



Explainable AI for Science and Medicine

Scott Lundberg

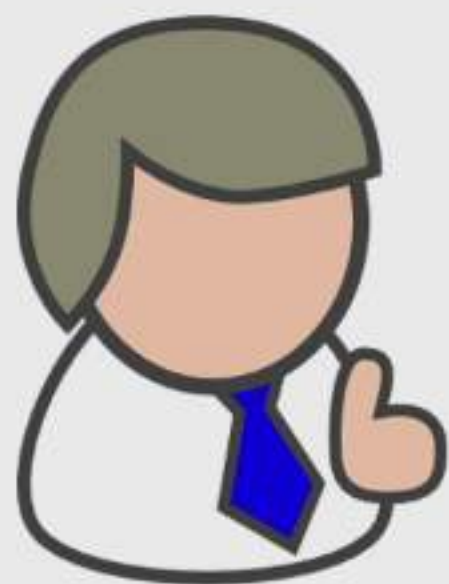
University of Washington



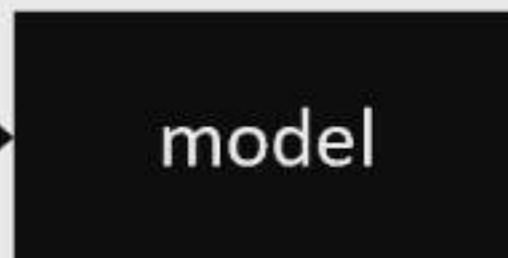
Why do we care so much about explainability
in machine learning?



John, a bank customer



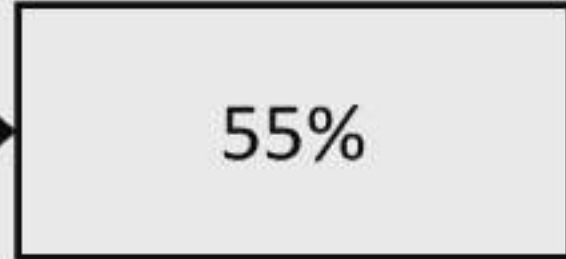
John, a bank customer



model



John, a bank customer

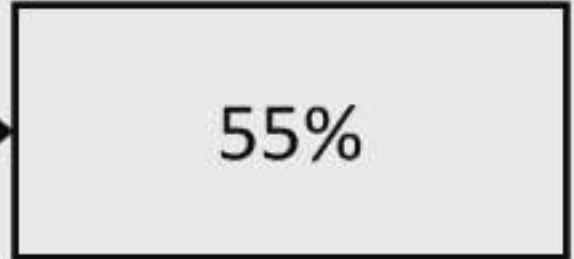


chance John will have
repayment problems

55%



John, a bank customer



chance John will have
repayment problems


No loan



Why?!



John, a bank customer



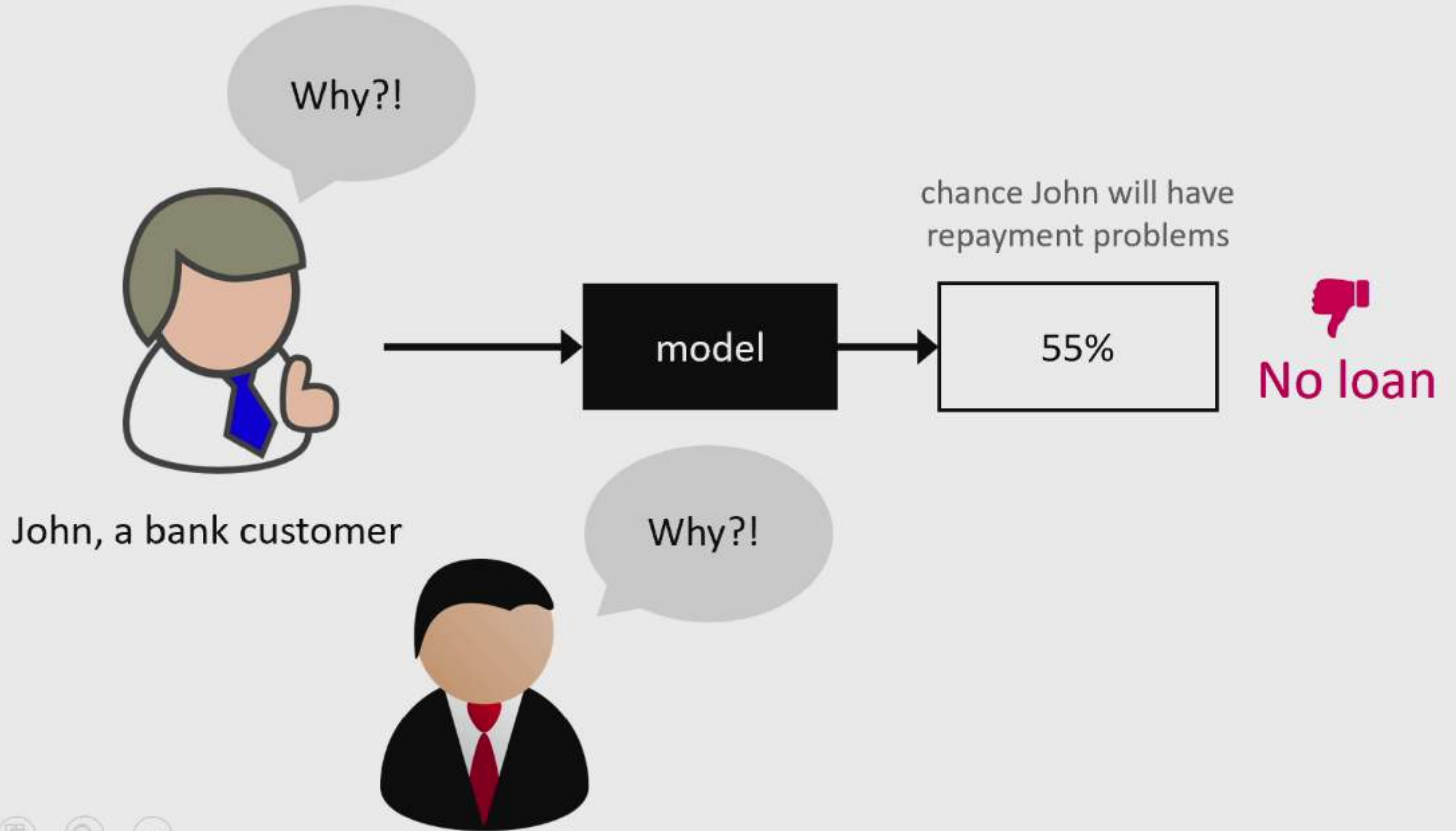
model

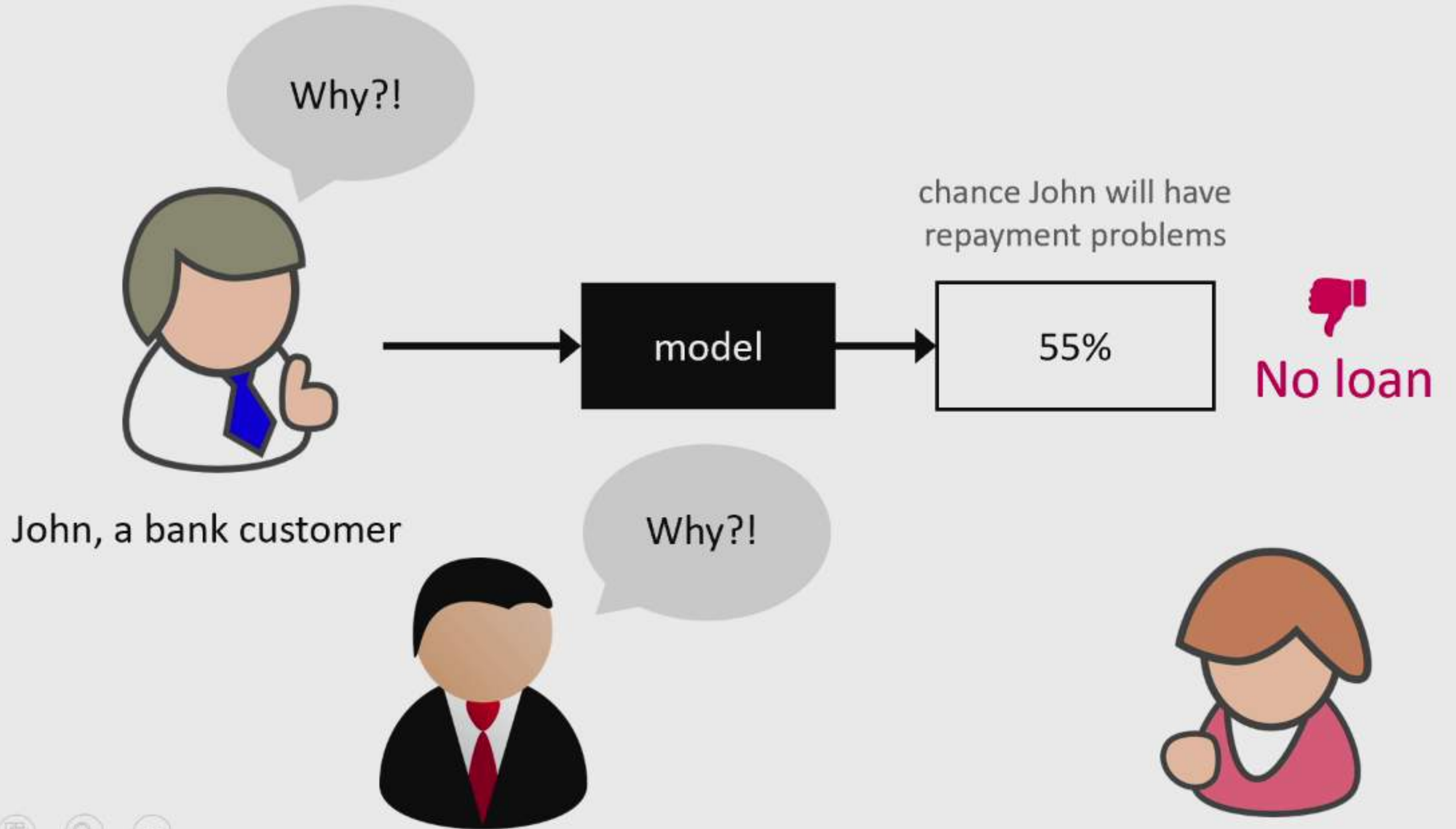


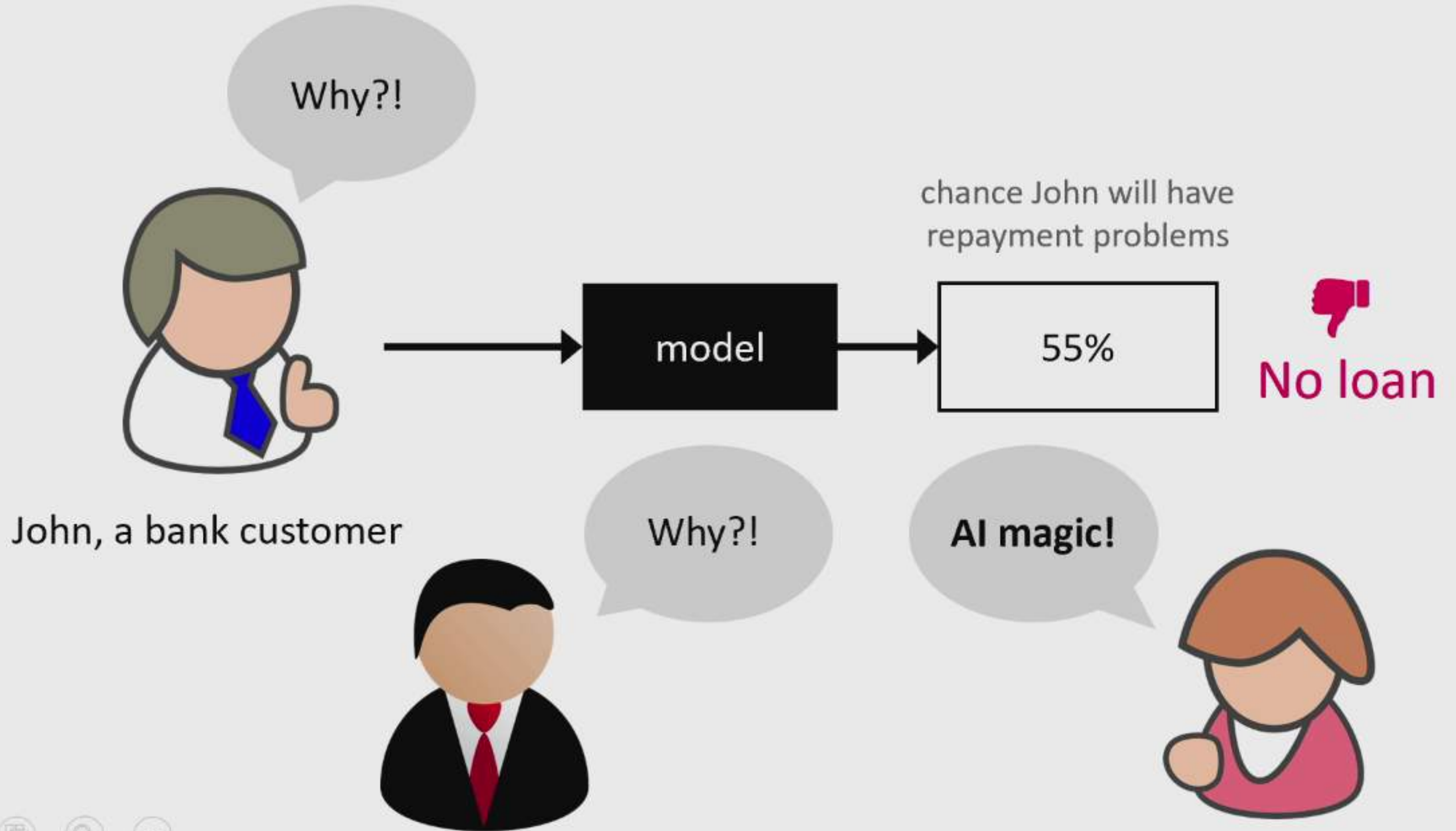
chance John will have
repayment problems

55%


No loan







Interpretable

Accurate

Complex model

Simple model



Interpretable

Accurate

Complex model



Simple model

	Interpretable	Accurate
Complex model	X	✓
Simple model	✓	X

Interpretable or accurate: **choose one.**

	Interpretable	Accurate
Complex model	X	✓
Simple model	✓	X

Interpretable or accurate: **choose one.**



	Interpretable	Accurate
Complex model	X	✓
Simple model	✓	X

Interpretable or accurate: **choose one.**



	Interpretable	Accurate
Complex model	X	✓
Simple model	✓ →	X

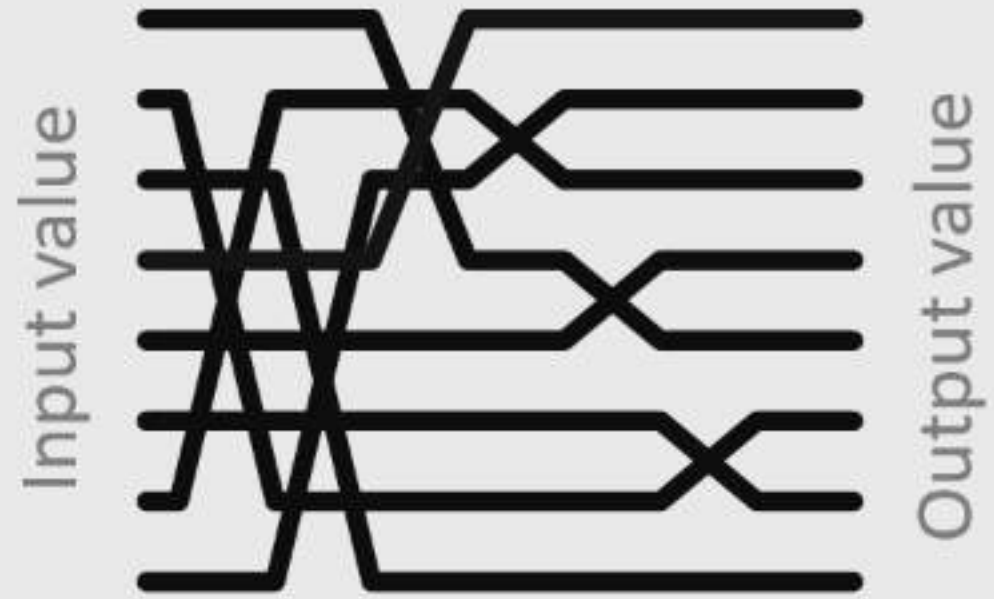
Interpretable or accurate: **choose one.**



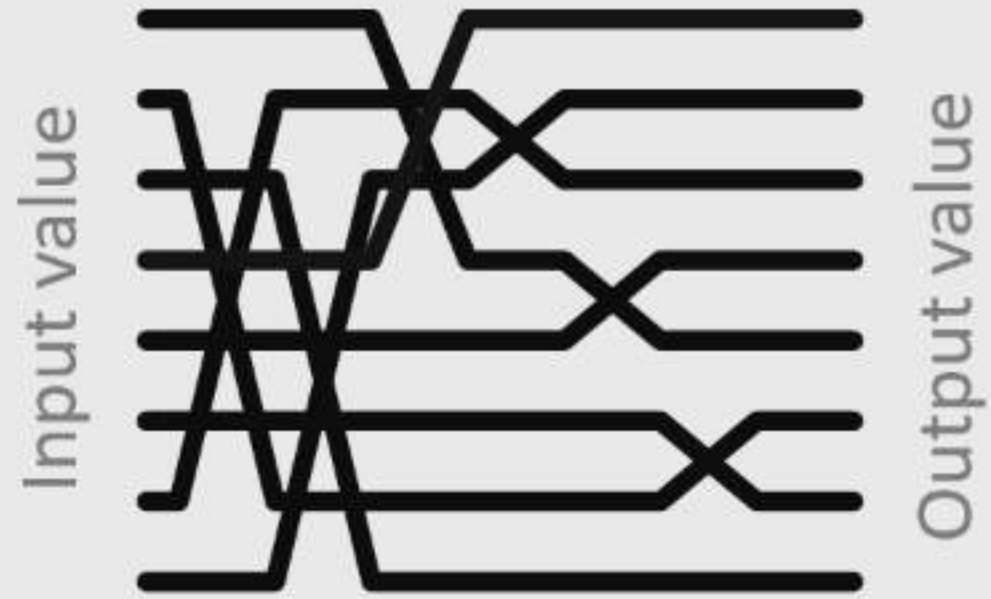
	Interpretable	Accurate
Complex model	X	✓
Simple model	✓	X

Interpretable or accurate: **choose one.**





Complex models are inherently complex!



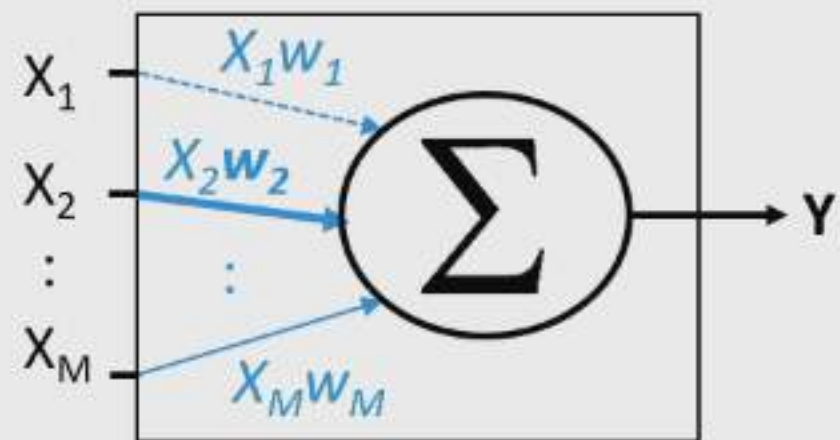
Complex models are inherently complex!



But a single prediction involves only a small piece of that complexity.

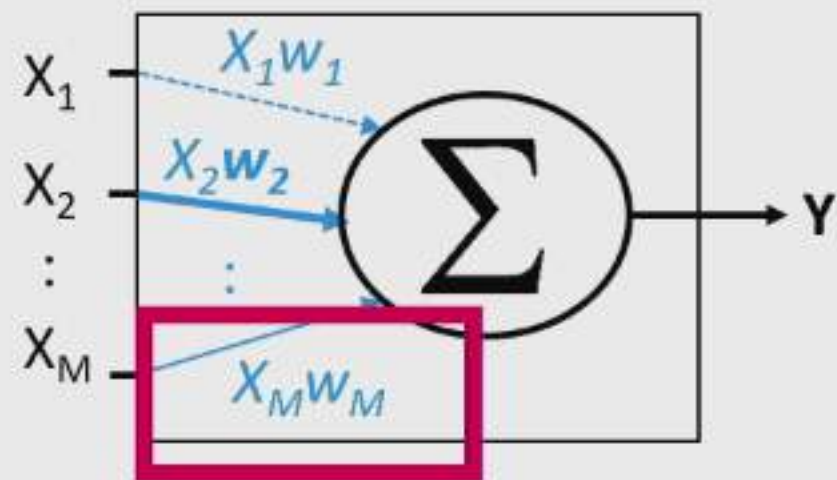
Linear model

X: Features **Y:** Outcome



Linear model

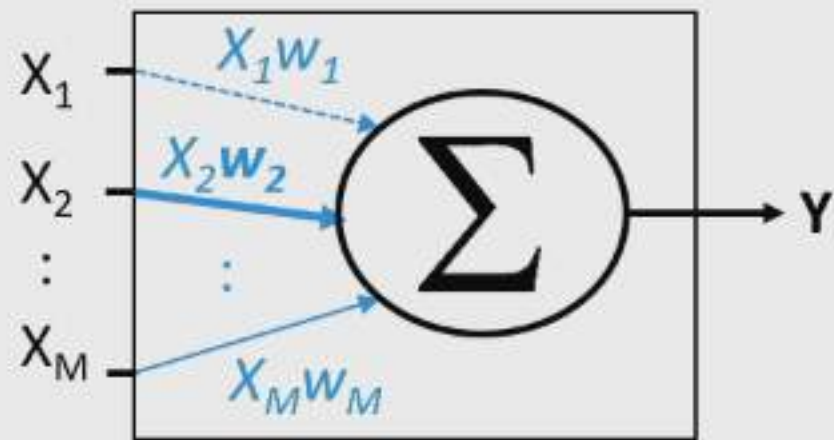
X: Features **Y:** Outcome



Credit attributed to feature X_M

Linear model

X: Features **Y:** Outcome



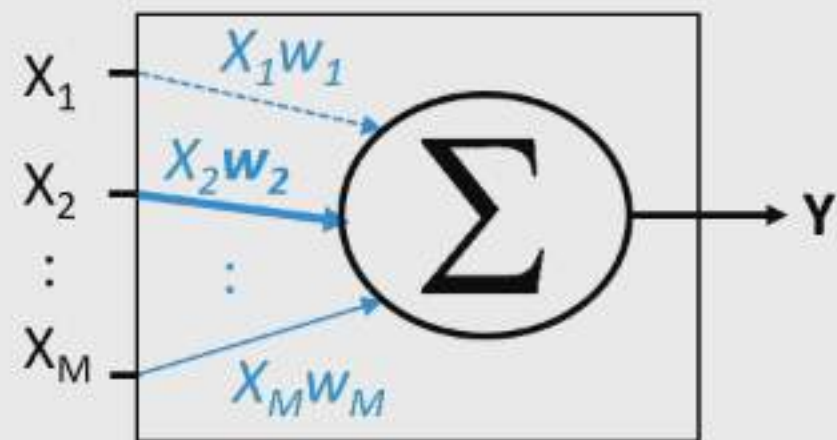
Complex model $f(\cdot)$

Black Box



Linear model

X: Features **Y:** Outcome



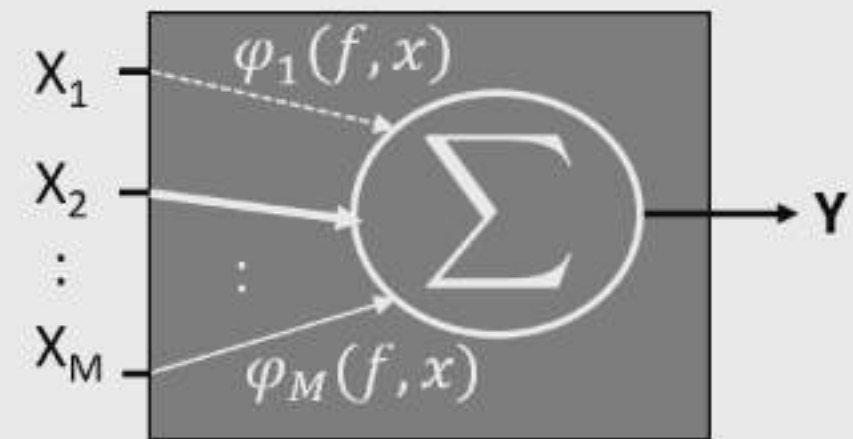
Complex model $f(\cdot)$

Black Box



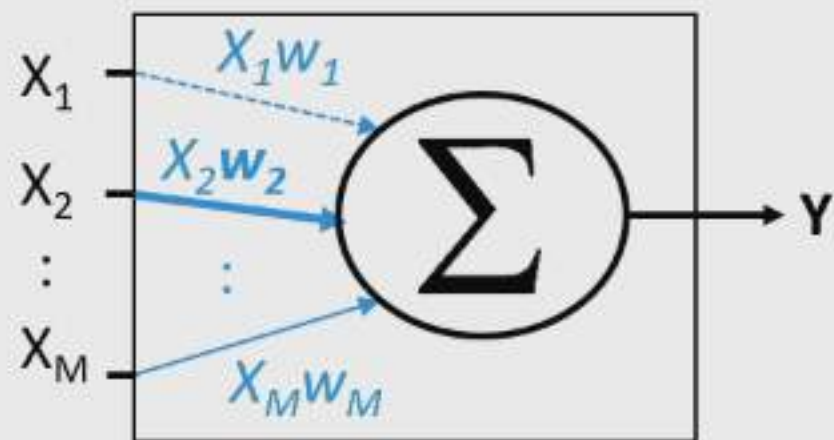
Additive feature attribution

For a particular prediction



Linear model

X: Features **Y:** Outcome



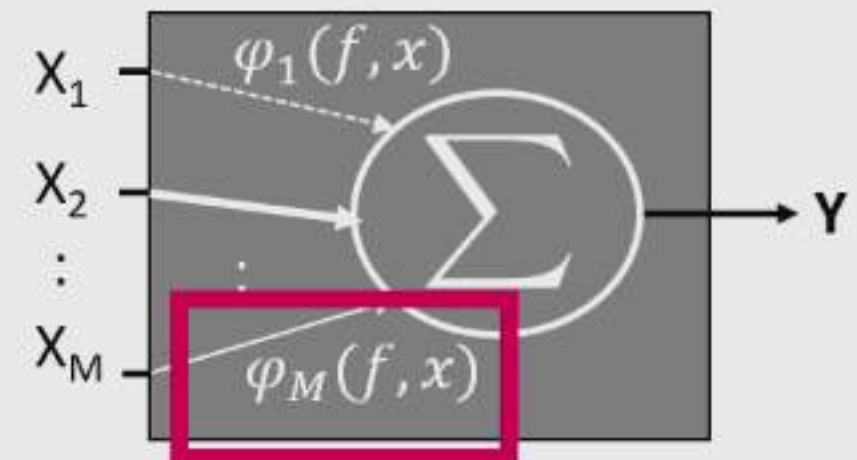
Complex model $f(\cdot)$

Black Box



Additive feature attribution

For a particular prediction



Credit attributed to feature X_M

LIME

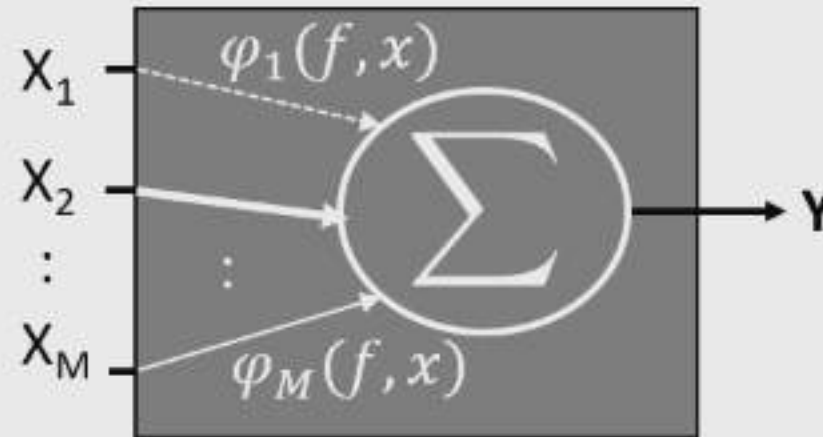
Ribeiro et al. 2016

Shapley reg. values

Lipovetsky et al. 2001

QII

Datta et al. 2016



Shapley sampling

Štrumbelj et al. 2011

DeepLIFT

Shrikumar et al. 2016

Relevance prop.

Bach et al. 2015

Saabas

Saabas 2014

LIME

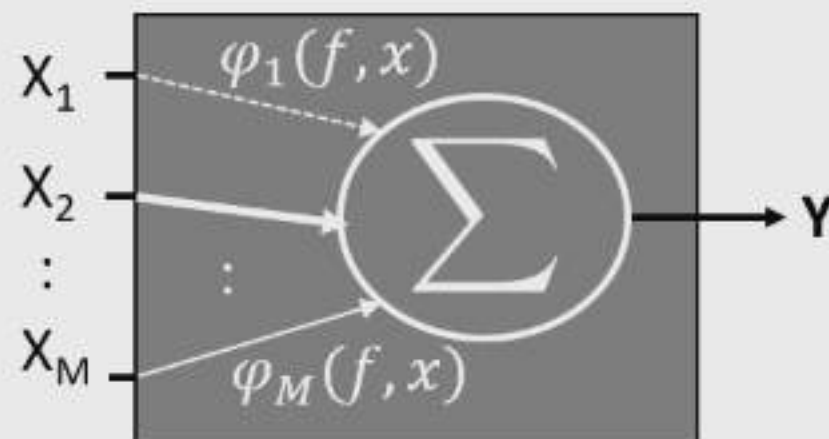
Ribeiro et al. 2016

Shapley reg. values

Lipovetsky et al. 2001

QII

Datta et al. 2016



Shapley sampling

Štrumbelj et al. 2011

DeepLIFT

Shrikumar et al. 2016

Relevance prop.

Bach et al. 2015

Saabas

Saabas 2014

LIME

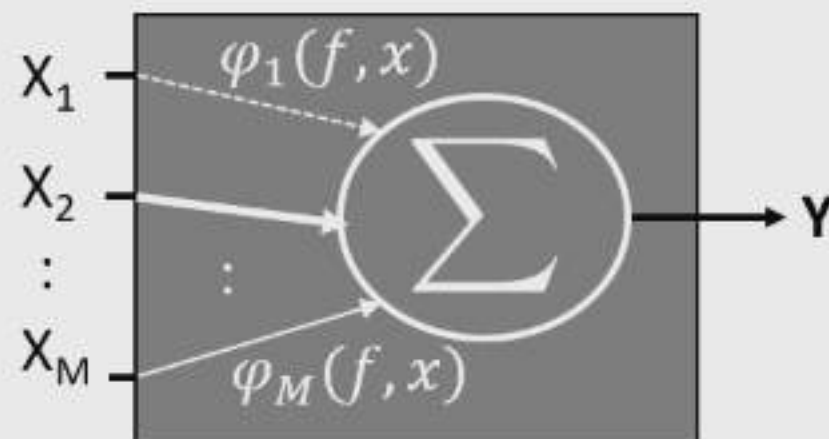
Ribeiro et al. 2016

Shapley reg. values

Lipovetsky et al. 2001

QII

Datta et al. 2016



Shapley sampling

Štrumbelj et al. 2011

DeepLIFT

Shrikumar et al. 2016

Relevance prop.

Bach et al. 2015

Saabas

Saabas 2014

LIME

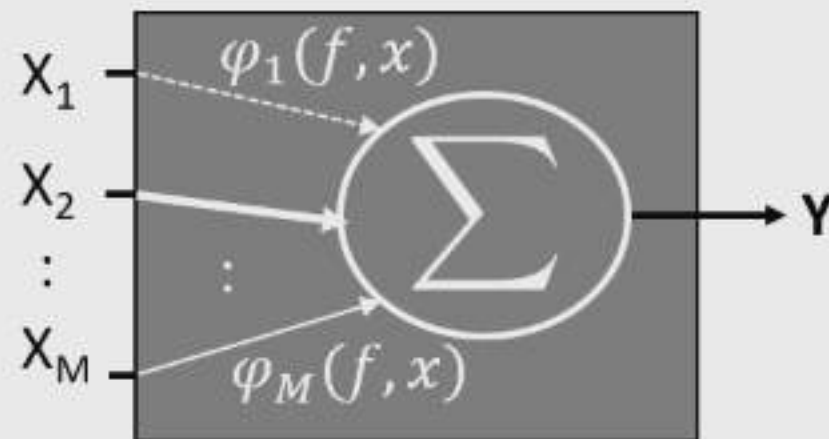
Ribeiro et al. 2016

Shapley reg. values

Lipovetsky et al. 2001

QII

Datta et al. 2016



Shapley sampling

Štrumbelj et al. 2011

DeepLIFT

Shrikumar et al. 2016

Relevance prop.

