Microsoft Research

Friends Don’t Let Friends Deploy Black Box Models:
The Importance of Intelligibility in Machine Learning

Rich Caruana

Microsoft

Microsoft Research

Friends Don’t Let Friends Deploy Black Box Models:
The Importance of Intelligibility in Machine Learning

Rich Caruana
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
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

Microsoft Research

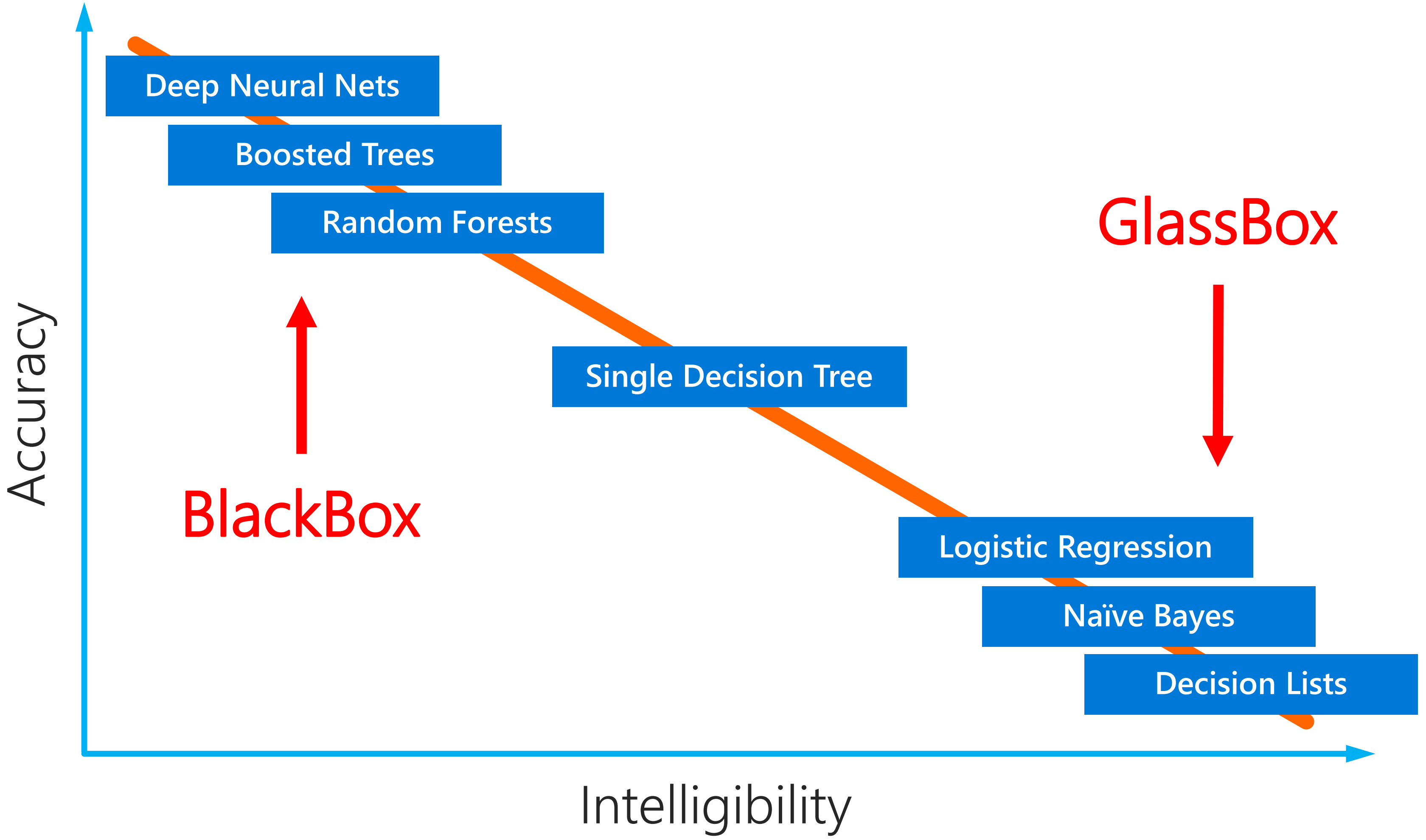
Friends Don't Let Friends Deploy Black Box Models:
The Importance of Intelligibility in Machine Learning

Rich Caruana

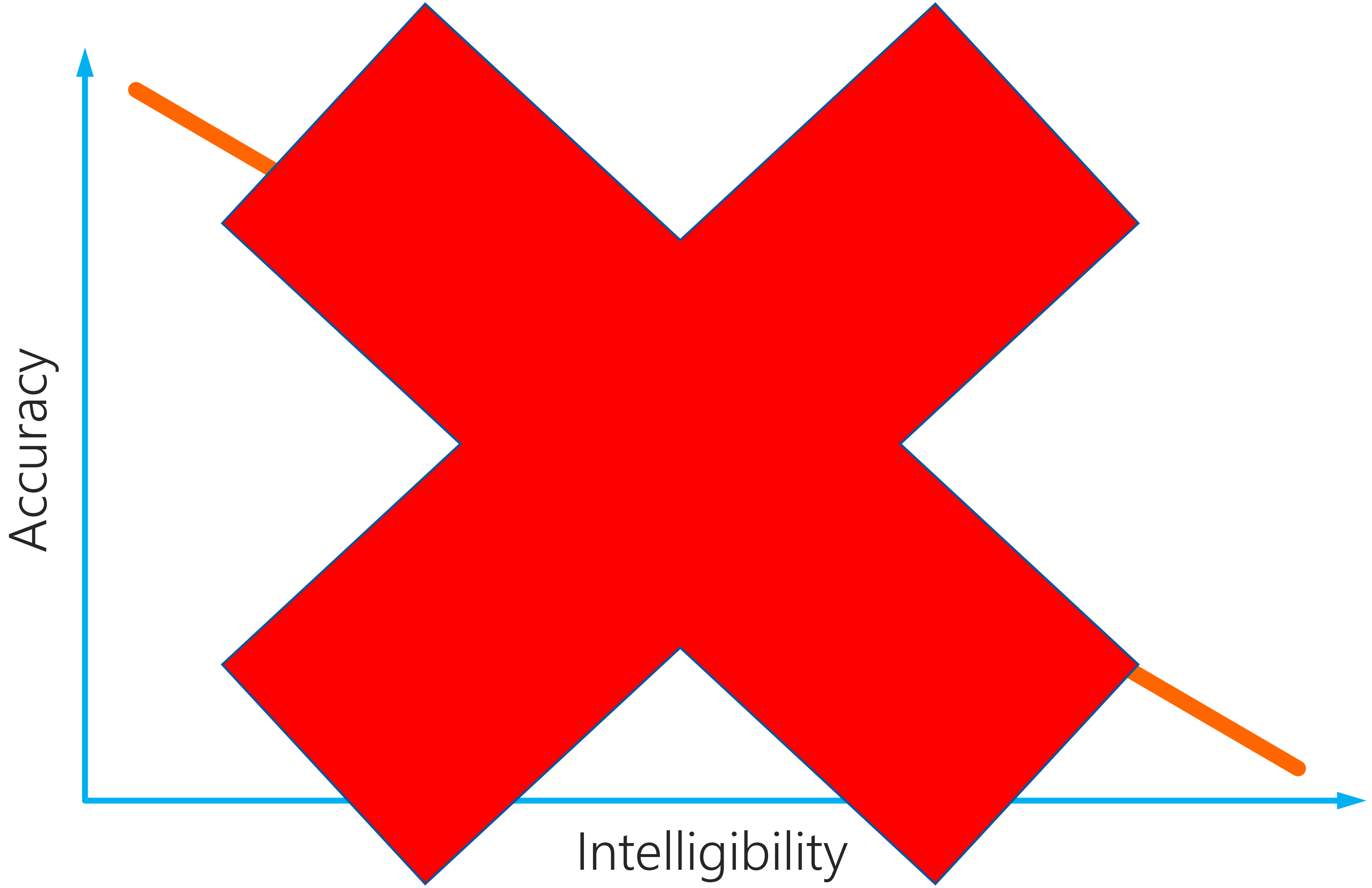
Yin Lou, Sarah Tan, Xuezhou Zhang, Ben Lengerich, Kingsley Chang, Jay Wang, Zhi Chen
Paul Koch, Harsha Nori, Sam Jenkins, Jessica Wolk, Luis França, Levi Melnick
Greg Cooper MD/PhD, Mike Fine MD, Vivienne Souter MD, Yin Aphinyanaphongs MD/PhD,
Giles Hooker, Johannes Gehrke, Tom Mitchell, Marc Sturm,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

Microsoft

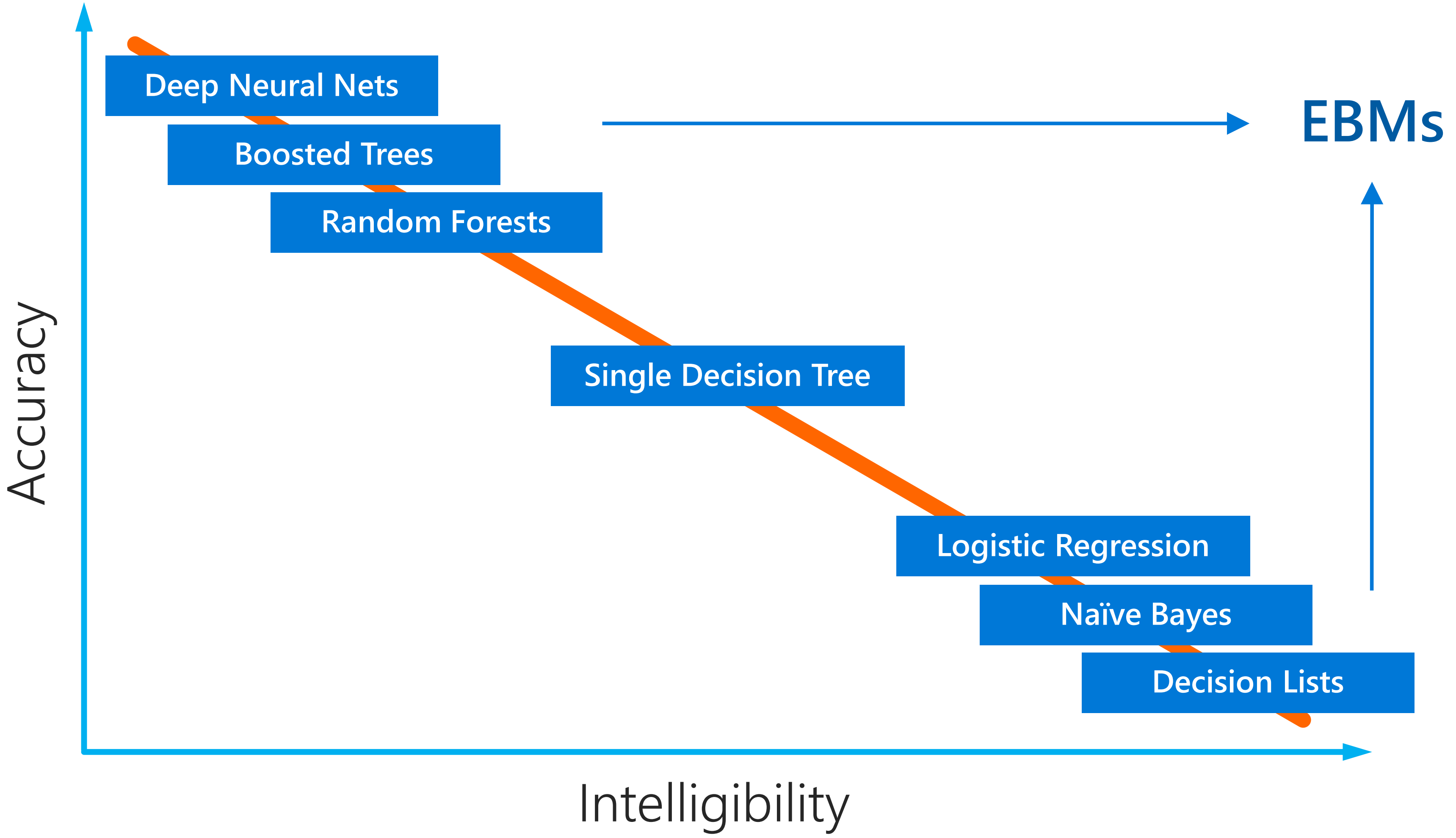
Accuracy vs. Intelligibility Tradeoff ???



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



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



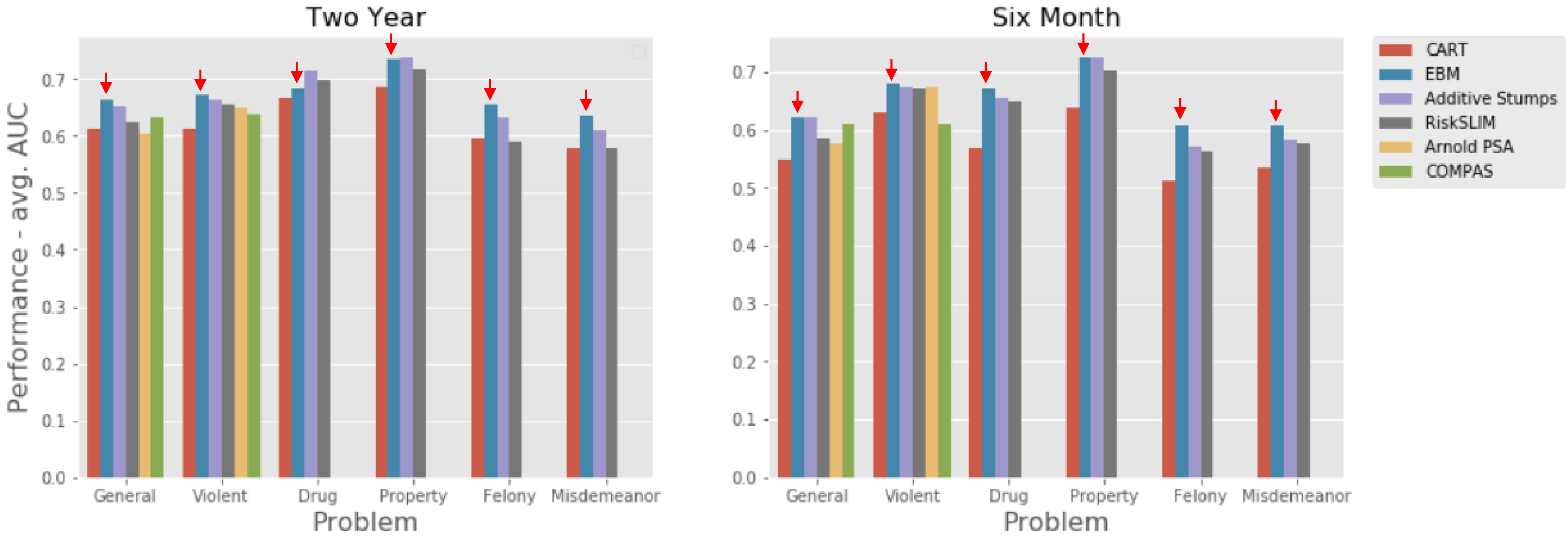
GlassBox

BlackBox

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 $\pm$ 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

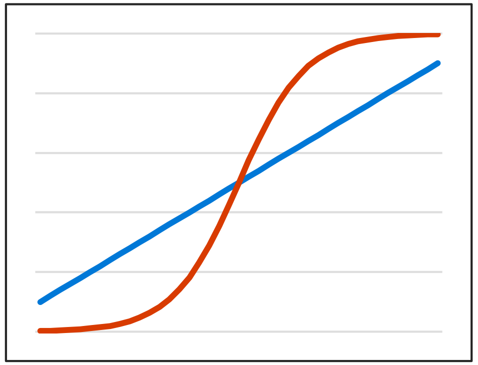
Table 2: RMSE on regression datasets for different learning methods. Each cell contains the mean RMSE $\pm$ one standard deviation obtained via 5-fold cross validation. Lower RMSE is better.

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.

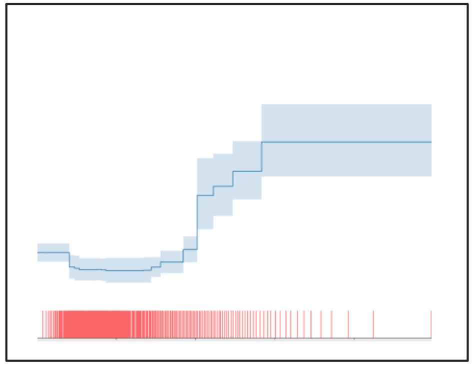
EBMs: Generalized Additive Models (GAMs)

Linear/Logistic Regression



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

GAMs/EBMs



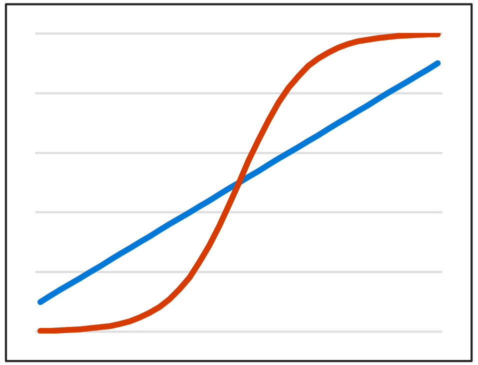
- More interpretable than linear/logistic
- Can be very accurate
- Can model nonlinearities
- Can model normal in middle
- More likely to show important effects
- Invented by Hastie & Tibshirani 1980's

BlackBox Machine Learning

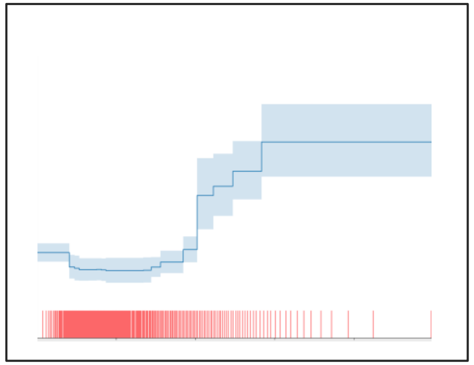
- Not interpretable (blackbox)
- Can be very accurate
- Can model nonlinearities
- Can model normal in middle
- More likely to learn spurious effects

EBMs: Generalized Additive Models (GAMs)

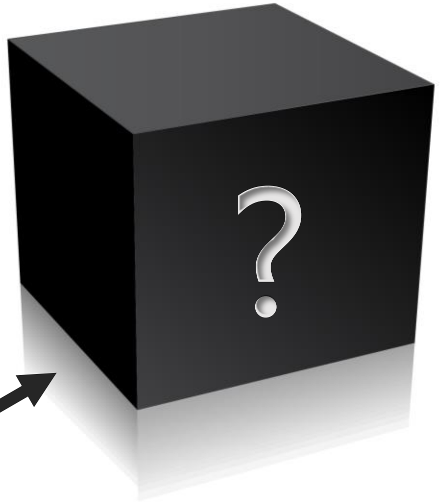
Linear/Logistic Regression



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

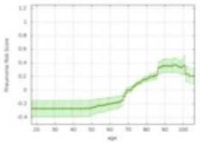
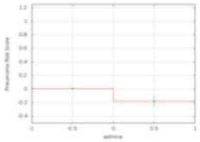
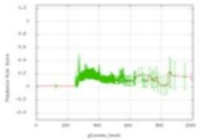
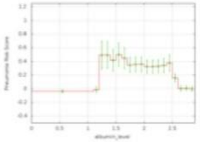
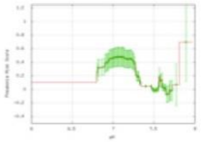
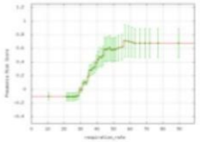
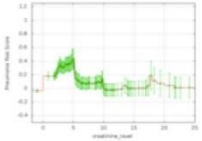
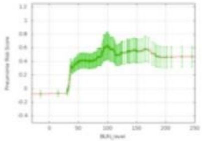
Pneumonia Dataset (collected 1989): 46 Features

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

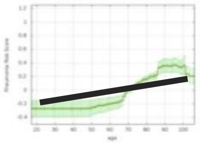
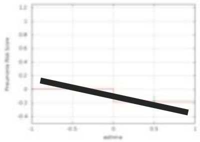
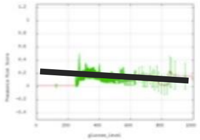
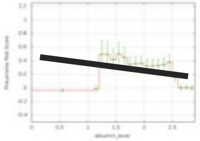
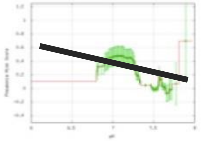
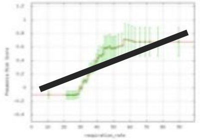
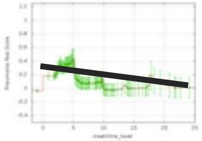
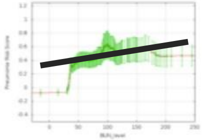
Laboratory findings
Sodium level (mEq/l)
Potassium level (mEq/l)
Creatinine level (mg/dl)
Glucose level (mg/dl)
BUN level (mg/dl)
Liver function tests (coded only as normal* or abnormal)
Albumin level (gm/dl)
Hematocrit
White blood cell count (1000 cells/μl)
Percentage bands
Blood pH
Blood pO_2 (mmHg)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

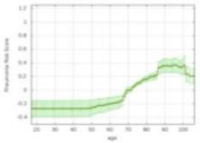
$$Score = baseline + \sum_{i=0}^{n} f_i(variable_i)$$

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

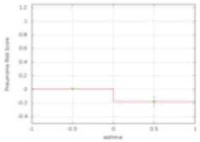
	Age	=>	-0.23
	Asthma	=>	-0.15
	Glucose	=>	+0.18
	Albumin	=>	+0.01
	Blood pH	=>	+0.38
	Respiration	=>	+0.21
	Creatinine	=>	-0.01
	BUN	=>	-0.21

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

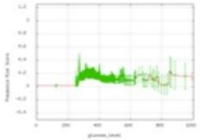
$$\text{POD} = \frac{1}{1+e^{-\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n}}$$



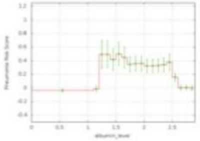
Age => -0.23



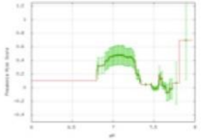
Asthma => -0.15



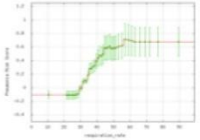
Glucose => +0.18



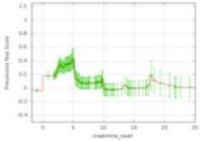
Albumin => +0.01



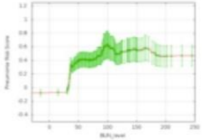
Blood pH => +0.38



Respiration => +0.21



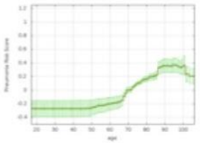
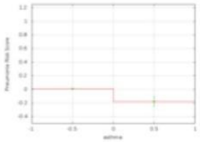
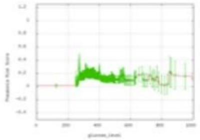
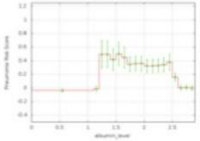
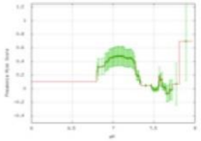
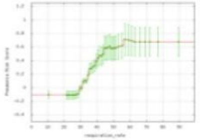
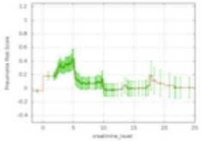
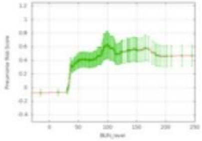
Creatinine => -0.01



