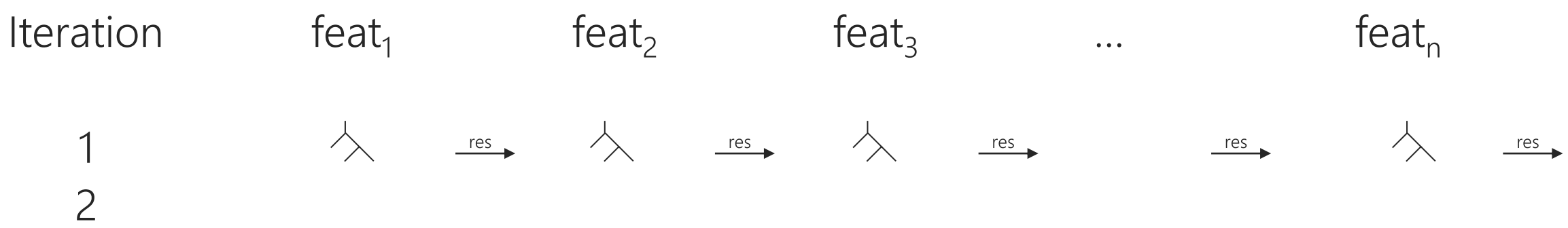
Iteration	$feat_1$	$feat_2$	$feat_3$	...	$feat_n$
1					

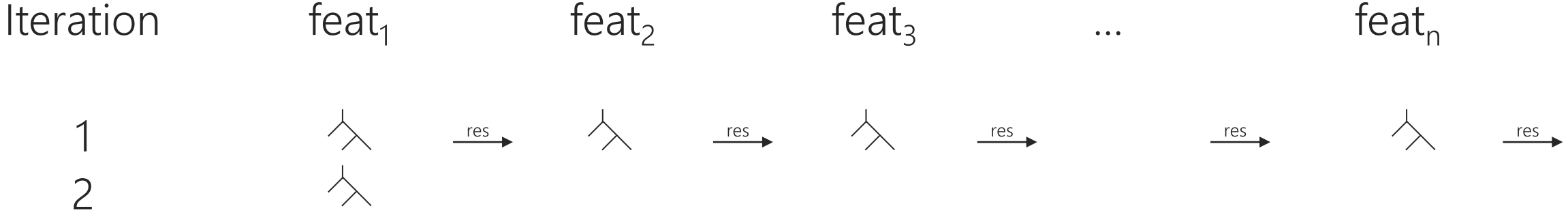
Iteration	$feat_1$	$feat_2$	$feat_3$	...	$feat_n$
1					

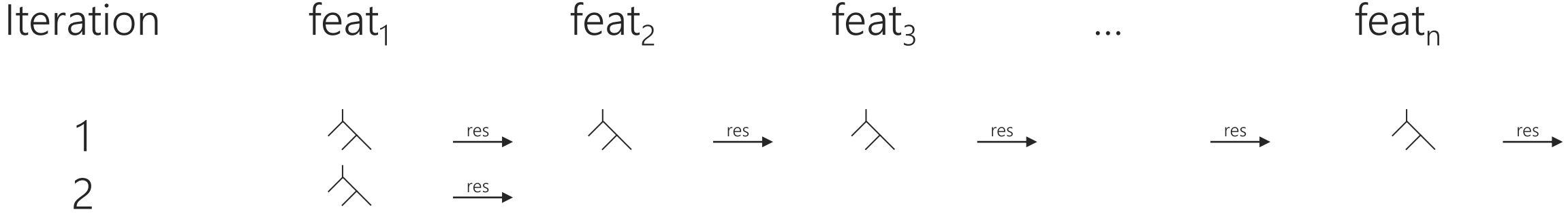
Iteration	$feat_1$		$feat_2$	$feat_3$	...	$feat_n$
1		res				

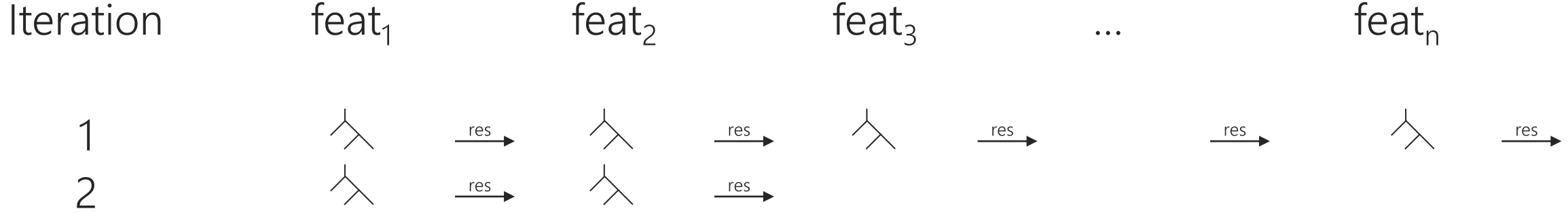
Iteration	$feat_1$	$feat_2$	$feat_3$	...	$feat_n$
1					

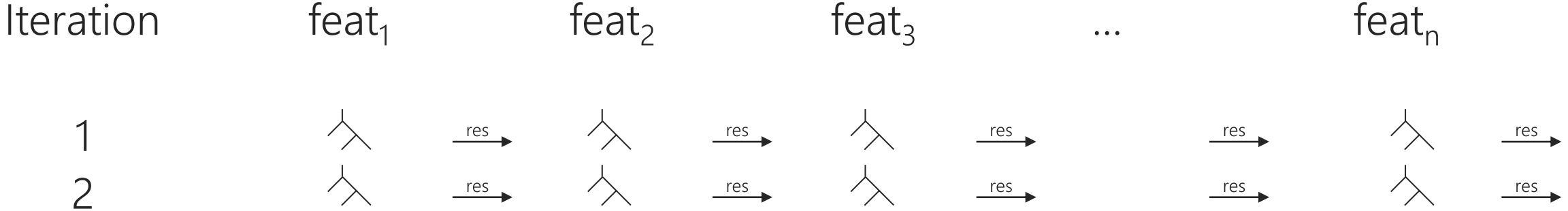

Iteration
feat1
feat2
feat3
...
featn
1
res
res
res
res
res

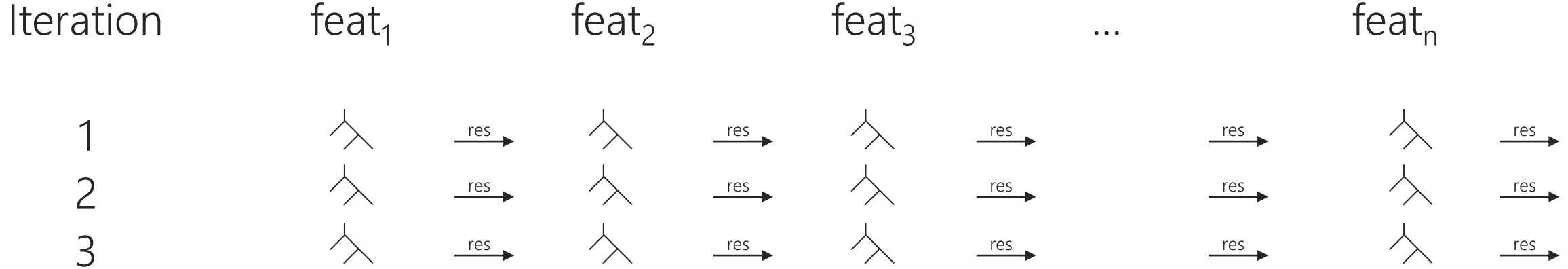

Iteration
feat1
feat2
feat3
...
featn
1
2
res
res
res
res
res

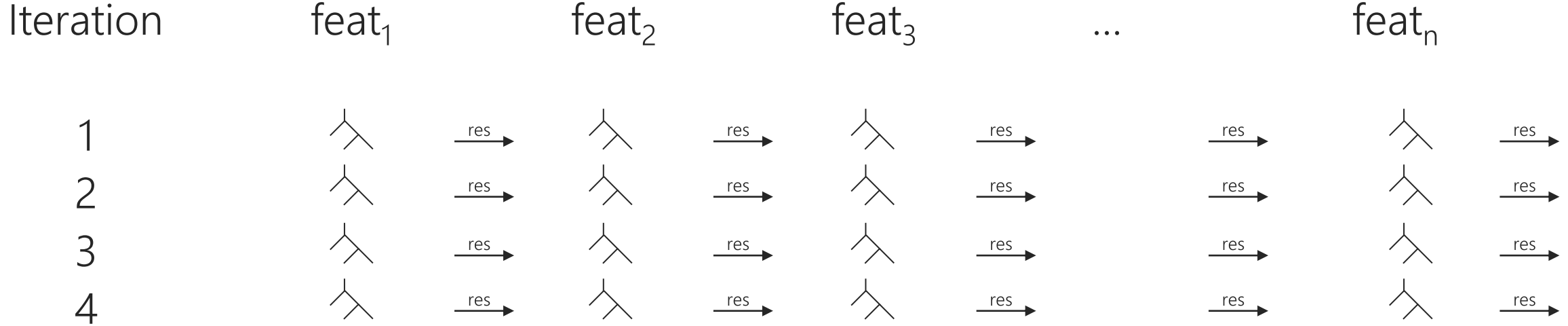

Iteration
feat1
feat2
feat3
...
featn
1
2
res
res
res
res
res

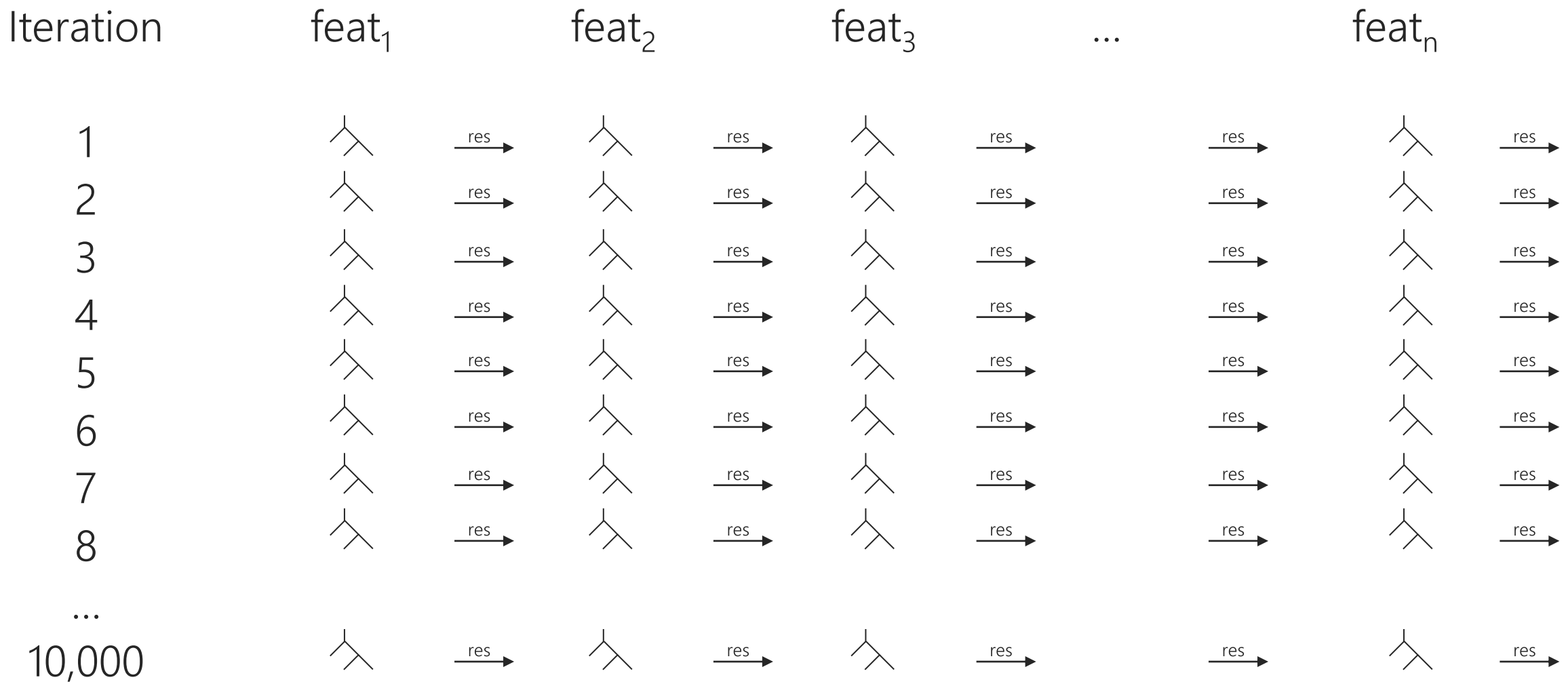

Iteration
feat1
feat2
feat3
...
featn
1
2
res
res
res
res
res
res
res

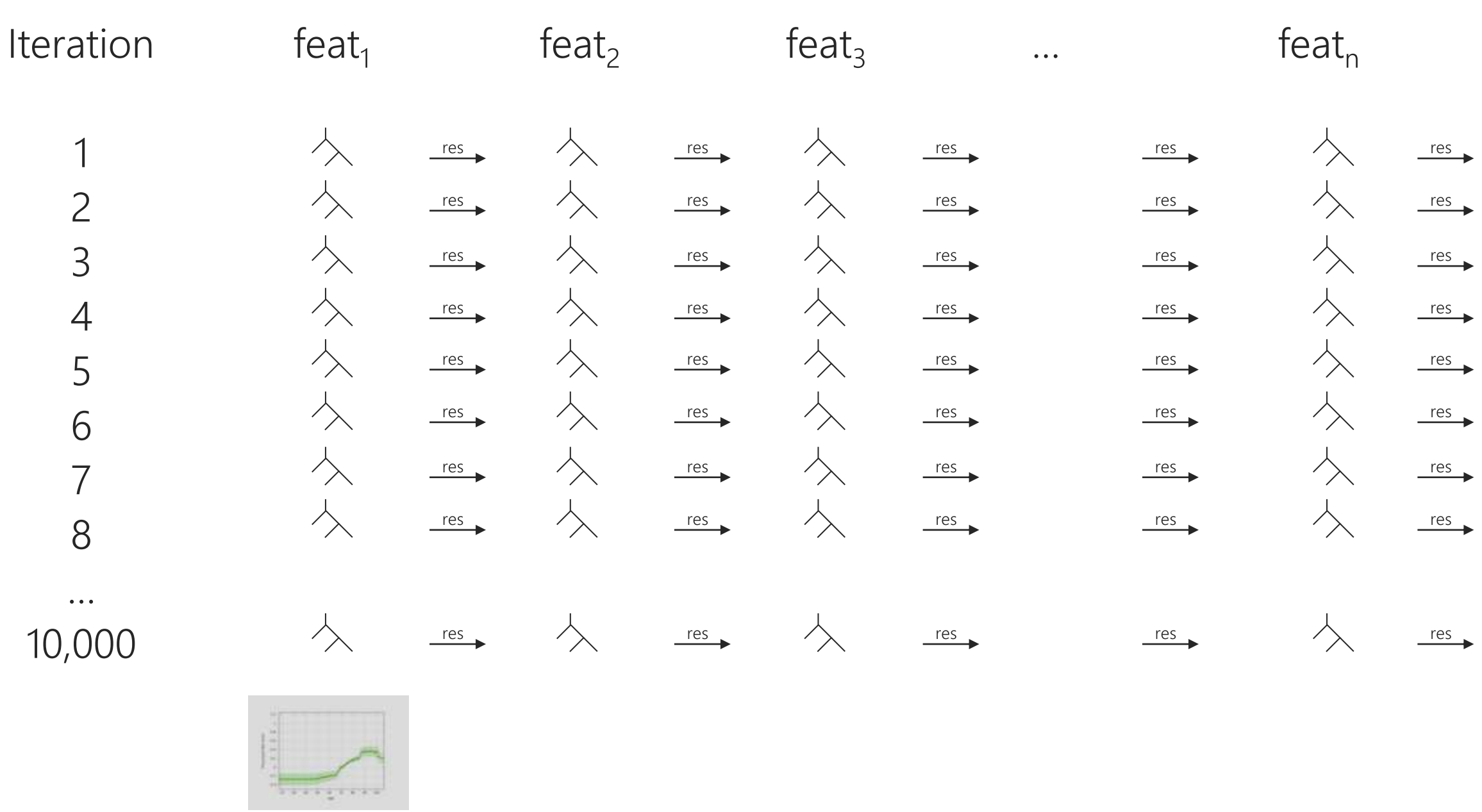

Iteration
feat1
feat2
feat3
...
featn
1
2
res
res
res
res
res
res
res

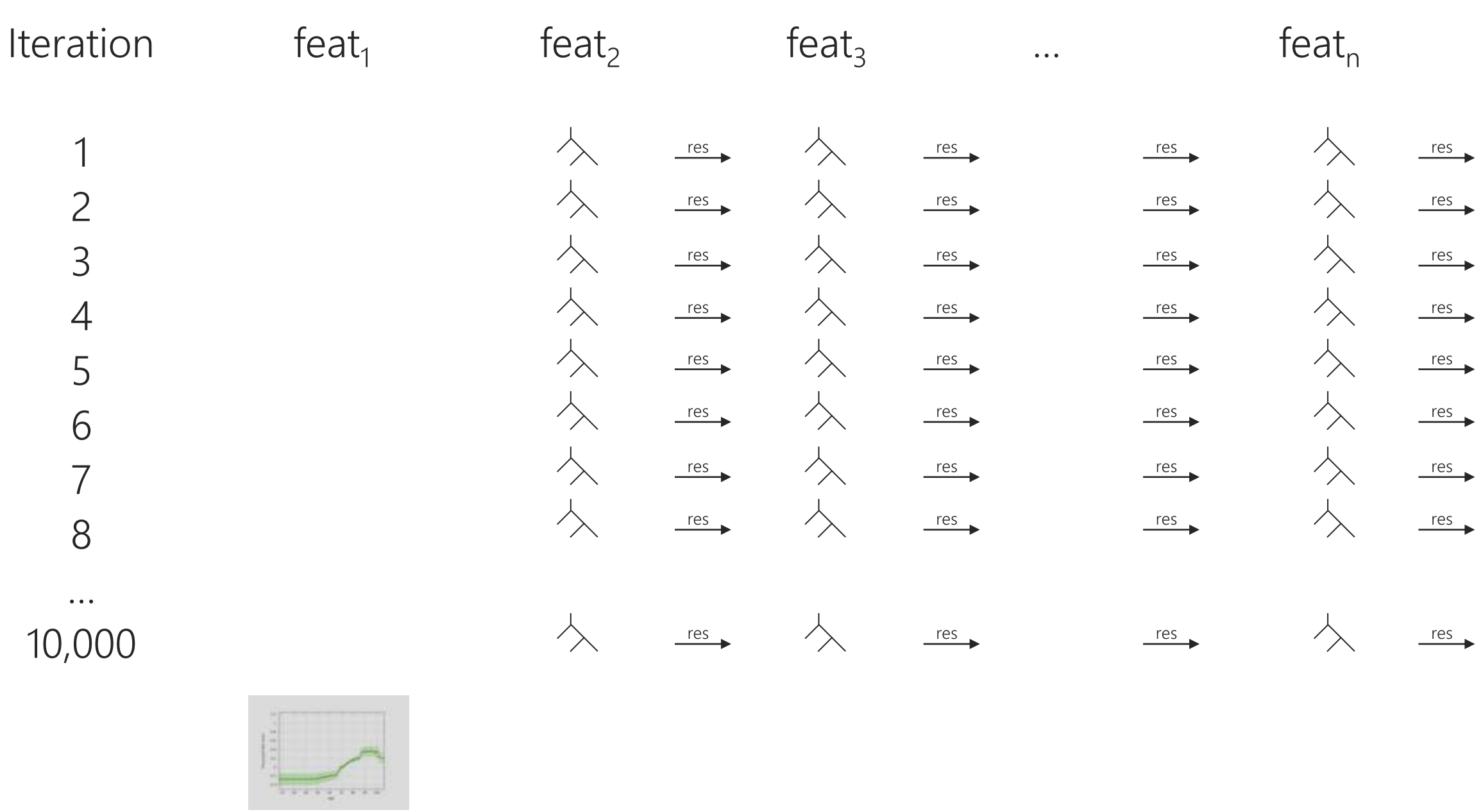

Iteration
feat1
feat2
feat3
...
featn
1
2
res

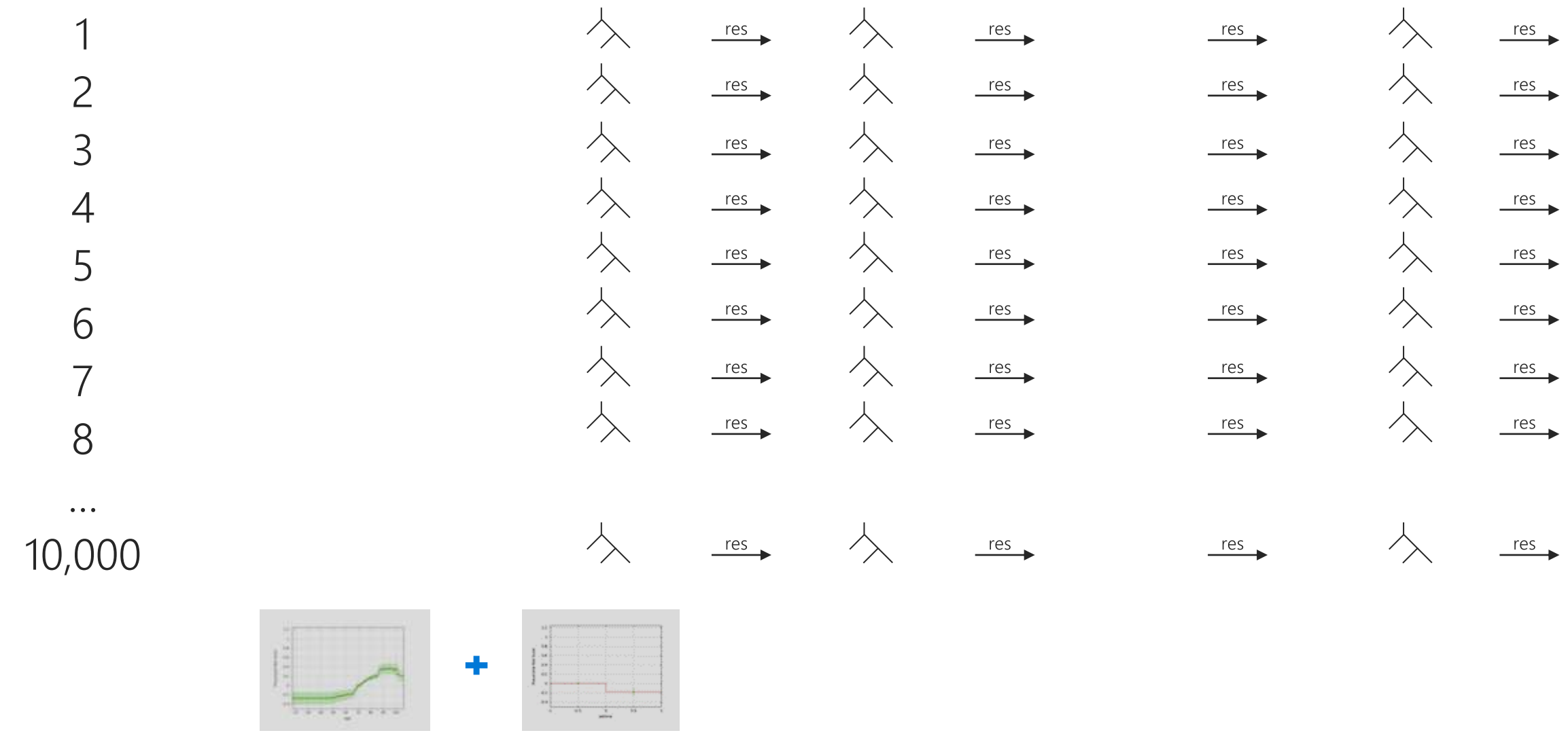

Iteration
feat1
feat2
feat3
...
featn
1
2
3
res


Iteration
feat1
feat2
feat3
...
featn
1
2
3
4
res


Iteration
feat1
feat2
feat3
...
featn
1
2
3
4
5
6
7
8
...
10,000
res

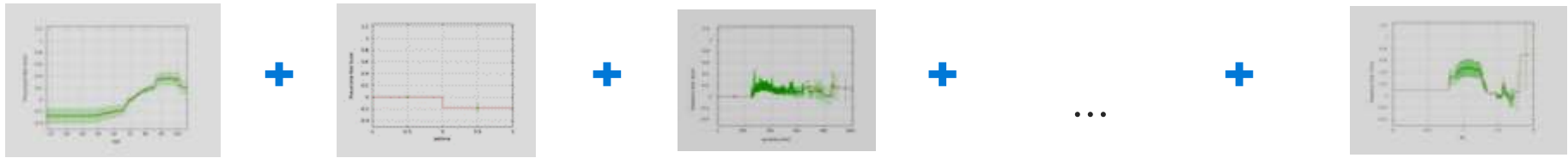

Iteration
feat1
feat2
feat3
...
featn
1
2
3
4
5
6
7
8
...
10,000
res


Iteration
feat1
feat2
feat3
...
featn
1
2
3
4
5
6
7
8
...
10,000
res

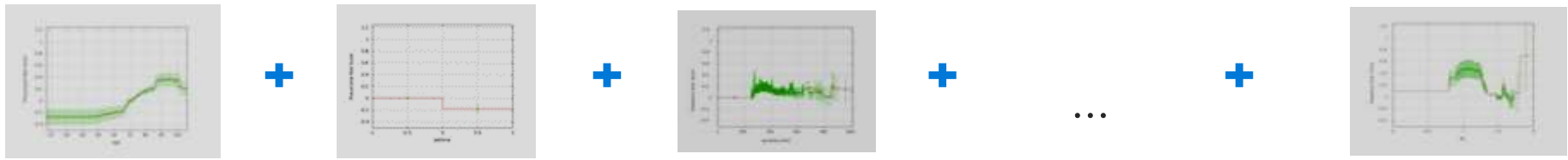

Iteration
feat1
feat2
feat3
...
featn
1
2
3
4
5
6
7
8
...
10,000
res

Iteration	$feat_1$	$feat_2$	$feat_3$	...	$feat_n$
1			res	res	res
2			res	res	res
3			res	res	res
4			res	res	res
5			res	res	res
6			res	res	res
7			res	res	res
8			res	res	res
...					
10,000			res	res	res

Iteration	$feat_1$	$feat_2$	$feat_3$	...	$feat_n$
1			res	res	res
2			res	res	res
3			res	res	res
4			res	res	res
5			res	res	res
6			res	res	res
7			res	res	res
8			res	res	res
...					
10,000			res	res	res



$feat_1$ $feat_2$ $feat_3$ ... $feat_n$



How to Fit Pairwise Interactions ?

