Microsoft Research

Explanation & Interpretability:

InterpretML: Explainable Boosting Machines (EBMs)

Rich Caruana

Microsoft

Microsoft Research

Explanation & Interpretability:

InterpretML: Explainable Boosting Machines (EBMs)

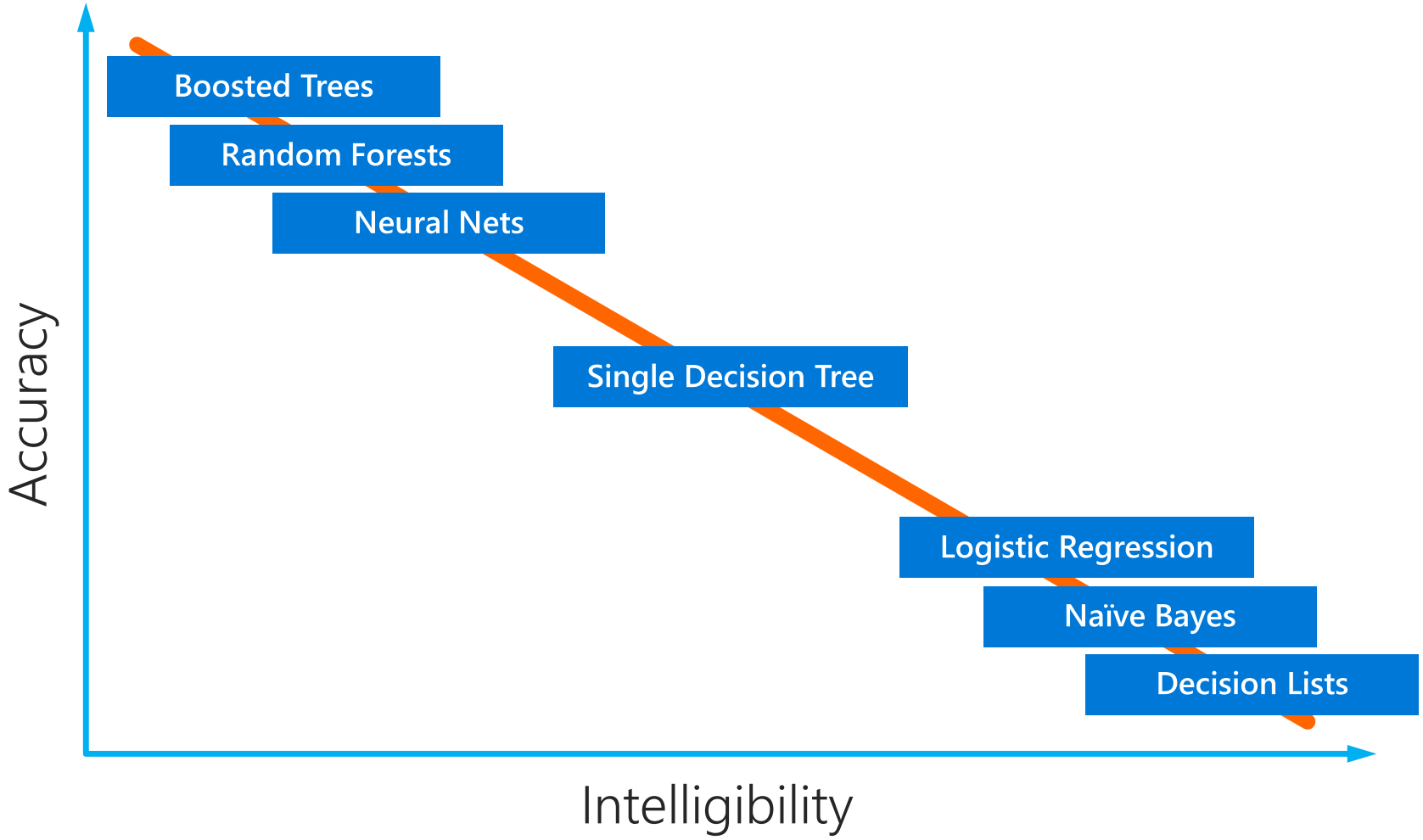
Rich Caruana
Yin Lou, Sarah Tan, Xuezhou Zhang, Ben Lengerich, Kingsley Chang, Paul Koch, Harsha Nori, Sam Jenkins, Giles Hooker, Johannes Gehrke, Tom Mitchell, Greg Cooper MD PhD, Mike Fine MD, Eric Horvitz MD PhD, Vivienne Souter MD, Nick Craswell, Marc Sturm, Noemie Elhadad, Jacob Bien, Noah Snavely

Microsoft

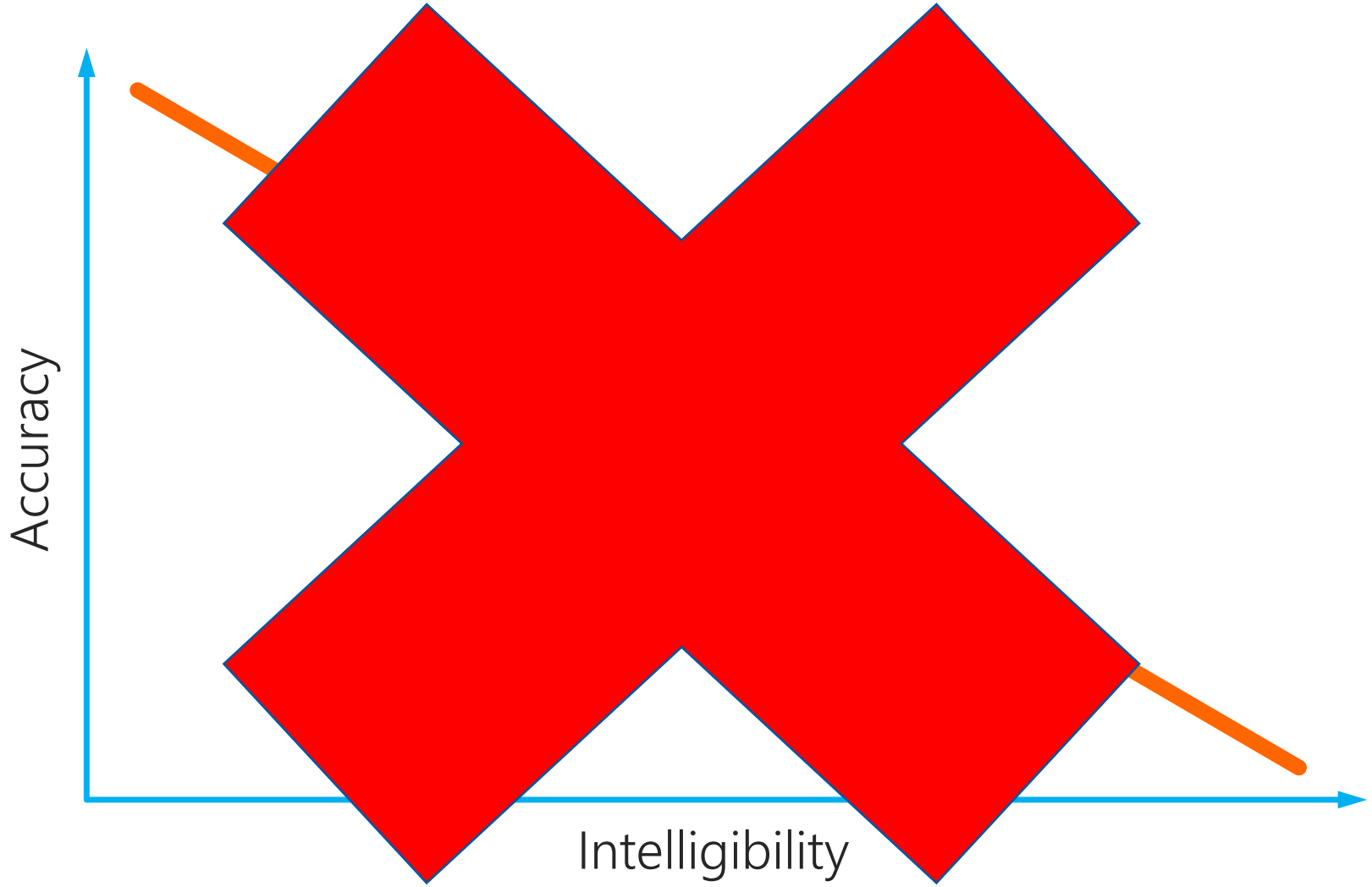
- Use Black-Box Explanation (LIME, SHAP, Partial Dependence, …) When:
  - You don't have access to the training data
  - Or model was pre-trained and given to you
  - Or a specific black-box model was required (neural net, boosted trees, random forests, …)
  - Or you're trying to understand a complex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

Accuracy vs. Intelligibility Tradeoff ???



Accuracy vs. Intelligibility Tradeoff --- Often Not True!



Accuracy vs. Intelligibility Tradeoff --- Often Not True!

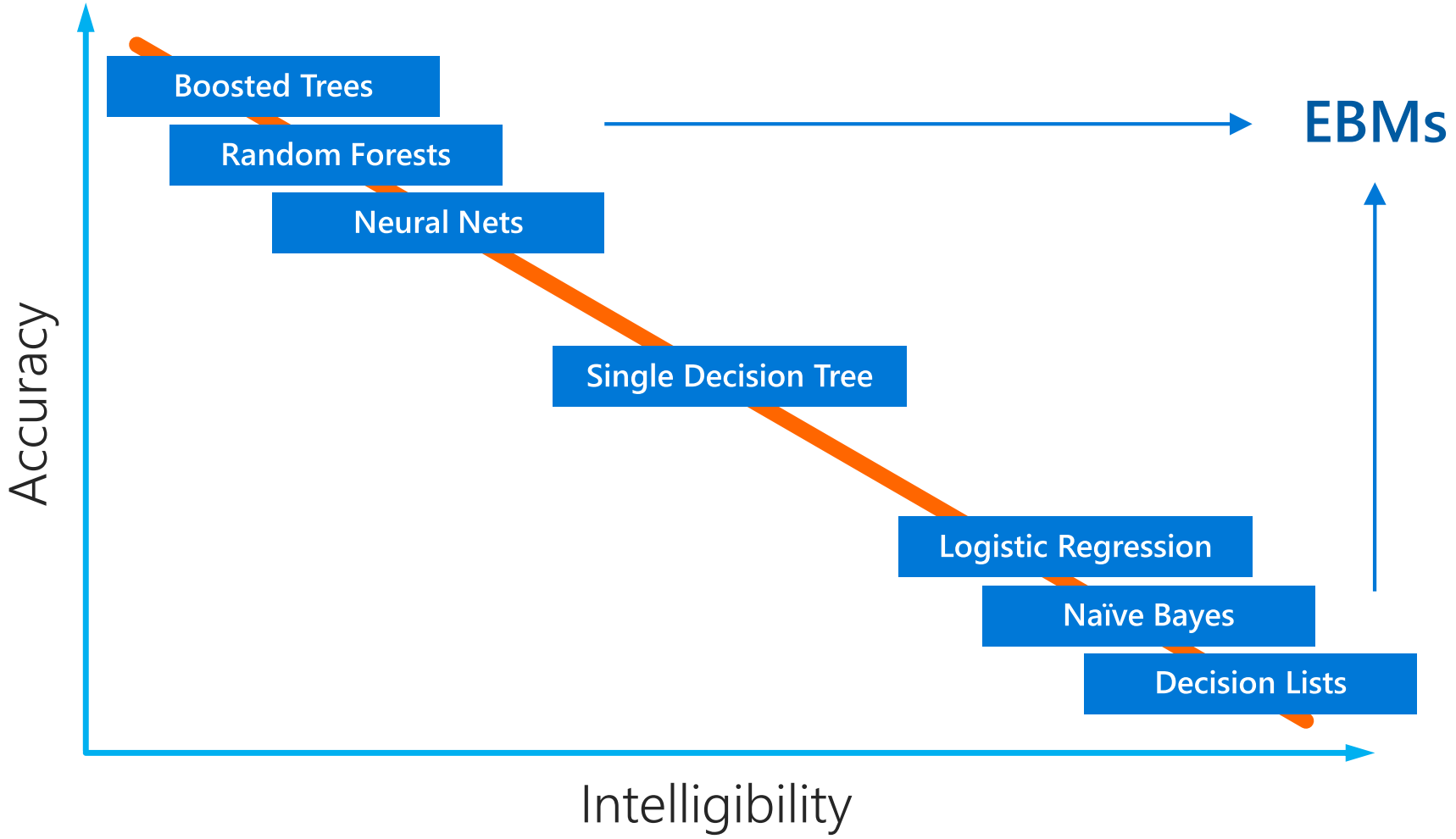
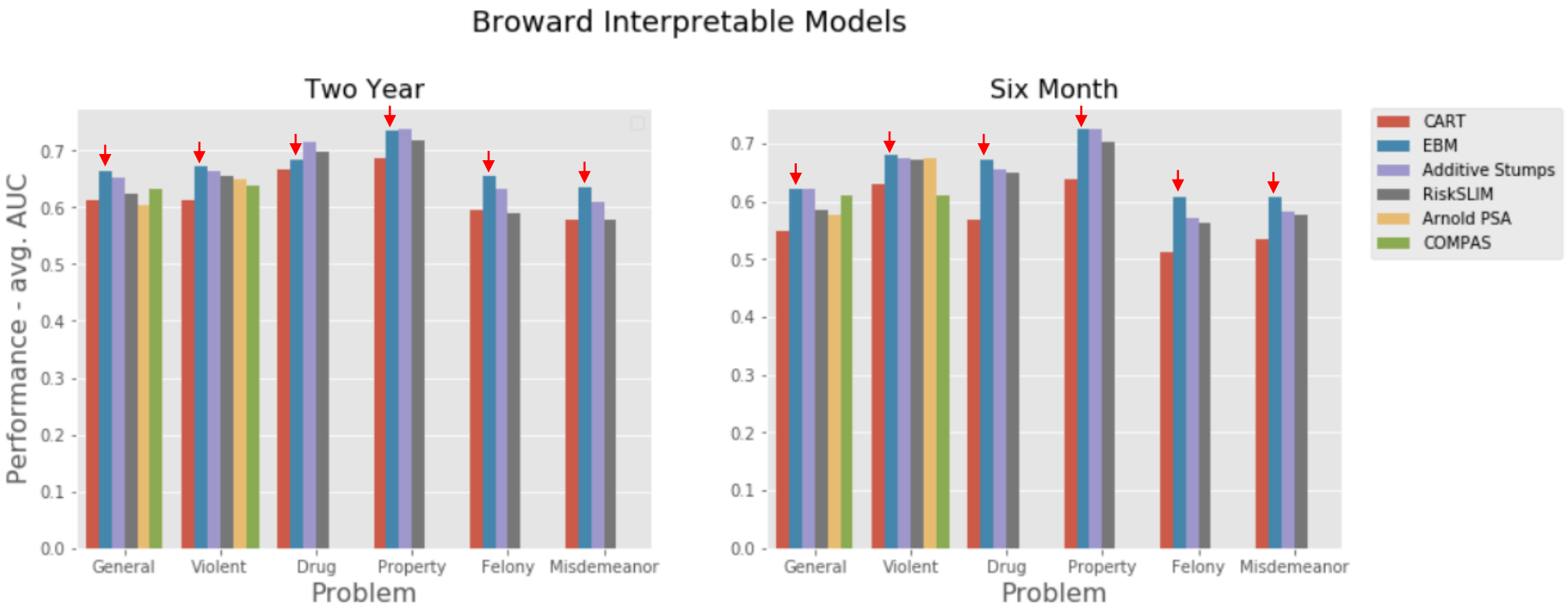


Table 1: Test set AUCs across 10 datasets. Best number in each row in **bold**.

	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., 2020. How Interpretable and Trustworthy are GAMs?. *arXiv preprint arXiv:2006.06466*.



"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., Frosst, N., Zhang, X., Caruana, R. and Hinton, G.E., 2020. Neural additive models: Interpretable machine learning with neural nets. *arXiv preprint arXiv:2004.13912*.

EBMs: Type of Generalized Additive Models (GAMS)

- Linear Model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$
- 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: $y = f(x_1, \ldots, x_n)$

- GAMs originally developed at Stanford in late 80's by Hastie & Tibshirani
  - But statisticians were too conservative fitting models: less accuracy and less intelligibility!

- Our contribution is to put EBMs (GAMs) on modern, machine learning steroids:
  - As machine learning people, we're not so conservative…
  - Improved accuracy, intelligibility and editability
  - Released a modern, easy-to-use open-source package: *https://github.com/interpretML/interpret*

Algorithm Sketch

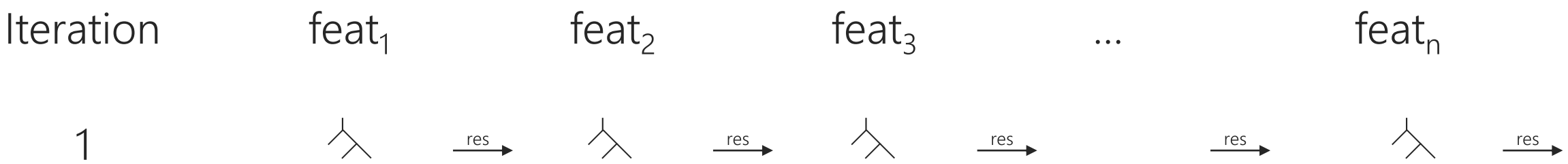
Iteration	$feat_1$	$feat_2$	$feat_3$	...	$feat_n$
1					

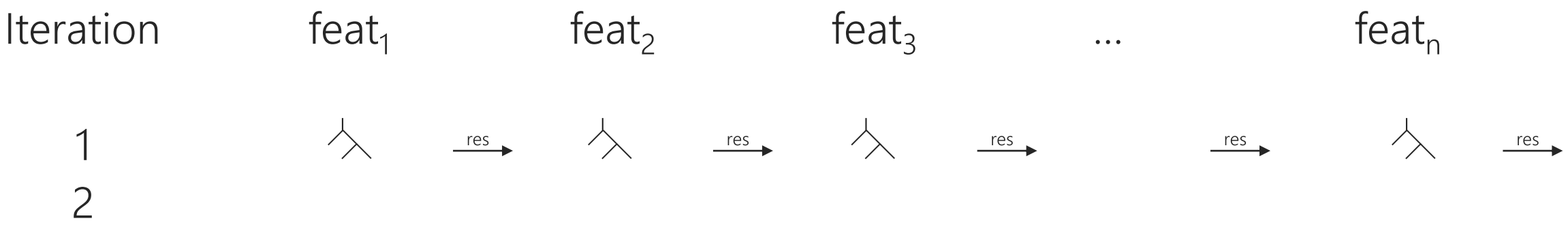
Iteration	$feat_1$	$feat_2$	$feat_3$	...	$feat_n$
1					

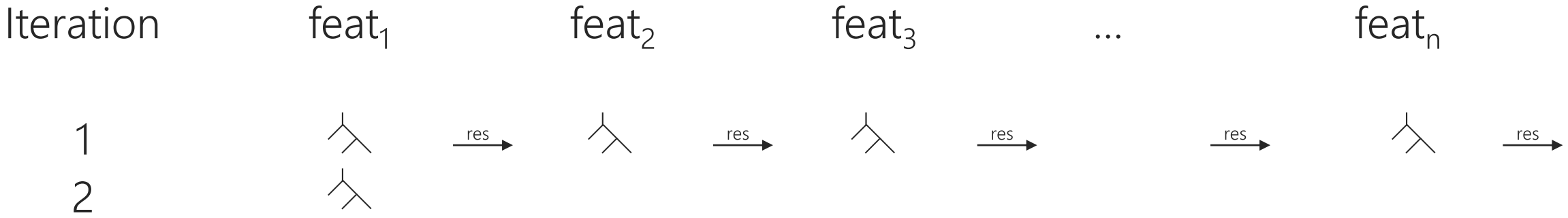
Iteration	$feat_1$	$feat_2$	$feat_3$	...	$feat_n$
1					

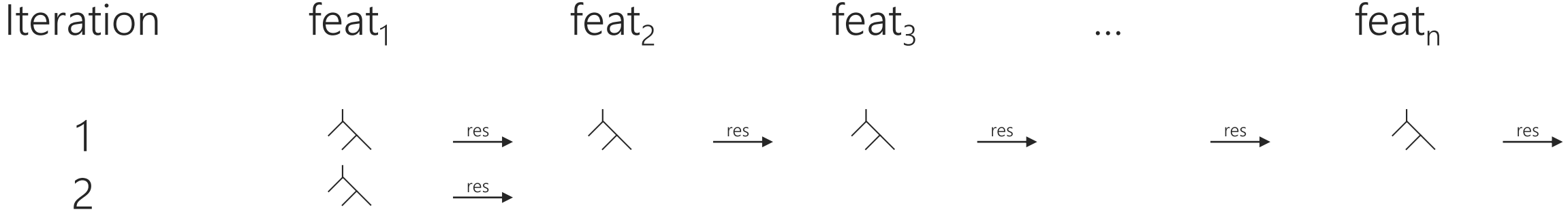
Iteration	$feat_1$		$feat_2$	$feat_3$	...	$feat_n$
1		res				

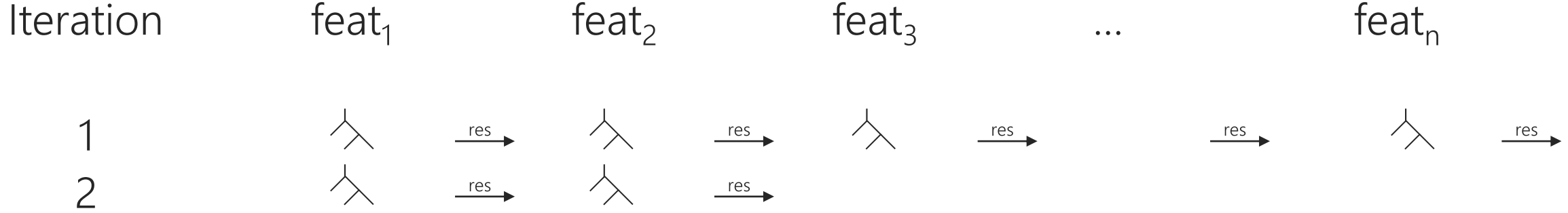
Iteration	$feat_1$	$feat_2$	$feat_3$	...	$feat_n$
1					