BUN => -0.21

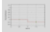
$$Score = baseline + \sum_{i=0}^{n} f_i(variable_i)$$

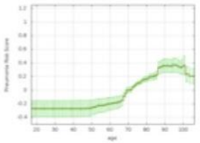
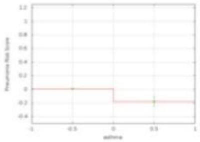
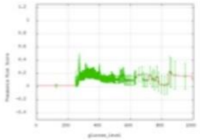
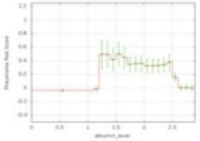
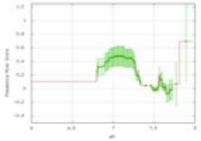
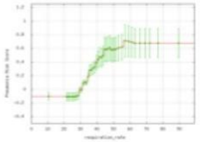
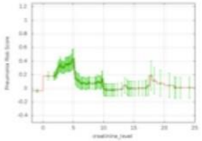
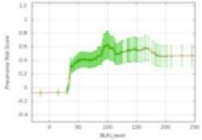
$$\text{POD} = \frac{1}{1+e^{-Score}}$$

	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

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

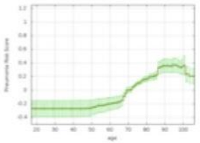
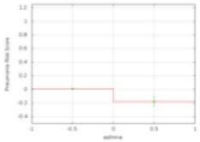
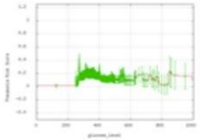
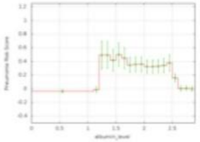
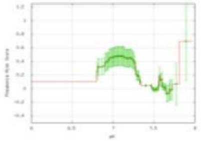
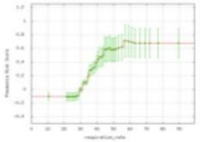
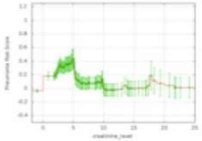
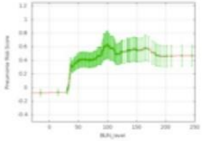
$$POD = \frac{1}{1+e^{-}}$$

 +  +  + … + 

	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

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

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

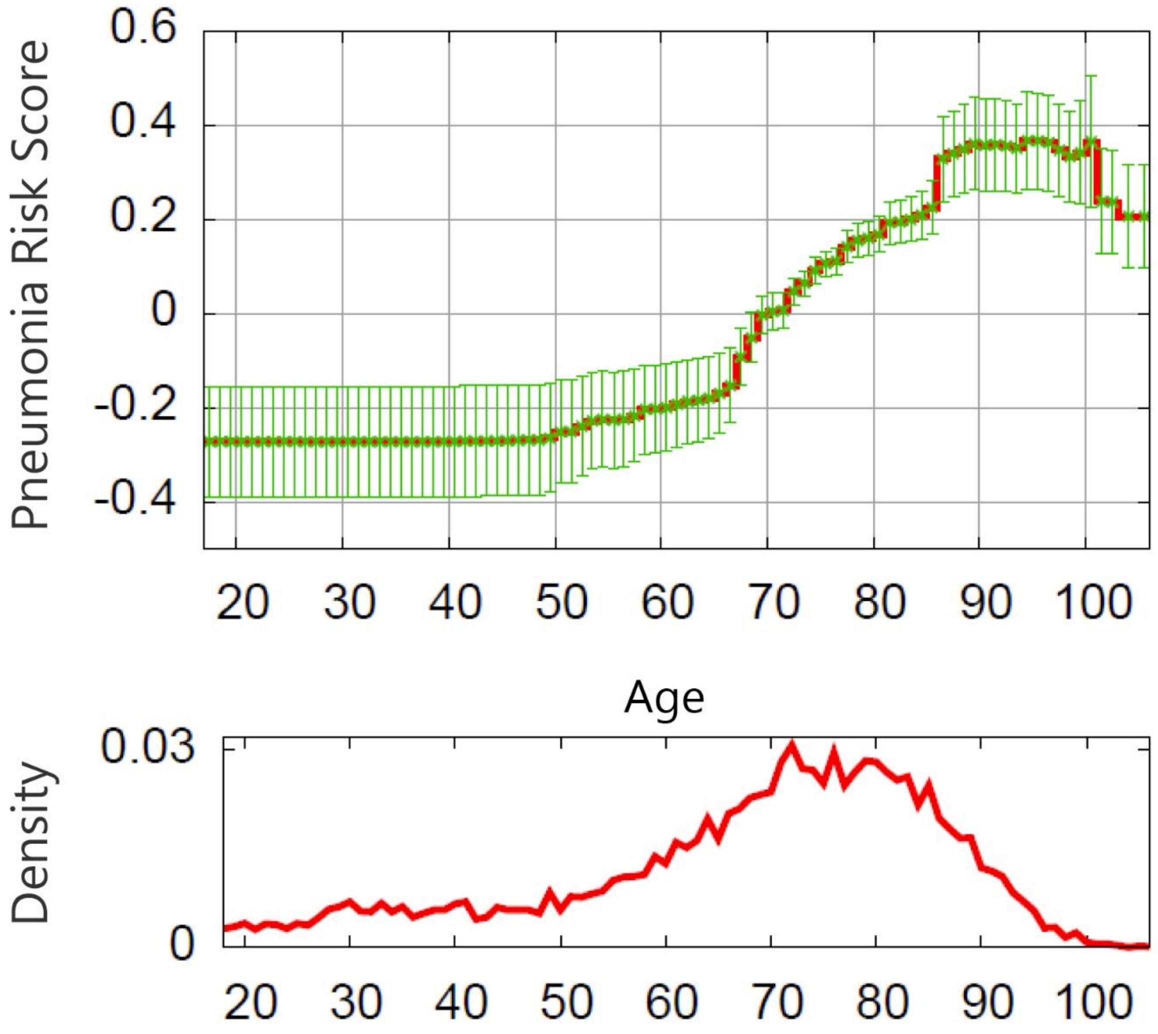
$$Score = -2.11 - 0.23 - 0.15 + 0.18 + 0.01 + 0.38 + \ \ldots$$

$$Score = -0.78$$

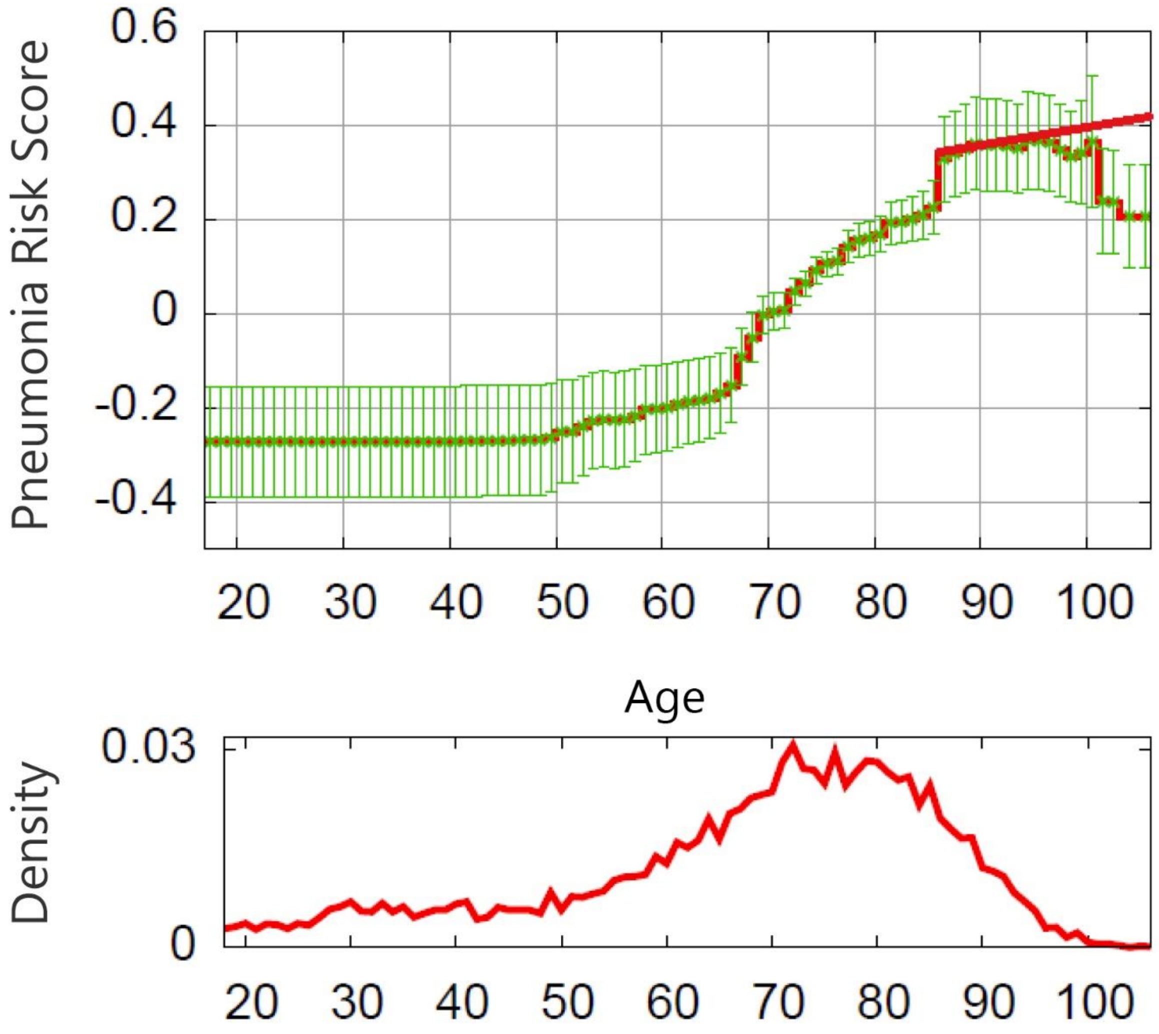
$$\text{POD} = \frac{1}{1+e^{-(-0.78)}}$$

$$\text{POD} = 0.3143$$

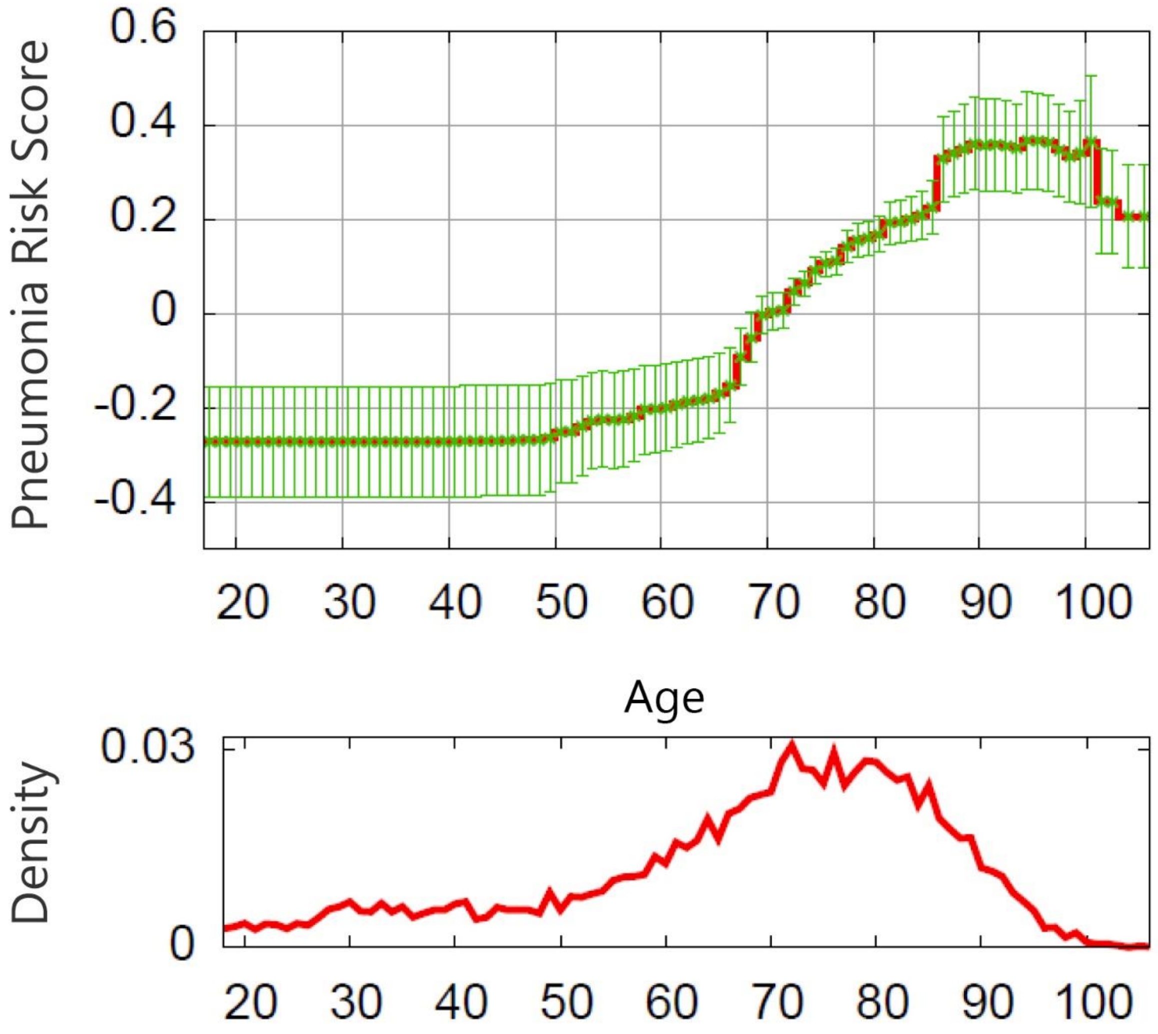
What EBMs Learn about Pneumonia Risk vs. Age



Fix Age > 100 Problem (Enforce Monotonicity)

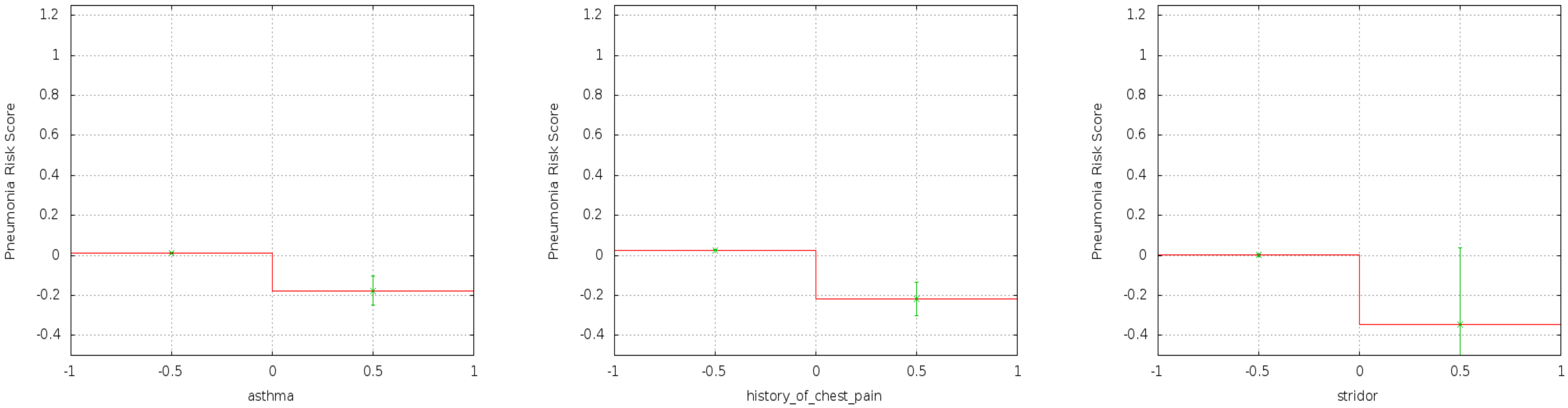


Original Model is Correct for Actuarial Use



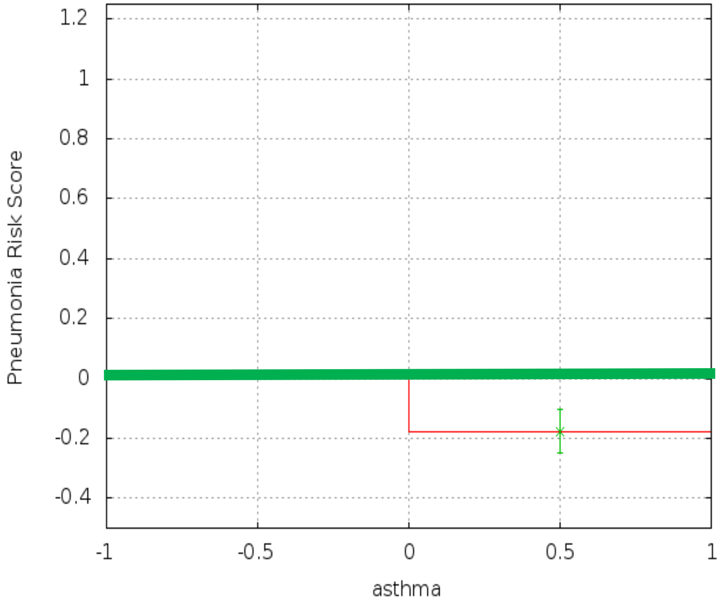
- Model correctness (and accuracy) depends on how model will be used
  - This is a good model for health insurance provides
  - But needs to be repaired before using for patient treatment decision

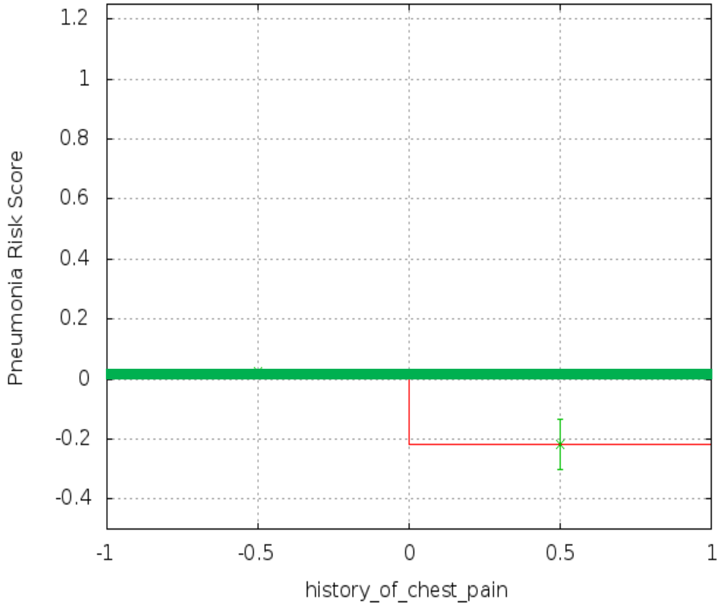
- A few things intelligible model learned:
  - Beware of jumps at round numbers --- almost always due to human/social/policy effects
  - Asthma => lower risk
  - History of chest pain => lower risk
  - History of heart disease => lower risk
  - Obstructed airway => lower risk

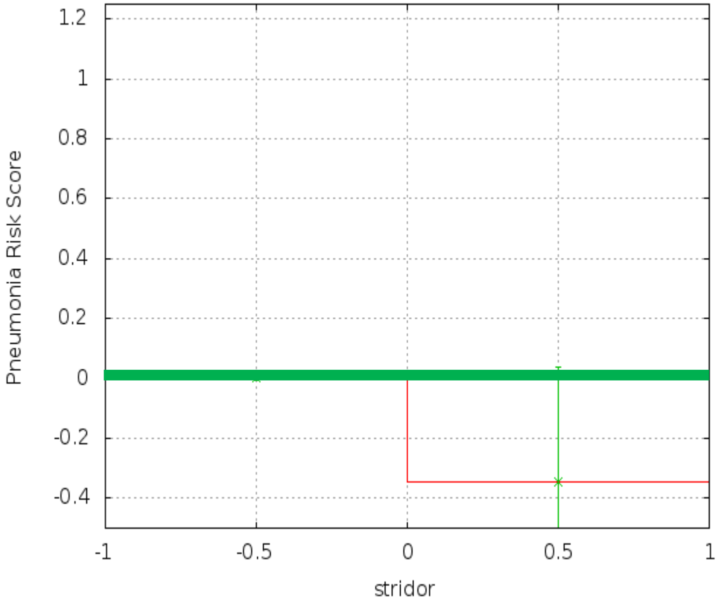


- Model correctness (and accuracy) depends on how model will be used
  - This is a good model for health insurance provides
  - But needs to be repaired before using for patient treatment decision

- 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

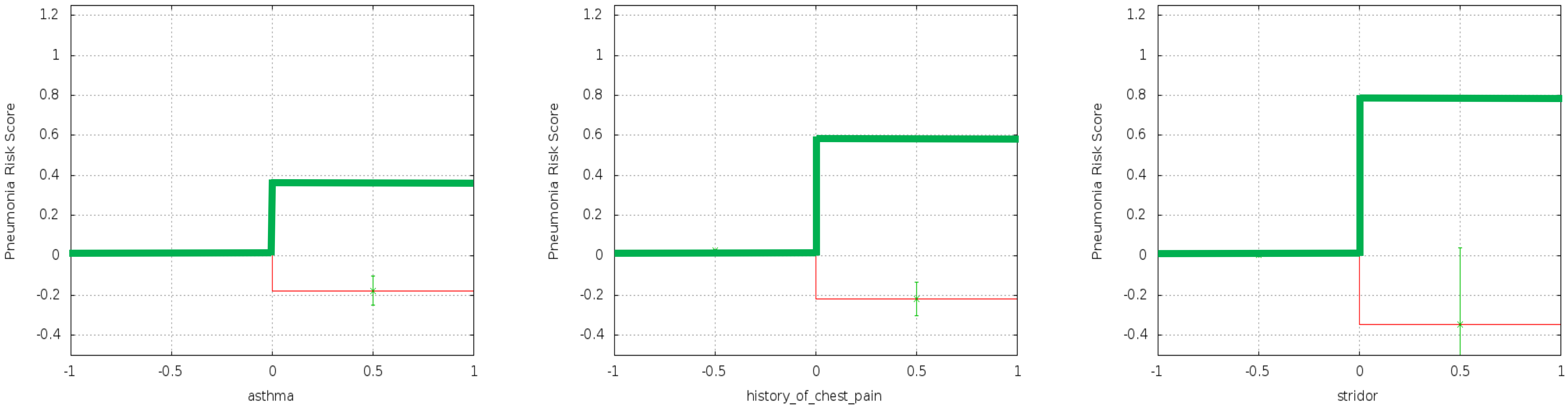






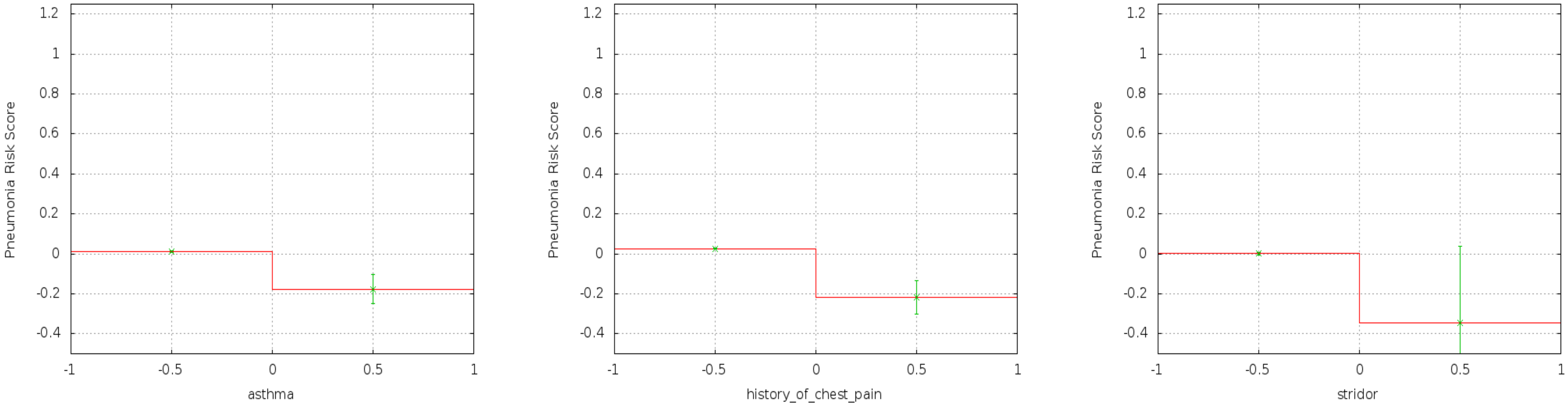
- Model correctness (and accuracy) depends on how model will be used
  - This is a good model for health insurance provides
  - But needs to be repaired before using for patient treatment decision