- FIT MAINS:
  - Fit main effects first
  - Freeze the main effects
  - Compute residual of main effects to original targets

- FIT PAIRS:
  - There are $O(N^2)$ possible pairs --- don't want to add that many terms to model
  - Use algorithm called FAST to heuristically sort $O(N^2)$ pairs by match to residual
  - User selects number of pairs to add to model
  - Run same round-robin boosting algorithm to fit K pairs

- Final Model = N Mains + K Pairs

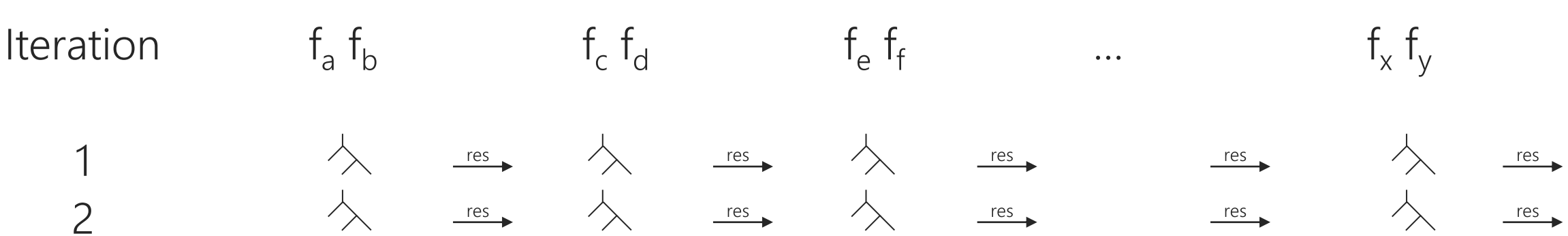
	$Pair_1$	$Pair_2$	$Pair_3$	...	$Pair_n$
Iteration	f_a f_b	f_c f_d	f_e f_f	...	f_x f_y
1					

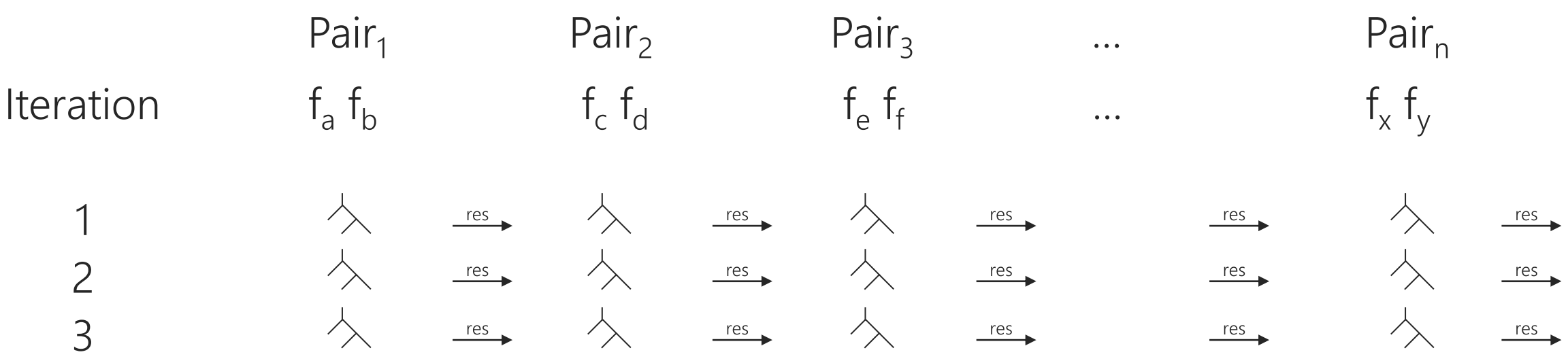
	$Pair_1$	$Pair_2$	$Pair_3$	...	$Pair_n$
Iteration	f_a f_b	f_c f_d	f_e f_f	...	f_x f_y
1					

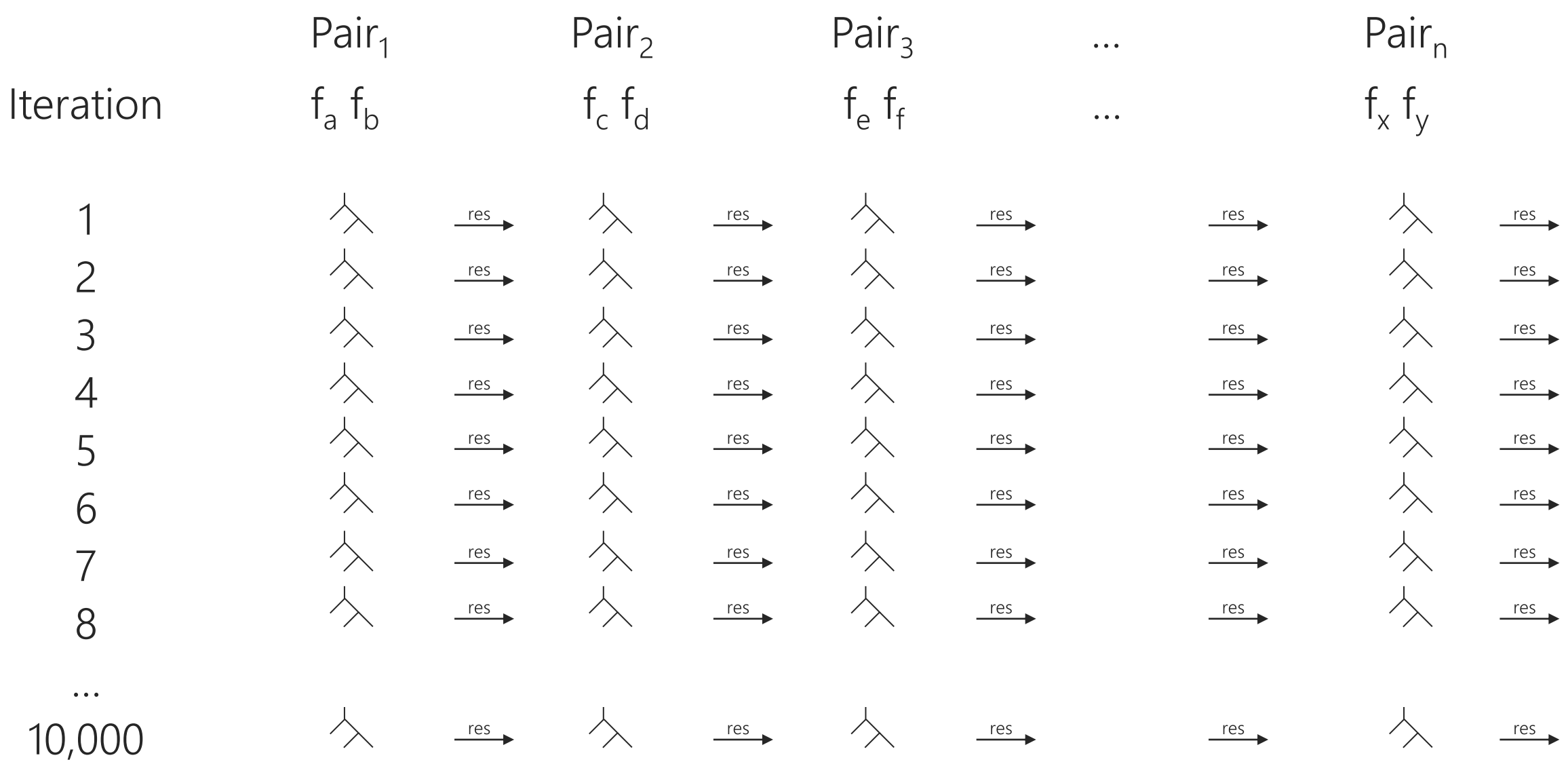
	$Pair_1$	$Pair_2$	$Pair_3$	...	$Pair_n$
Iteration	f_a f_b	f_c f_d	f_e f_f	...	f_x f_y
1					

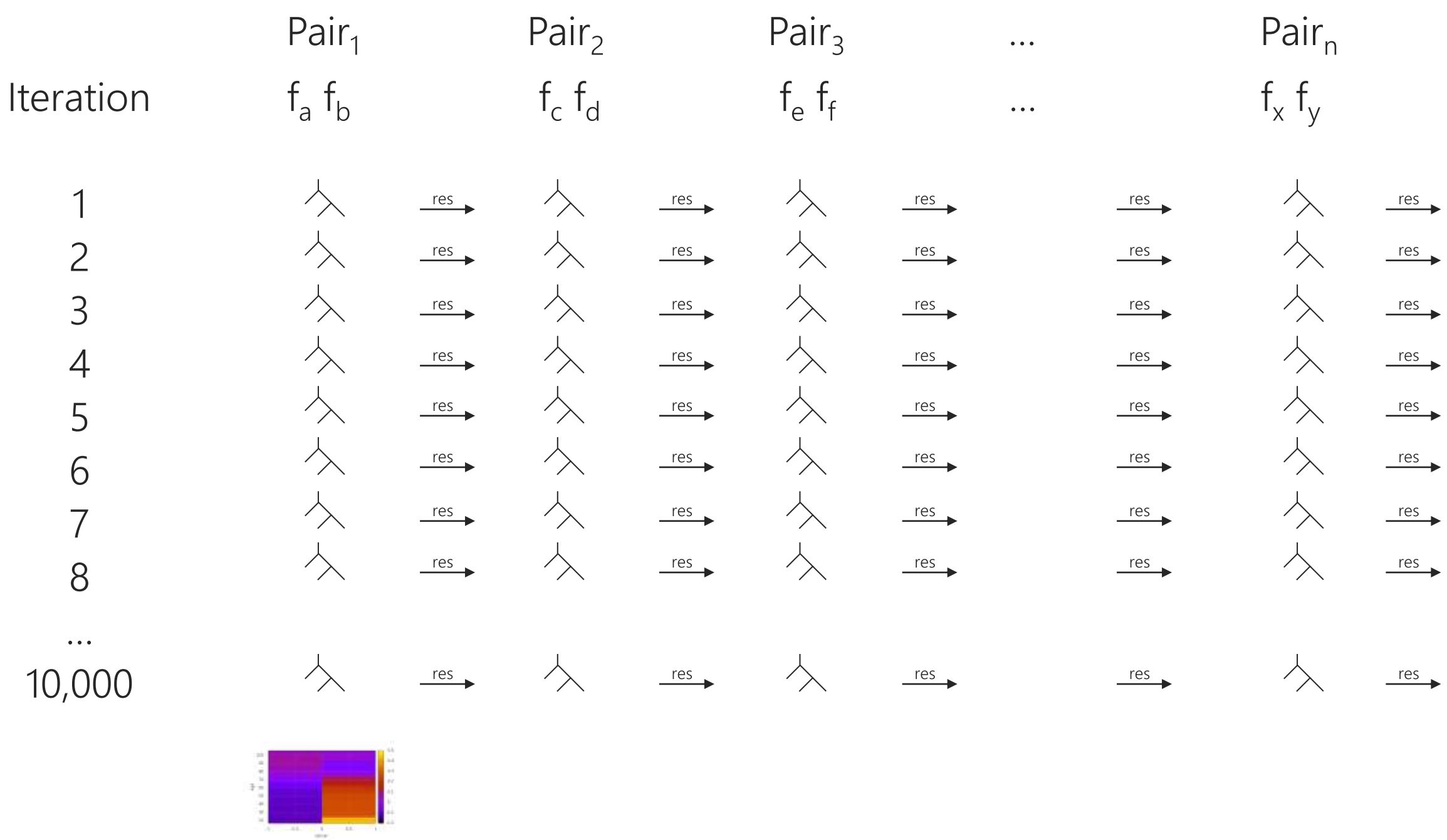
	$Pair_1$	$Pair_2$	$Pair_3$	...	$Pair_n$
Iteration	f_a f_b	f_c f_d	f_e f_f	...	f_x f_y
1					

	$Pair_1$	$Pair_2$	$Pair_3$	...	$Pair_n$
Iteration	f_a f_b	f_c f_d	f_e f_f	...	f_x f_y
1					

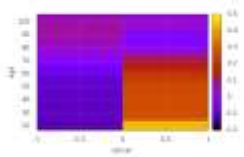

Pair1
Pair2
Pair3
...
Pairn
Iteration
fa fb
fc fd
fe ff
...
fx fy
1
2
res


Pair1
Pair2
Pair3
...
Pairn
Iteration
fa fb
fc fd
fe ff
...
fx fy
1
2
3
res


Pair1
Pair2
Pair3
...
Pairn
Iteration
fa fb
fc fd
fe ff
...
fx fy
1
2
3
4
5
6
7
8
...
10,000
res

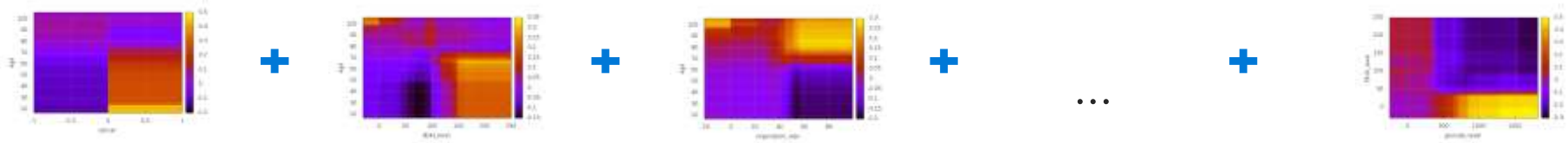

Pair1
Pair2
Pair3
...
Pairn
Iteration
fa fb
fc fd
fe ff
...
fx fy
1
2
3
4
5
6
7
8
...
10,000
res

Iteration	$Pair_1$	$Pair_2$	$Pair_3$	...	$Pair_n$
	f_a f_b	f_c f_d	f_e f_f	...	f_x f_y
1					
2					
3					
4					
5					
6					
7					
8					
...					
10,000					

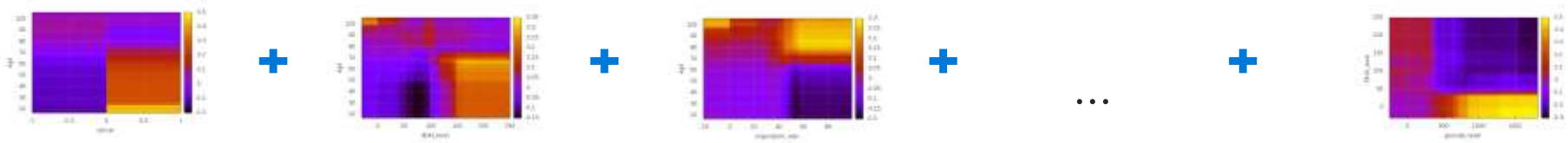


Iteration	$Pair_1$	$Pair_2$	$Pair_3$	...	$Pair_n$
	f_a f_b	f_c f_d	f_e f_f	...	f_x f_y
1			res	res	res
2			res	res	res
3			res	res	res
4			res	res	res
5			res	res	res
6			res	res	res
7			res	res	res
8			res	res	res
...					
10,000			res	res	res

Iteration	$Pair_1$	$Pair_2$	$Pair_3$	...	$Pair_n$
	f_a f_b	f_c f_d	f_e f_f	...	f_x f_y
1			res	res	res
2			res	res	res
3			res	res	res
4			res	res	res
5			res	res	res
6			res	res	res
7			res	res	res
8			res	res	res
...					
10,000			res	res	res

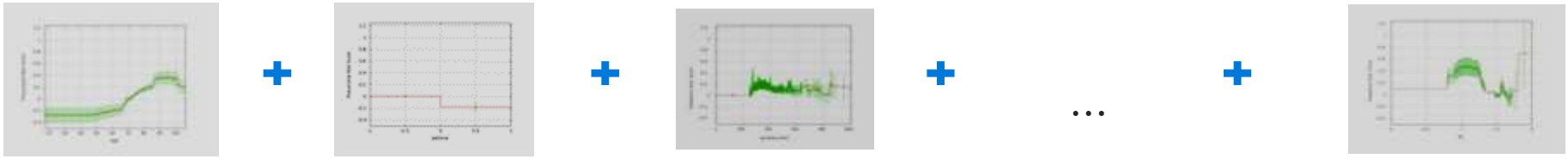


$Pair_1$	$Pair_2$	$Pair_3$	...	$Pair_n$
f_a f_b	f_c f_d	f_e f_f	...	f_x f_y



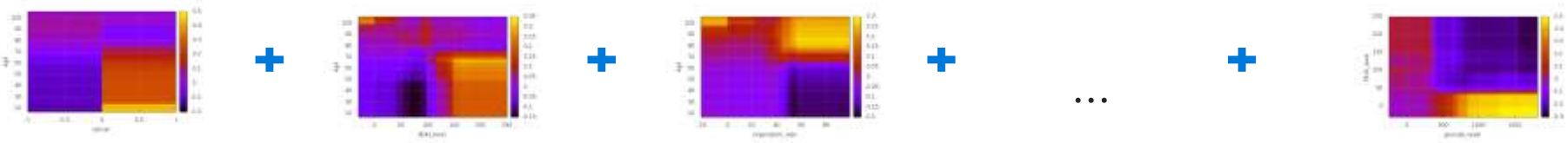
Final Model: Mains + Select Pairwise Interactions

$Main_1$ $feat_1$ + $Main_2$ $feat_2$ + $Main_3$ $feat_3$ + ... + $Main_m$ $feat_m$



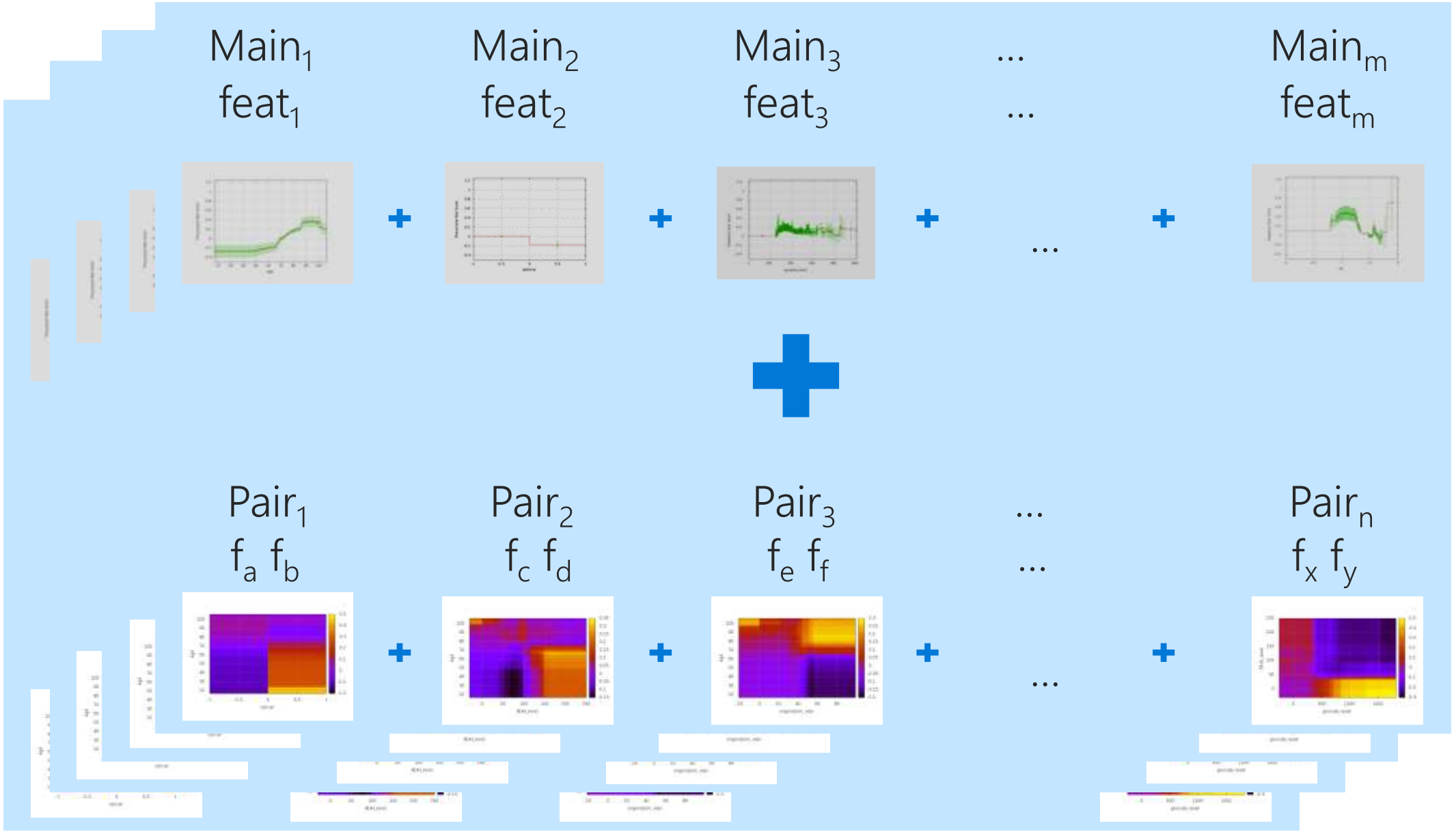
+

$Pair_1$ f_a f_b + $Pair_2$ f_c f_d + $Pair_3$ f_e f_f + ... + $Pair_n$ f_x f_y



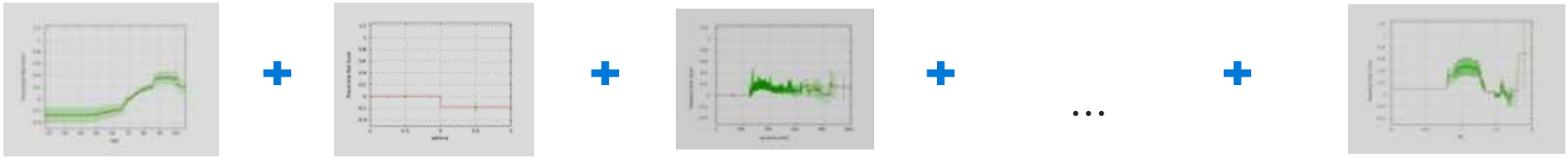
Final Model: Mains + Select Pairwise Interactions

Bagging 10X-100X



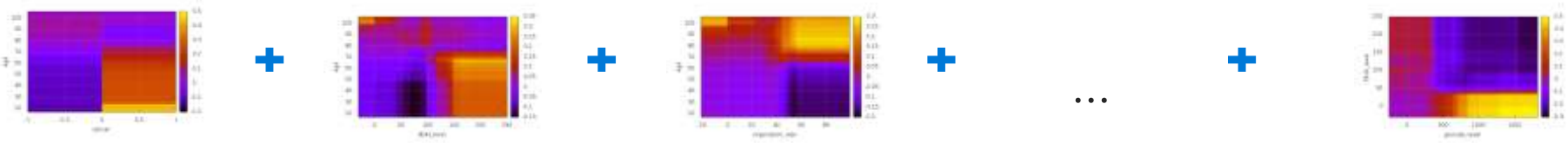
Final Model: Mains + Select Pairwise Interactions

$Main_1$ $feat_1$ + $Main_2$ $feat_2$ + $Main_3$ $feat_3$ + ... + $Main_m$ $feat_m$



+

$Pair_1$ f_a f_b + $Pair_2$ f_c f_d + $Pair_3$ f_e f_f + ... + $Pair_n$ f_x f_y



How Do We Fit GAMs with Neural Nets?

Why Bother to Fit GAMs with DNNs?

Limitations of EBMs

- EBMs have been state-of-the-art in glass-box learning for almost 10 years

- But…

- More than half of the ML community uses neural nets, not boosted trees, scikit-learn, …
- Algorithms based on boosted trees don't scale as easily as DNNs/CNNs trained on GPUs
- GAMs trained with boosted trees are not differentiable, which reduces flexibility
- Models trained with neural nets are more modular and flexible
- Hard to make some things like multitask learning work with boosted trees

Microsoft Research

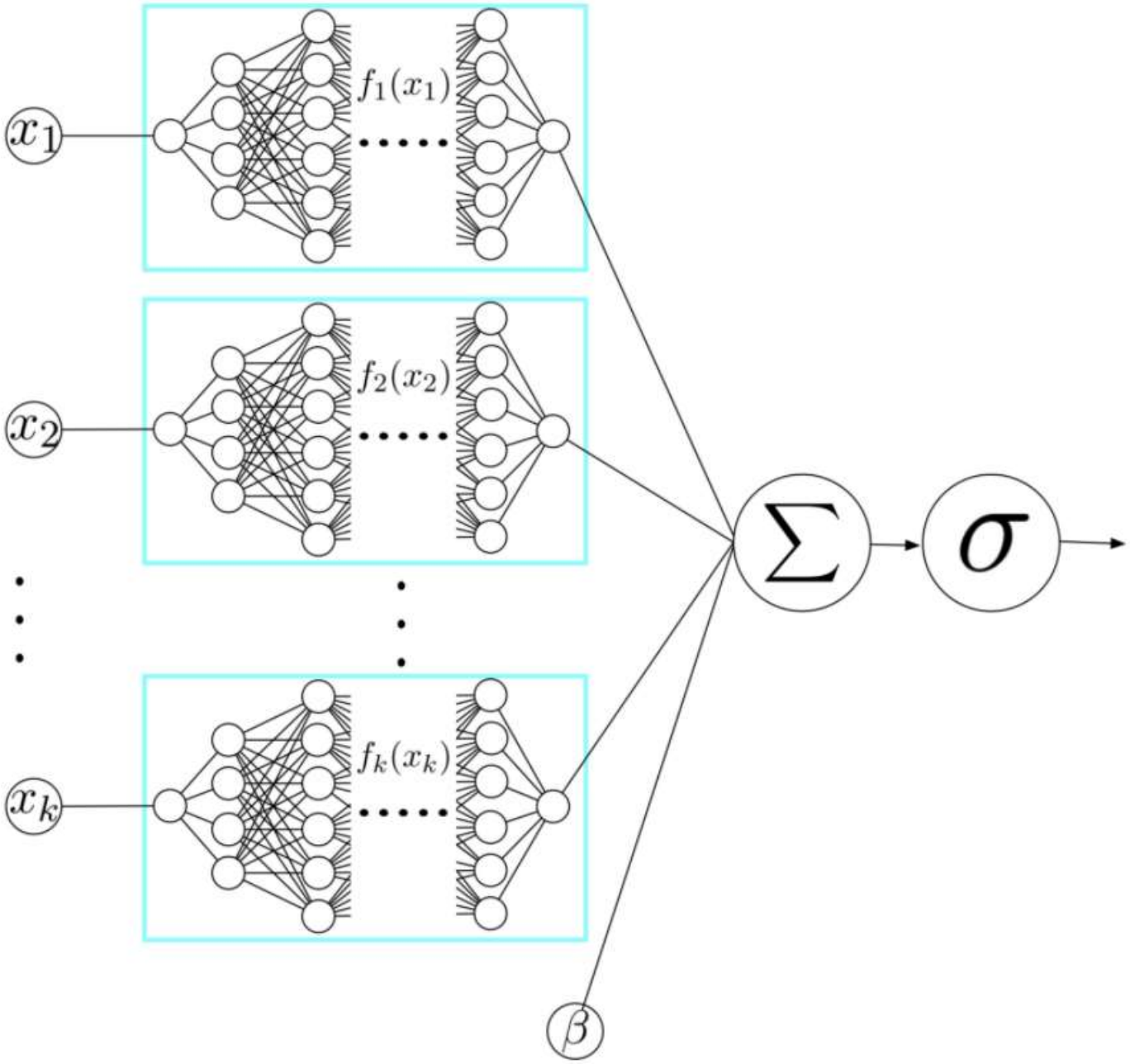
NeurIPS 2021
Spotlight
Paper

NAMs: Neural Additive Models
Interpretable Machine Learning With Neural Nets

Rishabh Agarwal, Levi Melnick, Ben Lengerich, Nicholas Frosst, Xuezhou Zhang, Rich Caruana, Geoffrey Hinton

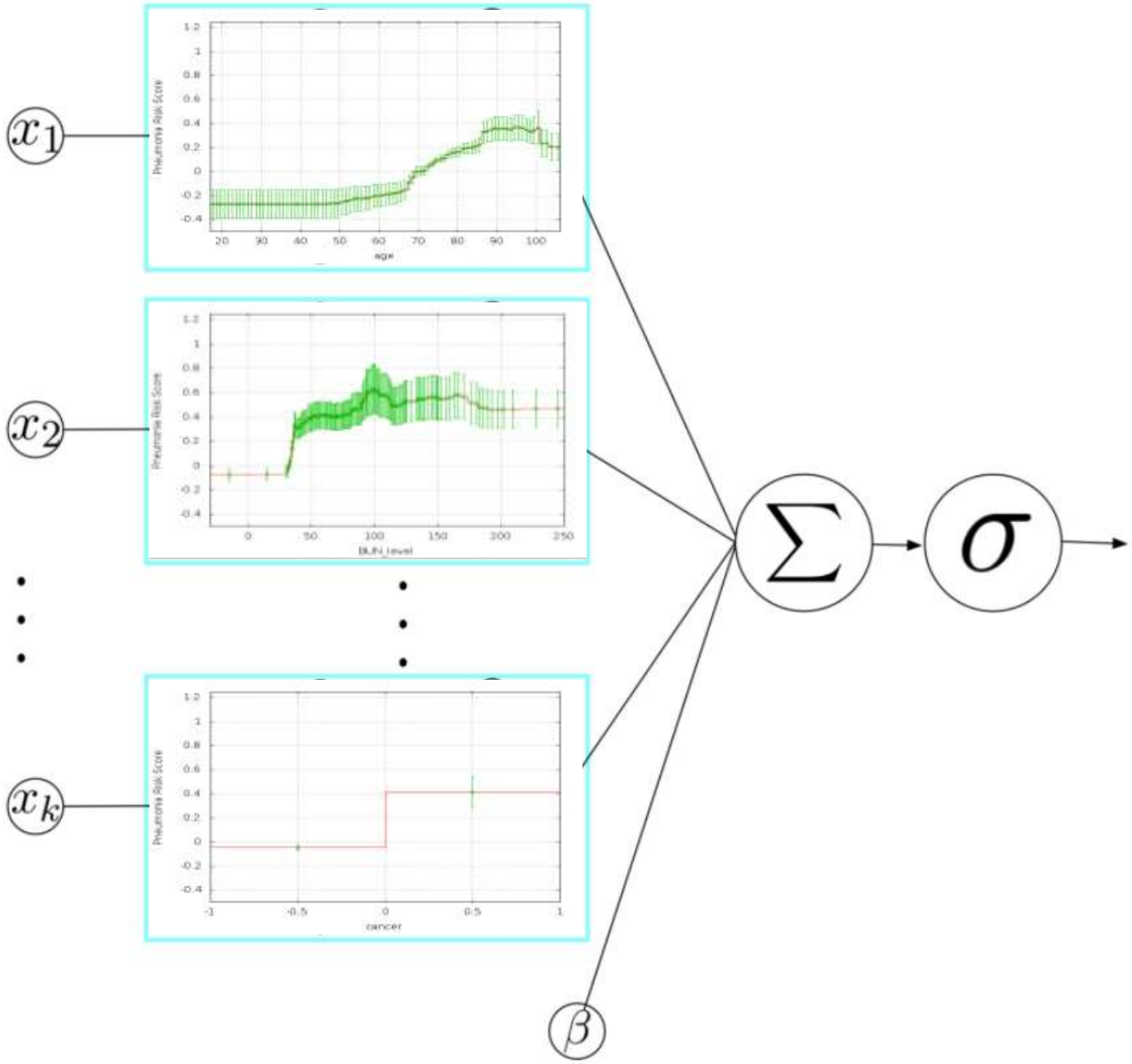
Microsoft

Deep Subnets



- Each feature feeds into a separate DNN subnet
- Subnets added at output layer
- Subnets learn separate additive models for each feature
- Sigmoid at output used for classification, not regression
- Subnets are learned in parallel
- Can be trained at massive scale on GPUs with standard software
- After training, subnets are replaced with graphs like EBMs

Deep Subnets → Feature Graphs

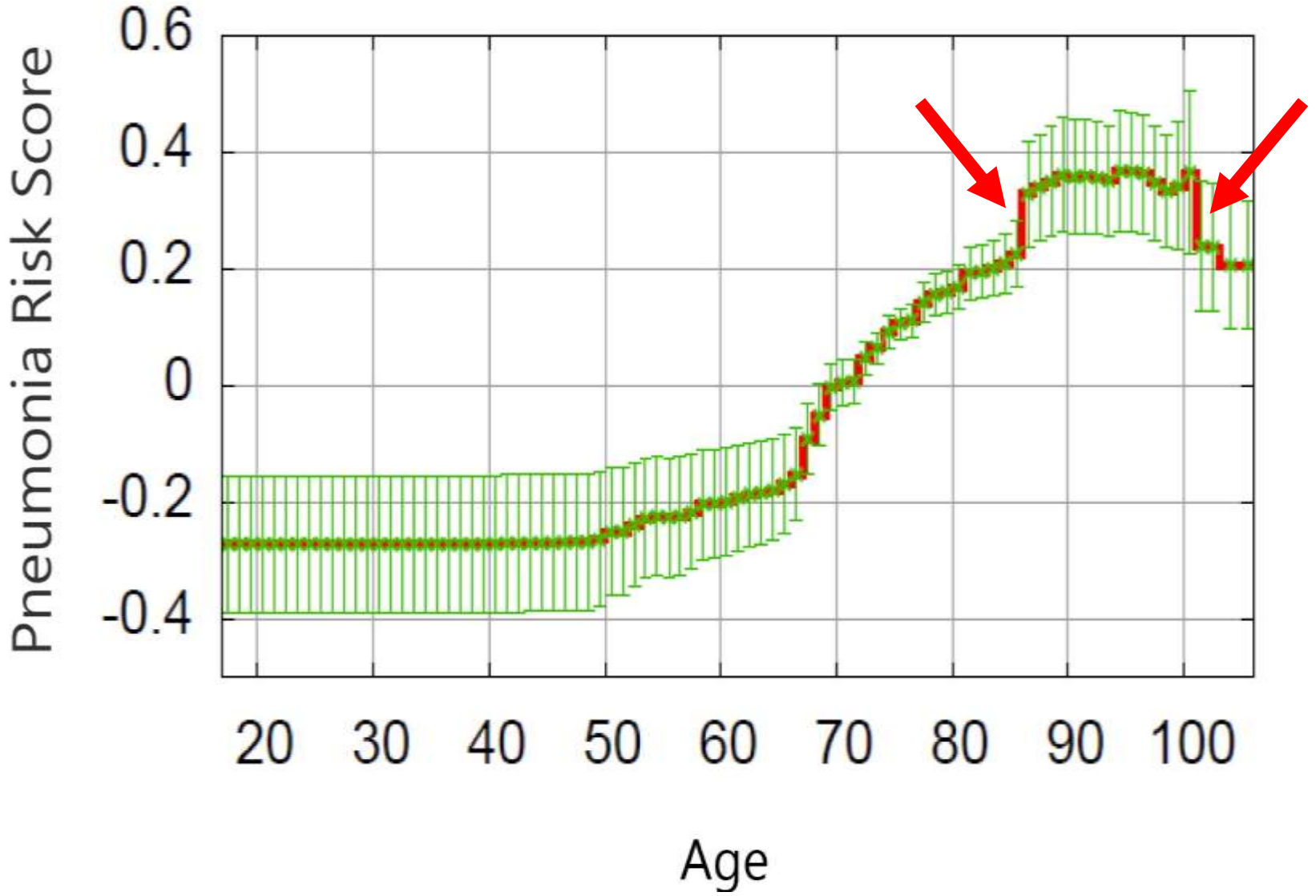


- Each feature feeds into a separate DNN subnet
- Subnets added at output layer
- Subnets learn separate additive models for each feature
- Sigmoid at output used for classification, not regression
- Subnets are learned in parallel
- Can be trained at massive scale on GPUs with standard software
- After training, subnets are replaced with feature graphs