Bach et al. 2015

Saabas

Saabas 2014

LIME

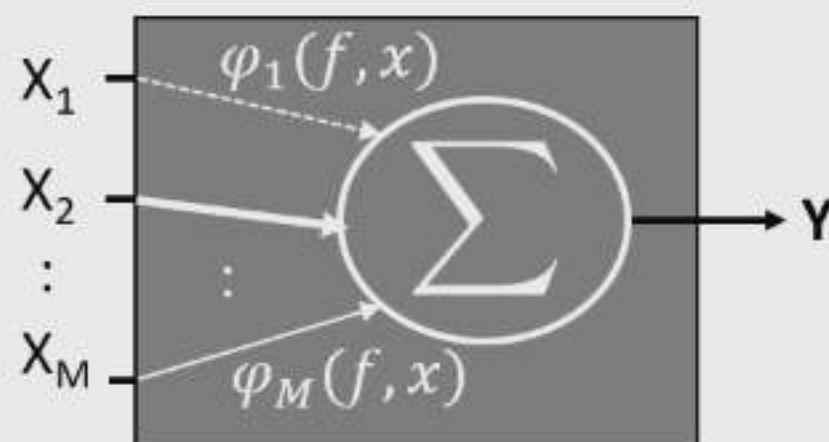
Ribeiro et al. 2016

Shapley reg. values

Lipovetsky et al. 2001

QII

Datta et al. 2016



Shapley sampling

Štrumbelj et al. 2011

DeepLIFT

Shrikumar et al. 2016

Relevance prop.

Bach et al. 2015

Saabas

Saabas 2014

Additive feature attribution methods

LIME

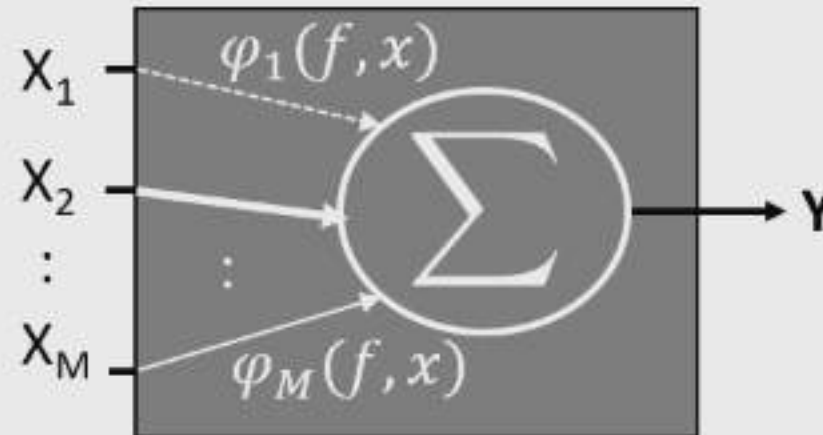
Ribeiro et al. 2016

Shapley reg. values

Lipovetsky et al. 2001

QII

Datta et al. 2016



Shapley sampling

Štrumbelj et al. 2011

DeepLIFT

Shrikumar et al. 2016

Relevance prop.

Bach et al. 2015

Saabas

Saabas 2014

Additive feature attribution methods

LIME

DeepLIFT

Shapley reg. values

Relevance prop.

QII

Shapley sampling

Saabas

Additive feature attribution methods

LIME

DeepLIFT

Shapley reg. values

Relevance prop.

QII

Shapley sampling

Saabas

Additive feature attribution methods

LIME

DeepLIFT

Shapley reg. values

Relevance prop.

QII

Shapley sampling

Saabas

Additive feature attribution methods

LIME

DeepLIFT

Shapley reg. values

Relevance prop.

QII

Shapley sampling

Saabas

SHAP

Lundberg and Lee. A unified approach to interpreting model predictions
NeurIPS 2017 (*oral presentation*)

Lundberg and Lee. An unexpected unity among methods for interpreting model predictions
NeurIPS Workshop on Interpretable Machine Learning in Complex Systems 2016 (*best paper award*)

How should we define $\varphi_i(f, x)$?
(the credit for the i 'th feature)





Base rate

20%

$E[f(x)]$

0





Base rate

20%

$E[f(x)]$

Prediction for John

55%

$f(x)$

0





Base rate

Prediction for John

20%

55%

0

$E[f(x)]$

$f(x)$



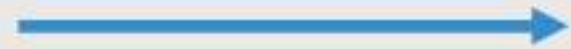
How did we get here?



20%

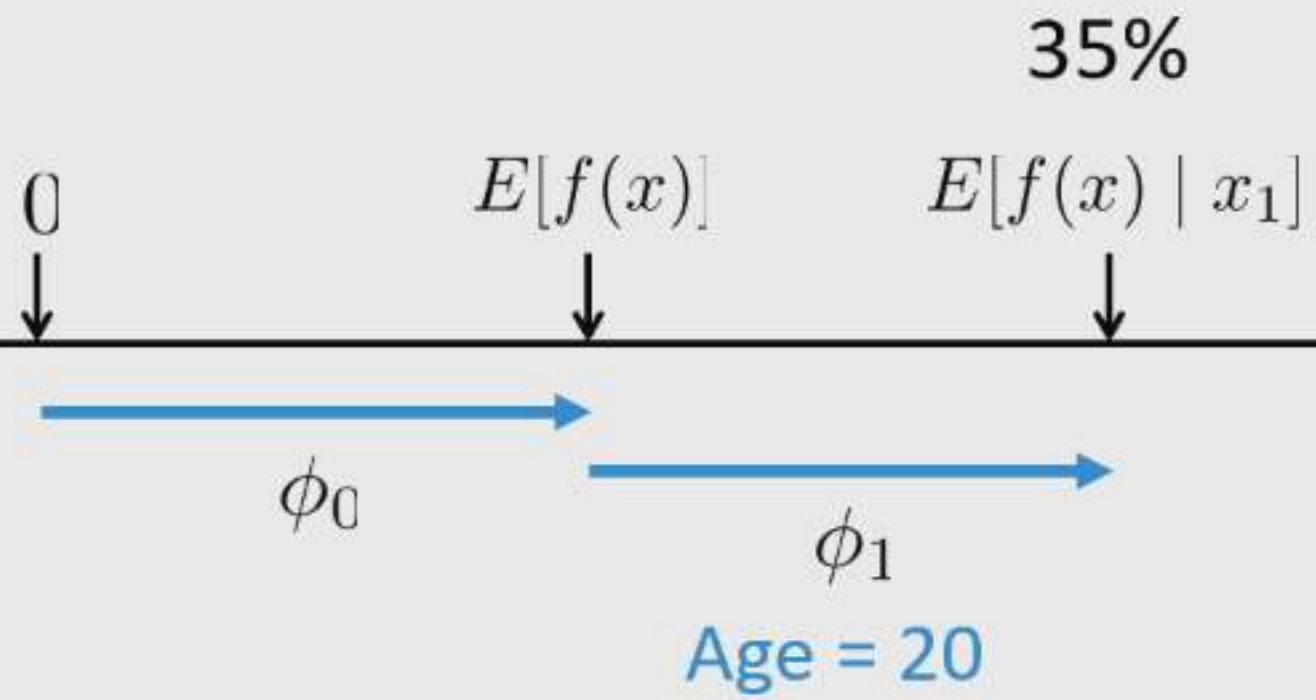
$E[f(x)]$

0



ϕ_0

Base rate





0

$E[f(x)]$

70%

$E[f(x) | x_1, x_2]$



ϕ_0



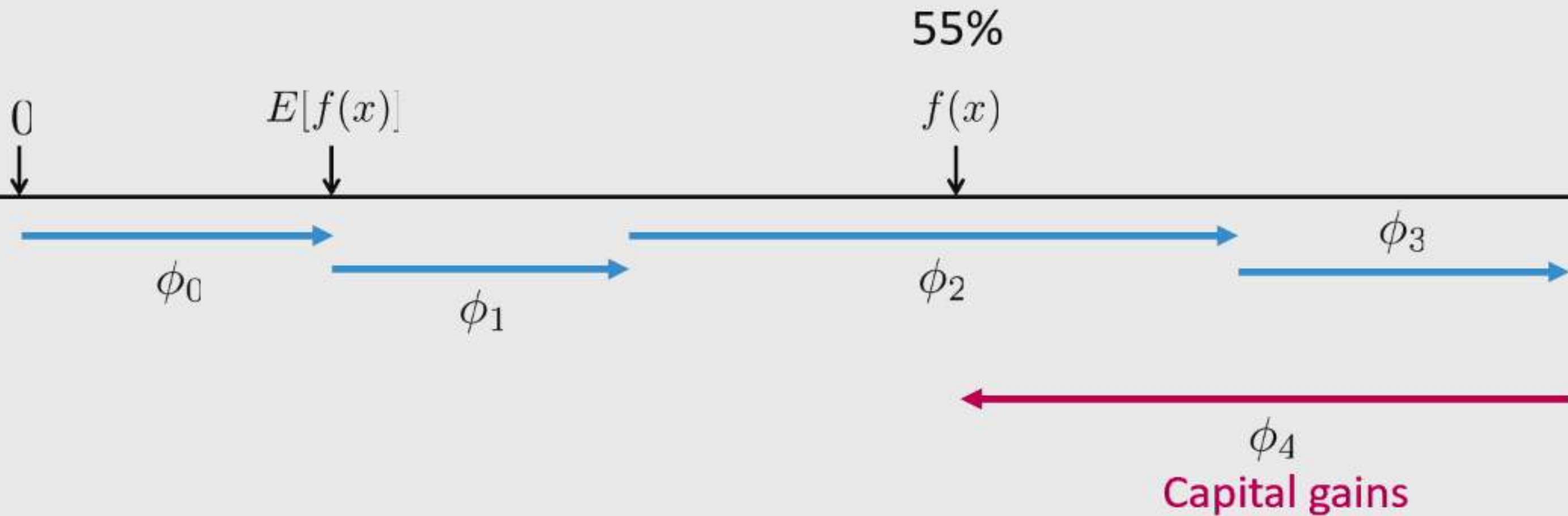
ϕ_1



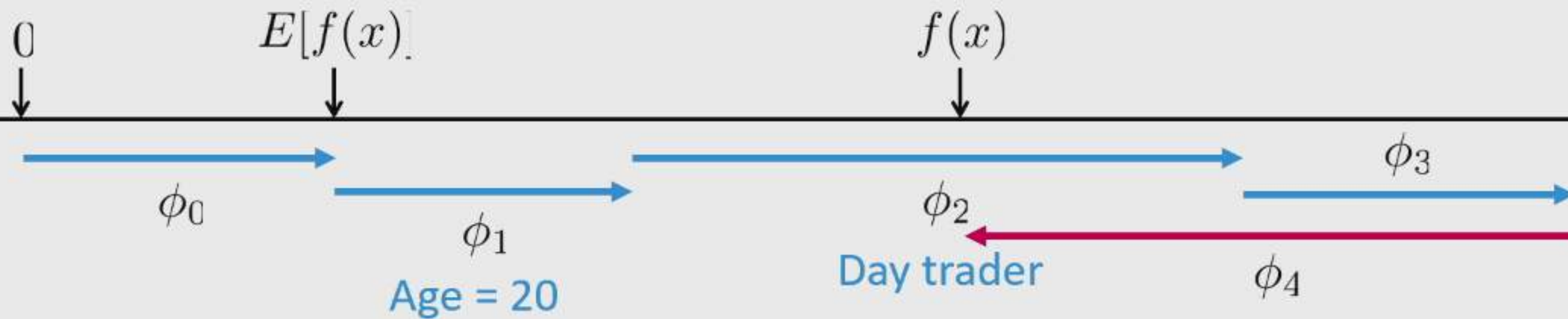
ϕ_2

Day trader

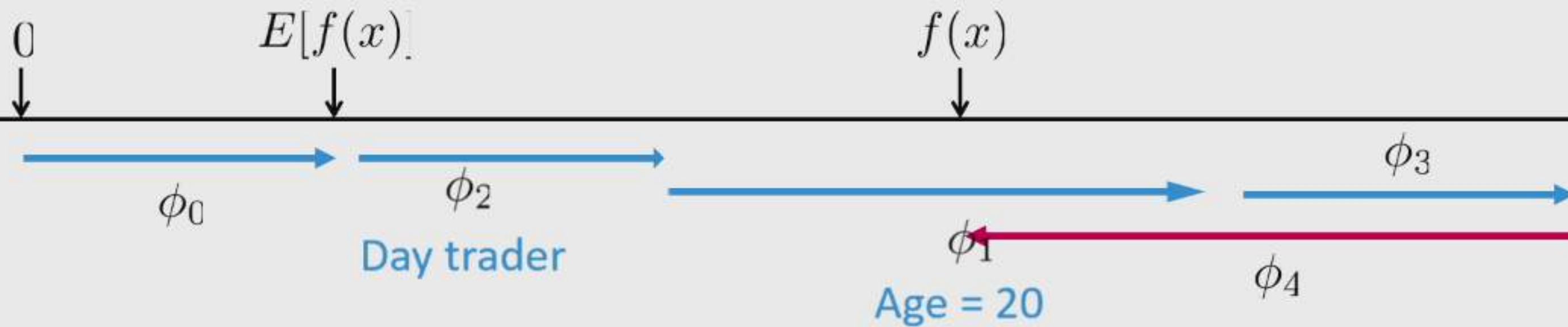


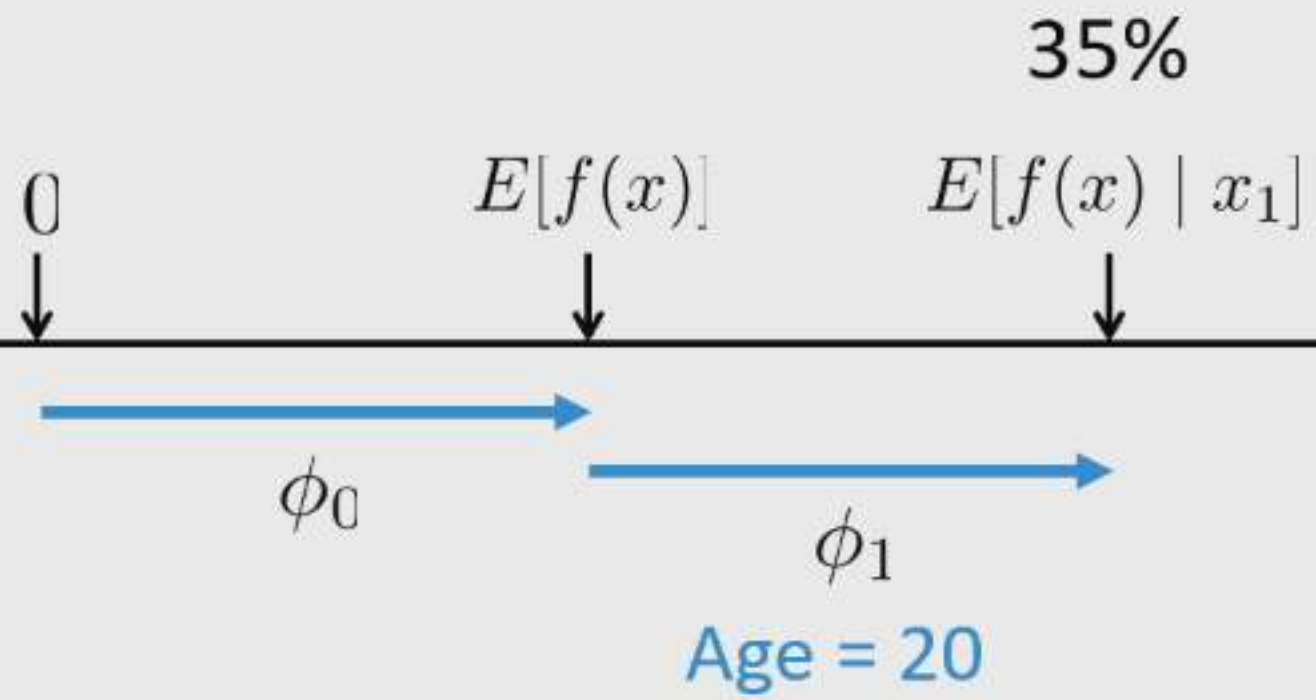


The order matters!



The order matters!







0

$E[f(x)]$

70%

$E[f(x) | x_1, x_2]$



ϕ_0



ϕ_1



ϕ_2

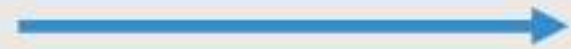
Day trader



20%

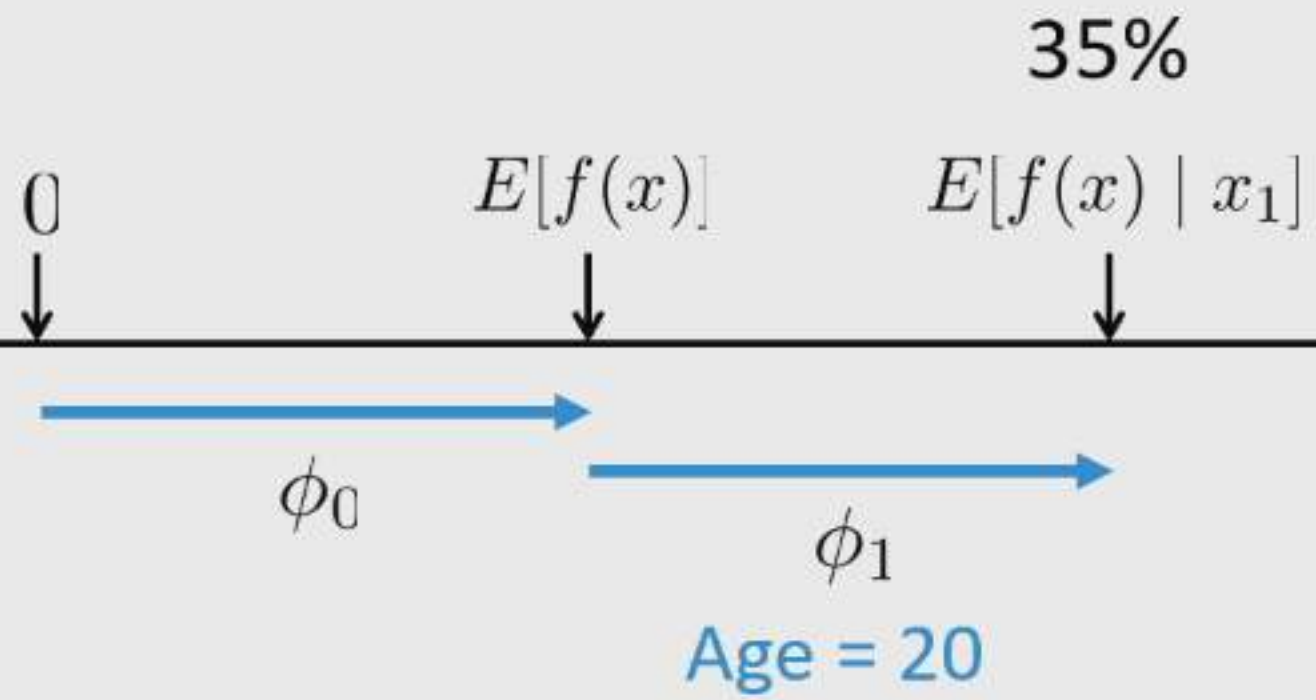
$E[f(x)]$

0



ϕ_0

Base rate





0

$E[f(x)]$

70%

$E[f(x) | x_1, x_2]$



ϕ_0



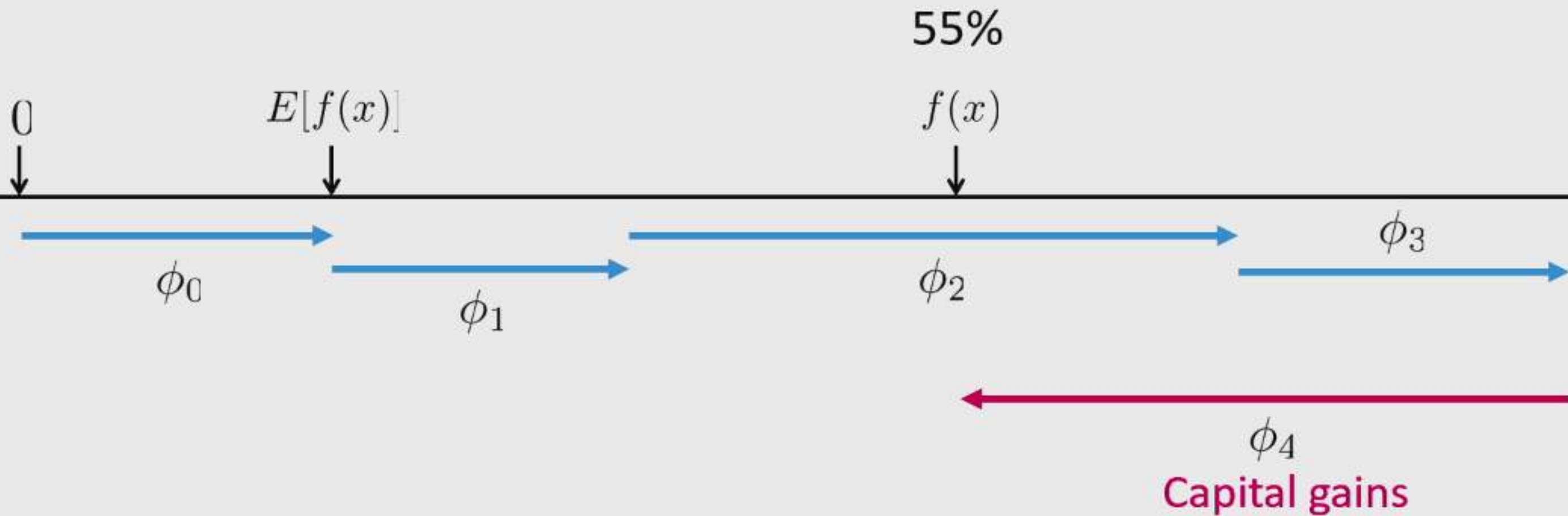
ϕ_1



ϕ_2

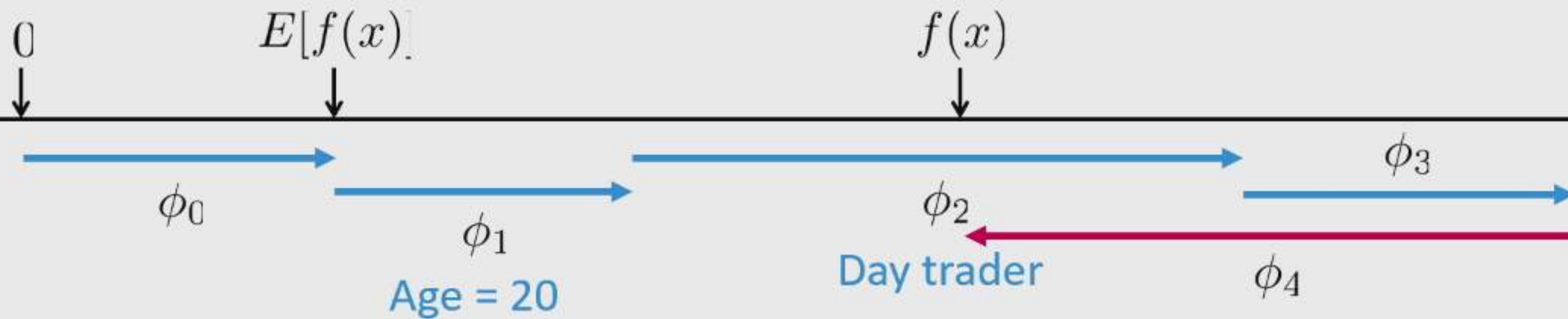
Day trader



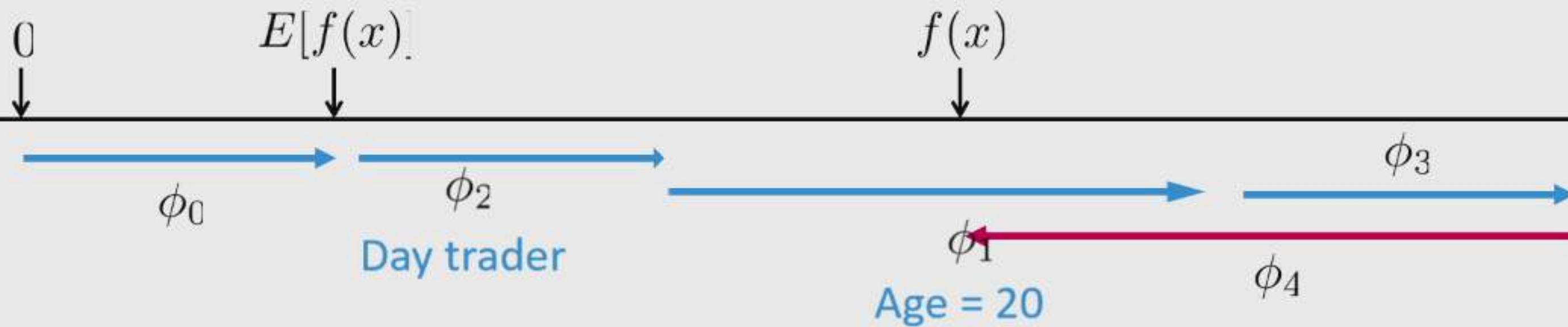




The order matters!



The order matters!

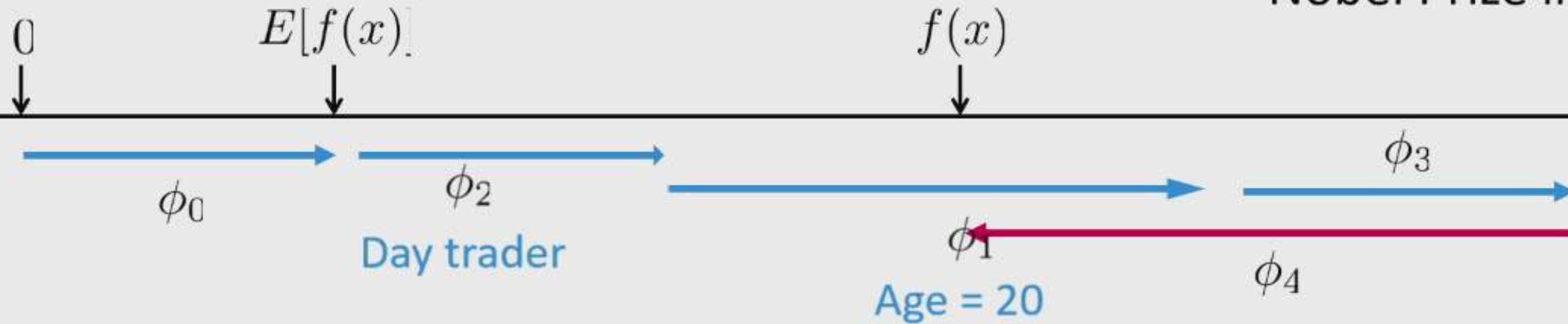


Lloyd Shapley



The order matters!

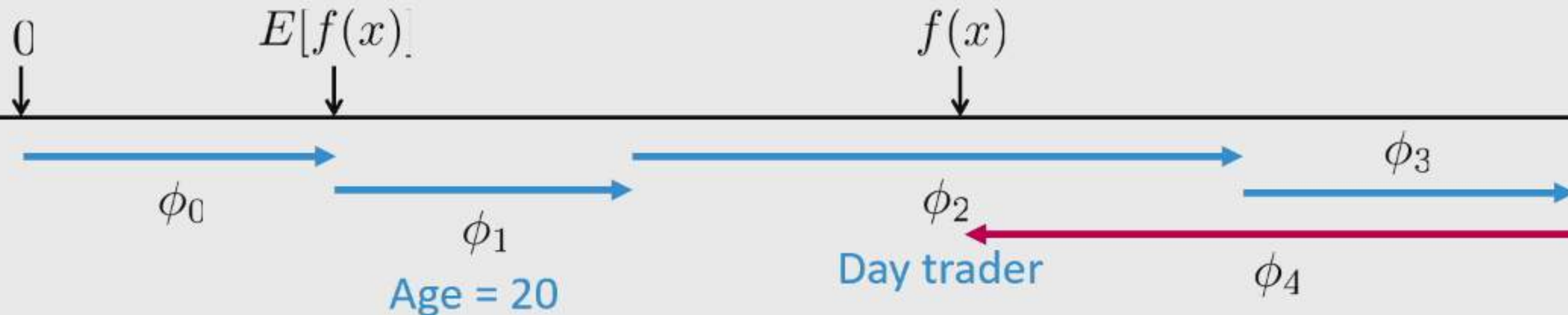
Nobel Prize in 2012



Shapley properties

1

Local accuracy (additivity) – The sum of the local feature attributions equals the difference between the base rate and the model output.

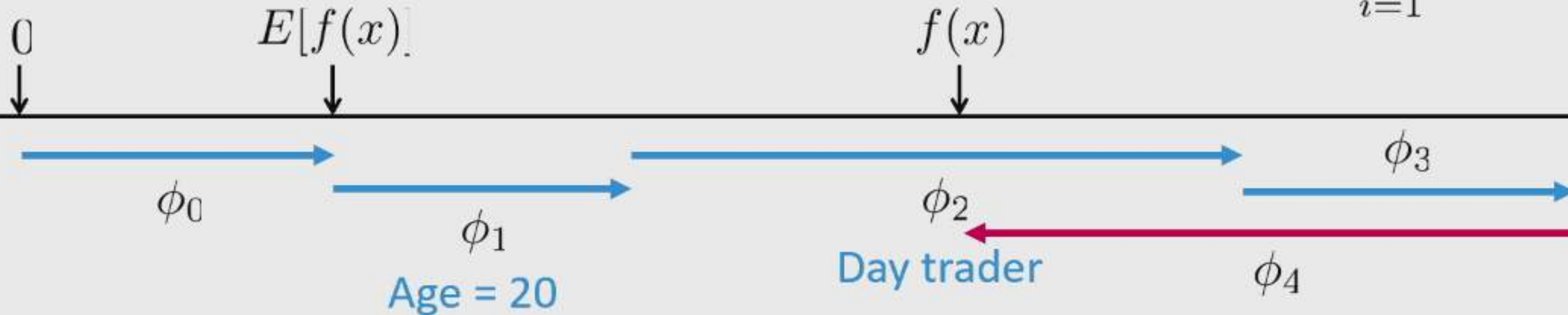


Shapley properties

1

Local accuracy (additivity) – The sum of the local feature attributions equals the difference between the base rate and the model output.

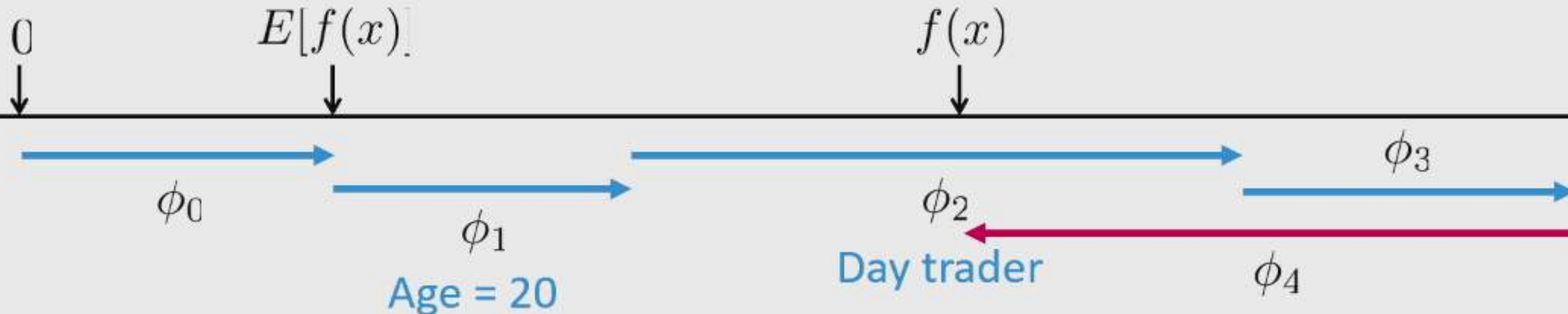
$$E[f(x)] + \sum_{i=1}^M \phi_i = f(x)$$



Shapley properties

2

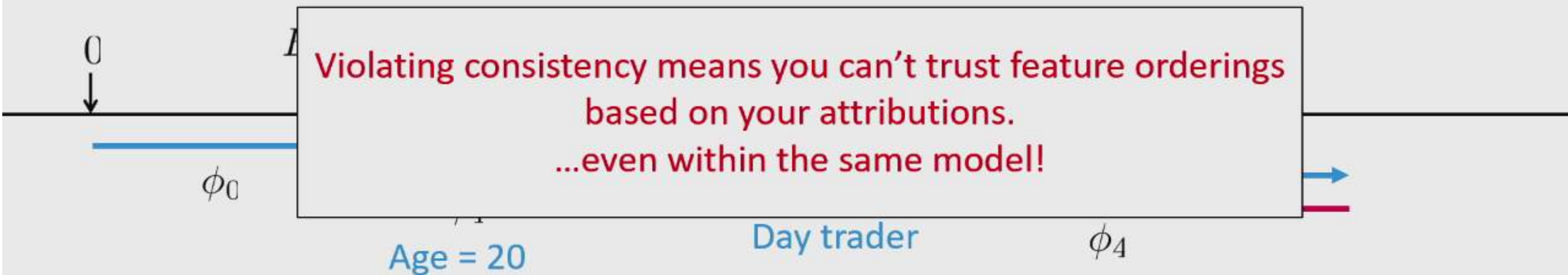
Consistency (monotonicity) – If you change the original model such that a feature has a larger impact in every possible ordering, then that input's attribution should not decrease.