- 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

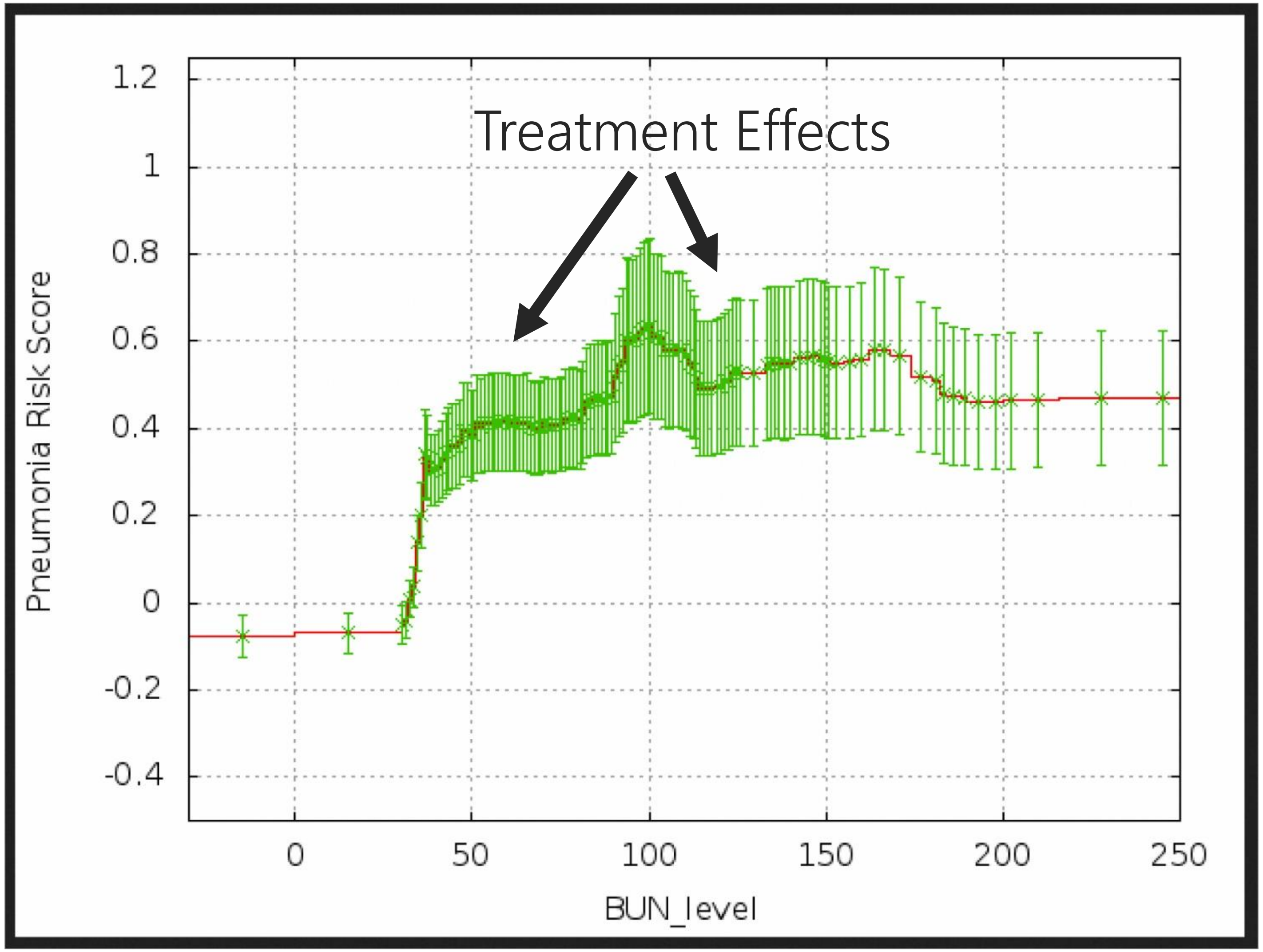


- Model correctness (and accuracy) depends on how model will be used
  - This is a good model for health insurance provides
  - But needs to be repaired before using for patient treatment decision

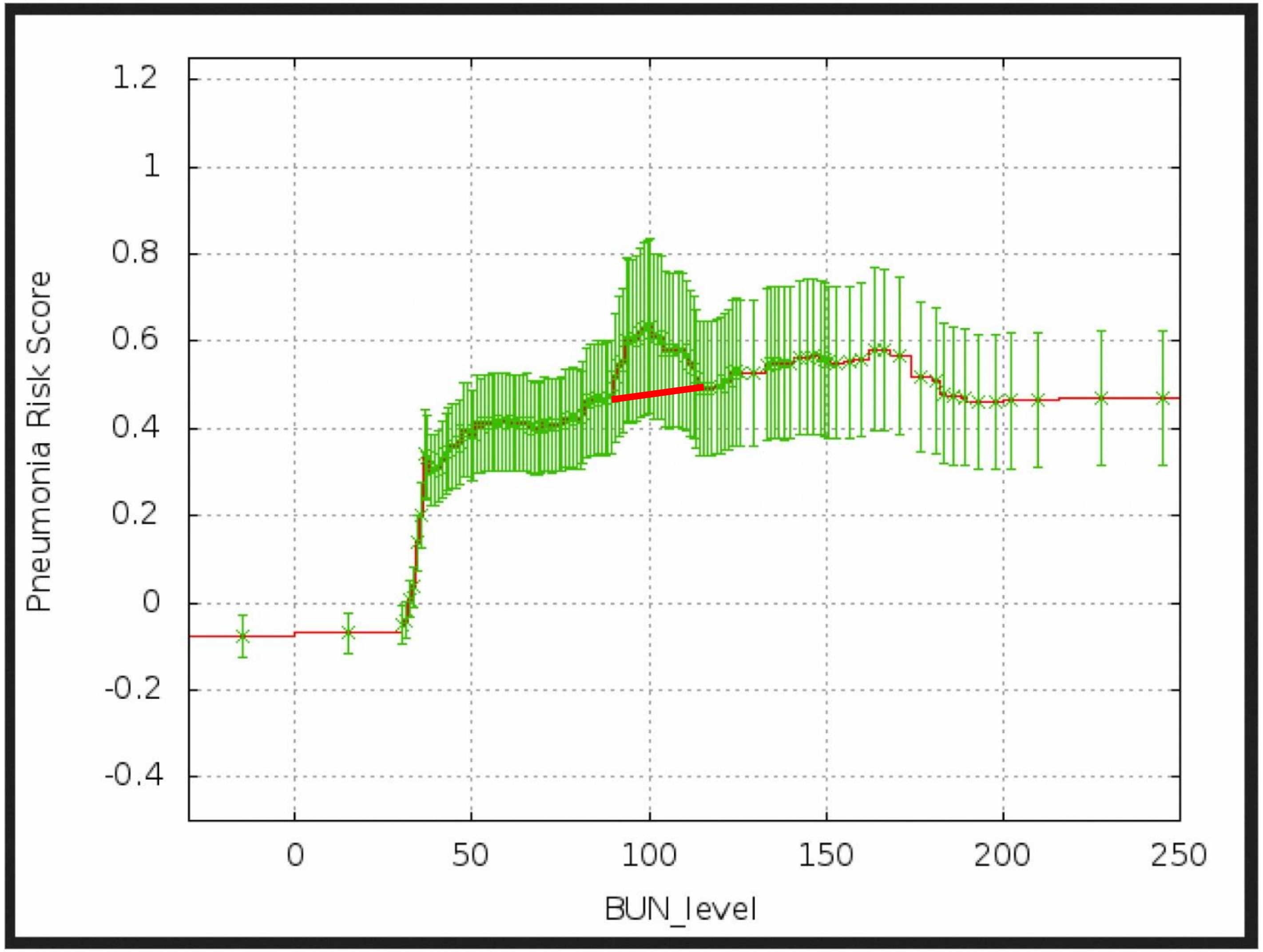
- A few things intelligible model learned:
  - Beware of jumps at round numbers --- almost always due to human/social/policy effects
  - Asthma => lower risk
  - History of chest pain => lower risk
  - History of heart disease => lower risk
  - Obstructed airway => lower risk
  - …
  - Model is rewarded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

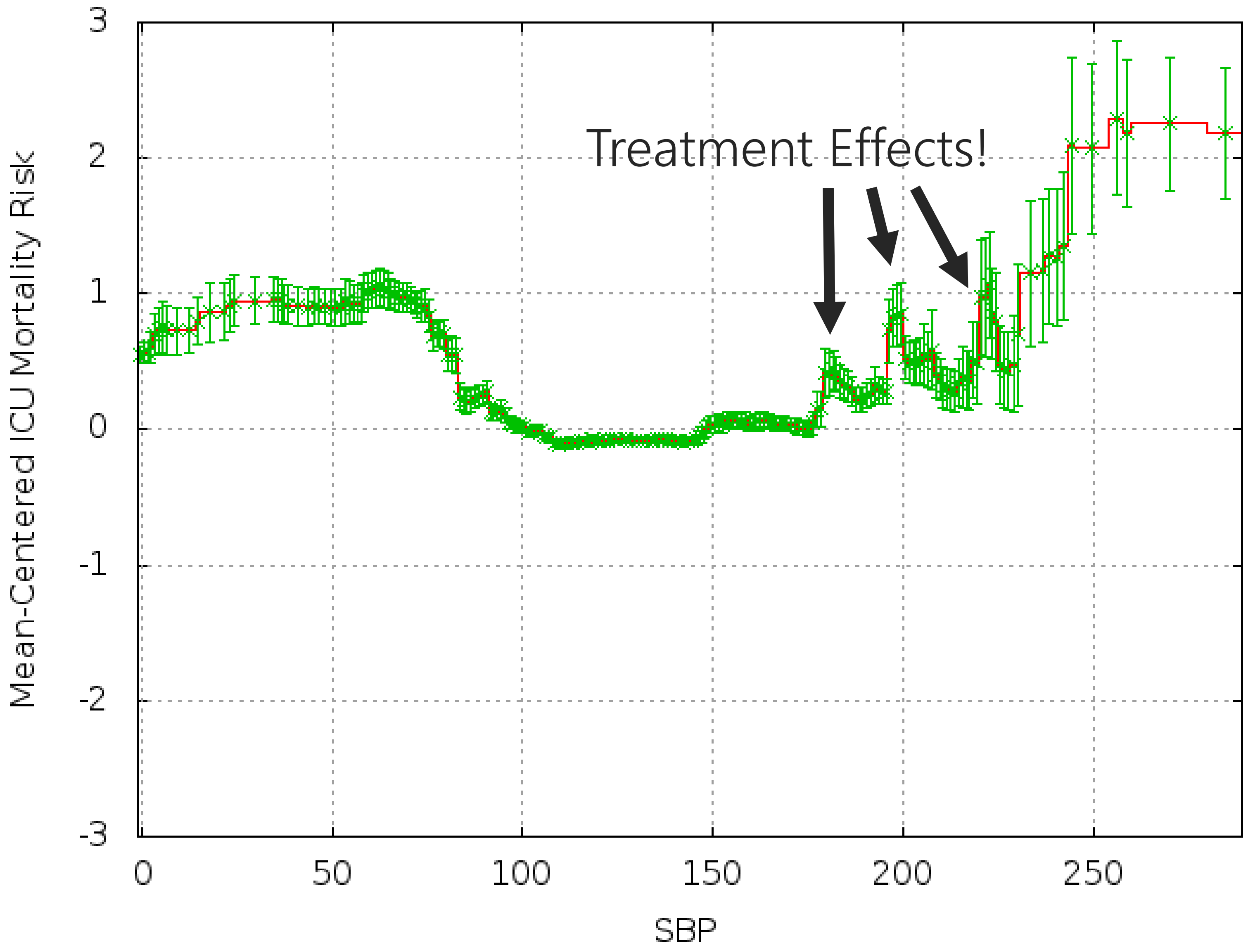


Intelligibility Can Create New Medical Science



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

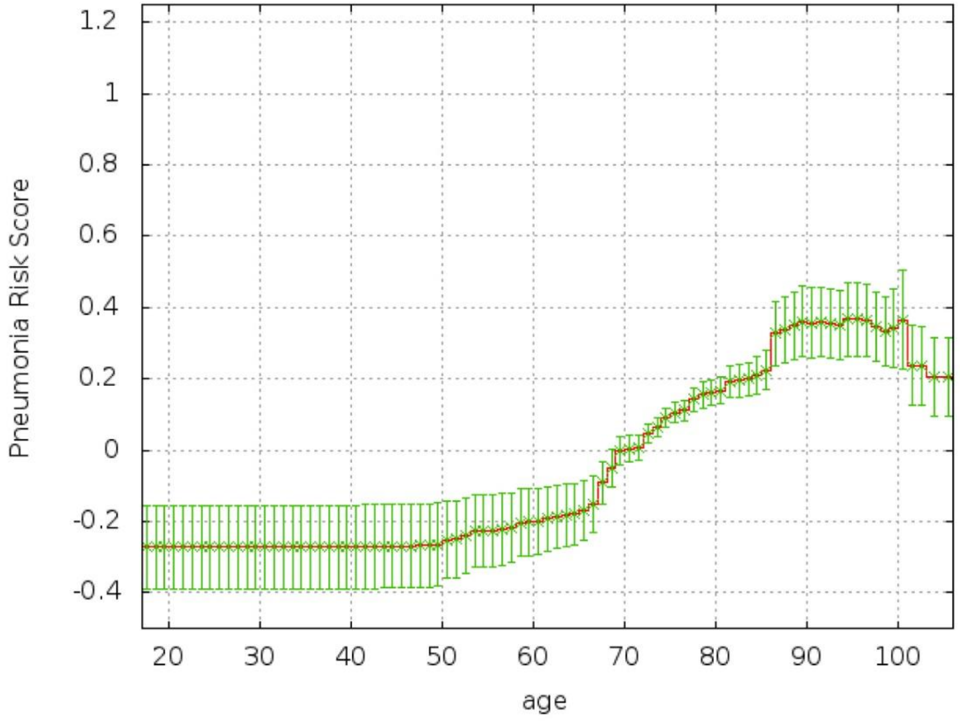
Treatment Effects Ubiquitous in All Medical Data

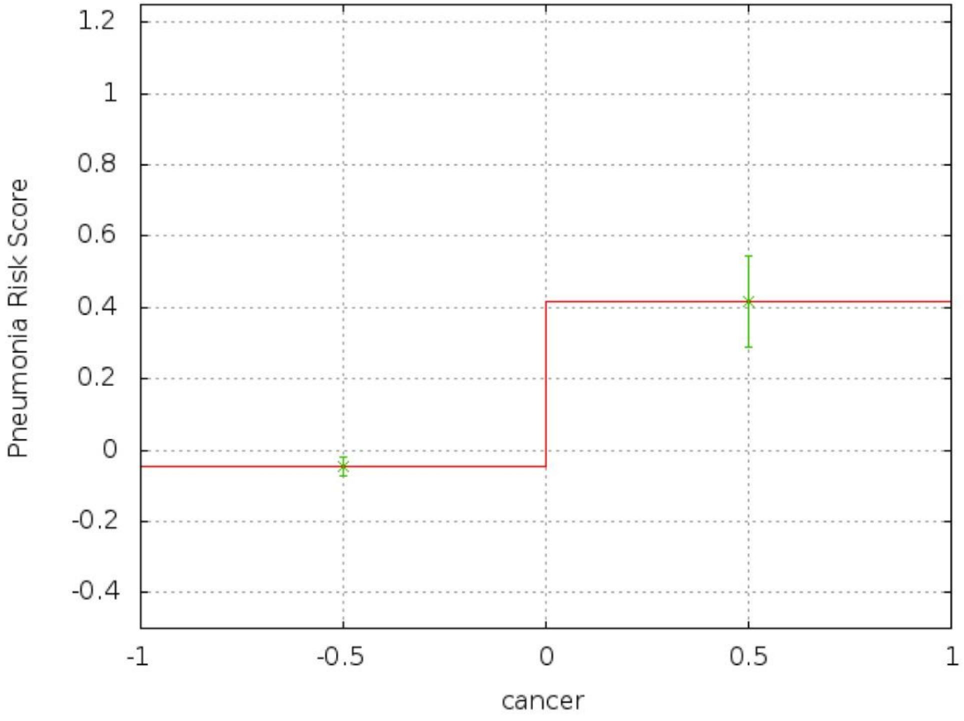


Pairwise Interactions?

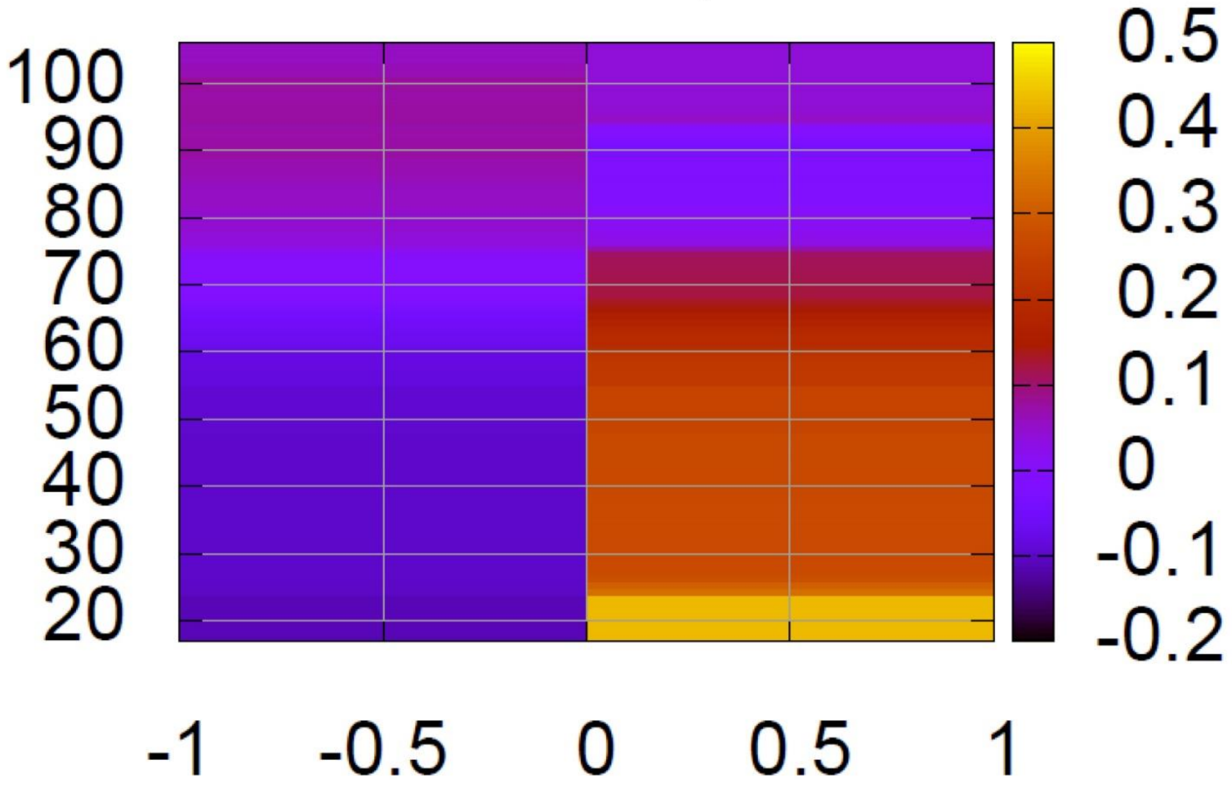
Like XOR (parity), interactions can't be modeled as a sum of independent effects:

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





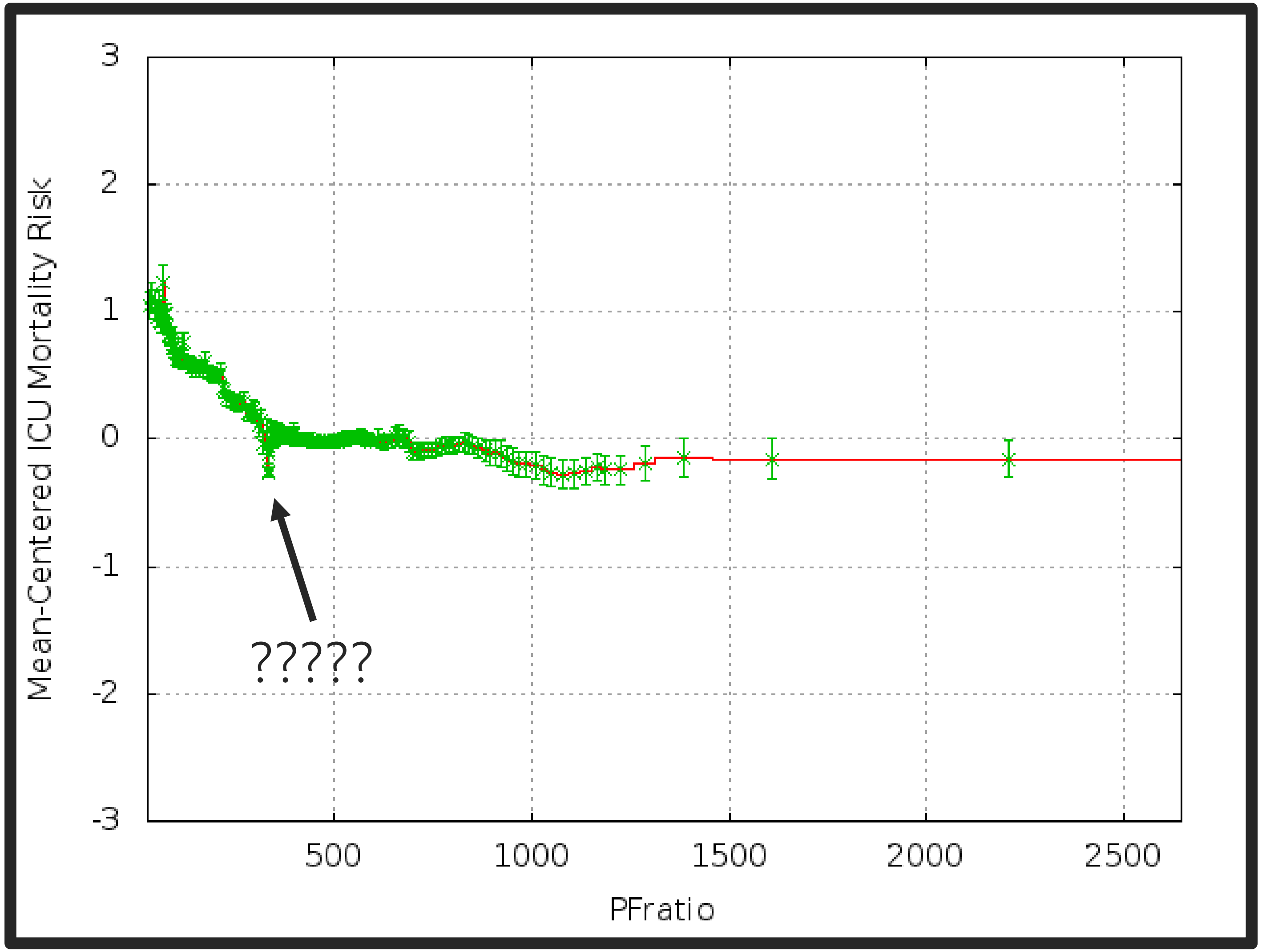
Pairwise Interaction: Age x Cancer (Pneumonia-95)