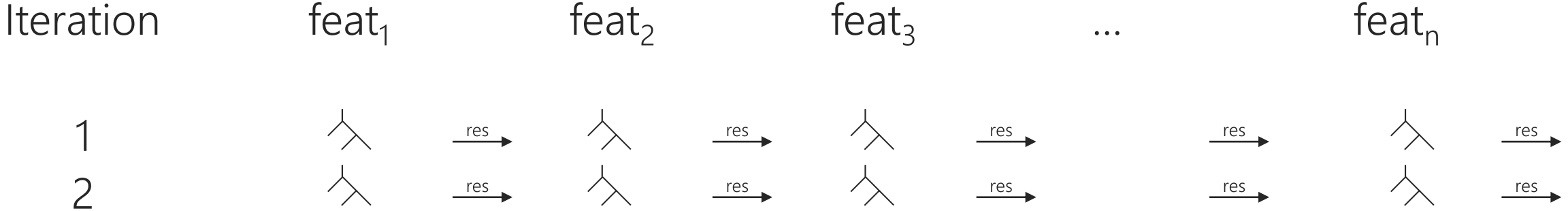

Iteration
feat1
feat2
feat3
...
featn
1
res
res
res
res
res

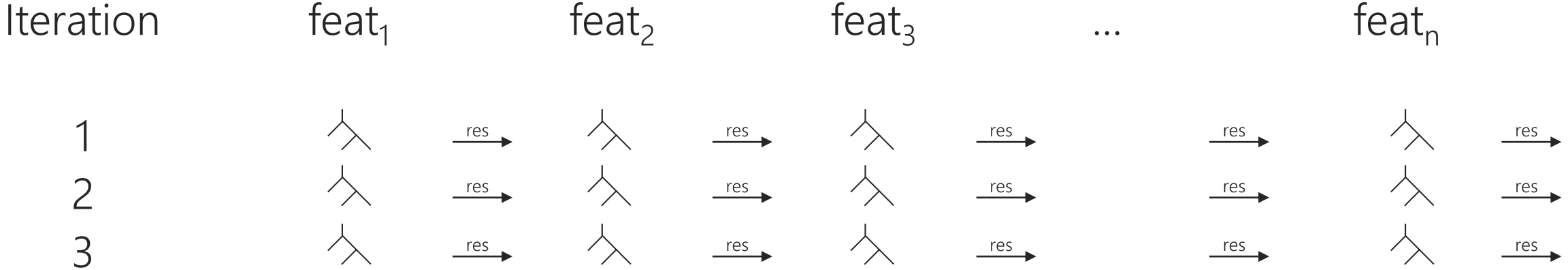

Iteration
feat1
feat2
feat3
...
featn
1
2
res
res
res
res
res

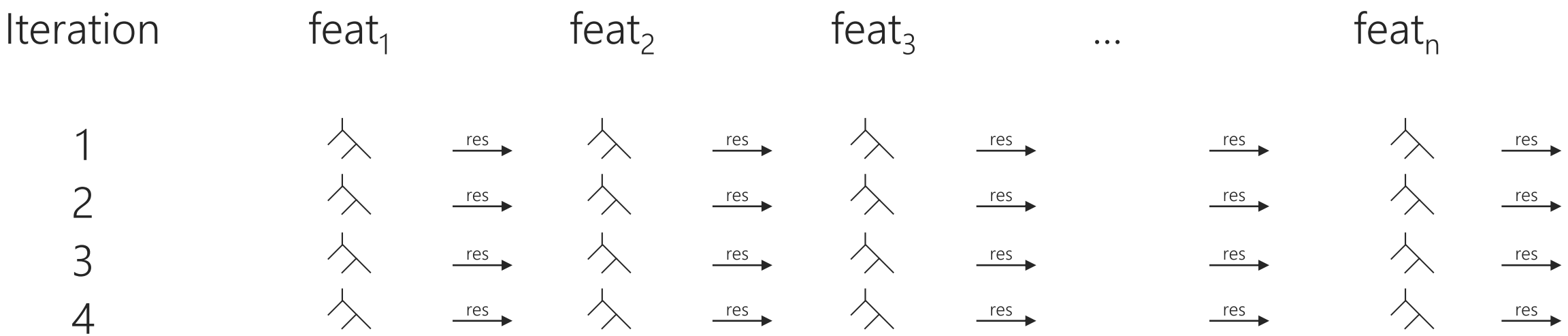

Iteration
feat1
feat2
feat3
...
featn
1
2
res
res
res
res
res

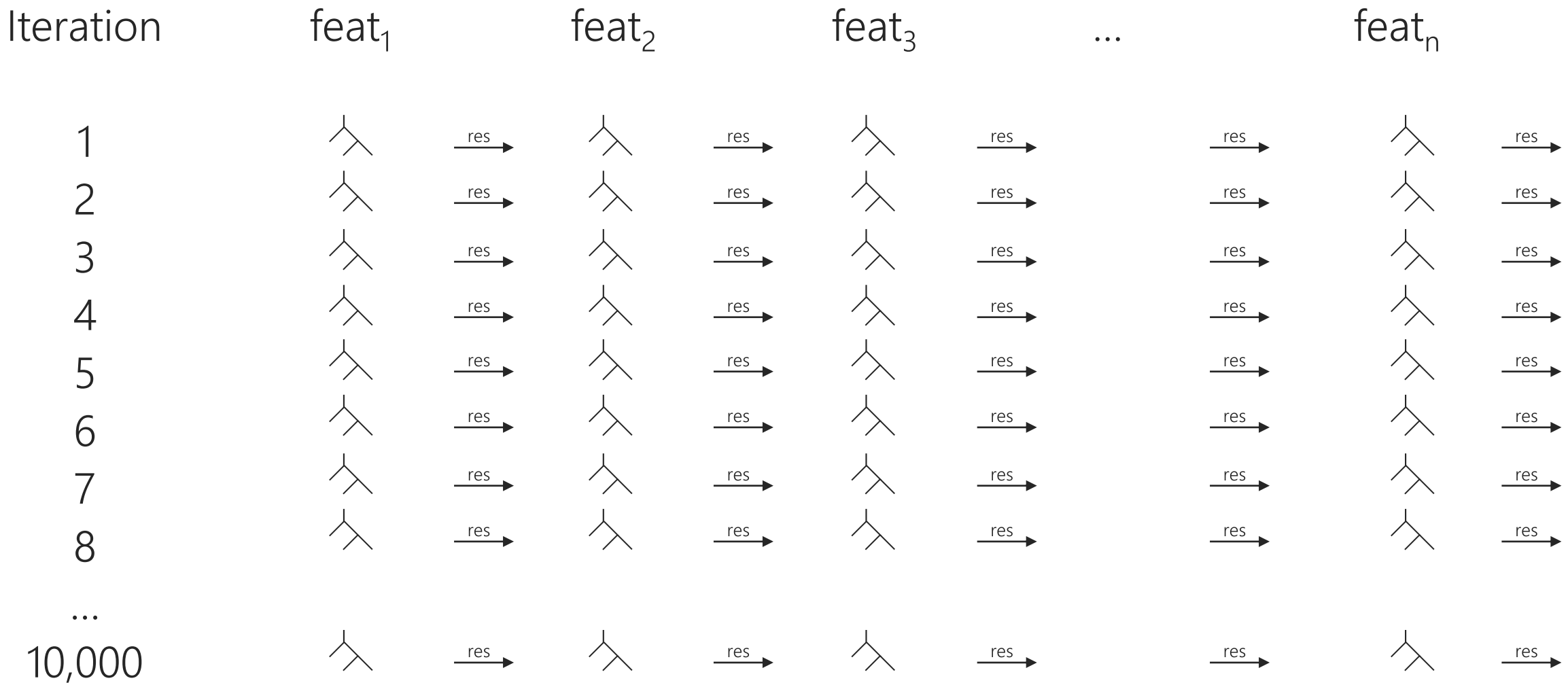

Iteration
feat1
feat2
feat3
...
featn
1
2
res
res
res
res
res
res

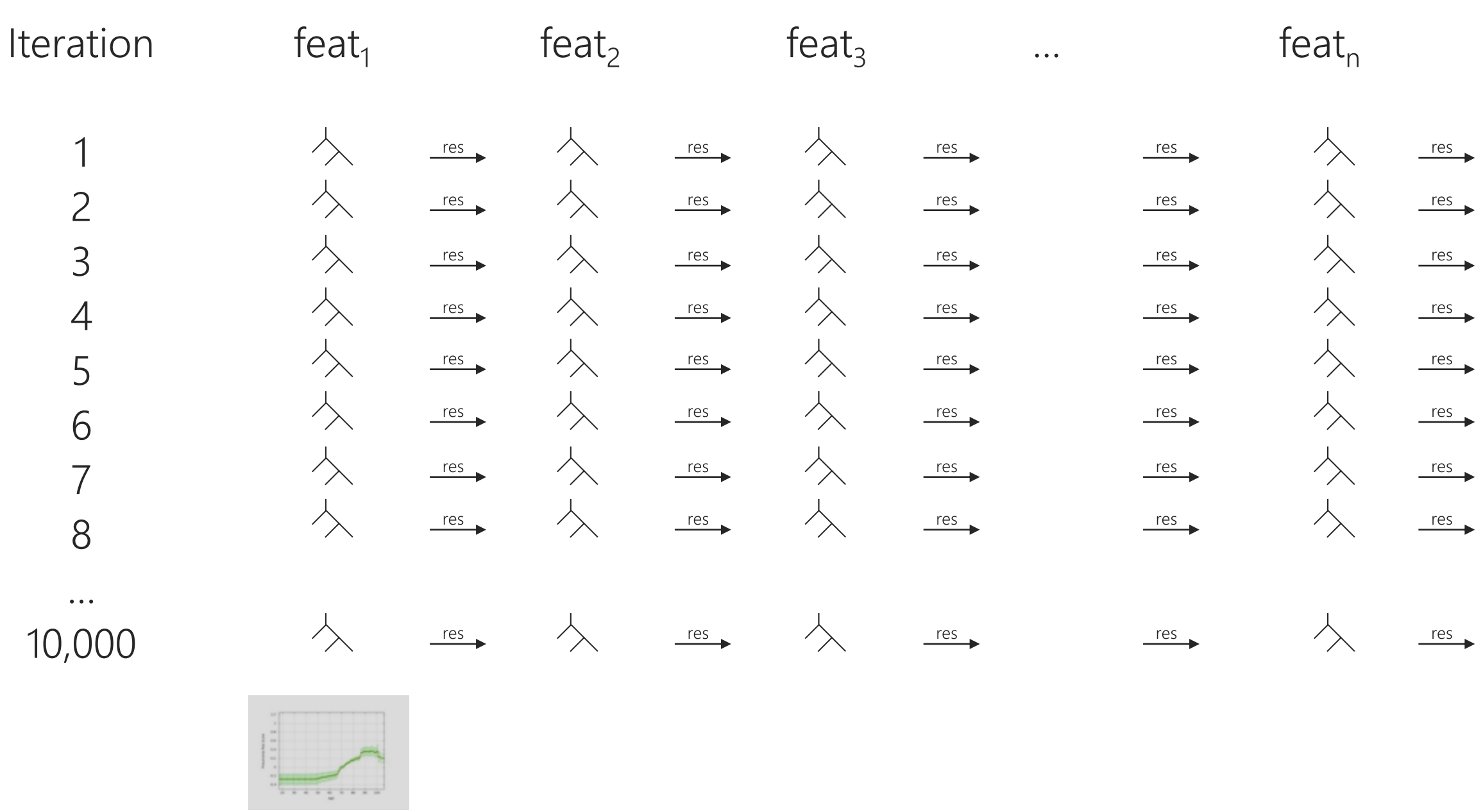

Iteration
feat1
feat2
feat3
...
featn
1
2
res
res
res
res
res
res
res

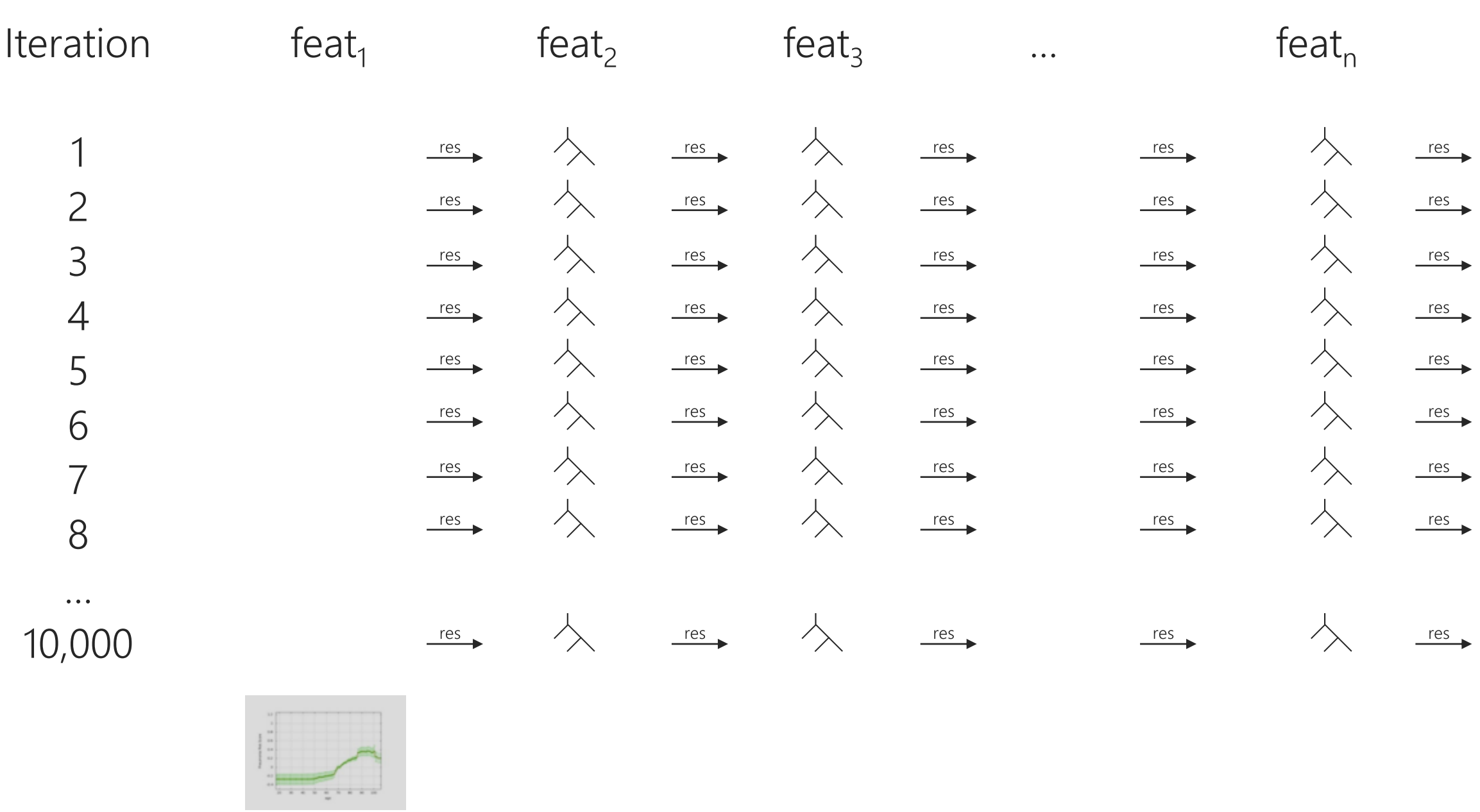

Iteration
feat1
feat2
feat3
...
featn
1
2
res

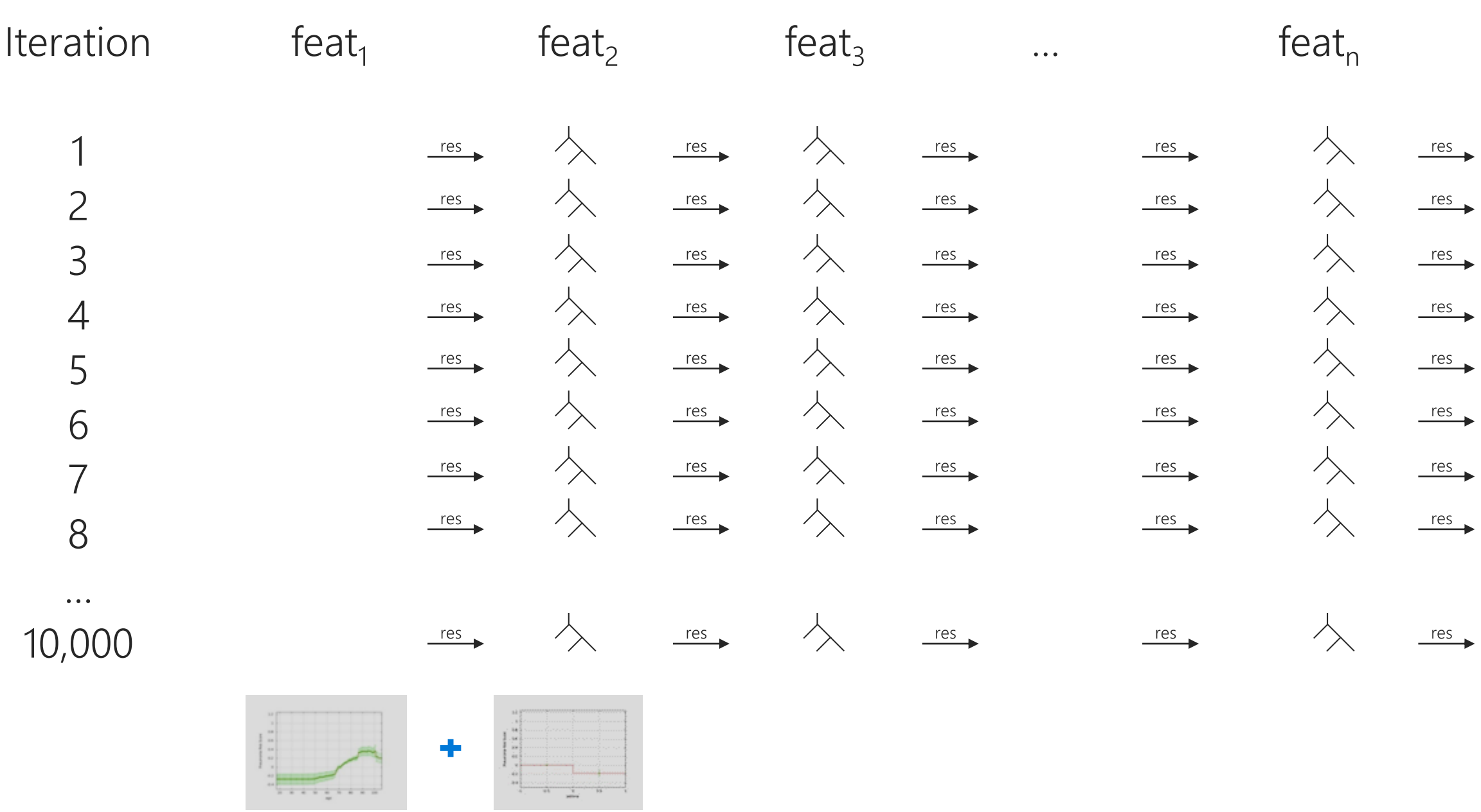

Iteration
feat1
feat2
feat3
...
featn
1
2
3
res

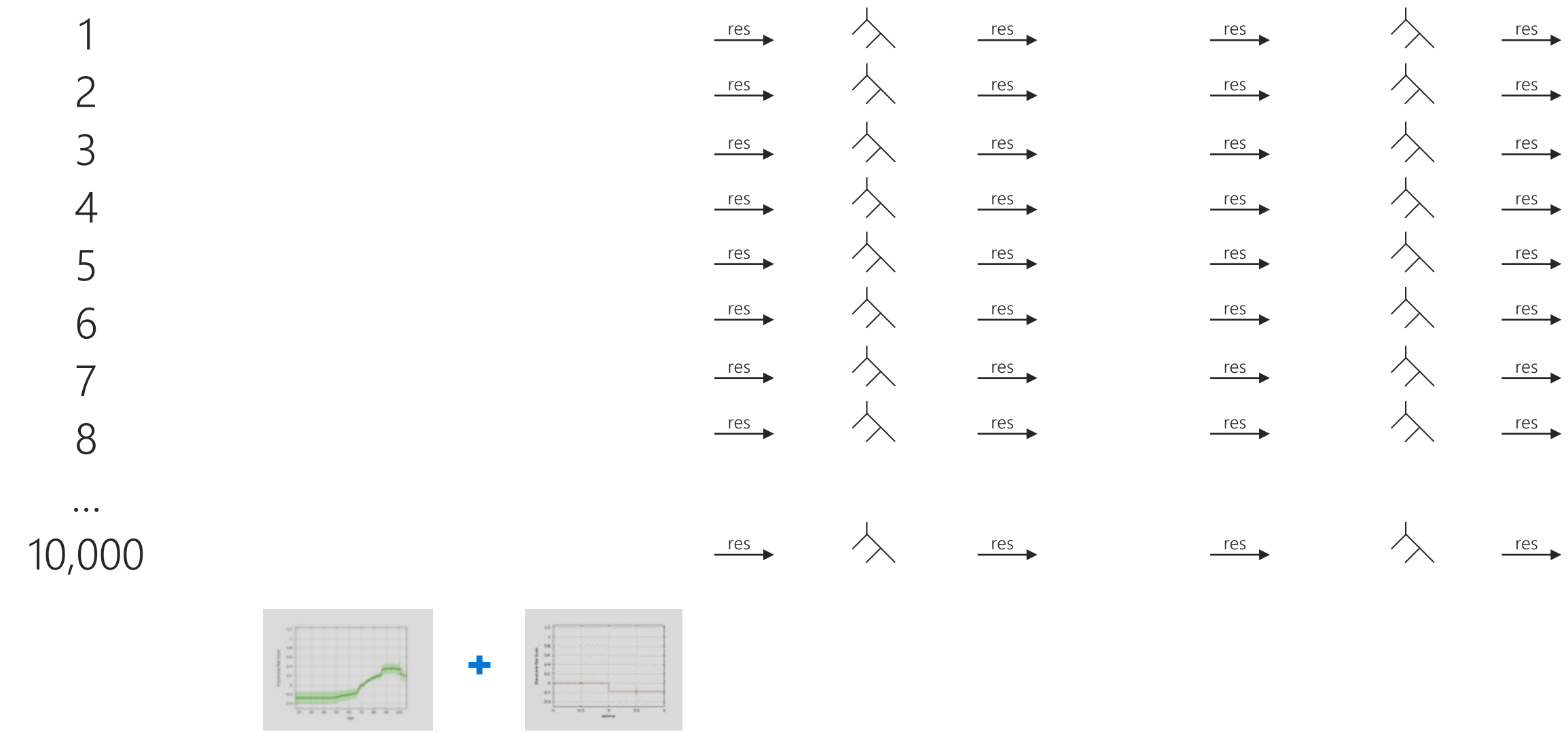

Iteration
feat1
feat2
feat3
...
featn
1
2
3
4
res

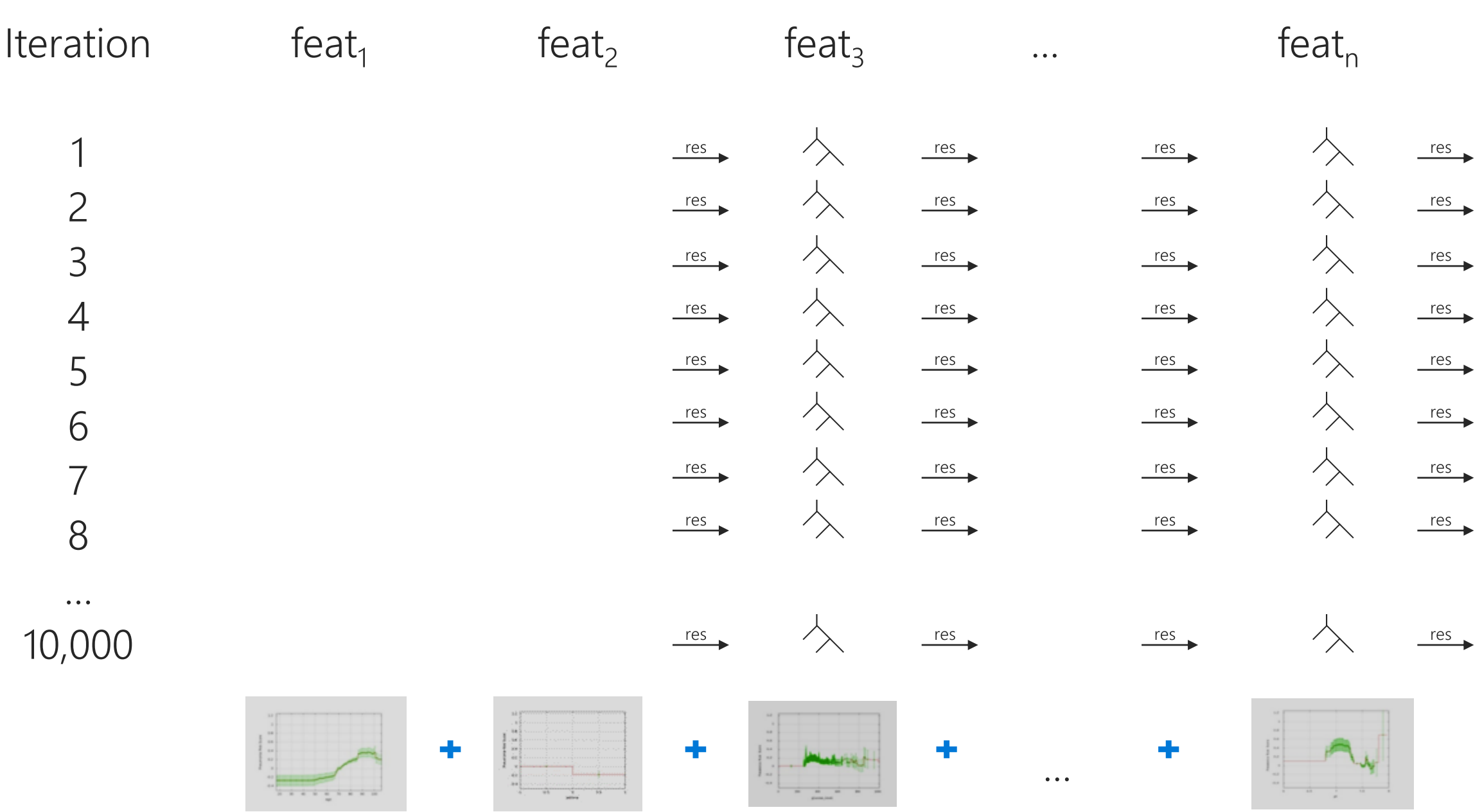

Iteration
feat1
feat2
feat3
...
featn
1
2
3
4
5
6
7
8
...
10,000
res


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

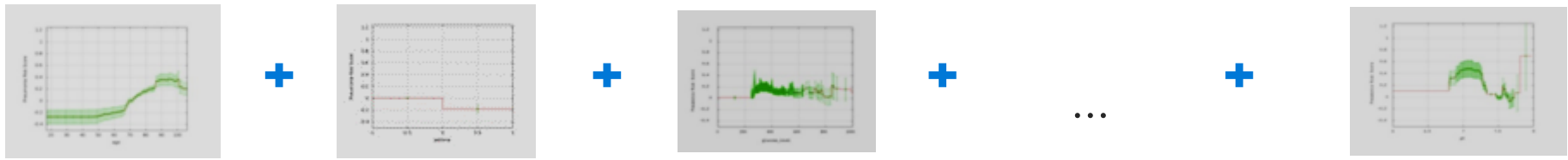

Iteration
feat1
feat2
feat3
...
featn
1
2
3
4
5
6
7
8
...
10,000
res


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


Iteration
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

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



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					

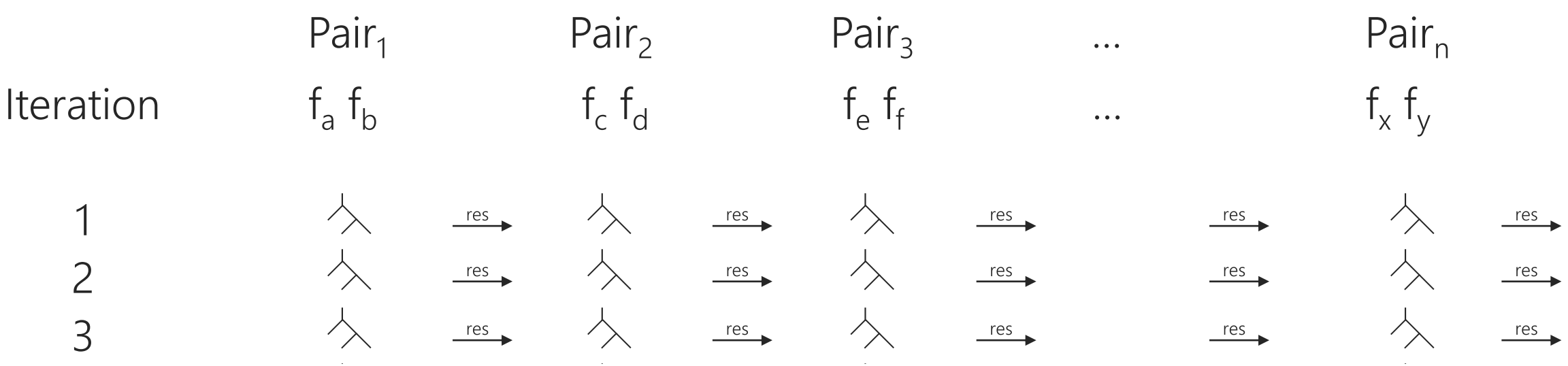
	$Pair_1$	$Pair_2$	$Pair_3$	...	$Pair_n$
Iteration	f_a f_b	f_c f_d	f_e f_f	...	f_x f_y
1					