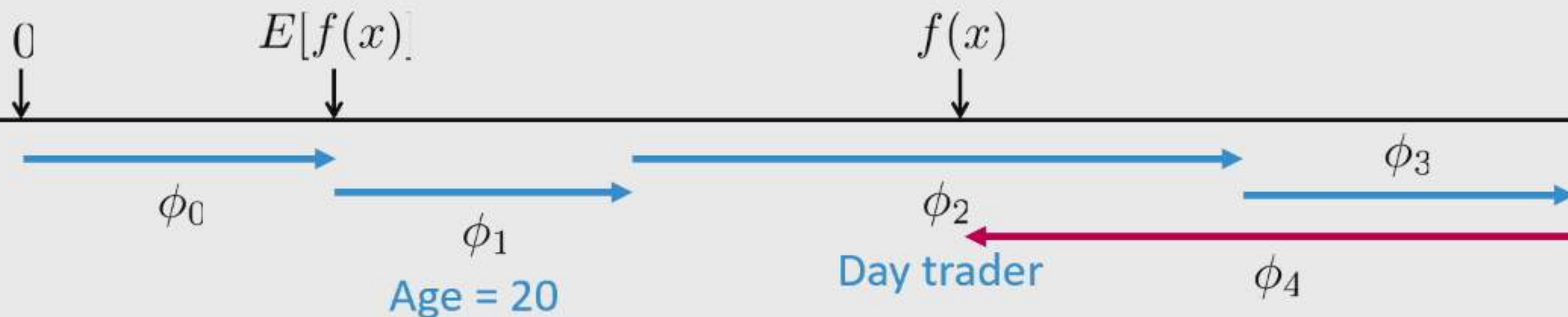
Shapley properties

2

Consistency (monotonicity) – If you change the original model such that a feature has a larger impact in every possible ordering, then that input's attribution should not decrease.

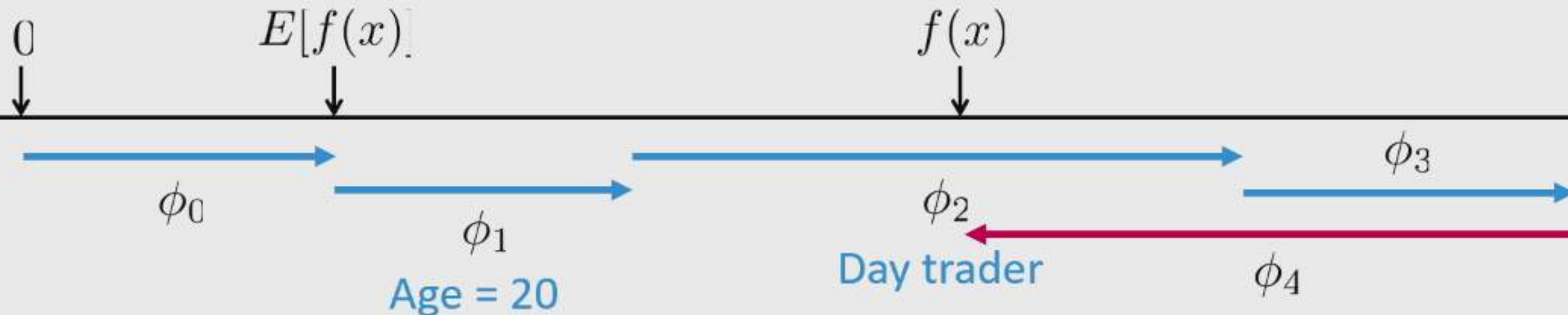


Shapley values result from **averaging over all $N!$ possible orderings.**



SHapley Additive exPlanation (SHAP) values

Shapley values result from **averaging over all $N!$ possible orderings.**



Options for NP-hard problems:

Options for NP-hard problems:

1. Prove that $P = NP$.

Options for NP-hard problems:

1. Prove that $P = NP$.
2. Find an approximate solution.

Options for NP-hard problems:

1. Prove that $P = NP$.

2. Find an approximate solution.

Options for NP-hard problems:

~~1. Prove that $P = NP$.~~

2. Find an approximate solution.

Options for NP-hard problems:

~~1. Prove that $P = NP$.~~

2. Find an approximate solution.

LIME

DeepLIFT

Shapley reg. values

SHAP

Relevance prop.

QII

Shapley sampling

Saabas



LIME

SHAP

LIME Objective

$$\xi = \arg \min_{g \in \mathcal{G}} L(f, g, \pi_{x'}) + \Omega(g)$$

LIME Objective

Loss function

Regularizer

$$\xi = \arg \min_{g \in \mathcal{G}} L(f, g, \pi_{x'}) + \Omega(g)$$

Local kernel

LIME Objective

Loss function

Regularizer

$$\xi = \arg \min_{g \in \mathcal{G}} L(f, g, \pi_{x'}) + \Omega(g)$$

Local kernel

The loss L , regularizer Ω , and local kernel $\pi_{x'}$ were all chosen heuristically...

L , Ω , and π_x , are forced under local accuracy and consistency !

L , Ω , and π_x , are forced under local accuracy and consistency !

$$L(f, g, \pi_{x^0}) = \int_{z^0}^z \left(f(h_x^{-1}(z^0)) - g(z^0) \right)^2 \pi_{x^0}(z^0)$$

$$\mathbb{E}(g) = 0$$

$$\pi_{x^0}(z^0) = \frac{(M-1)}{\binom{M-1}{|z^0|} |z^0| (M-|z^0|)}$$

L , Ω , and π_x , are forced under local accuracy and consistency !

This means we can now estimate the Shapley values using linear regression!

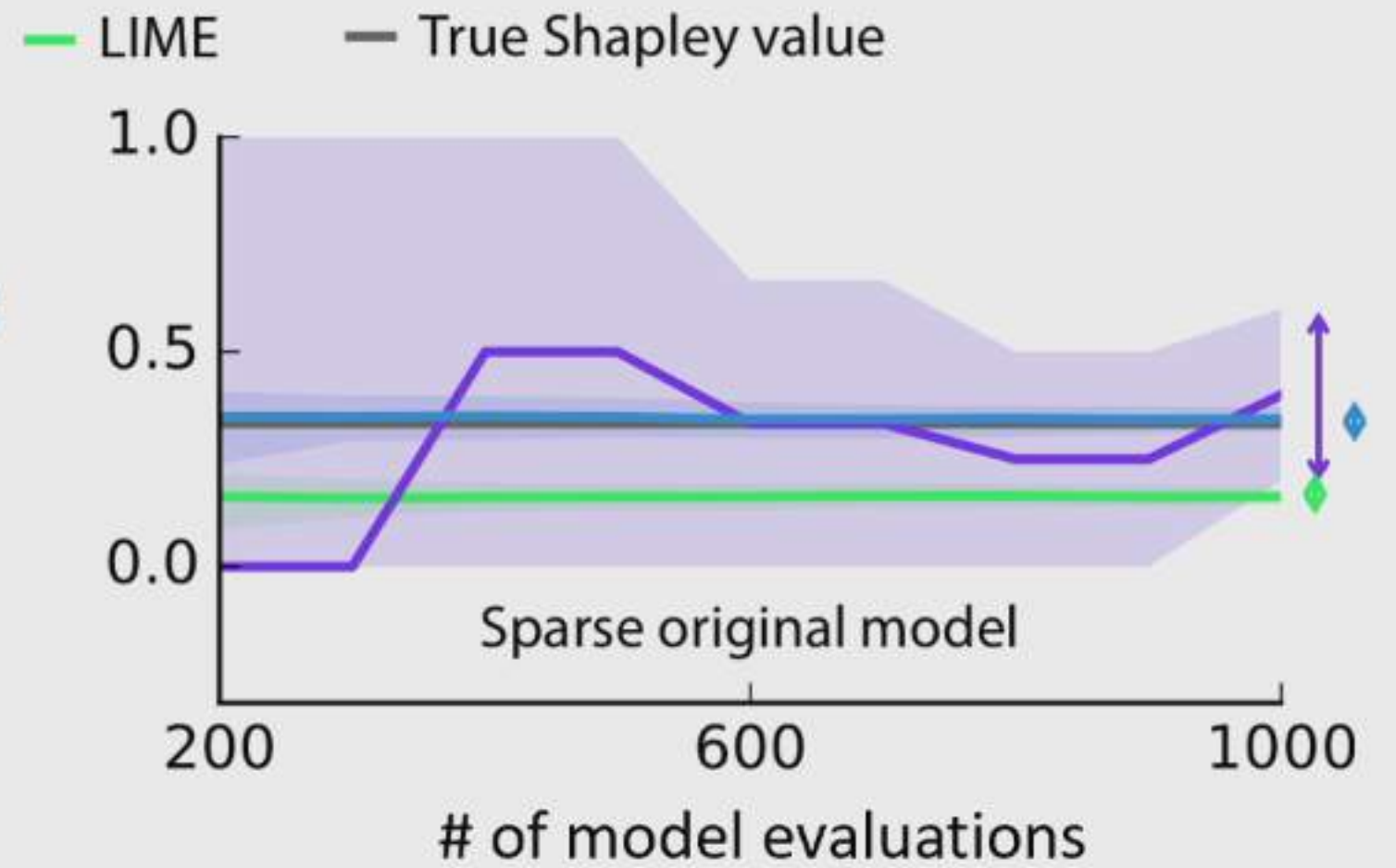
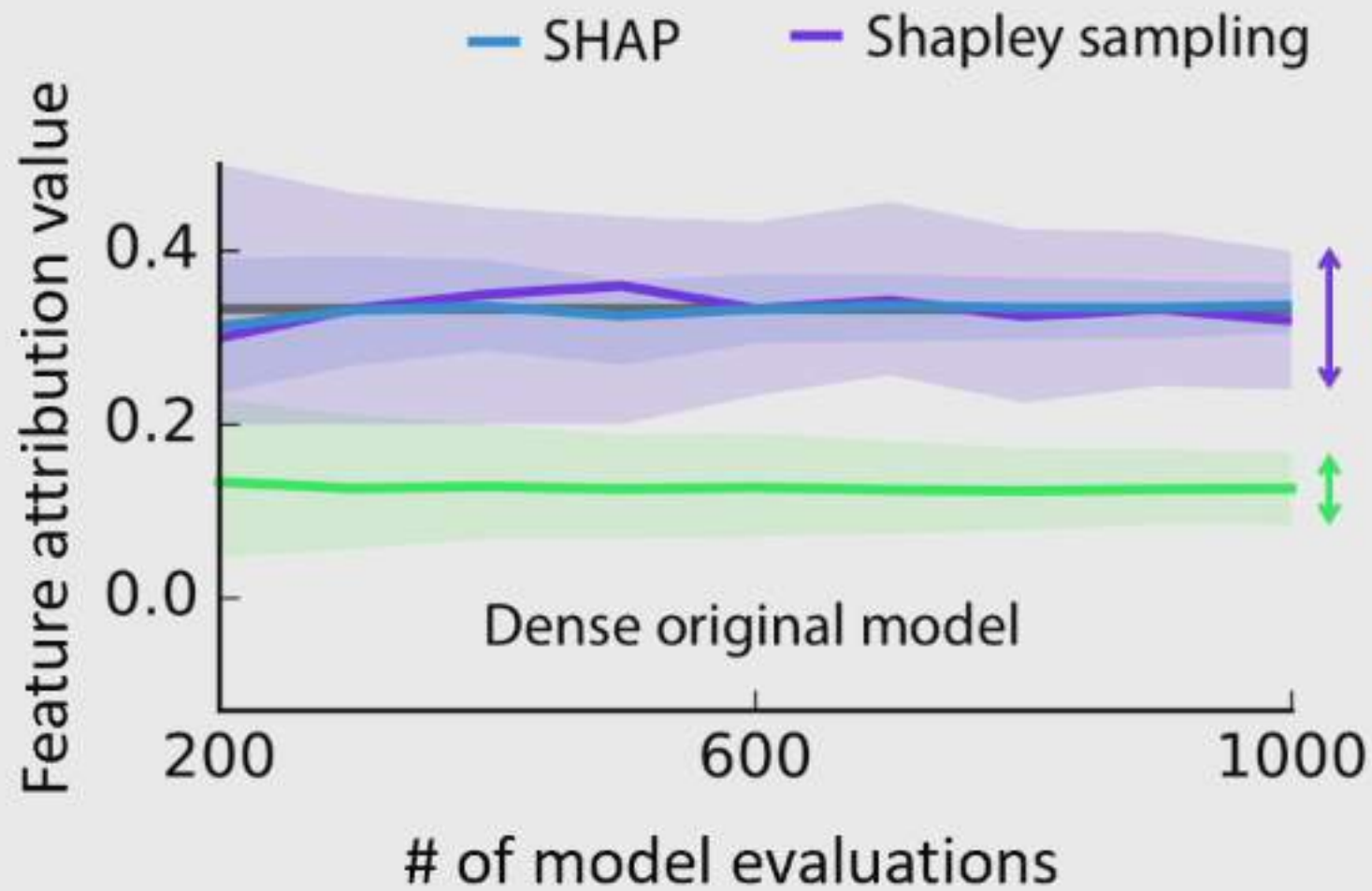
L , Ω , and π_x , are forced under local accuracy and consistency !

This means we can now estimate the Shapley values using linear regression!

(a fundamentally new way to estimate these classic values)

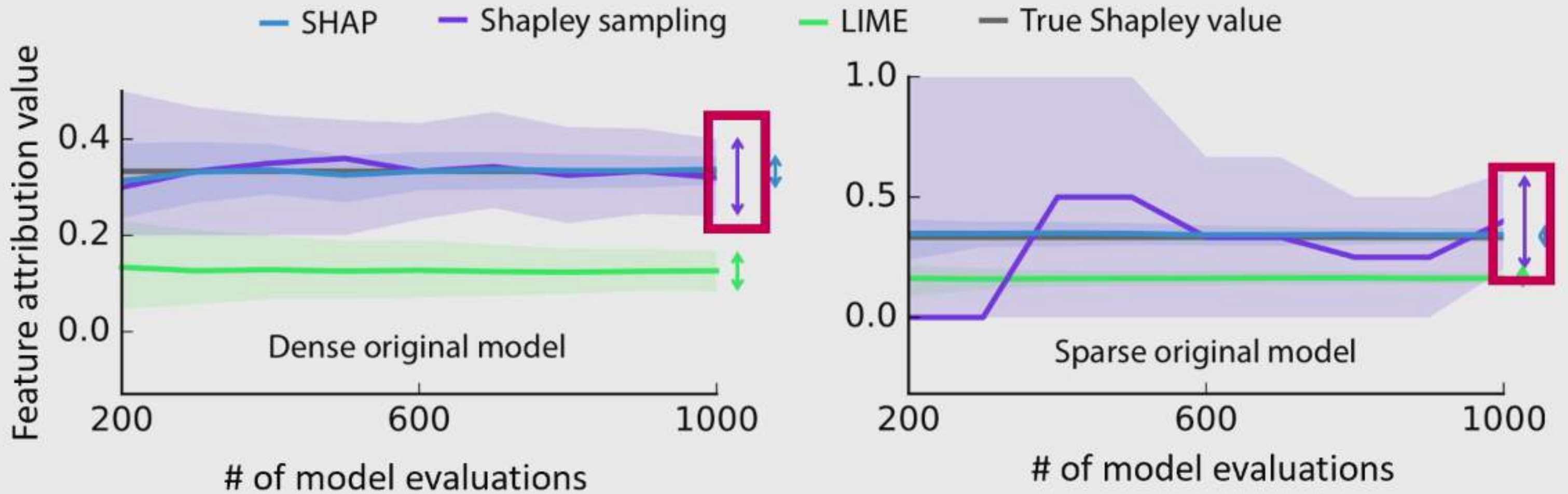


Faster estimation than classic Shapley methods





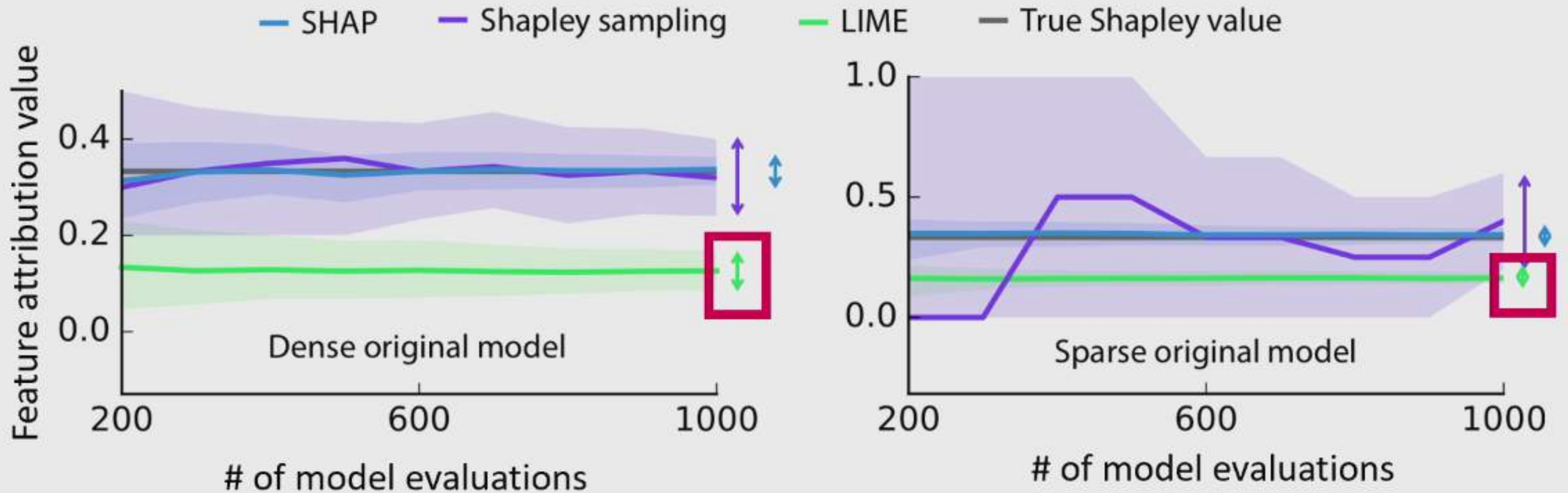
Faster estimation than classic Shapley methods



Permutation sampling has high variance



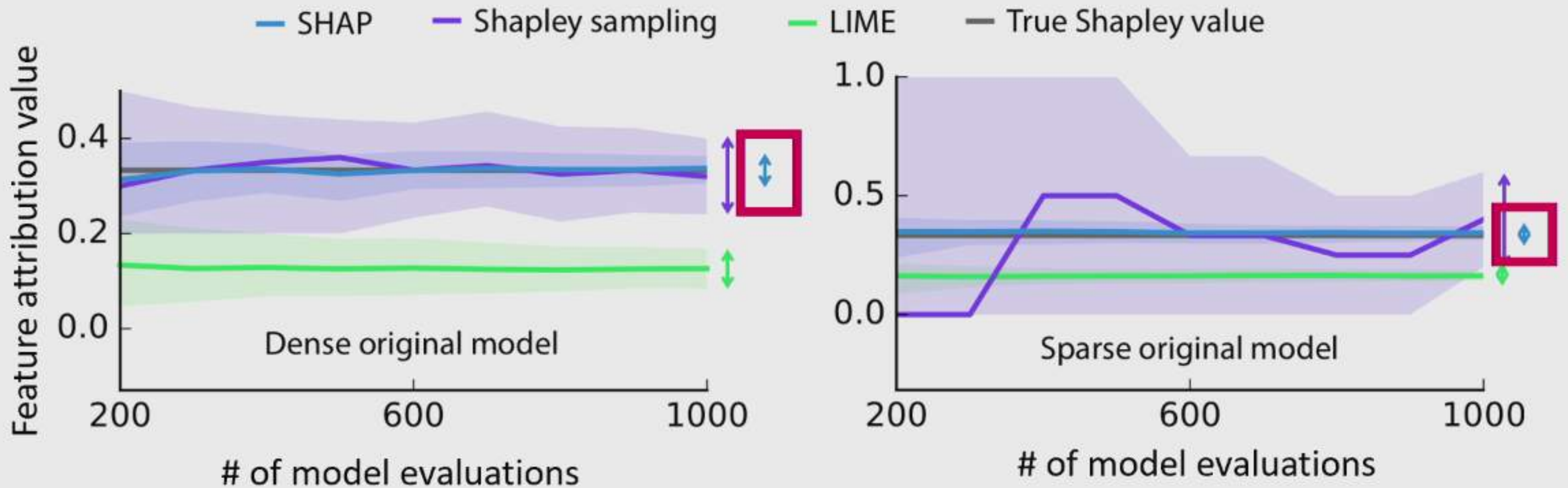
Faster estimation than classic Shapley methods



LIME has lower variance but does not converge to the Shapley values



Faster estimation than classic Shapley methods



SHAP retains the best of both (low variance and axiomatic agreement)

Explainable AI for Science and Medicine

Theory

Practice

Application

Explainable AI for Science and Medicine

Theory



Unification of explanation
methods

Practice

Application

Explainable AI for Science and Medicine

Theory



Unification of explanation
methods



Strong uniqueness results

Practice

Application

Explainable AI for Science and Medicine

Theory



Unification of explanation
methods



Strong uniqueness results

Practice



New estimation methods for
the classic Shapley values

Application

Explainable AI for Science and Medicine

Theory



Unification of explanation methods



Strong uniqueness results

Practice



New estimation methods for the classic Shapley values

Application



Anesthesia safety

Improving anesthesia safety through ML

Improving anesthesia safety through ML



The first public demonstration of Ether in 1846

The operating room is a data-rich environment

- High frequency measurements from many sensors

The operating room is a data-rich environment

- High frequency measurements from many sensors
- Predicting adverse events allows proactive intervention.

The operating room is a data-rich environment

- High frequency measurements from many sensors
- Predicting adverse events allows proactive intervention.
- Hypoxemia (low blood oxygen)

The operating room is a data-rich environment

- High frequency measurements from many sensors
- Predicting adverse events allows proactive intervention.
- Hypoxemia (low blood oxygen)
- Prescience predicts hypoxemia within the next 5 minutes.

Prescience predicts hypoxemia and explains why



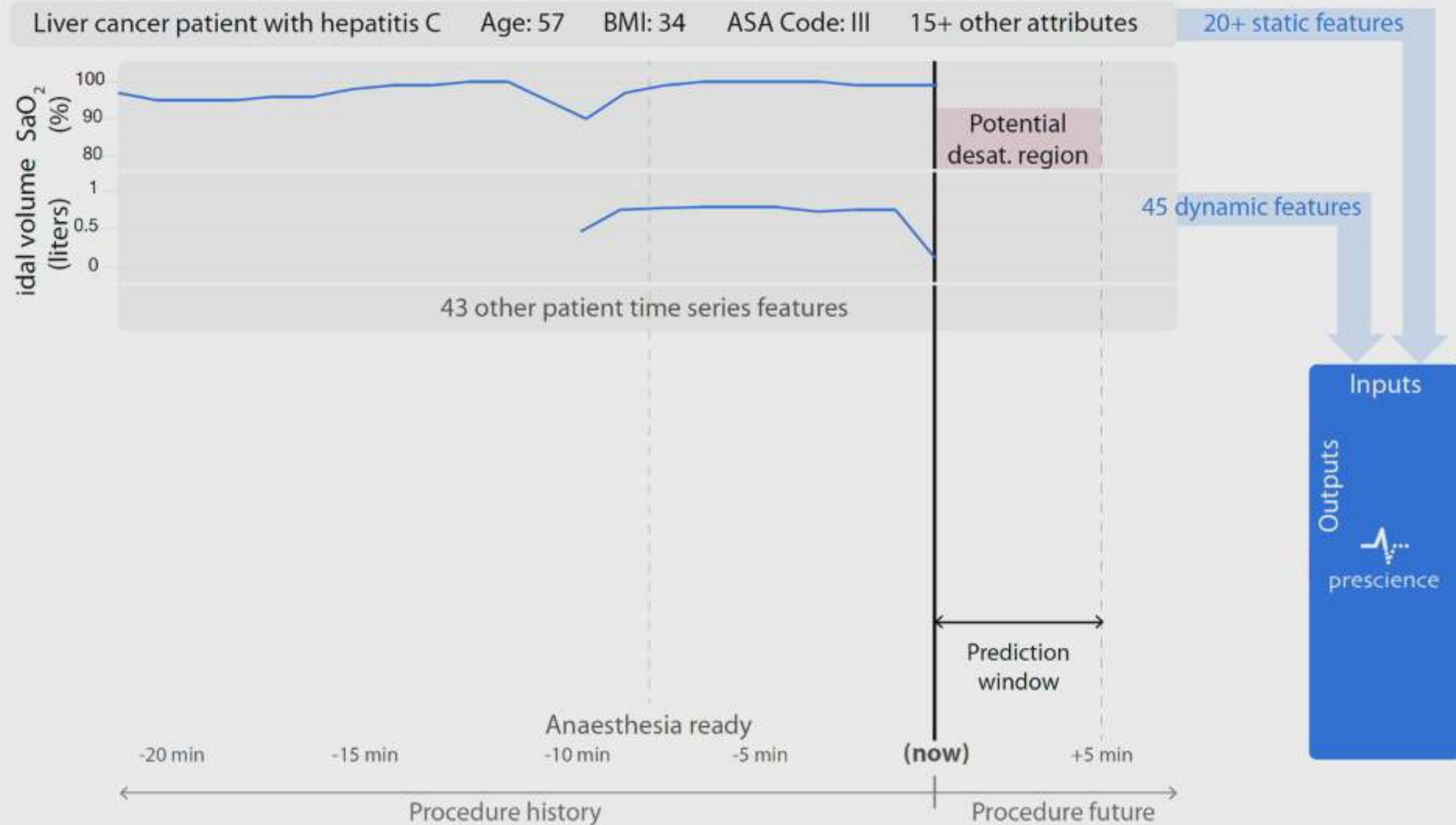
Prescience predicts hypoxemia and explains why

Liver cancer patient with hepatitis C Age: 57 BMI: 34 ASA Code: III 15+ other attributes

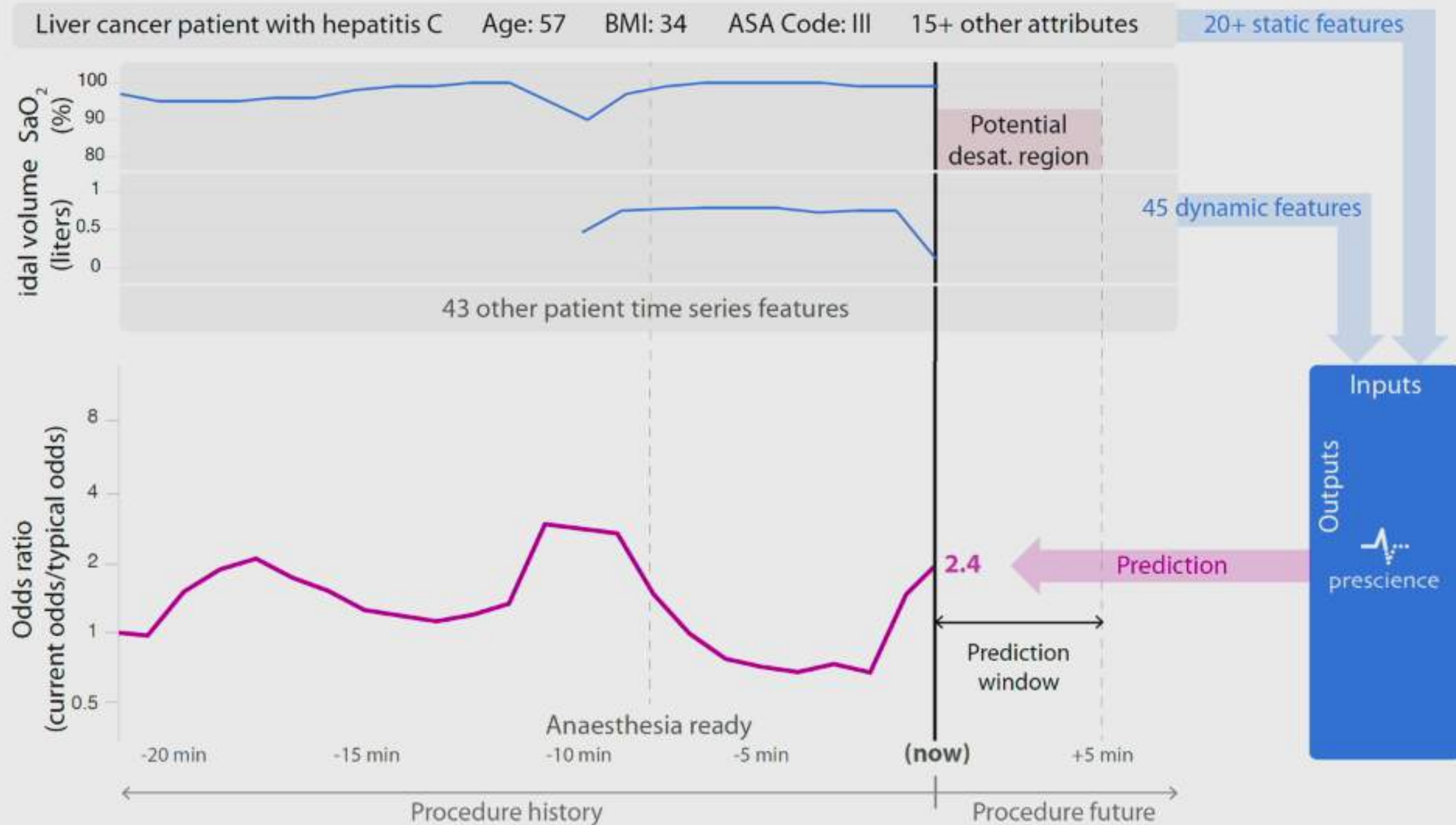
20+ static features



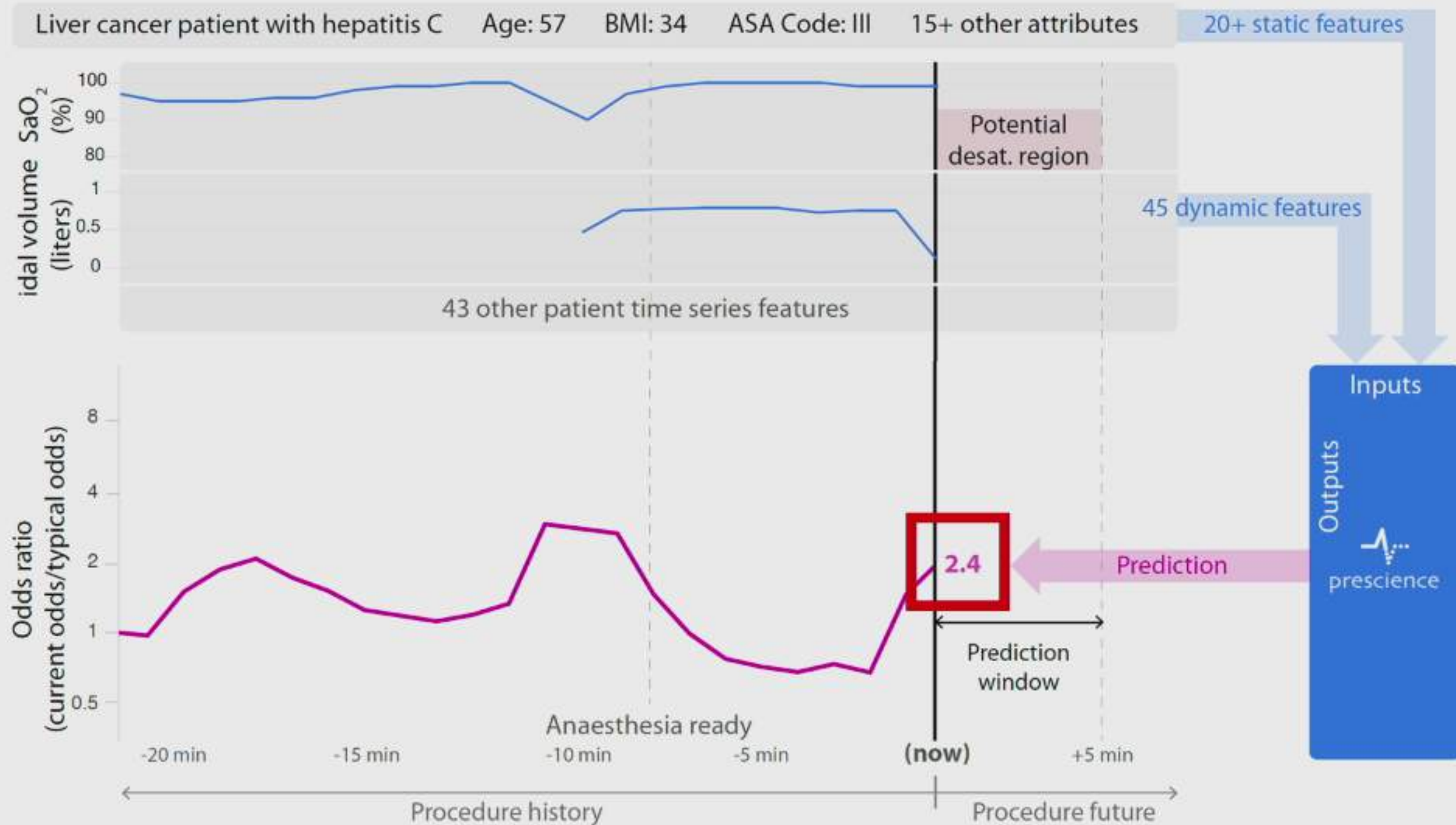
Prescience predicts hypoxemia and explains why