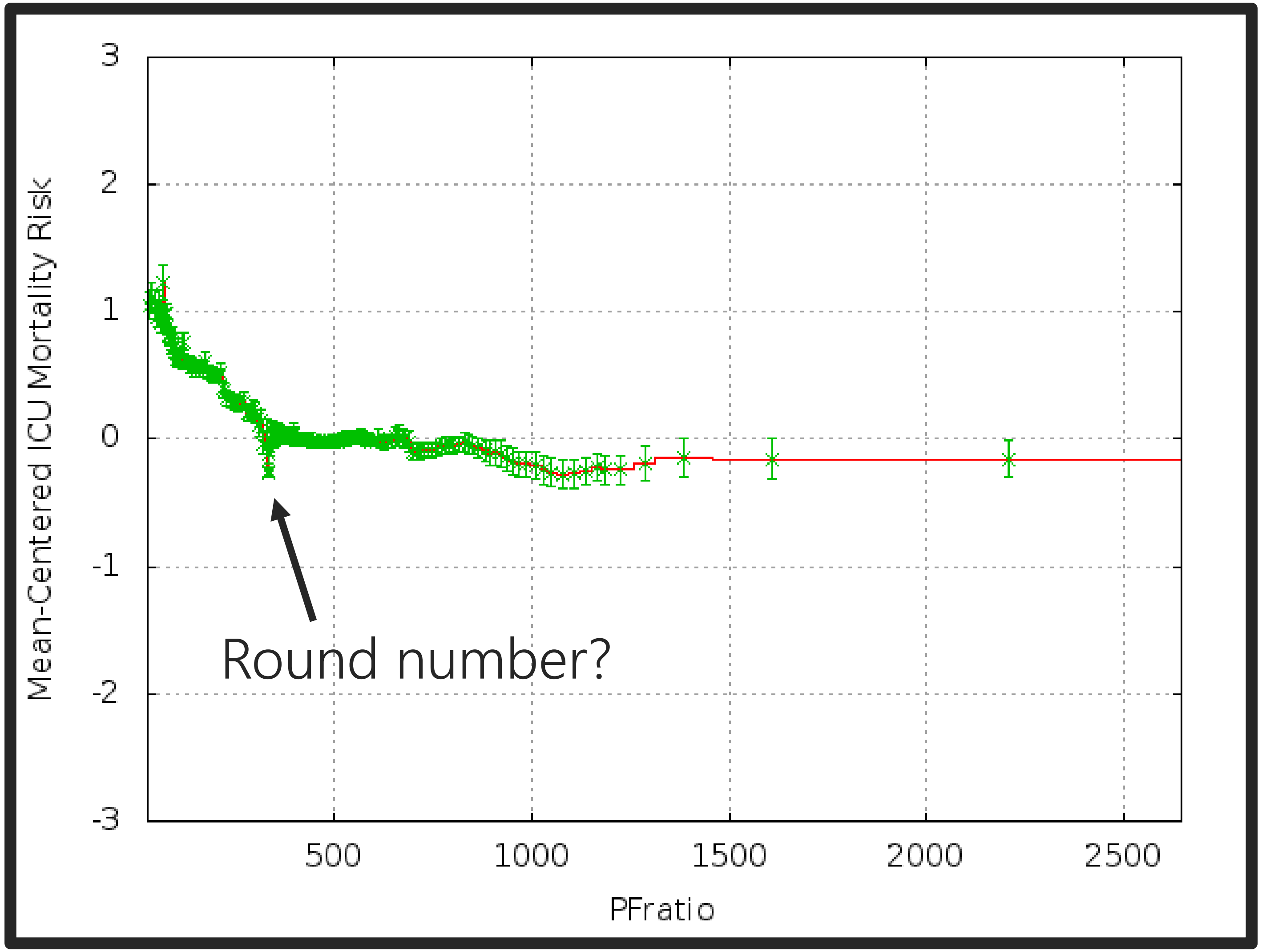
Age vs. Cancer

Example 2: ICU Mortality

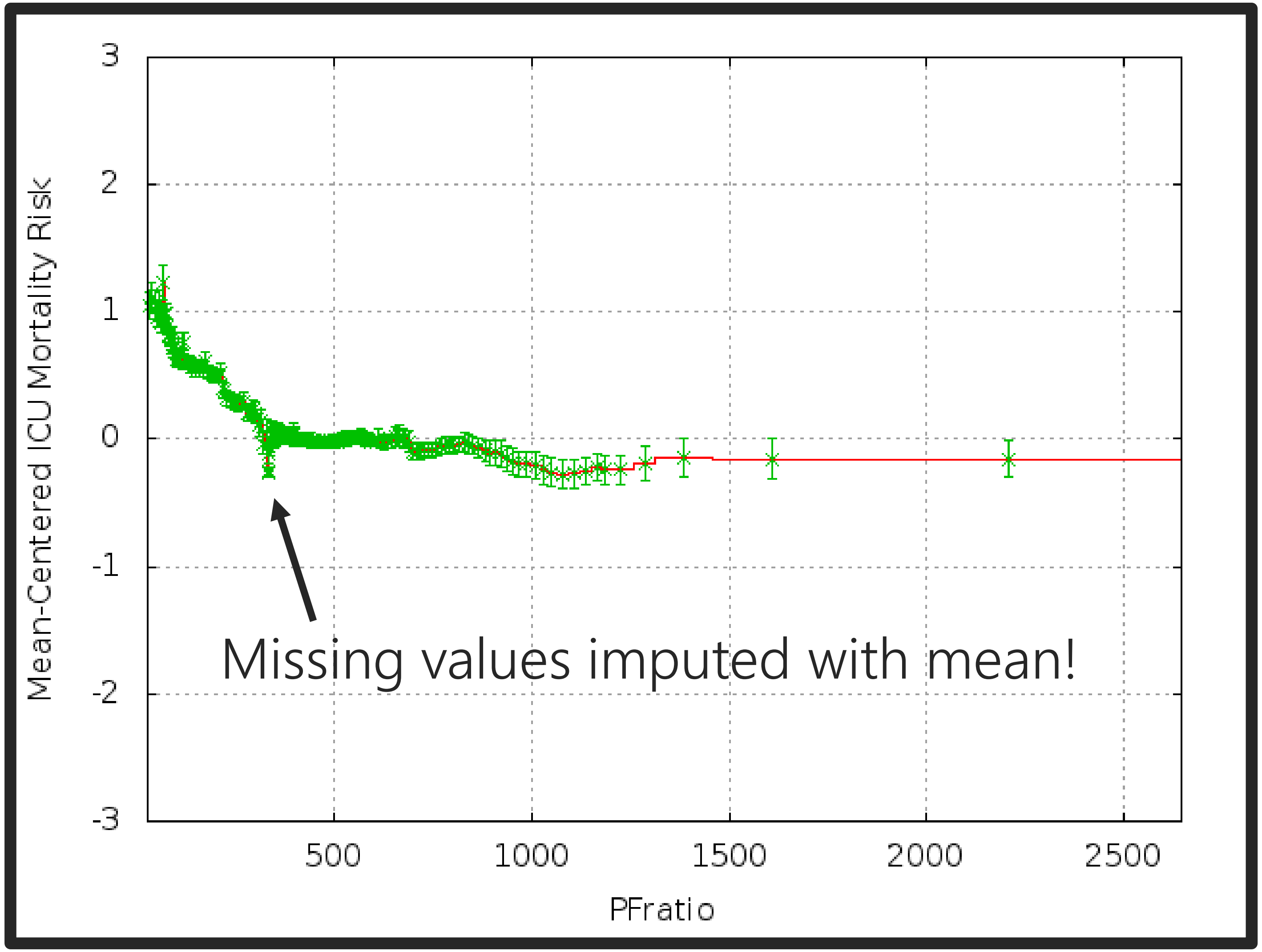
Intelligibility Helps Debug Data: PaO2/FiO2 Ratio



Intelligibility Helps Debug Data: PaO2/FiO2 Ratio



Intelligibility Helps Debug Data: PaO2/FiO2 Ratio



Intelligibility Has Completely Changed How We Think About and Handle Missing Values

Example 3: 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 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

Impact on Real Policy

Wikimedia Policy
@wikimediapolicy

Replying to @wikimediapolicy

Re: algo. transparency, MSFT's Rich Caruana gives an example of glass box methods: "You see a decrease in malicious editing [of Wikipedia] at 30 days because that is when [it] automatically logs you out. If you remember your password, you're less likely to do malicious editing."

7:18 PM · Jul 30, 2020 · Twitter Web App

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

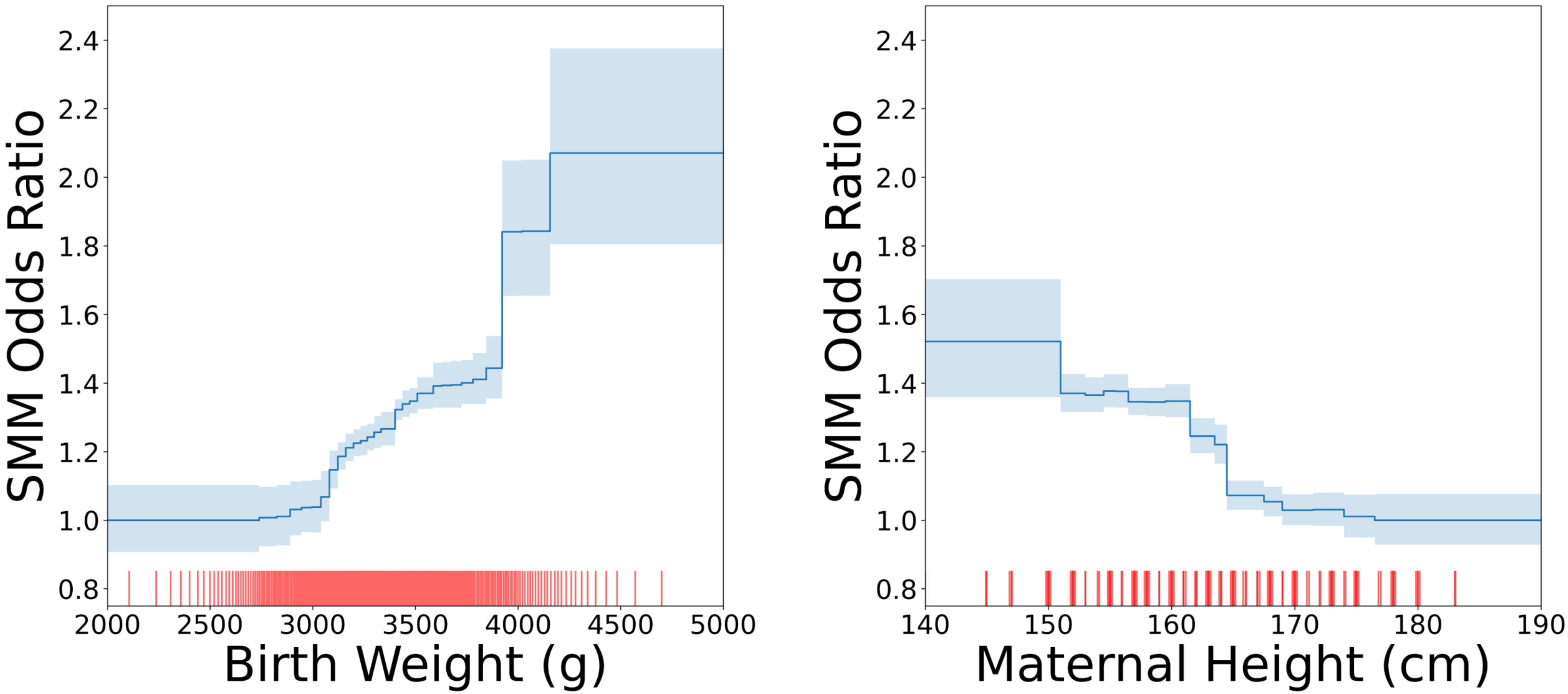
Example 5: Severe Maternal Morbidity

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)

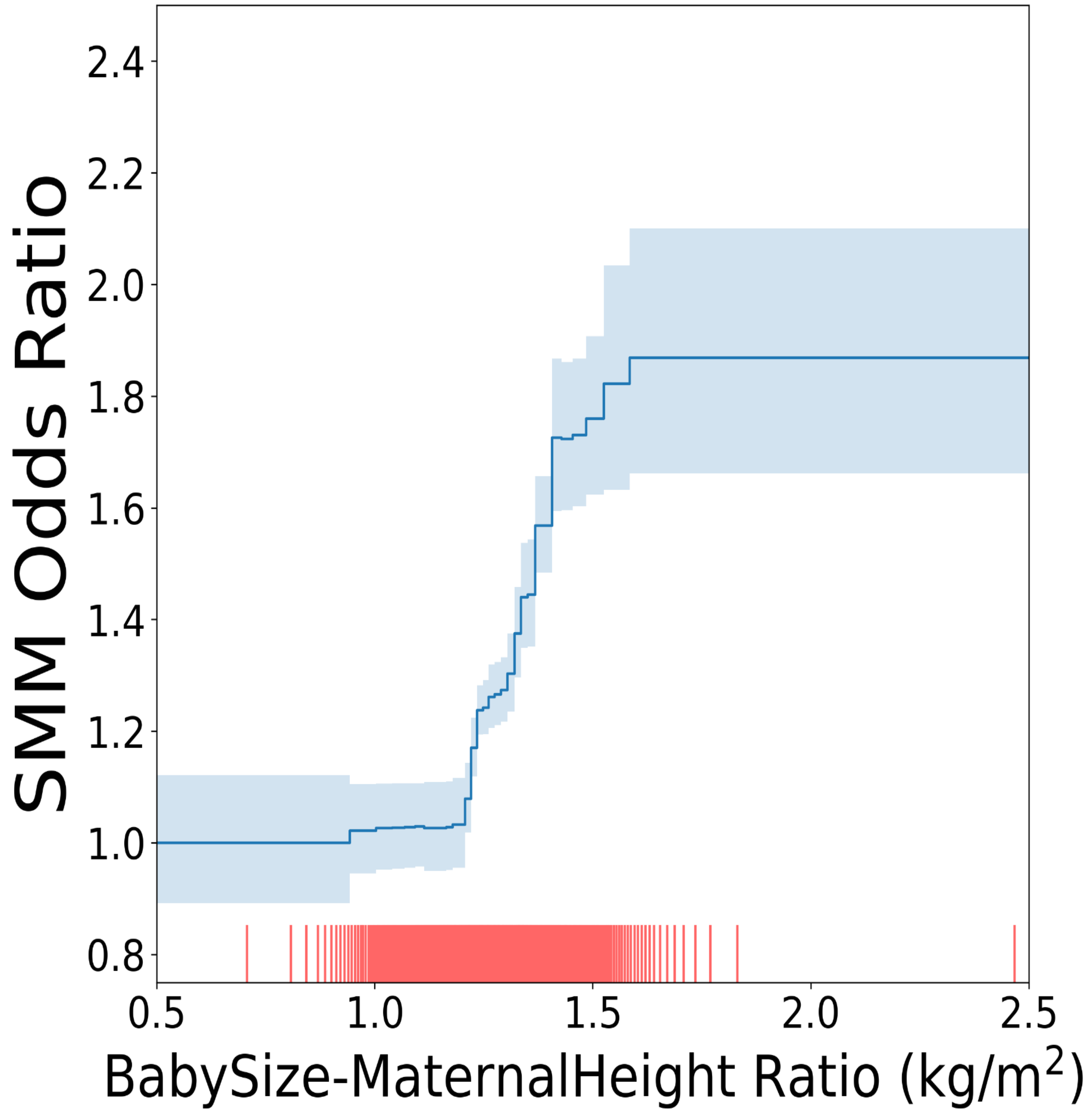
Intelligible ML Says Most Important Factors Are...



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

"BMI" for Pregnancy

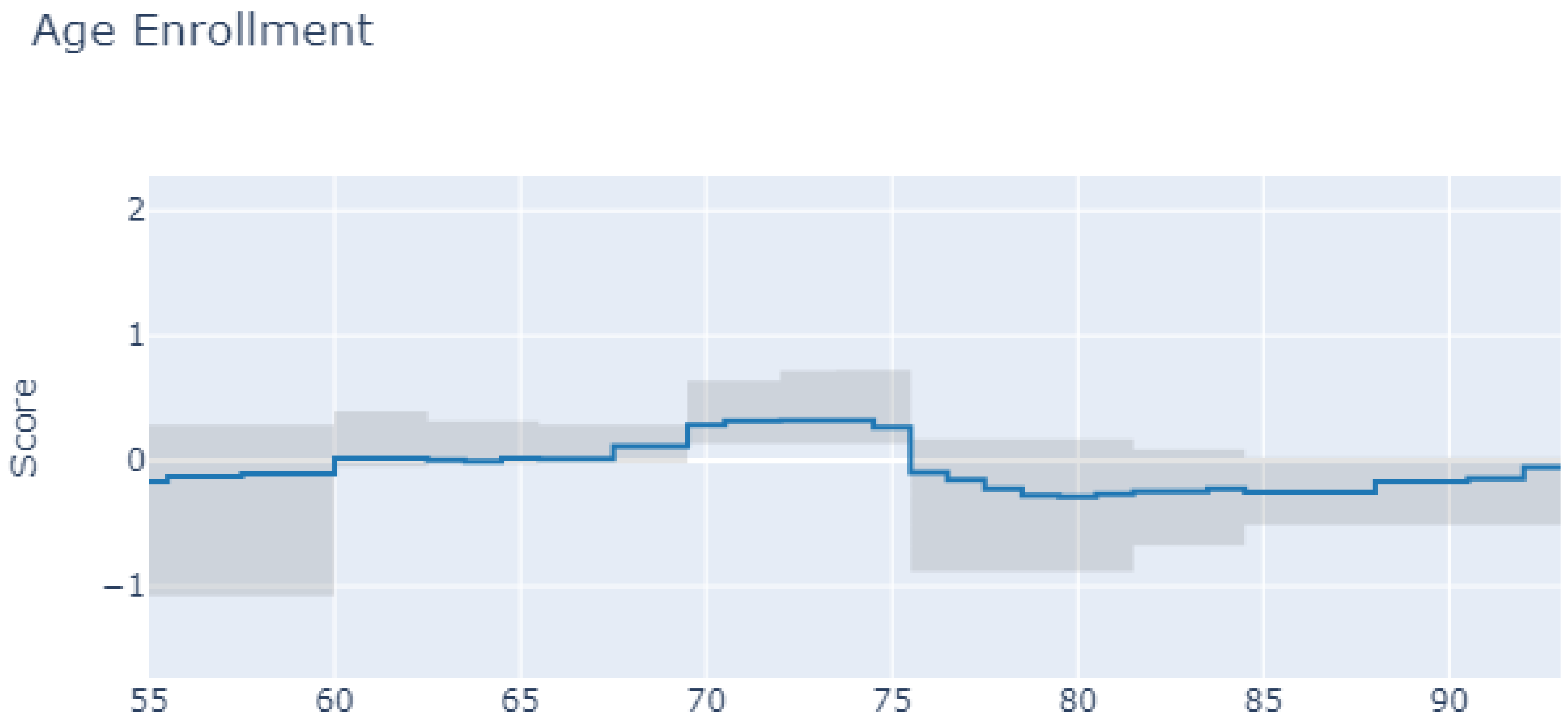
$$\frac{BabyBirthWeight}{MaternalHeight^2}$$