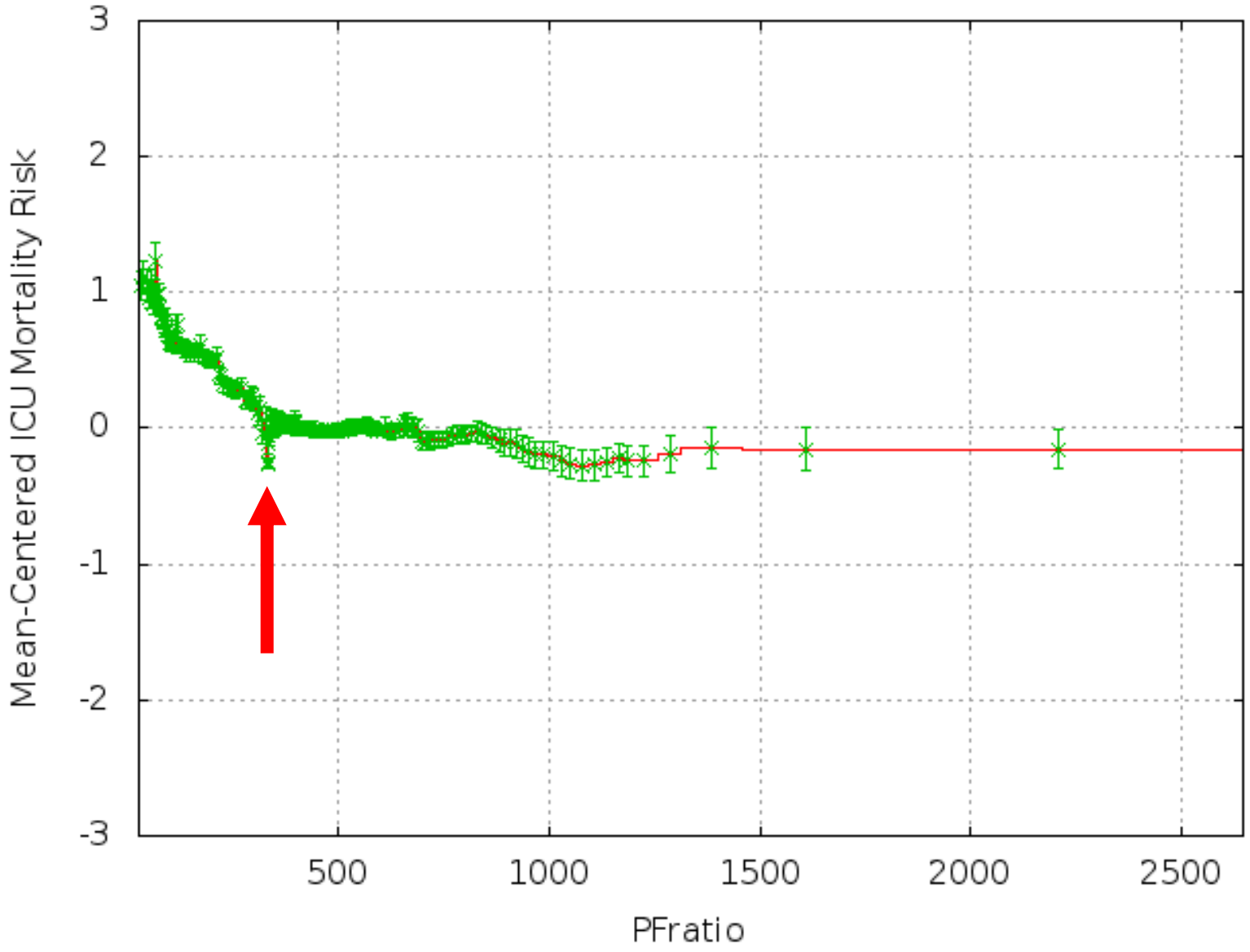
But there’s a problem…

Work with EBMs Show Jumps in Graphs Are Important

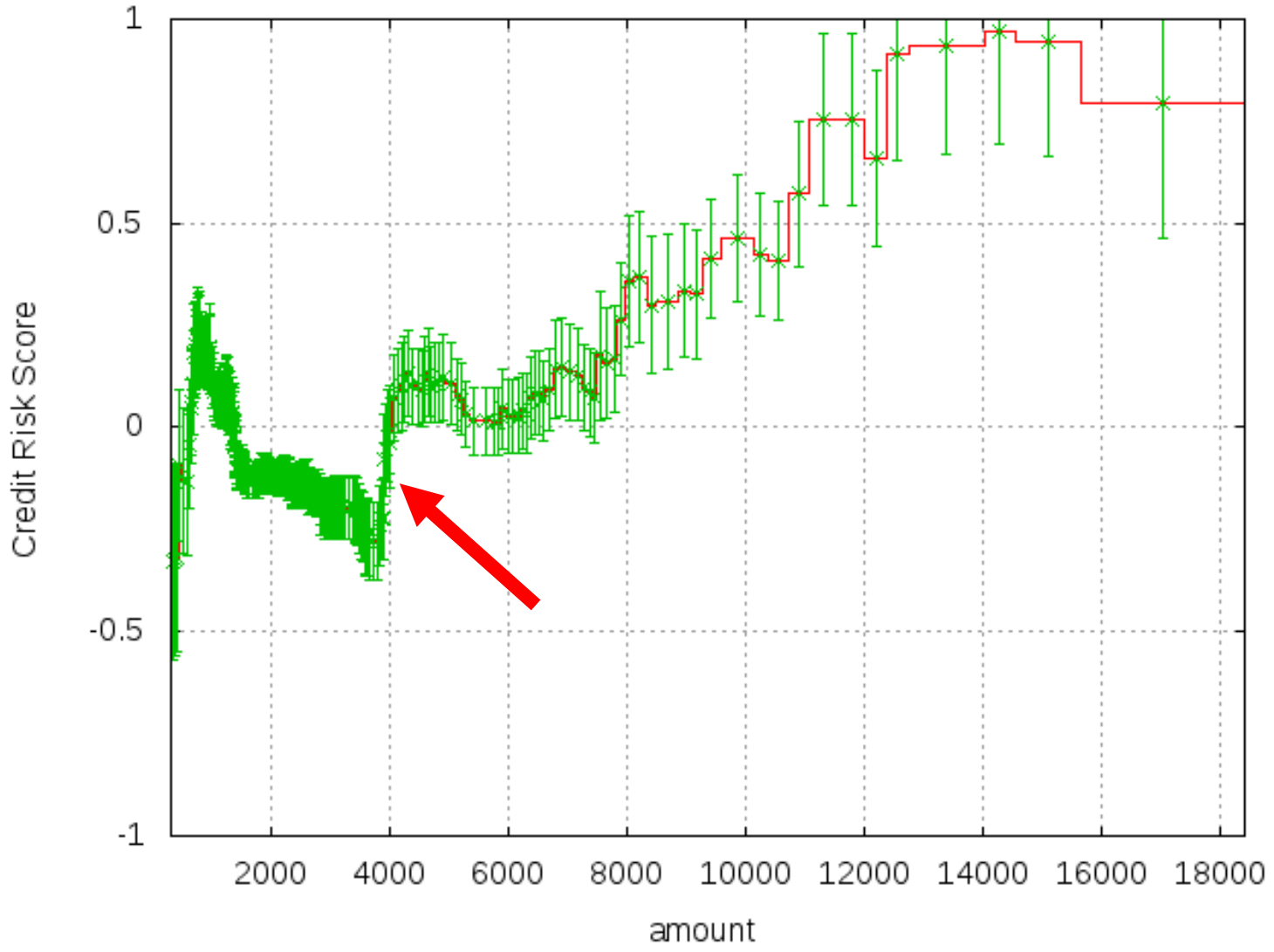
Work with EBMs Show Jumps in Graphs Are Important



Work with EBMs Show Jumps in Graphs Are Important

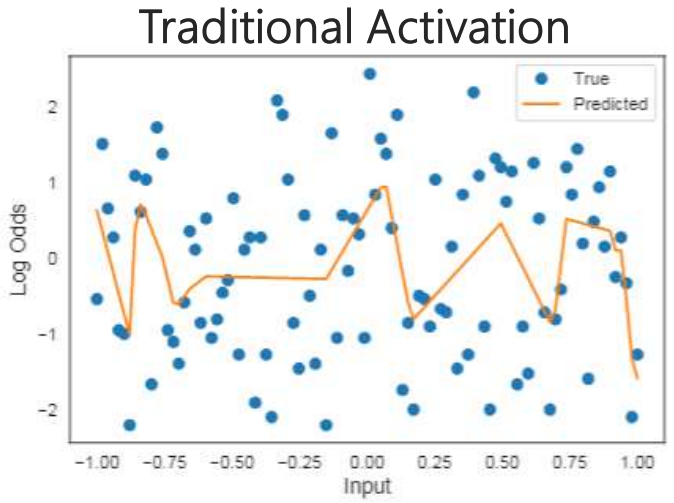


Work with EBMs Show Jumps in Graphs Are Important



DNNs Tend to Be Too Smooth to Learn Jumps Well

- How do we make DNNs "jumpier" without driving the entire model into overfitting?
- Trick is a special activation function: ExU: $h(x) = f\left(e^{w} * (x - b)\right)$
  - slope of activation function can be very steep so small changes in input => large changes in output



- Although overfitting is less of an issue in additive models like NAMs
  - To further reduce overfitting, we apply dropout, weight decay, capped ReLU activations, and also bag the NAM model 25-100 times to form an ensemble

Empirical Results

Accuracy of NAMs

Classification

Model	COMPAS	MIMIC-II	Credit Fraud
Logistic Regression	0.730± 0.014	0.791 ± 0.007	0.975 ± 0.010
Decision Trees	0.723 ± 0.010	0.768 ± 0.008	0.956 ± 0.004
NAMs	0.741 ± 0.009	0.830 ± 0.008	0.980 ± 0.002
EBMs	0.740 ± 0.012	0.835 ± 0.007	0.976 ± 0.009
XGBoost	0.742 ± 0.009	0.844 ± 0.006	0.981 ± 0.008
DNNs	0.735 ± 0.006	0.832 ± 0.009	0.978 ± 0.003

AUC on classification datasets.
Higher is better.

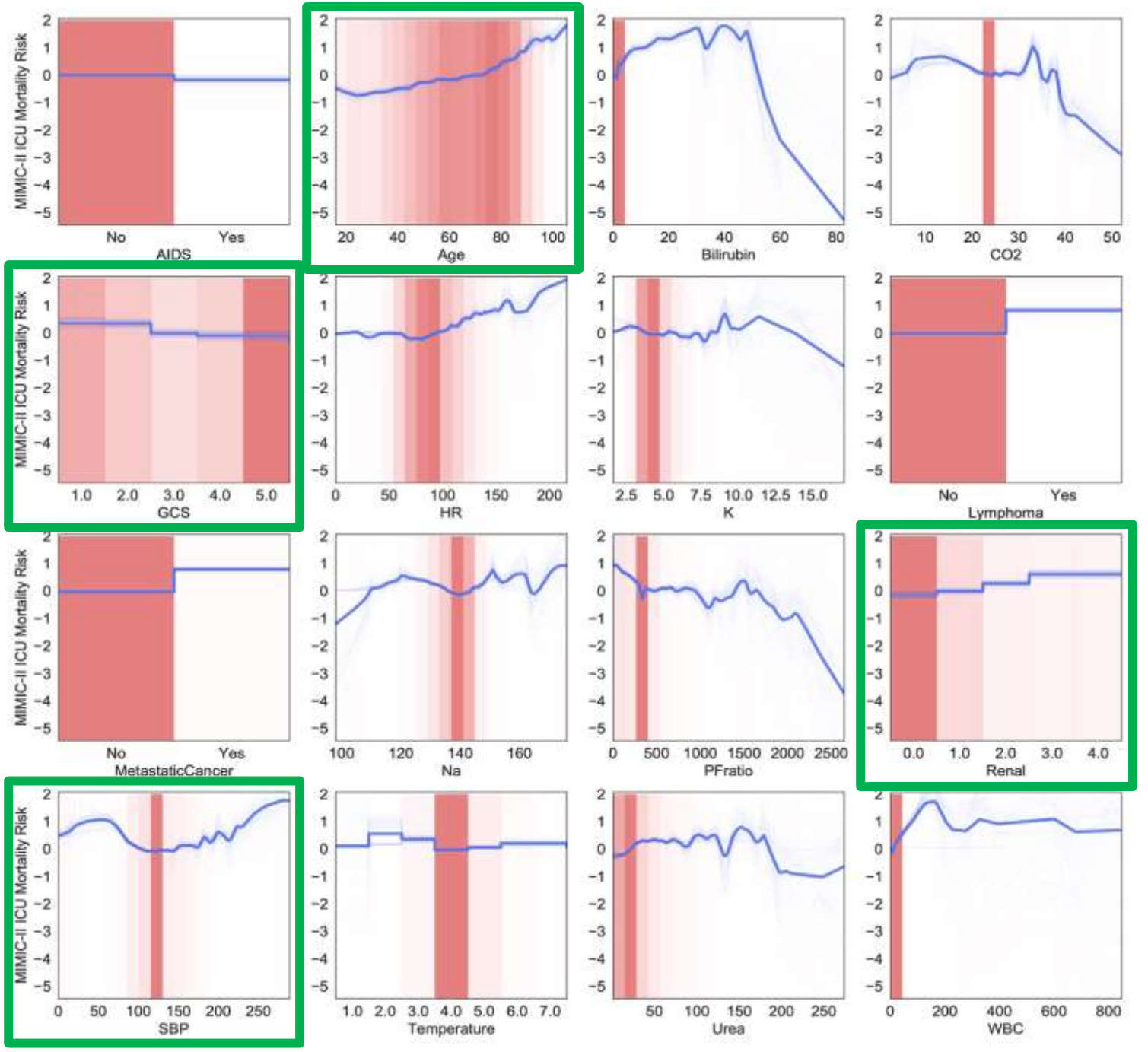
Regression

Model	California Housing	FICO Score
Linear Regression	0.728 ± 0.015	4.344 ± 0.056
Decision Trees	0.720 ± 0.006	4.900 ± 0.113
NAMs	0.562 ± 0.007	3.490 ± 0.081
EBMs	0.557 ± 0.009	3.512 ± 0.095
XGBoost	0.532 ± 0.014	3.345 ± 0.071
DNNs	0.492 ± 0.009	3.324 ± 0.092

RMSE on regression datasets.
Lower is better.

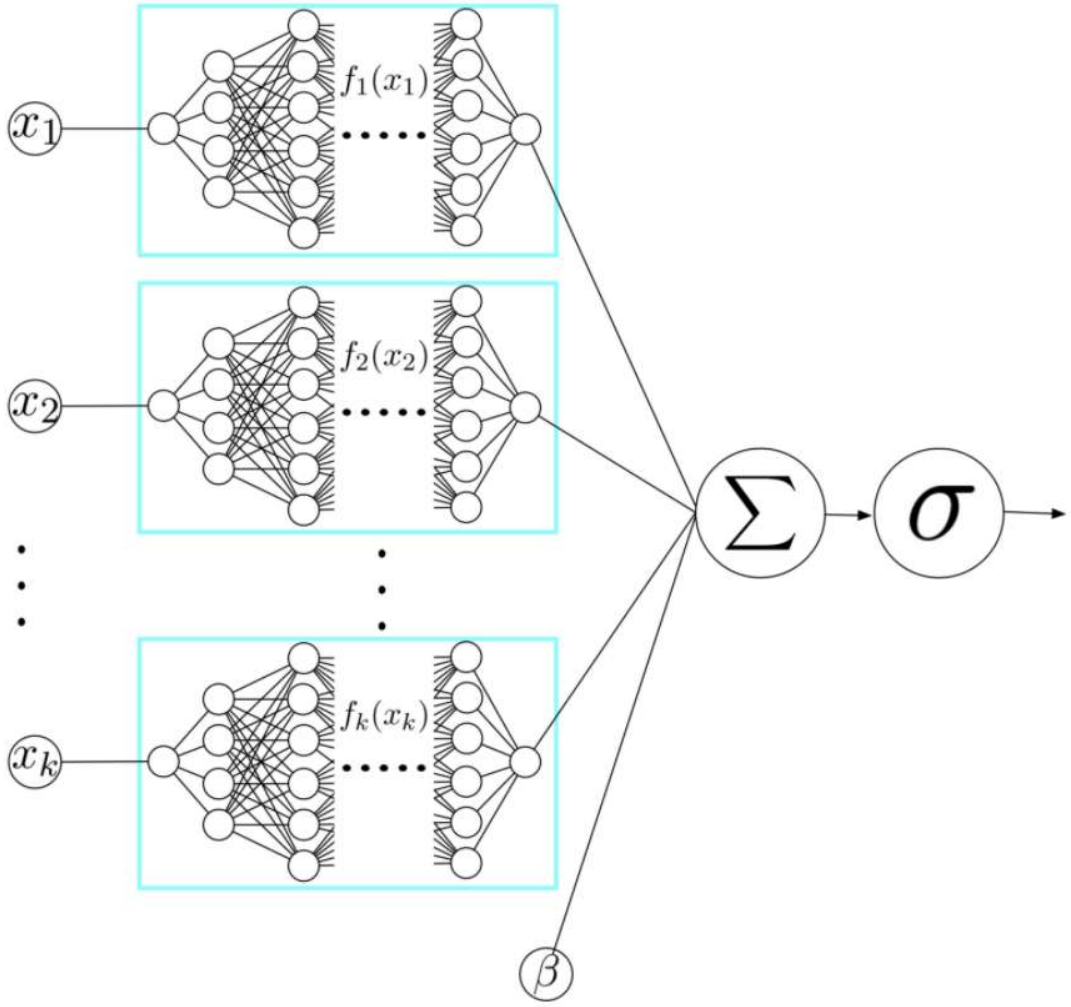
Little or no loss in accuracy for NAMs
compared to DNNs on tabular data!

MIMIC-II ICU Mortality Prediction

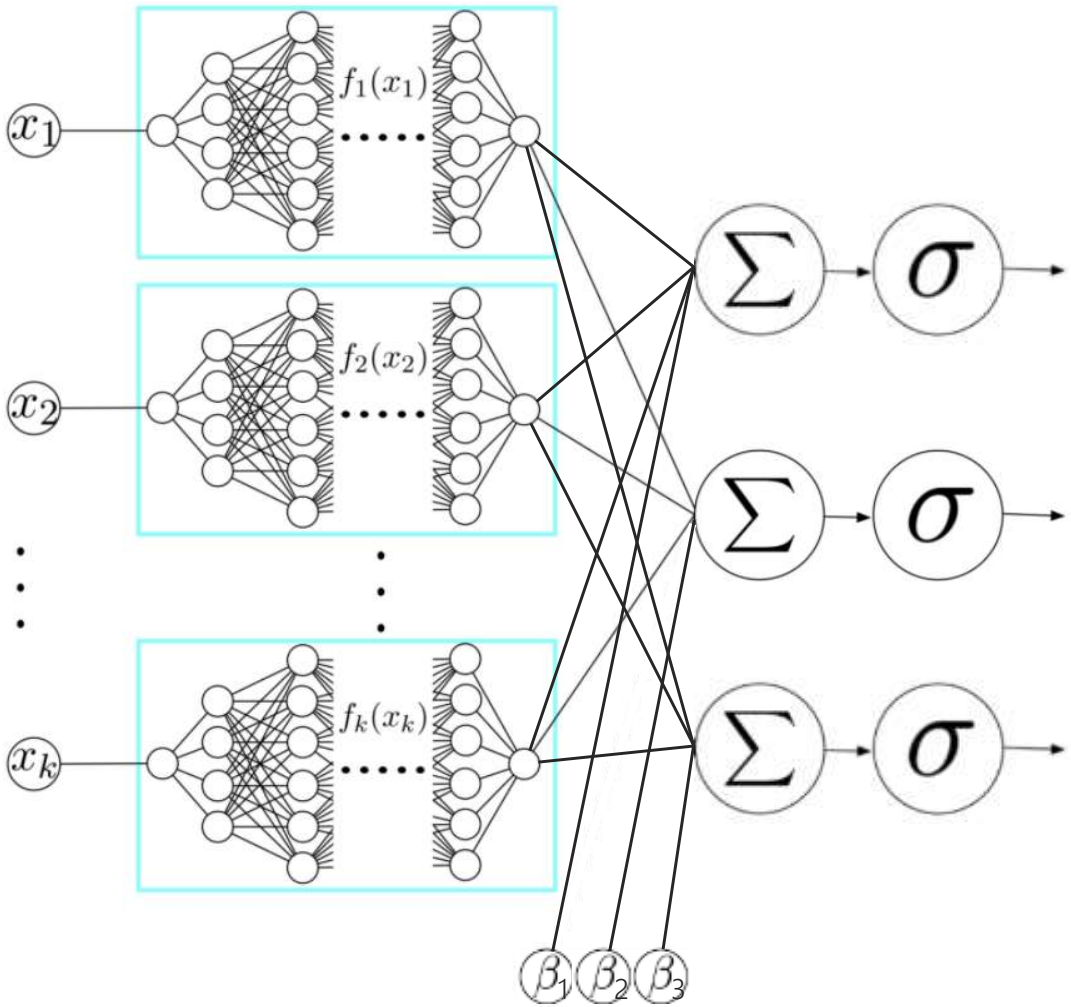


Multitask Learning with NAMs

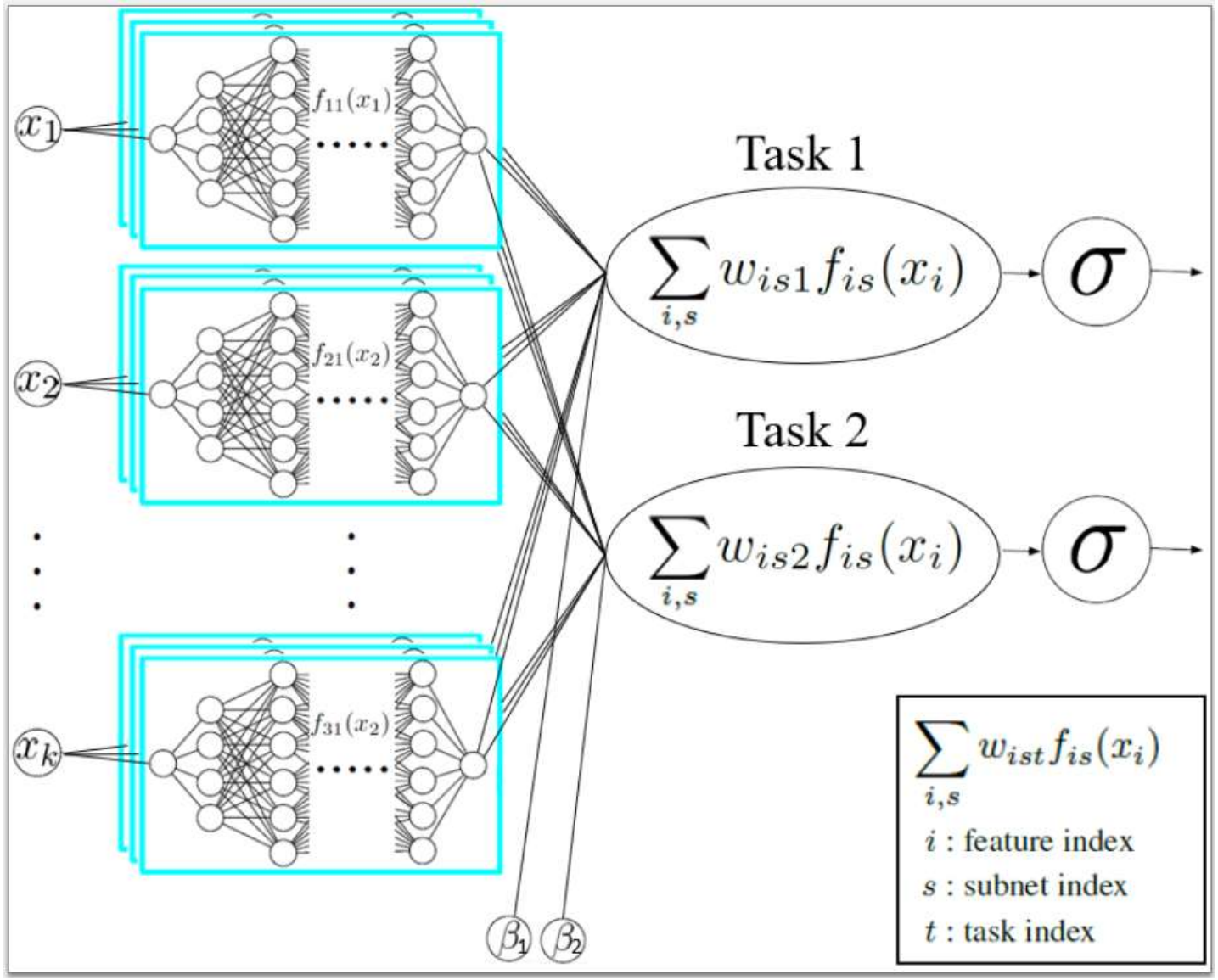
Single Task NAM



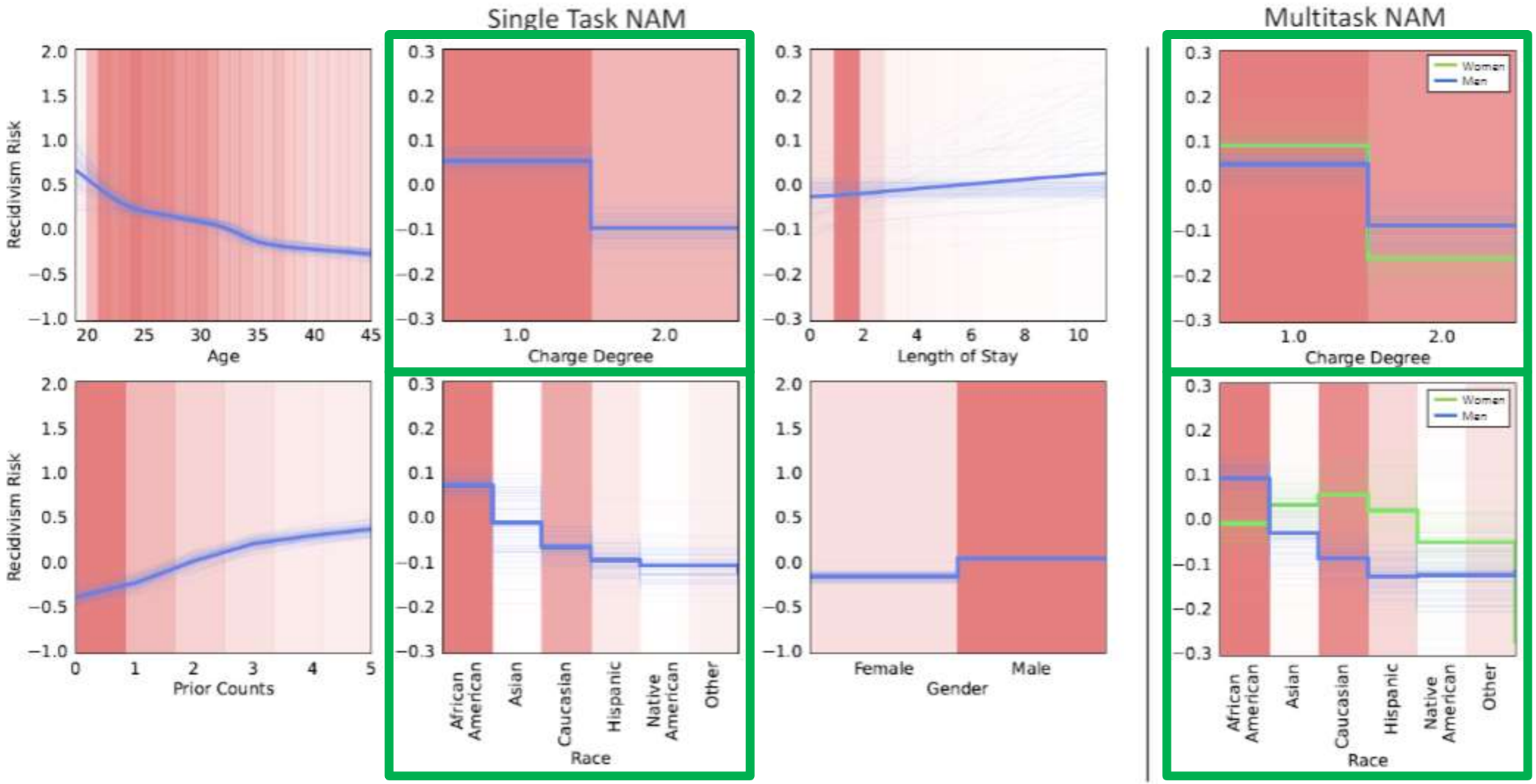
MultiTask NAM



More Flexible MultiTask NAM: Multiple SubNets per Feature



Model	COMPAS Women	COMPAS Men	COMPAS Combined
Single Task NAM	0.716 ± 0.026	0.735 ± 0.009	0.737 ± 0.010
Multitask NAM	0.723 ± 0.019	0.737 ± 0.009	0.739 ± 0.010