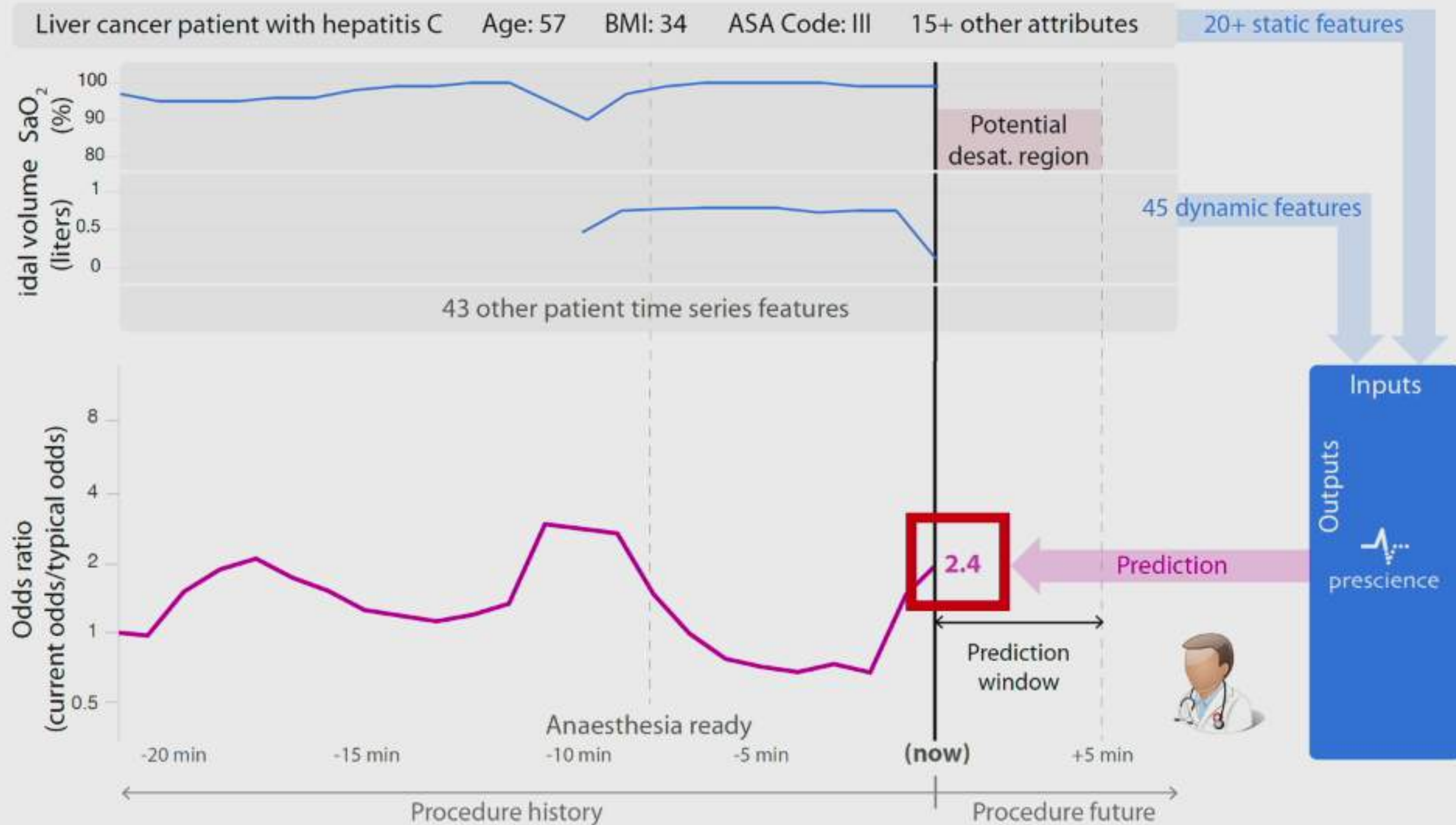
Prescience predicts hypoxemia and explains why



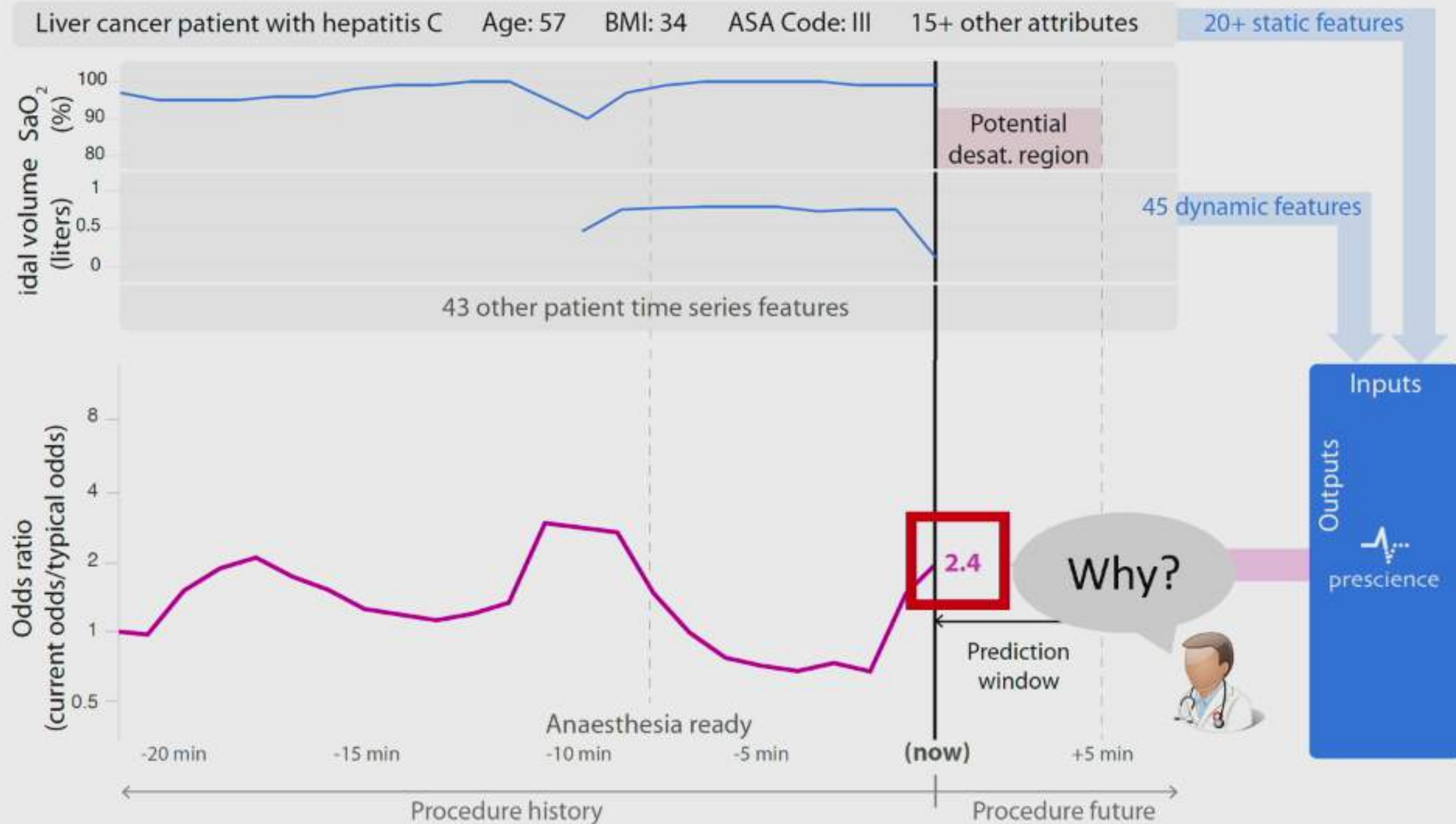
Prescience predicts hypoxemia and explains why



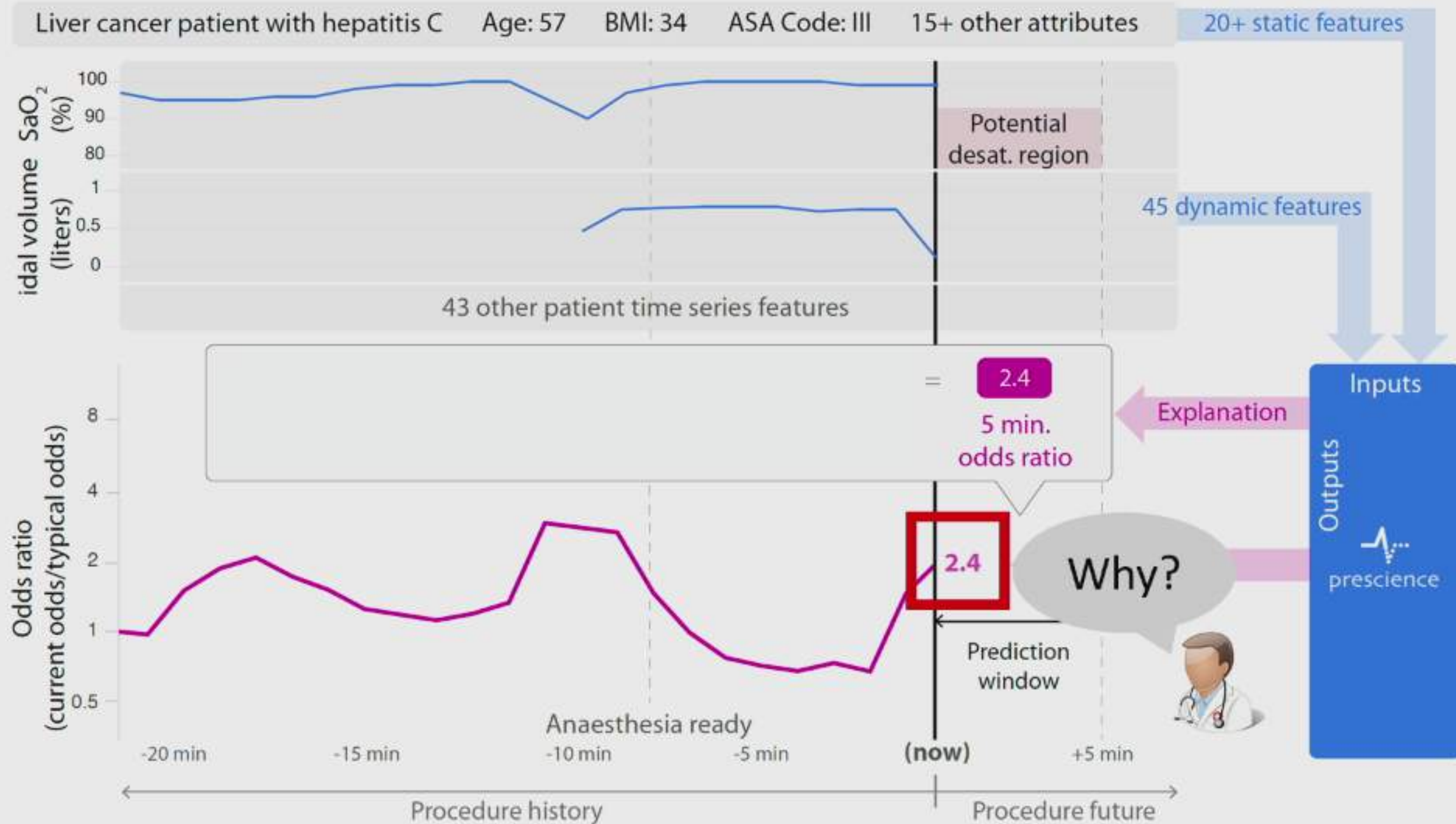
Prescience predicts hypoxemia and explains why



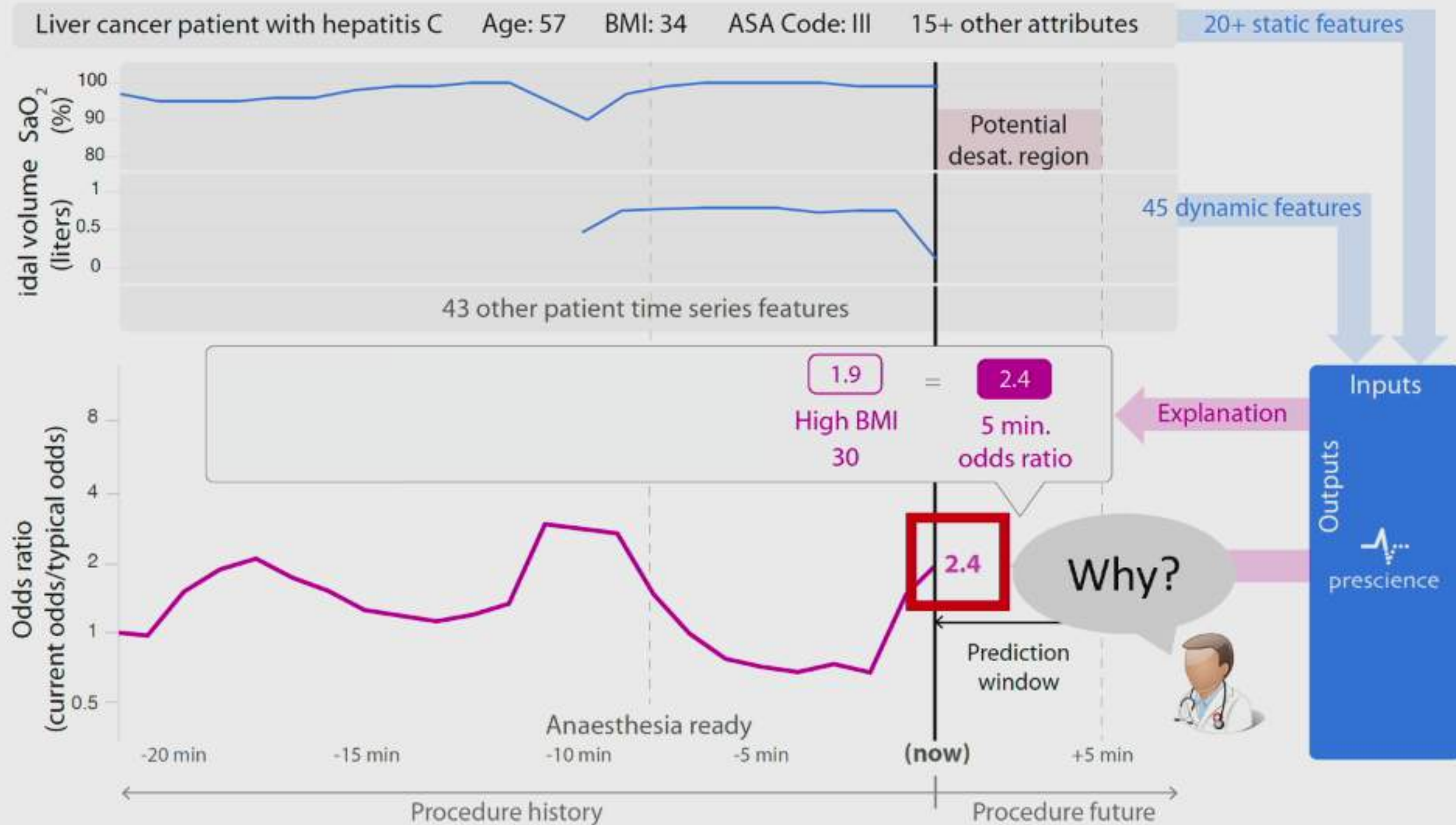
Prescience predicts hypoxemia and explains why



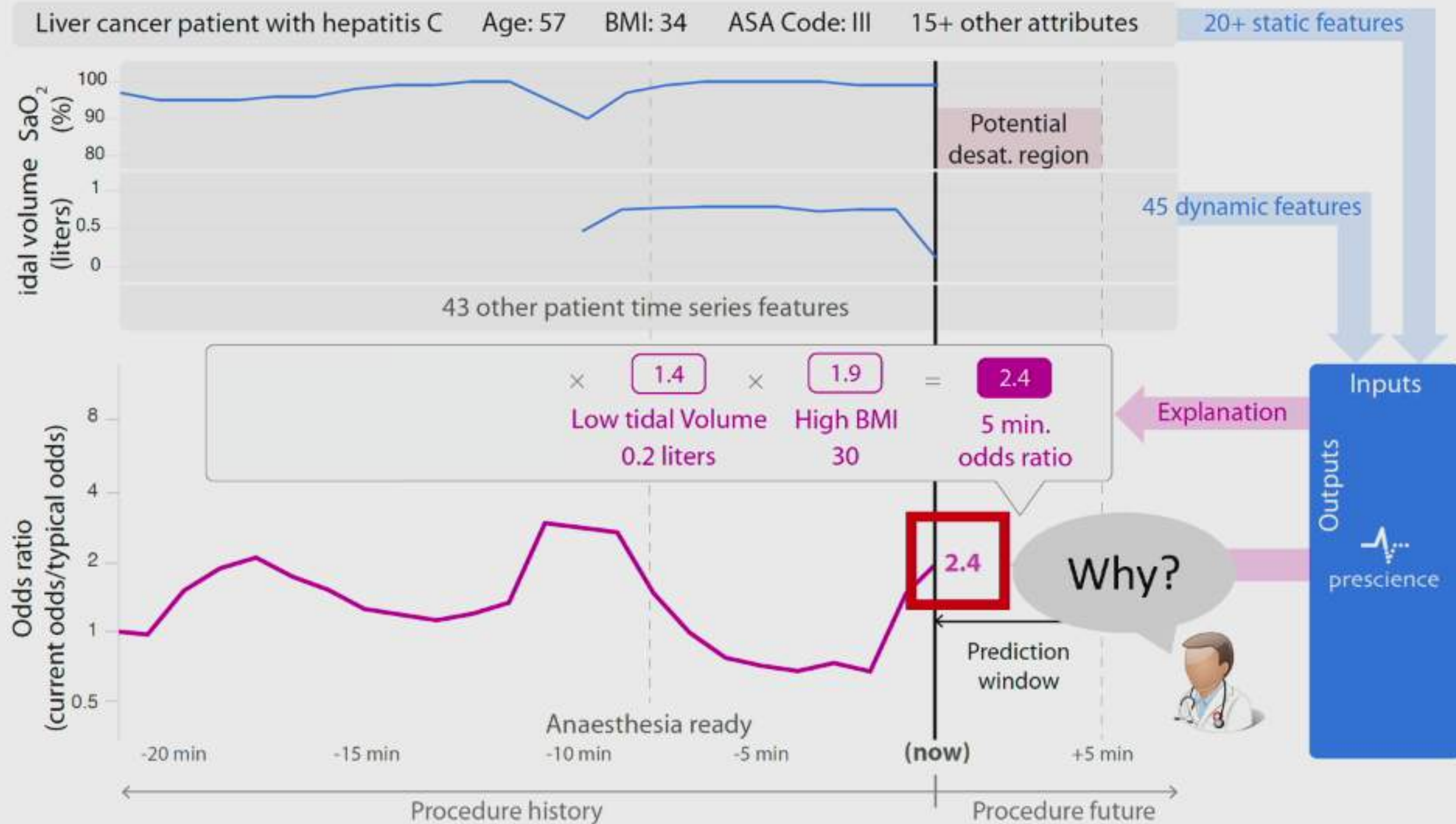
Prescience predicts hypoxemia and explains why



Prescience predicts hypoxemia and explains why

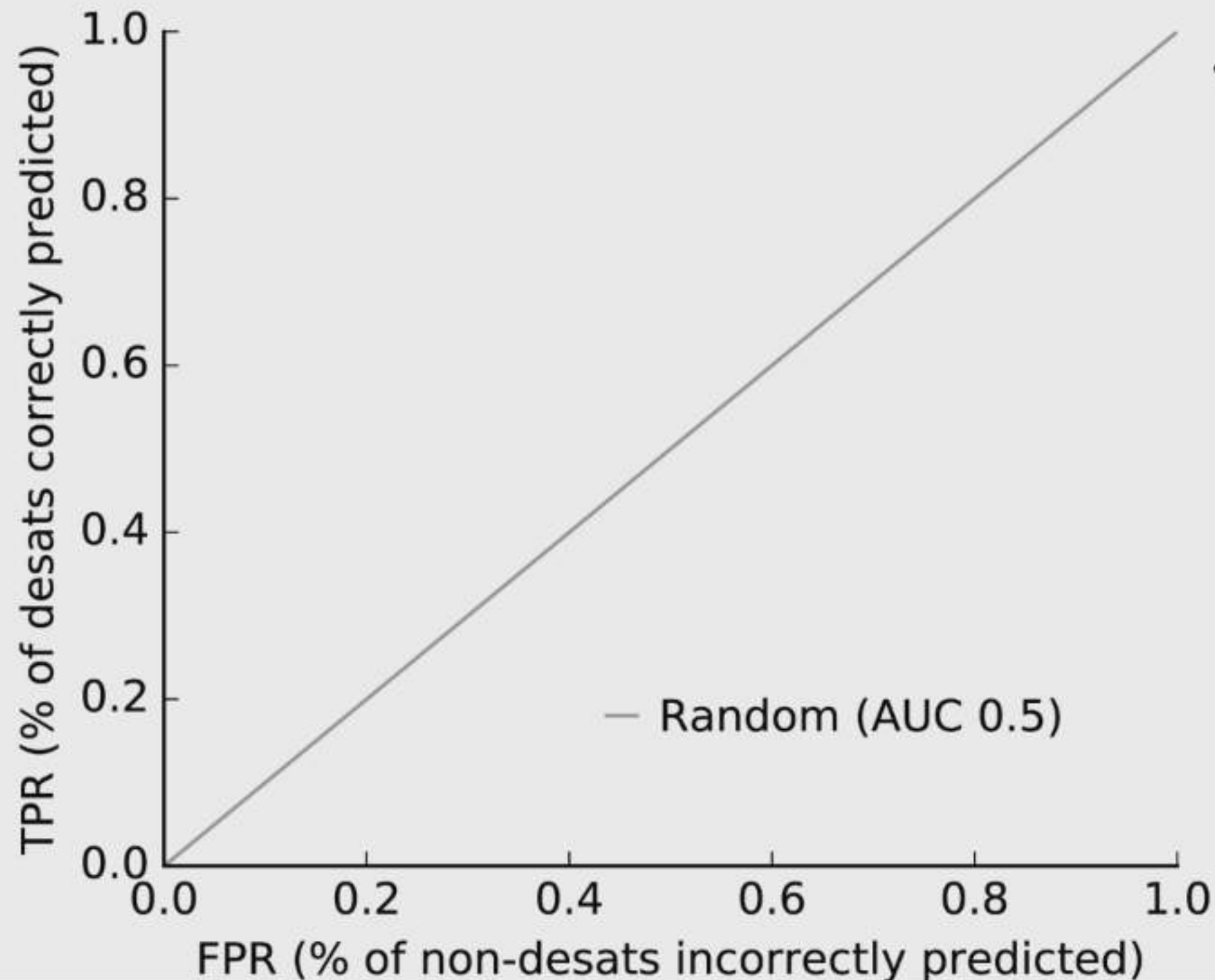


Prescience predicts hypoxemia and explains why



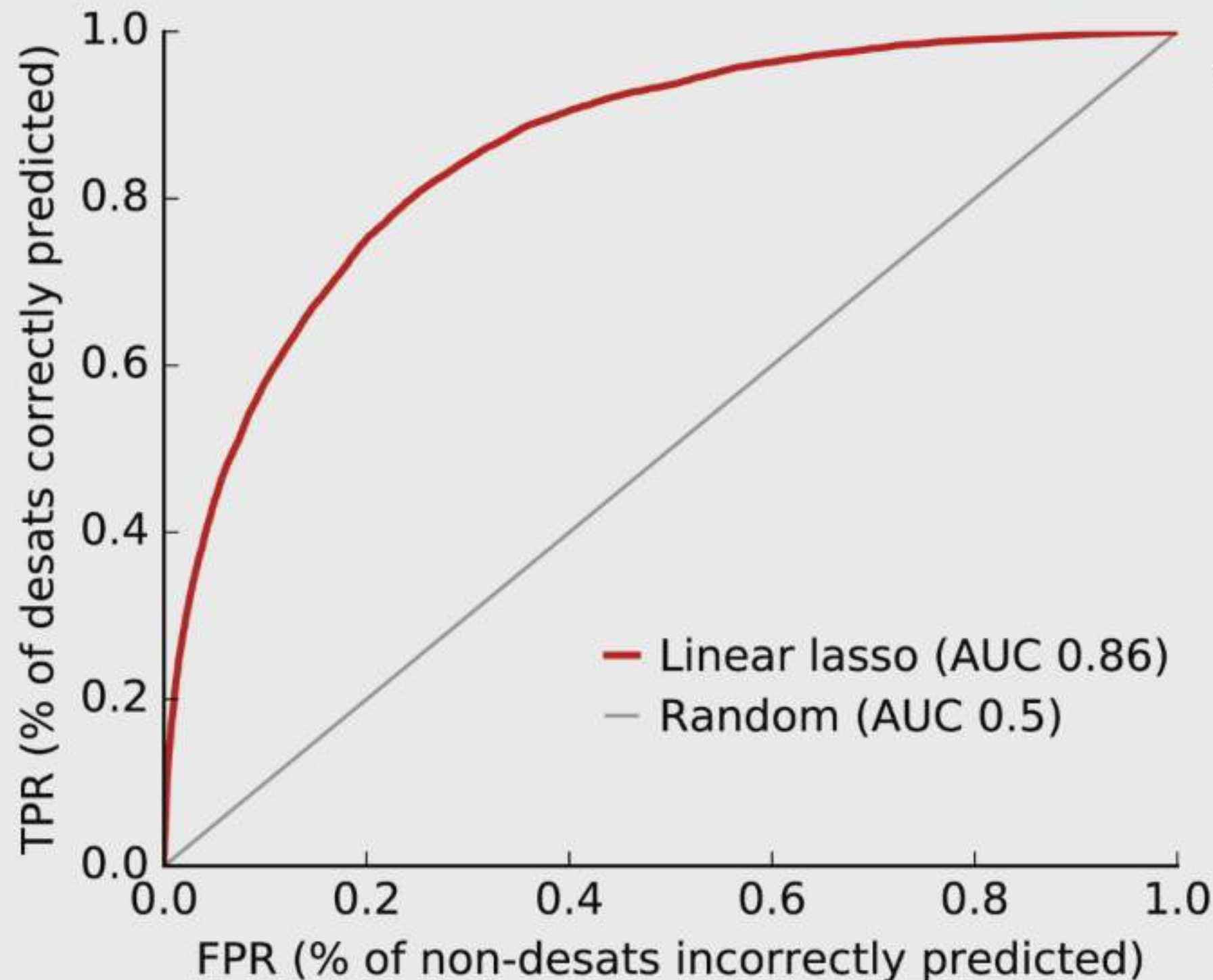
An interpretability vs. accuracy tradeoff

An interpretability vs. accuracy tradeoff



- Receiver operating characteristic (ROC) curves on a held out test set.

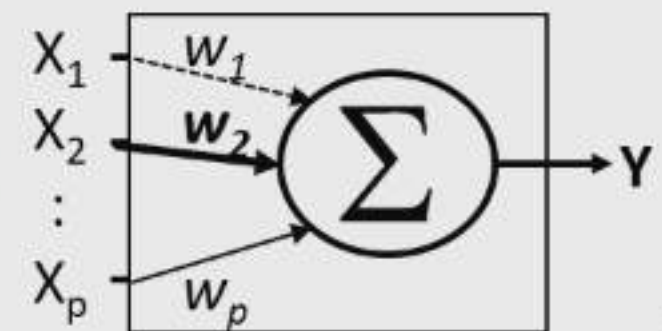
An interpretability vs. accuracy tradeoff



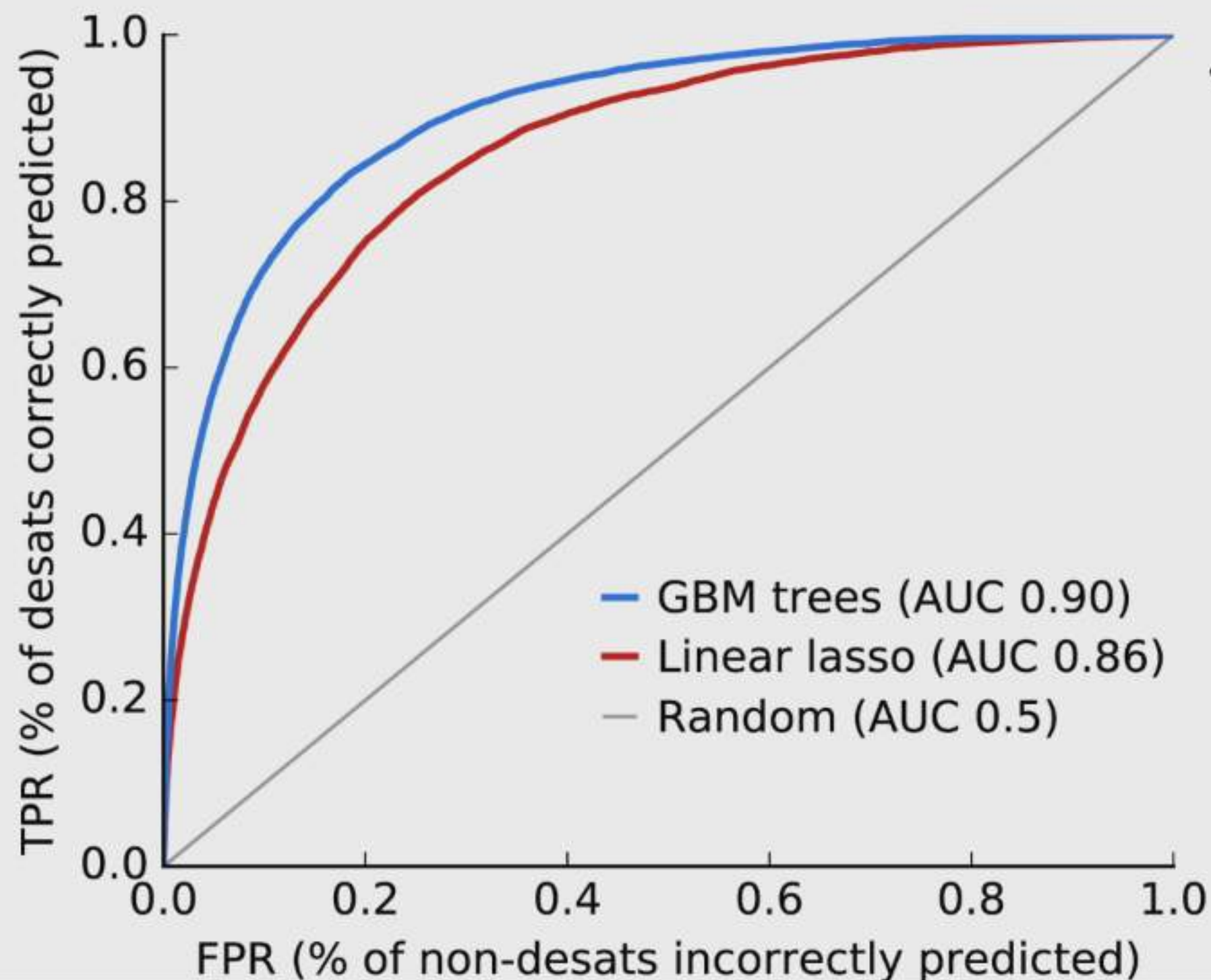
- Receiver operating characteristic (ROC) curves on a held out test set.