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

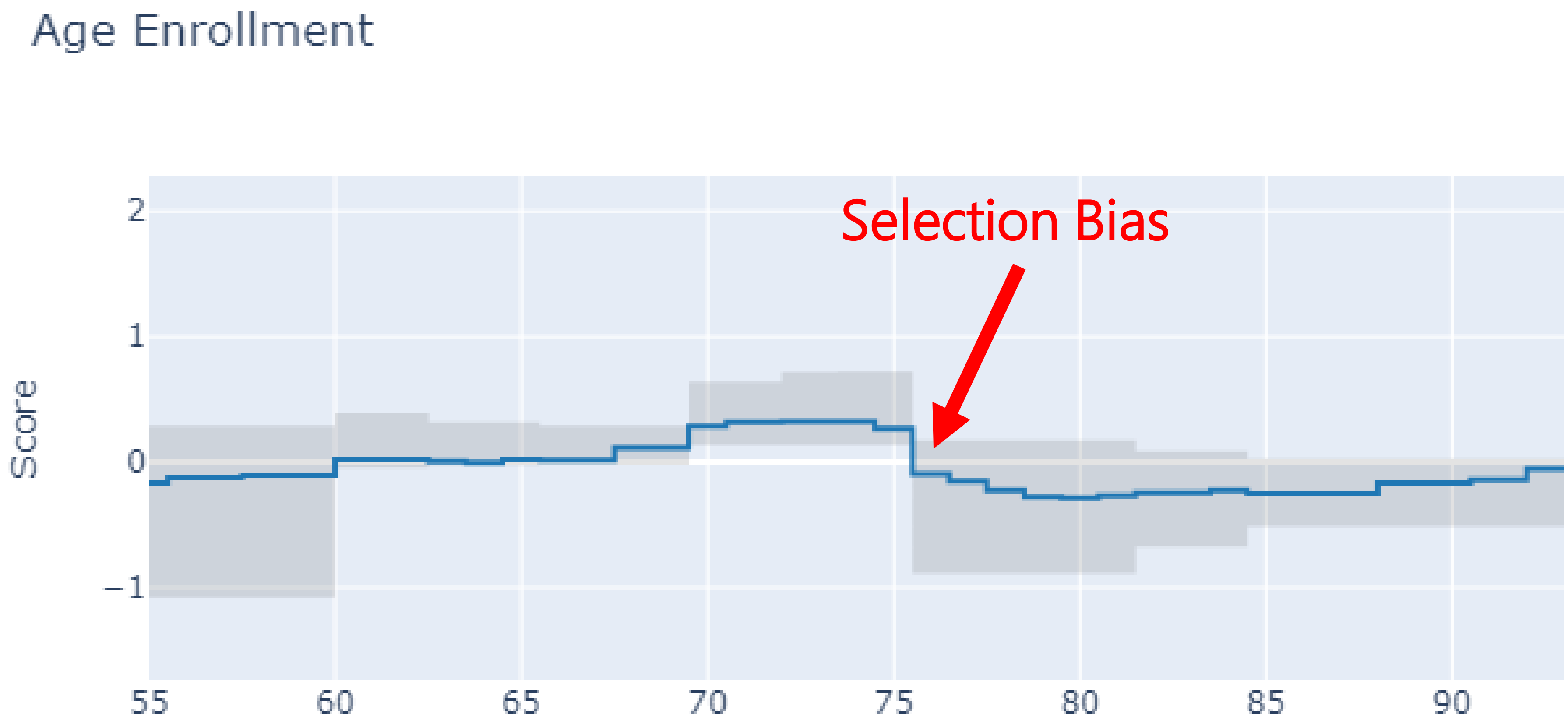
Example 6: Cancer Treatment

Cancer Mortality Risk vs. Age



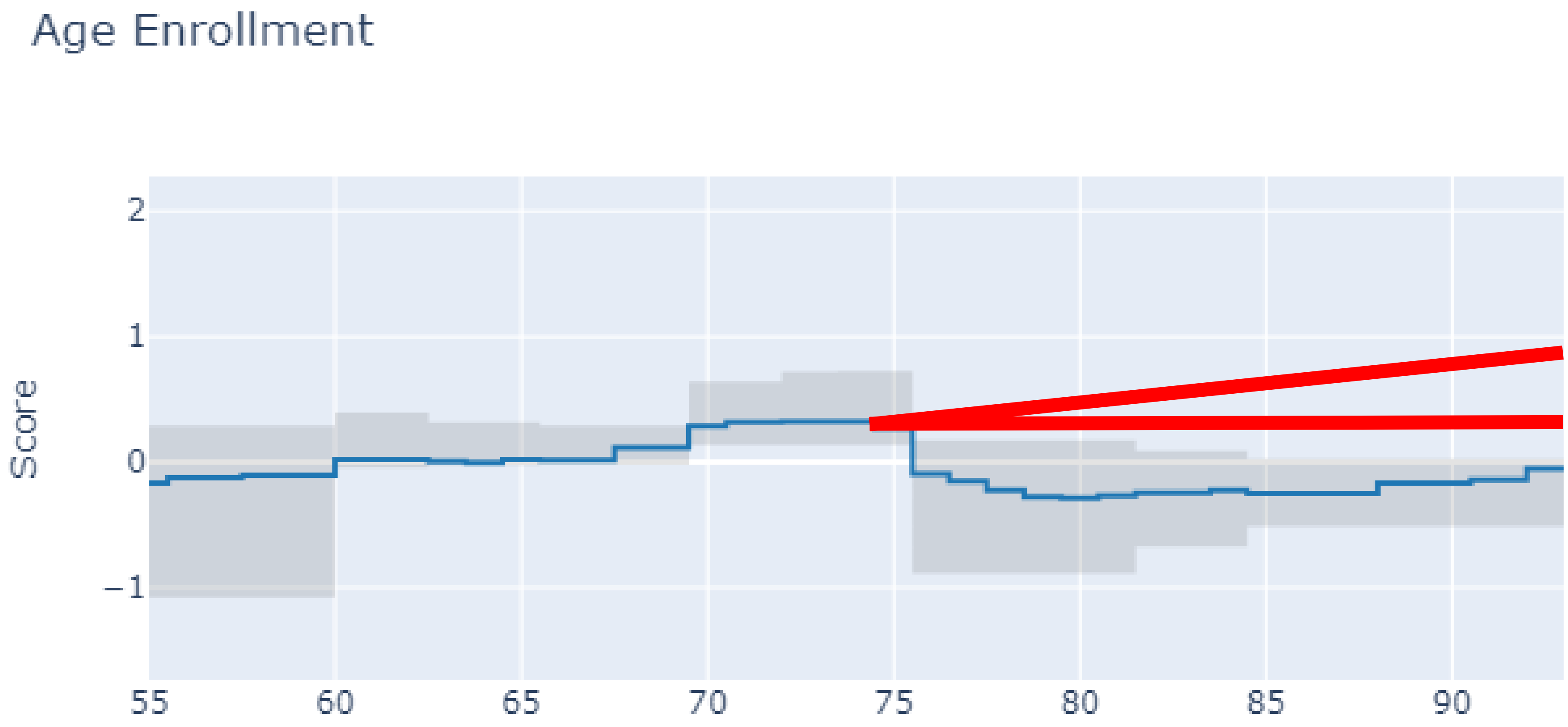
Zheng Zhang, Ying Xiao M.D., Sang Ho Lee (University of Pennsylvania), Rich Caruana (Microsoft)

Cancer Mortality Risk vs. Age



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

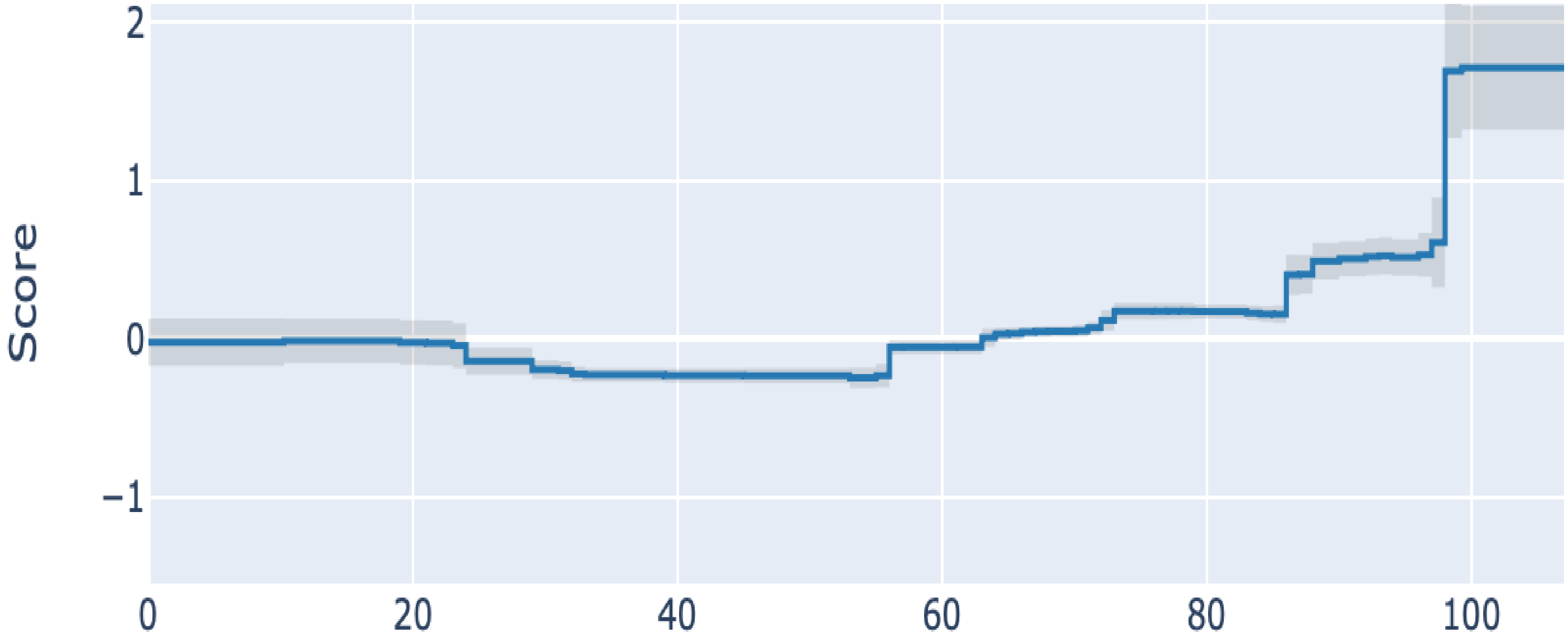
Cancer Mortality Risk vs. Age



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

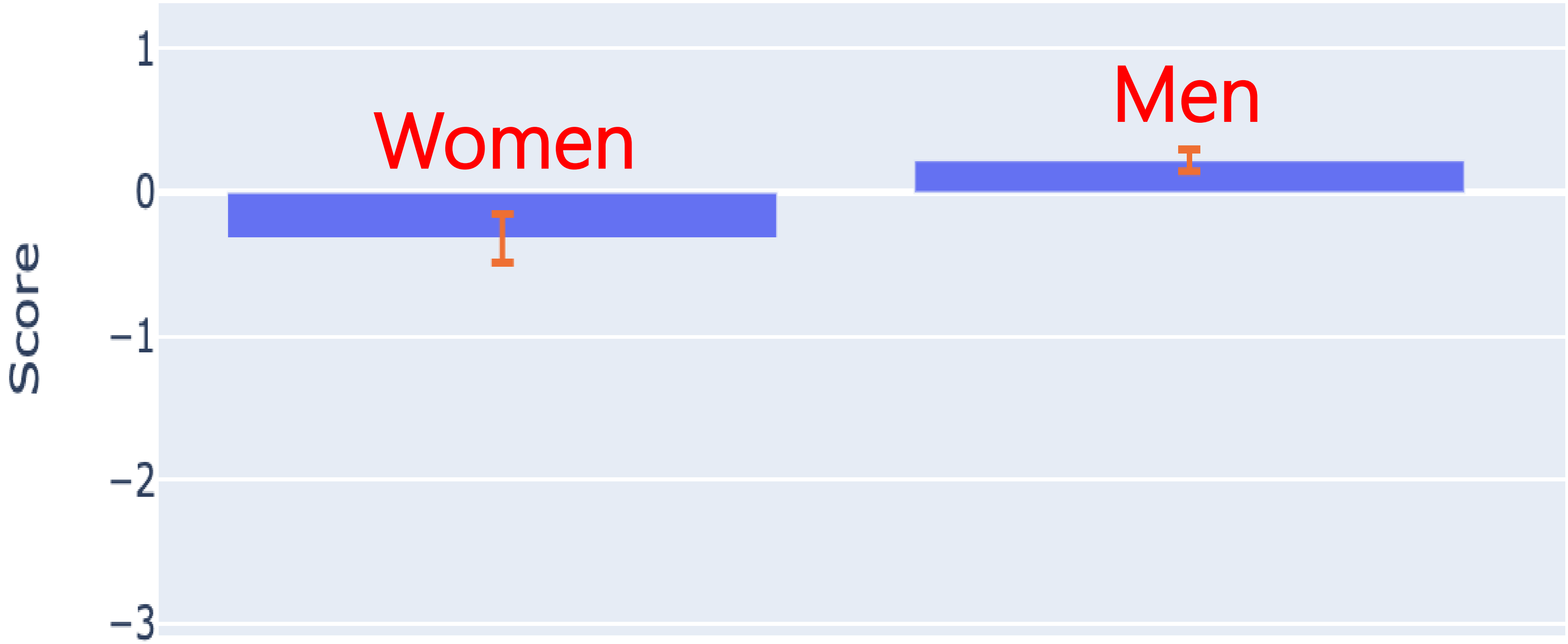
Example 7: COVID-19 Mortality

COVID-19 Mortality Risk vs. Age



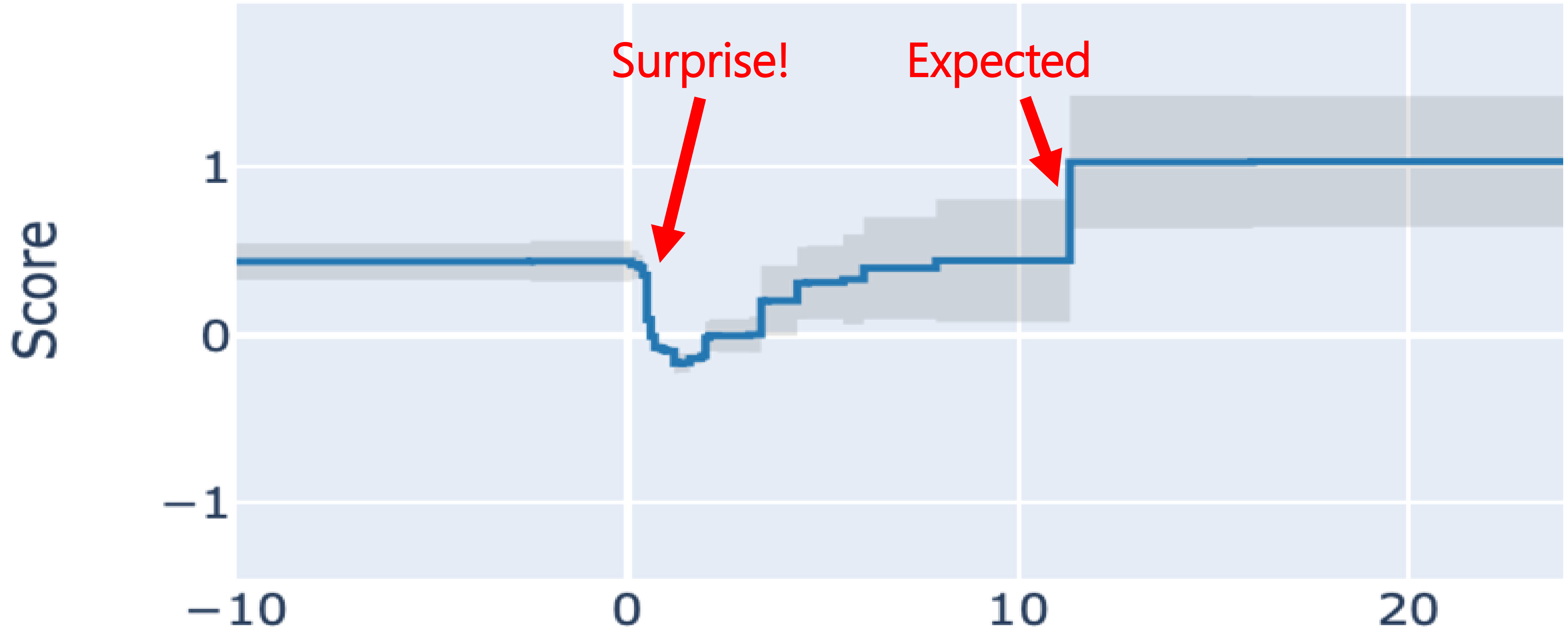
Ben Lengerich (CMU/MIT), Rich Caruana (Microsoft), Aphinyanaphongs Yindalon (NYU)

COVID-19 Mortality Risk vs Gender



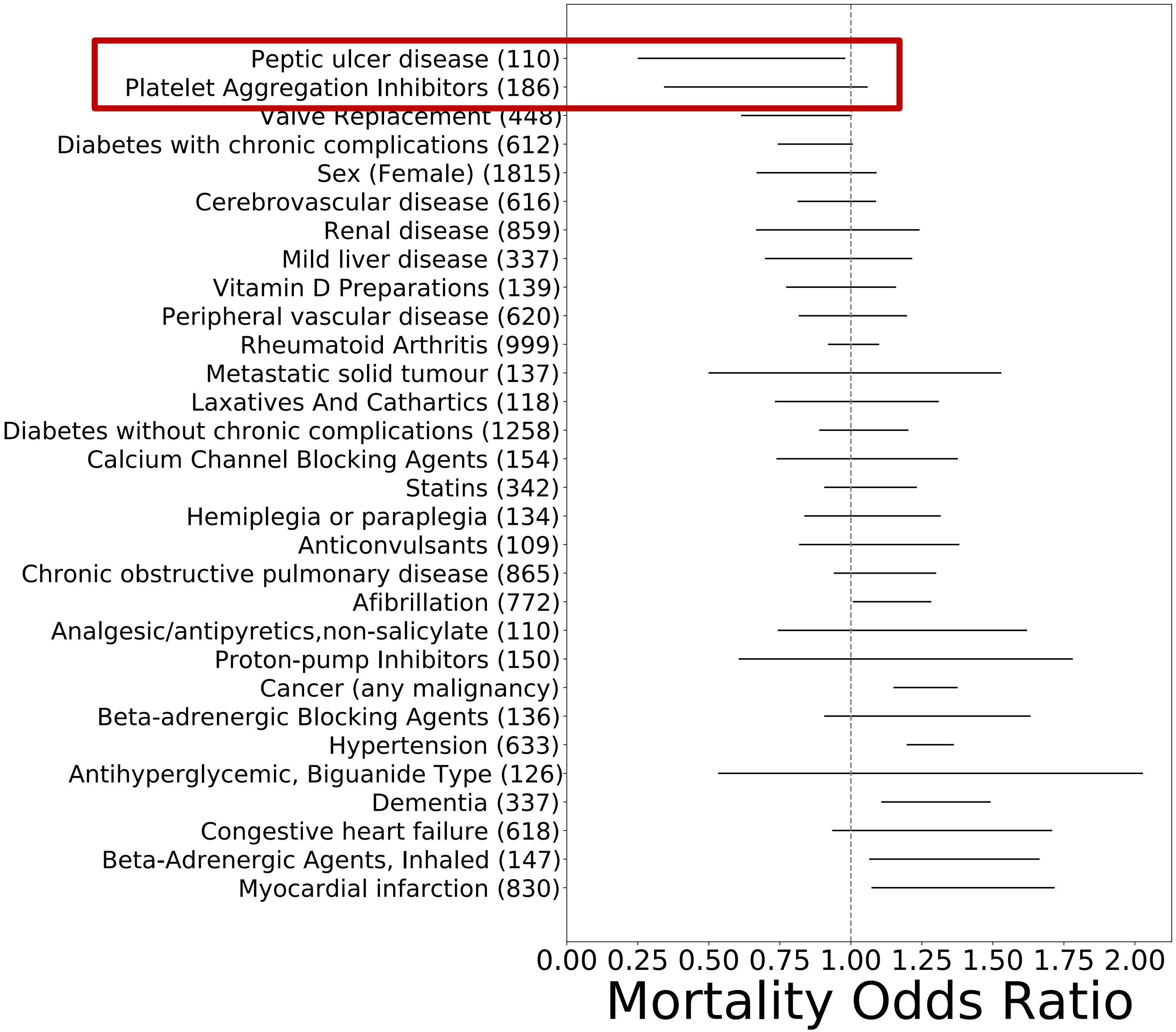
Ben Lengerich (CMU/MIT), Rich Caruana (Microsoft), Aphinyanaphongs Yindalon (NYU)

Surprising Discovery: Lymphocytes_Absolute_a

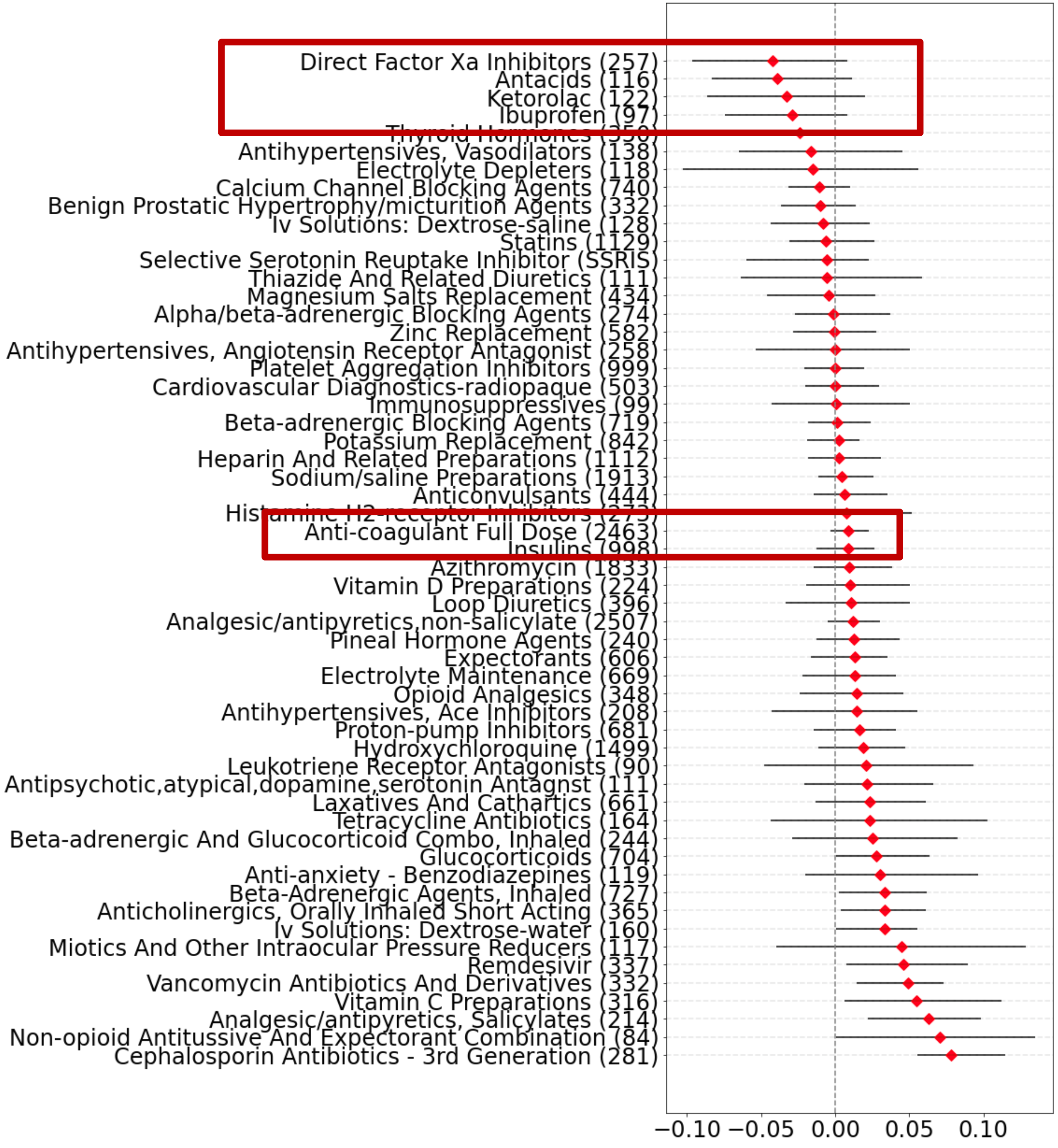


Joint work with Aphinyanaphongs Yindalon (NYU)

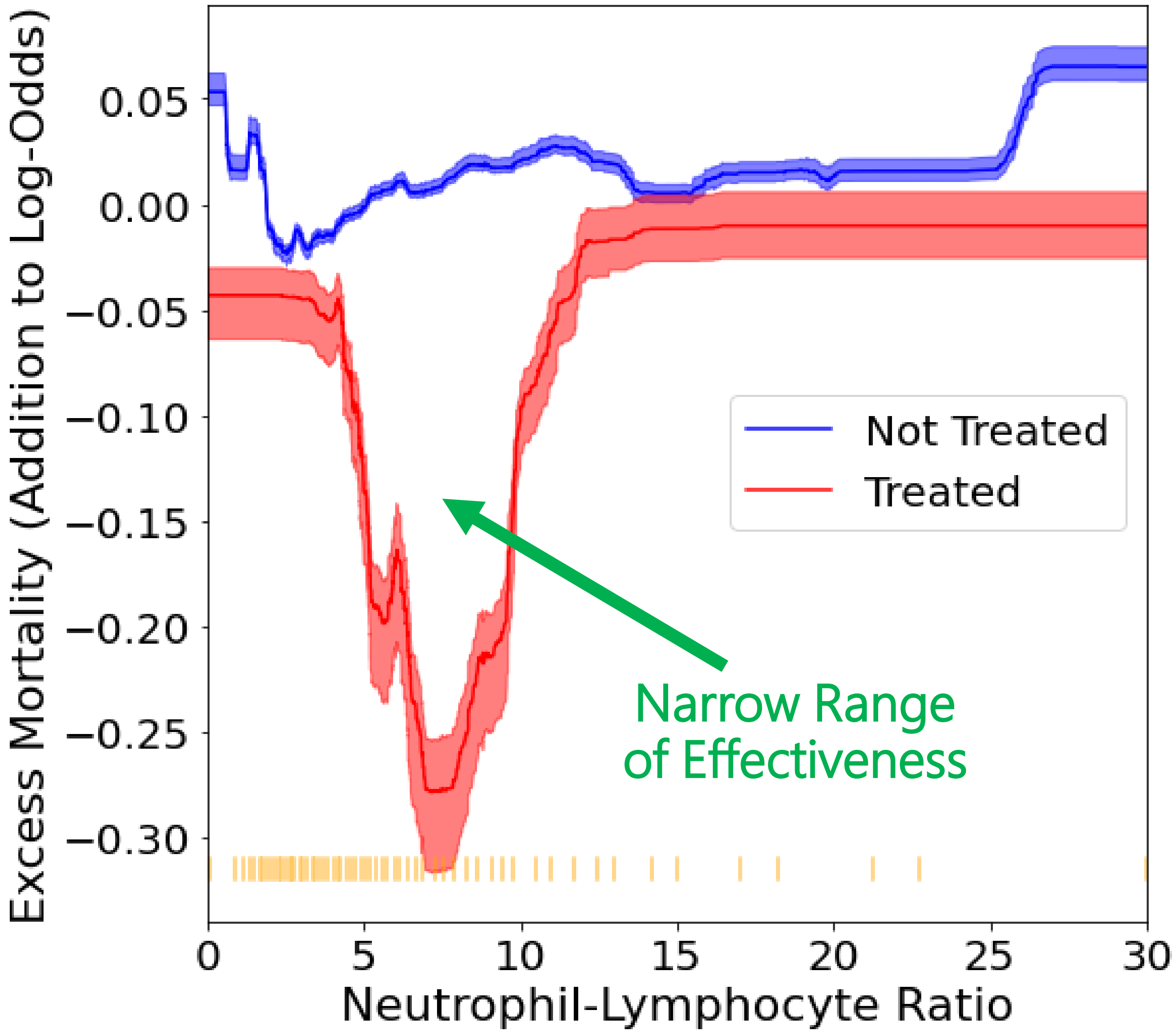
Mortality Risk from Comorbidities and Out-patient Meds