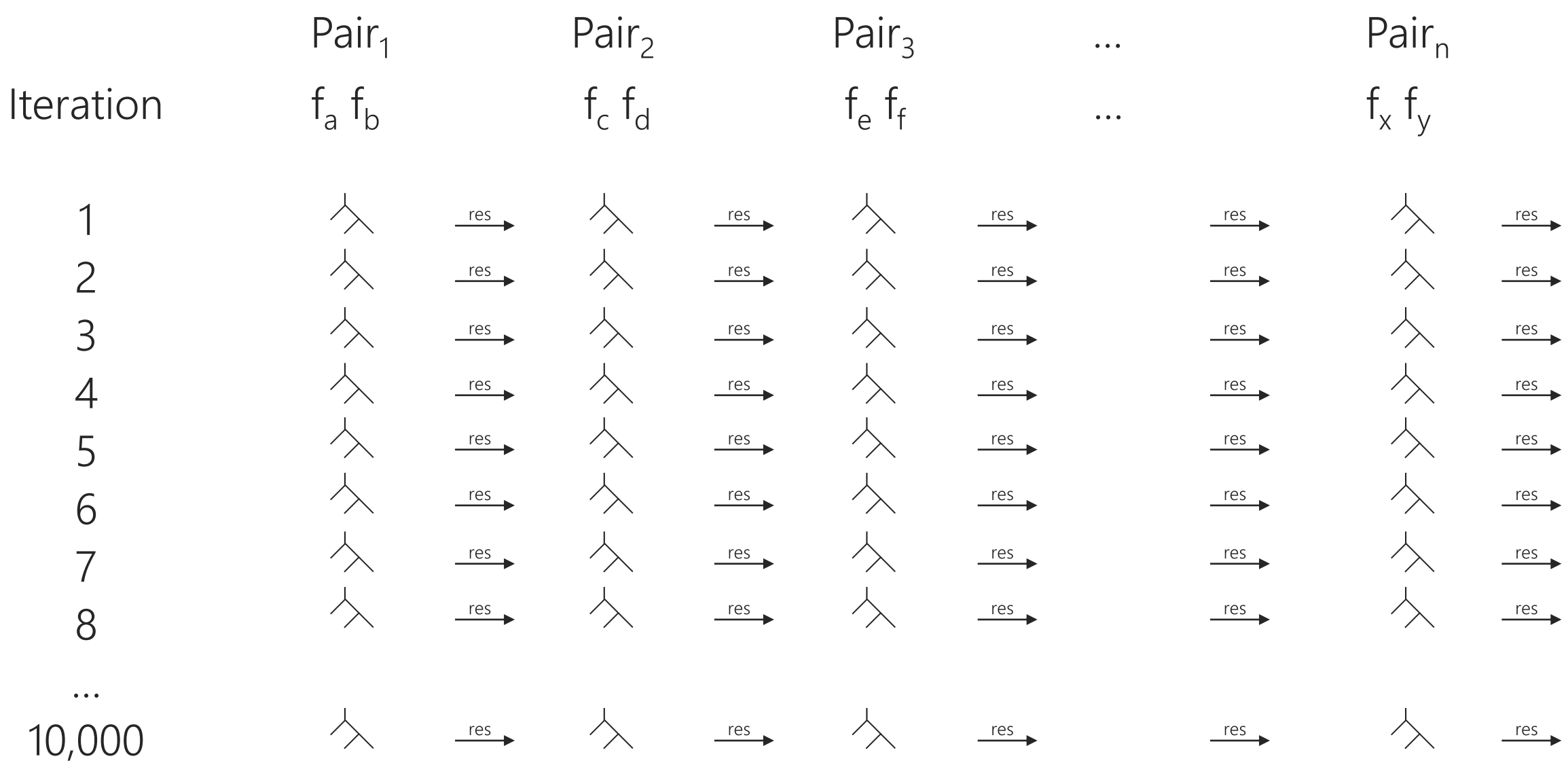
	$Pair_1$	$Pair_2$	$Pair_3$	...	$Pair_n$
Iteration	f_a f_b	f_c f_d	f_e f_f	...	f_x f_y
1					

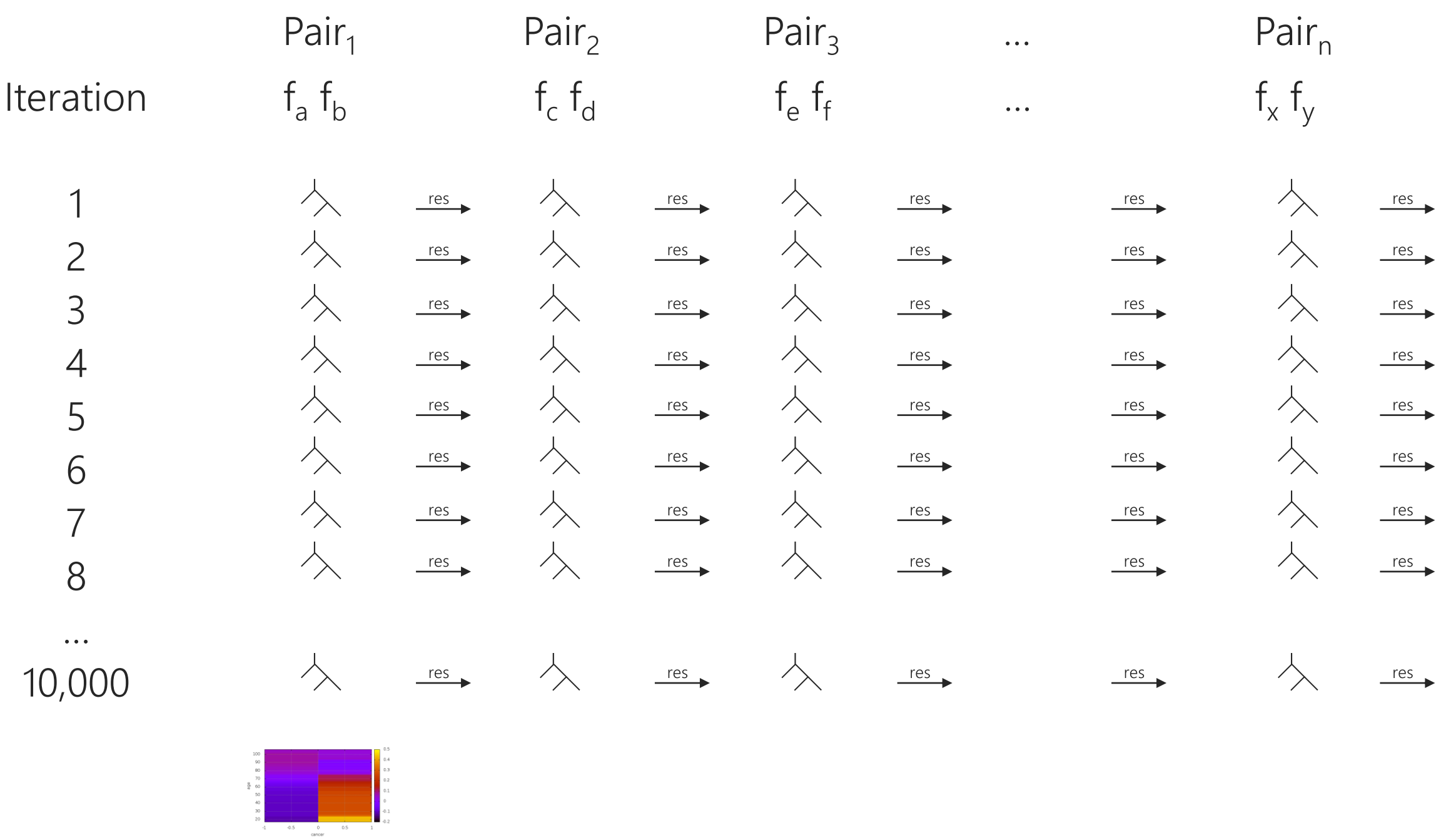
	$Pair_1$	$Pair_2$	$Pair_3$	...	$Pair_n$
Iteration	f_a f_b	f_c f_d	f_e f_f	...	f_x f_y
1					

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

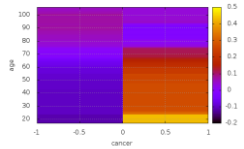
Iteration	$Pair_1$	$Pair_2$	$Pair_3$	...	$Pair_n$
	f_a f_b	f_c f_d	f_e f_f	...	f_x f_y
1					
2					


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

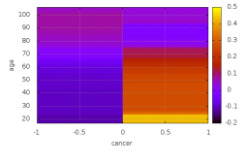

Pair1
Pair2
Pair3
...
Pairn
Iteration
fa fb
fc fd
fe ff
...
fx fy
1
2
3
4
5
6
7
8
...
10,000
res

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		 res	 res	res	 res
2		 res	 res	res	 res
3		 res	 res	res	 res
4		 res	 res	res	 res
5		 res	 res	res	 res
6		 res	 res	res	 res
7		 res	 res	res	 res
8		 res	 res	res	 res
...					
10,000		 res	 res	res	 res

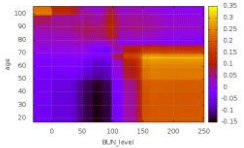


Iteration	$Pair_1$	$Pair_2$	$Pair_3$	...	$Pair_n$
	f_a f_b	f_c f_d	f_e f_f	...	f_x f_y
1			res	res	res
2			res	res	res
3			res	res	res
4			res	res	res
5			res	res	res
6			res	res	res
7			res	res	res
8			res	res	res
...					
10,000			res	res	res

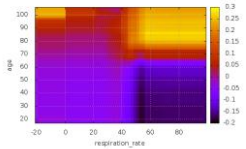
Iteration	$Pair_1$	$Pair_2$	$Pair_3$	...	$Pair_n$
	f_a f_b	f_c f_d	f_e f_f	...	f_x f_y
1			res	res	res
2			res	res	res
3			res	res	res
4			res	res	res
5			res	res	res
6			res	res	res
7			res	res	res
8			res	res	res
...					
10,000			res	res	res



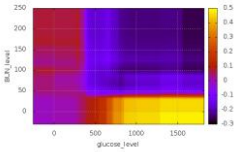
+



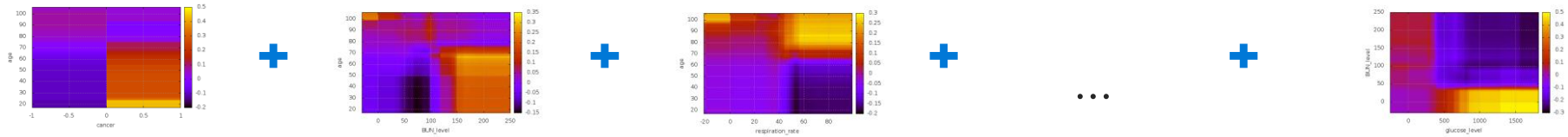
+



+ ... +

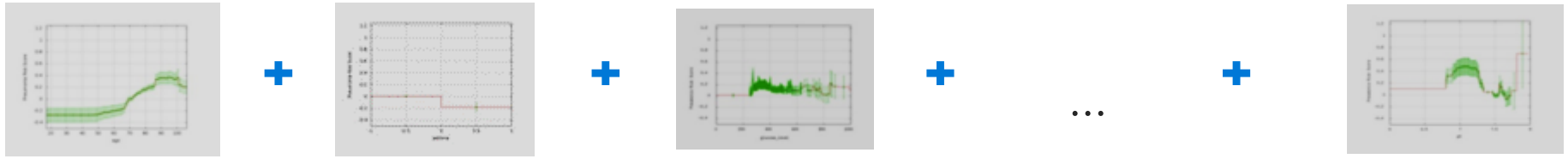


$Pair_1$	$Pair_2$	$Pair_3$	...	$Pair_n$
f_a f_b	f_c f_d	f_e f_f	...	f_x f_y



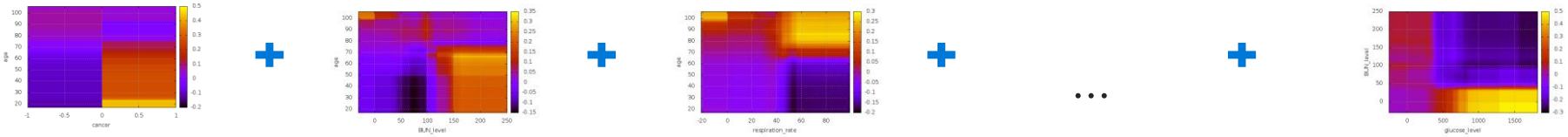
Final Model: Mains + Select Pairwise Interactions

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



+

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



Explainable Boosting Machine (EBM) Case Studies

1. Pneumonia Mortality
2. MIMIC-II: ICU Mortality
3. German Credit
4. Wikipedia Malicious Edits
5. Pregnancy: Severe Maternal Morbidity
6. COVID-19 Mortality
7. 30-Day Hospital Readmission
8. Bias & Recidivism Prediction

Case Study 1: Pneumonia Mortality

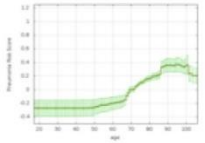
Pneumonia Dataset (collected 1989): 46 Features

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

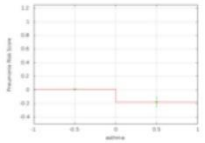
Physical examination findings
Respiration rate (resps/min)
Heart rate (beats/min)
Systolic blood pressure (mmHg)
Temperature (°C)
Altered mental status (disorientation, lethargy, or coma)
Wheezing
Stridor
Heart murmur
Gastrointestinal bleeding

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)

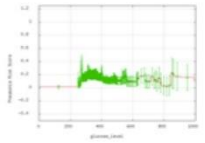
Chest X-ray findings
Positive chest X-ray
Lung infiltrate
Pleural effusion
Pneumothorax
Cavitation/empyema
Lobe or lung collapse
Chest mass



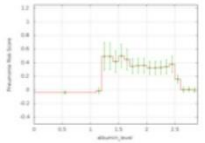
Age => -0.23



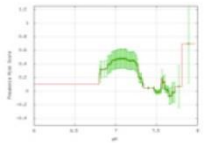
Asthma => -0.15



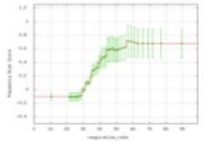
Glucose => +0.18



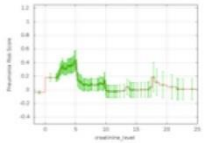
Albumin => +0.01



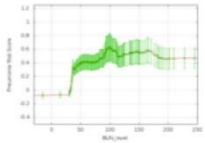
Blood pH => +0.38



Respiration => +0.21



Creatinine => -0.01

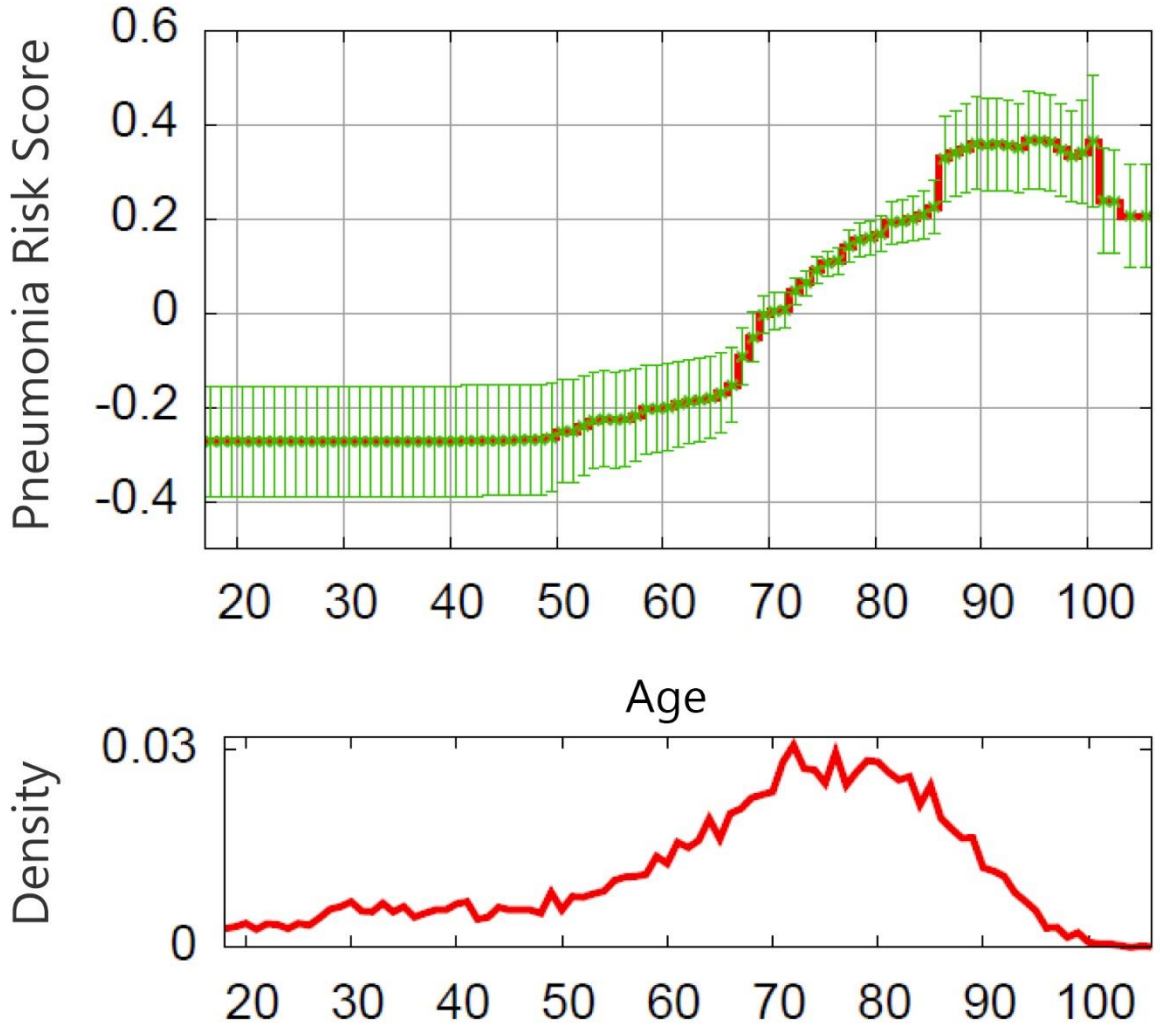


BUN => -0.21

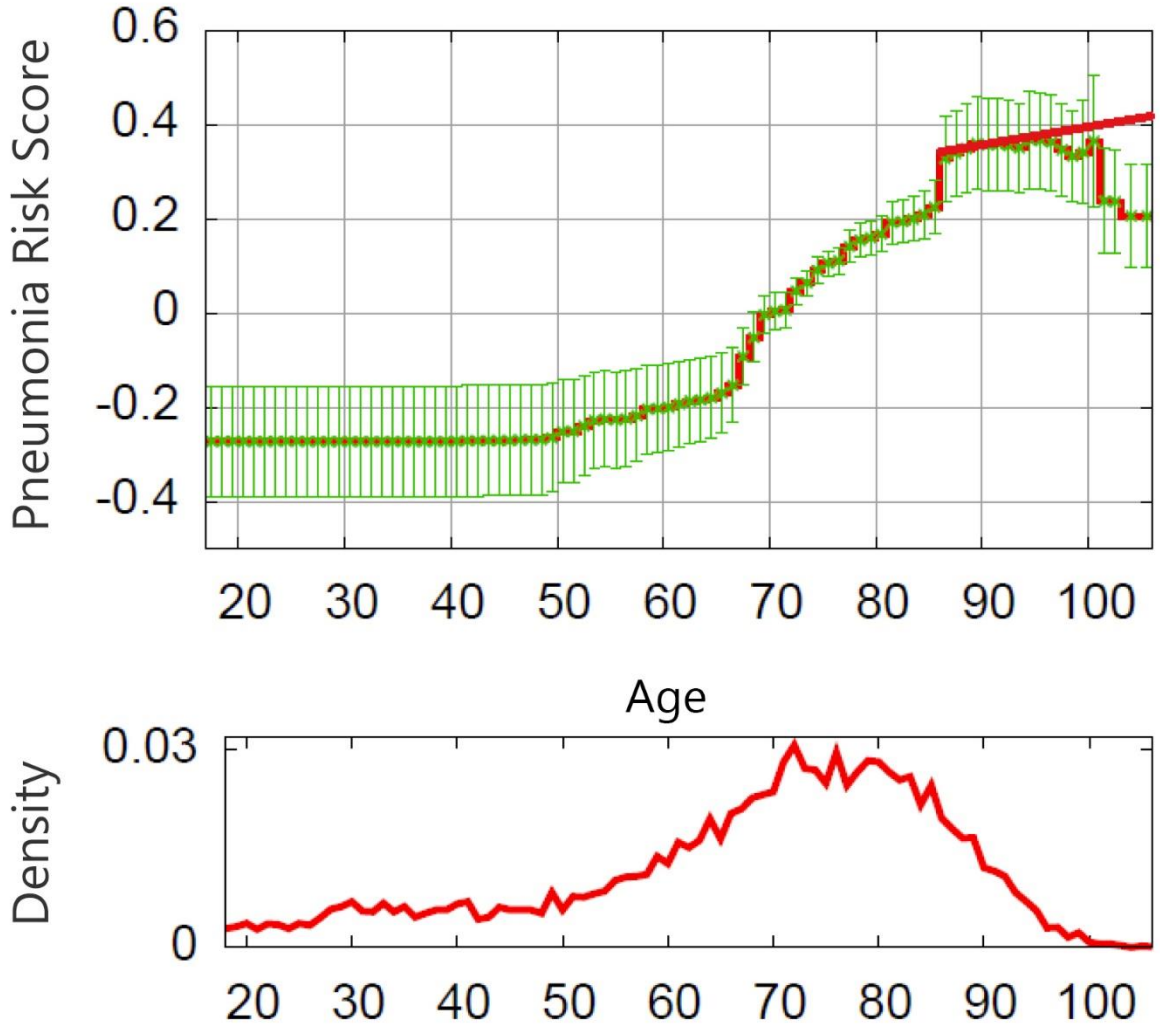
$$Score = baseline + \sum_{i=0}^{n} f_i(variable_i)$$

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

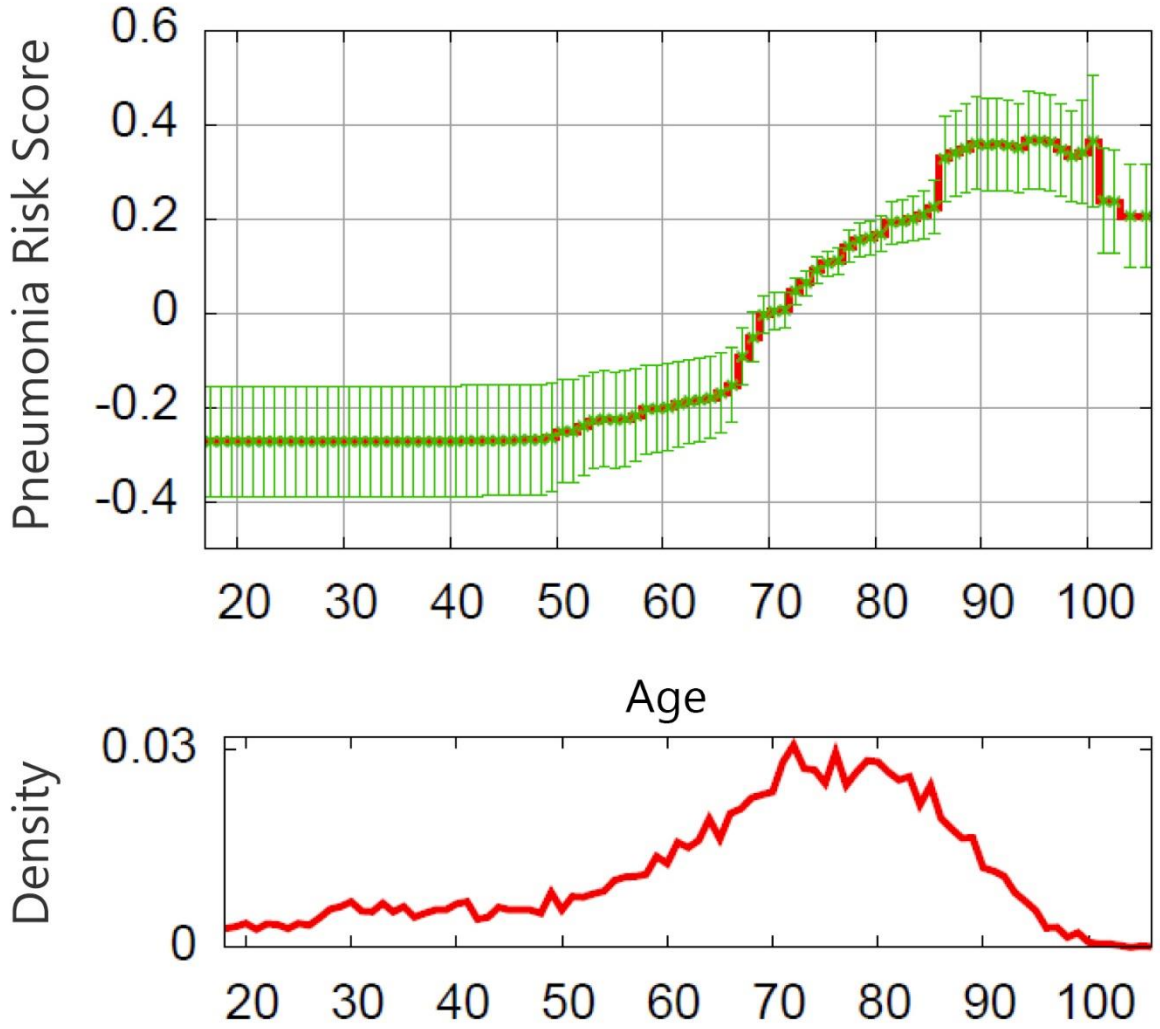
What EBMs Learn about Pneumonia Risk vs. Age