**Generalized
linear model**

X: Features Y: Outcome



An interpretability vs. accuracy tradeoff

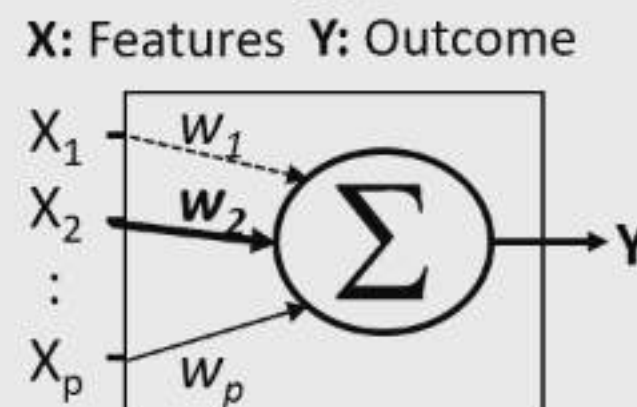


- Receiver operating characteristic (ROC) curves on a held out test set.

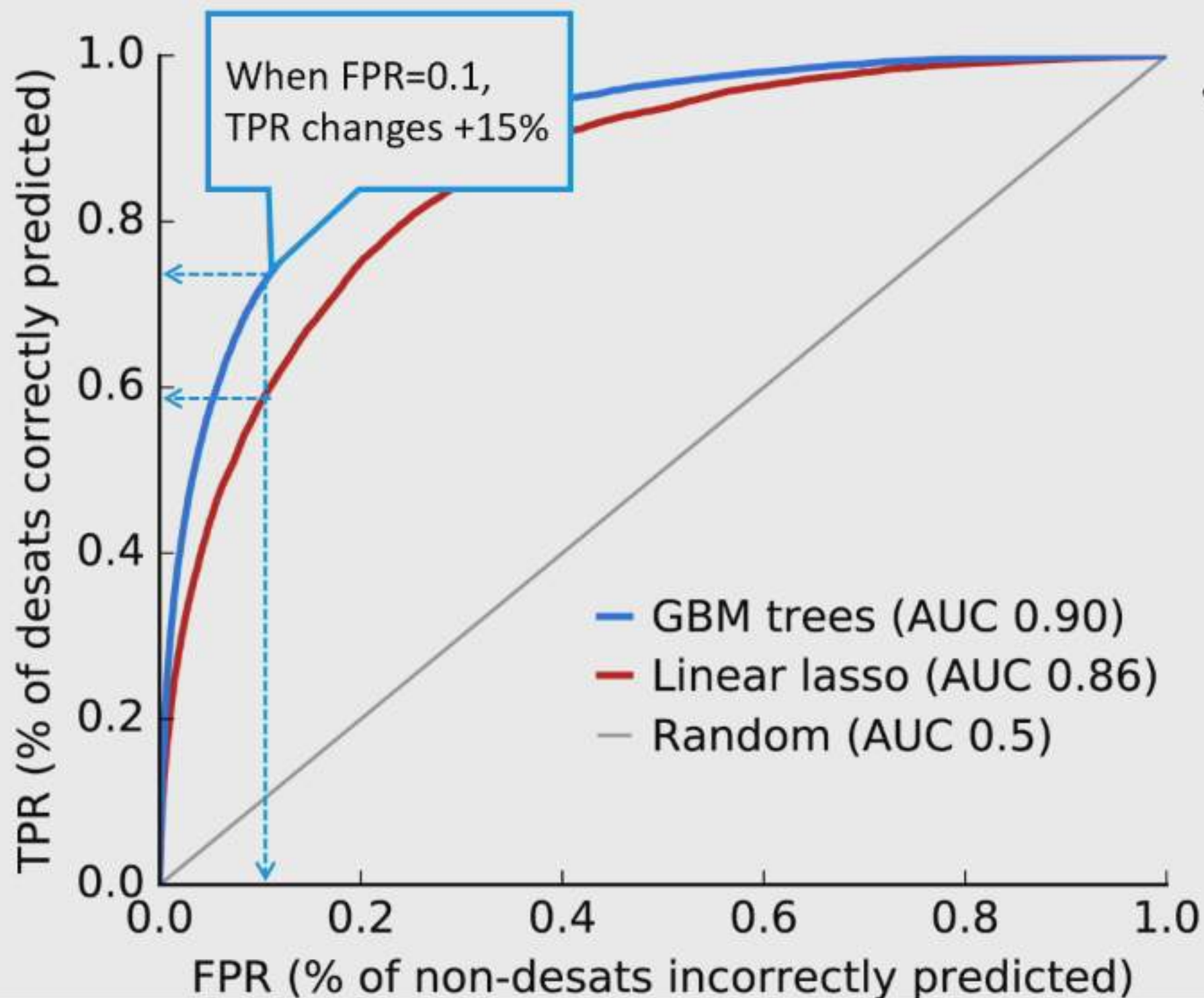
Complex model $f(\cdot)$



Generalized linear model



An interpretability vs. accuracy tradeoff

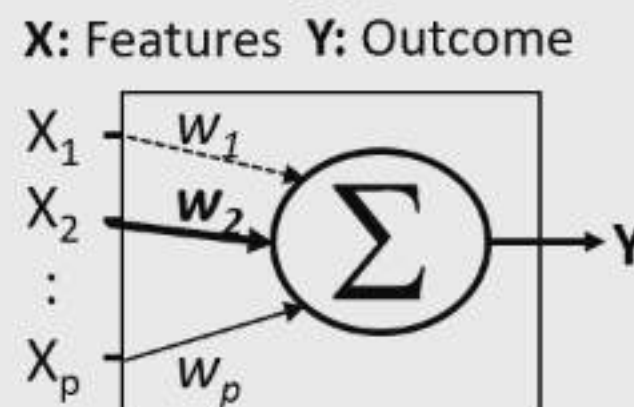


- Receiver operating characteristic (ROC) curves on a held out test set.

Complex model $f(\cdot)$



Generalized linear model



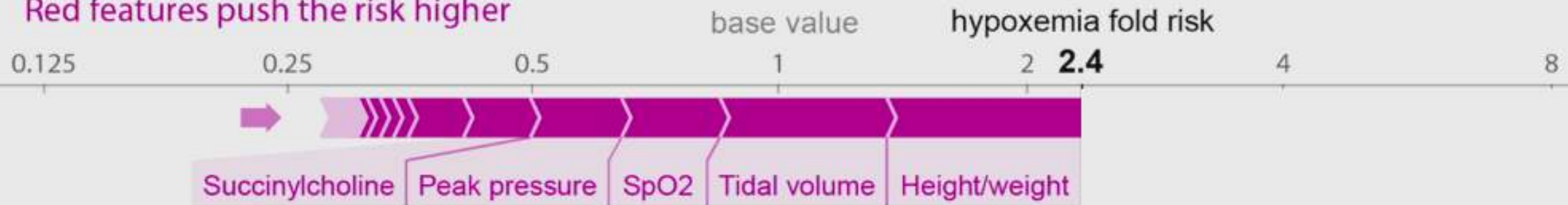
Using SHAP values in the operating room

Using SHAP values in the operating room



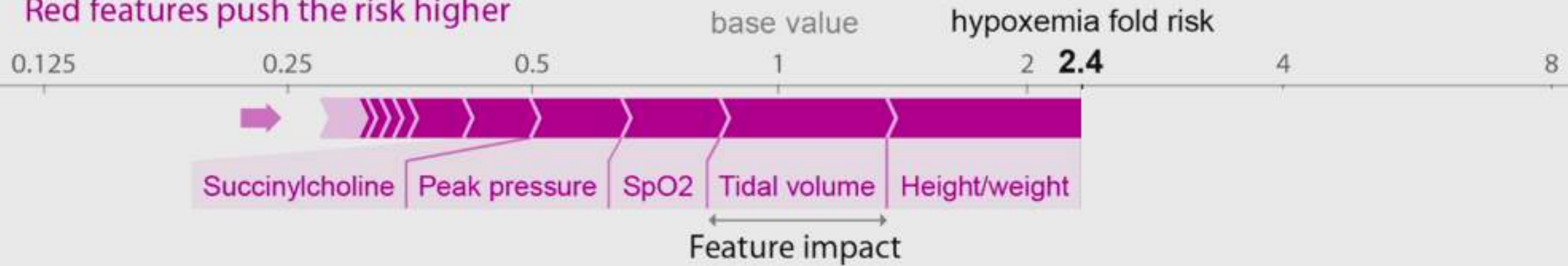
Using SHAP values in the operating room

Red features push the risk higher



Using SHAP values in the operating room

Red features push the risk higher



Using SHAP values in the operating room

Red features push the risk higher

base value

hypoxemia fold risk

Green features push the risk lower

0.125

0.25

0.5

1

2

2.4

4

8

Succinylcholine

Peak pressure

SpO2

Tidal volume

Height/weight

Pulse

Sevoflurane

Respiration rate

ablation in proc text

Feature impact

Using SHAP values in the operating room

Red features push the risk higher

base value

hypoxemia fold risk

Green features push the risk lower

0.125

0.25

0.5

1

2

2.4

4

8

Succinylcholine

Peak pressure

SpO2

Tidal volume

Height/weight

Pulse

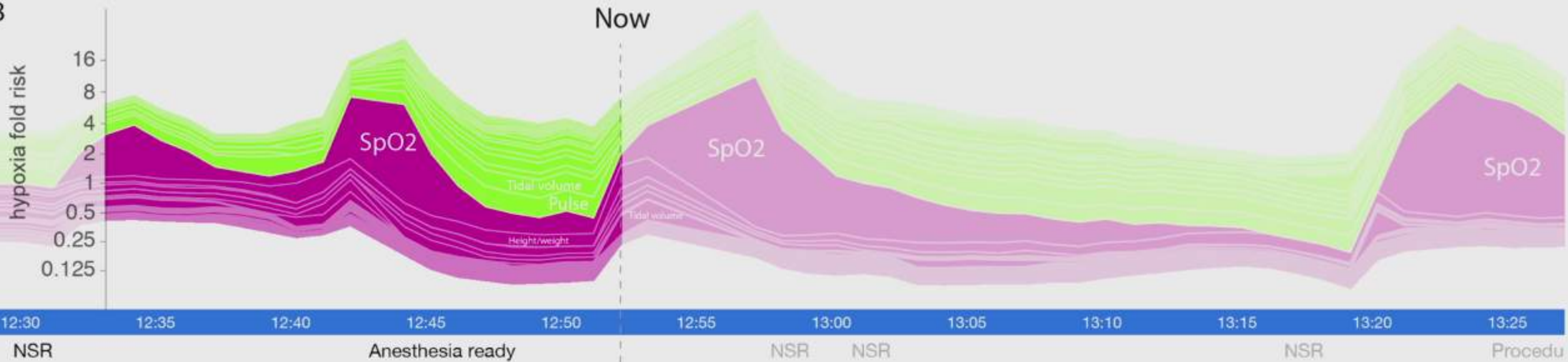
Sevoflurane

Respiration rate

ablation in proc text

Feature impact

B



Using SHAP values in the operating room

Red features push the risk higher

base value

hypoxemia fold risk

Green features push the risk lower

0.125

0.25

0.5

1

2

2.4

4

8

Succinylcholine

Peak pressure

SpO2

Tidal volume

Height/weight

Pulse

Sevoflurane

Respiration rate

ablation in proc text

Feature impact

Using SHAP values in the operating room

Red features push the risk higher

base value

hypoxemia fold risk

Green features push the risk lower

0.125

0.25

0.5

1

2

2.4

4

8

Succinylcholine

Peak pressure

SpO2

Tidal volume

Height/weight

Pulse

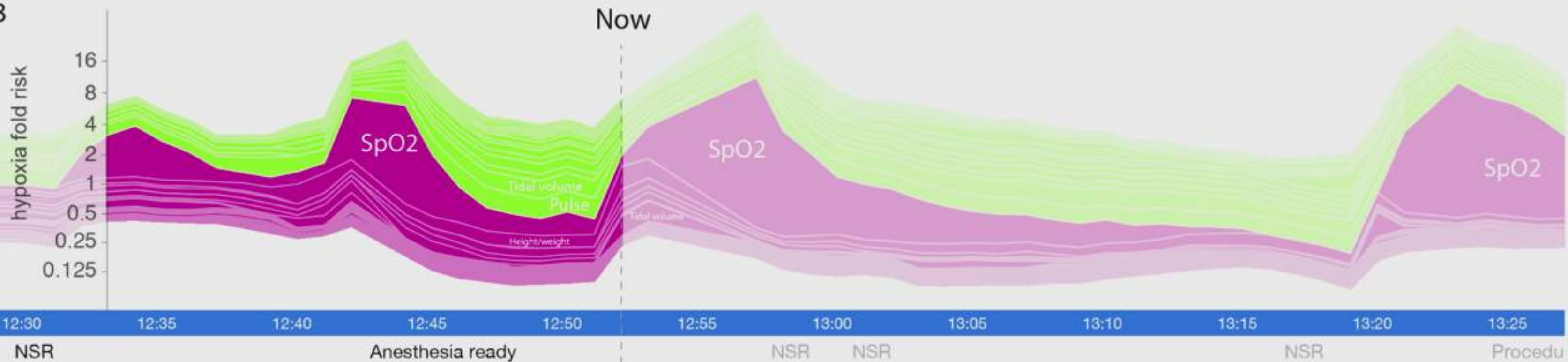
Sevoflurane

Respiration rate

ablation in proc text

Feature impact

B



Using SHAP values in the operating room

Red features push the risk higher

base value

hypoxemia fold risk

Green features push the risk lower

0.125

0.25

0.5

1

2

2.4

4

8

Succinylcholine

Peak pressure

SpO2

Tidal volume

Height/weight

Pulse

Sevoflurane

Respiration rate

ablation in proc text

Feature impact

Using SHAP values in the operating room

Red features push the risk higher

base value

hypoxemia fold risk

Green features push the risk lower

0.125

0.25

0.5

1

2

2.4

4

8

Succinylcholine

Peak pressure

SpO2

Tidal volume

Height/weight

Pulse

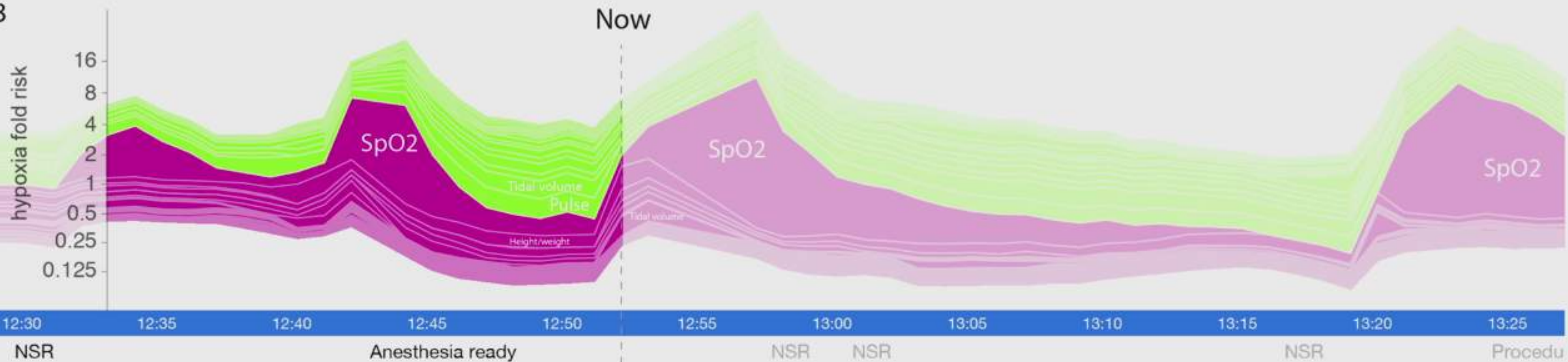
Sevoflurane

Respiration rate

ablation in proc text

Feature impact

B



Using SHAP values in the operating room

Red features push the risk higher

base value

hypoxemia fold risk

Green features push the risk lower

0.125

0.25

0.5

1

2

2.4

4

8

Succinylcholine

Peak pressure

SpO₂

Tidal volume

Height/weight

Pulse

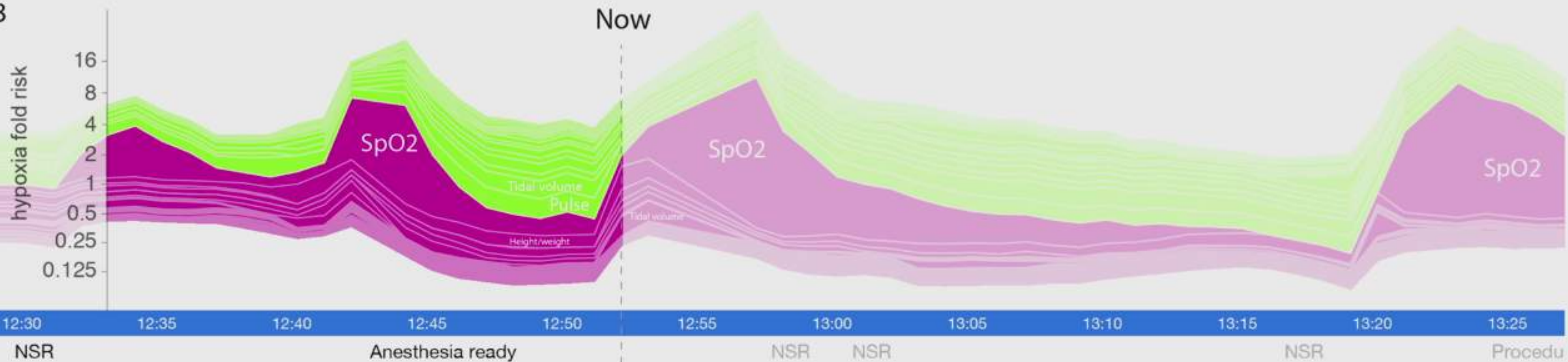
Sevoflurane

Respiration rate

ablation in proc text

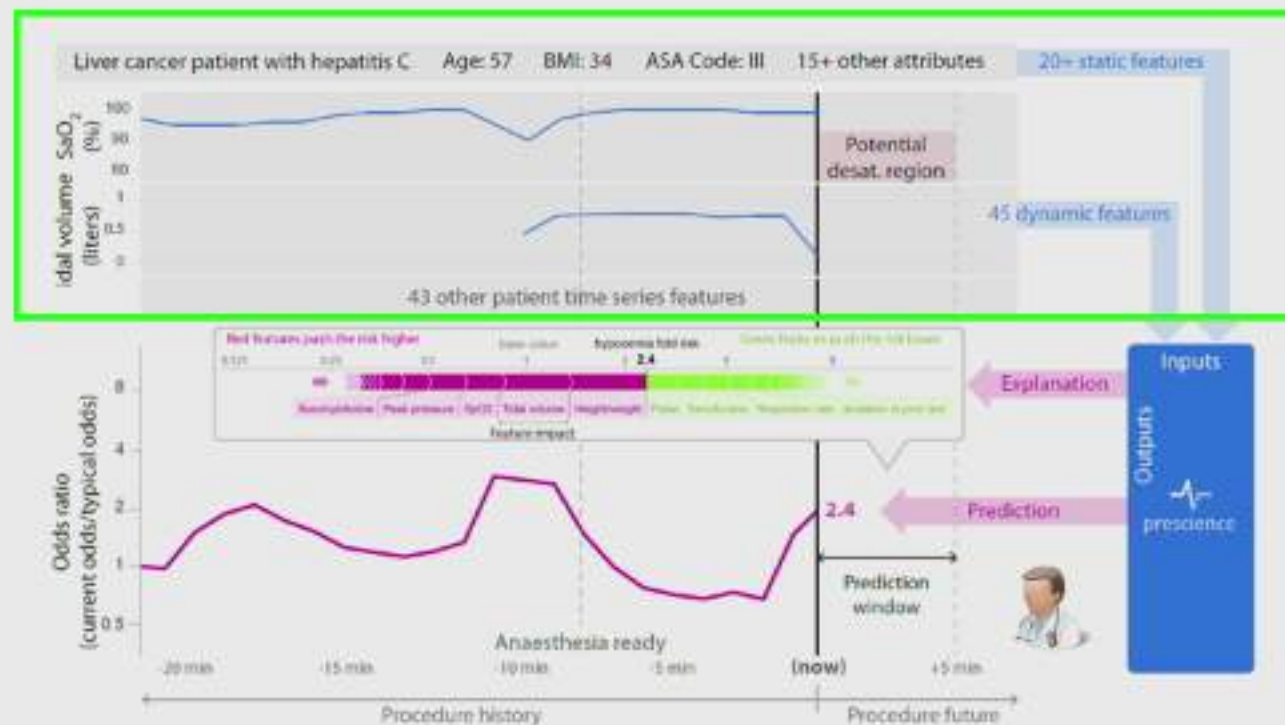
Feature impact

B



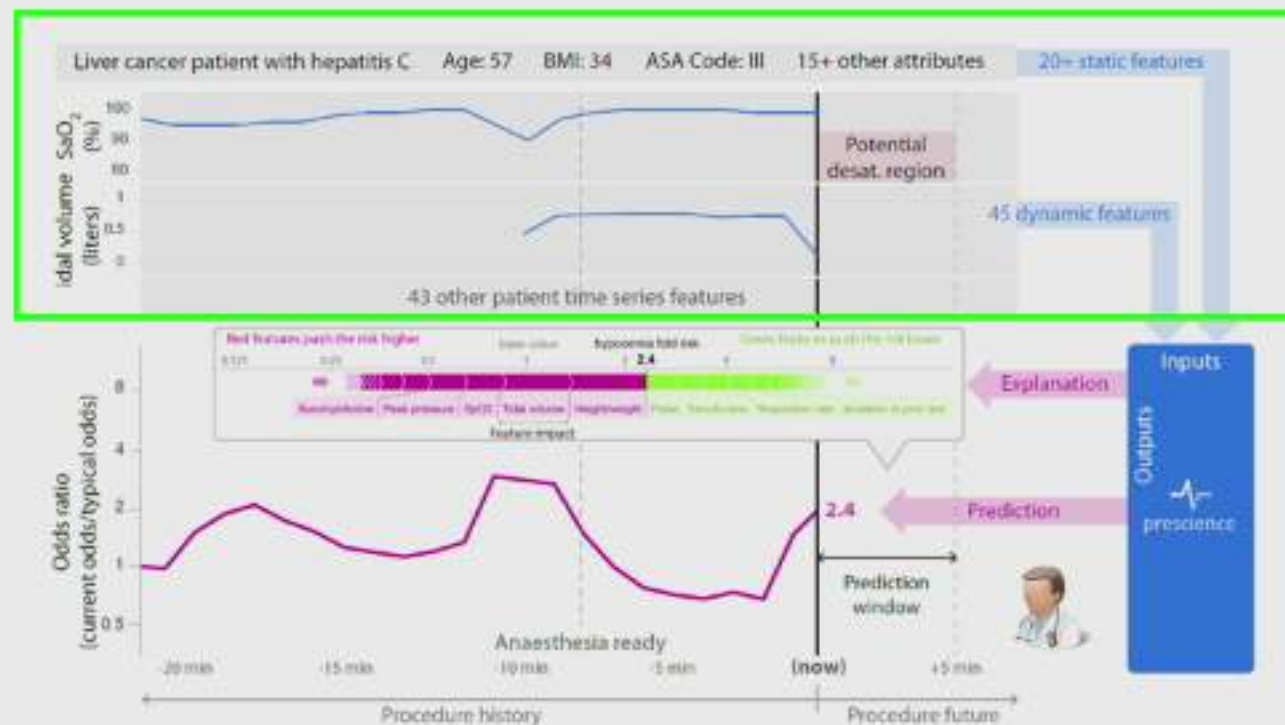
Prescience improves anesthesiologist's ability to predict hypoxemia

- We replayed prerecorded surgery data in a web-based visualization to 5 anesthesiologists.



Prescience improves anesthesiologist's ability to predict hypoxemia

- We replayed prerecorded surgery data in a web-based visualization to 5 anesthesiologists.
- Each anesthesiologist provided a relative risk of hypoxemia for ~270 cases **without** or **with** the aid of Prescience.



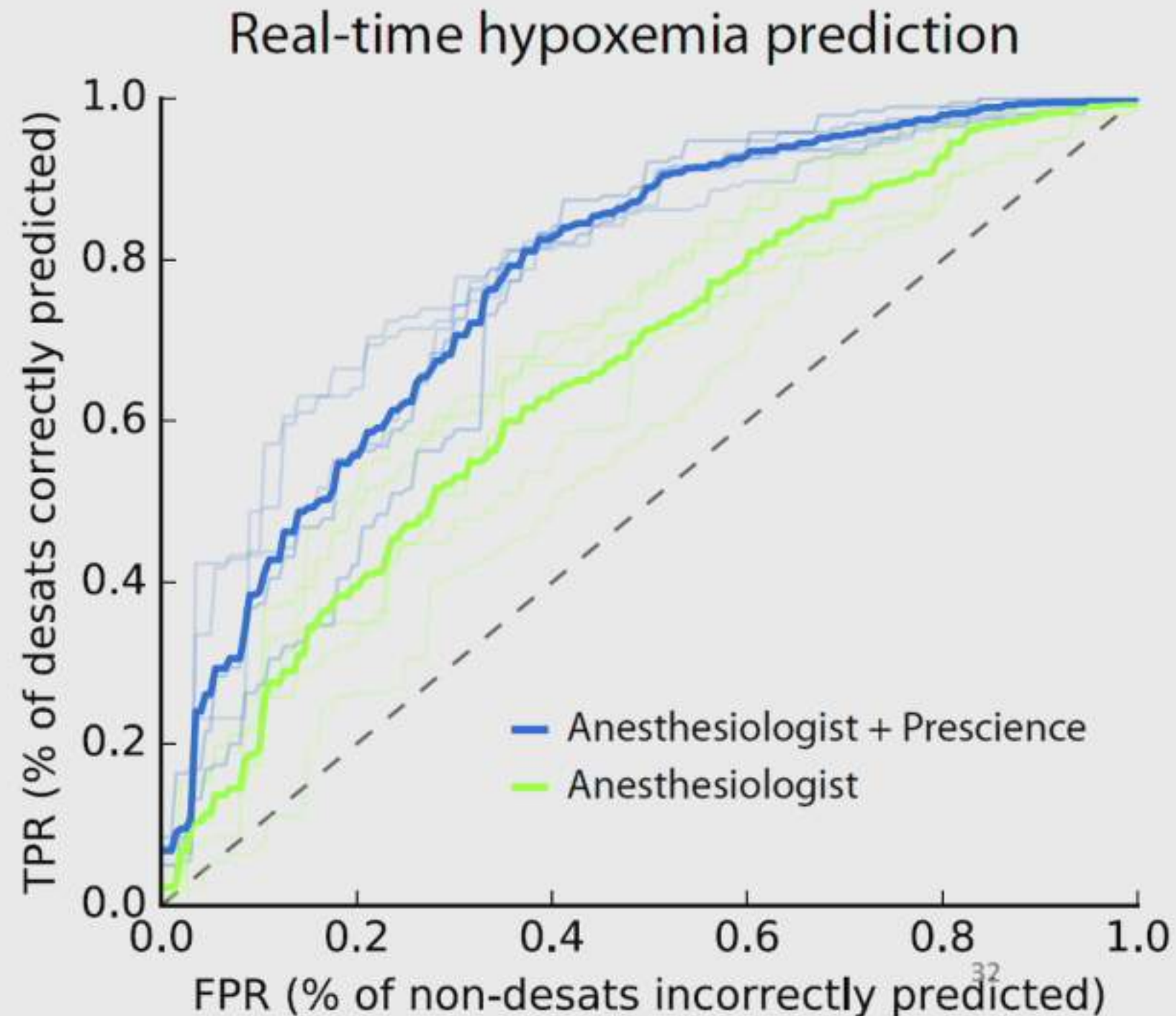
Prescience improves anesthesiologist's ability to predict hypoxemia

- We replayed prerecorded surgery data in a web-based visualization to 5 anesthesiologists.
- Each anesthesiologist provided a relative risk of hypoxemia for ~270 cases **without** or **with** the aid of Prescience.



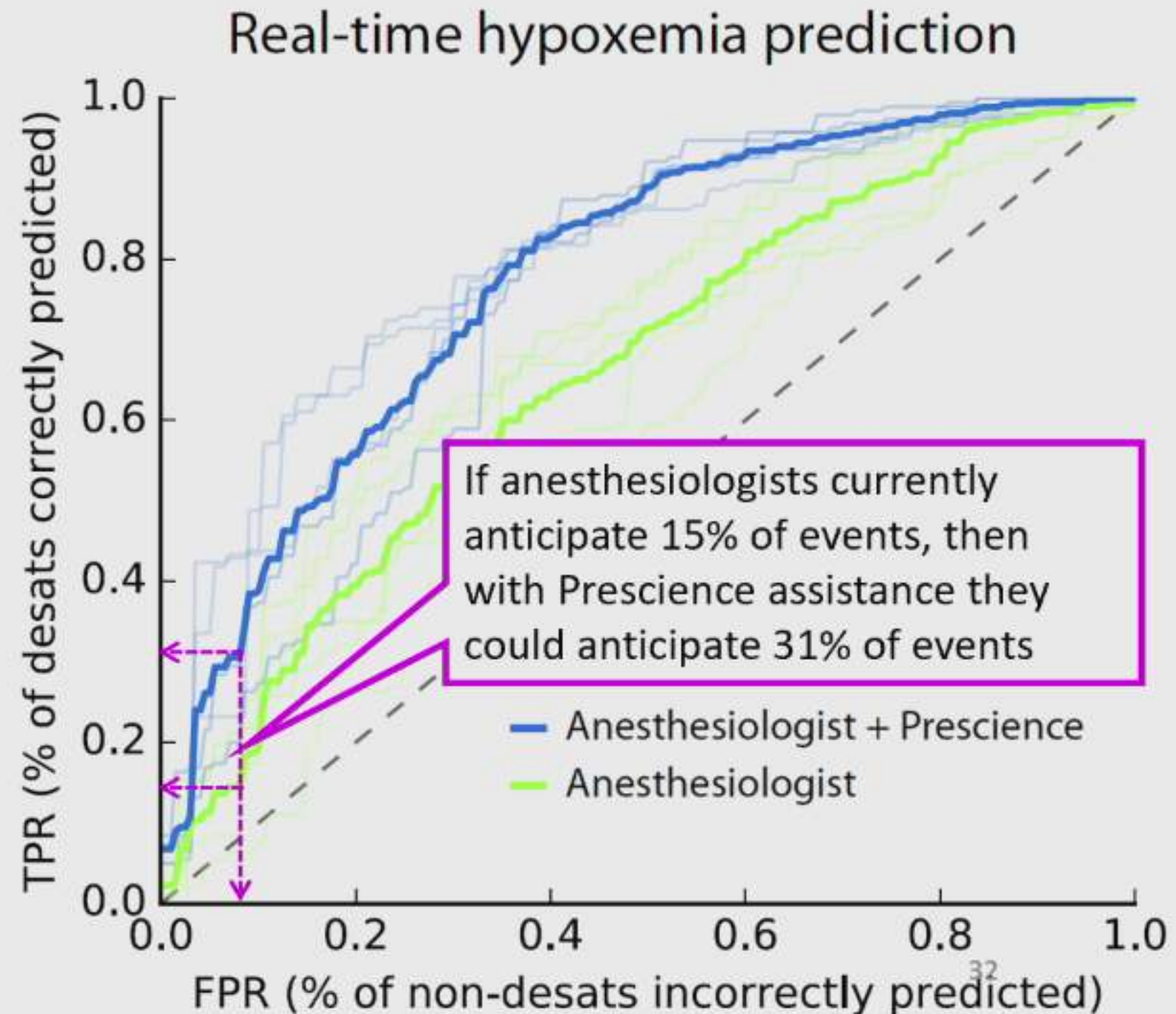
Prescience improves anesthesiologist's ability to predict hypoxemia

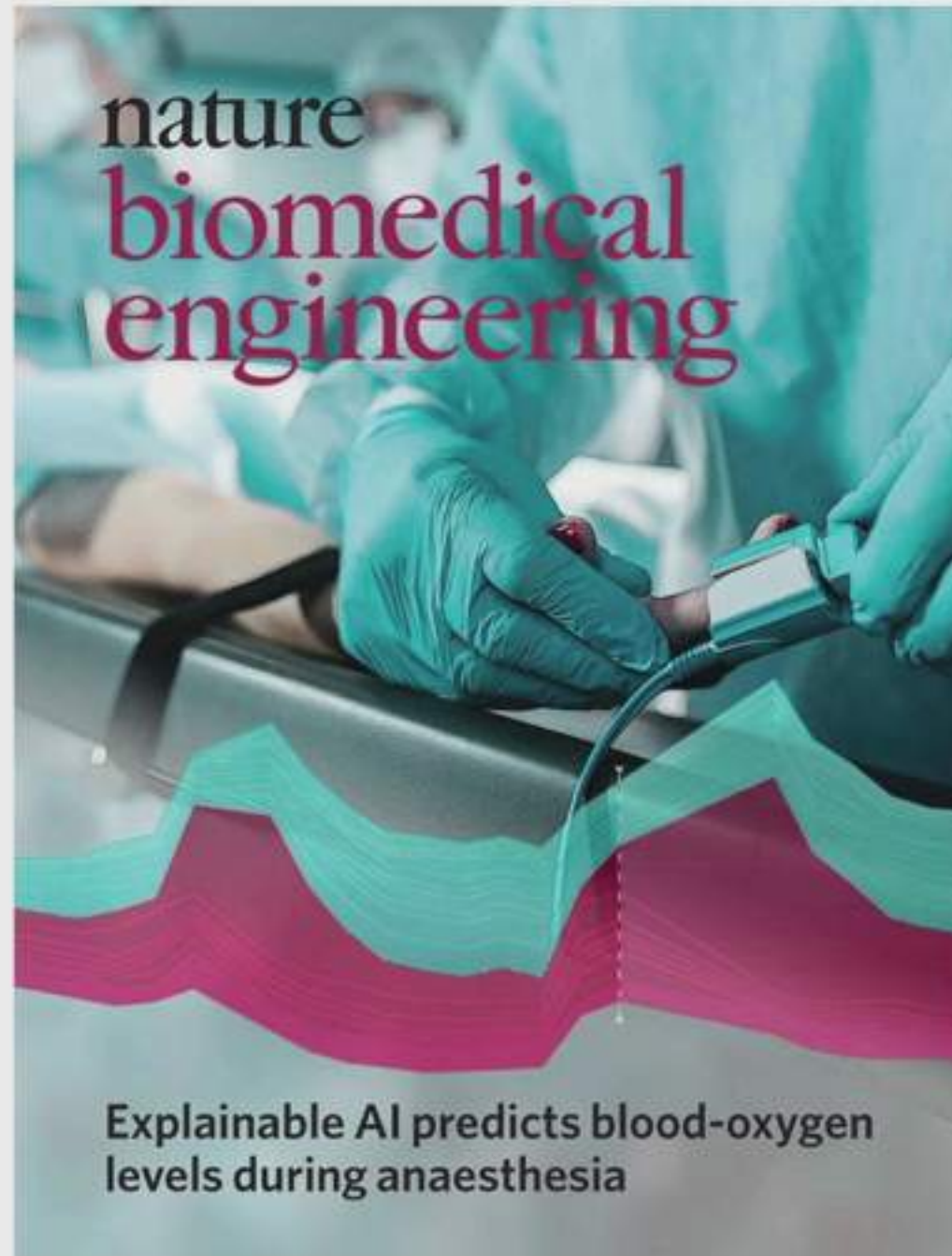
- We replayed prerecorded surgery data in a web-based visualization to 5 anesthesiologists.
- Each anesthesiologist provided a relative risk of hypoxemia for ~270 cases **without** or **with** the aid of Prescience.