Mortality Risk from In-patient Meds



Glucocorticoid Steroids



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

Differential Privacy via EBMs

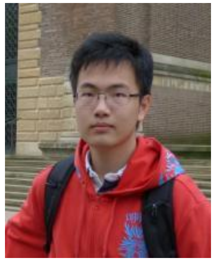
ICML 2021

Harsha Nori

MSR Aether

Rich Caruana

MSR AI



Zhiqi Bu

U. Penn

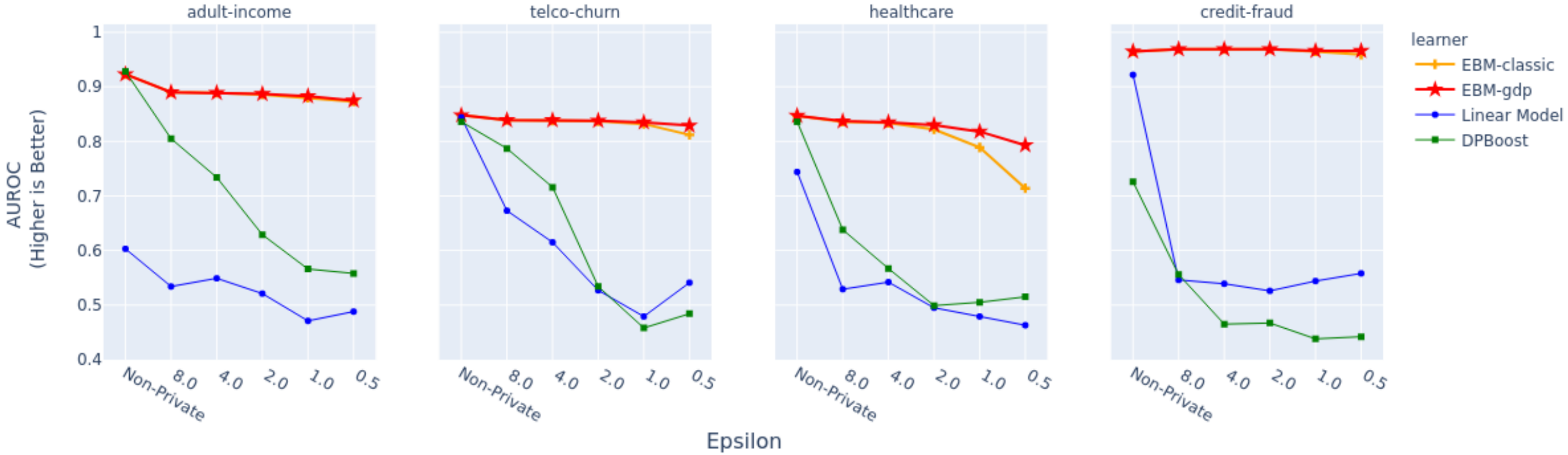
Judy Shen

Stanford

Janardhan Kulkarni

MSR Algorithms

High Accuracy, Perfect Interpretability, Strong Privacy



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

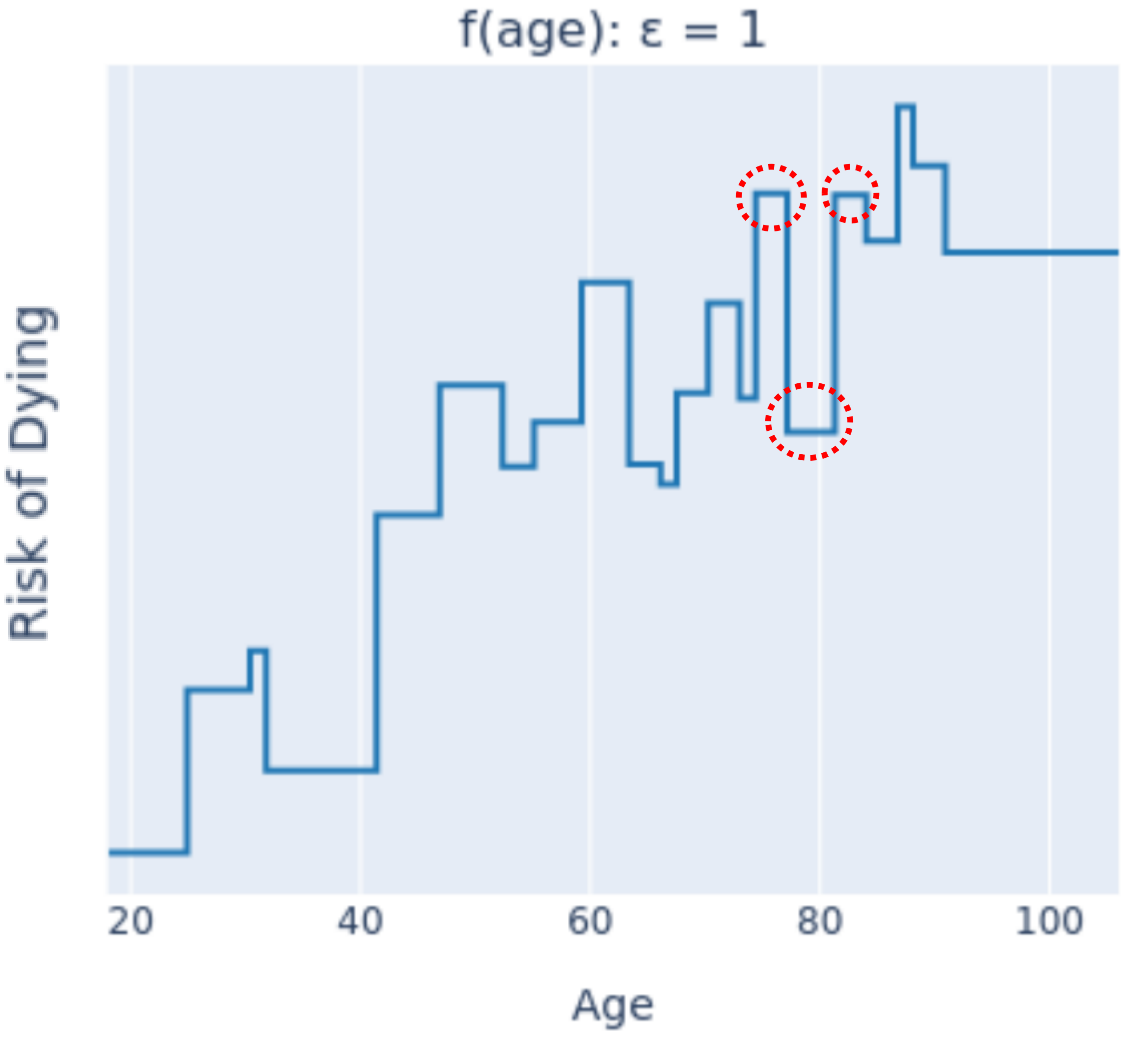
Editing Unwanted Bias

- 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

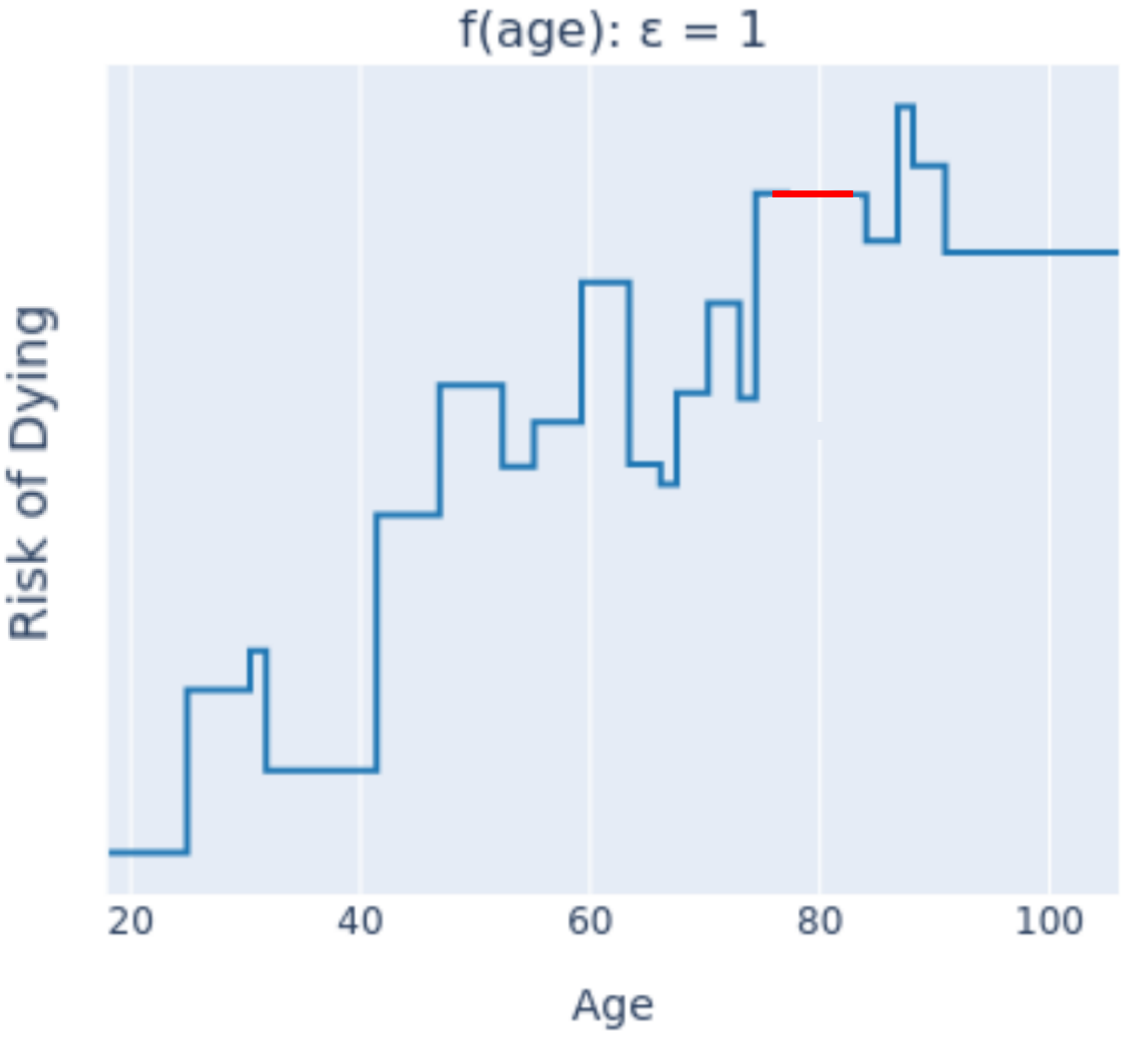
- Differential privacy can introduce noise and unwanted bias
  - Is 80 less risky than 77 and 82?

- 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"

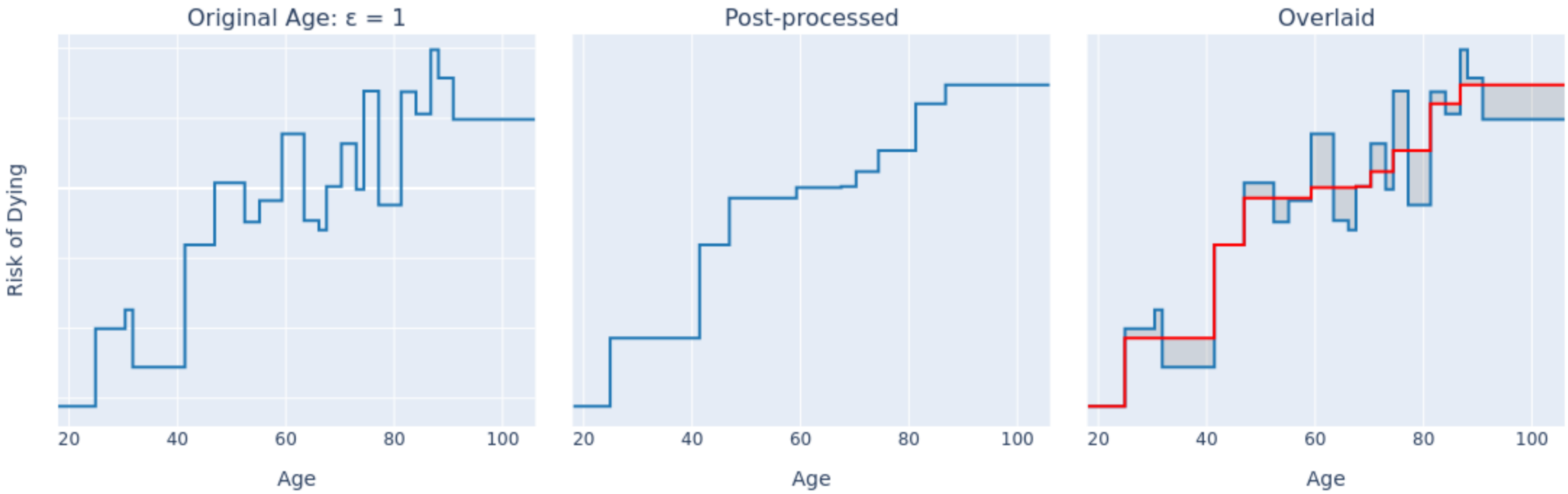
  - Intuitively makes sense – need more noise to protect smaller populations.

We can fix this!



Monotonicity for Free

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



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

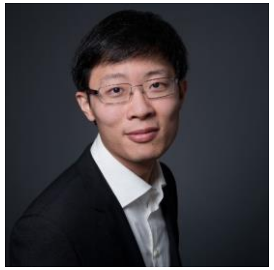
NeurIPS 2021
Spotlight
Paper

NAMs: Neural Additive Models
Interpretable Machine Learning With Neural Nets

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

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

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 using it clinically
  - Learn from your data to improve healthcare

- 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, …
  - If you work in healthcare, don't use linear/logistic regression --- they're not accurate and they lie!

InterpretML

Open-Source Tool for Intelligibility

github.com/interpretml/interpret

Microsoft Research

Thank You!

Microsoft

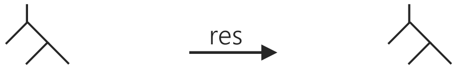
Algorithm Sketch

Iteration	$feat_1$	$feat_2$	$feat_3$	…	$feat_n$
1					

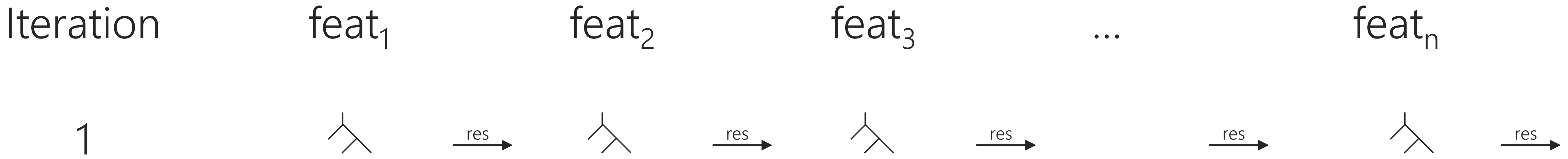
Iteration	$feat_1$	$feat_2$	$feat_3$	…	$feat_n$
1					

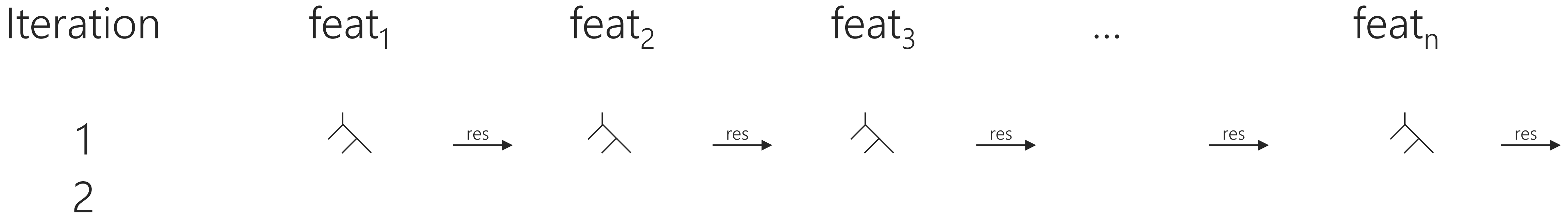
Iteration	$feat_1$		$feat_2$	$feat_3$	…	$feat_n$
1		$\xrightarrow{\text{res}}$				

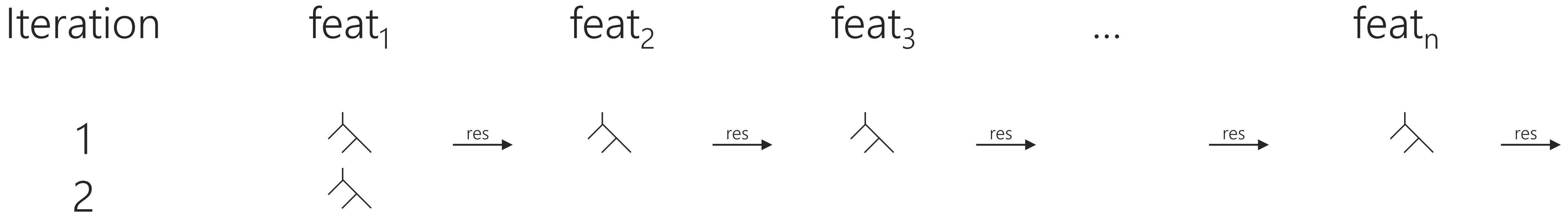
Iteration	$feat_1$	$feat_2$	$feat_3$	…	$feat_n$
1					



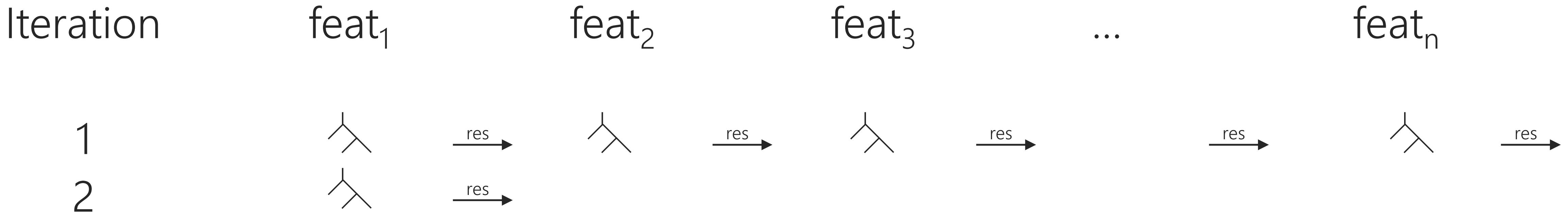
Iteration	$feat_1$	$feat_2$	$feat_3$	…	$feat_n$
1					


Iteration
feat$_1$
feat$_2$
feat$_3$
...
feat$_n$
1
res
res
res
res
res

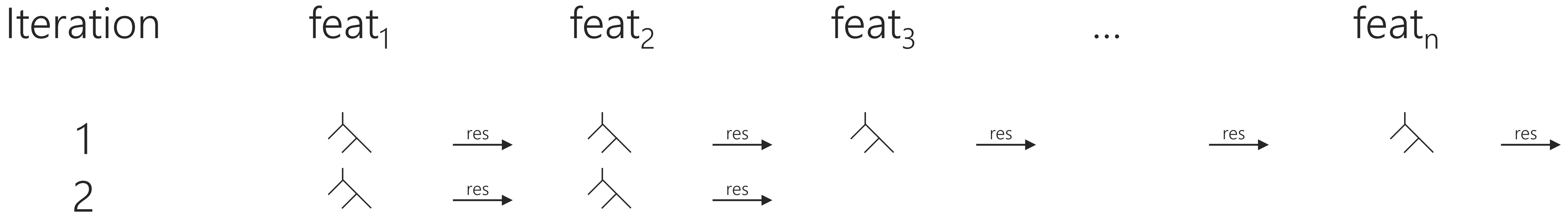

Iteration
feat1
feat2
feat3
...
featn
1
2
res
res
res
res
res

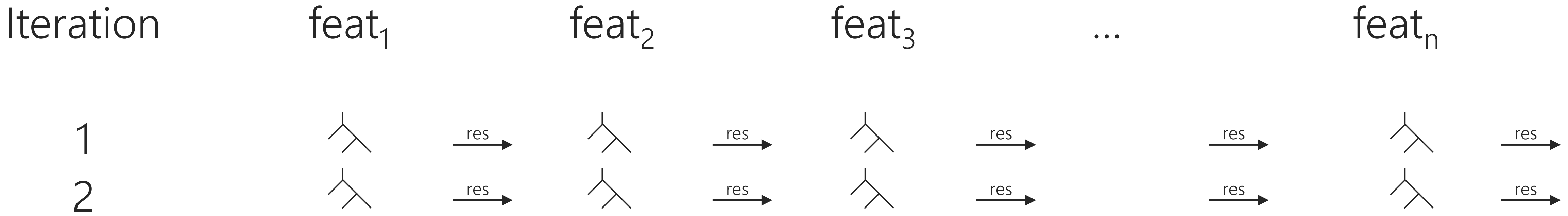


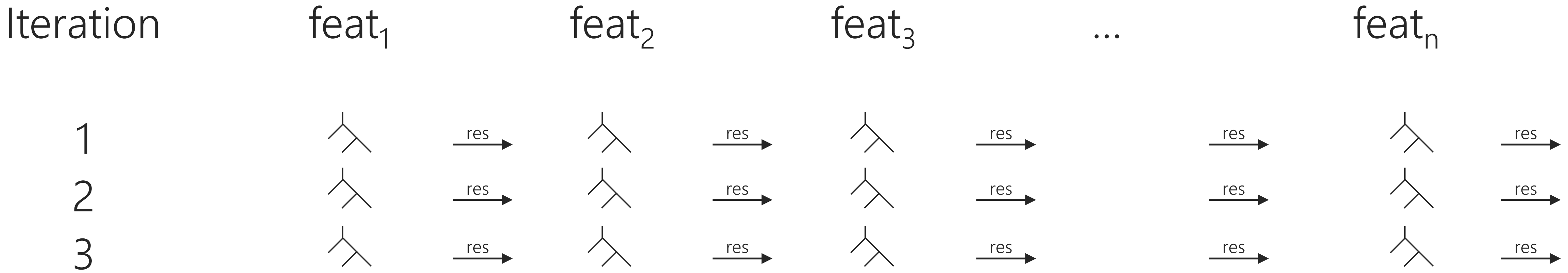
Iteration
feat1
feat2
feat3
...
featn
1
2
res
res
res
res
res