Glass-Box EBMs vs. Missing Values

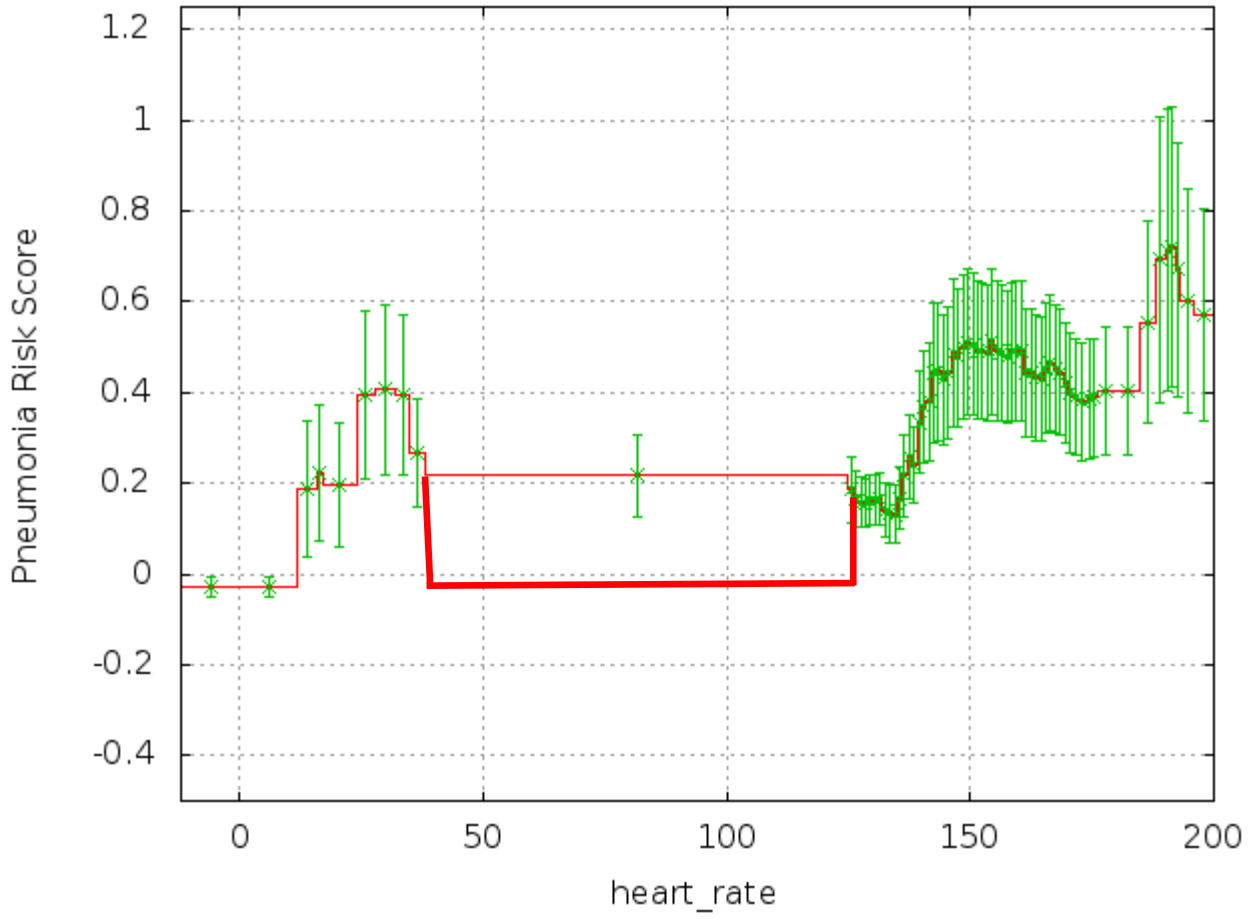
Missing Values

- MAR: Missing at Random
  - No correlation between missing and outcome (i.e., no leakage)
  - Great if true, but often not true
- MCAR: Missing Completely at Random
  - No correlation between missing and outcome or other features
  - Even better than MAR!
- MNAM: Missing Not At Random
  - Most common case
  - **Missing Assumed Normal** (common in healthcare)
  - If value likely normal, don't bother to test/report the value --- if heart rate normal, don't record it

Missing Assumed Normal

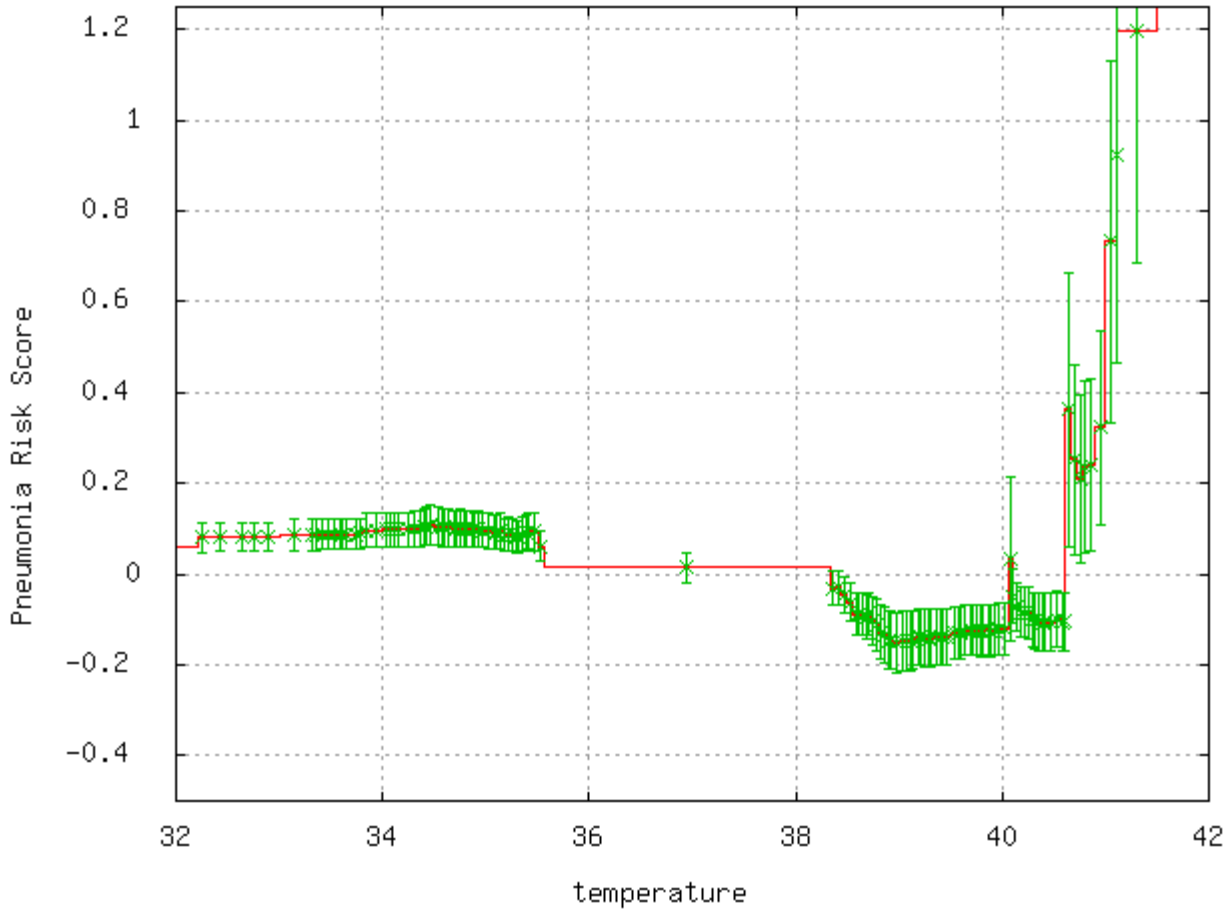
Case Study: Pneumonia Risk

Pneumonia Dataset: Heart Rate (Pulse)



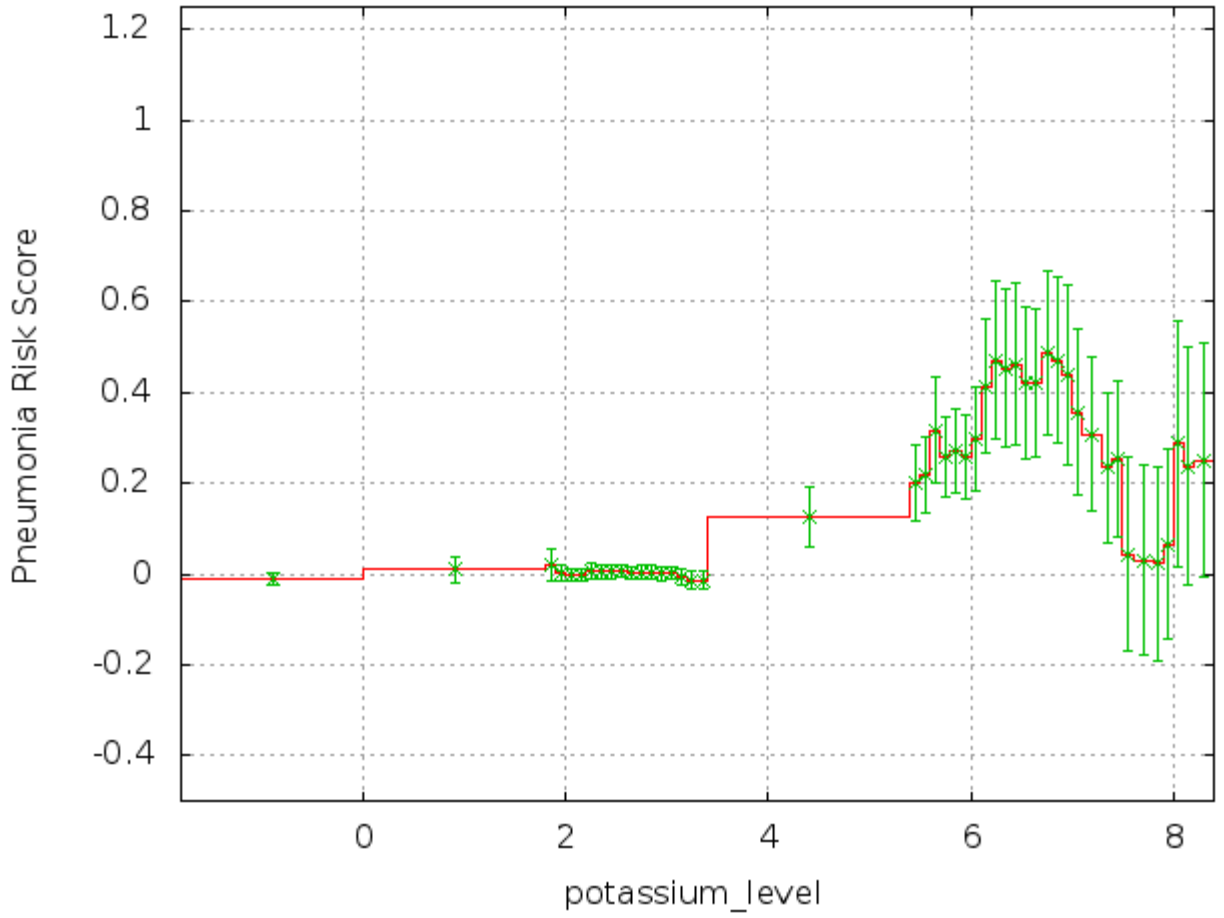
- What is the flat spot in middle for pulse = 40-125?
- 91% patients missing heart rate!
- Missing assumed normal
- Missing coded as 0
- Model sees no data 40-125
- Model interpolates between HR=39 and HR=126
- Would yield bad predictions for normal patients if HR collected!
- Can edit EBM graph to repair

Pneumonia Dataset: Body Temperature (Fever…)



- What is the flat spot in middle for temperature = 35.5C-38.5C?
- 62% patients missing temperature!
- Missing assumed normal
- Missing coded as 0
- Model sees no data 96F-101F
- Model interpolates over the missing data
- In this case, would yield reasonable predictions for normal patients if body temperature was collected
- No edit needed

Pneumonia Dataset: Blood Potassium Level

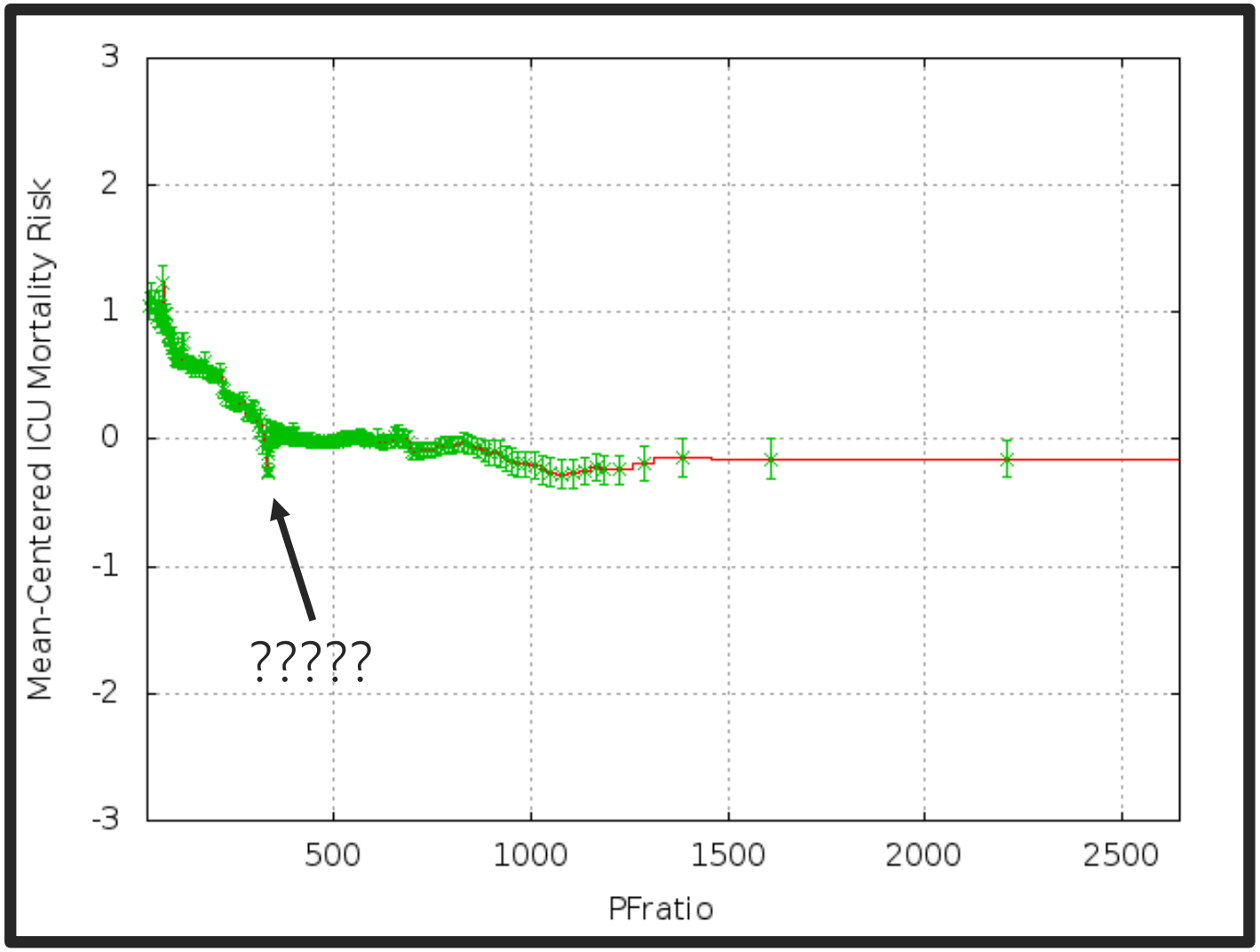


- What is the flat spot in middle for potassium = 3.5-5.2?
- 78% patients missing potassium!
- Missing assumed normal
- Missing coded as 0
- Model sees no data 3.5-5.2
- Model interpolates over missing
- In this case, model will yield mildly incorrect predictions for patients with normal potassium if collected
- Not clear what the right repair is

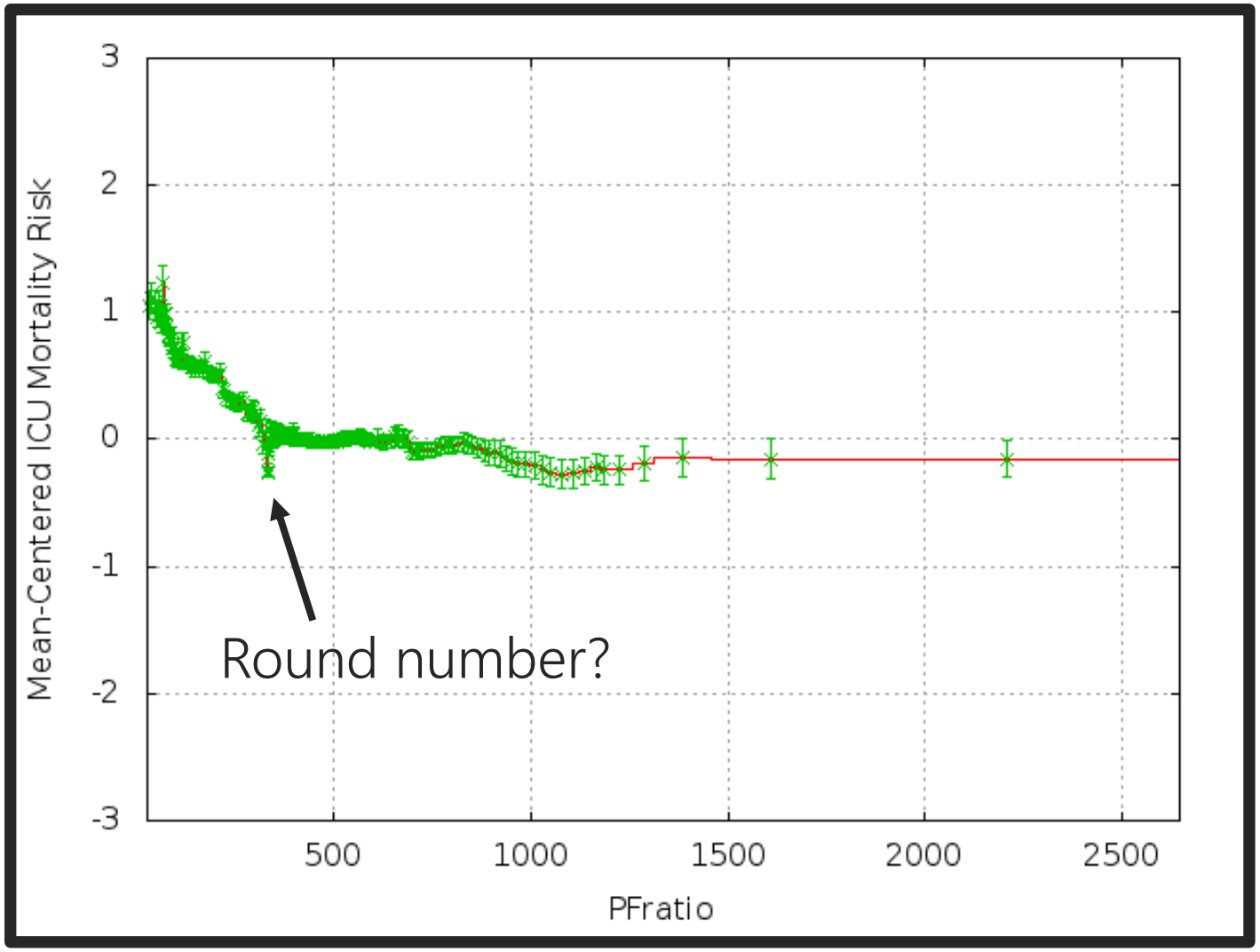
Imputing Missing Values

Case Study: ICU Mortality

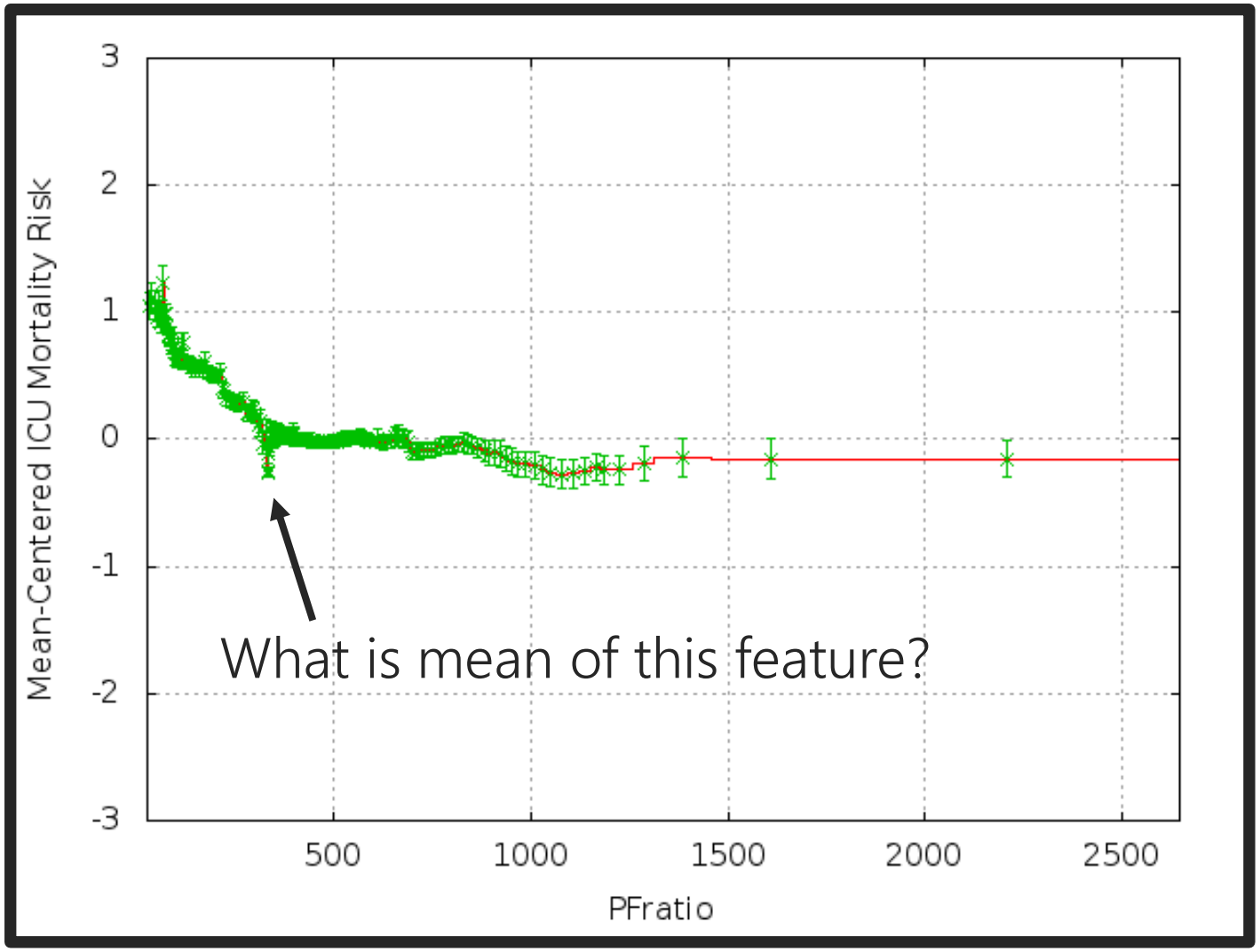
Intelligibility Helps Debug Data: PaO2/FiO2 Ratio



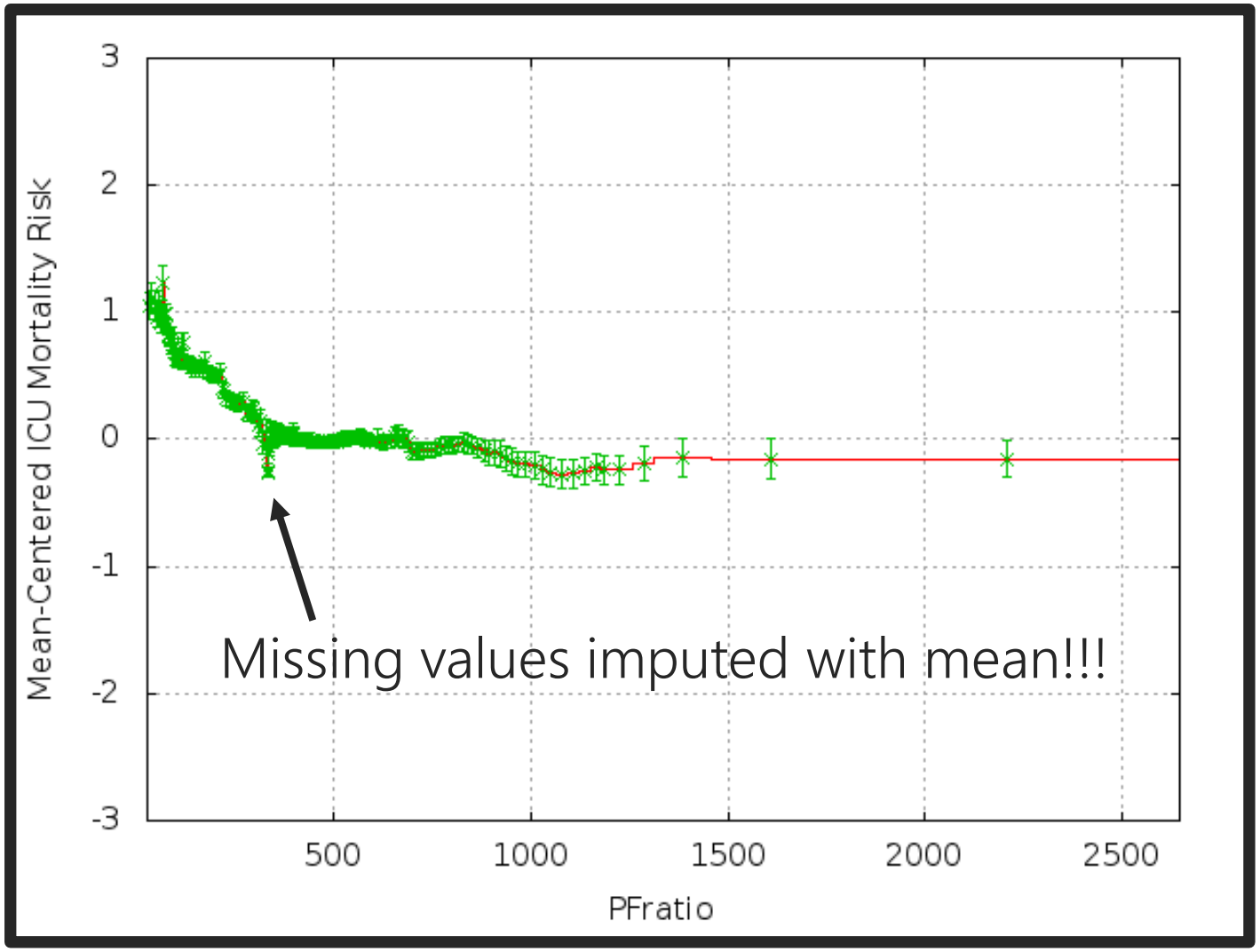
Intelligibility Helps Debug Data: PaO2/FiO2 Ratio



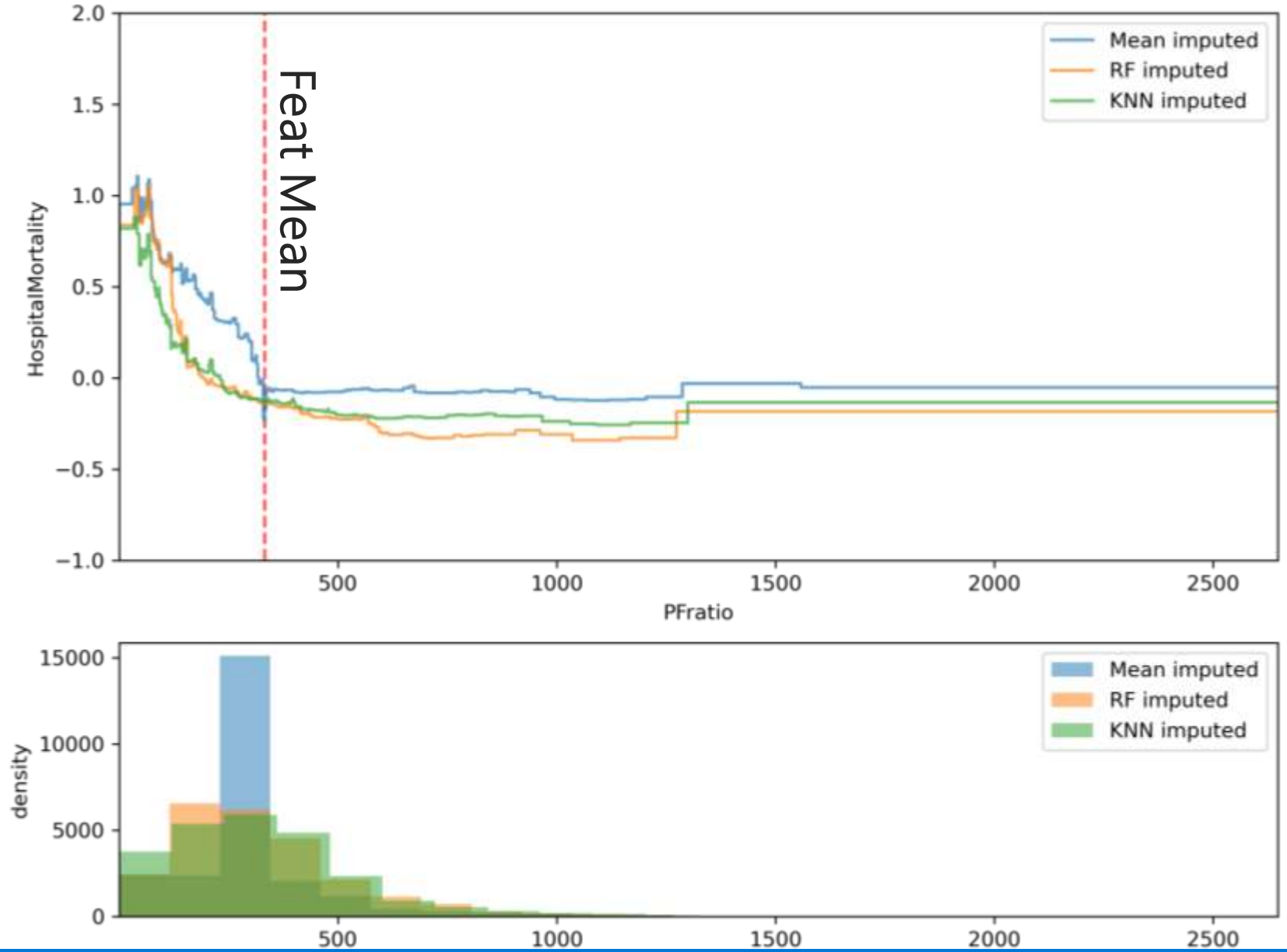
Intelligibility Helps Debug Data: PaO2/FiO2 Ratio