Prescience improves anesthesiologist's ability to predict hypoxemia

- We replayed prerecorded surgery data in a web-based visualization to 5 anesthesiologists.
- Each anesthesiologist provided a relative risk of hypoxemia for ~270 cases **without** or **with** the aid of Prescience.



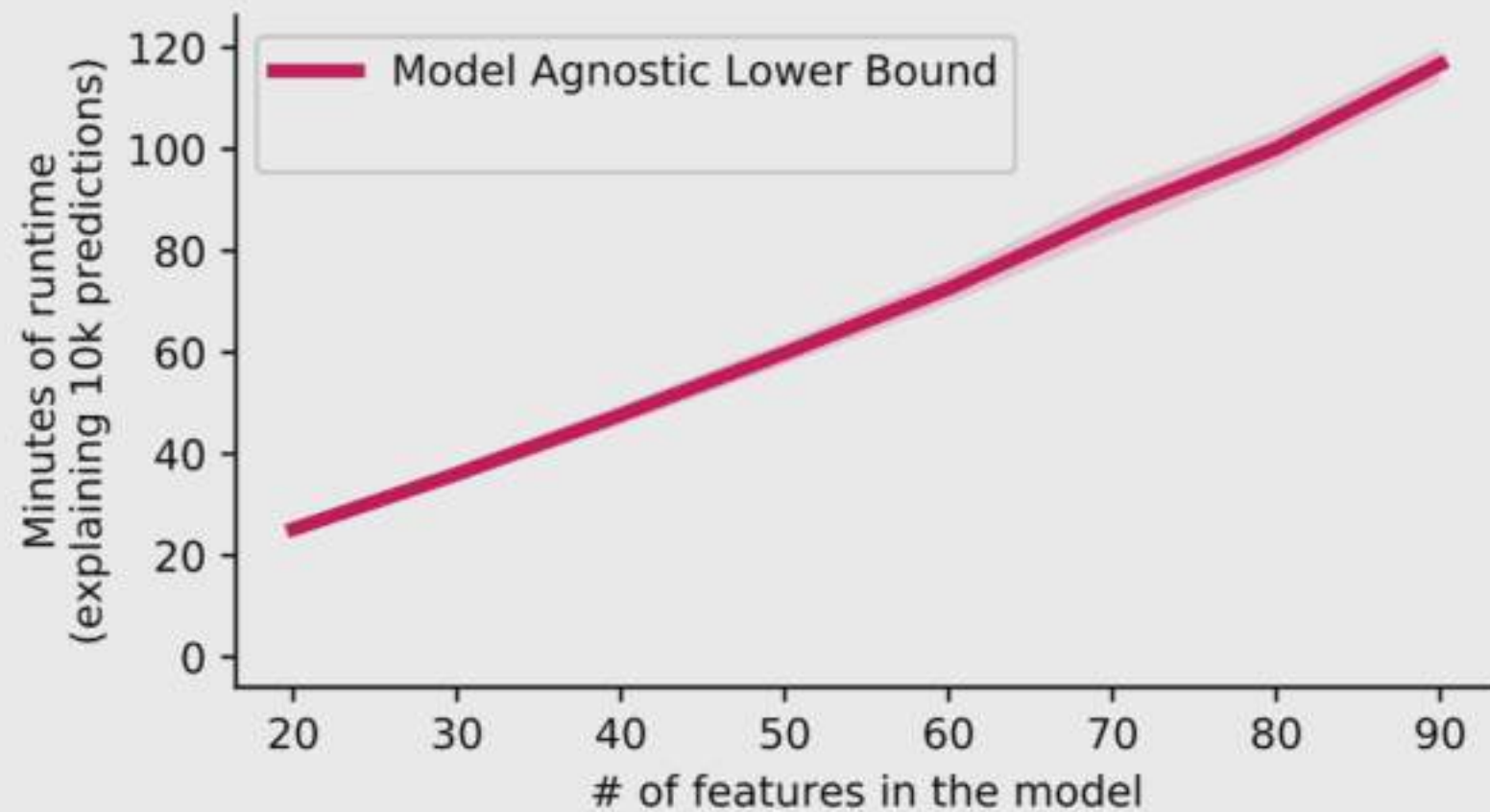


Lundberg et al., Explainable machine-learning predictions for the prevention of hypoxemia during surgery, Nature Biomedical Engineering 2018 (*cover article*)

Room for improvement: Model agnostic approaches can be slow and variable

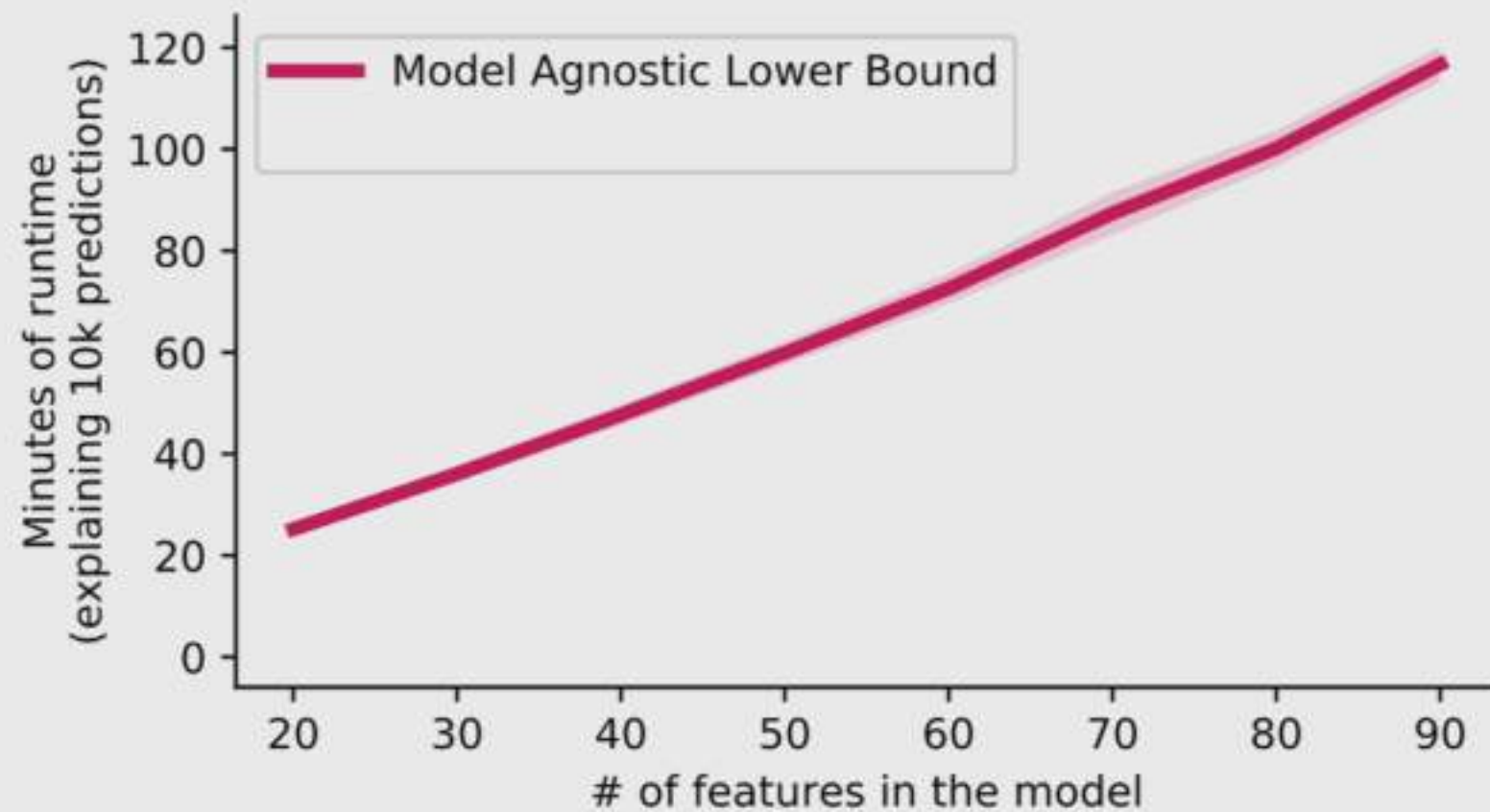
Room for improvement: Model agnostic approaches can be slow and variable

Explanation runtime (simulated data)

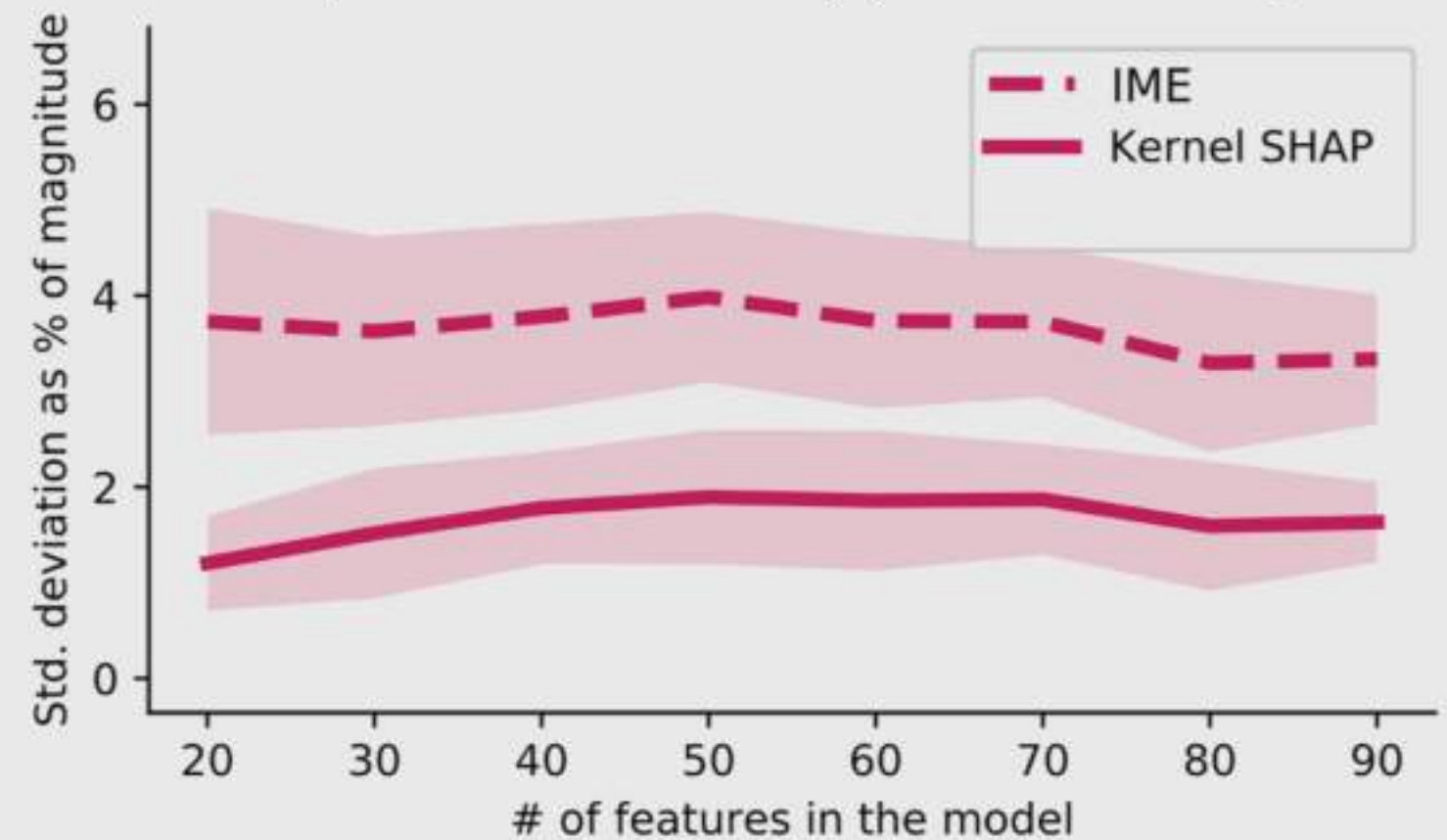


Room for improvement: Model agnostic approaches can be slow and variable

Explanation runtime (simulated data)



Explanation variability (simulated data)



Options for NP-hard problems:

~~1. Prove that $P = NP$.~~

2. Find an approximate solution.

Options for NP-hard problems:

~~1. Prove that $P = NP$.~~

2. Find an approximate solution.

3. Restrict the problem definition.

Explainable AI for Science and Medicine

Theory



Unification of explanation methods



Strong uniqueness results

Practice



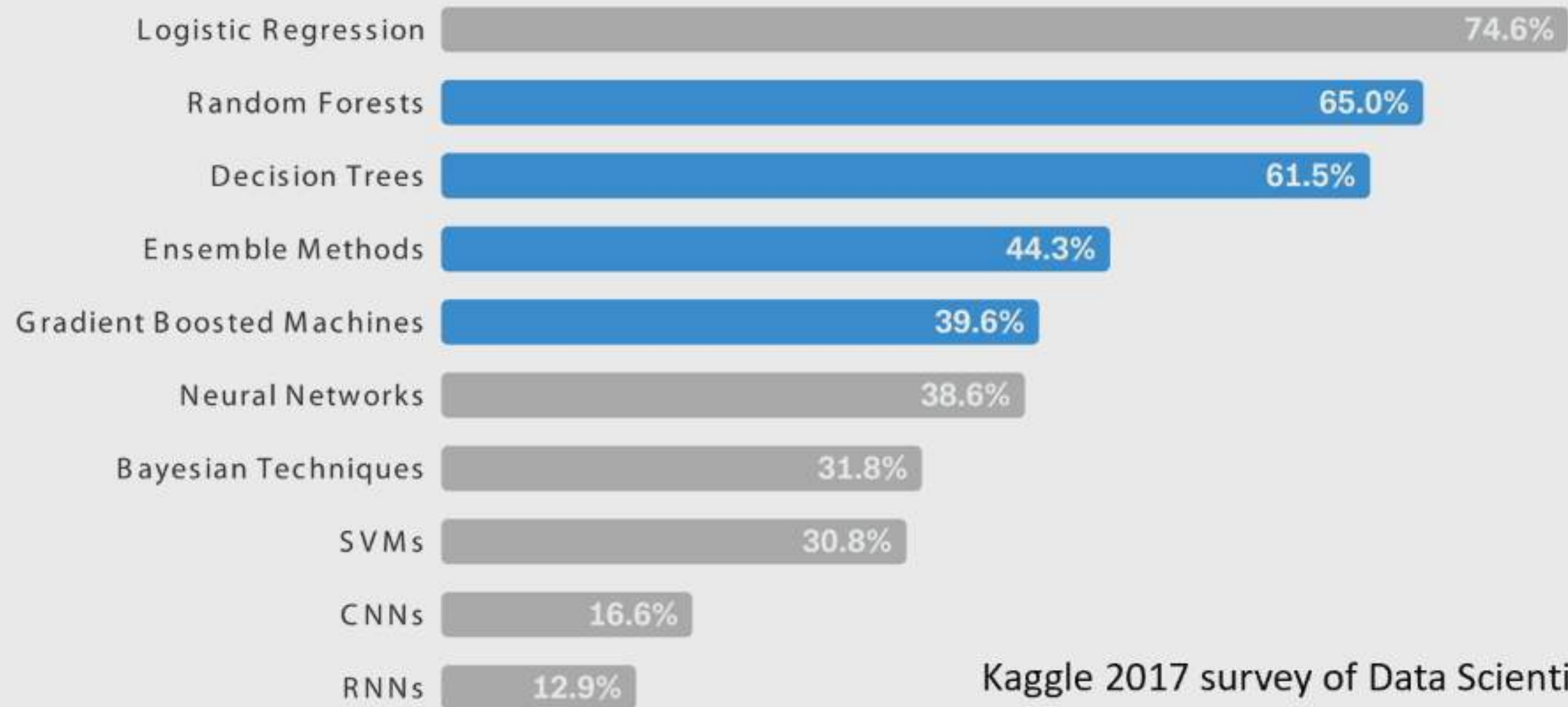
New estimation methods for the classic Shapley values

Application



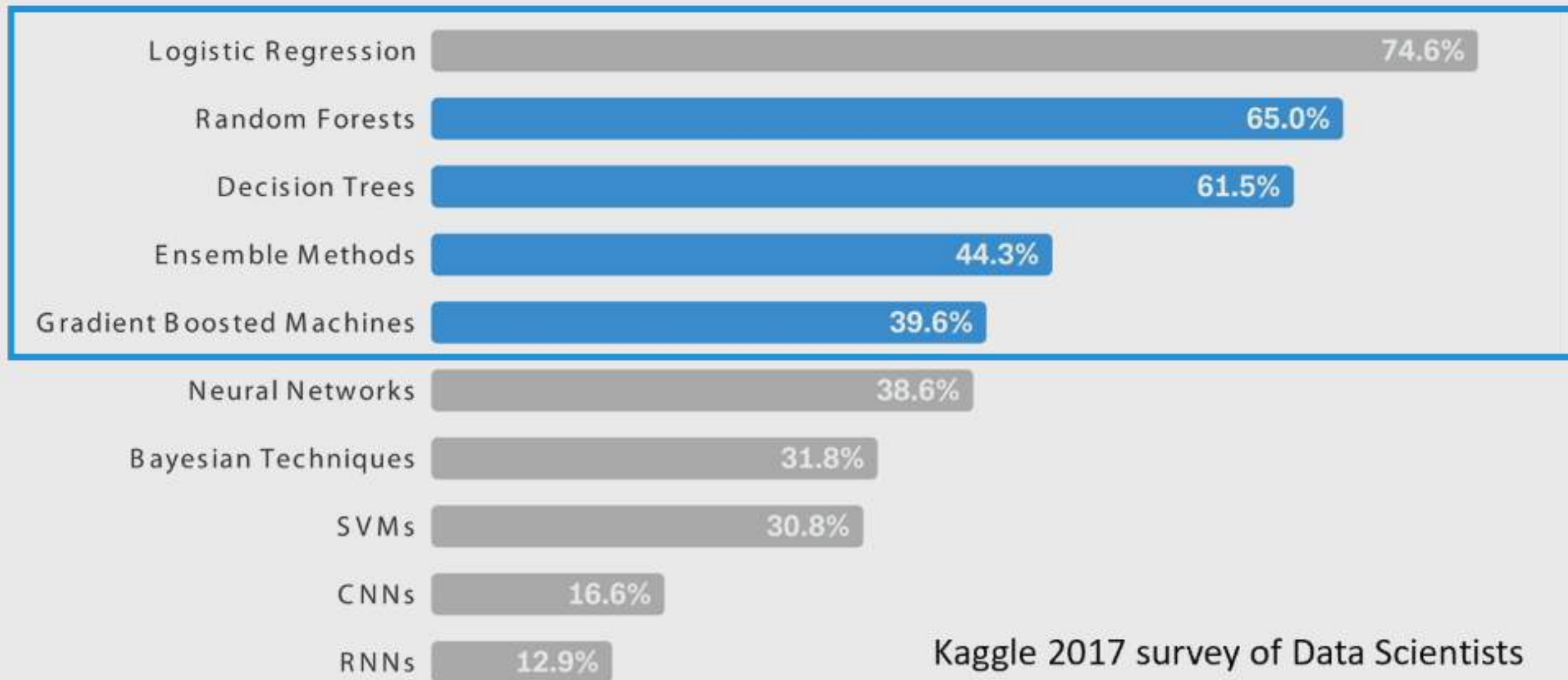
Anesthesia safety

Tree-based models are the most popular complex models used in industry



Kaggle 2017 survey of Data Scientists

Tree-based models are the most popular complex models used in industry



Kaggle 2017 survey of Data Scientists

SHAP values for trees

Direct Solution $O(TLM!N)$ Factorial

SHAP values for trees

Direct Solution

$O(TLM!N)$ Factorial

$O(TL2^M N)$ Exponential

SHAP values for trees

Direct Solution

$O(TLM!N)$ Factorial

$O(TL2^M N)$ Exponential

The solution depends on an exponential number of expected values!

SHAP values for trees

Direct Solution

$O(TLM!N)$ **Factorial**

$O(TL2^M N)$ **Exponential**

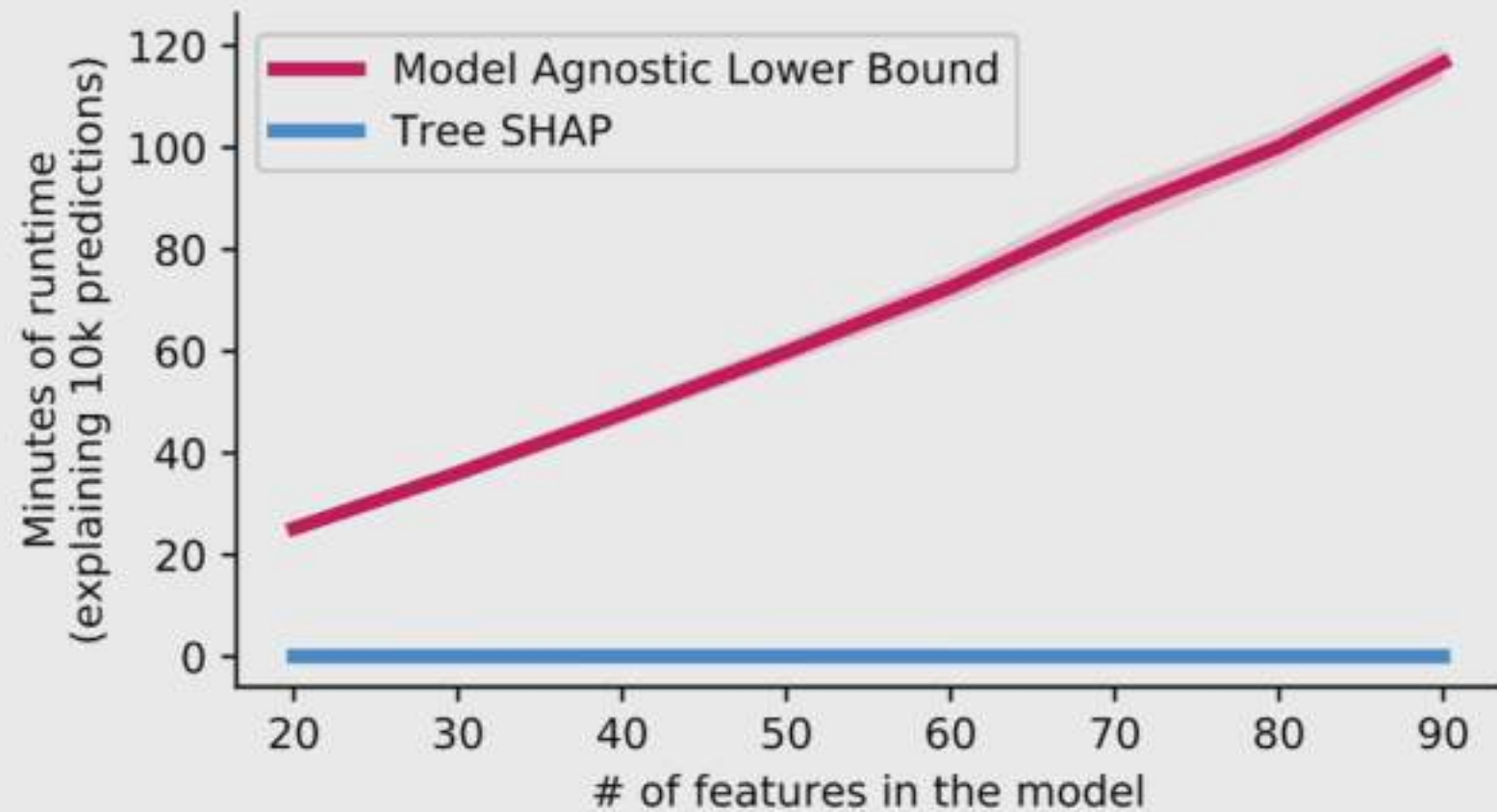
to

Tree SHAP

$O(TLD^2)$ **Polynomial**

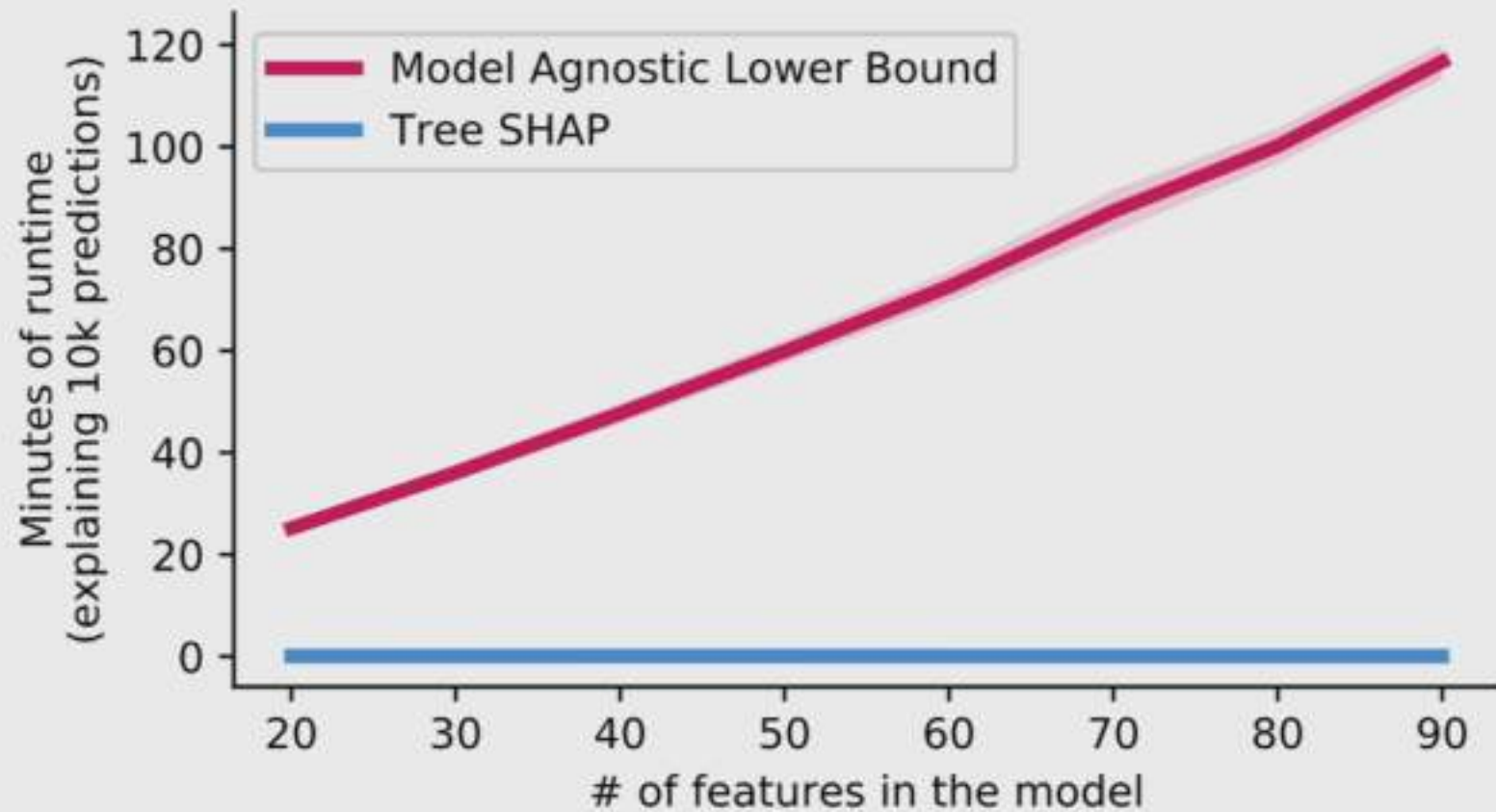
Tree SHAP is fast and exact

Explanation runtime (simulated data)