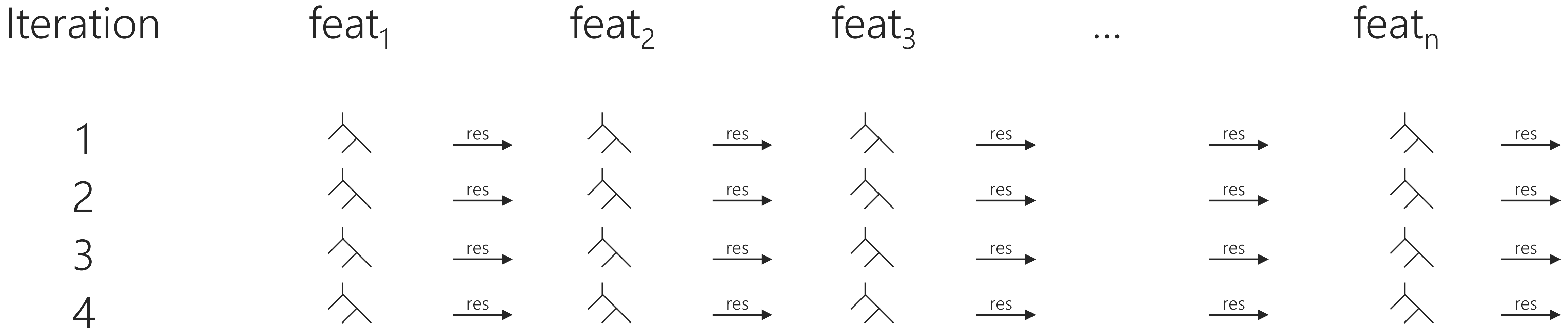


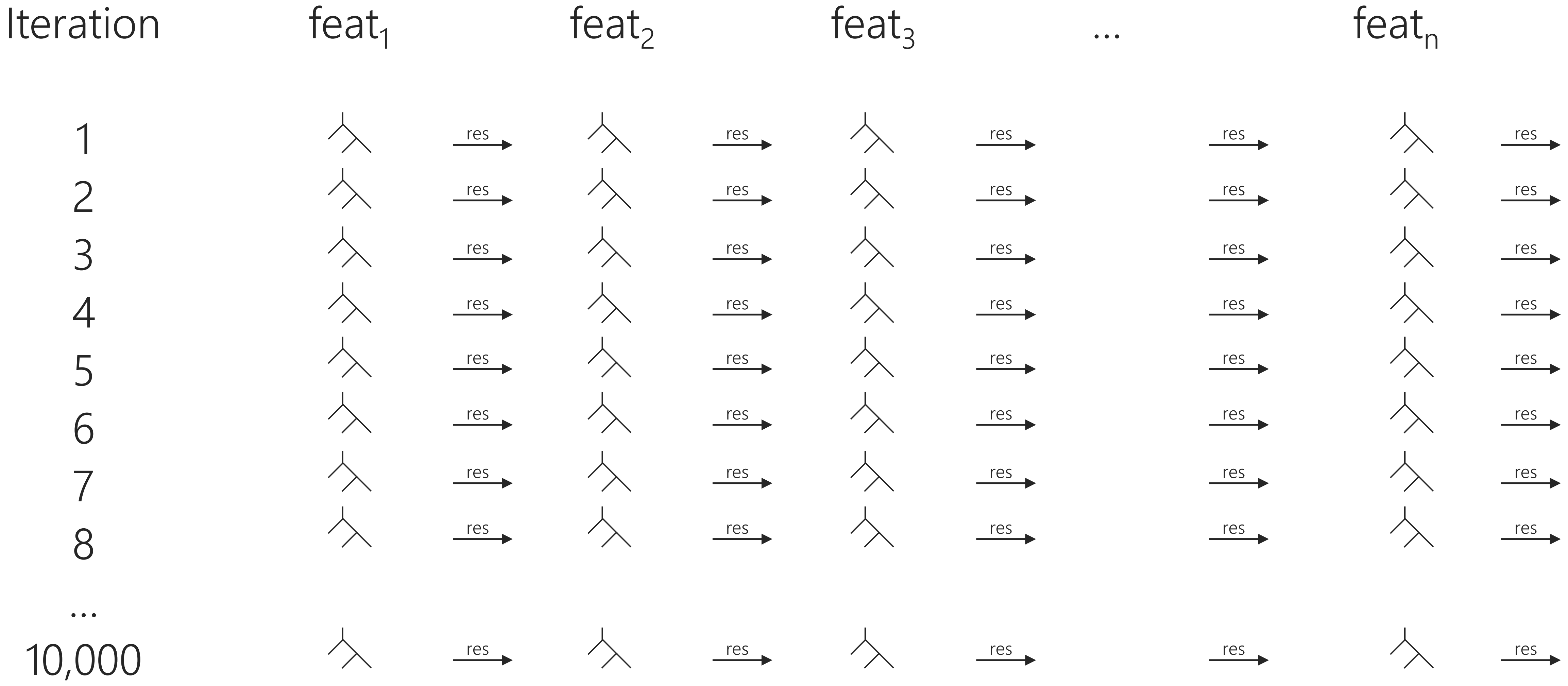
Iteration
feat$_1$
feat$_2$
feat$_3$
...
feat$_n$
1
2
res
res
res
res
res
res

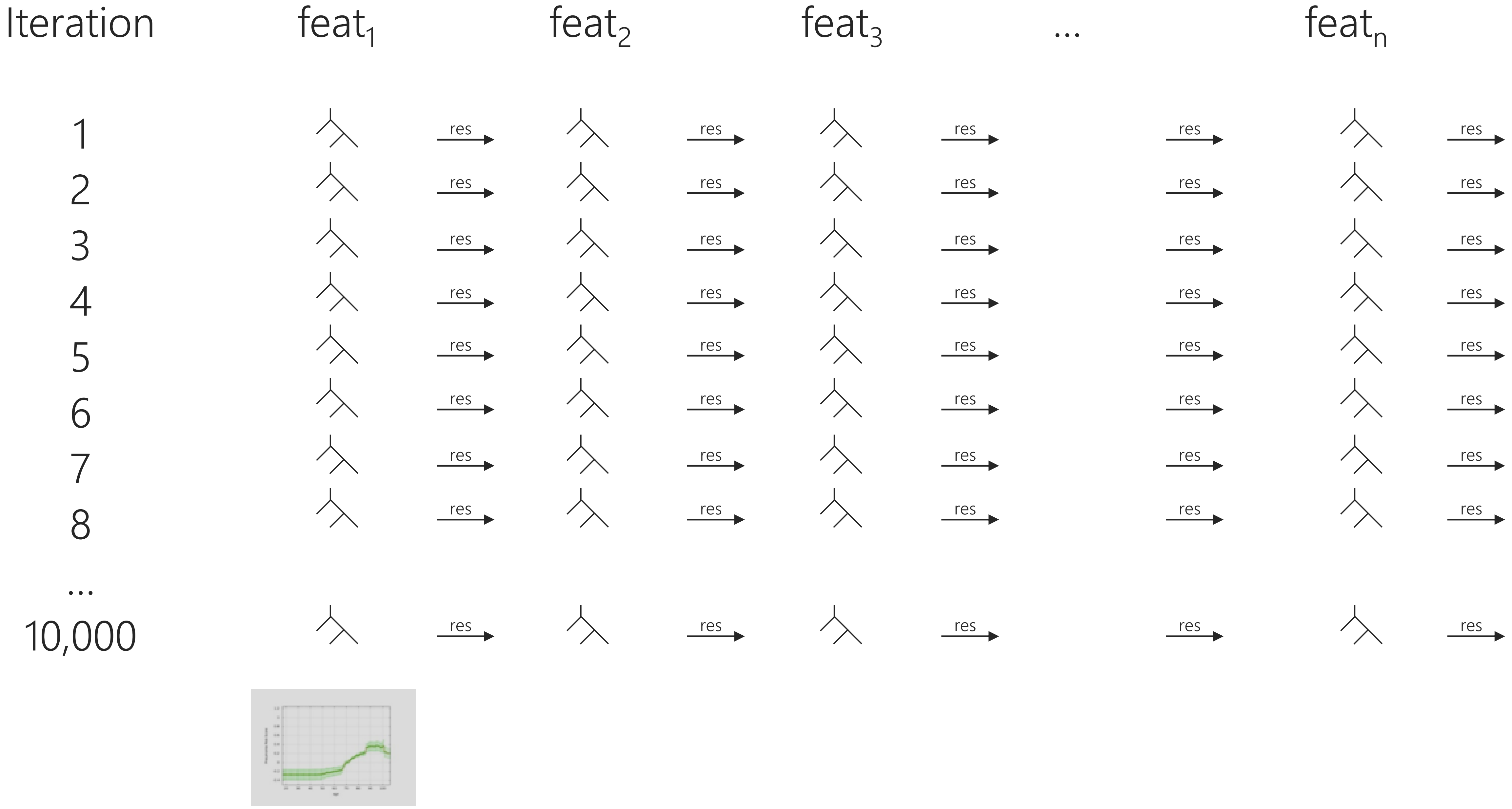

Iteration
feat1
feat2
feat3
...
featn
1
2
res


Iteration
feat1
feat2
feat3
...
featn
1
2
res

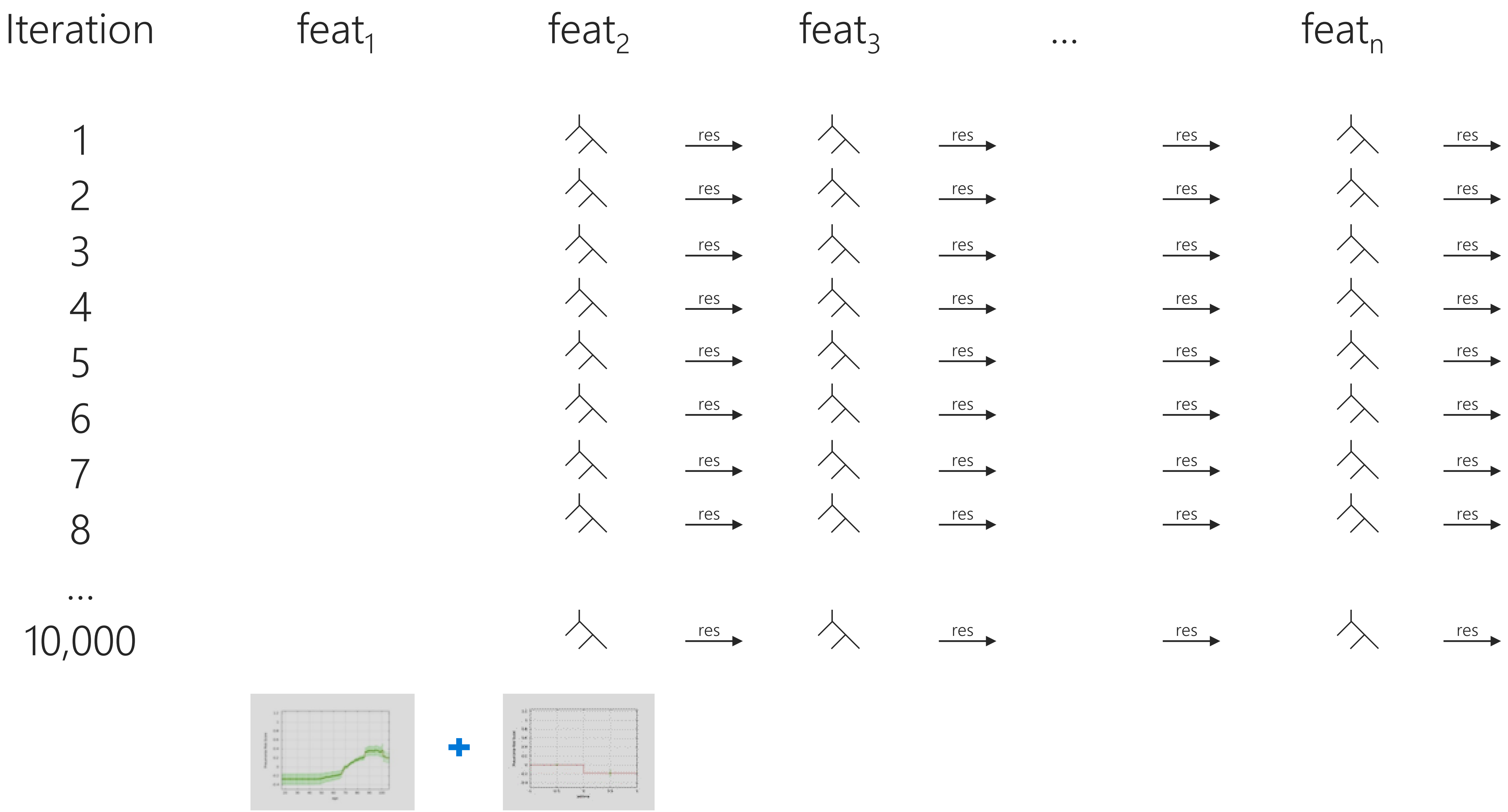

Iteration
feat1
feat2
feat3
...
featn
1
2
3
res


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

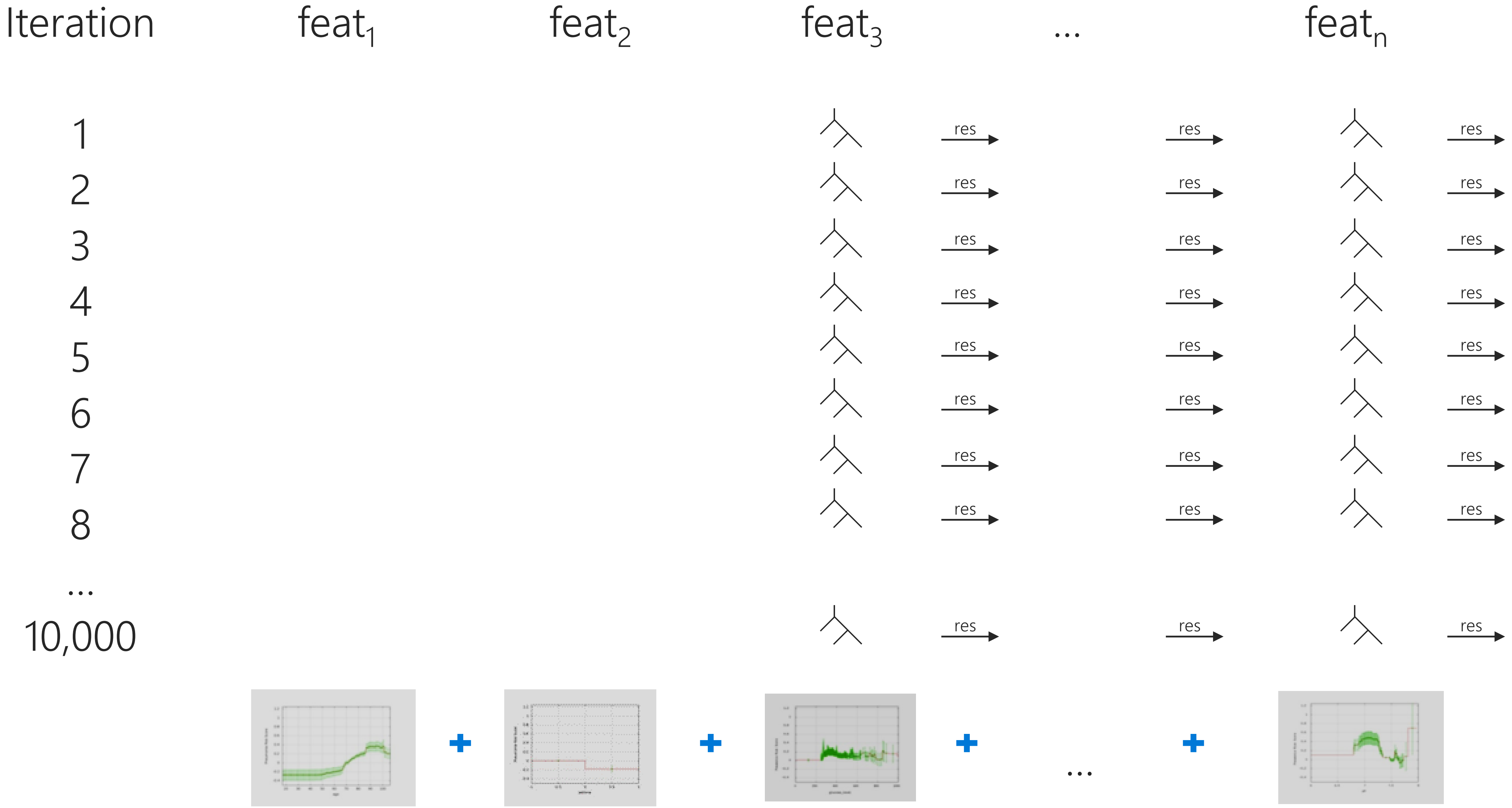

Iteration
feat1
feat2
feat3
...
featn
1
2
3
4
5
6
7
8
...
10,000
res


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

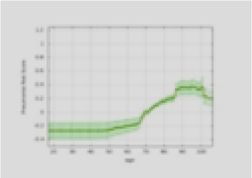
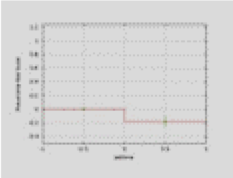
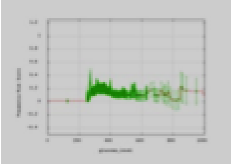
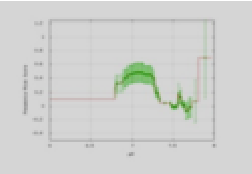

Iteration
feat$_1$
feat$_2$
feat$_3$
...
feat$_n$
1
2
3
4
5
6
7
8
...
10,000
res


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

$feat_1$ $feat_2$ $feat_3$ … $feat_n$

 +  +  + … + 

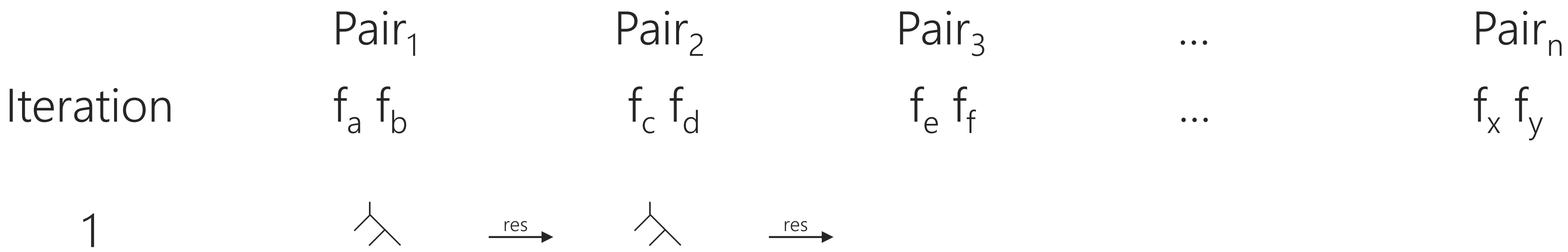
How to Fit Pairwise Interactions ?

- 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					

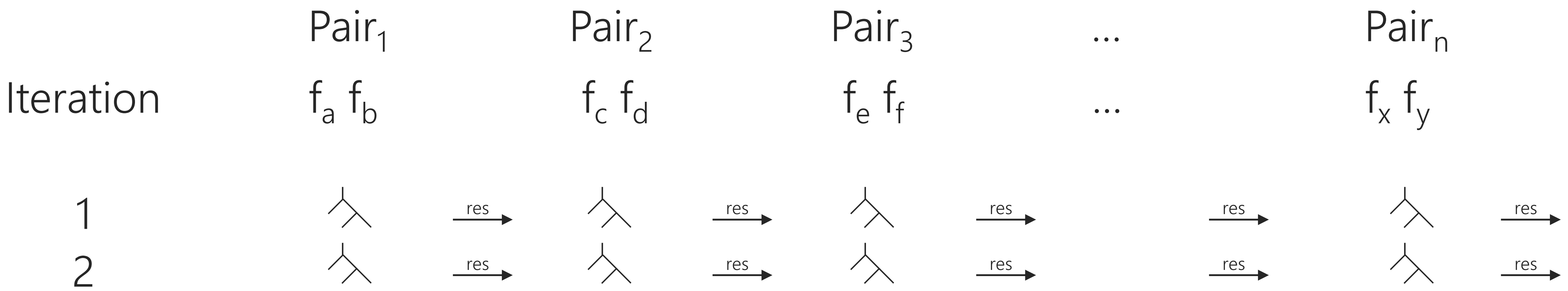
	$Pair_1$	$Pair_2$	$Pair_3$	…	$Pair_n$
Iteration	f_a f_b	f_c f_d	f_e f_f	…	f_x f_y
1					

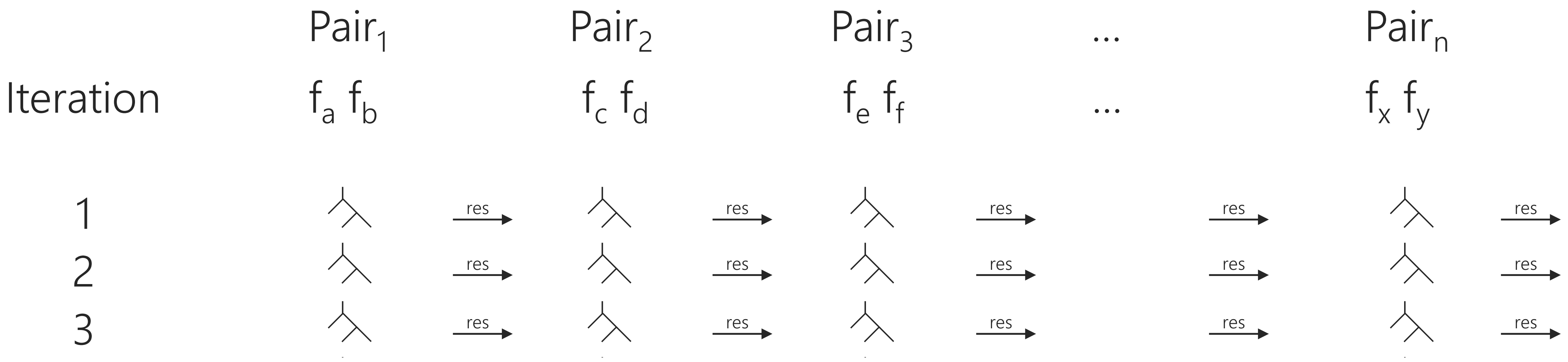
	$Pair_1$	$Pair_2$	$Pair_3$	…	$Pair_n$
Iteration	f_a f_b	f_c f_d	f_e f_f	…	f_x f_y
1	 $\xrightarrow{res}$				