Fix Age > 100 Problem (Enforce Monotonicity)

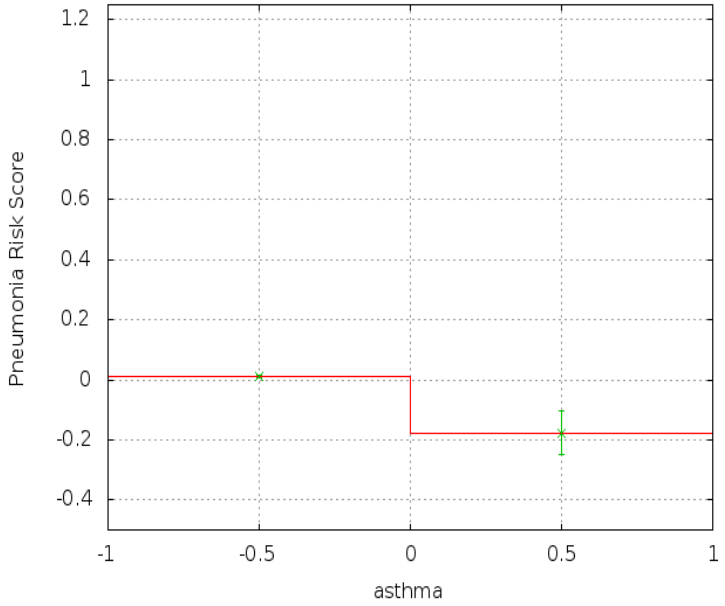


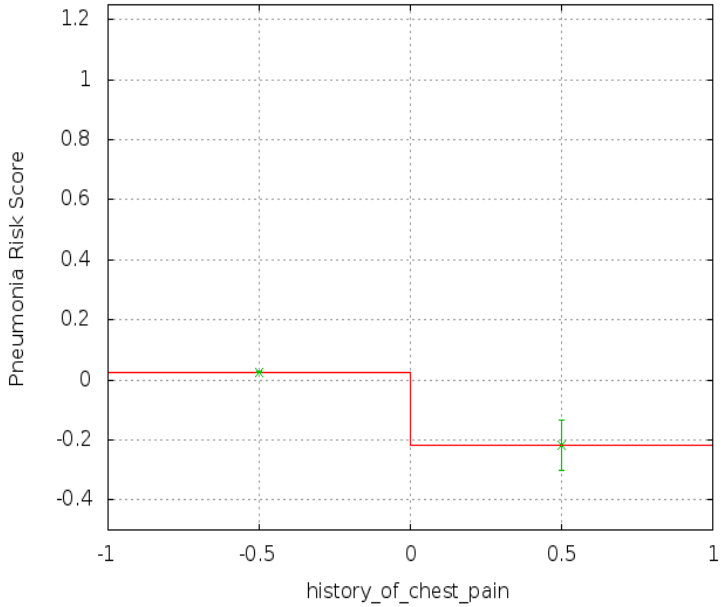
Original Model is Correct for Actuarial Use

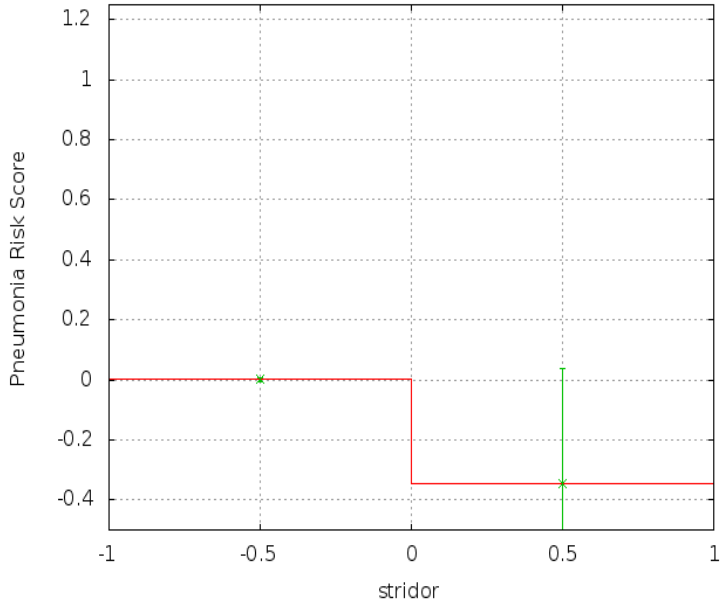


- Model correctness (and accuracy) depends on how model will be used
  - This is a good model for health insurance provides
  - But needs to be repaired before using for patient treatment decision

- 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

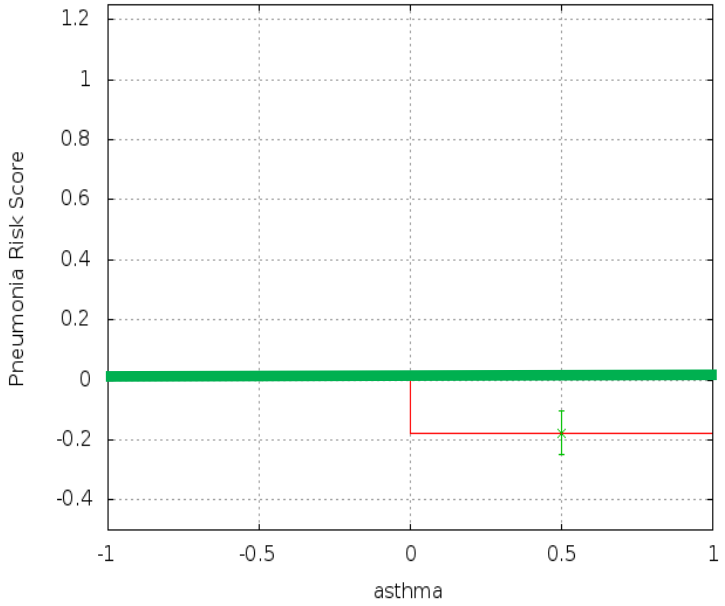


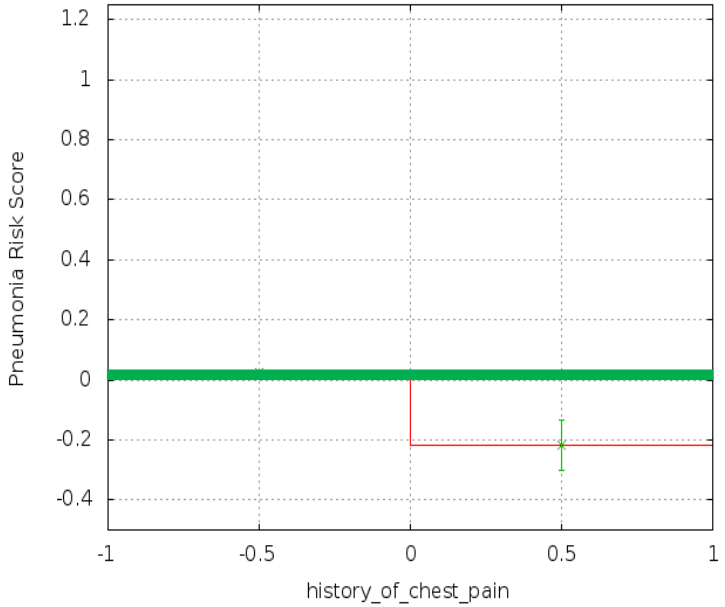


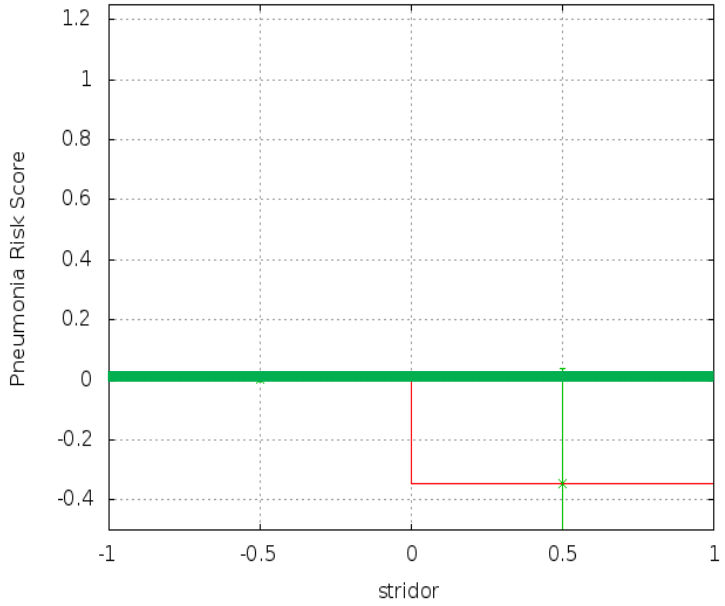


- Model correctness (and accuracy) depends on how model will be used
  - This is a good model for health insurance provides
  - But needs to be repaired before using for patient treatment decision

- 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

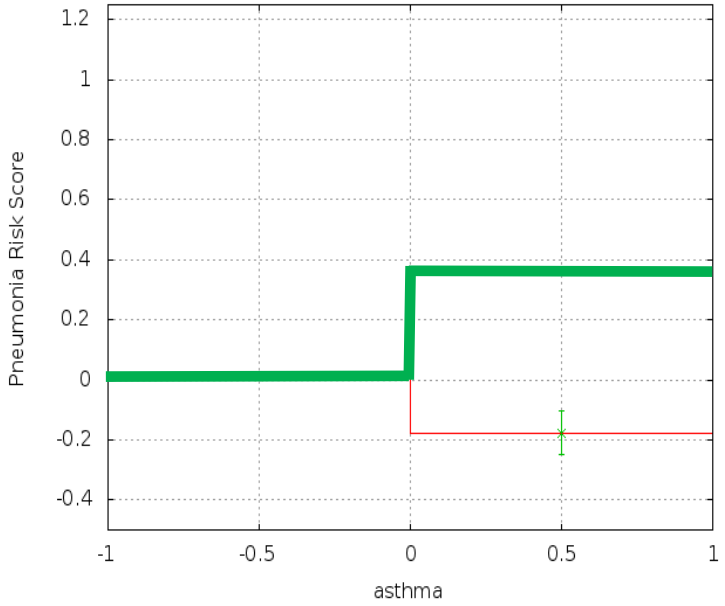


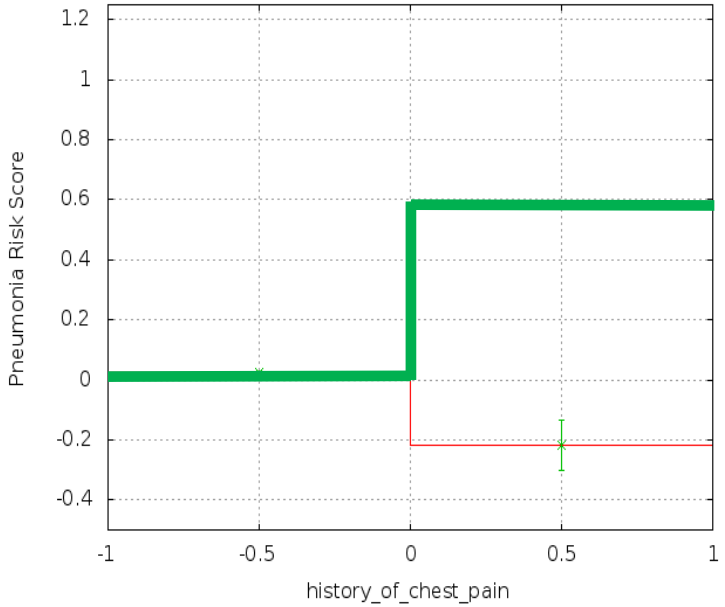


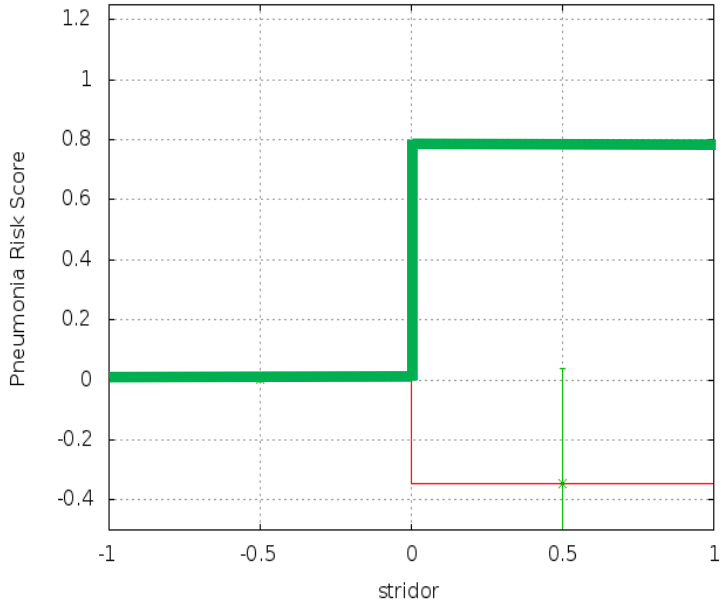


- Model correctness (and accuracy) depends on how model will be used
  - This is a good model for health insurance provides
  - But needs to be repaired before using for patient treatment decision

- 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





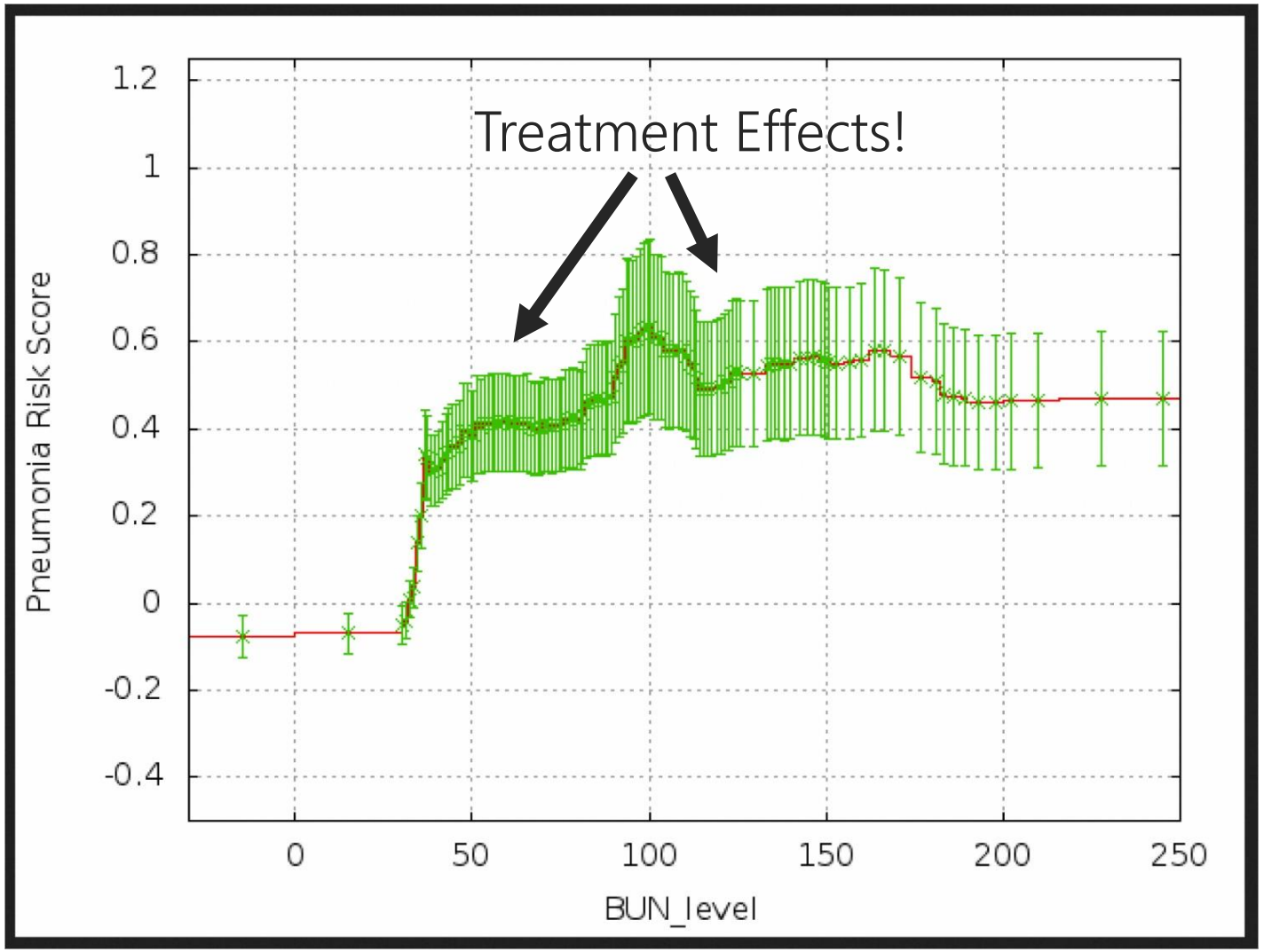


- Model correctness (and accuracy) depends on how model will be used
  - This is a good model for health insurance provides
  - But needs to be repaired before using for patient treatment decision

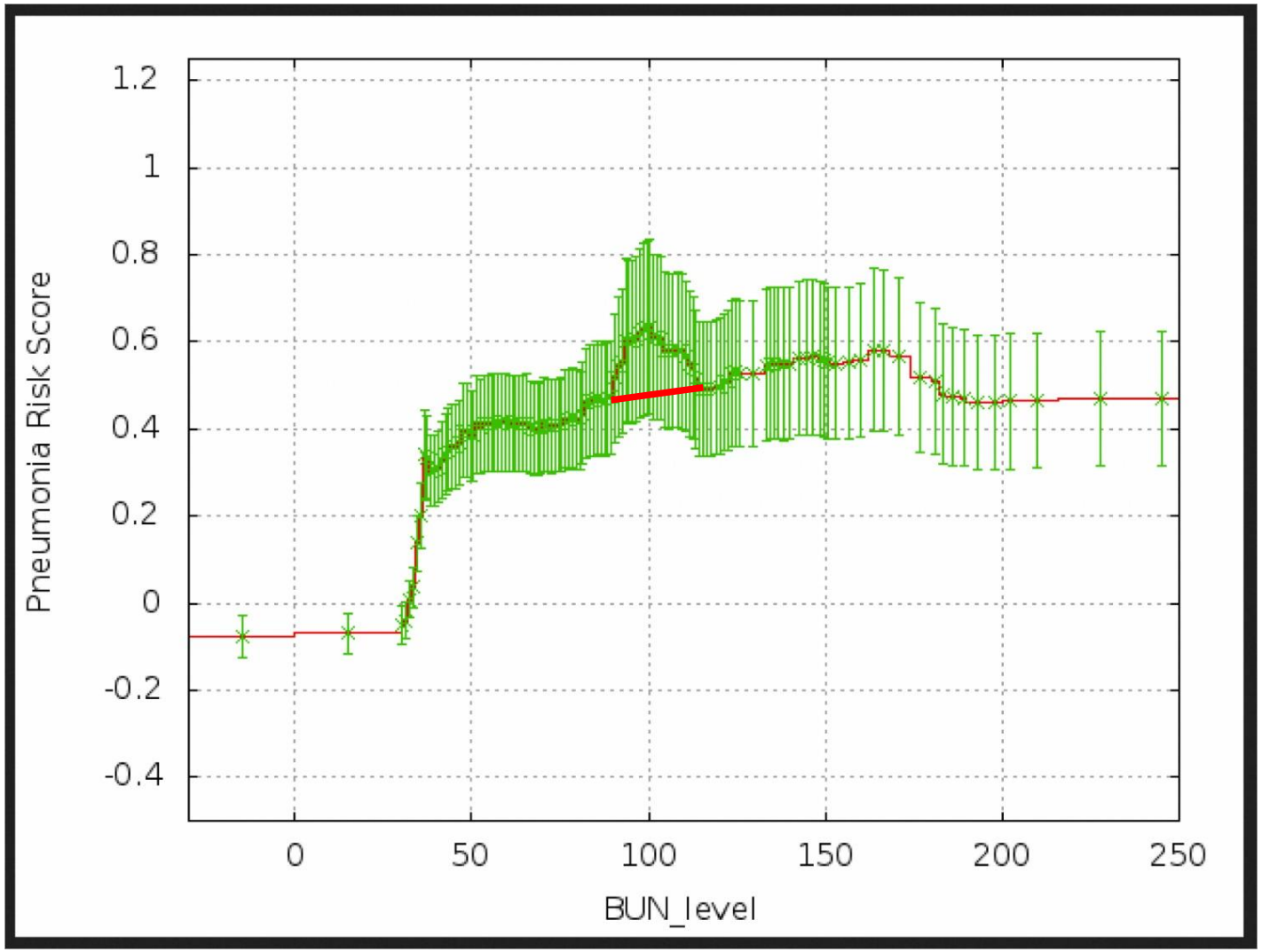
- A few things intelligible model learned:
  - Beware of jumps at round numbers --- almost always due to human/social/policy effects
  - Asthma => lower risk
  - History of chest pain => lower risk
  - History of heart disease => lower risk
  - Obstructed airway => lower risk
  - …
  - Models are rewarded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