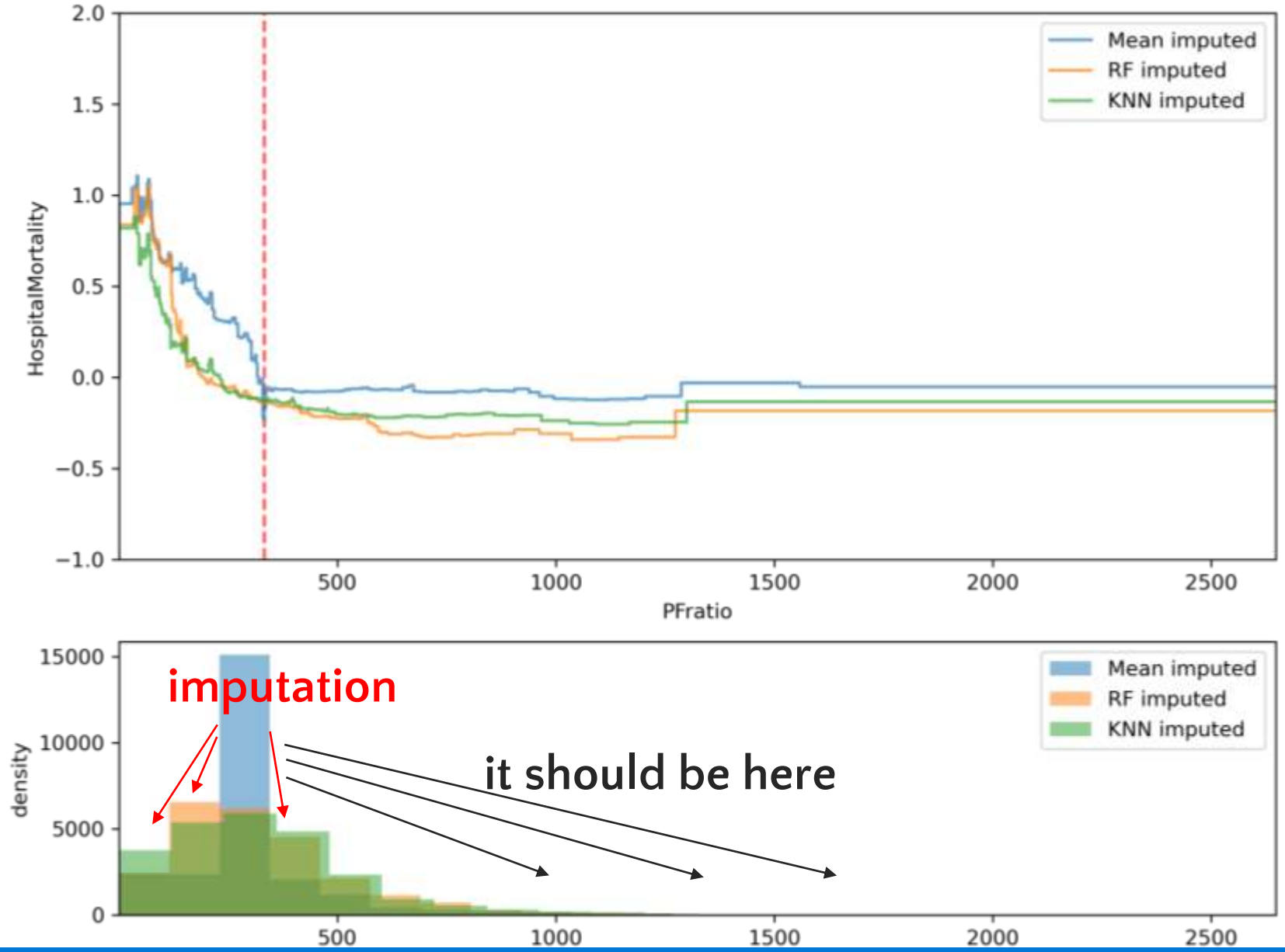
Intelligibility Helps Debug Data: PaO2/FiO2 Ratio



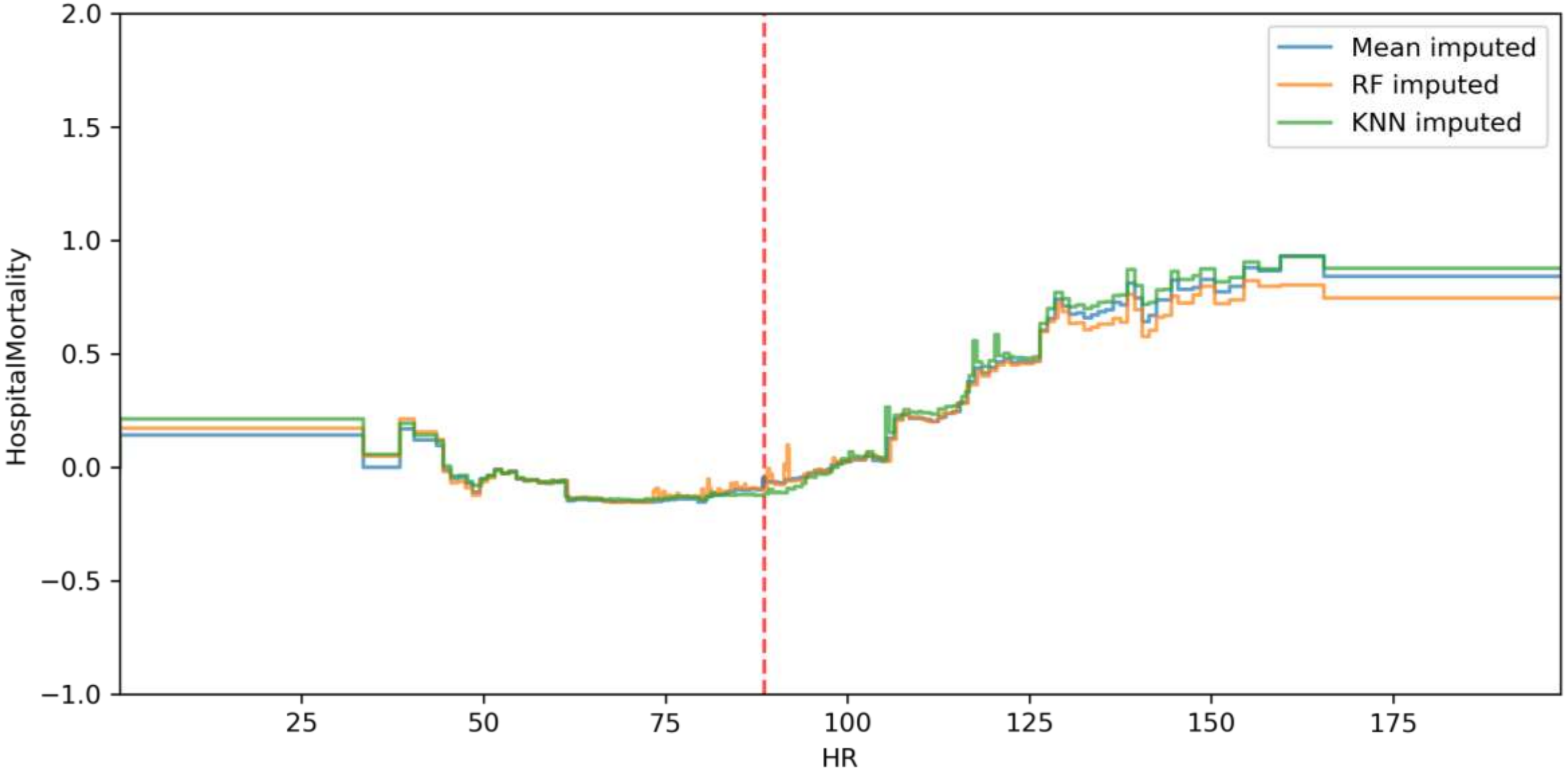
Will More Advanced Imputation Solve Problem?



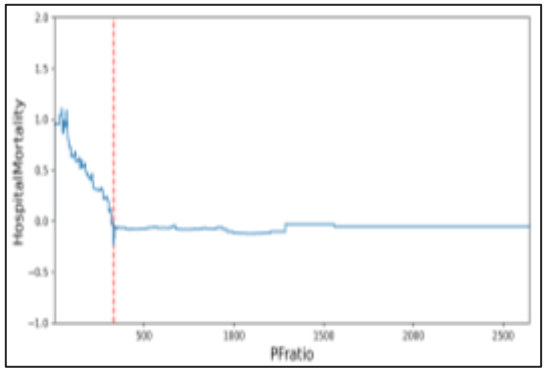
Will More Advanced Imputation Solve Problem?



Will More Advanced Imputation Solve Problem?

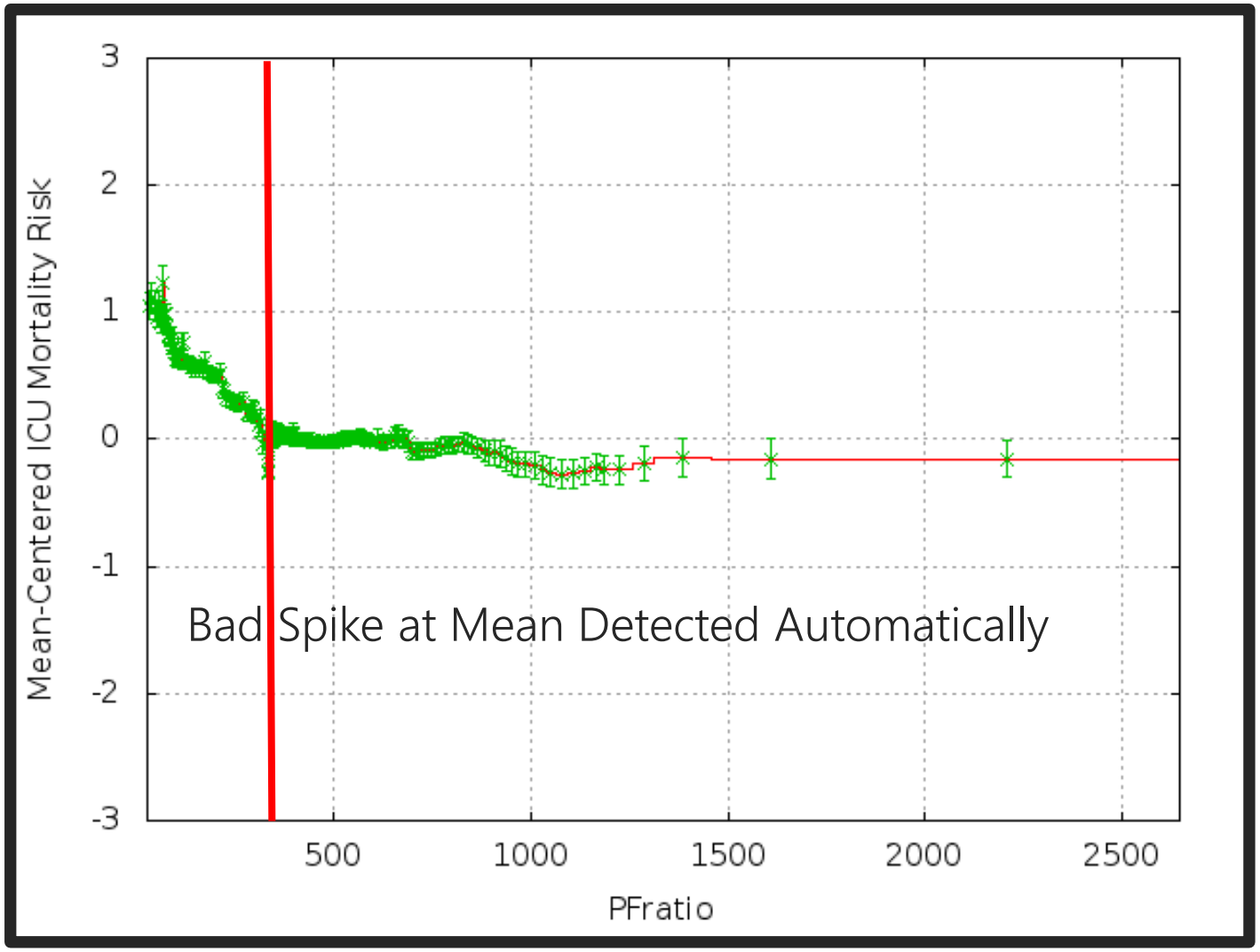


Imputation Often Fails --- Auto Detect Failures?



→ 2nd order derivative of bins → outlier detector → is mean/median an outlier?

Automatically Warns About Bad Mean Impute for Pfratio



Missing Not At Random

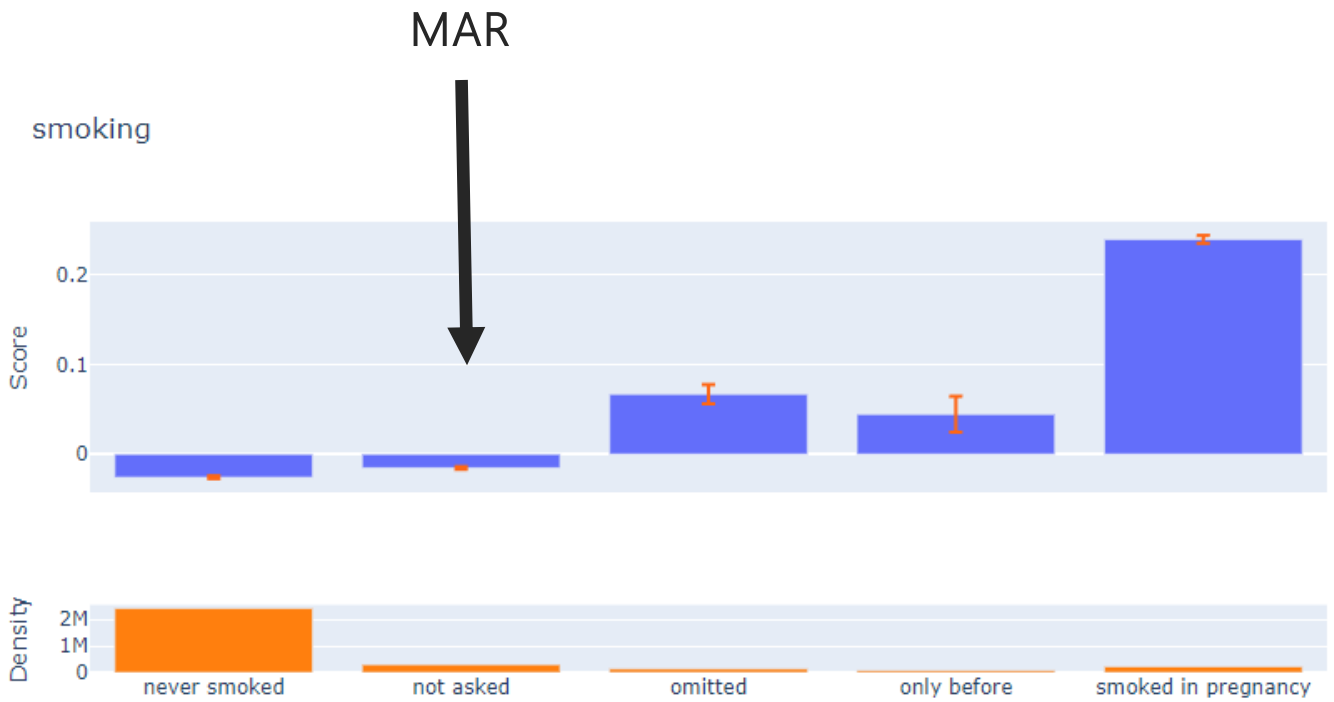
Case Study: Infant Mortality

Effect of smoking in pregnancy

'**Not asked**' – mothers whose states didn't adopt the newer questionnaire that includes this question
Missing at random

'**Omitted**' – mothers who decided not to answer the question
Missing by choice

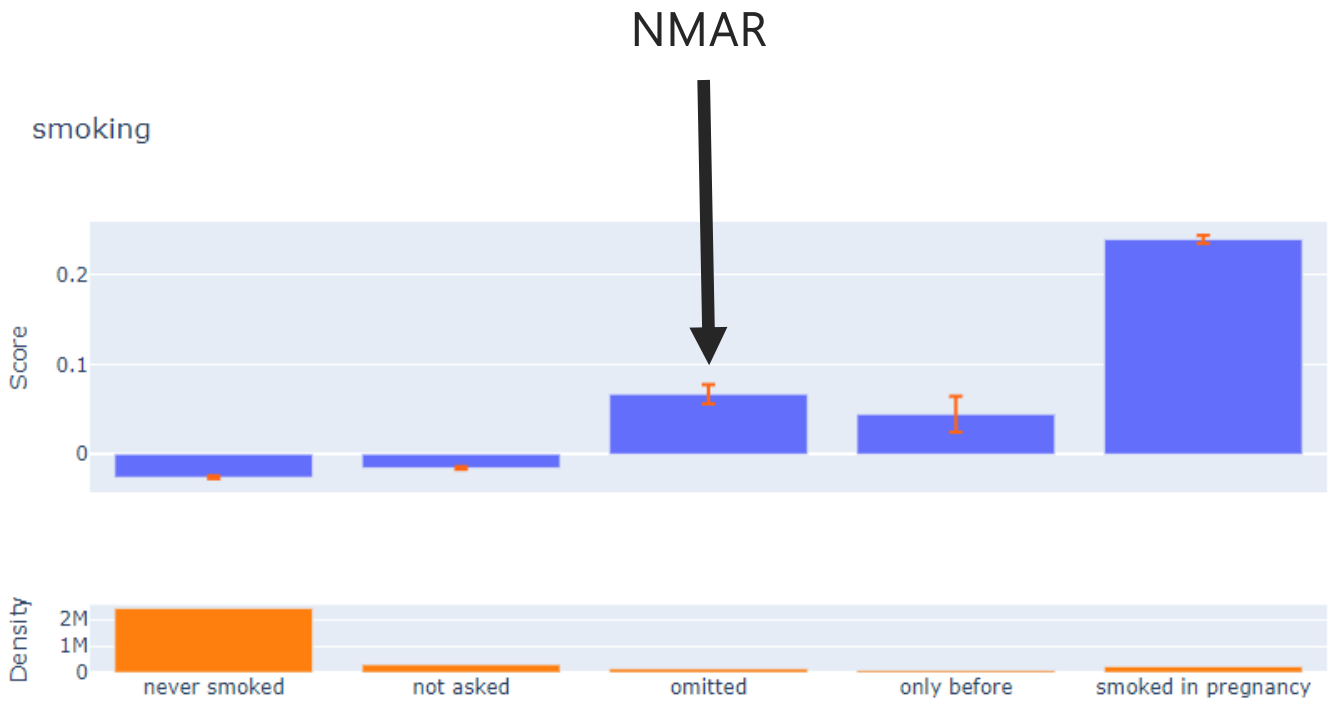
Effect of smoking in pregnancy



'**Not asked**' – mothers whose states didn't adopt the newer questionnaire that includes this question
Missing at random

'Omitted' – mothers who decided not to answer the question
Missing by choice

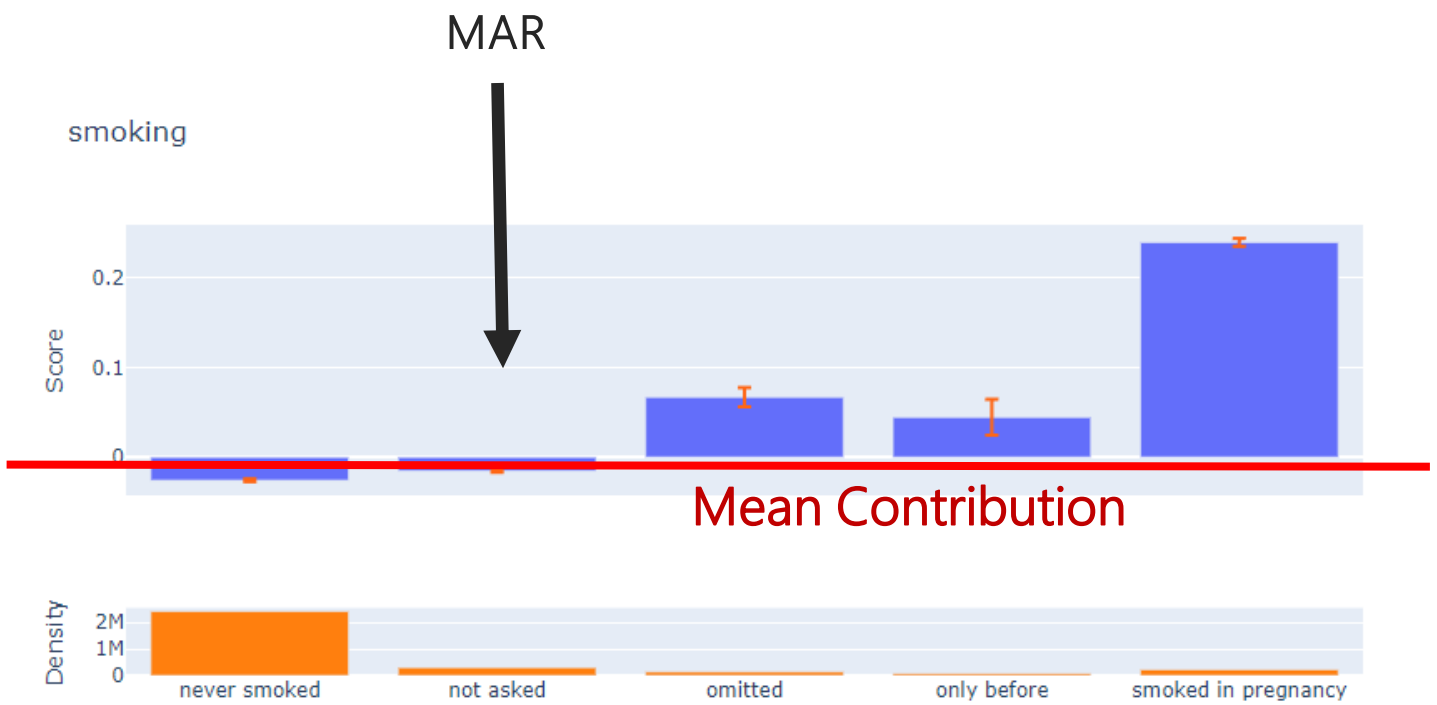
Effect of smoking in pregnancy



'Not asked' – mothers whose states didn't adopt the newer questionnaire that includes this question
Missing at random

'**Omitted**' – mothers who decided not to answer the question
Missing by choice

Effect of smoking in pregnancy



'**Not asked**' – mothers whose states didn't adopt the newer questionnaire that includes this question
Missing at random

'**Omitted**' – mothers who decided not to answer the question
Missing by choice

Risk for missing values should be the same as the average risk for the dataset.

When it is higher (lower), the values are not missing at random

Key Missing Value Learnings and Insights

- Missing values are ubiquitous in all real datasets
- Imputation with the mean/median usually a bad idea --- stop it!
- Surprisingly, more advanced imputation often fails, too!
  - Now that we can see what RF & KNN imputation do, they often don't' do what you hoped
- Sometimes the best way to handle missing is by coding as a unique value
  - With glassbox learning, this is where we start
  - But be careful: a unique value can leak information when missing is correlated with outcome
- Without interpretable, glassbox learning, you're flying blind
  - All models, glassbox or blackbox, learn similar things from data, with glassbox you can see it
- In an interpretable model, if missing isn't predicted as average, it's not MAR

Key Missing Value Learnings and Insights

- Missing values are ubiquitous in all real datasets
- Imputation with the mean/median usually a bad idea --- stop it!
- Surprisingly, more advanced imputation often fails, too!
  - Now that we can see what RF & KNN imputation do, they often don't' do what you hoped
- Sometimes the best way to handle missing is by coding as a unique value
  - With glassbox learning, this is where we start
  - But be careful: a unique value can leak information when missing is correlated with outcome
- Without interpretable, glassbox learning, you're flying blind
  - All models, glassbox or blackbox, learn similar things from data, with glassbox you can see it
- In an interpretable model, if missing isn't predicted as average, it's not MAR
- Sometimes editing the model can help solve missing value problems

Differential Privacy via EBMs

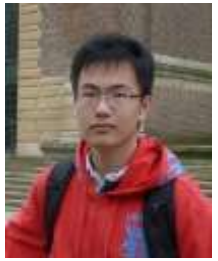
ICML 2021

Harsha Nori

MSR Aether

Rich Caruana

MSR AI



Zhiqi Bu

U. Penn

Judy Shen

Stanford

Janardhan Kulkarni

MSR Algorithms

Motivations

- Customers and regulators care about data privacy
- Anonymization methods are easy to break
- Data powers business
- Differential privacy is the gold standard
  - Provides formal, *mathematical* privacy guarantees
  - Still lets us do data science!

The New York Times

TECHNOLOGY

A Face Is Exposed for AOL Searcher No. 4417749

By MICHAEL BARBARO and TOM ZELLER Jr. AUG. 9, 2006

Thelma Arnold's identity was betrayed by AOL records of her Web searches, like ones for her dog, Dudley, who clearly has a problem.

Erik S. Lesser for The New York Times

NETFLIX

Researchers reverse Netflix anonymization

Robert Lemos, SecurityFocus 2007-12-04

Example Scenario

Government wants to know:
What % of people took illegal drugs in high school?

Traditional Analysis:

-Survey kids

-Calculate Summary

-Publish Results

First Name	Last Name	Age	Race	Answer
Harsha	Nori	23	Asian	No
Carol	Williams	27	Mixed	Yes
John	Johnson	33	White	Yes
Bob	Smith	22	White	No
Alice	Allen	60	Black	No

Answer:
40%

Dropping Identifiers?

Government wants to know:
What % of people took illegal drugs in high school?

~~First Name~~	~~Last Name~~	Age	Race	Answer
~~Harsha~~	~~Nori~~	23	Asian	No
~~Carol~~	~~Williams~~	27	Mixed	Yes
~~John~~	~~Johnson~~	33	White	Yes
~~Bob~~	~~Smith~~	22	White	No
~~Alice~~	~~Allen~~	60	Black	No

We need attributes to slice data – can't always drop them too!

No names left, but *any* prior information breaks this.

Ex: Knowing Carol is mixed race immediately betrays her answer.

Dimensionality:
Even with billions of records, combinations of attributes will tie to unique individuals.

A better mechanism: Randomized Response

Government wants to know:
What % of people took illegal drugs in high school?

First Name	Last Name	Age	Race	Answer
Harsha	Nori	23	Asian	No->No
Carol	Williams	27	Mixed	Yes
John	Johnson	33	White	Yes->No
Bob	Smith	22	White	No
Alice	Allen	60	Black	No->YES

Before responding, tell each user to flip a (private) coin.

If heads: Report the truth.

If tails: Coin Decides -- Flip it again . If heads, say YES. If tails, NO.

-Every individual now has plausible deniability. (“I said yes because of the coin!”)
-In the aggregate, we can *subtract out* the noise, and get a great estimate!

Takeaways

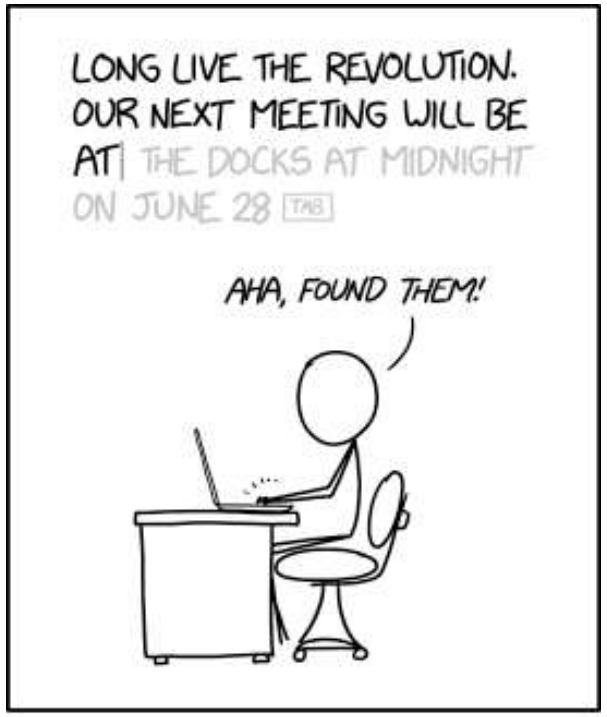
Government wants to know:
What % of people took illegal drugs in high school?

First Name	Last Name	Age	Race	Answer
Harsha	Nori	23	Asian	No->No
Carol	Williams	27	Mixed	Yes
John	Johnson	33	White	Yes->No
Bob	Smith	22	White	No
Alice	Allen	60	Black	No->YES

Any dataset OR algorithm run on a dataset leaks information.

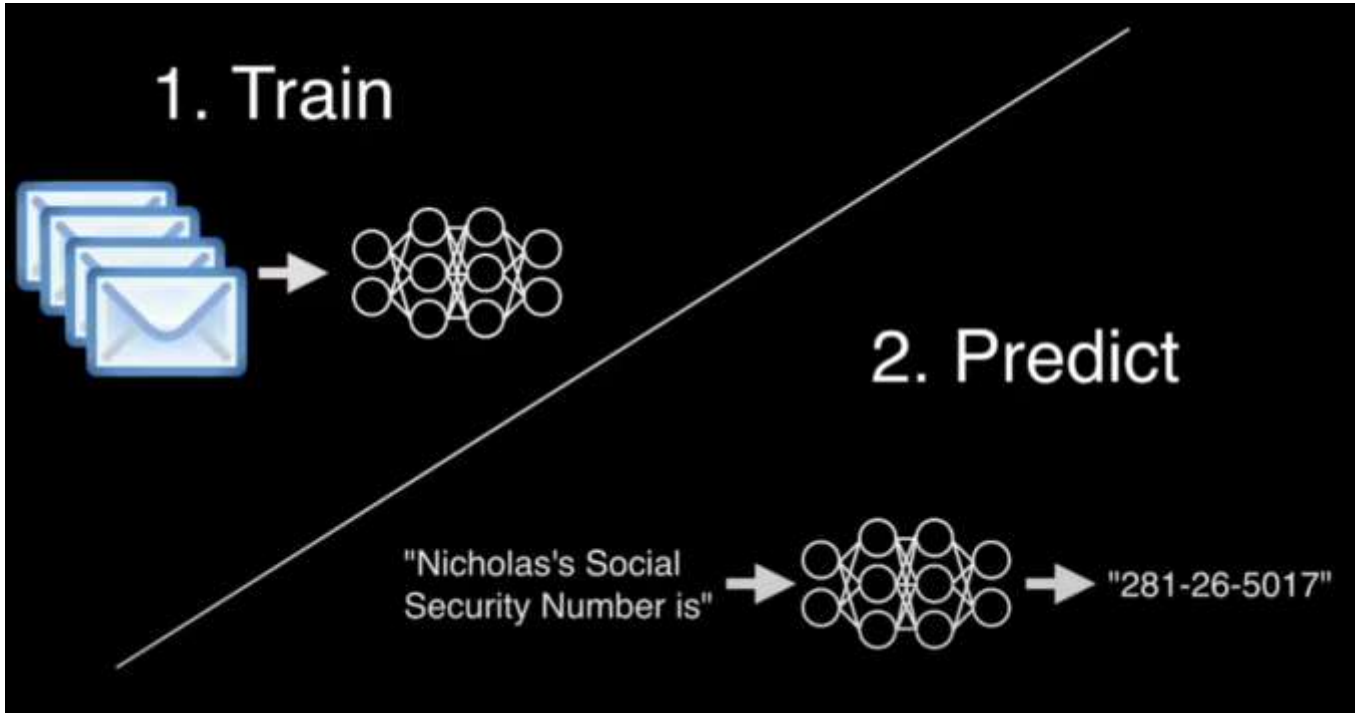
Randomized algorithms can help protect privacy through plausible deniability.

ML Algorithms Leak Information!



WHEN YOU TRAIN PREDICTIVE MODELS ON INPUT FROM YOUR USERS, IT CAN LEAK INFORMATION IN UNEXPECTED WAYS.

[HTTPS://XKCD.COM/2169/](https://xkcd.com/2169/)



CARLINI, NICHOLAS, ET AL. "THE SECRET SHARER: EVALUATING AND TESTING UNINTENDED MEMORIZATION IN NEURAL NETWORKS." *28TH USENIX SECURITY SYMPOSIUM (USENIX SECURITY 19)*. 2019.