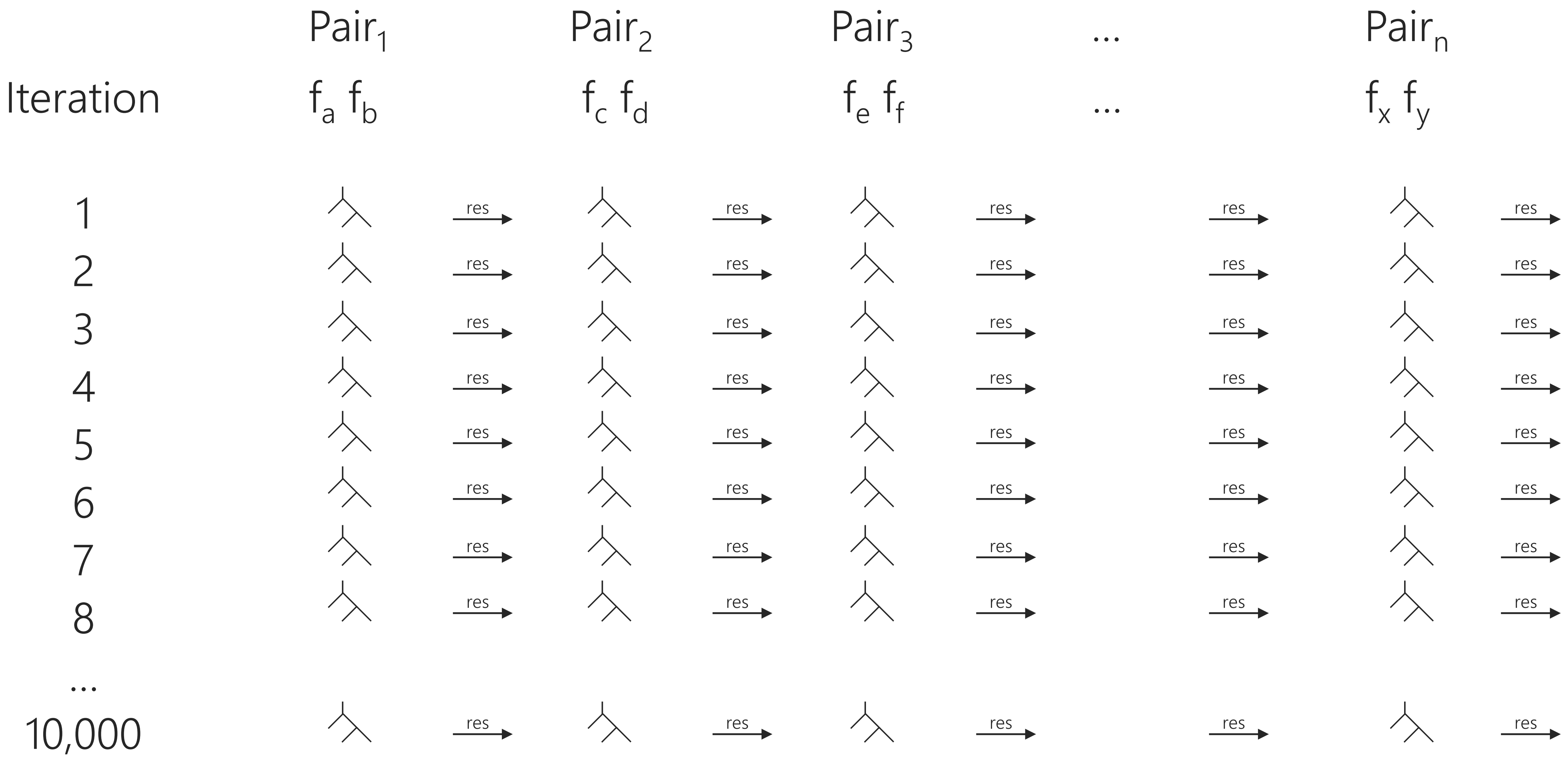
Pair$_1$
Pair$_2$
Pair$_3$
...
Pair$_n$
Iteration
f_a f_b
f_c f_d
f_e f_f
...
f_x f_y
1
res
res

	$Pair_1$		$Pair_2$		$Pair_3$		…		$Pair_n$	
Iteration	f_a f_b		f_c f_d		f_e f_f		…		f_x f_y	
1		$\xrightarrow{\text{res}}$		$\xrightarrow{\text{res}}$		$\xrightarrow{\text{res}}$		$\xrightarrow{\text{res}}$		$\xrightarrow{\text{res}}$

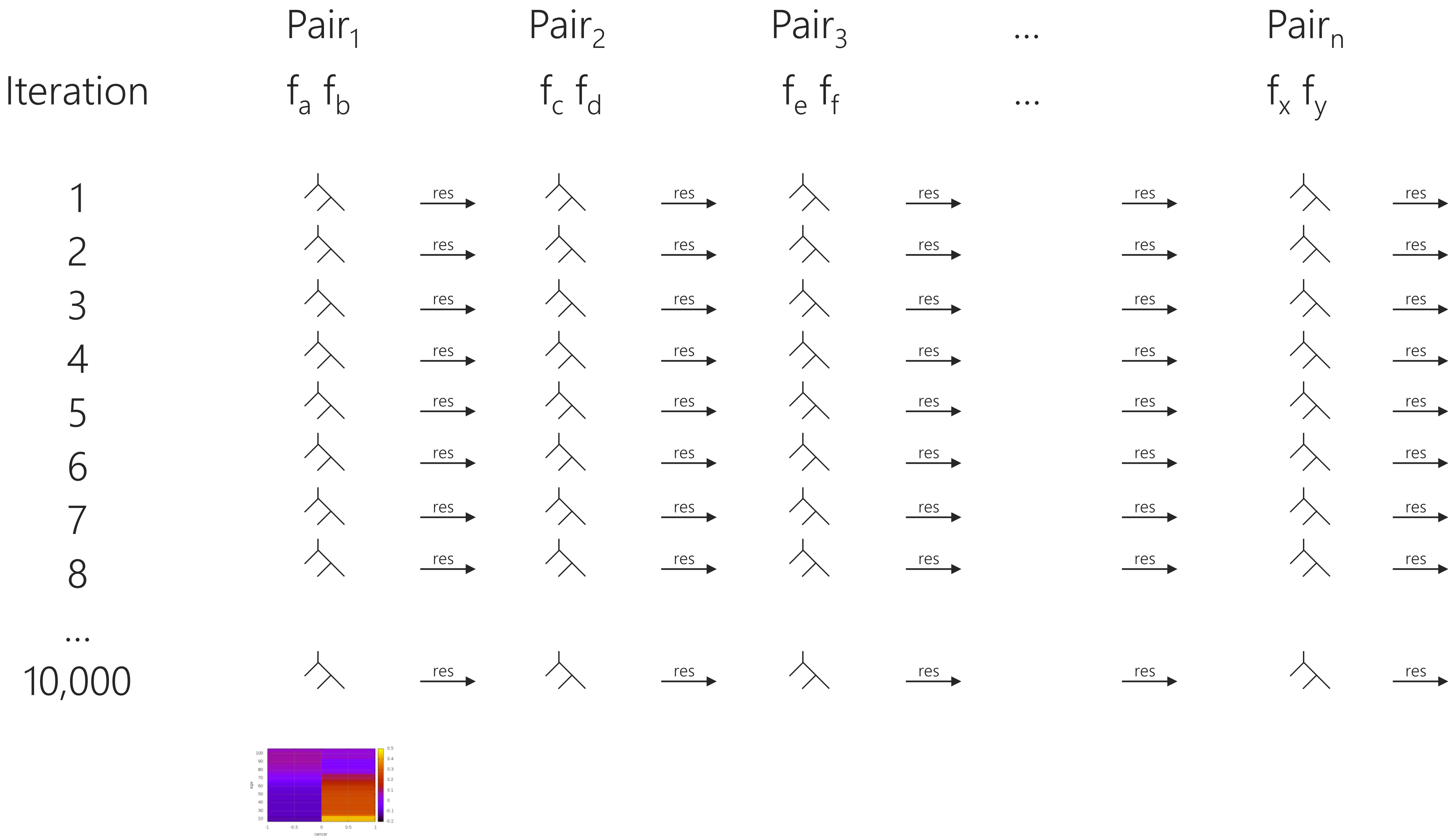

Pair1
Pair2
Pair3
...
Pairn
Iteration
fa fb
fc fd
fe ff
...
fx fy
1
2
res

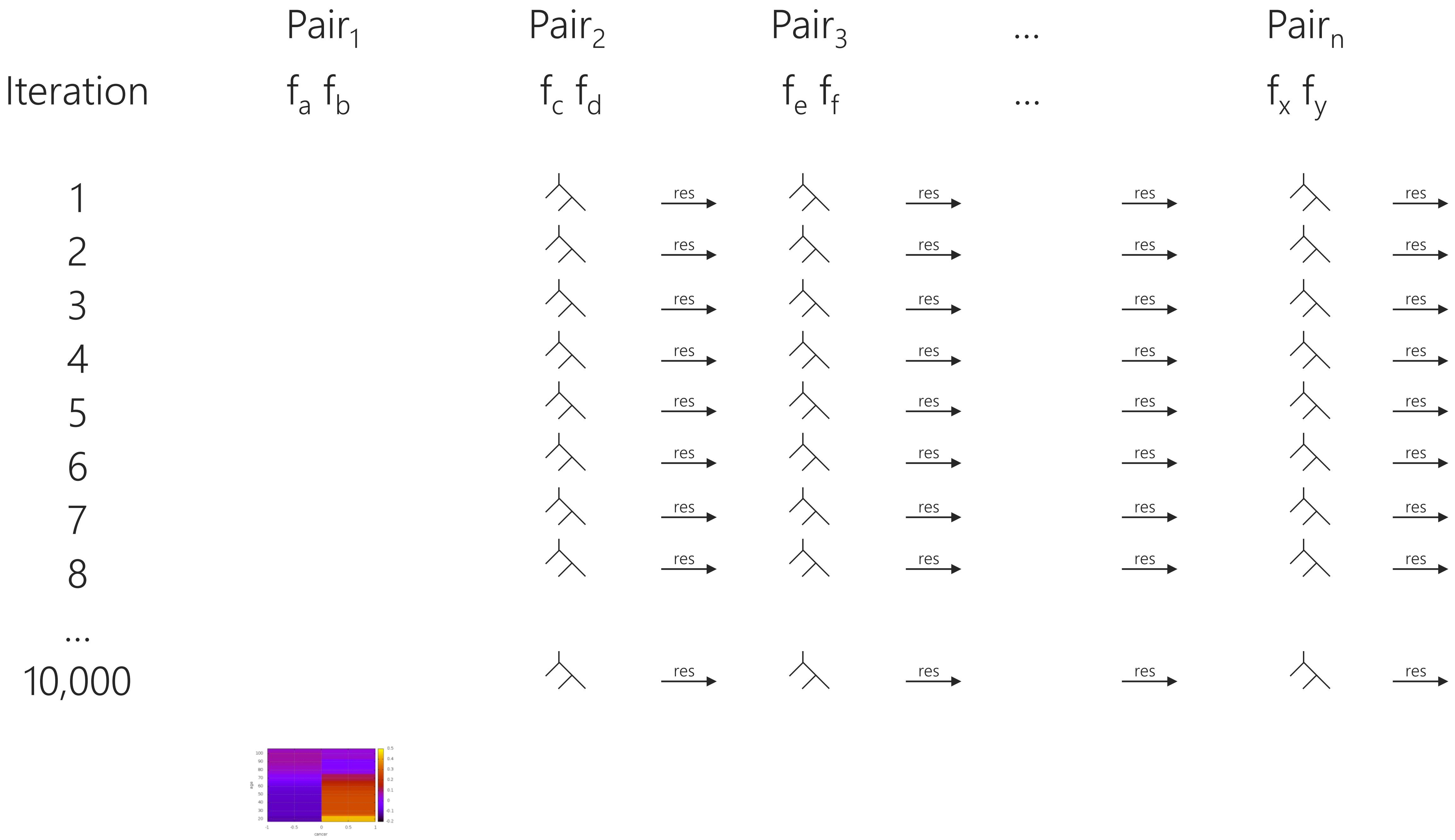


Pair1
Pair2
Pair3
...
Pairn
Iteration
fa fb
fc fd
fe ff
...
fx fy
1
2
3
res



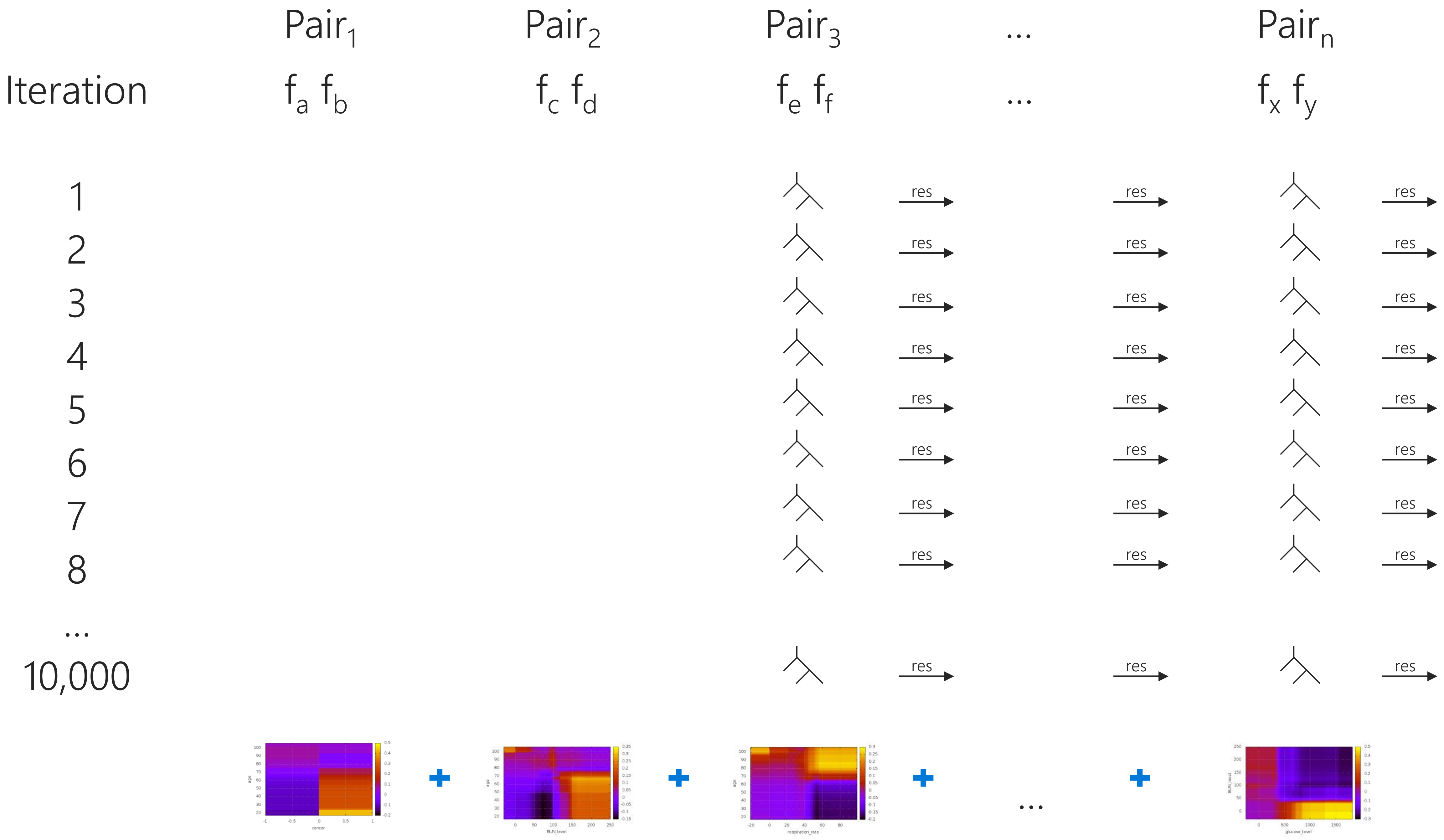
Pair$_1$ Pair$_2$ Pair$_3$ … Pair$_n$
Iteration
f_a f_b f_c f_d f_e f_f … f_x f_y
1
2
3
4
5
6
7
8
…
10,000
res


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
2
3
4
5
6
7
8
…
10,000
res



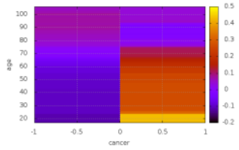
$Pair_1$ $Pair_2$ $Pair_3$ … $Pair_n$
Iteration
f_a f_b f_c f_d f_e f_f … f_x f_y
1
2
3
4
5
6
7
8
…
10,000
res
age
cancer

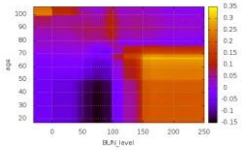
	$Pair_1$	$Pair_2$	$Pair_3$		…		$Pair_n$	
Iteration	f_a f_b	f_c f_d	f_e f_f		…		f_x f_y	
1				$\xrightarrow{res}$		$\xrightarrow{res}$		$\xrightarrow{res}$
2				$\xrightarrow{res}$		$\xrightarrow{res}$		$\xrightarrow{res}$
3				$\xrightarrow{res}$		$\xrightarrow{res}$		$\xrightarrow{res}$
4				$\xrightarrow{res}$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

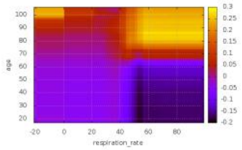


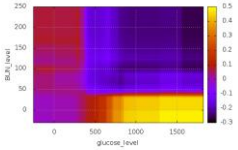
Pair$_1$
Pair$_2$
Pair$_3$
...
Pair$_n$
Iteration
f_a f_b
f_c f_d
f_e f_f
...
f_x f_y
1
2
3
4
5
6
7
8
...
10,000
res
cancer
BUN_level
respiration_rate
glucose_level

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



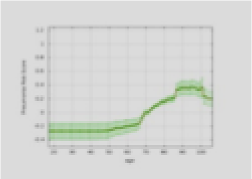
+


+


+ … +


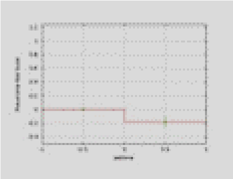
Final Model: Mains + Select Pairwise Interactions

Main_1 feat_1



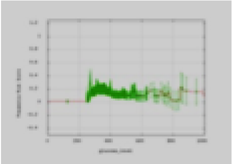
+

Main_2 feat_2



+

Main_3 feat_3



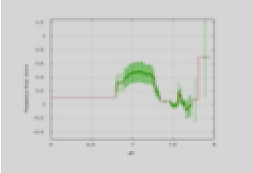
+

…
…

…

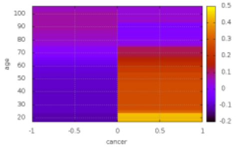
+

Main_m feat_m



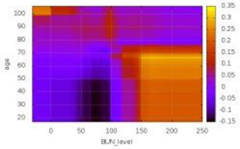
+

Pair_1 $f_a\ f_b$



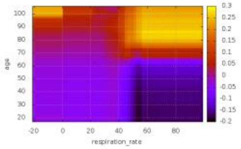
+

Pair_2 $f_c\ f_d$



+

Pair_3 $f_e\ f_f$



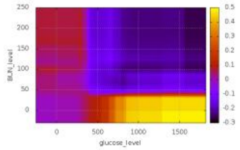
+

…
…

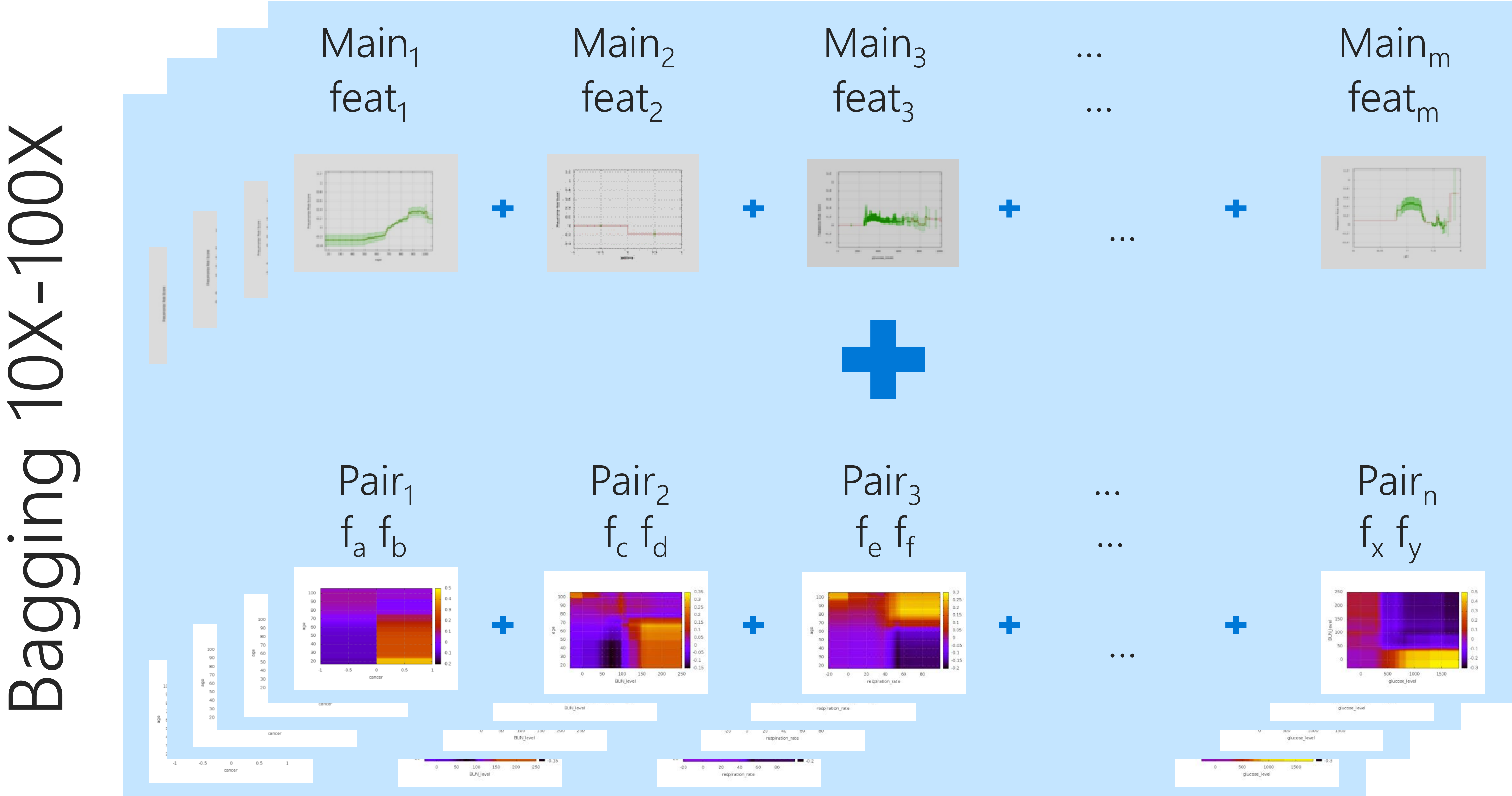
…

+

Pair_n $f_x\ f_y$

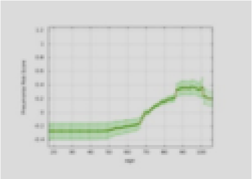
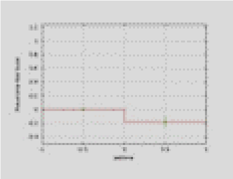
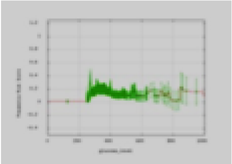
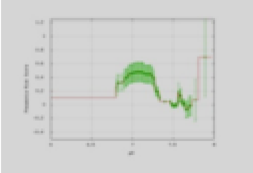


Final Model: Mains + Select Pairwise Interactions



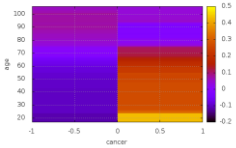
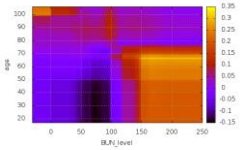
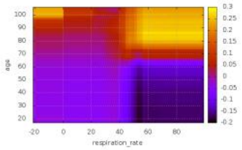
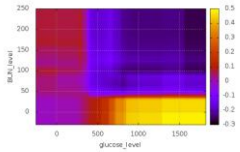
Final Model: Mains + Select Pairwise Interactions

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

 +  +  + ... + 

+

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

 +  +  + ... + 

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, ...)