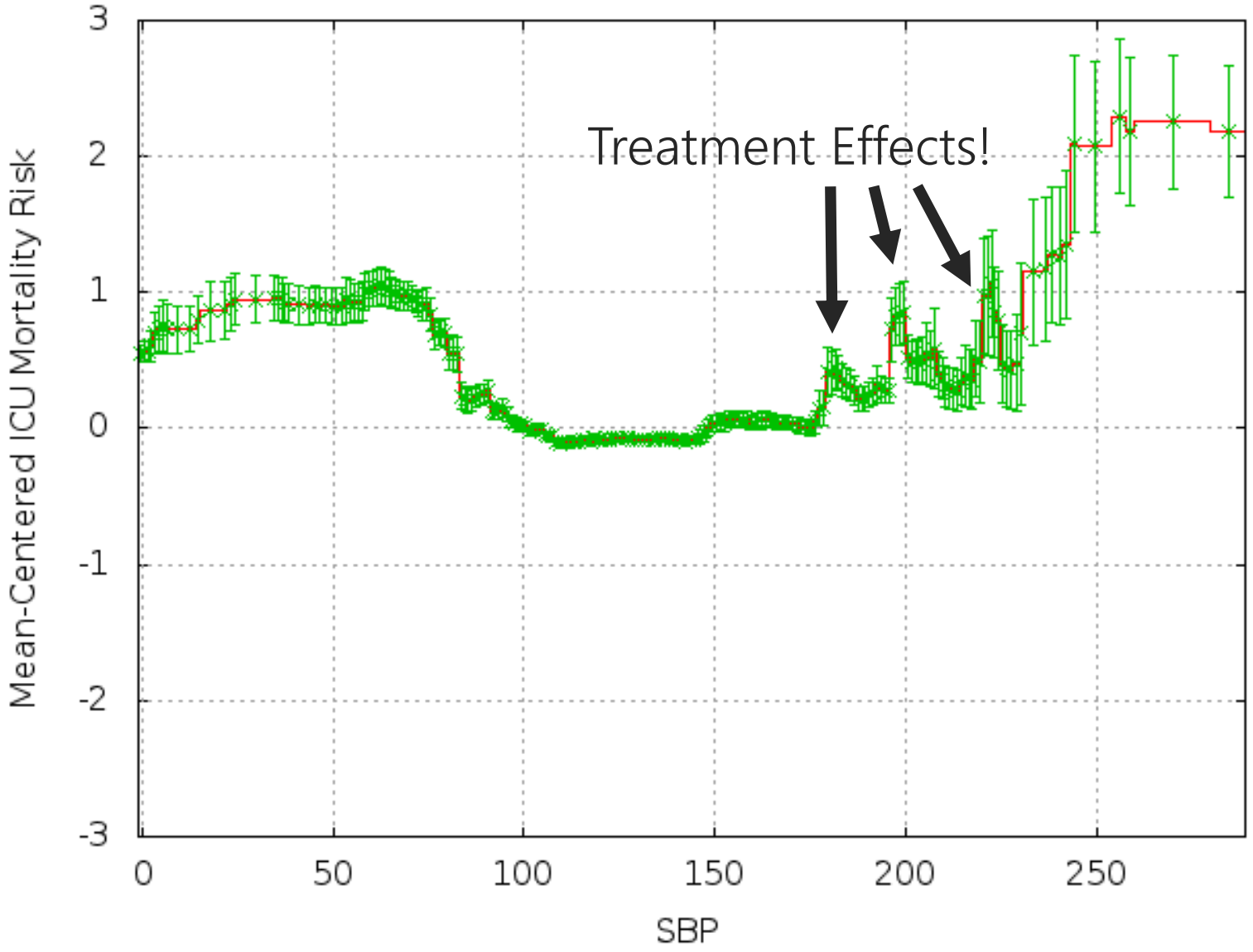


Intelligibility Can Create New Medical Science



Can save 2500 lives per year

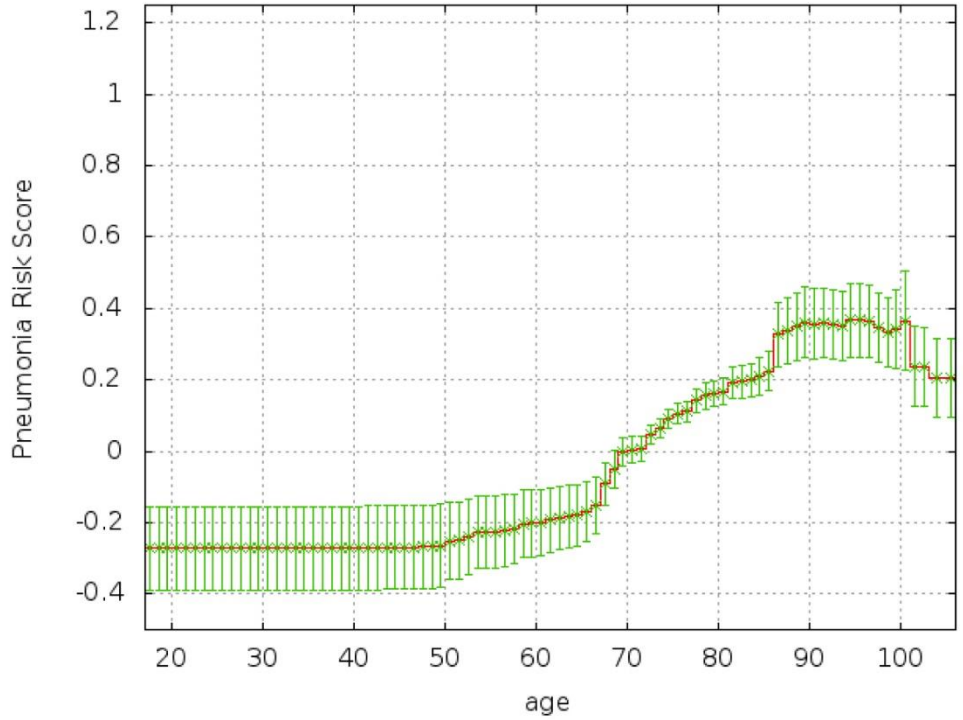
Treatment Effects Ubiquitous in All Medical Data

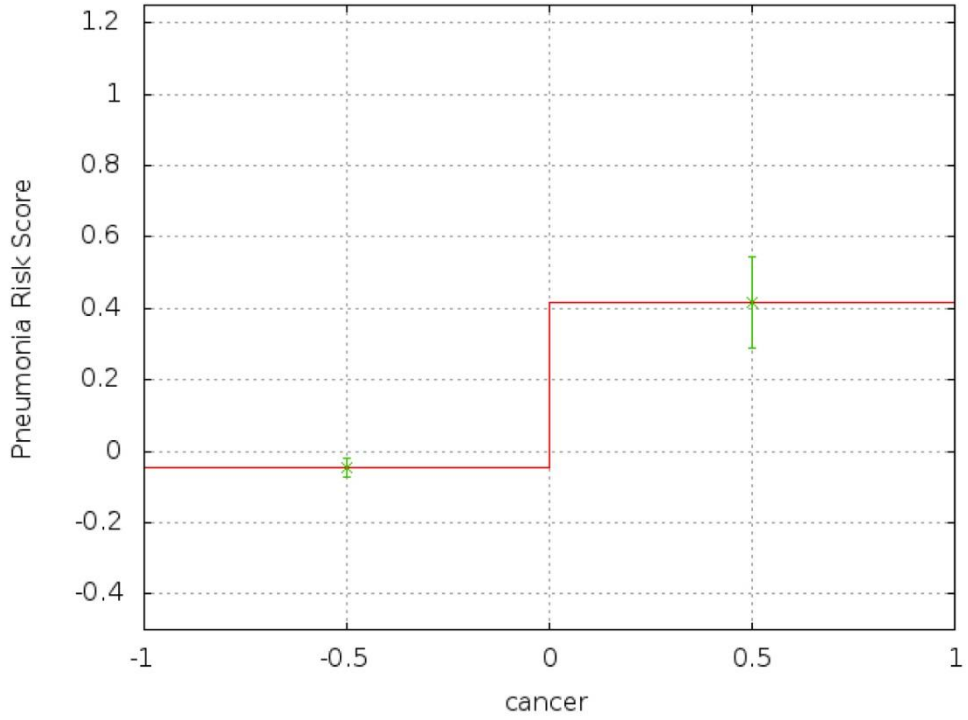


Pairwise Interactions?

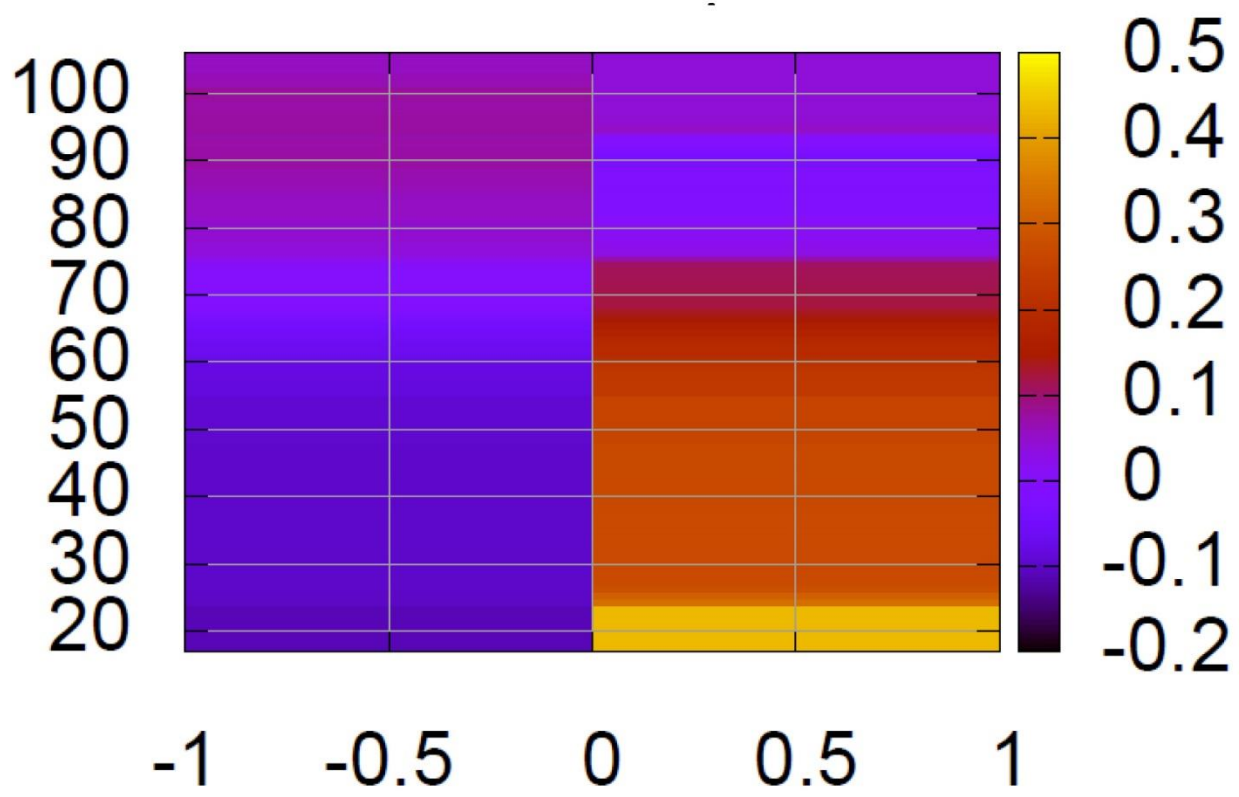
Like XOR (parity), interactions can't be modeled as a sum of independent effects:

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



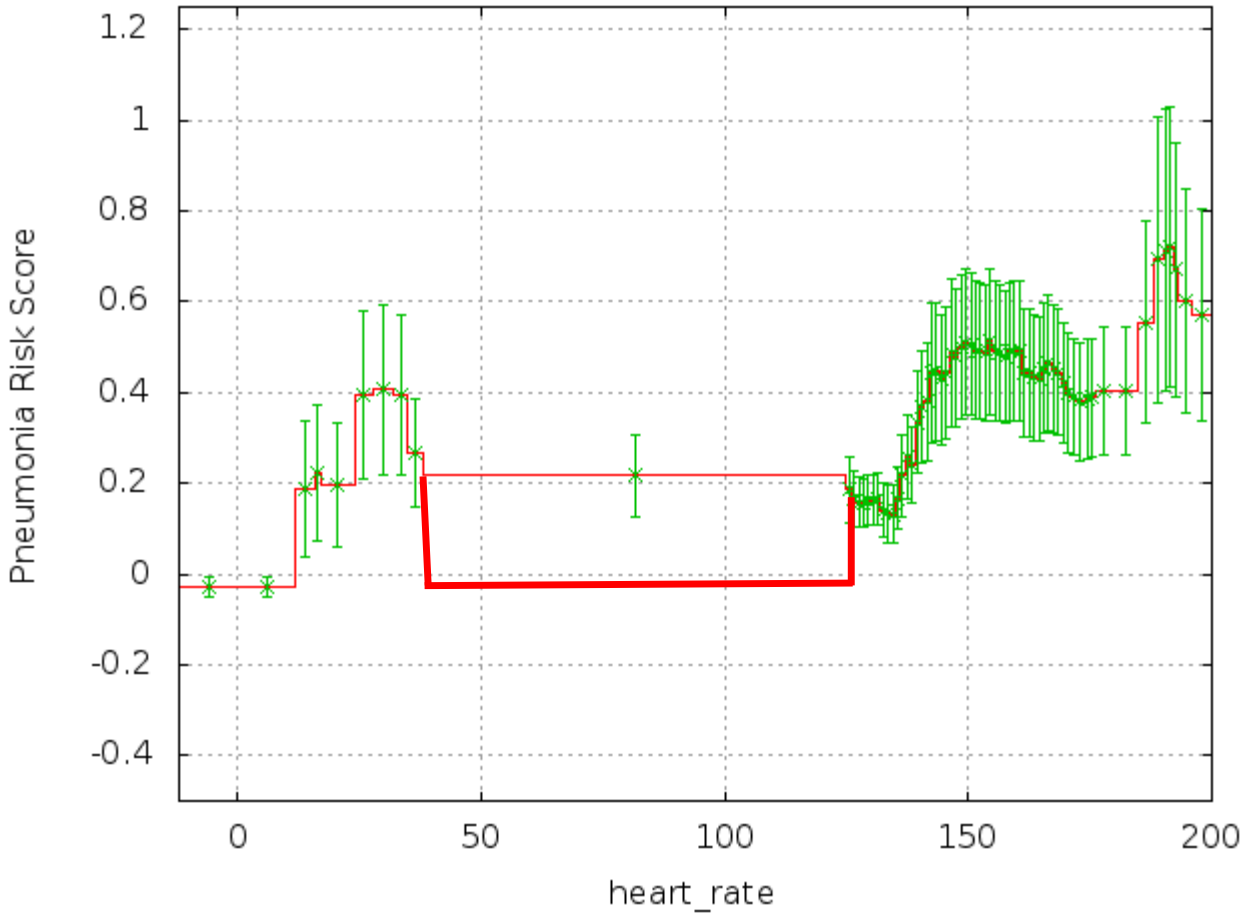


Pairwise Interaction: Age x Cancer (Pneumonia-95)



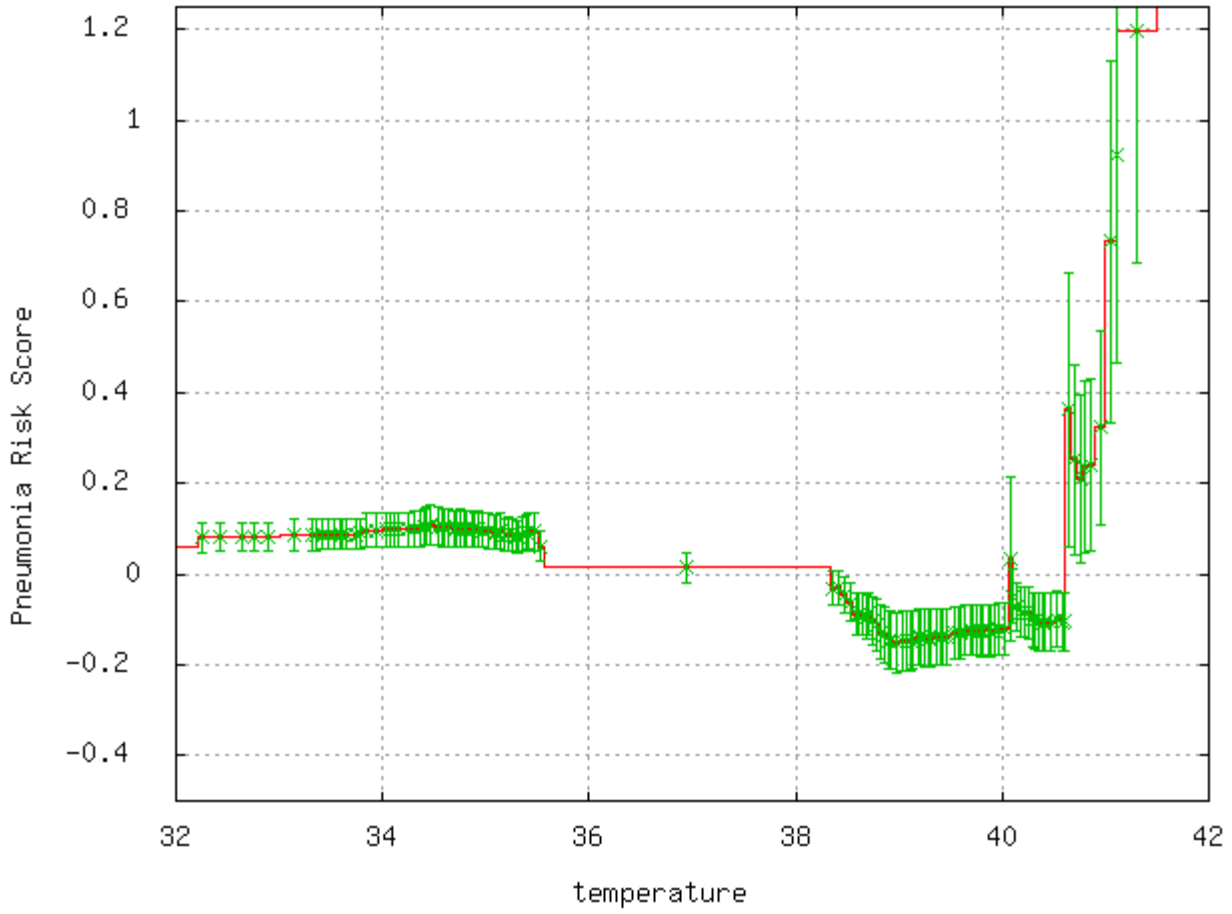
Age vs. Cancer

Pneumonia Dataset: Heart Rate (Pulse)



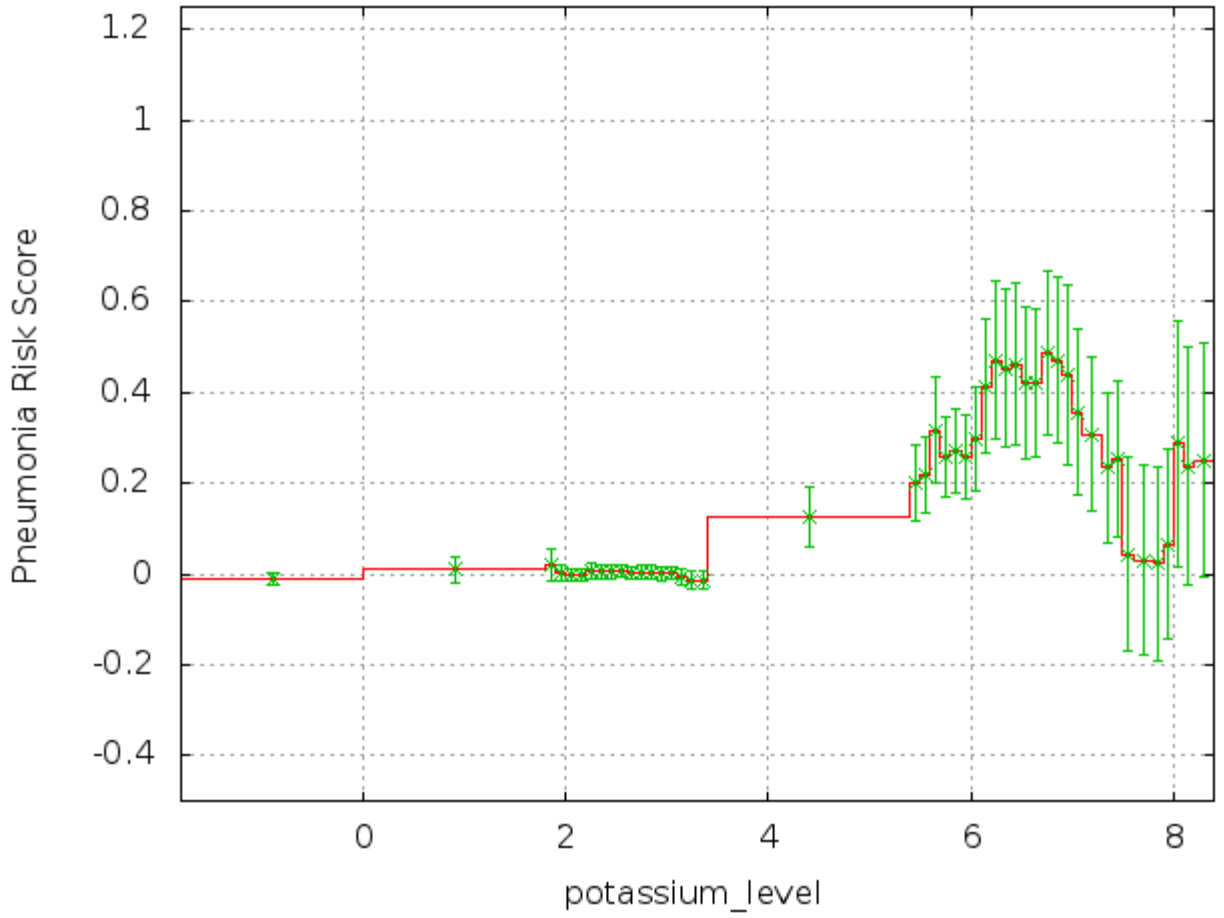
- What is the flat spot in middle for pulse = 40-125?
- 91% patients missing heart rate!
- Missing assumed normal
- Missing coded as 0
- Model sees no data 40-125
- Model interpolates between HR=39 and HR=126
- Would yield bad predictions for normal patients if HR collected!
- Can edit EBM graph to repair

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

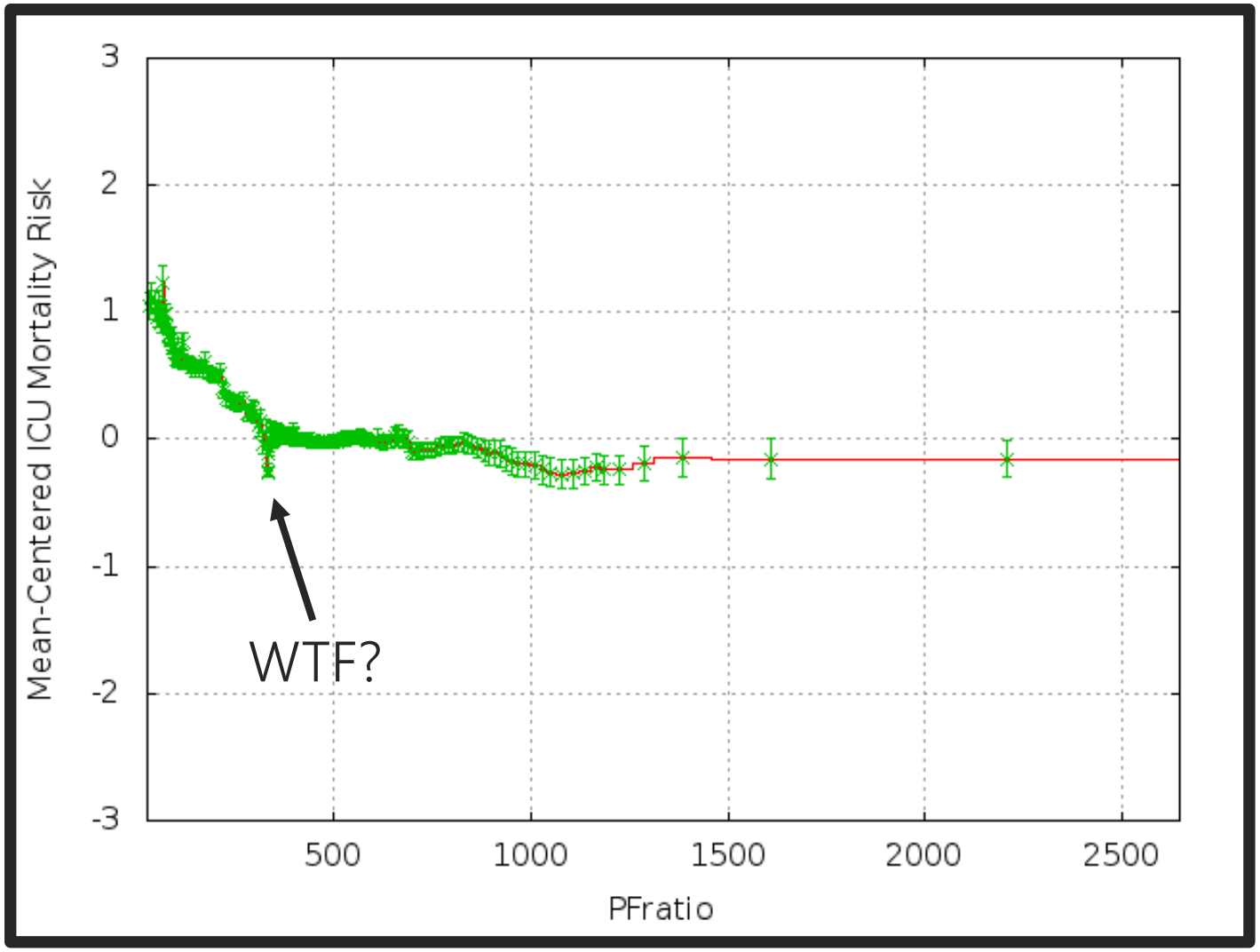
Pneumonia Dataset: Blood Potassium Level



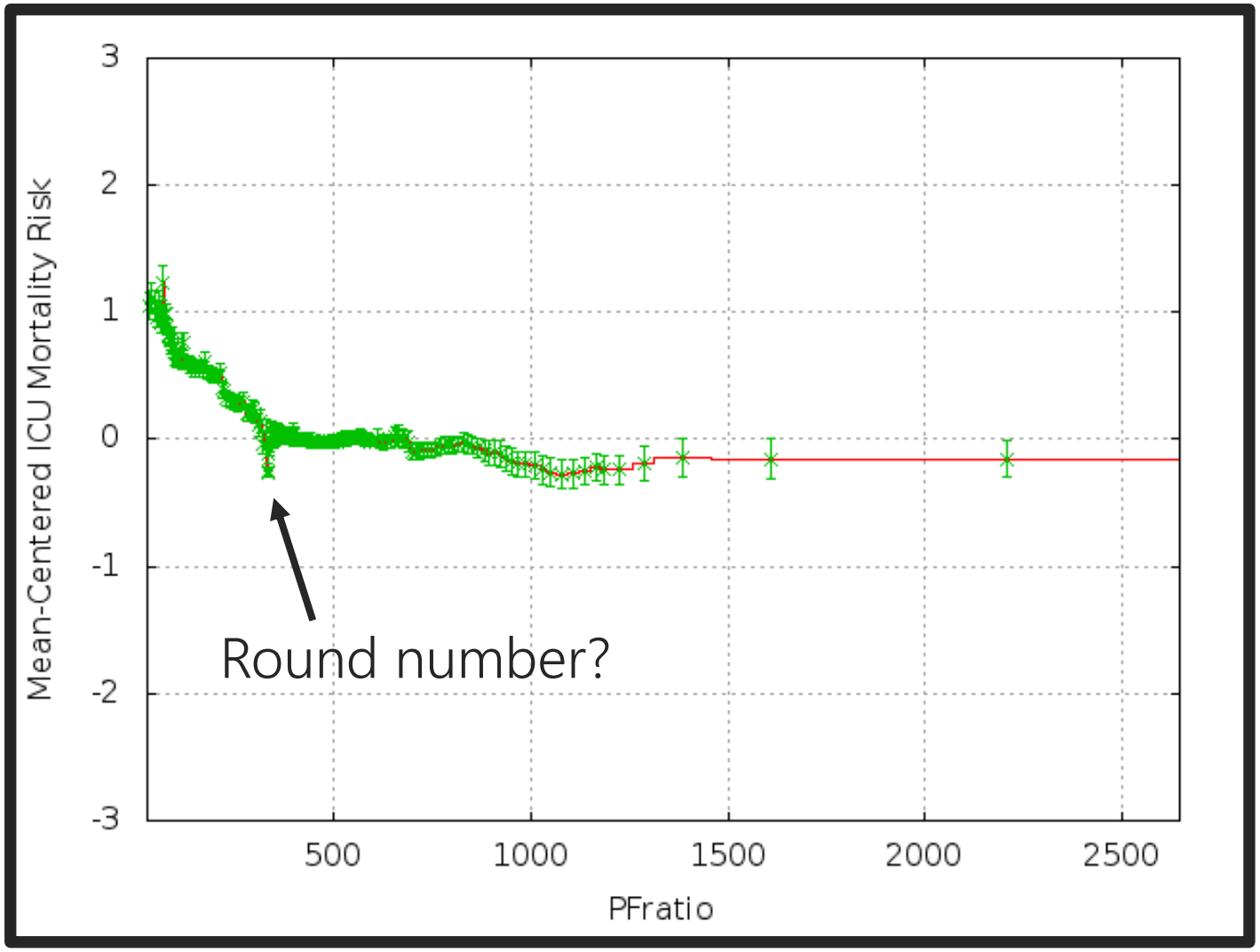
- What is the flat spot in middle for potassium = 3.5-5.2?
- 78% patients missing potassium!
- Missing assumed normal
- Missing coded as 0
- Model sees no data 3.5-5.2
- Model interpolates over missing
- In this case, model will yield mildly incorrect predictions for patients with normal potassium if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