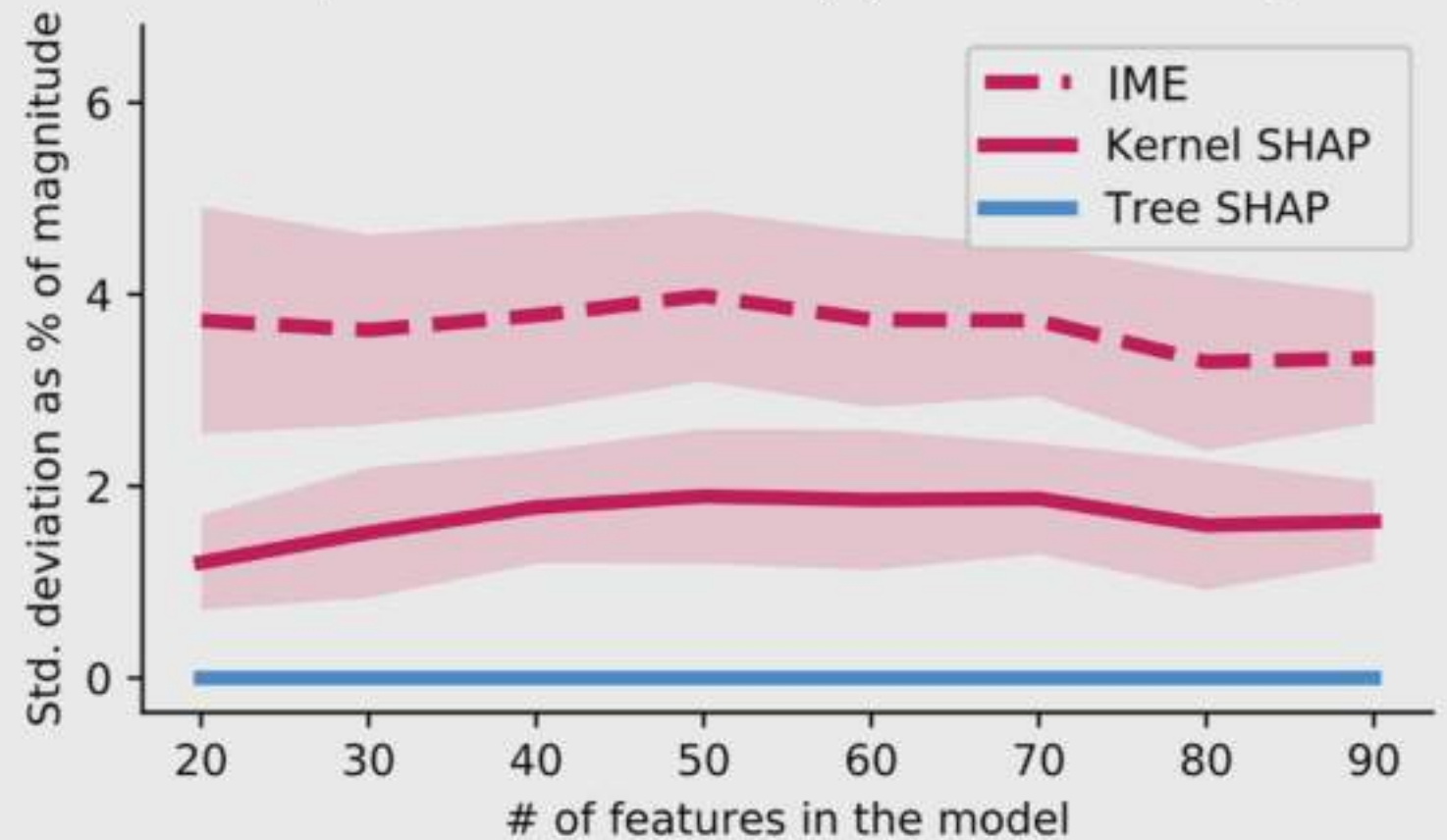


Tree SHAP is fast and exact

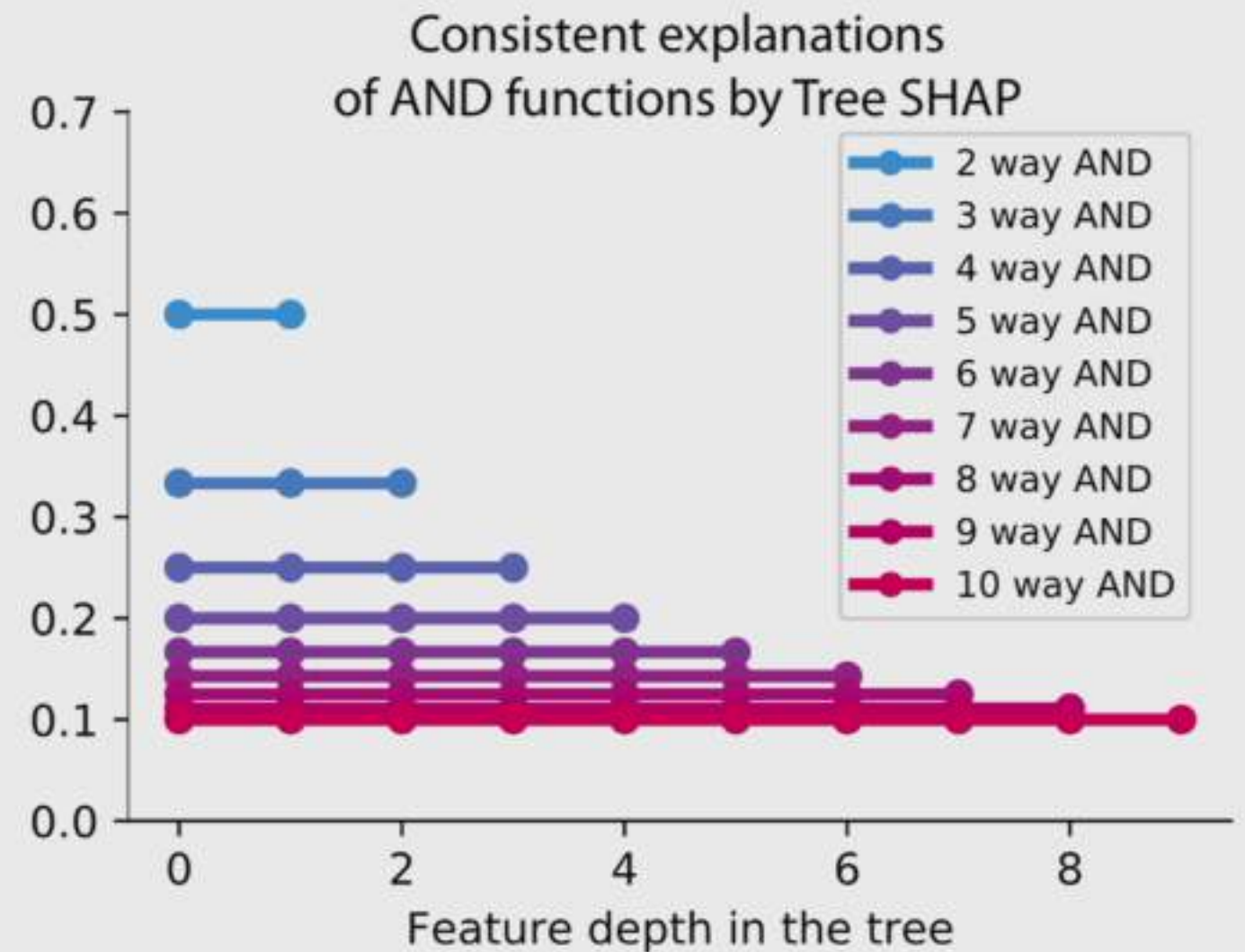
Explanation runtime (simulated data)



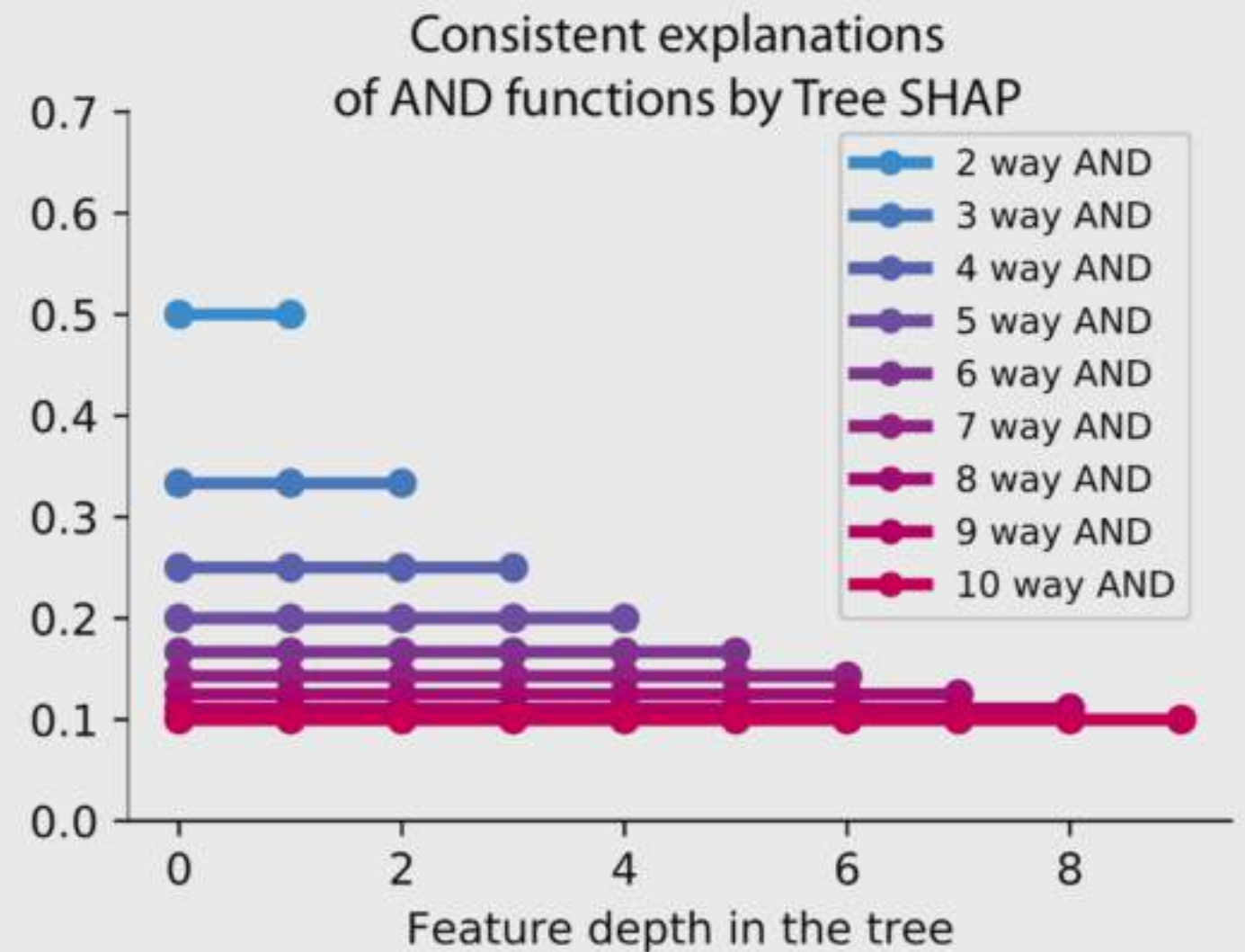
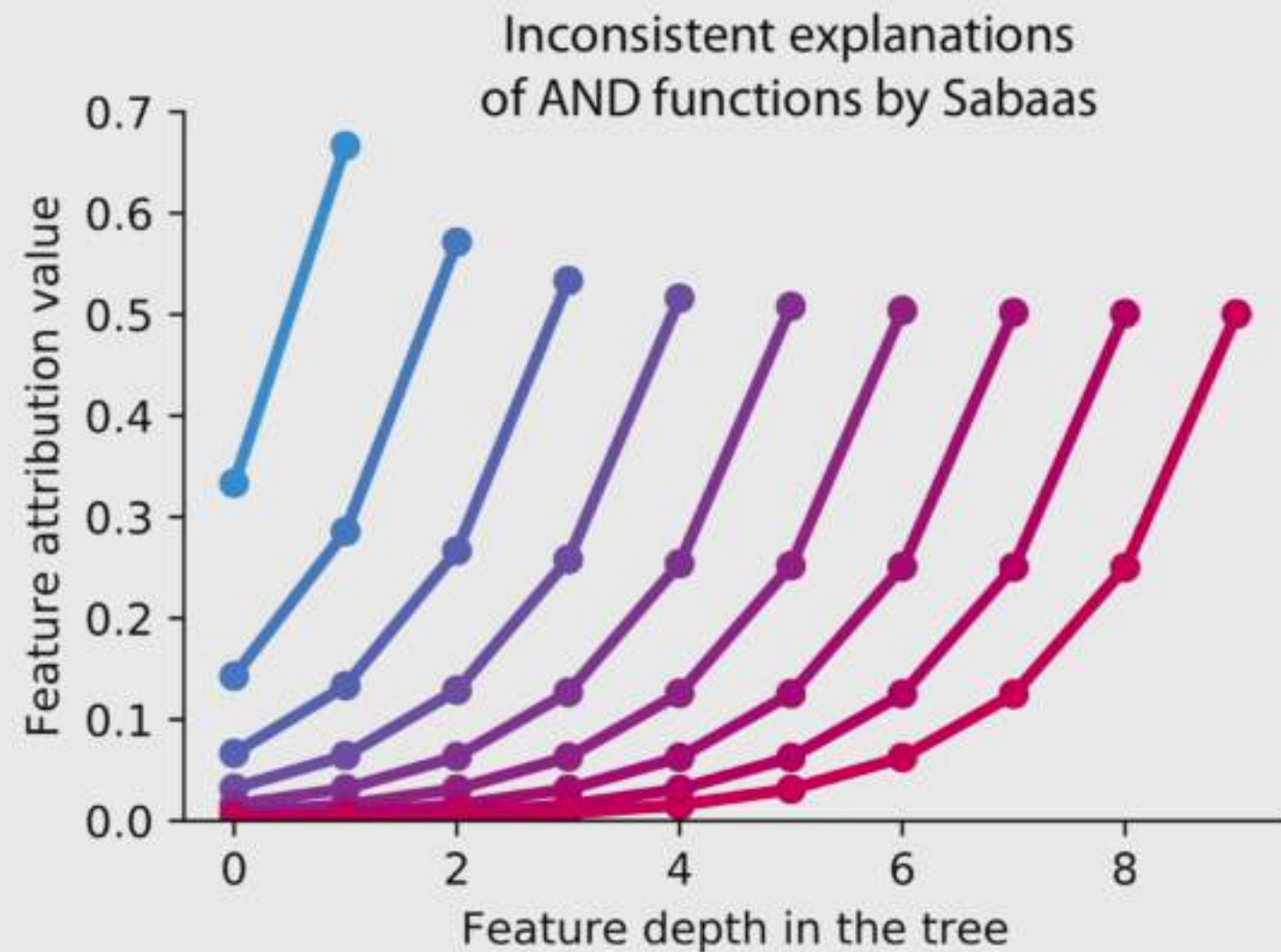
Explanation variability (simulated data)



Current tree explanation methods are inconsistent



Current tree explanation methods are inconsistent



Different evaluation metrics

Different evaluation metrics

- Runtime
- Local Accuracy
- Consistency Guarantees
- Keep Positive (mask)
- Keep Positive (resample)
- Keep Positive (impute)
- Keep Negative (mask)
- Keep Negative (resample)
- Keep Negative (impute)
- Keep Absolute (mask)
- Keep Absolute (resample)
- Keep Absolute (impute)
- Remove Positive (mask)
- Remove Positive (resample)
- Remove Positive (impute)
- Remove Negative (mask)
- Remove Negative (resample)
- Remove Negative (impute)
- Remove Absolute (mask)
- Remove Absolute (resample)
- Remove Absolute (impute)

Different explanation methods for trees

Different evaluation metrics

TreeExplainer (independent)

TreeExplainer

Saabas

Kernel SHAP 1000 mean ref.

IME 1000

mean(|TreeExplainer|)

Gain/Gini Importance

Random

TreeExplainer (independent)

TreeExplainer

Saabas

Kernel SHAP 1000 mean ref.

IME 1000

mean(|TreeExplainer|)

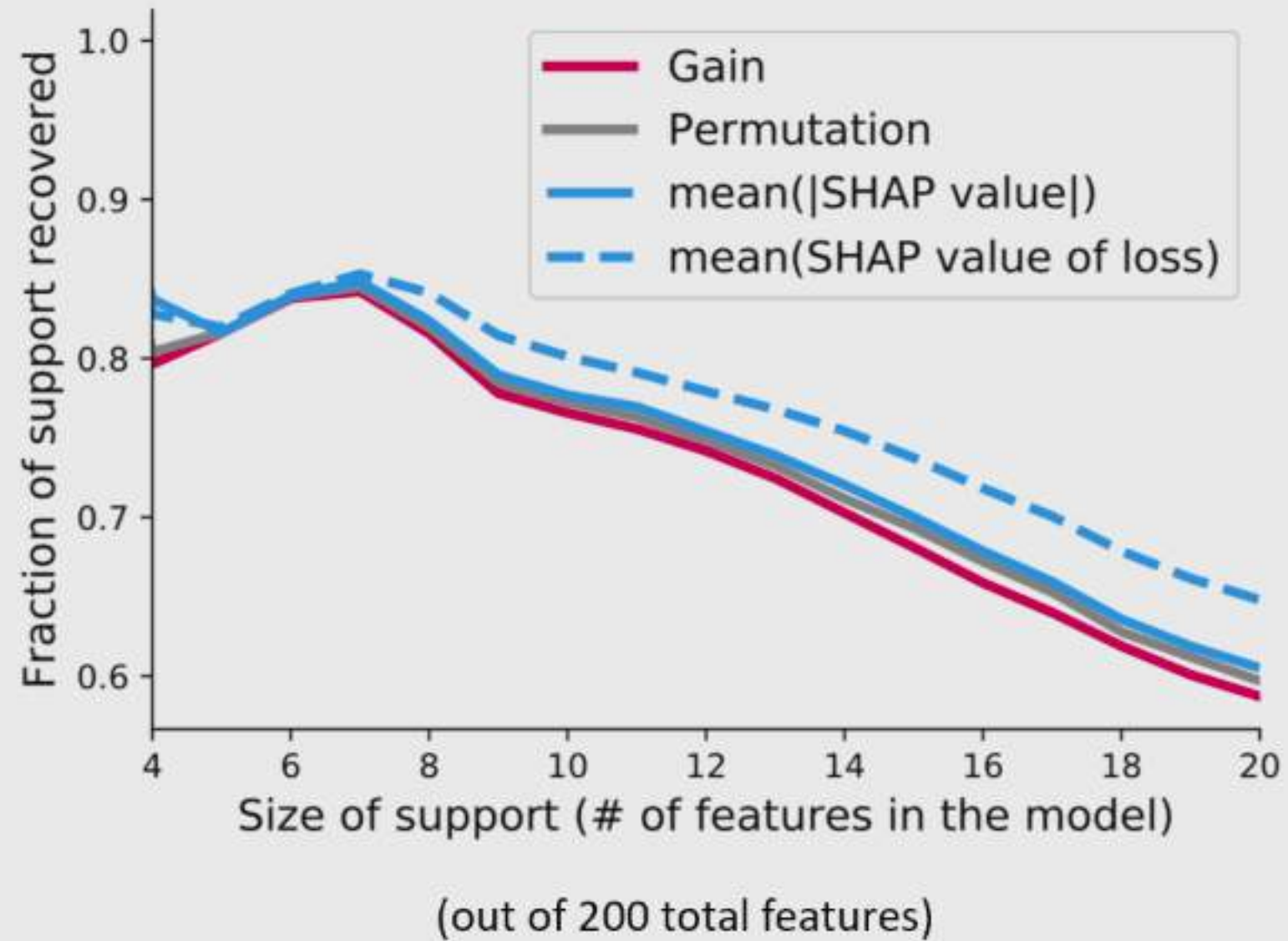
Gain/Gini Importance

Random

- Runtime
- Local Accuracy
- Consistency Guarantees
- Keep Positive (mask)
- Keep Positive (resample)
- Keep Positive (impute)
- Keep Negative (mask)
- Keep Negative (resample)
- Keep Negative (impute)
- Keep Absolute (mask)
- Keep Absolute (resample)
- Remove Positive (mask)
- Remove Positive (resample)
- Remove Positive (impute)
- Remove Negative (mask)
- Remove Negative (resample)
- Remove Negative (impute)
- Remove Absolute (mask)
- Remove Absolute (resample)
- Remove Absolute (impute)

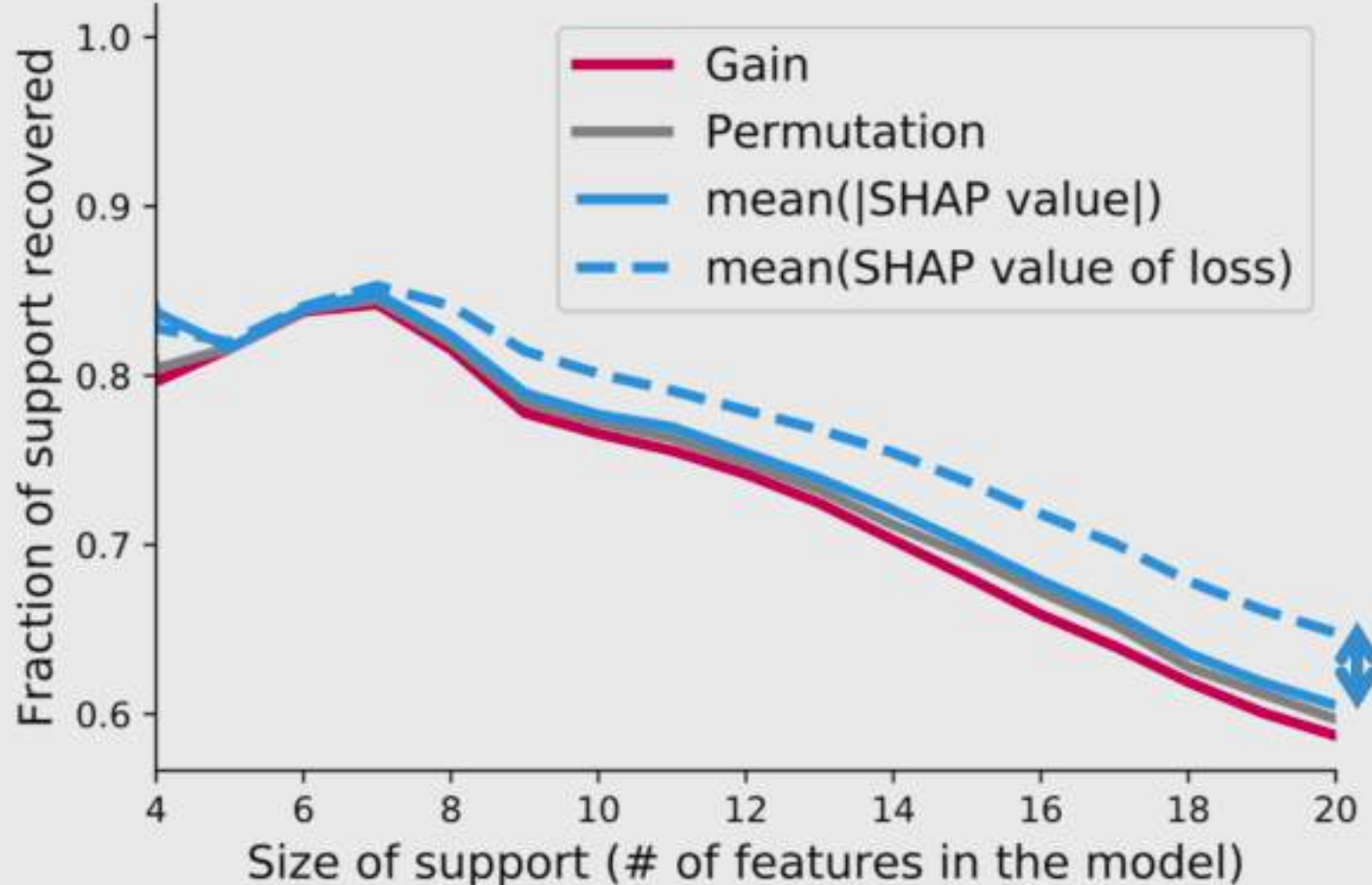
Improved feature selection power

Single decision tree
(minimum interactions)



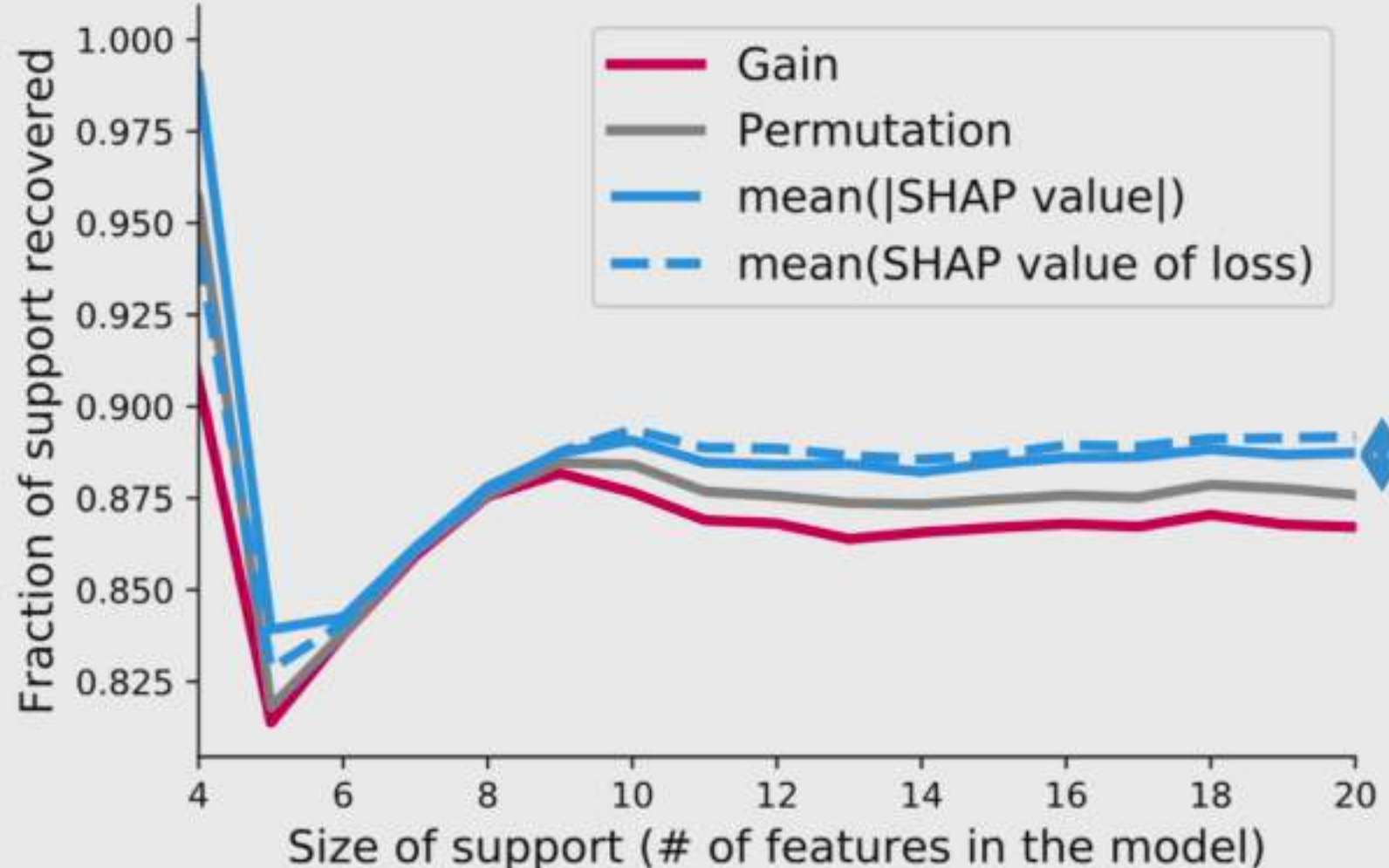
Improved feature selection power

Single decision tree
(minimum interactions)



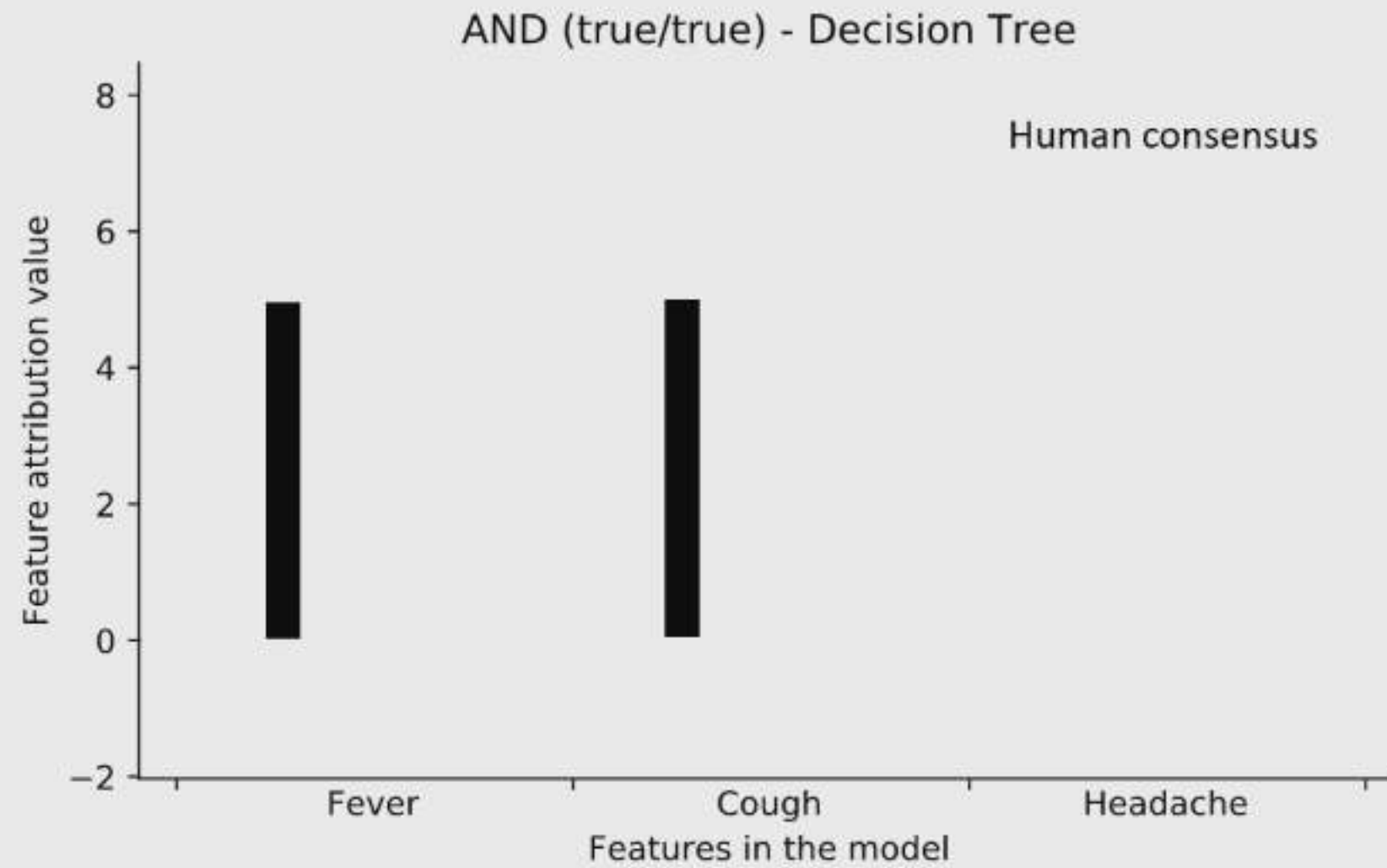
(out of 200 total features)

Random forest with 10 trees
(minimum interactions)

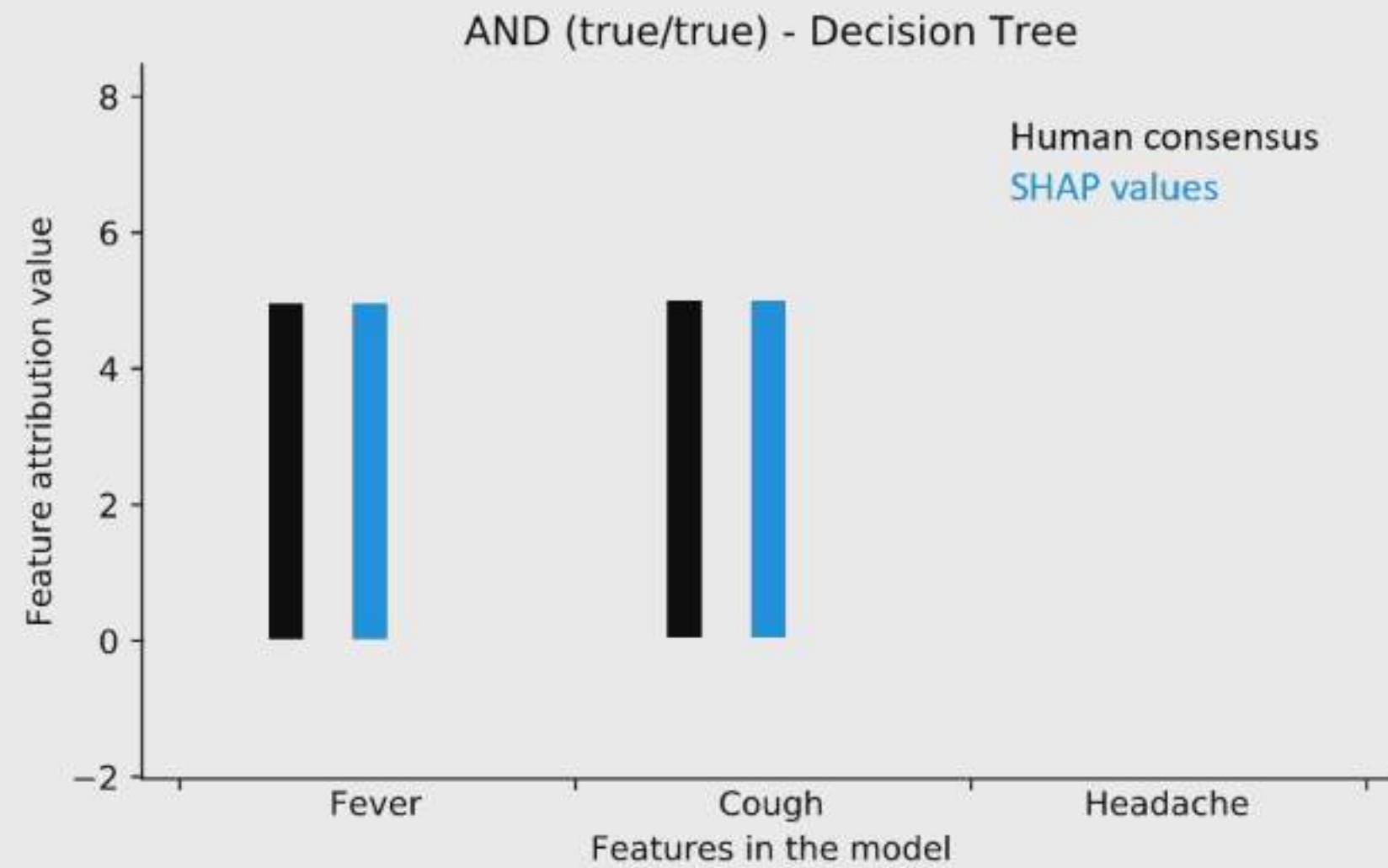


(out of 200 total features)