Differential Privacy

- Formal, mathematical standard for privacy:

$$\Pr[A(D_1) \in S] \leq e^{\varepsilon} * \Pr[A(D_2) \in S] + \delta$$

- Plain English: the output of algorithms run *with* and *without* any individual's data should be hard to distinguish across *all* users.

- (ε, δ) privacy parameters. Making these smaller increases ***privacy*** but weakens ***utility***.

Dataset D1

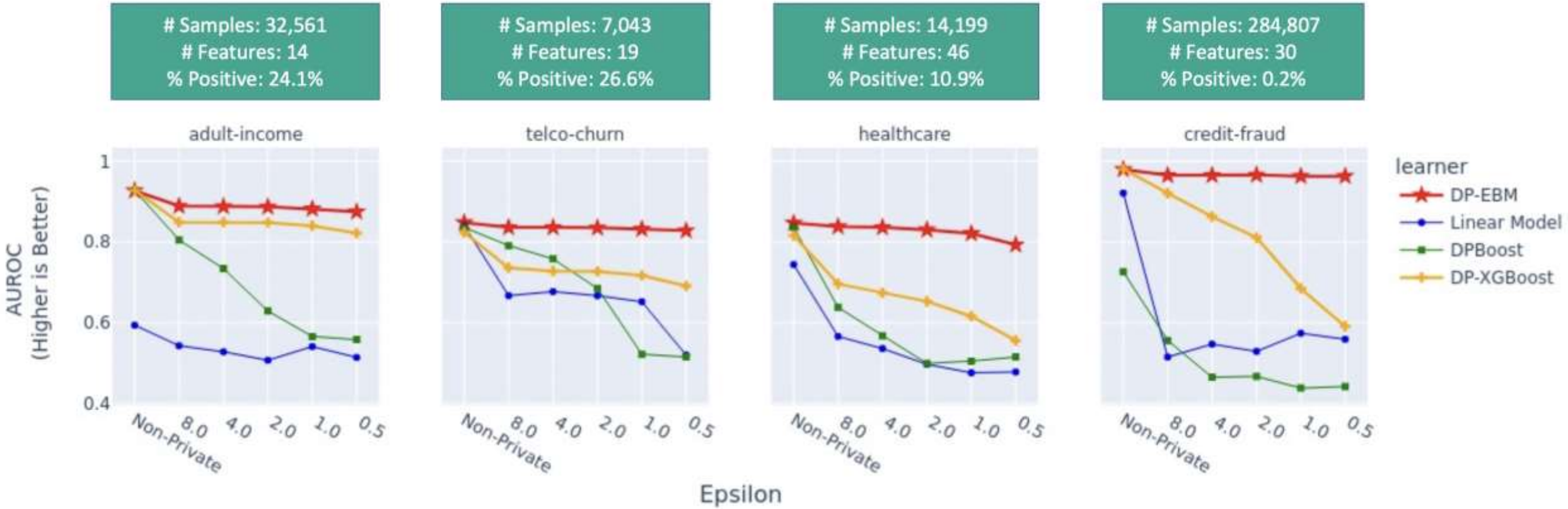
Name	Net Worth
Alice	$2 Million
Bob	$50,000
Carol	$850,000
Bill G.	~$100 Billion

Dataset D2

Name	Net Worth
Alice	$2 Million
Bob	$50,000
Carol	$850,000

A differentially private algorithm "A" must produce "approximately equal" results for D1 and D2.

High Accuracy, Perfect Interpretability, Strong Privacy



Comparison of DP-EBM with DP Logistic Regression, DPBoost, and DP-XGBoost.
Average AUROC of 25 folds of cross validation at varying privacy guarantees.

Making EBMs Differentially Private

Creating differentially private algorithms

SENSITIVITY

COMPOSITION

Creating differentially private algorithms

Sensitivity

- How much can changing a single entry in your dataset affect the output?

Composition

Creating differentially private algorithms

SENSITIVITY

- How much can changing a single entry in your dataset affect the output?

$$\Delta f = \max |f(x) - f(y)|$$

  - Ex: Count | Sensitivity = 1
  - Ex: Sum | Sensitivity = max() – min()
  - More difficult with ML algorithms

COMPOSITION

Creating differentially private algorithms

SENSITIVITY

- How much can changing a single entry in your dataset affect the output?

$$\Delta f = \max |f(x) - f(y)|$$

  - Ex: Count | Sensitivity = 1
  - Ex: Sum | Sensitivity = max() – min()
  - More difficult with ML algorithms

- Laplace: $\mathcal{M}_{\text{Lap}}(x, f, \epsilon) = f(x) + \text{Lap}\left(\mu = 0, b = \frac{\Delta f}{\epsilon}\right)$
- Gaussian:

$$\mathcal{M}_{\text{Gauss}}(x, f, \epsilon, \delta) = f(x) + \mathcal{N}\left(\mu = 0, \sigma^2 = \frac{2\ln(1.25/\delta) \cdot (\Delta f)^2}{\epsilon^2}\right)$$

COMPOSITION

Creating differentially private algorithms

Sensitivity

- How much can changing a single entry in your dataset affect the output?

$$\Delta f = \max |f(x) - f(y)|$$

  - Ex: Count | Sensitivity = 1
  - Ex: Sum | Sensitivity = max() – min()
  - More difficult with ML algorithms

- Laplace: $\mathcal{M}_{\text{Lap}}(x, f, \epsilon) = f(x) + \text{Lap}\left(\mu = 0, b = \frac{\Delta f}{\epsilon}\right)$
- Gaussian: $\mathcal{M}_{\text{Gauss}}(x, f, \epsilon, \delta) = f(x) + \mathcal{N}\left(\mu = 0, \sigma^2 = \frac{2\ln(1.25/\delta) \cdot (\Delta f)^2}{\epsilon^2}\right)$

Composition

Creating differentially private algorithms

Sensitivity

- How much can changing a single entry in your dataset affect the output?

$$\Delta f = \max |f(x) - f(y)|$$

  - Ex: Count | Sensitivity = 1
  - Ex: Sum | Sensitivity = max() – min()
  - More difficult with ML algorithms

- Laplace: $\mathcal{M}_{\text{Lap}}(x, f, \epsilon) = f(x) + \text{Lap}\left(\mu = 0, b = \frac{\Delta f}{\epsilon}\right)$
- Gaussian: $\mathcal{M}_{\text{Gauss}}(x, f, \epsilon, \delta) = f(x) + \mathcal{N}\left(\mu = 0, \sigma^2 = \frac{2\ln(1.25/\delta) \cdot (\Delta f)^2}{\epsilon^2}\right)$

Composition

- What's the final privacy guarantee if we apply *k* different DP mechanisms?

Creating differentially private algorithms

Sensitivity

- How much can changing a single entry in your dataset affect the output?

$$\Delta f = \max |f(x) - f(y)|$$

  - Ex: Count | Sensitivity = 1
  - Ex: Sum | Sensitivity = max() – min()
  - More difficult with ML algorithms

- Laplace: $\mathcal{M}_{\text{Lap}}(x, f, \epsilon) = f(x) + \text{Lap}\left(\mu = 0, b = \frac{\Delta f}{\epsilon}\right)$
- Gaussian: $\mathcal{M}_{\text{Gauss}}(x, f, \epsilon, \delta) = f(x) + \mathcal{N}\left(\mu = 0, \sigma^2 = \frac{2\ln(1.25/\delta) \cdot (\Delta f)^2}{\epsilon^2}\right)$

Composition

- What's the final privacy guarantee if we apply *k* different DP mechanisms?
  - Naïve: $\varepsilon_{final} = \varepsilon_1 + \varepsilon_2 + \ldots + \varepsilon_k$

Creating differentially private algorithms

Sensitivity

- How much can changing a single entry in your dataset affect the output?

$$\Delta f = \max |f(x) - f(y)|$$

  - Ex: Count | Sensitivity = 1
  - Ex: Sum | Sensitivity = max() – min()
  - More difficult with ML algorithms

- Laplace: $\mathcal{M}_{\text{Lap}}(x, f, \epsilon) = f(x) + \text{Lap}\left(\mu = 0, b = \frac{\Delta f}{\epsilon}\right)$
- Gaussian: $\mathcal{M}_{\text{Gauss}}(x, f, \epsilon, \delta) = f(x) + \mathcal{N}\left(\mu = 0, \sigma^2 = \frac{2\ln(1.25/\delta) \cdot (\Delta f)^2}{\epsilon^2}\right)$

Composition

- What's the final privacy guarantee if we apply *k* different DP mechanisms?
  - Naïve: $\varepsilon_{final} = \varepsilon_1 + \varepsilon_2 + \ \ldots \ + \ \varepsilon_k$

- Turns out we can do better!
  - Strong: $\varepsilon_{final} \approx \sqrt{(\varepsilon_1 + \varepsilon_2 + \ \ldots \ + \ \varepsilon_k)}$

Creating differentially private algorithms

Sensitivity

- How much can changing a single entry in your dataset affect the output?

$$\Delta f = \max |f(x) - f(y)|$$

  - Ex: Count | Sensitivity = 1
  - Ex: Sum | Sensitivity = max() – min()
  - More difficult with ML algorithms

- Laplace: $\mathcal{M}_{\text{Lap}}(x, f, \epsilon) = f(x) + \text{Lap}\left(\mu = 0, b = \frac{\Delta f}{\epsilon}\right)$
- Gaussian: $\mathcal{M}_{\text{Gauss}}(x, f, \epsilon, \delta) = f(x) + \mathcal{N}\left(\mu = 0, \sigma^2 = \frac{2\ln(1.25/\delta) \cdot (\Delta f)^2}{\epsilon^2}\right)$

Composition

- What's the final privacy guarantee if we apply *k* different DP mechanisms?
  - Naïve: $\varepsilon_{final} = \varepsilon_1 + \varepsilon_2 + \ \ldots \ + \ \varepsilon_k$
- Turns out we can do better!
  - Strong: $\varepsilon_{final} \approx \sqrt{(\varepsilon_1 + \varepsilon_2 + \ \ldots \ + \ \varepsilon_k)}$
- Gaussian Differential Privacy
  - [Dong, Roth, Su 2019]
  - Best known composition theorem for Gaussian Mechanisms

Creating differentially private algorithms

Sensitivity

- How much can changing a single entry in your dataset affect the output?

$$\Delta f = \max |f(x) - f(y)|$$

  - Ex: Count | Sensitivity = 1
  - Ex: Sum | Sensitivity = max() – min()
  - More difficult with ML algorithms

- Laplace: $\mathcal{M}_{\text{Lap}}(x, f, \epsilon) = f(x) + \text{Lap}\left(\mu = 0, b = \frac{\Delta f}{\epsilon}\right)$
- Gaussian:

$$\mathcal{M}_{\text{Gauss}}(x, f, \epsilon, \delta) = f(x) + \mathcal{N}\left(\mu = 0, \sigma^2 = \frac{2\ln(1.25/\delta) \cdot (\Delta f)^2}{\epsilon^2}\right)$$

Composition

- What's the final privacy guarantee if we apply *k* different DP mechanisms?
  - Naïve: $\varepsilon_{final} = \varepsilon_1 + \varepsilon_2 + \ \ldots \ + \ \varepsilon_k$
- Turns out we can do better!
  - Strong: $\varepsilon_{final} \approx \sqrt{(\varepsilon_1 + \varepsilon_2 + \ \ldots \ + \ \varepsilon_k)}$
- Gaussian Differential Privacy
  - [Dong, Roth, Su 2019]
  - Best known composition theorem for Gaussian Mechanisms

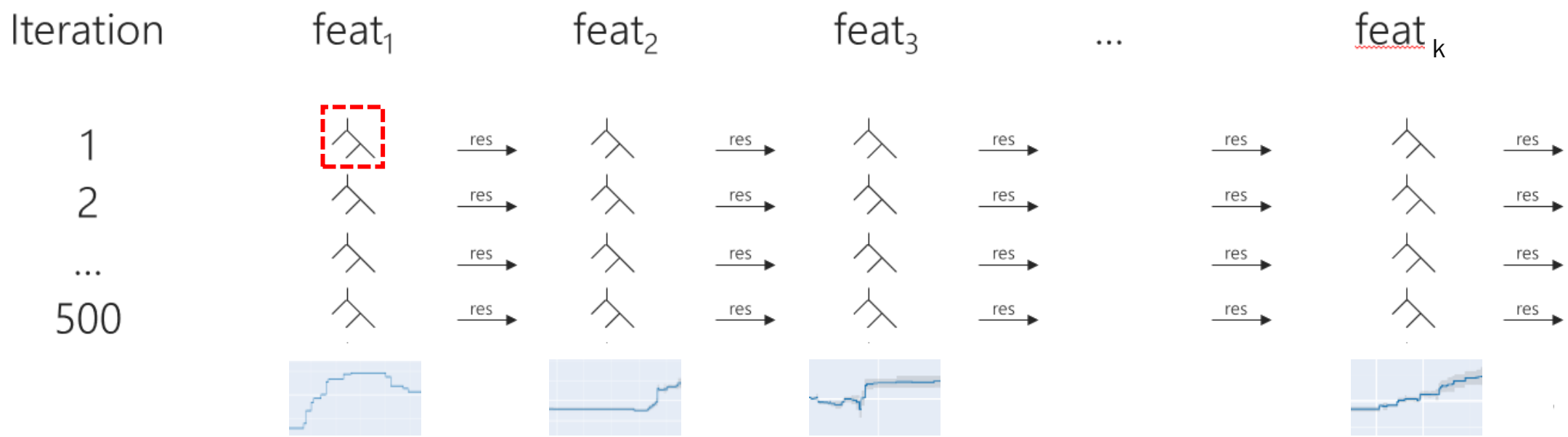
Differentially Private Explainable Boosting Machines

SENSITIVITY

Composition

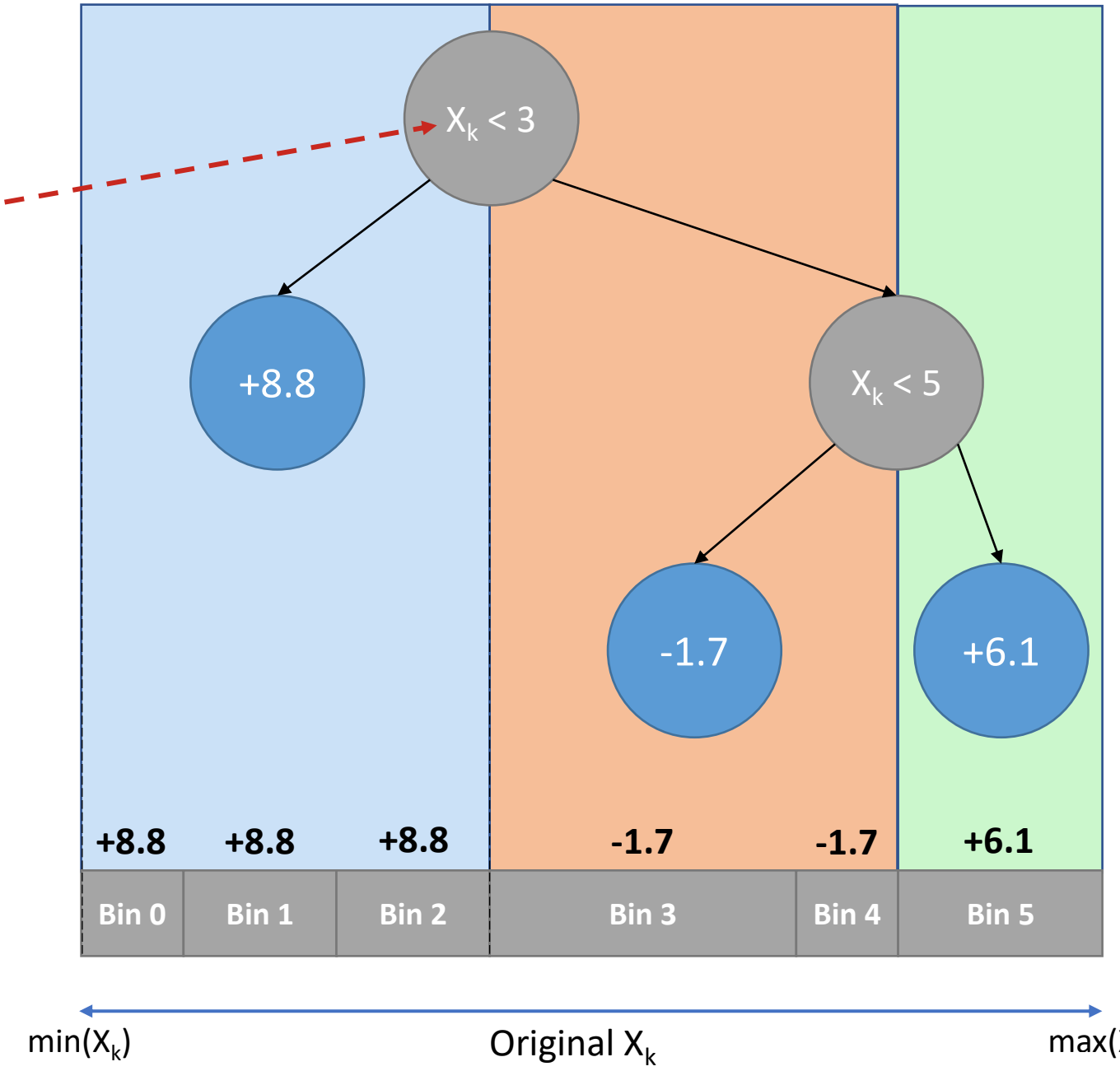
$$\Delta f = \max |f(x) - f(y)|$$

EBM: GAM with Gradient Boosted Decision Trees



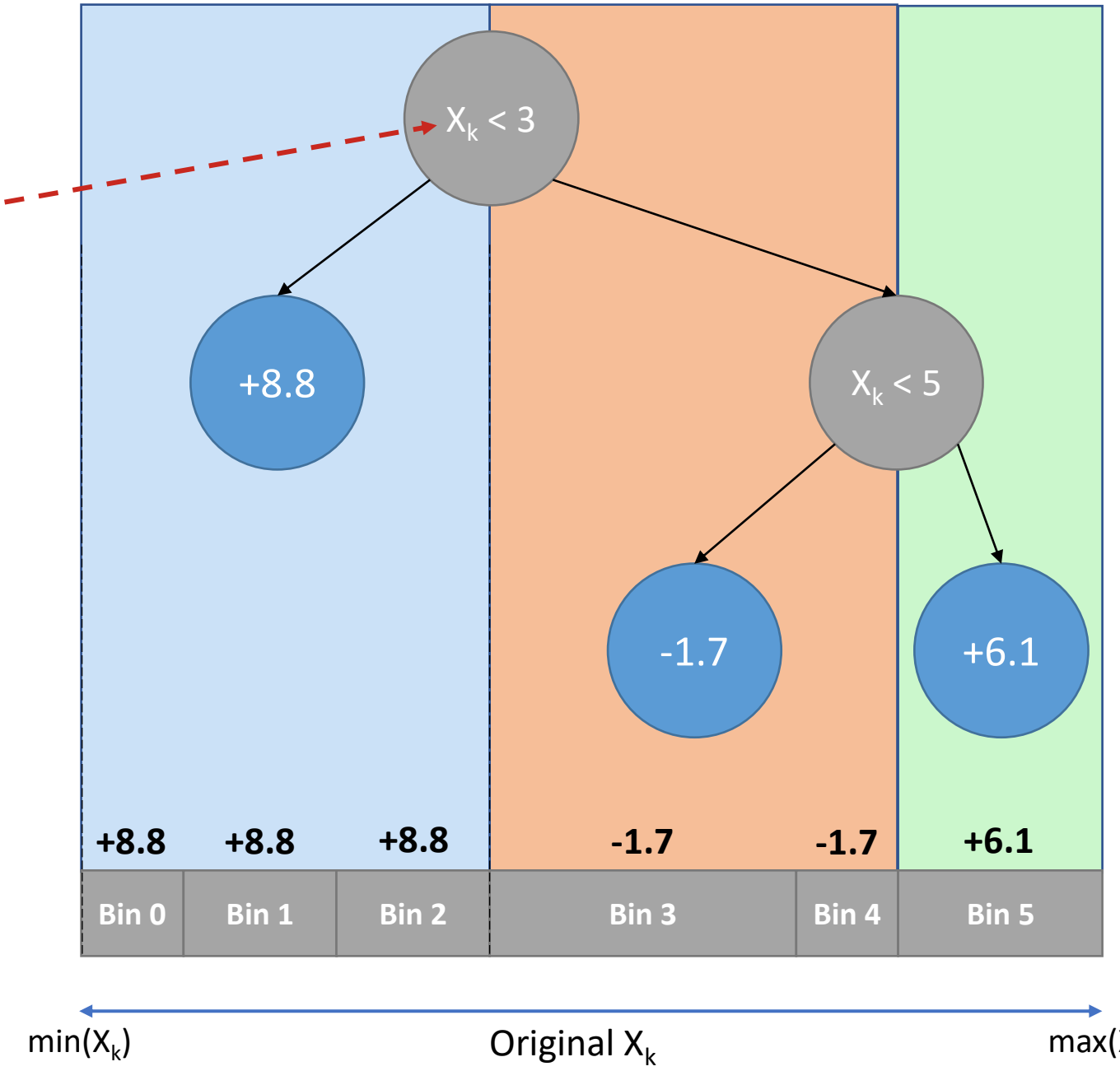
DP-EBM: Single Tree

- Sources of Privacy Leakage:
  - What feature to use at each node?



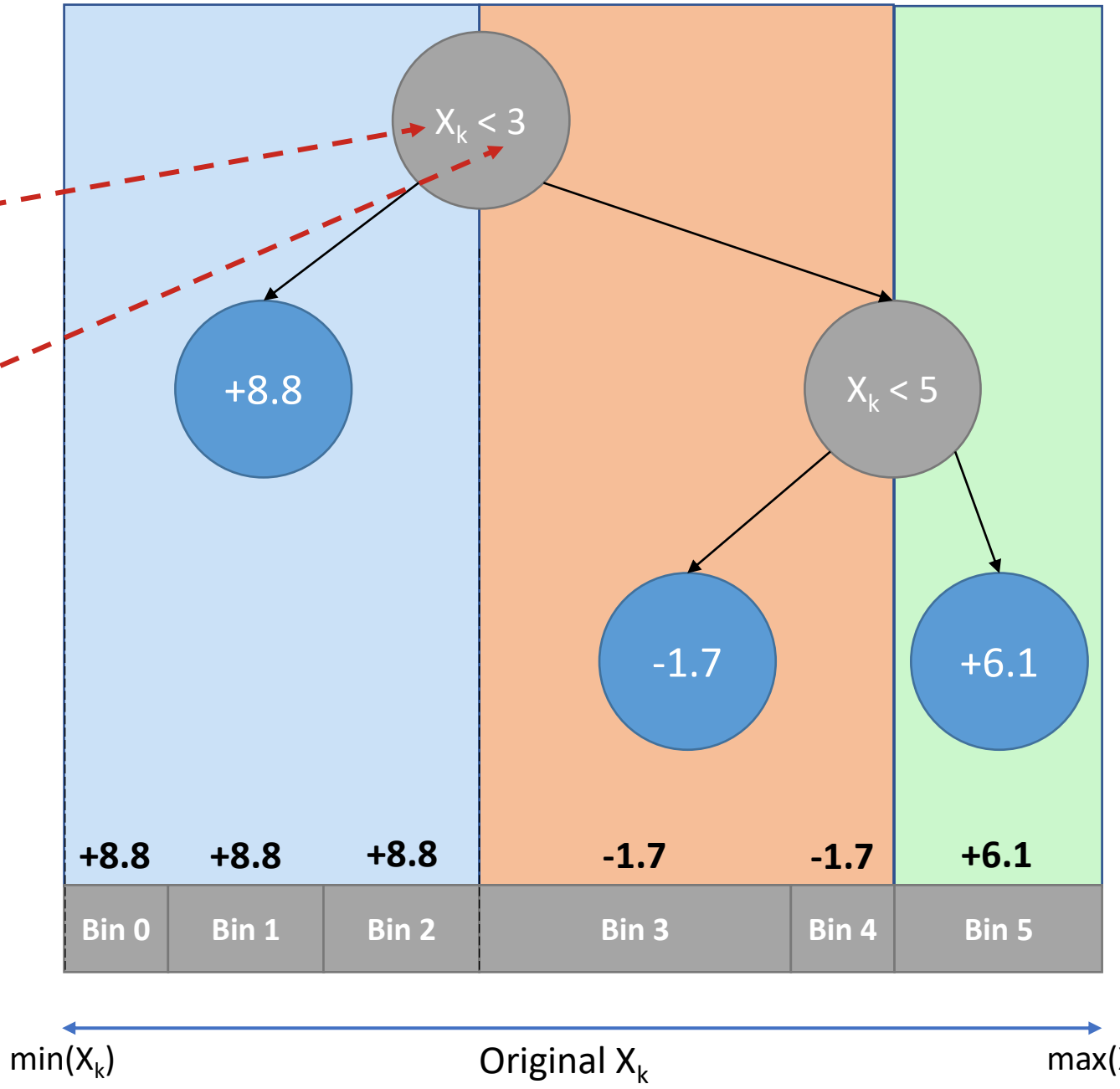
DP-EBM: Single Tree

- Sources of Privacy Leakage:
  - What feature to use at each node?
    - Trad: Search all features
    - EBM: Visit on a schedule ✓



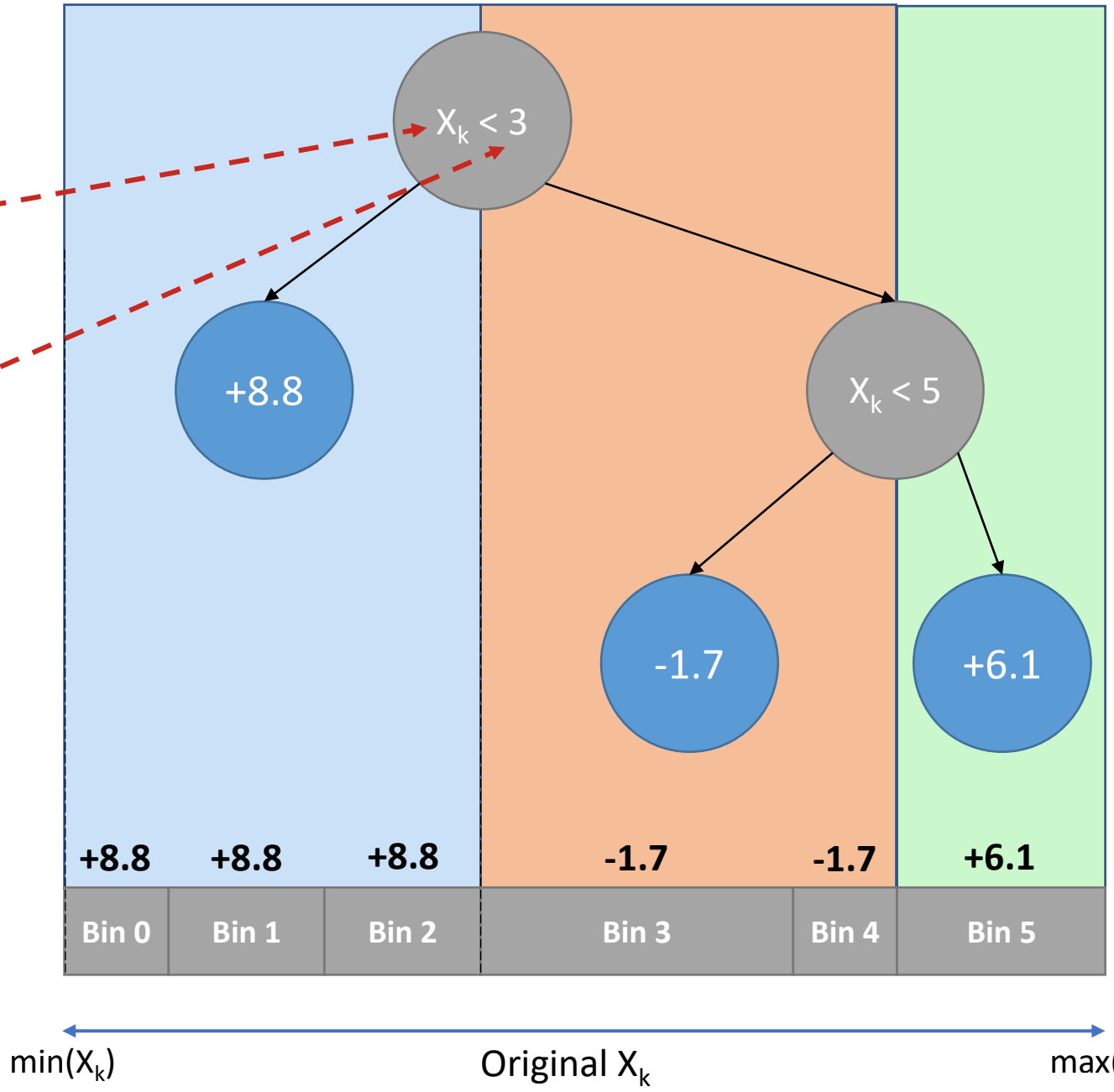
DP-EBM: Single Tree

- Sources of Privacy Leakage:
  - What feature to use at each node?
    - Trad: Search all features
    - EBM: Visit on a schedule ✓
  - What threshold to select at each node?