Case Study 2: MIMIC-II ICU Mortality

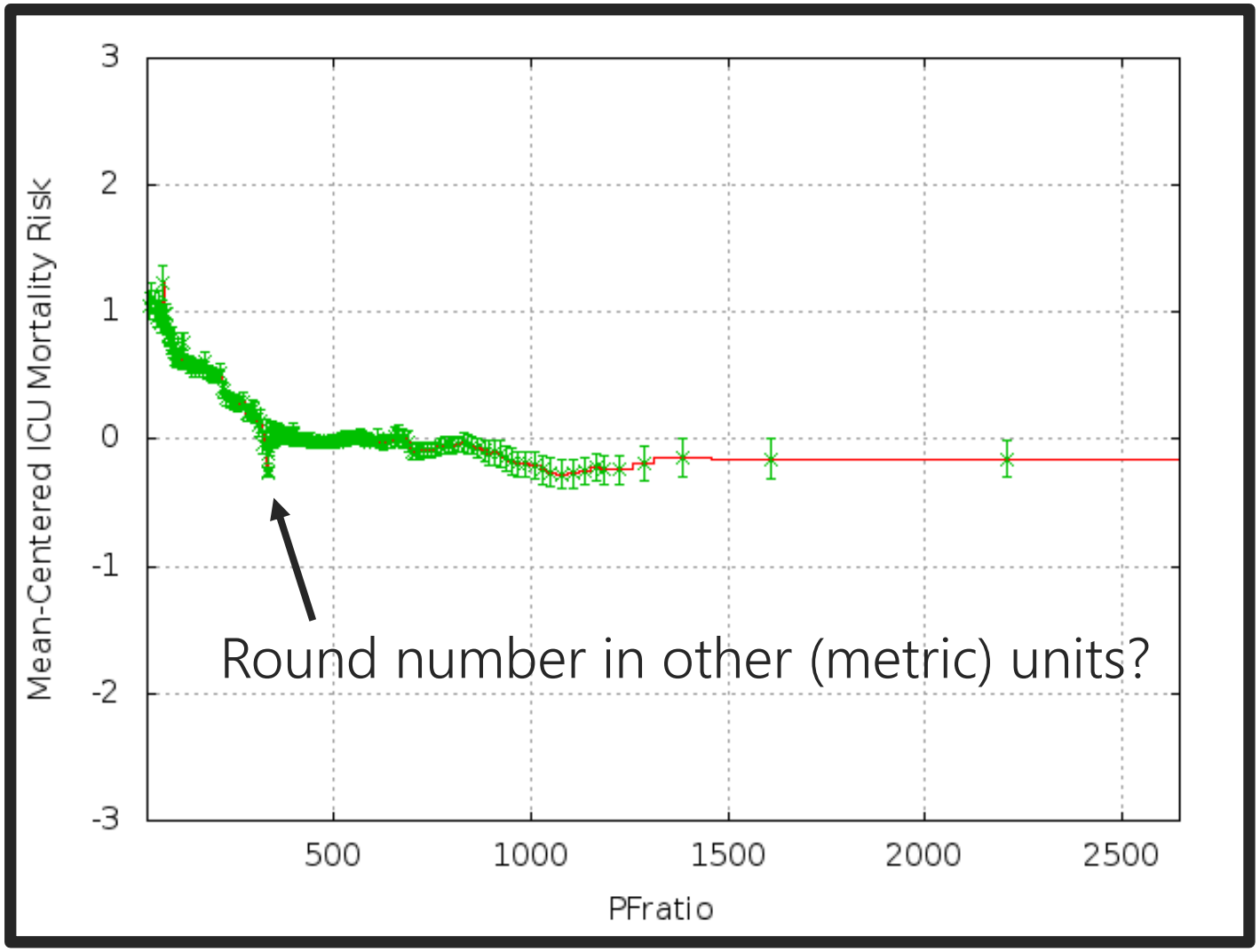
Intelligibility Helps Debug Data: PaO2/FiO2 Ratio



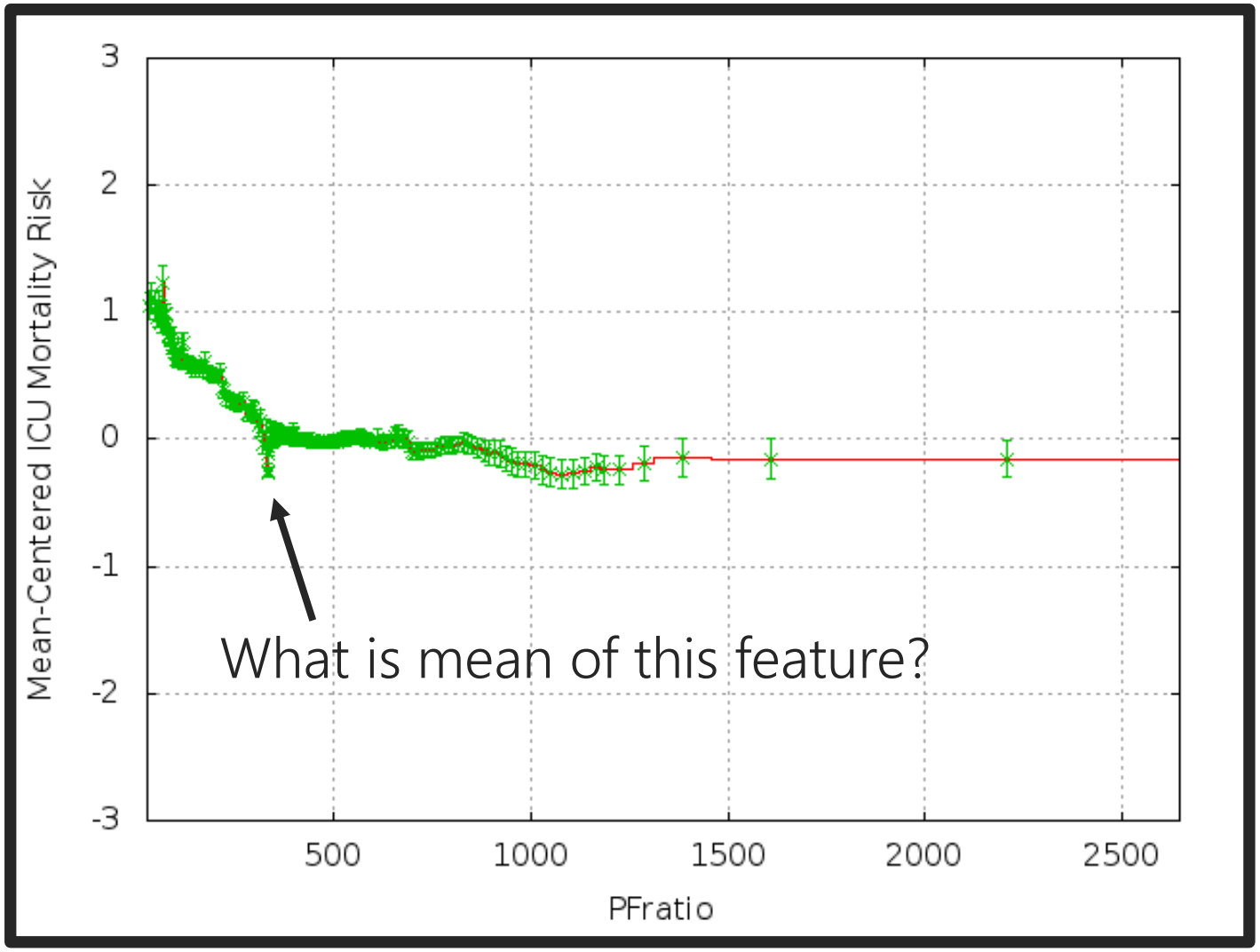
Intelligibility Helps Debug Data: PaO2/FiO2 Ratio



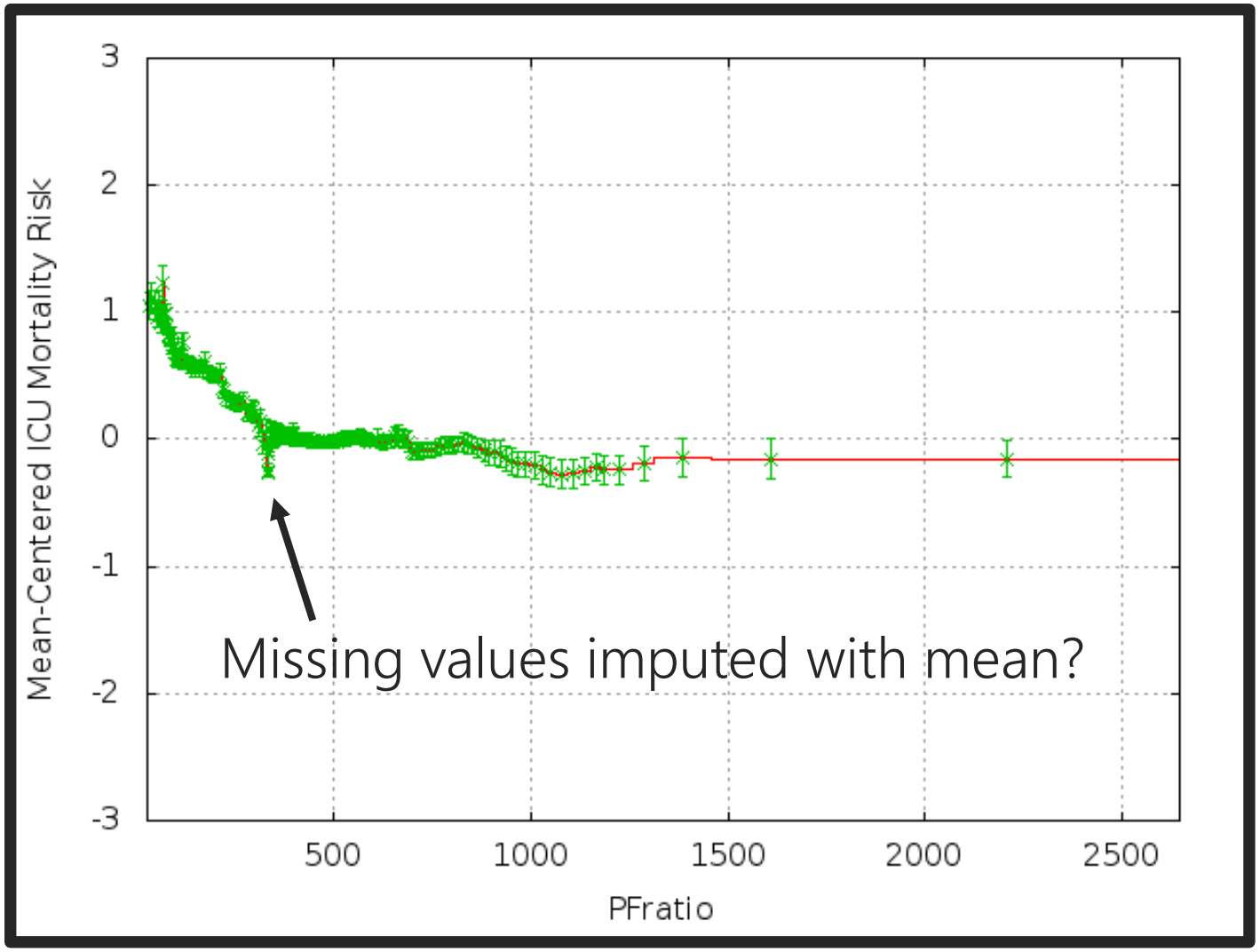
Intelligibility Helps Debug Data: PaO2/FiO2 Ratio



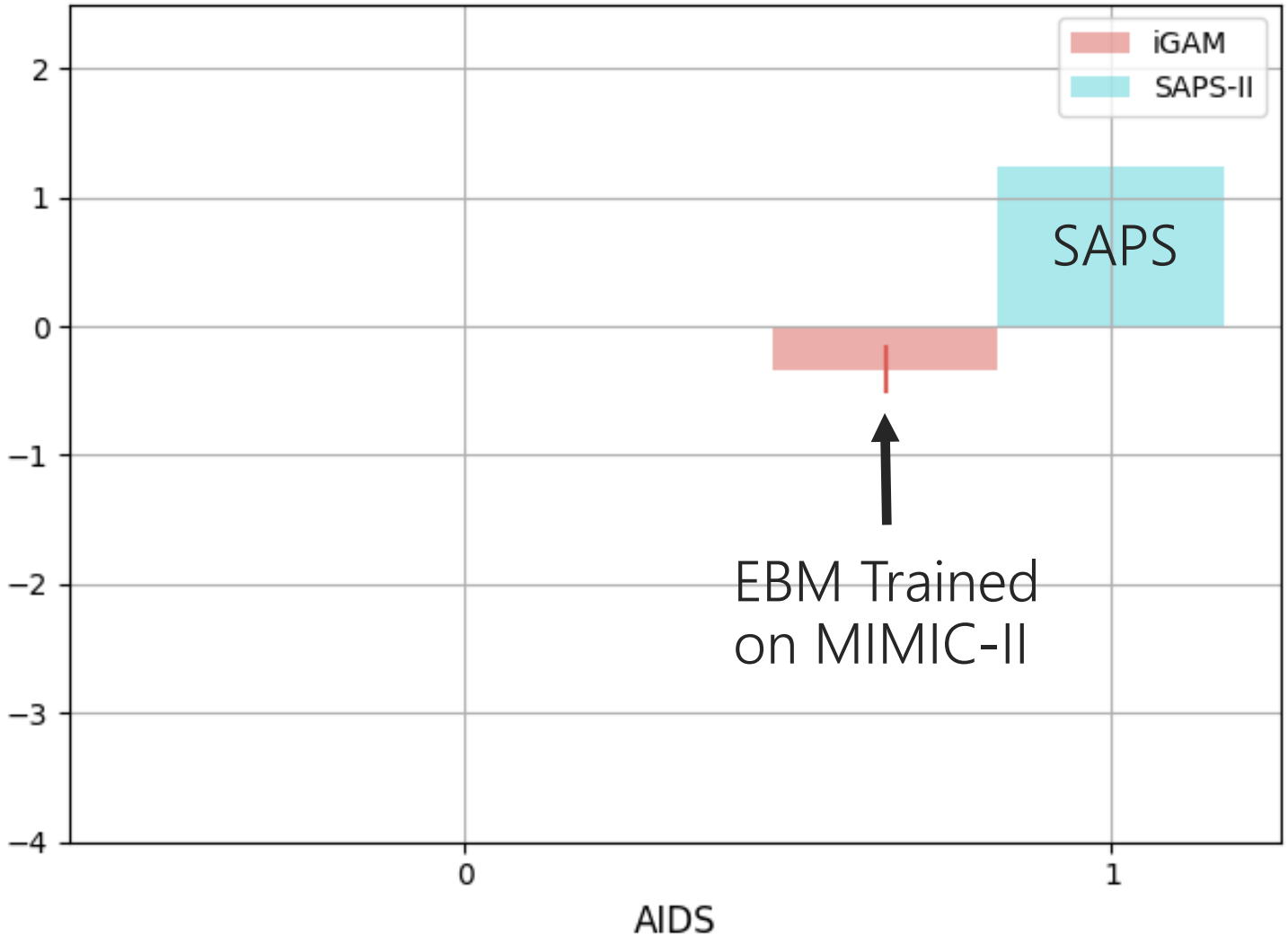
Intelligibility Helps Debug Data: PaO2/FiO2 Ratio



Intelligibility Helps Debug Data: PaO2/FiO2 Ratio

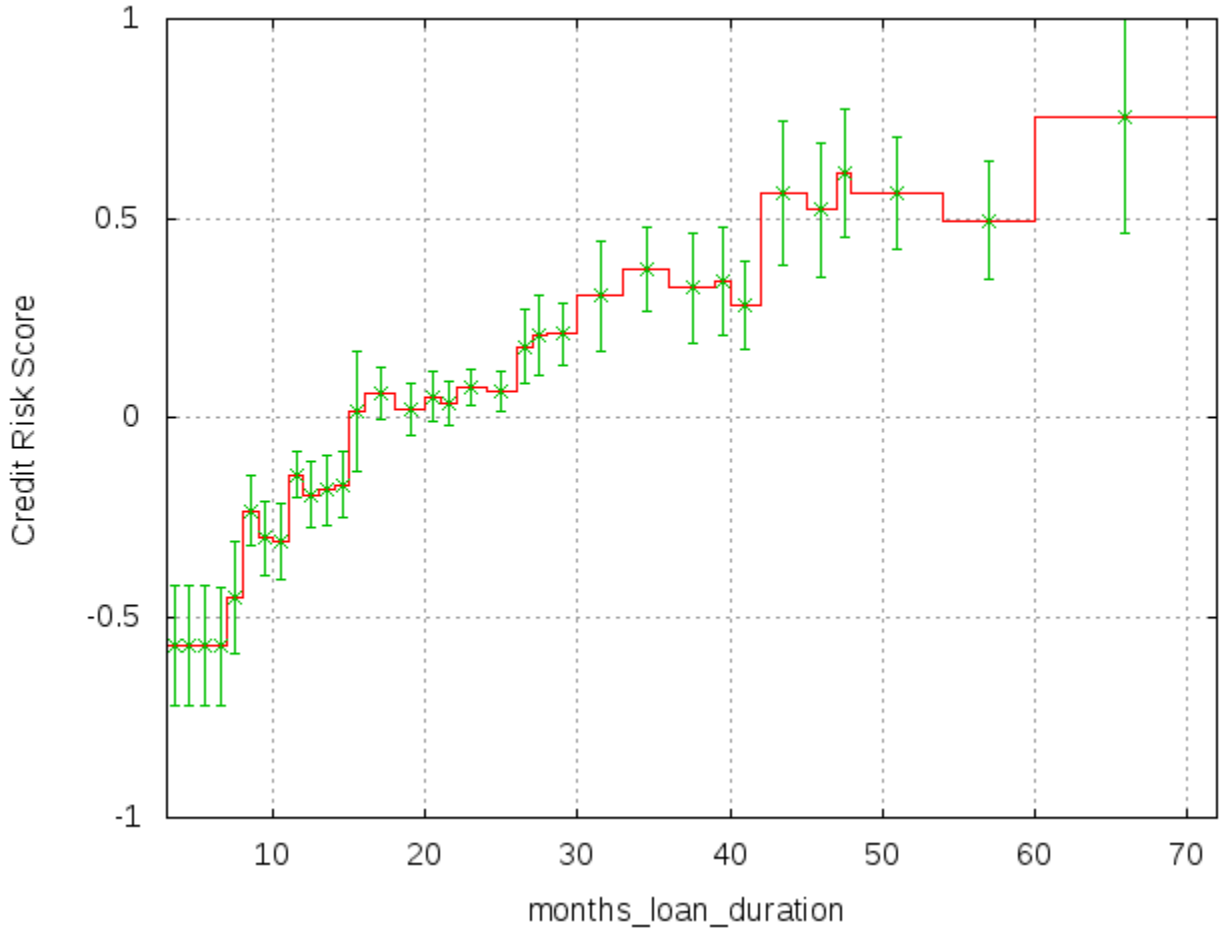


SAPSII Calculator vs EBMs: HIV/AIDS ???

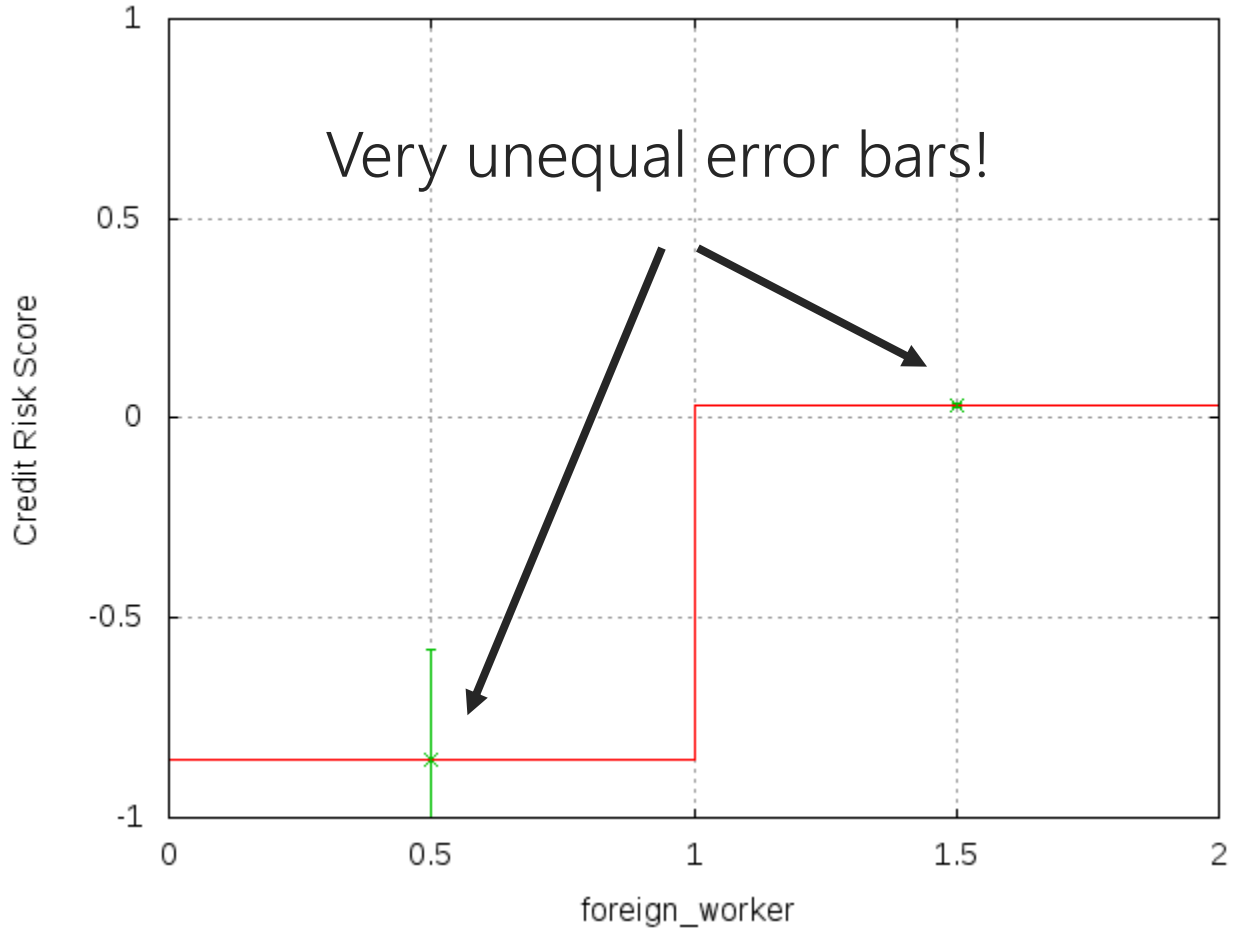


Case Study 3: German Credit Dataset

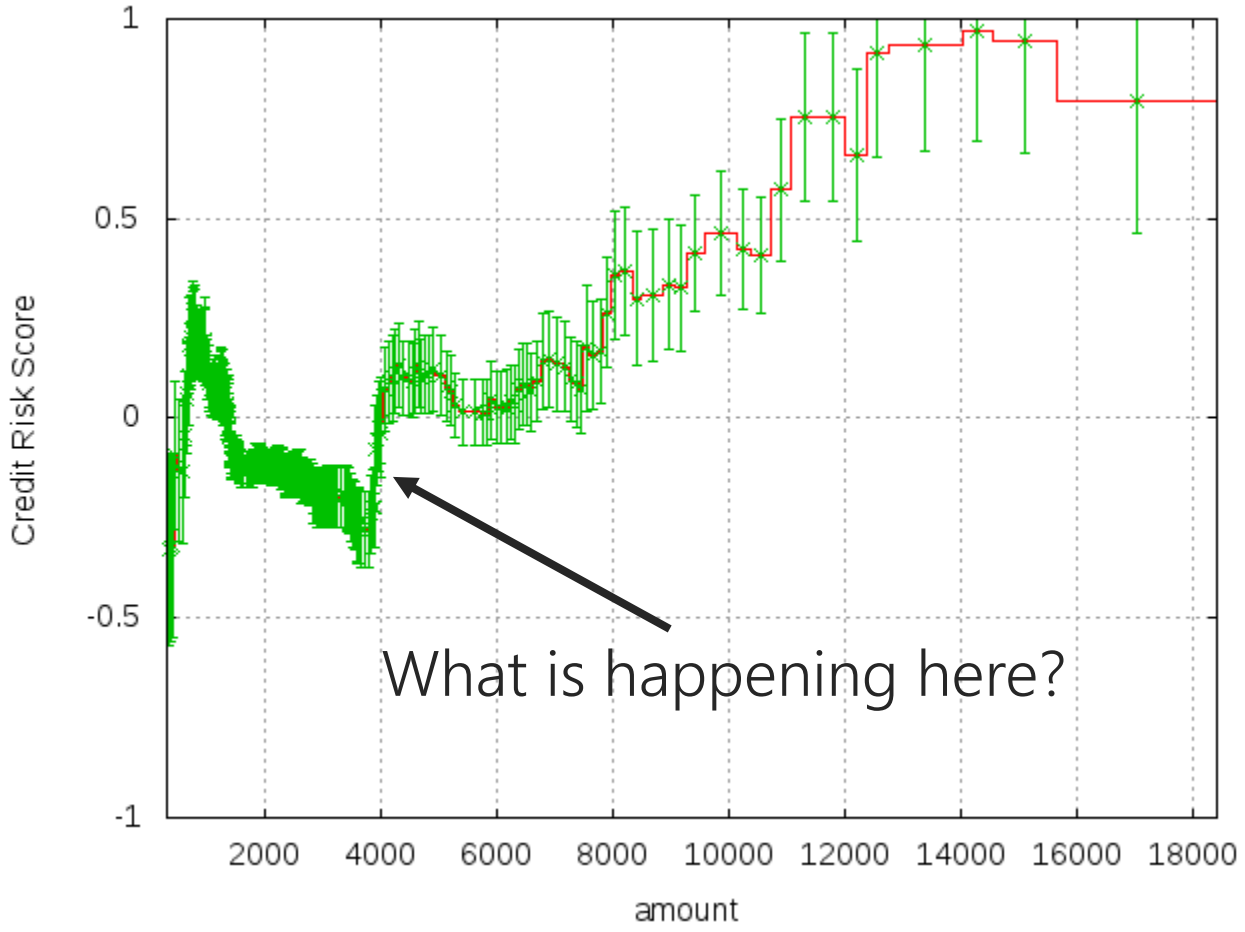
German Credit Dataset



German Credit Dataset: 1st Hint of Problems



German Credit Dataset



This dataset should never be used for research in fairness, bias and transparency