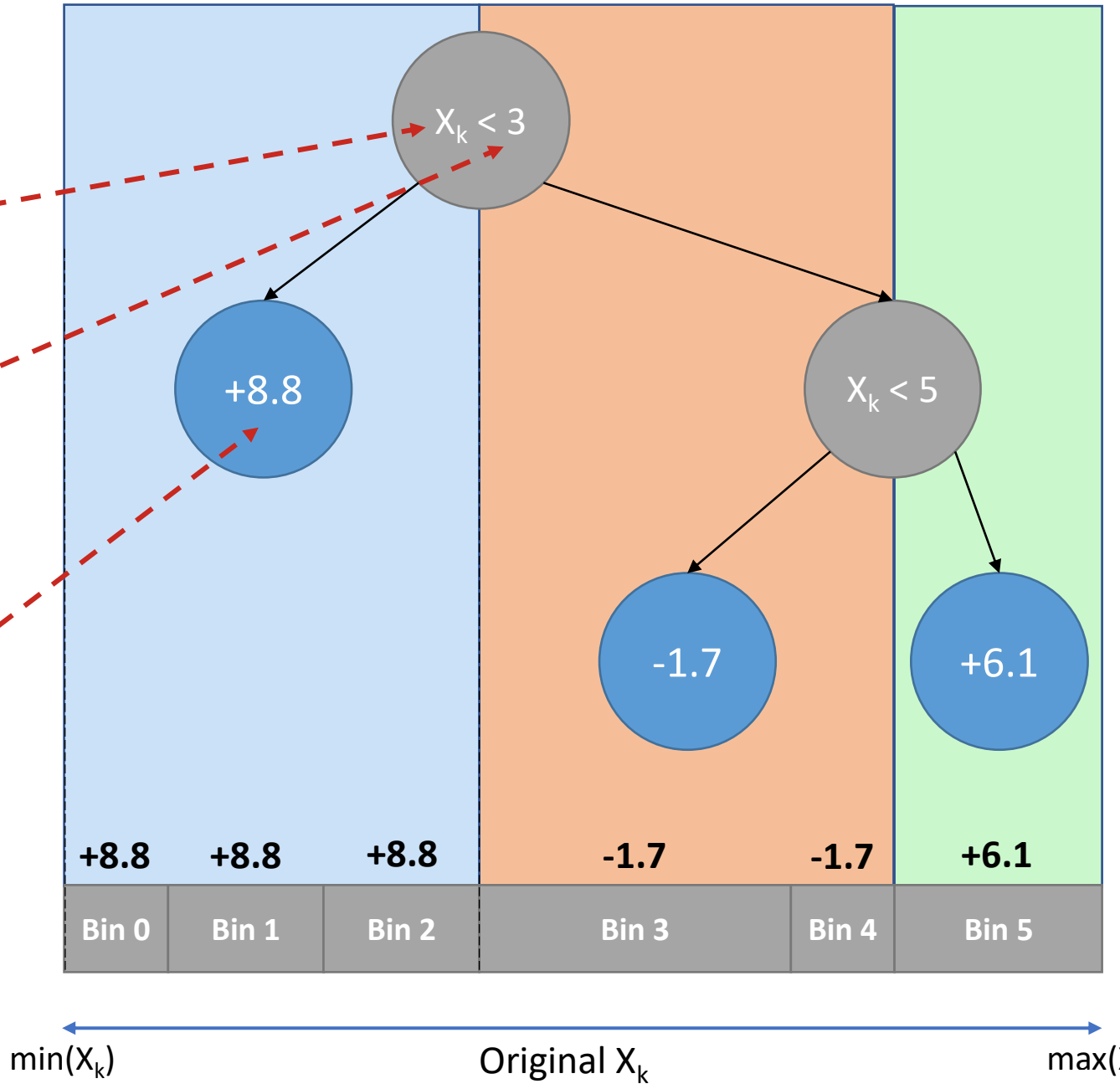
DP-EBM: Single Tree

- Sources of Privacy Leakage:
  - What feature to use at each node?
    - Trad: Search all features
    - EBM: Visit on a schedule ✓
  - What threshold to select at each node?
    - Trad: Search all split points
    - EBM: Choose ***randomly*** ✓



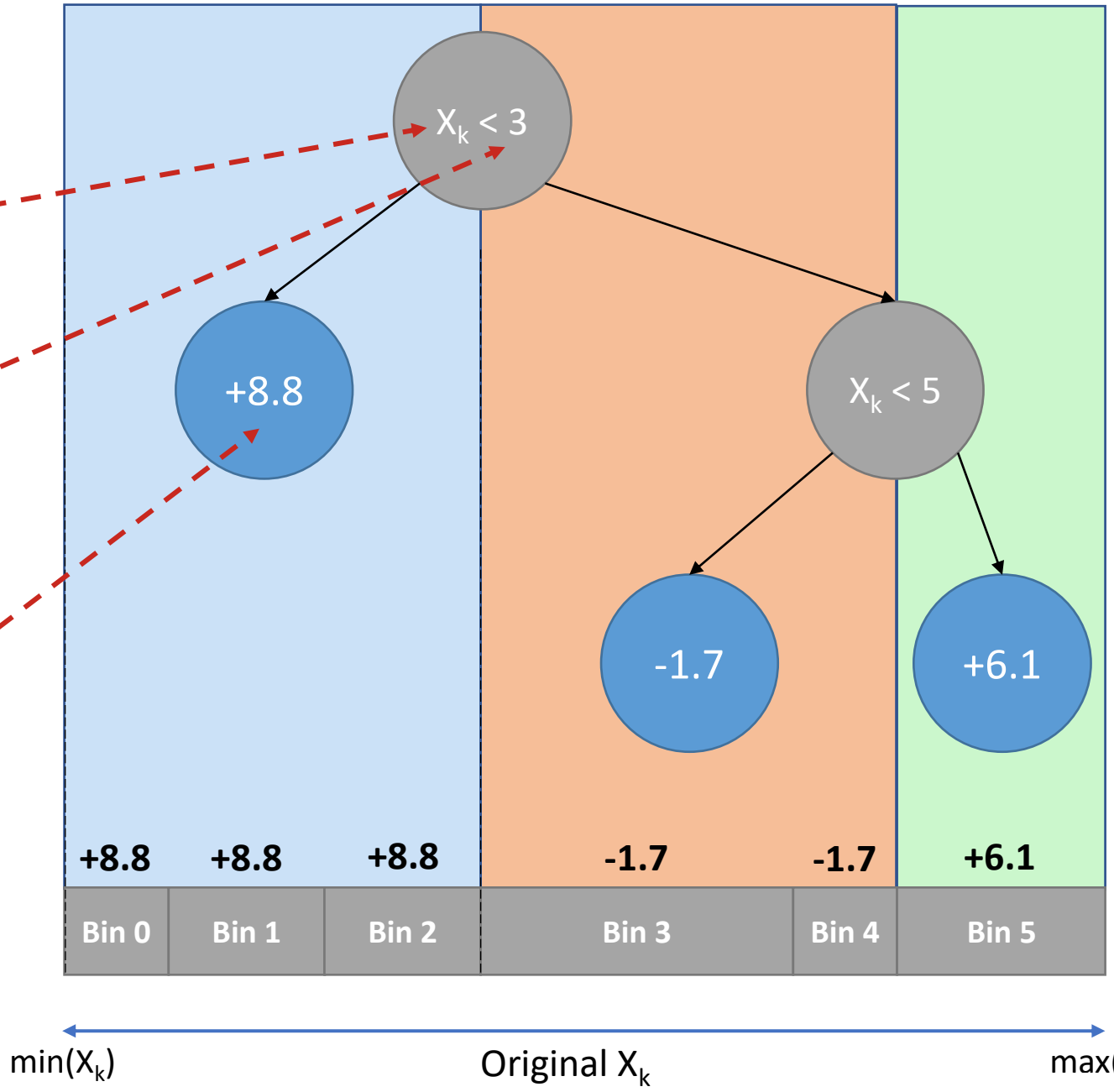
DP-EBM: Single Tree

- Sources of Privacy Leakage:
  - What feature to use at each node?
    - Trad: Search all features
    - EBM: Visit on a schedule ✓
  - What threshold to select at each node?
    - Trad: Search all split points
    - EBM: Choose ***randomly*** ✓
  - What value to learn for leaf nodes?



DP-EBM: Single Tree

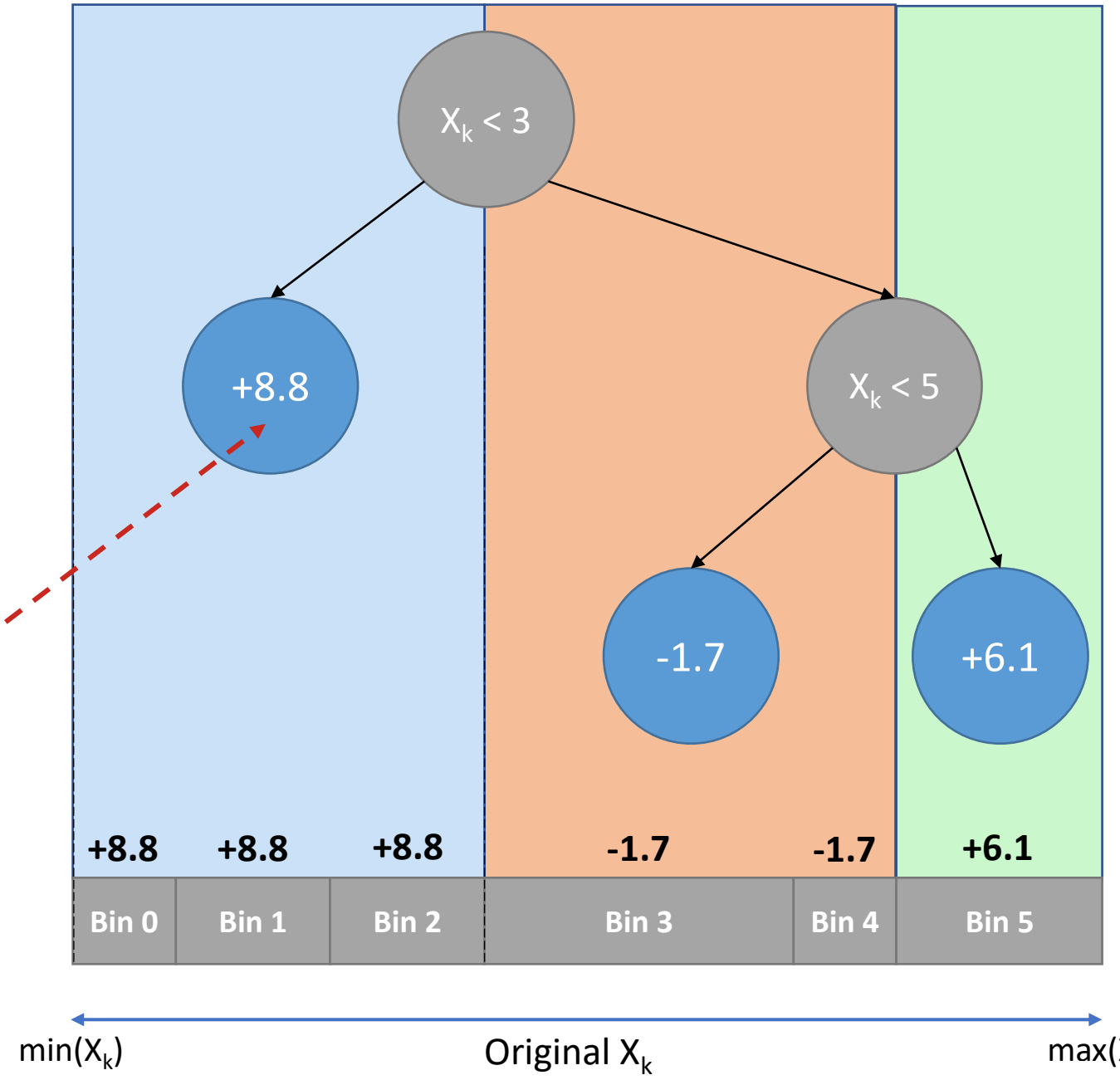
- Sources of Privacy Leakage:
  - What feature to use at each node?
    - Trad: Search all features
    - EBM: Visit on a schedule ✓
  - What threshold to select at each node?
    - Trad: Search all split points
    - EBM: Choose ***randomly*** ✓
  - What value to learn for leaf nodes?
    - Trad: Average of data in leaf node
    - EBM: Private Average ✓



DP-EBM: Single Tree

- Sources of Privacy Leakage:
  - What feature to use at each node?
    - Trad: Sea[illegible] features
    - EBM: Visit[illegible] schedule
  - What threshold to select at each node?
    - Trad: Sea[illegible] split points
    - EBM: Cho[illegible]ndomly
  - What value to learn for leaf nodes?
    - Trad: Average of data in leaf node
    - EBM: Private Average ✓

- **DP-EBMs: 100% of budget** is used on leaf node values

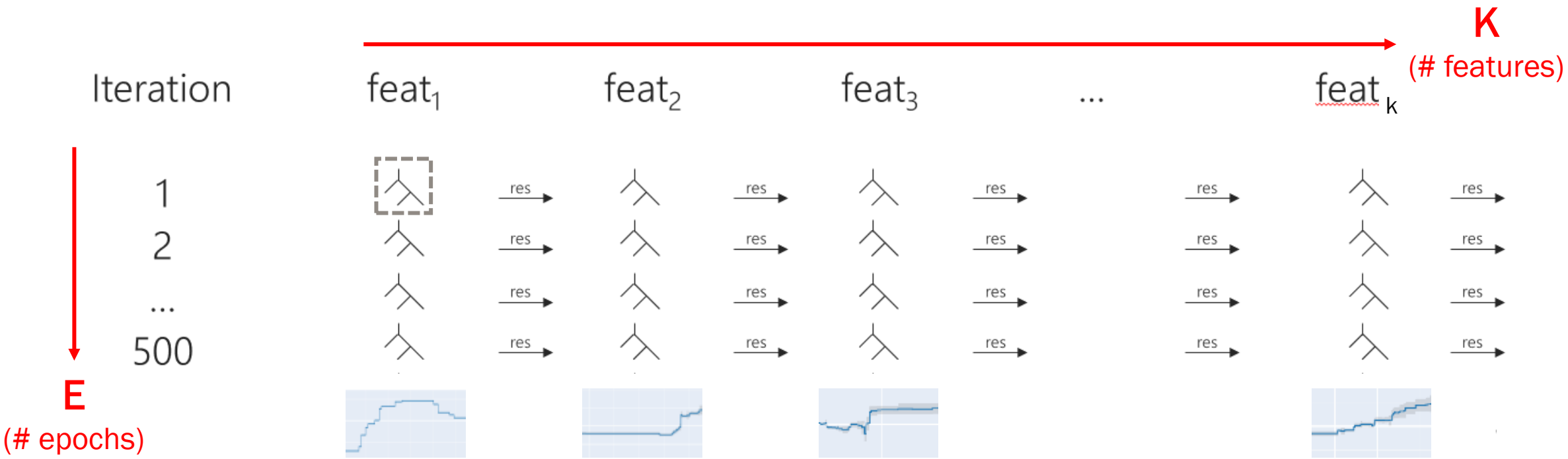


Explainable Boosting Machines

SENSITIVITY

Composition

EBM: GAM with Gradient Boosted Decision Trees

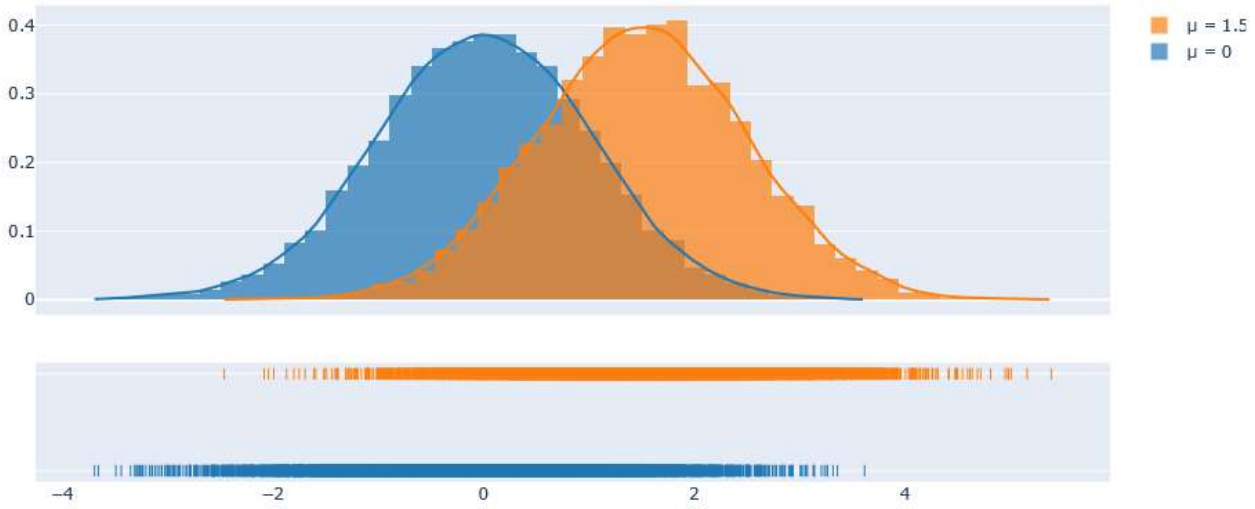


Gaussian Differential Privacy[1]

- Builds on hypothesis testing interpretation of differential privacy.
  - H0 : underlying dataset = D1
  - H1 : underlying dataset = D2
  - The upper bound of power for a test at any for significance level $0 < \alpha < 1$ is: $e^{\varepsilon}\alpha + \delta$

- However, (ε, δ)-DP is under-parametrized; with composition, upper bound of testing power is not tight.

- GDP introduces single parameter hypothesis testing interpretation: μ-GDP.
  - Uses type I and type II tradeoffs to distinguish $N(0, 1)$ from $N(\mu, 1)$.



Leaking privacy from a (1.5)-GDP algorithm is at least as hard as distinguishing between these two distributions with one draw.

Theorem 4. *The k-fold composition of μ_i-GDP mechanisms is $\sqrt{\mu_1^2 + \mu_2^2 + \ldots \mu_k^2}$-GDP.*

Finally, one can convert GDP guarantees to the standard (ϵ, δ)-DP guarantee using the following theorem:

Theorem 5. *A mechanism is μ-GDP if and only if it is (ε, δ)-DP where*

$$\delta = \phi(-\frac{\varepsilon}{\mu} + \frac{\mu}{2}) - e^{\varepsilon}\phi(-\frac{\varepsilon}{\mu} - \frac{\mu}{2})$$

[1] Dong, Jinshuo, Aaron Roth, and Weijie J. Su. "Gaussian differential privacy." arXiv preprint arXiv:1905.02383 (2019).

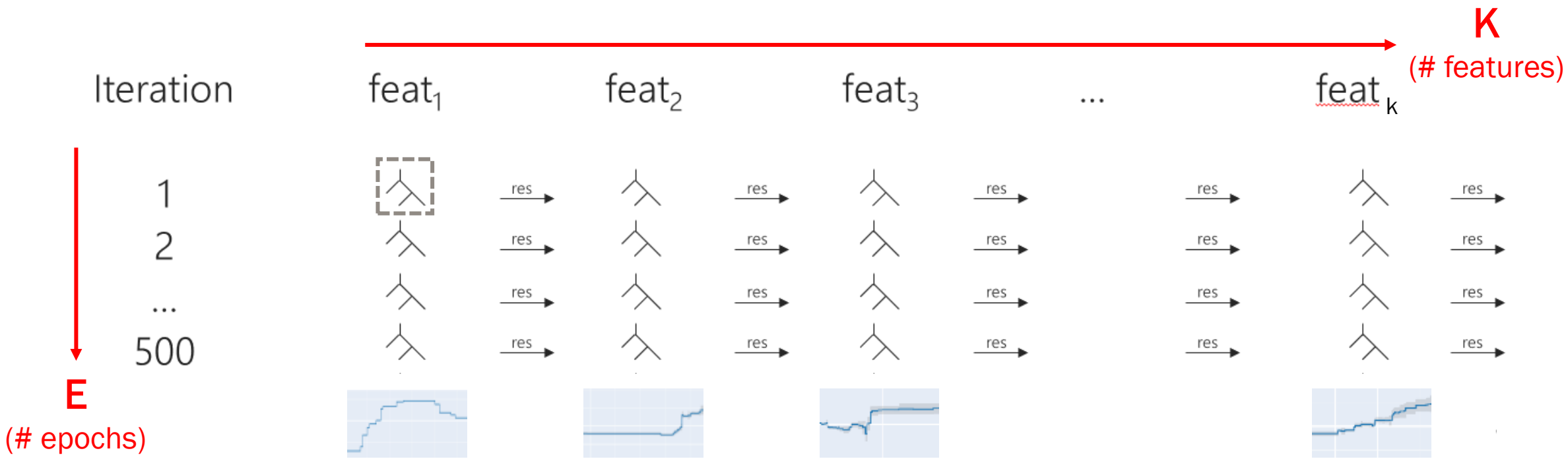
Explainable Boosting Machines

SENSITIVITY

Composition

Corollary 3.3. *The n-fold composition of μ_i-GDP mechanisms is $\sqrt{\mu_1^2 + \cdots + \mu_n^2}$-GDP.*

EBM: GAM with Gradient Boosted Decision Trees



What does privacy do to models?

Differential Privacy Adds Noise and Unwanted Bias to Models

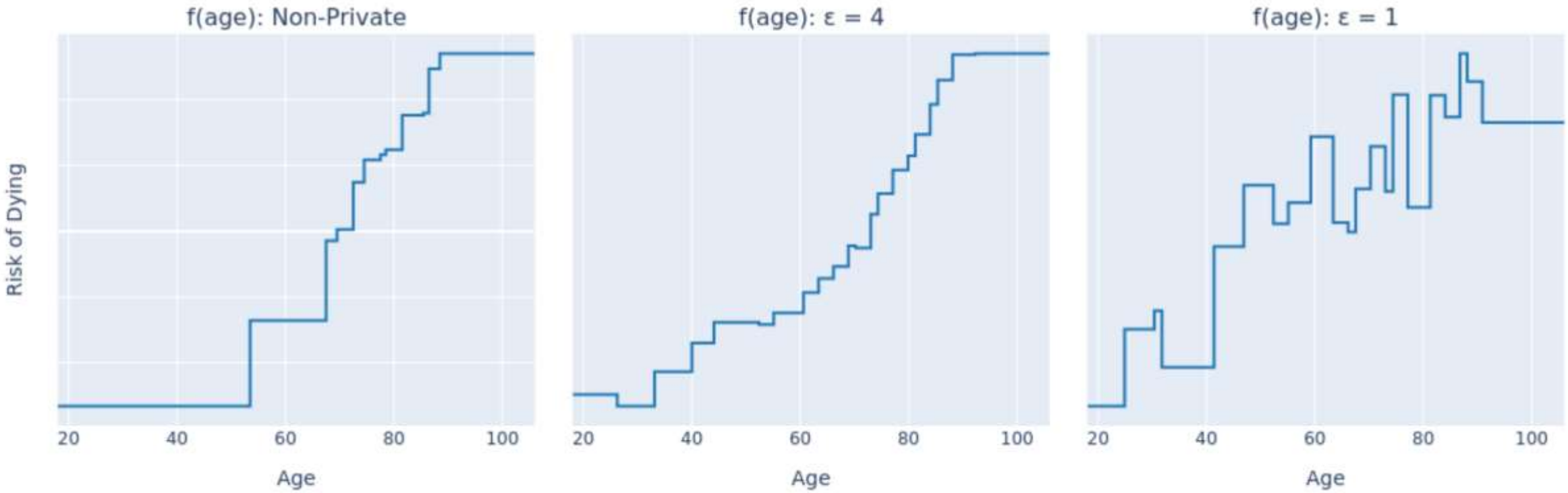


Figure 3. Risk of dying as a function of age from three EBMs trained on the healthcare dataset with varying privacy guarantees.

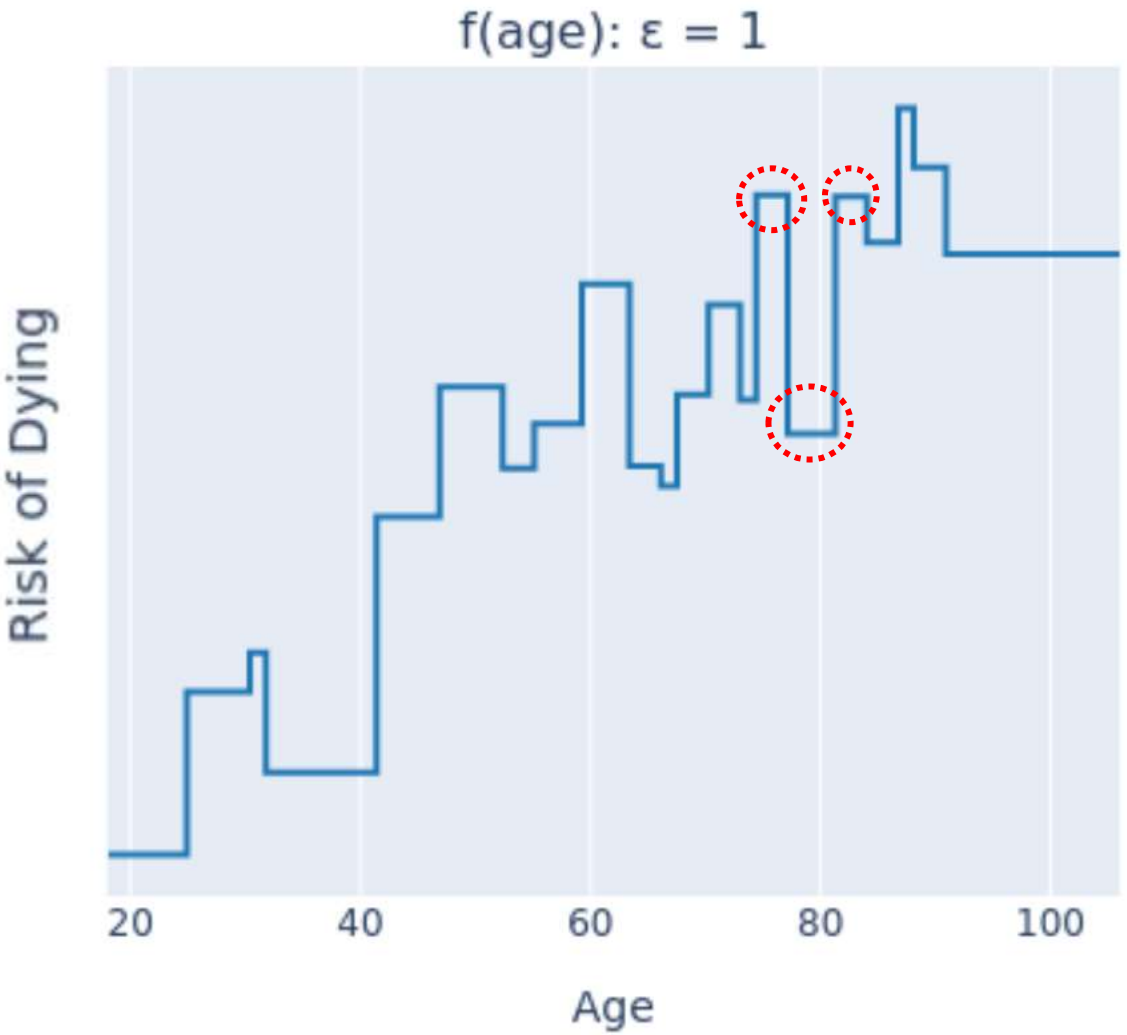
Editing Unwanted Bias

- Differential privacy can introduce noise and unwanted bias
  - Is 80 less risky than 77 and 82?

- Bias will impact minorities more
  - Impossibility Results in Fairness + DP: [Cummings, Gupta, Kimpara, Morgenstern]

    "We show that it is impossible to achieve both differential privacy and exact fairness while maintaining non-trivial accuracy"

  - Intuitively makes sense – need more noise to protect smaller populations.



Editing Unwanted Bias

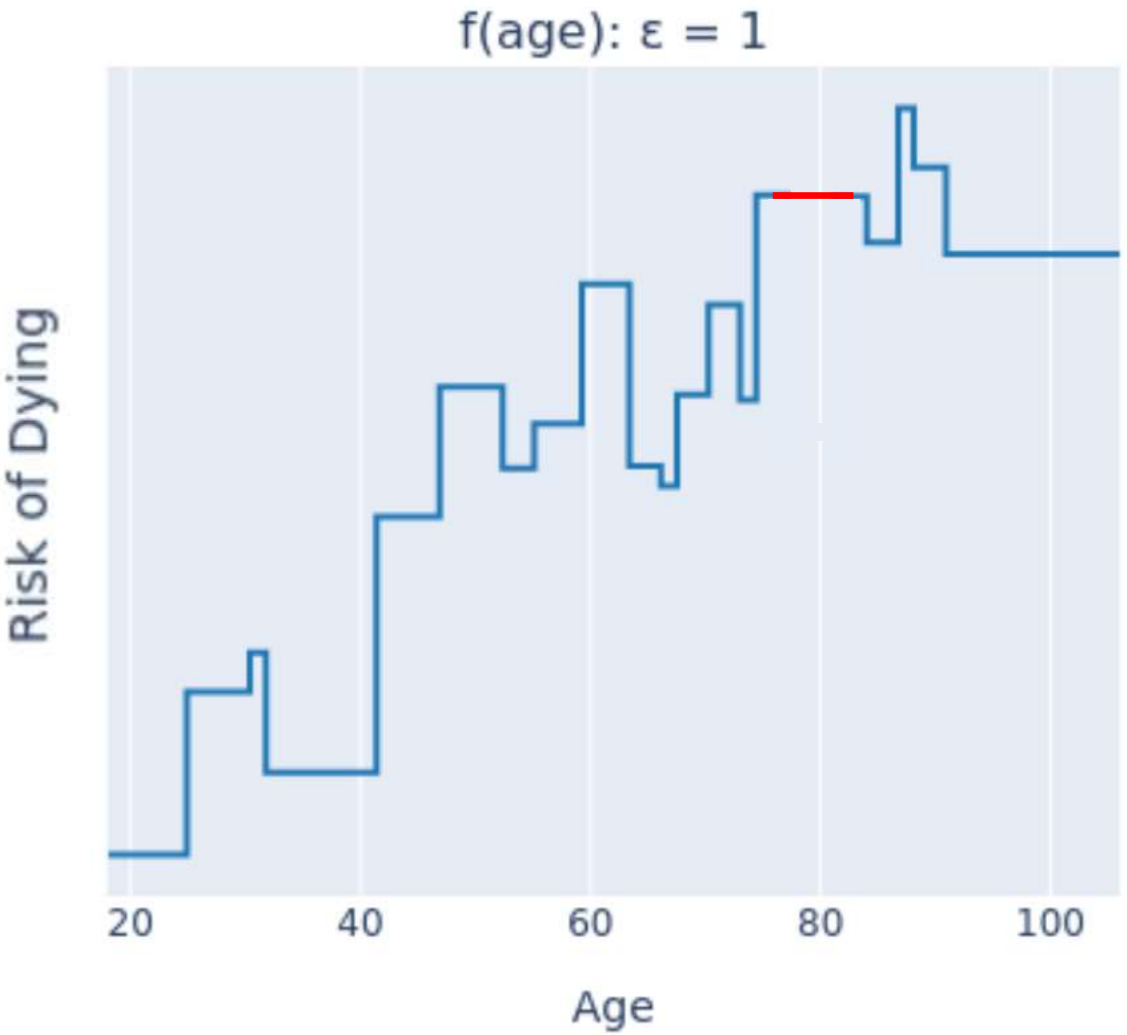
- Differential privacy can introduce noise and unwanted bias
  - Is 80 less risky than 77 and 82?