Case Study 4: 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

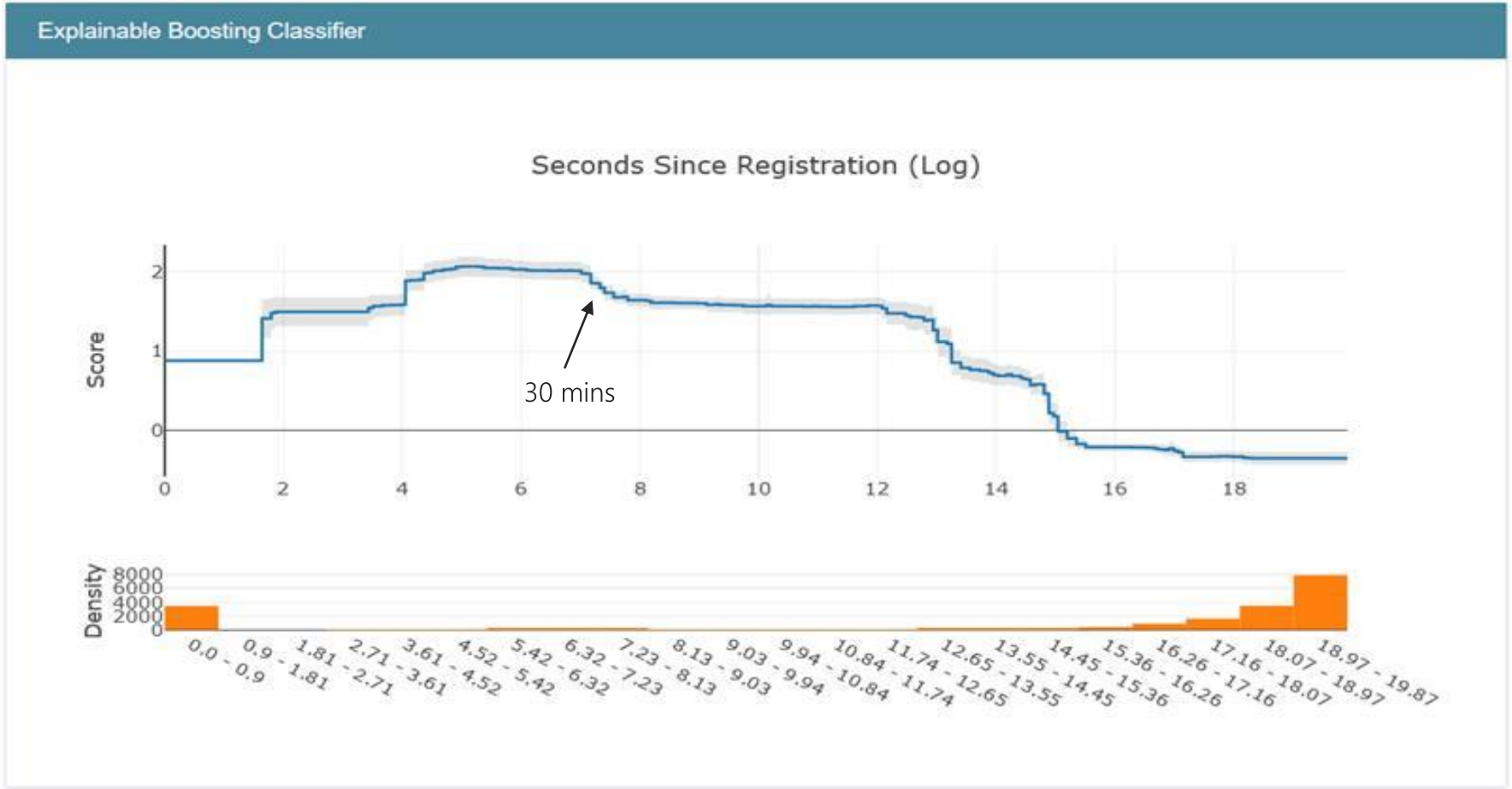
- 160,000 edits per day, 10-15% of which are flagged as damaging (e.g., malicious)
- Current ML tools are not intelligible, do not give help or explanations to editors

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

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

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

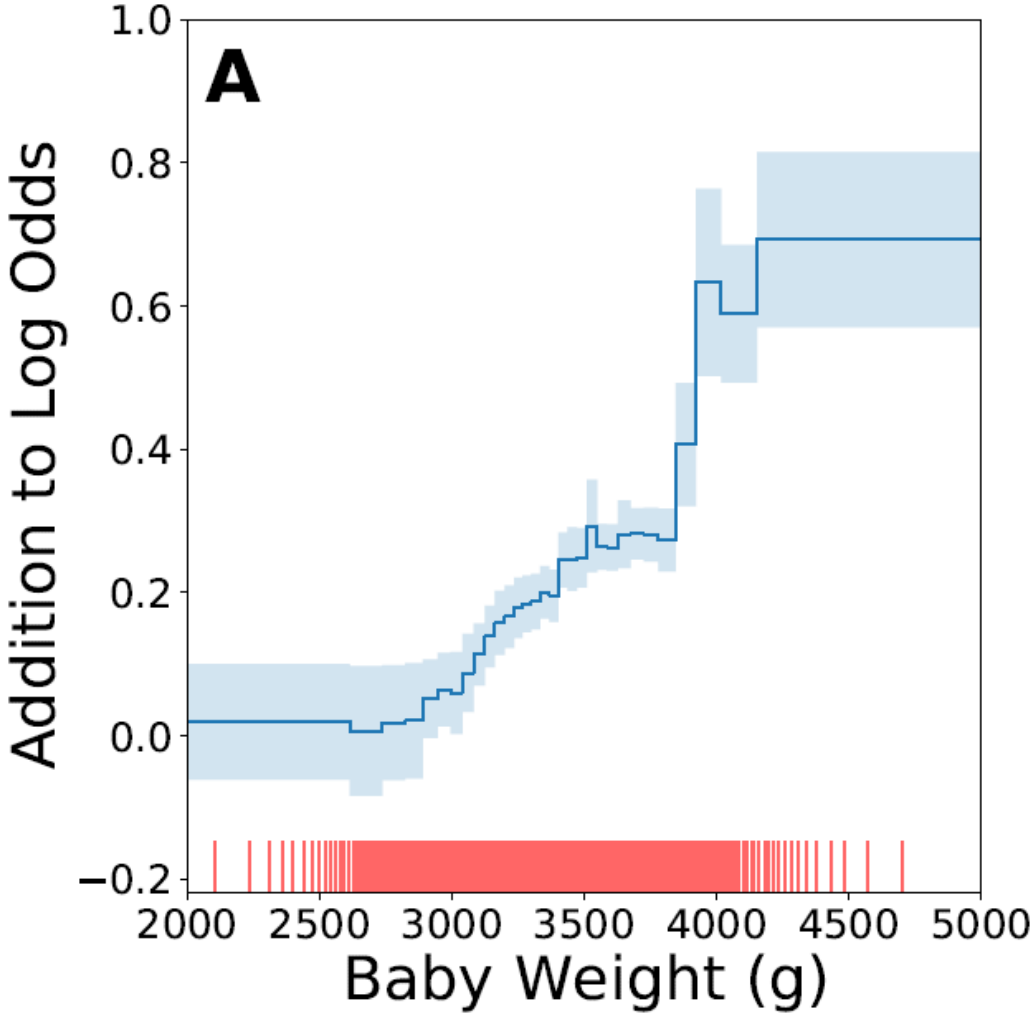
Case Study 5: Maternal Morbidity

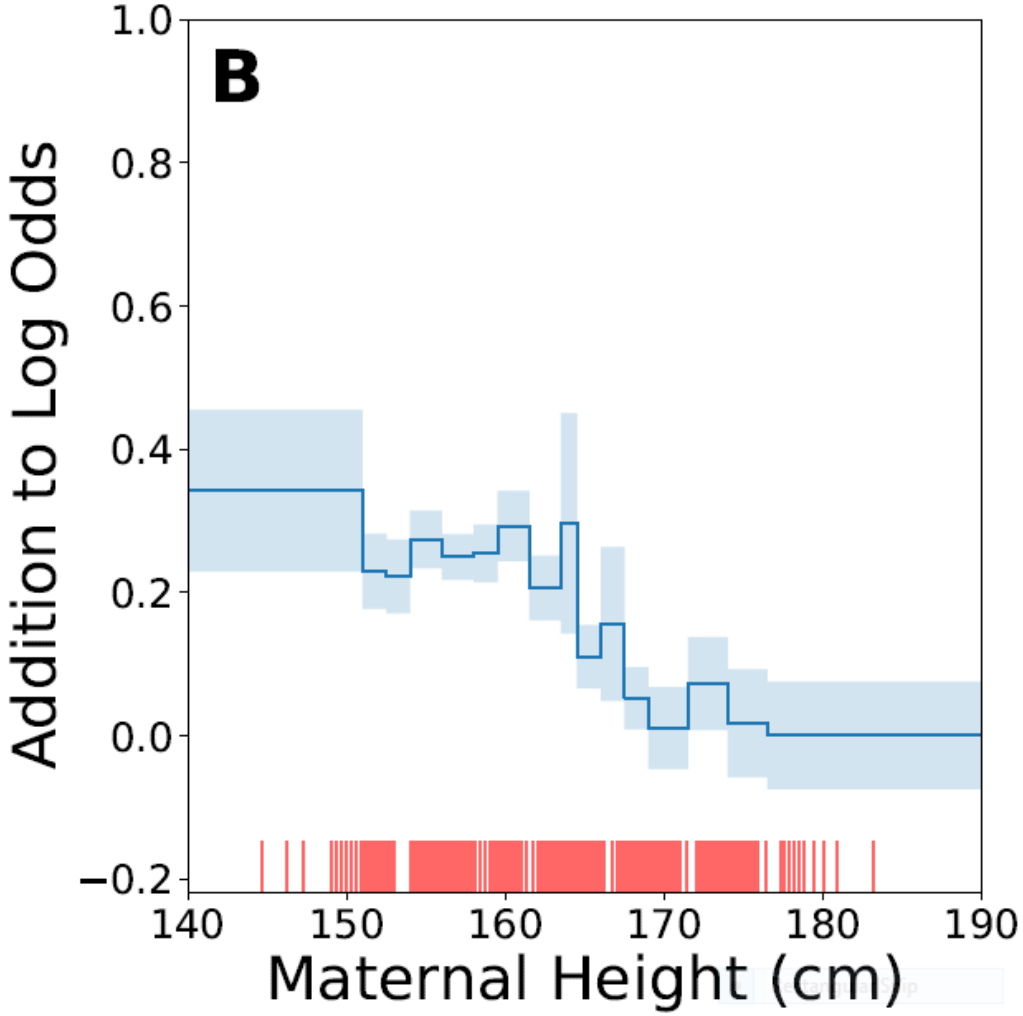
Predicting Severe Maternal Morbidity (SMM)

- SMM is about predicting maternal risk:
  - Hemorrhage, hysterectomy, thromboembolism, need for blood transfusion, eclampsia, …
  - Usual factors considered are factors about maternal health
    - Pre-eclampsia (maternal hypertension caused by pregnancy)
    - Diabetes
    - Maternal BMI
    - …

- But when we train EBMs of SMM data, most important factors are very different…
  - Baby weight (i.e., size of baby)
  - Maternal height (i.e., frame size of mother)
  - Pre-eclampsia

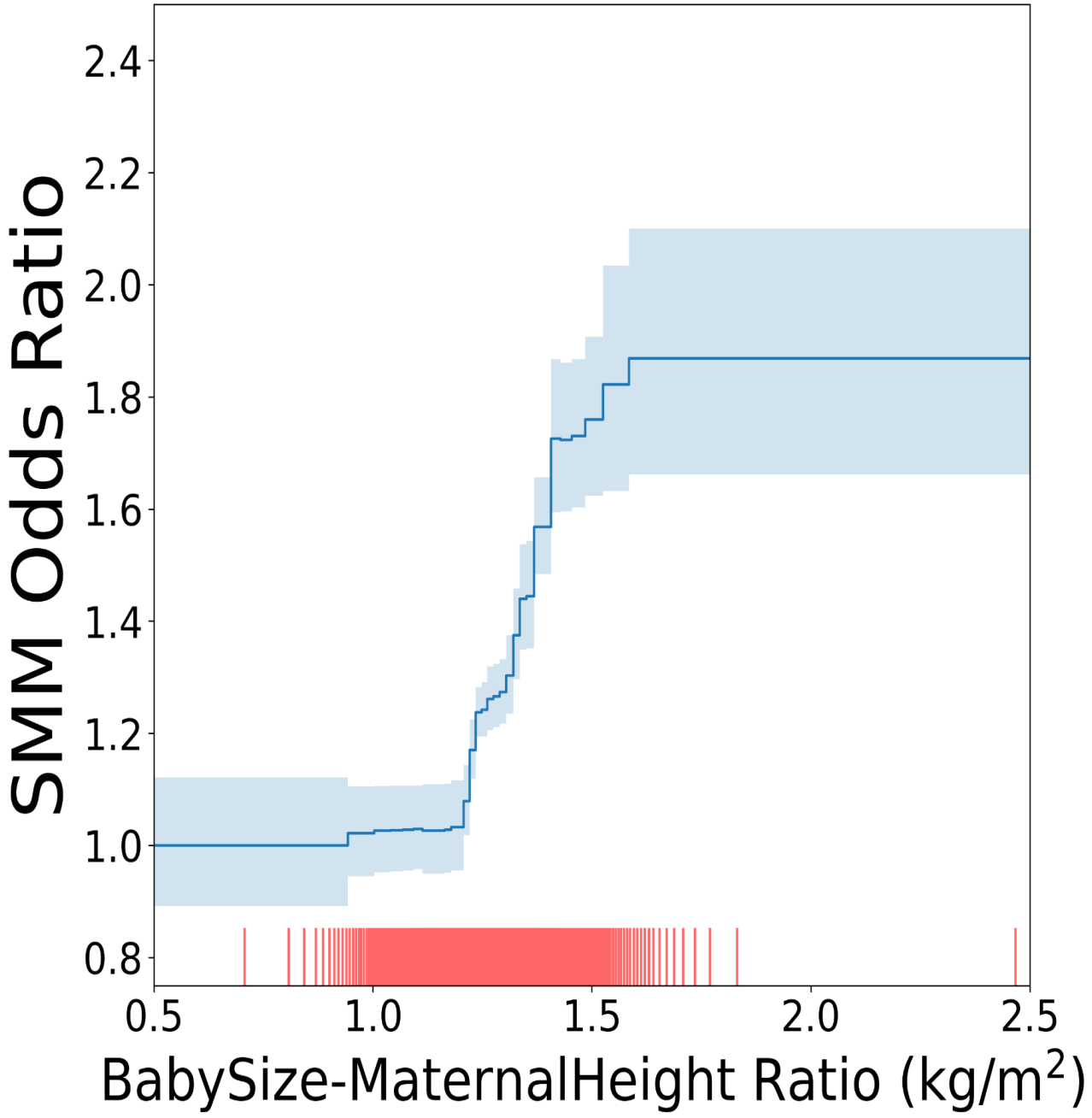
Predicting Severe Maternal Morbidity (SMM)