- Bias will impact minorities more
  - Impossibility Results in Fairness + DP: [Cummings, Gupta, Kimpara, Morgenstern]

    "We show that it is impossible to achieve both differential privacy and exact fairness while maintaining non-trivial accuracy"

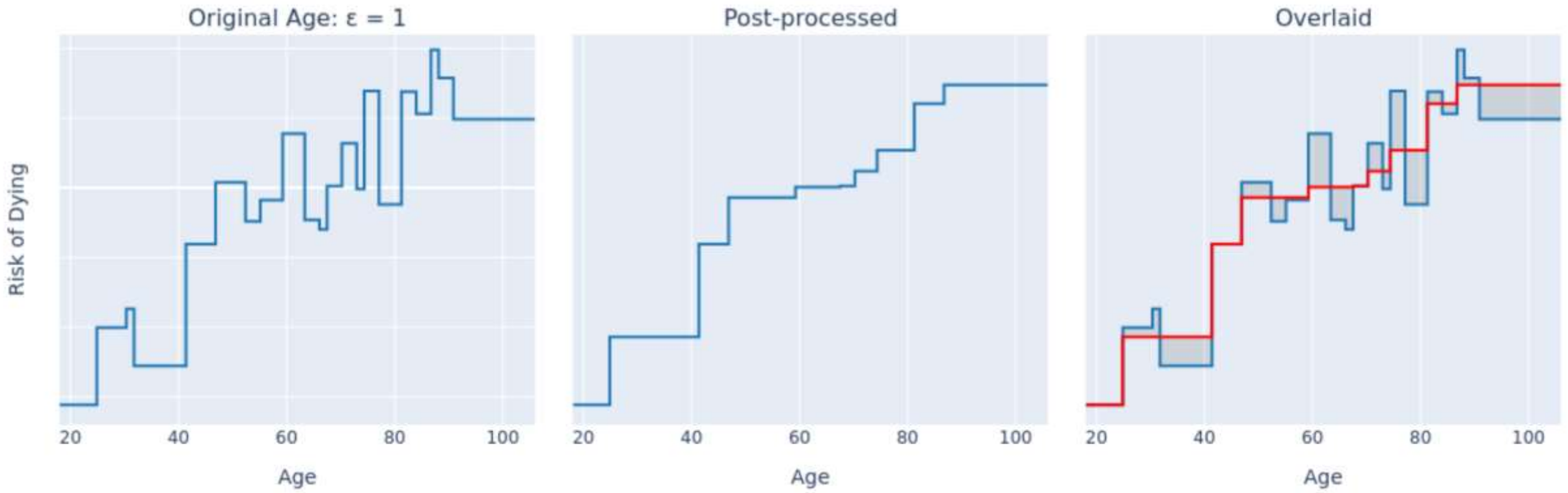
  - Intuitively makes sense – need more noise to protect smaller populations.

We can fix this!



Monotonicity for Free

Optimal Monotonicity via Postprocessing: Pool Adjacent Violators Algorithm (PAV)



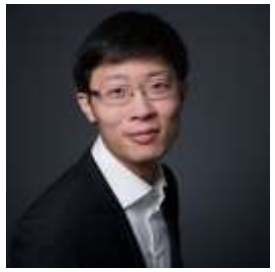
Privacy <3 Interpretability

- DP-EBMs are **the** most accurate algorithm for privacy-preserving machine learning on tabular data.

- Interpretability lets us detect and mitigate unwanted artifacts of noise before deployment.

Differential Privacy and Interpretability in Causal Modeling

CLEAR '22: Oral Presentation



Fengshi Niu

Stanford

Harsha Nori

MSR Aether

Brian Quistorff

Bureau of Economic Analysis

Rich Caruana

MSR AI

Donald Ngwe

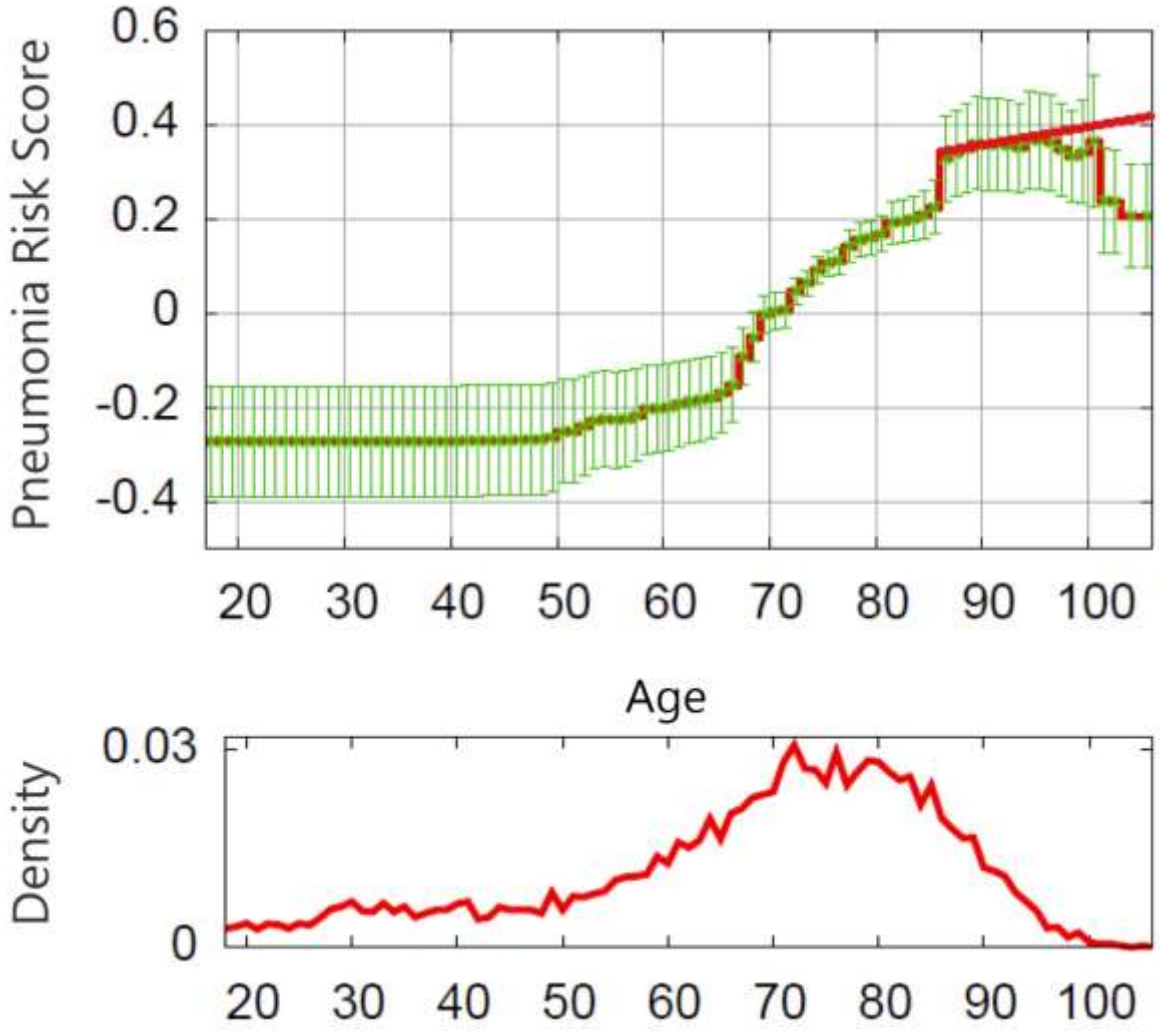
Office of Chief Economist

Aadharsh Kannan

Office of Chief Economist

Editing Glass-Box Models

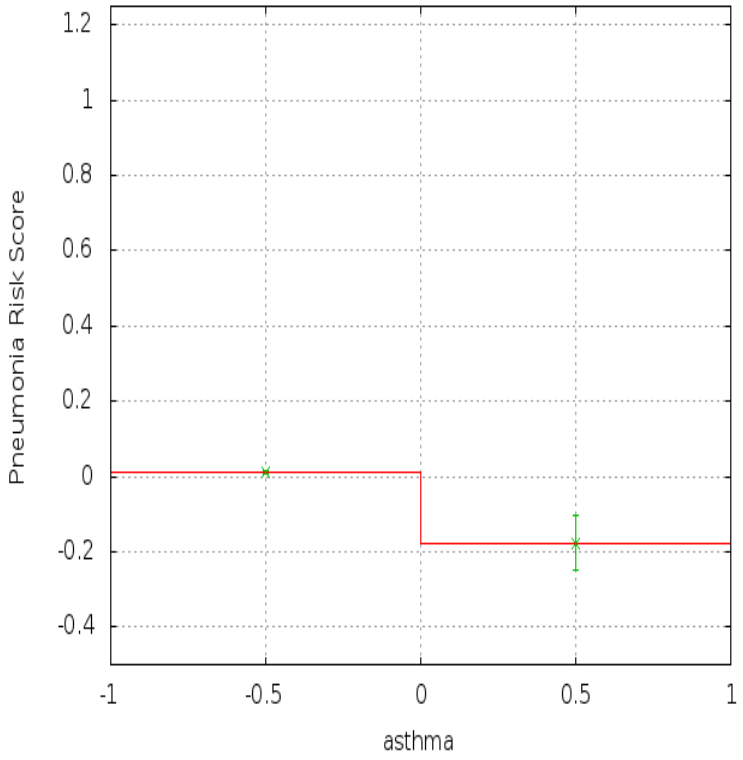
Fix Age > 100 Problem (Enforce Monotonicity)



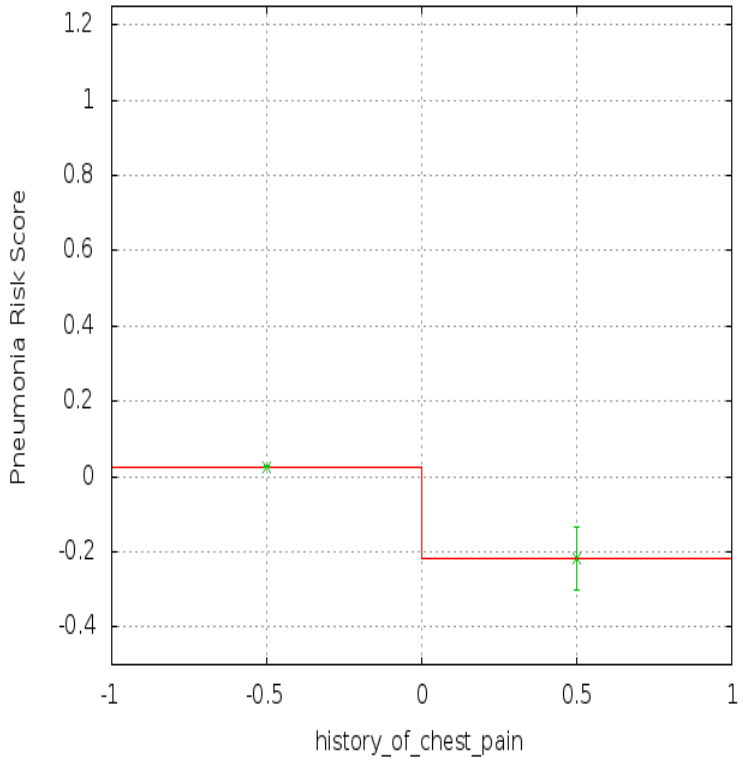
Surprising Statistical "Facts" About Pneumonia

Asthma, Chest Pain, Heart Disease, and Stridor Good for Your Pneumonia!

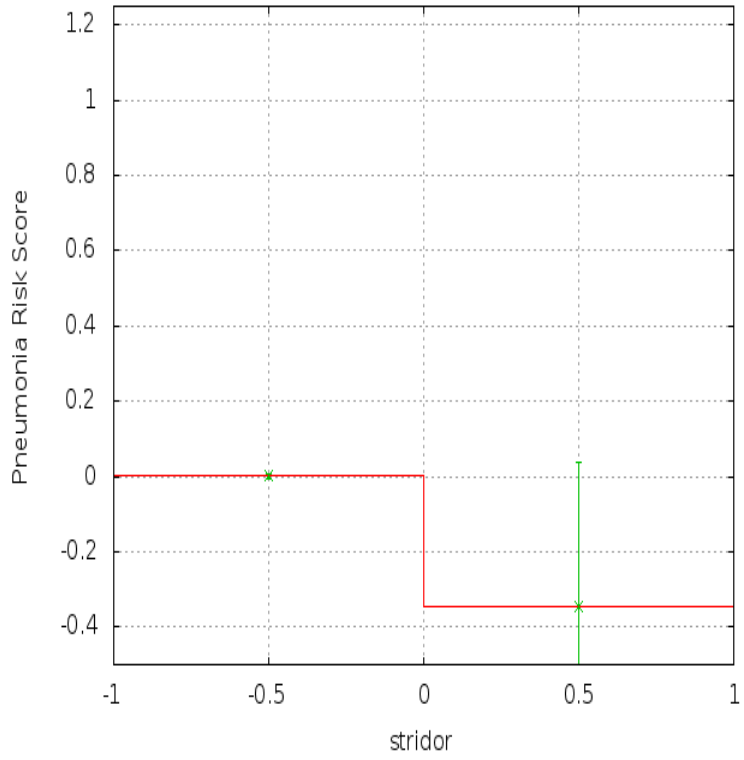
ASTHMA



CHEST PAIN



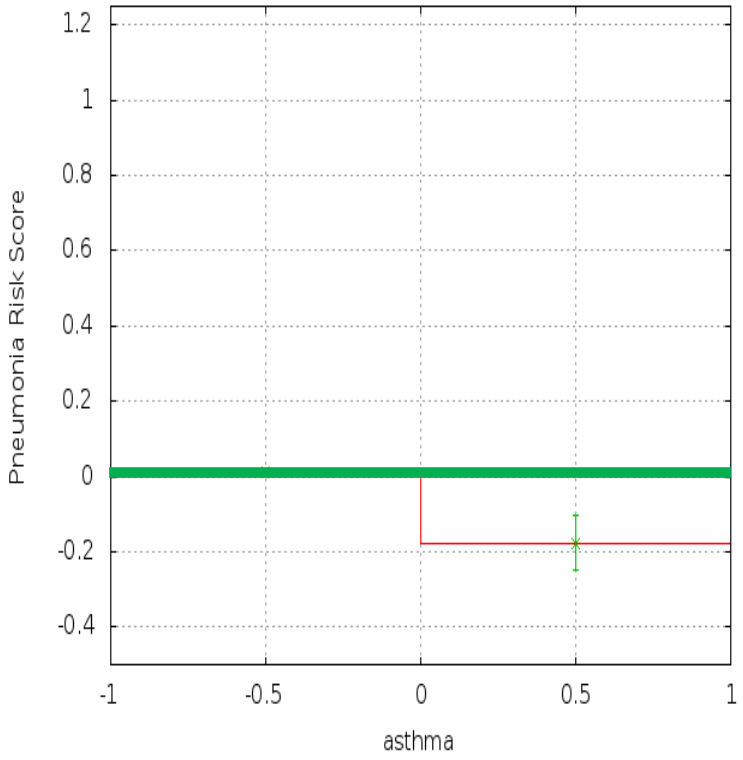
OBSTRUCTED AIRWAY



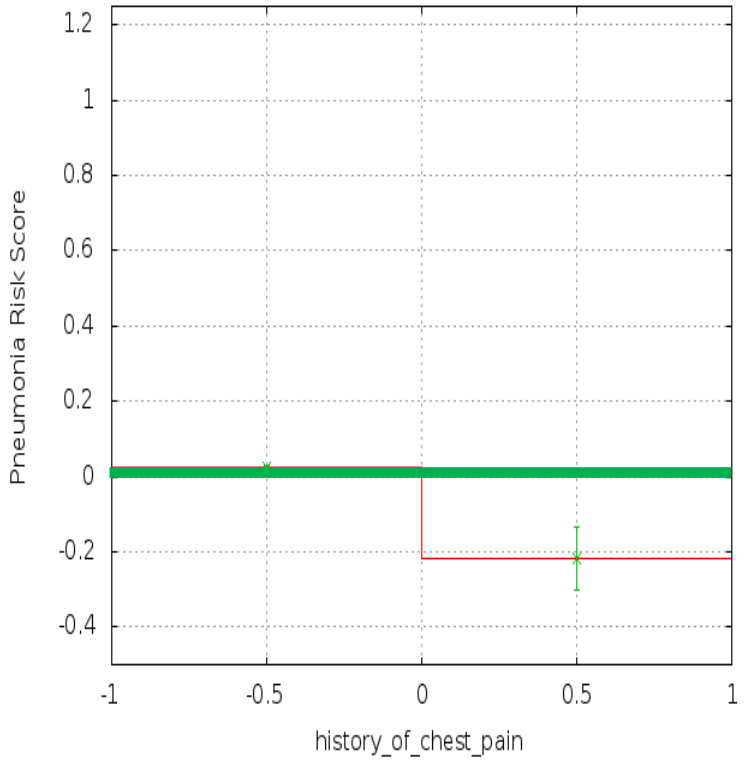
Surprising Statistical "Facts" About Pneumonia

Asthma, Chest Pain, Heart Disease, and Stridor Good for Your Pneumonia!

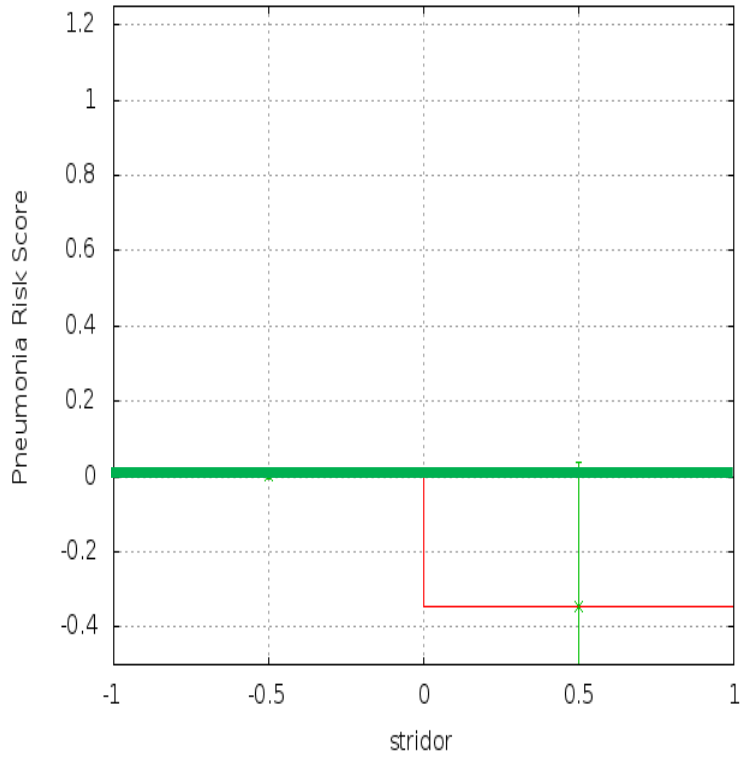
ASTHMA



CHEST PAIN



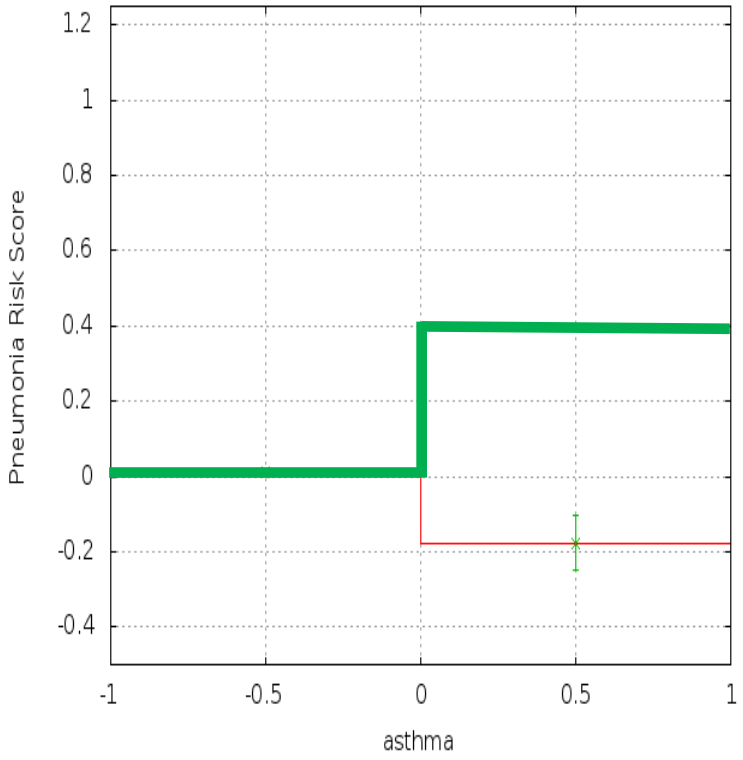
OBSTRUCTED AIRWAY



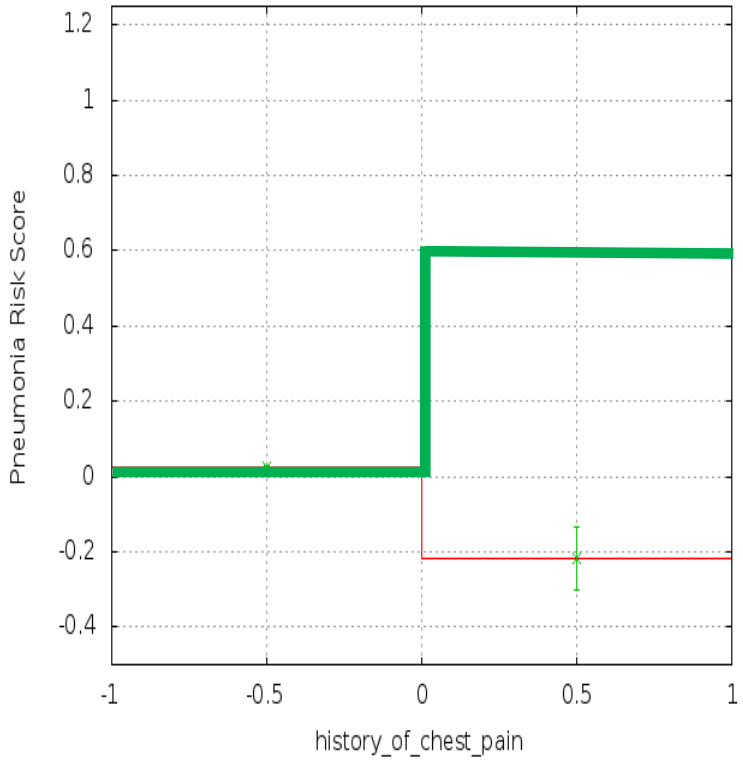
Surprising Statistical "Facts" About Pneumonia

Asthma, Chest Pain, Heart Disease, and Stridor Good for Your Pneumonia!

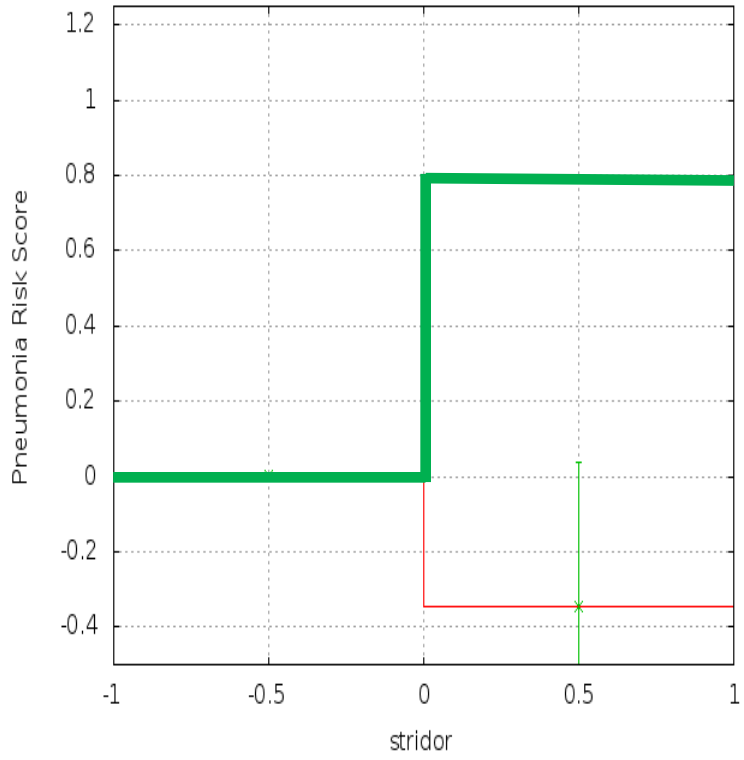
ASTHMA



CHEST PAIN



OBSTRUCTED AIRWAY



Housing Pricing Data

Pneumonia Dataset: Heart Rate (Pulse)

- Model sees no data 40-125
- Model interpolates between HR=39 and HR=126
- Would yield bad predictions for normal patients if HR collected!
- Can edit EBM graph to repair

Align ML Model Behaviors with Human Users' Knowledge

GAM CHANGER

NeurIPS Research2Clinics Workshop: Best Paper Award

Jay Wang
Georgia Tech

Alex Kale
University of Washington

Harsha Nori
Microsoft

Peter Stella
NYU Langone Health

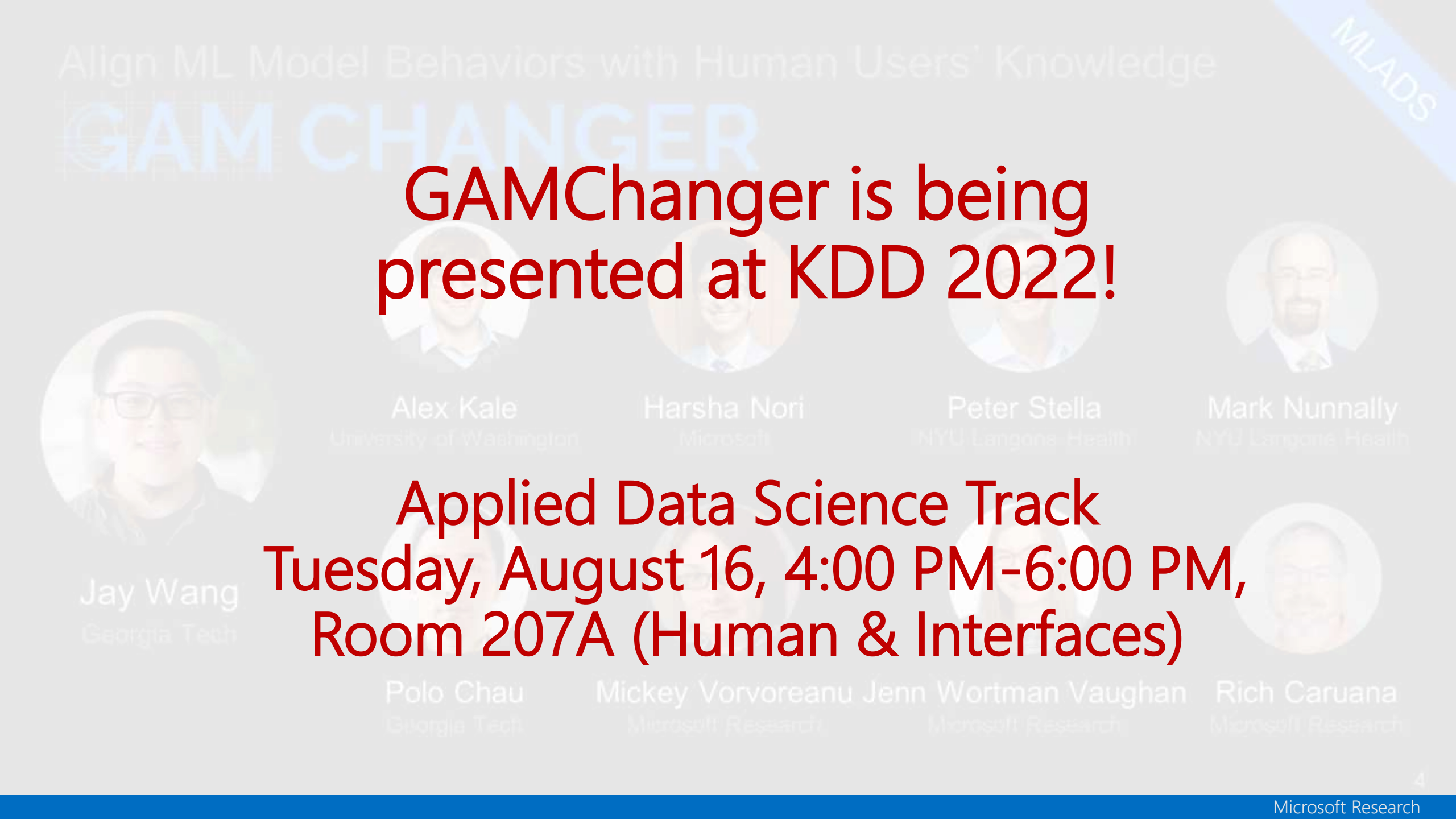
Mark Nunnally
NYU Langone Health

Polo Chau
Georgia Tech

Mickey Vorvoreanu
Microsoft Research

Jenn Wortman Vaughan
Microsoft Research

Rich Caruana
Microsoft Research



GAMChanger is being presented at KDD 2022!

Applied Data Science Track
Tuesday, August 16, 4:00 PM-6:00 PM, Room 207A (Human & Interfaces)

Hands-On Section:

https://github.com/interpretml/kdd2022-tutorial

Summary

- Every dataset has flaws
  - Every time we apply glass-box ML to a new dataset we find these kinds of problems
  - High accuracy not sufficient --- models are rewarded with high accuracy for predicting wrong things
  - Without intelligibility and explanation you're flying blind --- that's dangerous!

- Glass-Box ML models like EBMs and NAMs give you the tools to need:
  - To understand, vet and edit your model before deploying it
  - Learn from your data to improve engineering and science
  - Better deal with missing values and imputation
  - Detect and mitigate problems with bias and fairness

- EBMs & NAMs are currently the most accurate glass-box learning methods available
  - Easy to use open-source package: github.com/interpretml/interpret
  - Can now train glass-box EBM models just as easily as XGBoost, GBT, RF, …

InterpretML

Open-Source Tool for Intelligibility

github.com/interpretml/interpret 

Thank You!

Do's and Don'ts for EBMs

- Don't do feature selection --- first train EBM model on all available features
- Don't do feature engineering --- first train EBM model using raw features
- Don't impute missing values --- first train EBM model using unique codes for missing

- Do compare accuracy of EBM to other blackbox models such as DNN, GBT, and RF
- Do look at graphs --- there's gold (and secrets) hidden in those graphs
- Do detective work to understand anomalies --- data scientists + domain experts
- Do fix problems --- either edit graphs, clean data, or get new data
- Do compare graphs trained on this data to graphs from other data (other years, …)

- Use Black-Box Explanation (LIME, SHAP, Partial Dependence, …) When:
  - You don't have access to the training data
  - Or model was pre-trained and given to you
  - Or a specific black-box model was required (neural net, boosted trees, random forests, …)
  - Or you're trying to understand a complex pipeline from beginning to end
  - ➢ **Must use black-box explanation methods**

- Use Black-Box Explanation (LIME, SHAP, Partial Dependence, …) When:
  - You don't have access to the training data
  - Or model was pre-trained and given to you
  - Or a specific black-box model was required (neural net, boosted trees, random forests, …)
  - Or you're trying to understand a complex pipeline from beginning to end
  - ➢ Must use black-box explanation methods

- But Use Glass-Box Machine Learning (EBM: Explainable Boosting Machine) When:
  - You have access to the training data and you're the one training the model
  - You're the one who needs to debug the model, retrain the model, improve model accuracy, …
  - ➢ Should use Glass-Box ML methods such as EBMs
    - Exact intelligibility, not approximate as with black-box explanation methods
    - Better intelligibility leads to faster debugging and model development/improvement
    - Models are editable to correct bias and errors

Thank You!