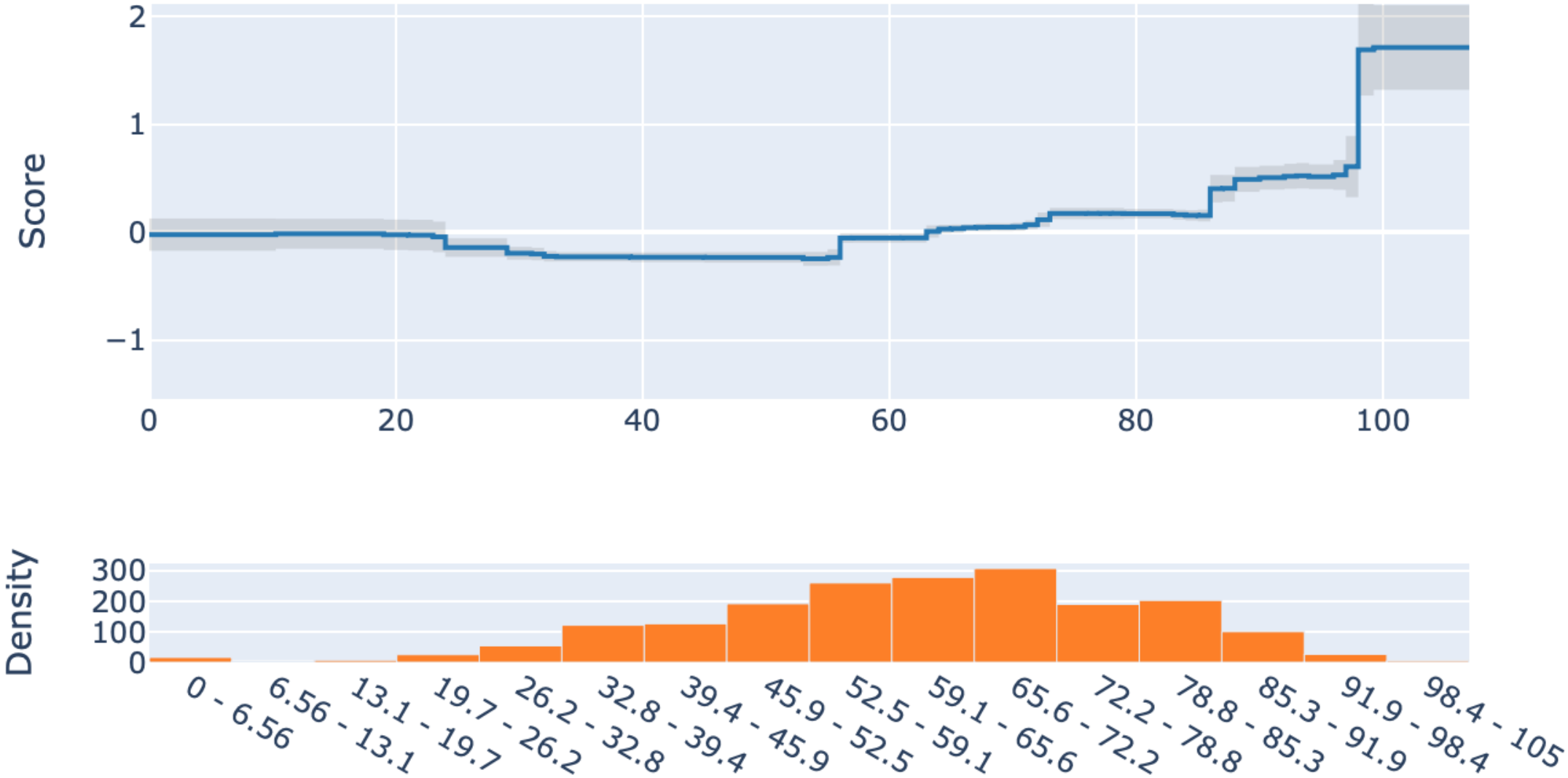
“BMI” for Pregnancy

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



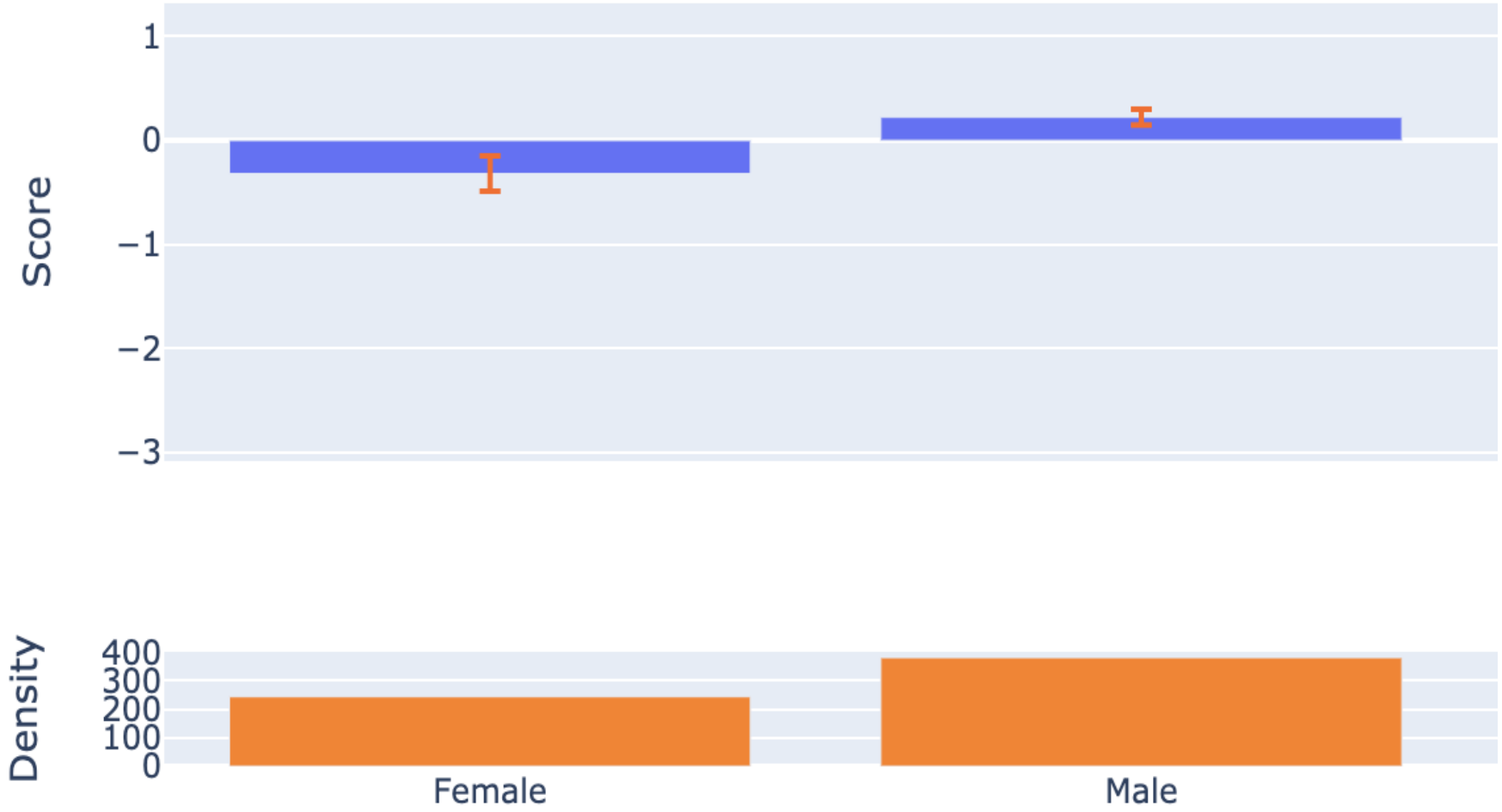
Case Study 6: COVID-19

COVID-19: Mortality Risk vs Age



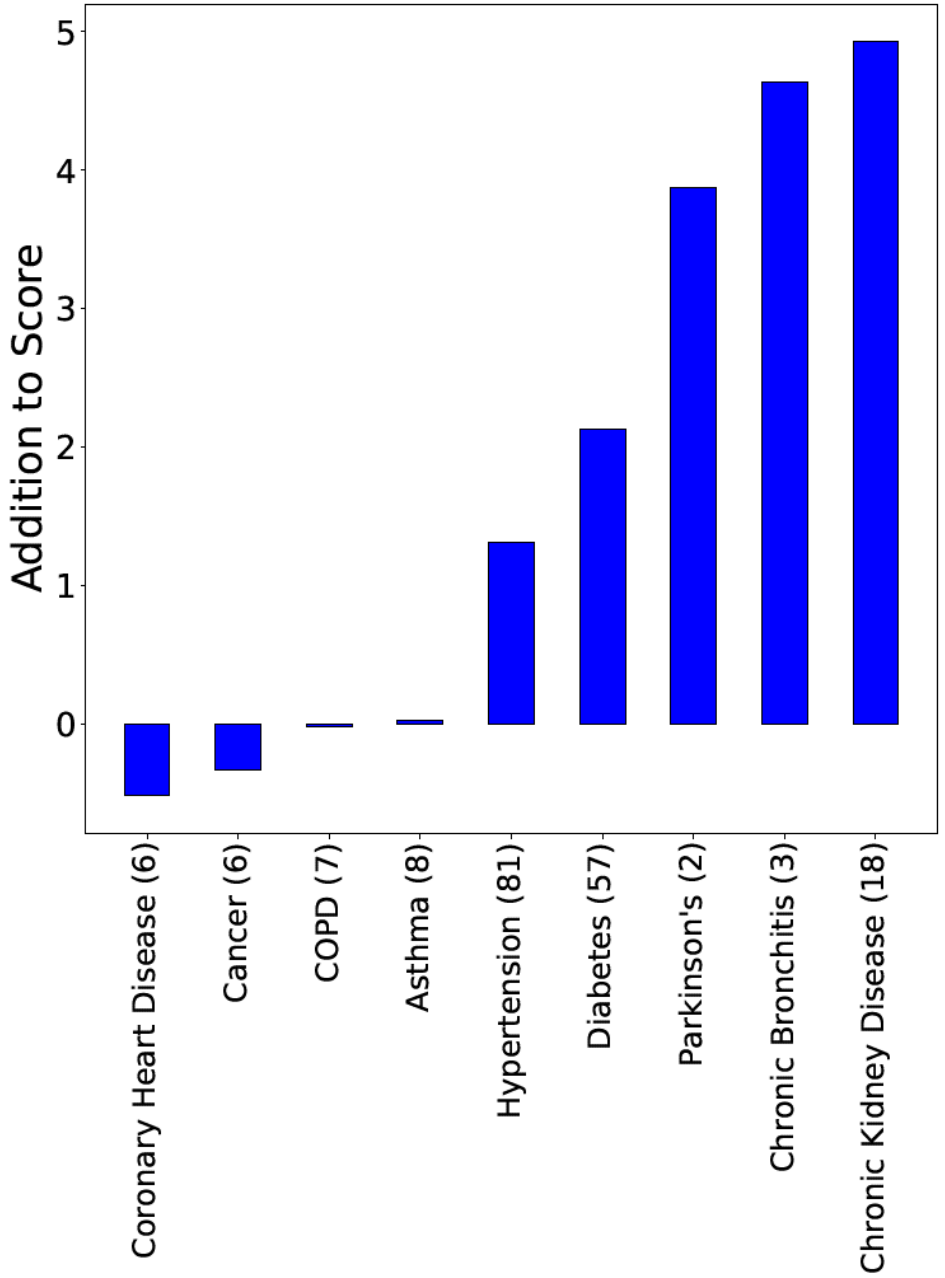
Joint work with Ben Lengerich (CMU)

COVID-19 Mortality Risk vs Gender



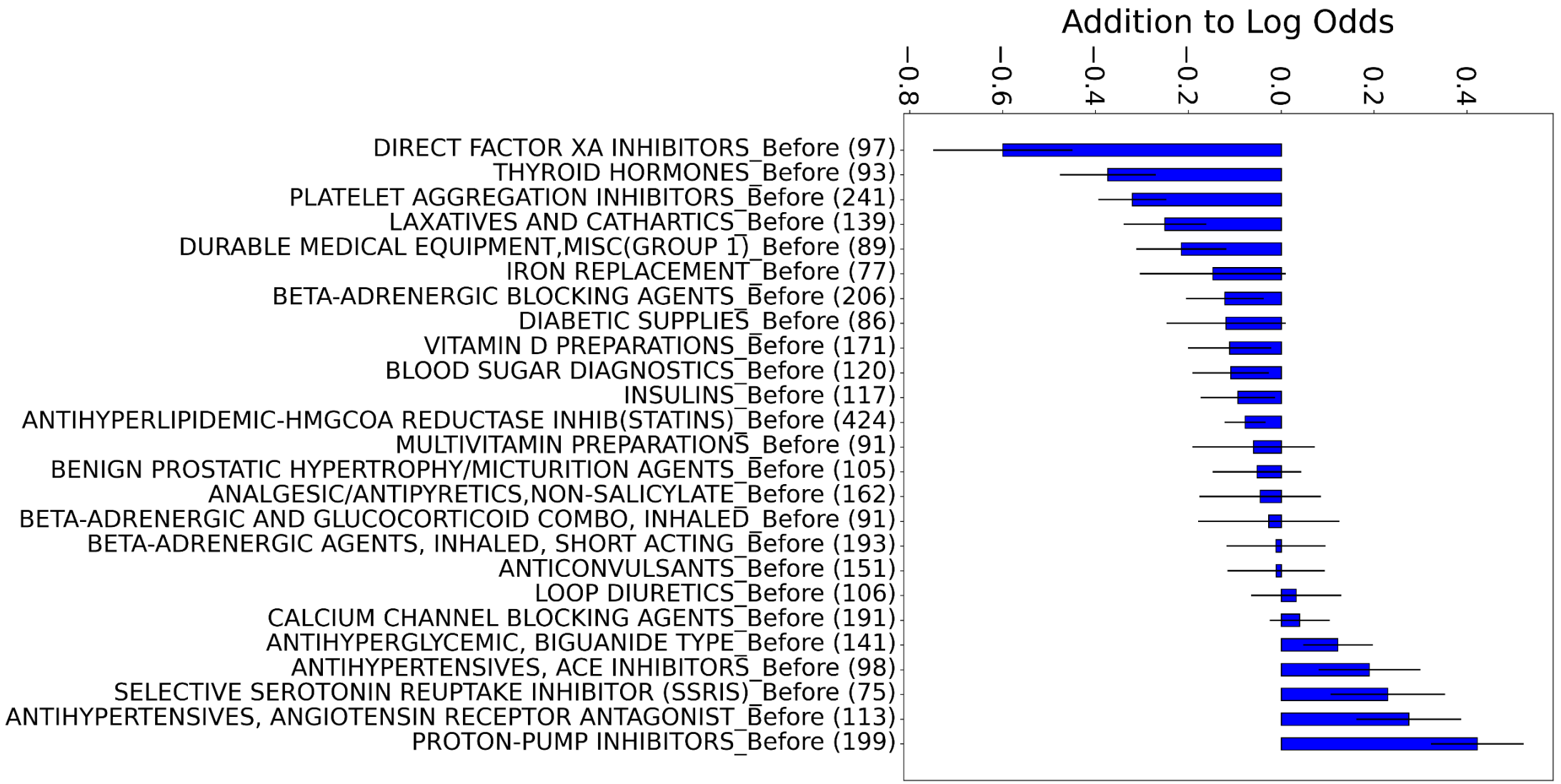
Joint work with Ben Lengerich (CMU)

COVID-19 Risk Factors



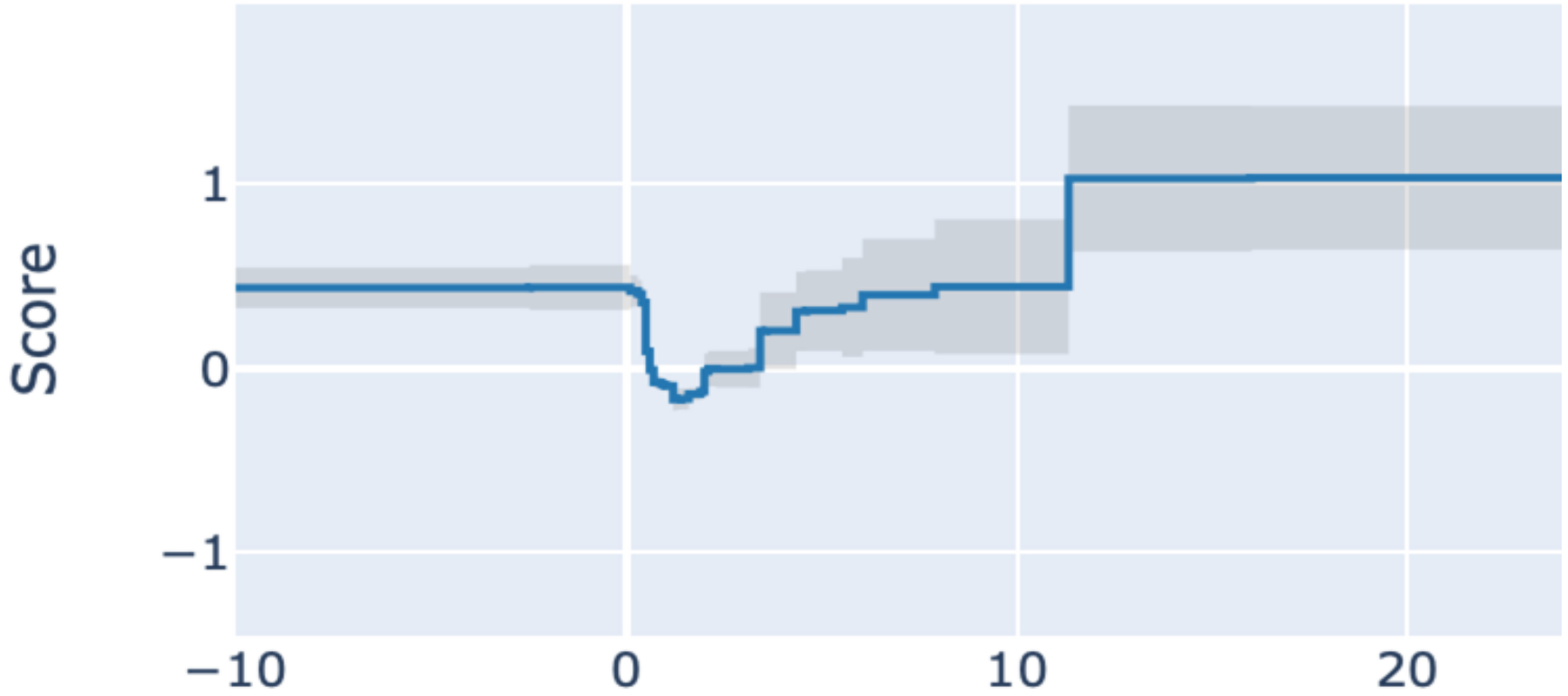
Joint work with Ben Lengerich (CMU)

P(Mortality), adjusting for lab tests, comorbidities, drugs before.



Joint work with Ben Lengerich (CMU)

First Discovery: LYMPHOCYTES_ABSOLUTE_a



Joint work with Ben Lengerich (CMU)

Case Study 7: 30-Day Hospital 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 EBMs 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

30-day Hospital Readmission Example

Case Study 8: Bias & Recidivism Prediction

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

FAT*/ML: ProPublica COMPAS Recidivism Data

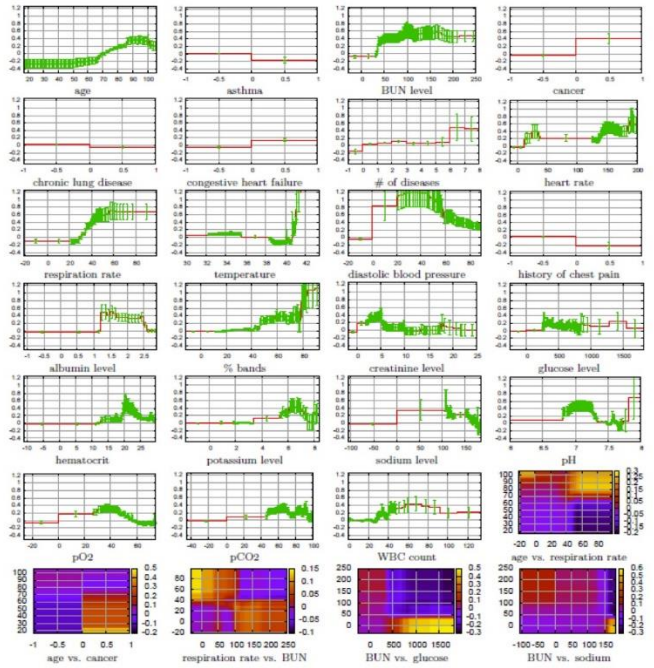
- COMPAS is a black-box model used to predict future criminal behavior
  - Model is black-box because it is protected by IP, not because it is a deep net
  - Criminal justice officials use risk prediction to inform bail, sentencing and parole decisions

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

Distillation Trick to Open Up Black-Box Models

Soft labels from Teacher

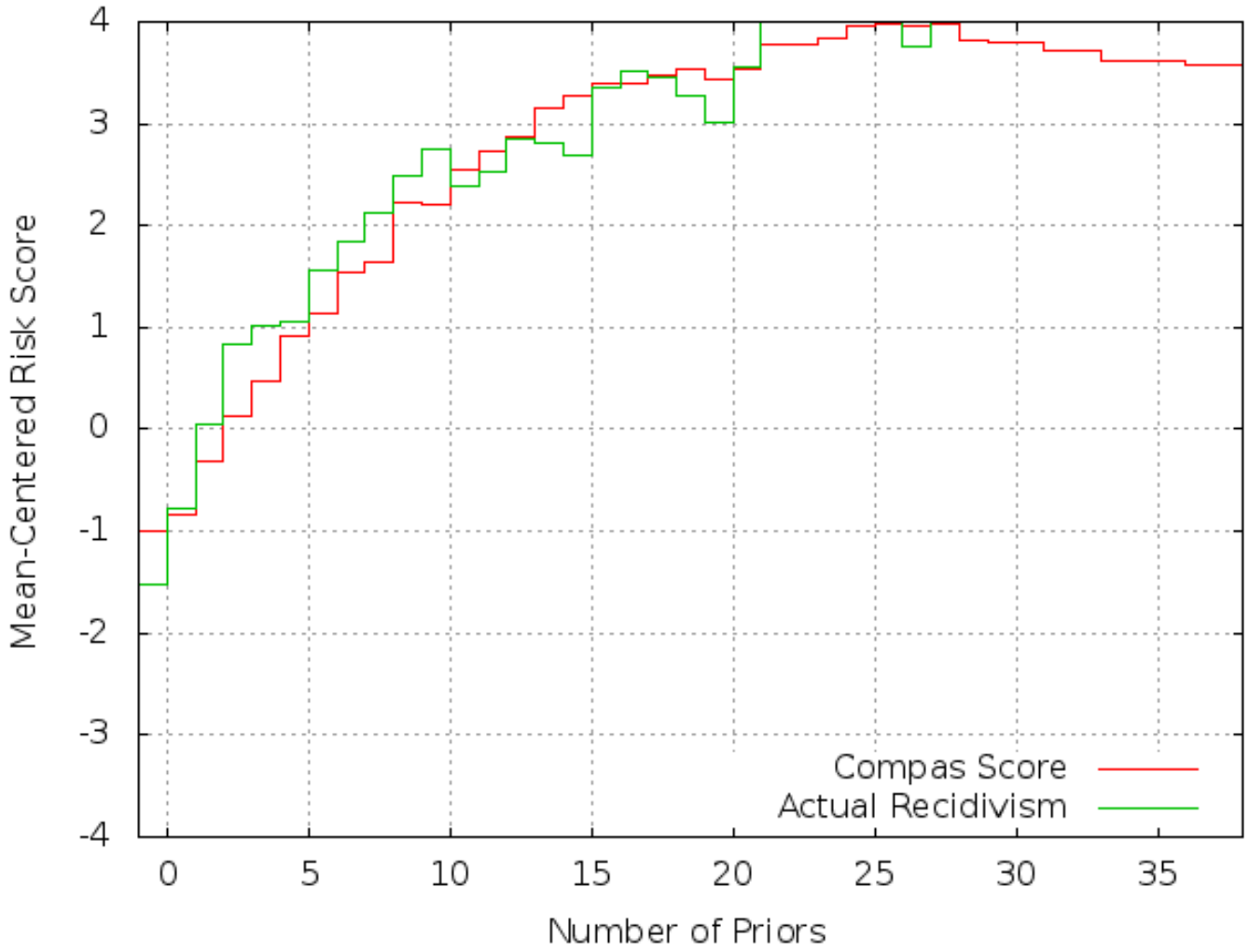
Black-Box Compas Model



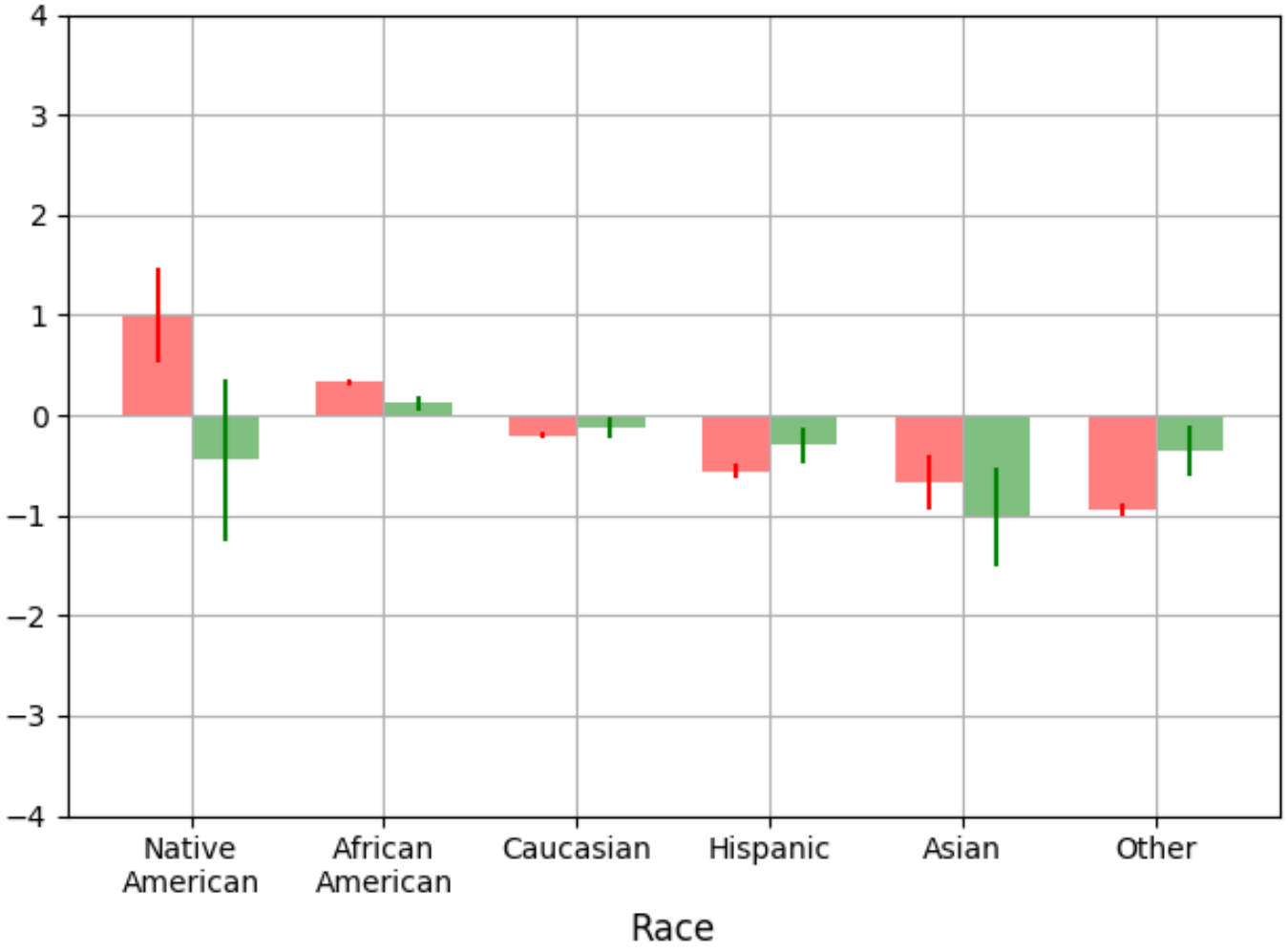
Black-Box Teacher Model

Glass-Box EBM Student Model

Recidivism Risk vs. Number of Prior Convictions



Recidivism Risk vs. Race



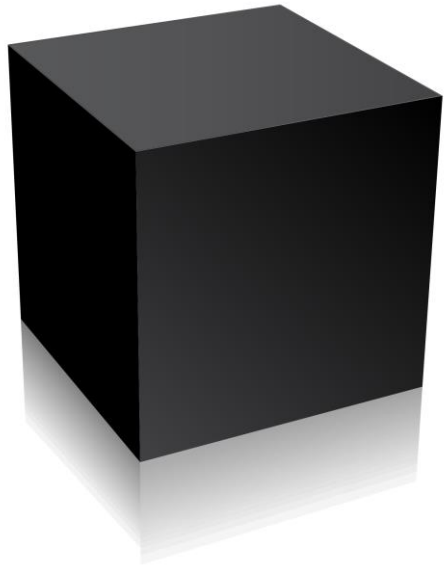
Summary

- Intelligibility & Explanation are critical in many domains
- Can also help debug data, build trust in model, and improve model accuracy

- If you must use/understand black-box models or complex pipelines:
  - LIME, SHAP, Partial Dependence, … or Distillation with EBMs

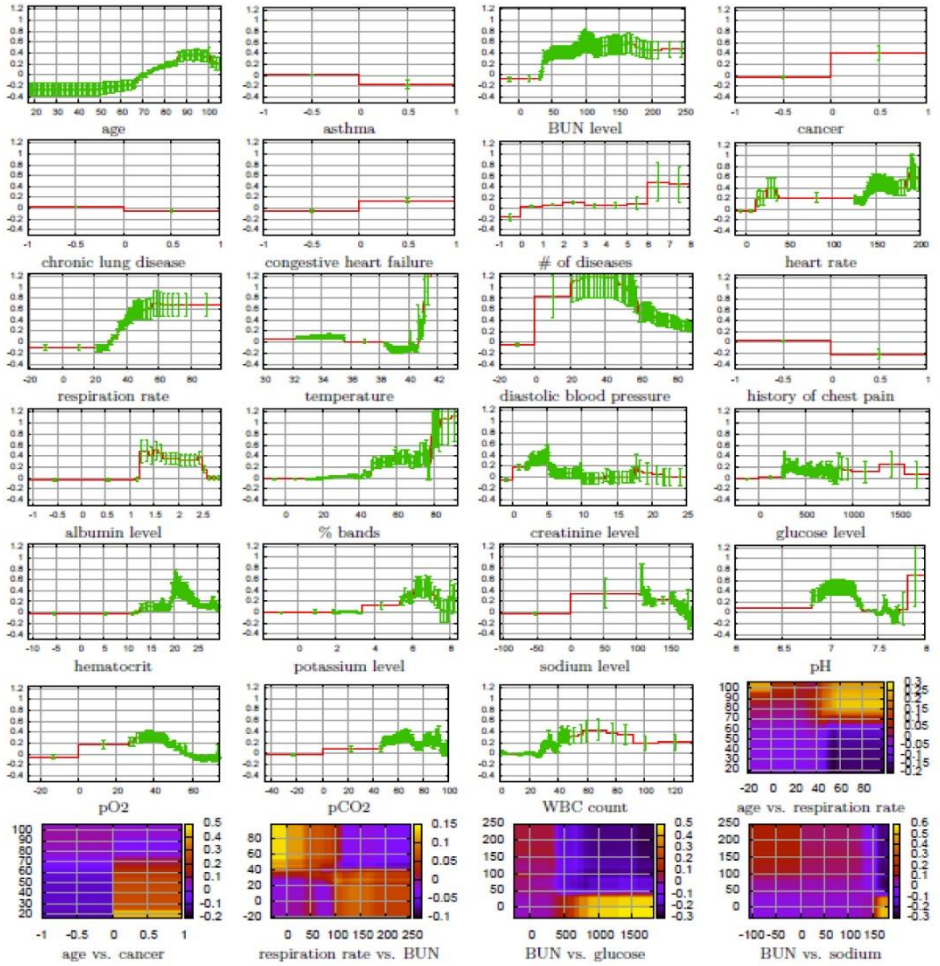
- If you are the one training the model, use glass-box models instead!
  - Usually little or no loss in accuracy (sometimes glass-box accuracy is better)
  - Explanations are exact --- no approximations
  - Models are editable --- easy way to repair models when they learn to do something stupid
  - EBMs are the current state-of-the-art in glass-box ML
  - Also very compact and fast at runtime

Black Box



or

EBM Intelligible Model



Thank You!

InterpretML

Open-Source Tool for Intelligibility

github.com/interpretml/interpret