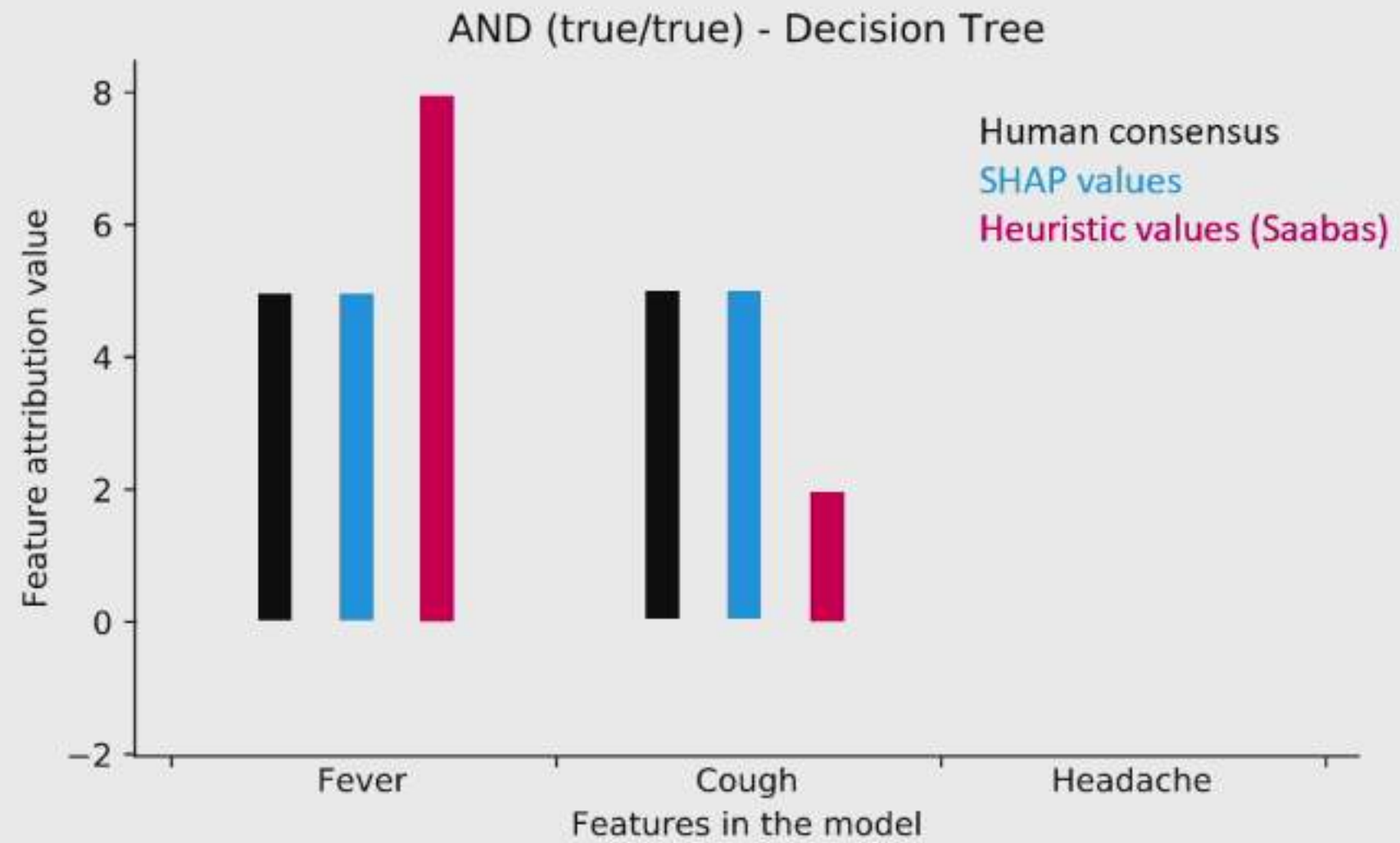
Consistency with human intuition



Consistency with human intuition



Consistency with human intuition





Fast exact computation



Fast exact computation



Attractive theoretical guarantees



Fast exact computation



Attractive theoretical guarantees



Excellent performance on XAI metrics



Fast exact computation



Attractive theoretical guarantees



Excellent performance on XAI metrics



Improves global feature selection power



Fast exact computation



Attractive theoretical guarantees



Excellent performance on XAI metrics



Improves global feature selection power



Consistent with human intuition

Explainable AI for Science and Medicine

Theory



Unification of explanation methods



Strong uniqueness results

Practice



New estimation methods for the classic Shapley values



Explainable AI tools

Application



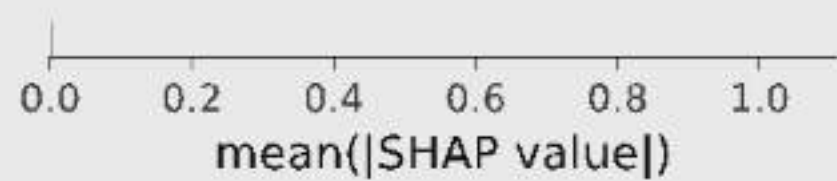
Anesthesia safety



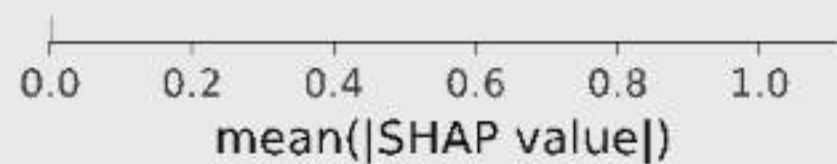
Mortality risk +
Hospital scheduling

Mortality risk model

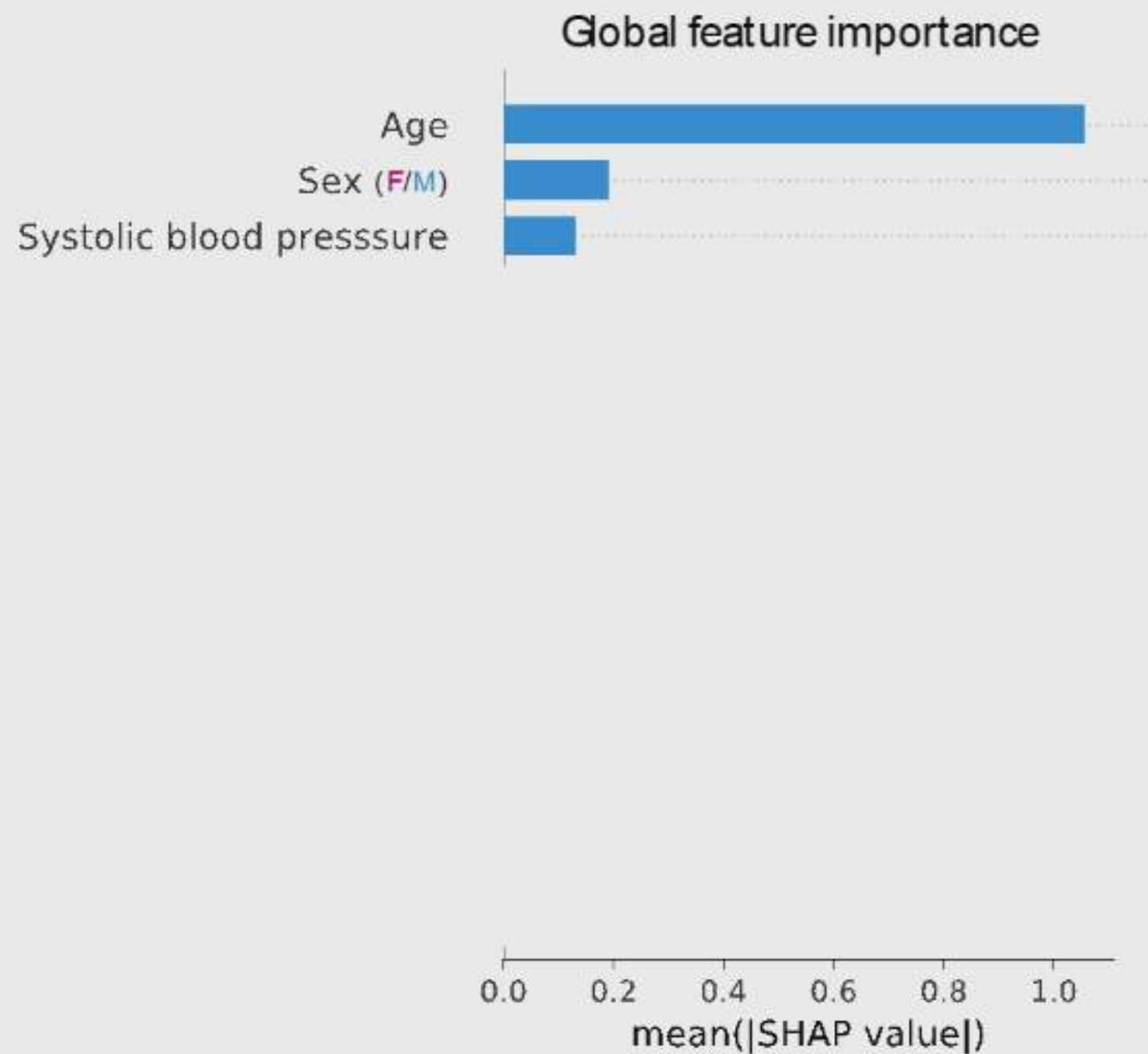
Global feature importance



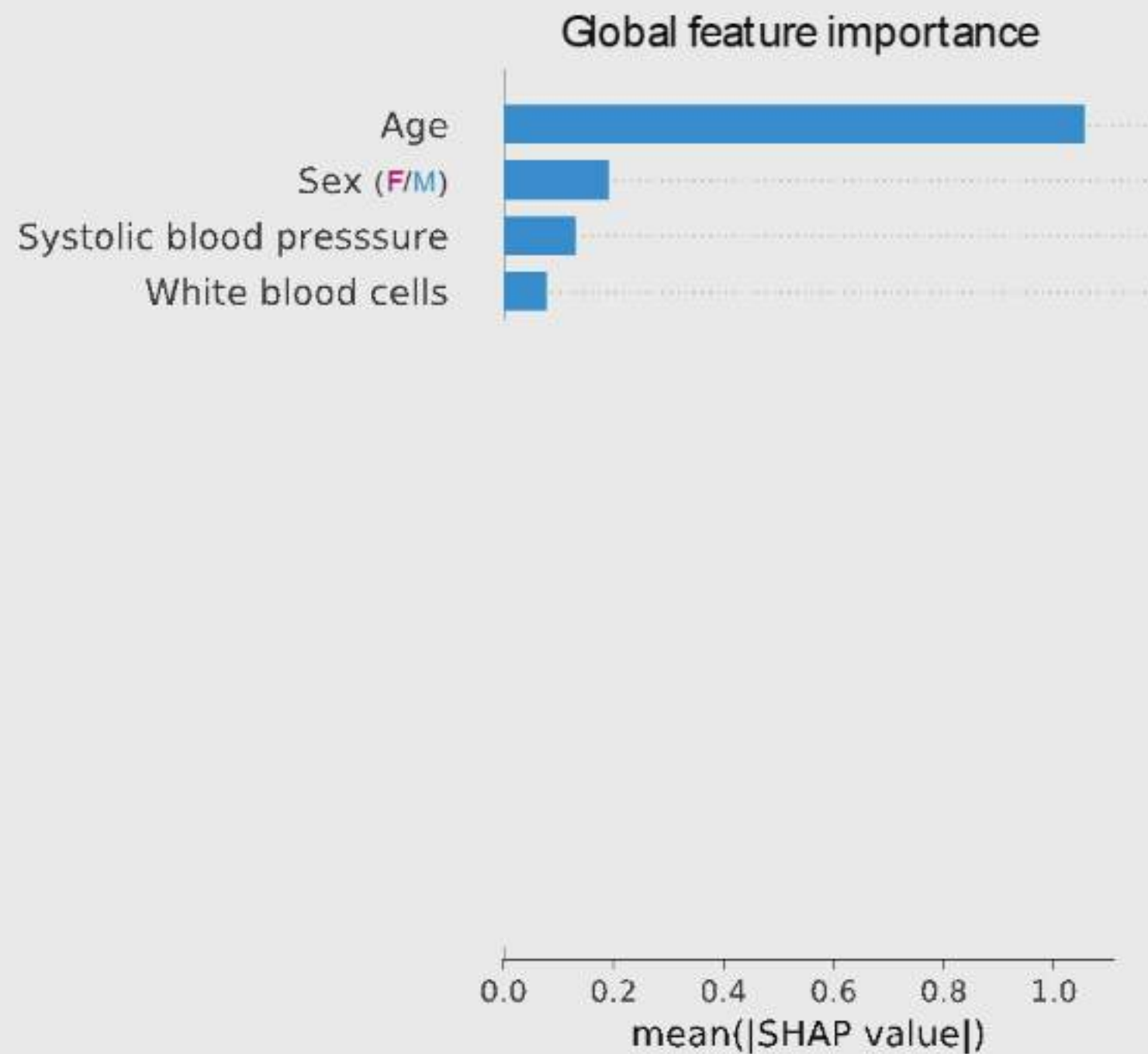
Mortality risk model



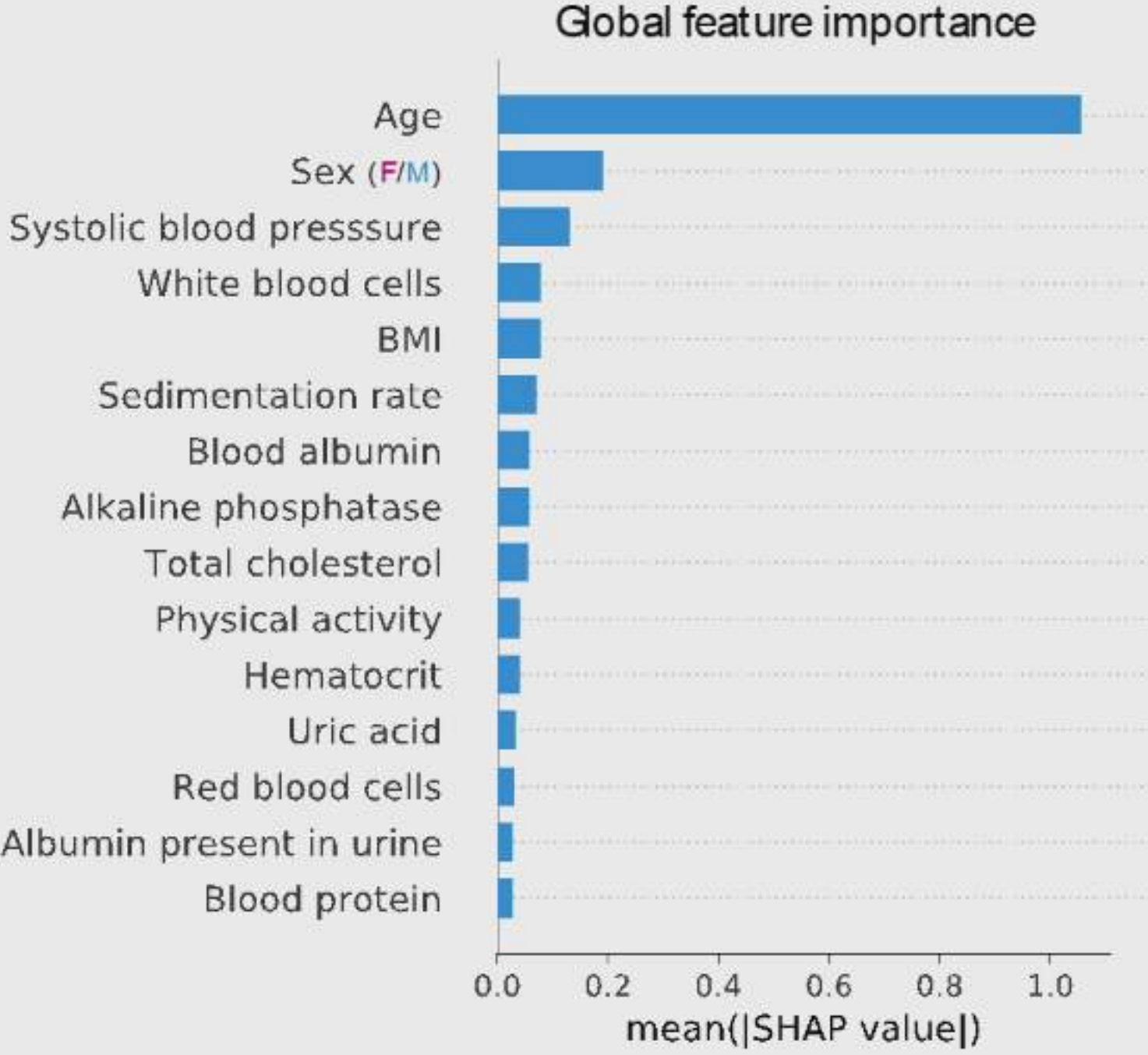
Mortality risk model



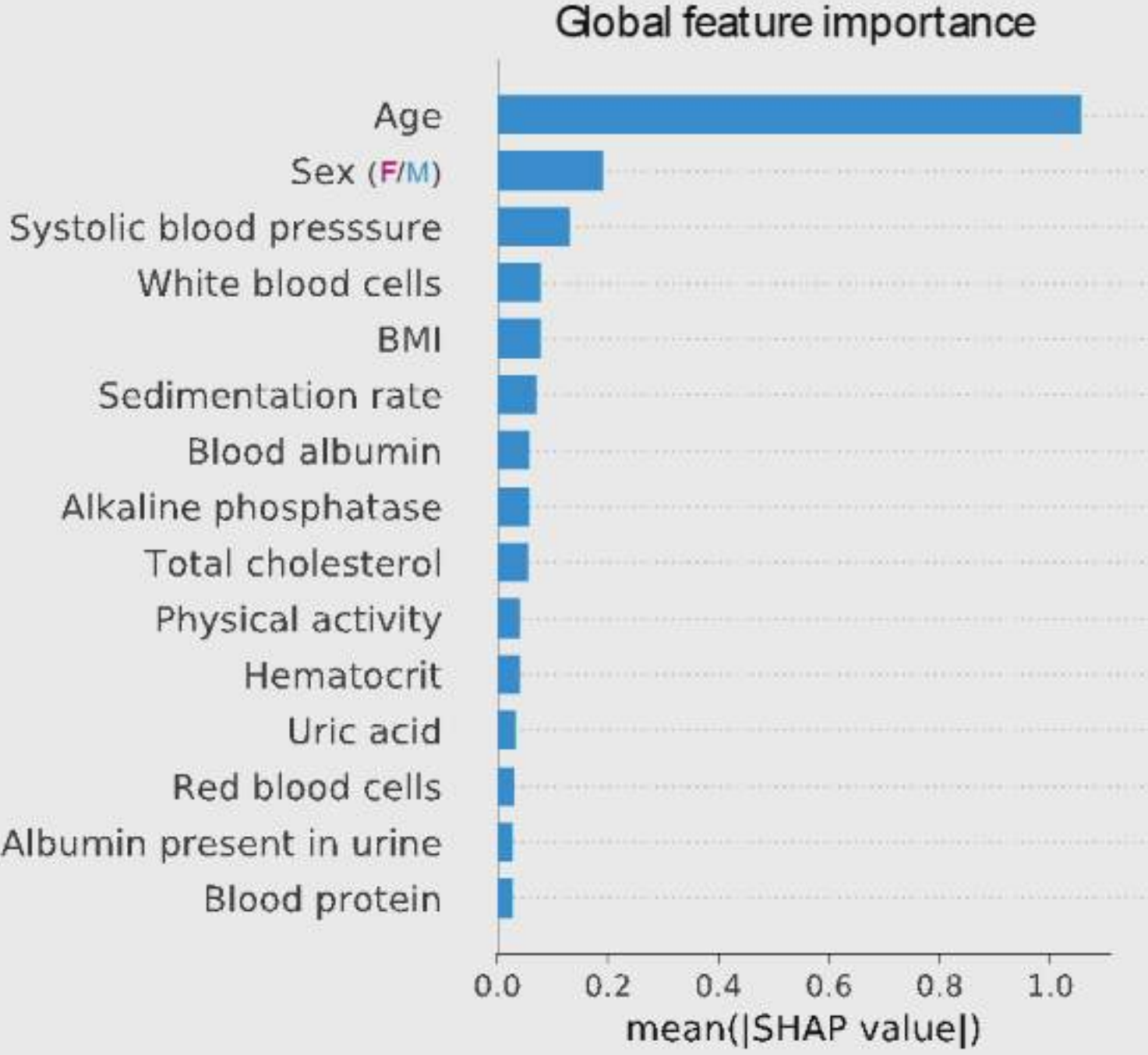
Mortality risk model



Mortality risk model

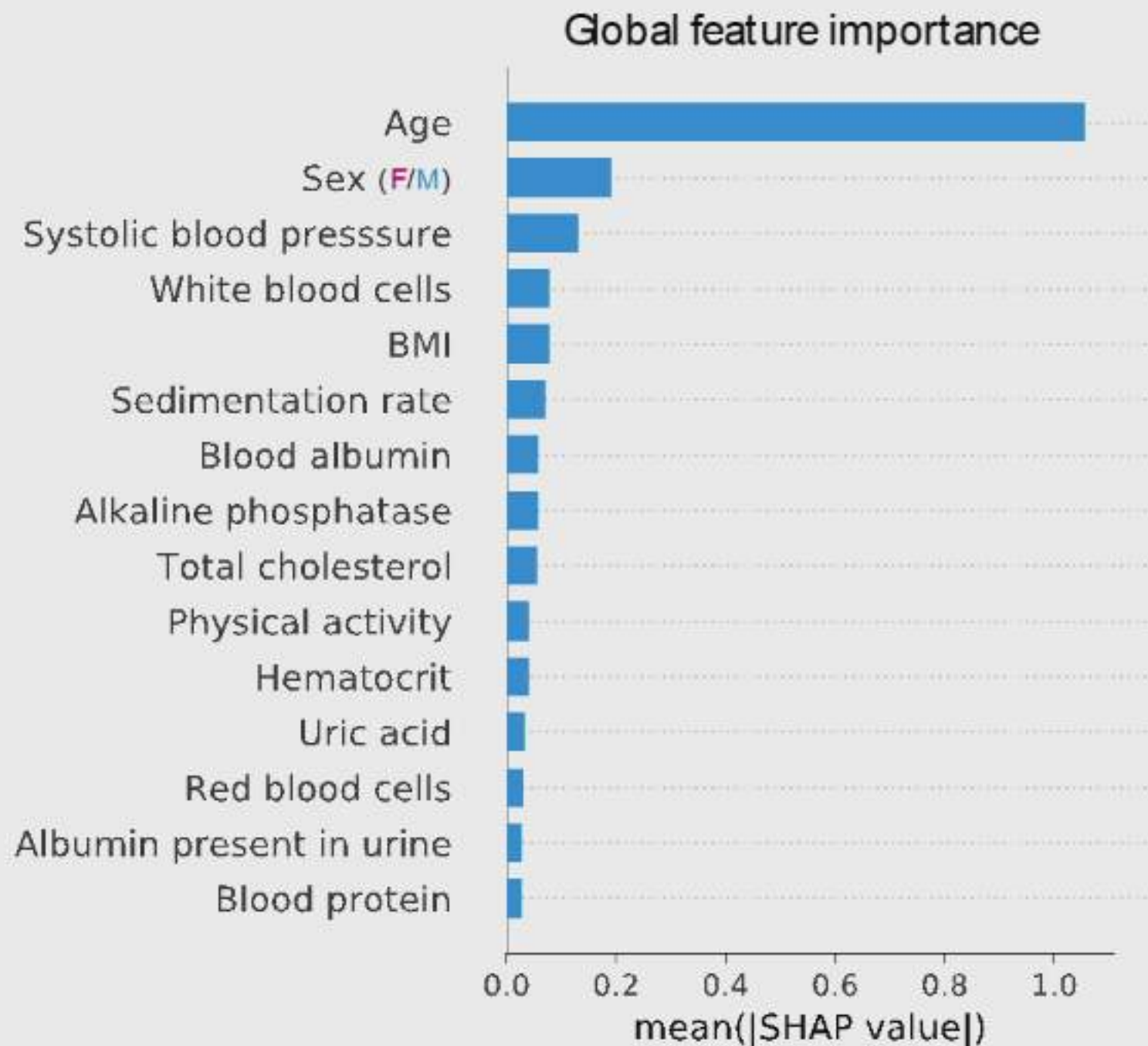


Mortality risk model



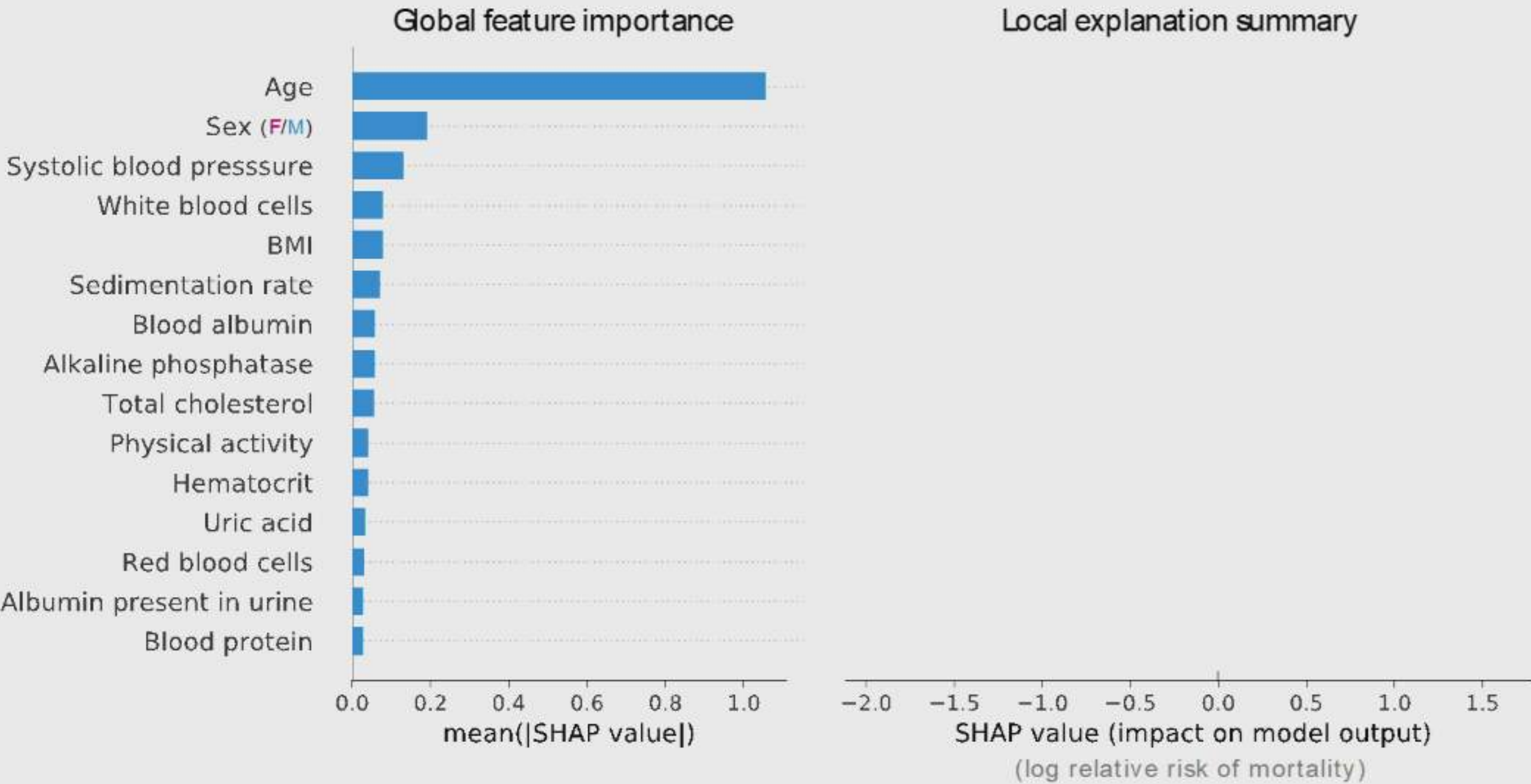
Conflates the prevalence of an effect with the magnitude of an effect

Reveal rare high-magnitude mortality effects

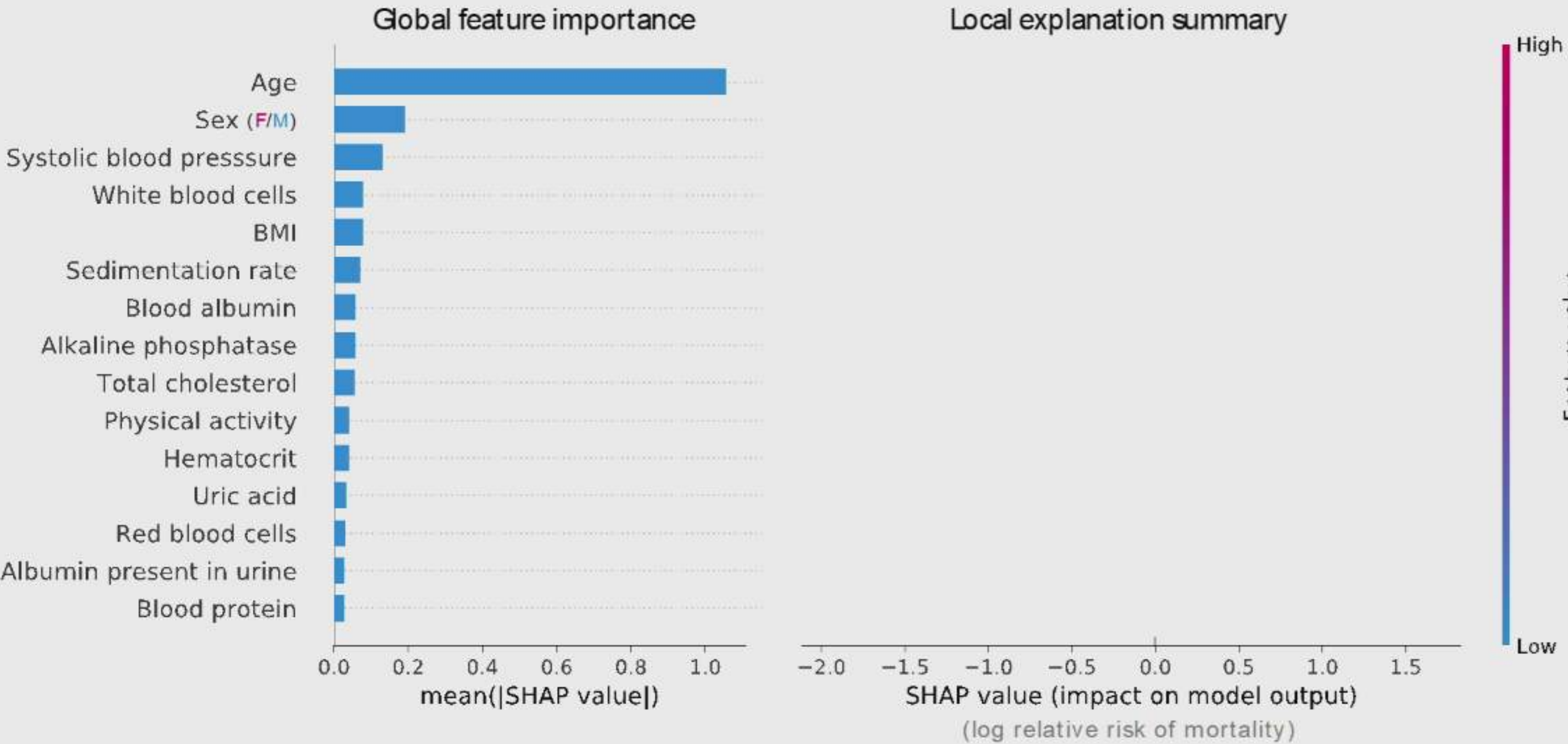


Conflates the
prevalence of an effect
with the
magnitude of an effect

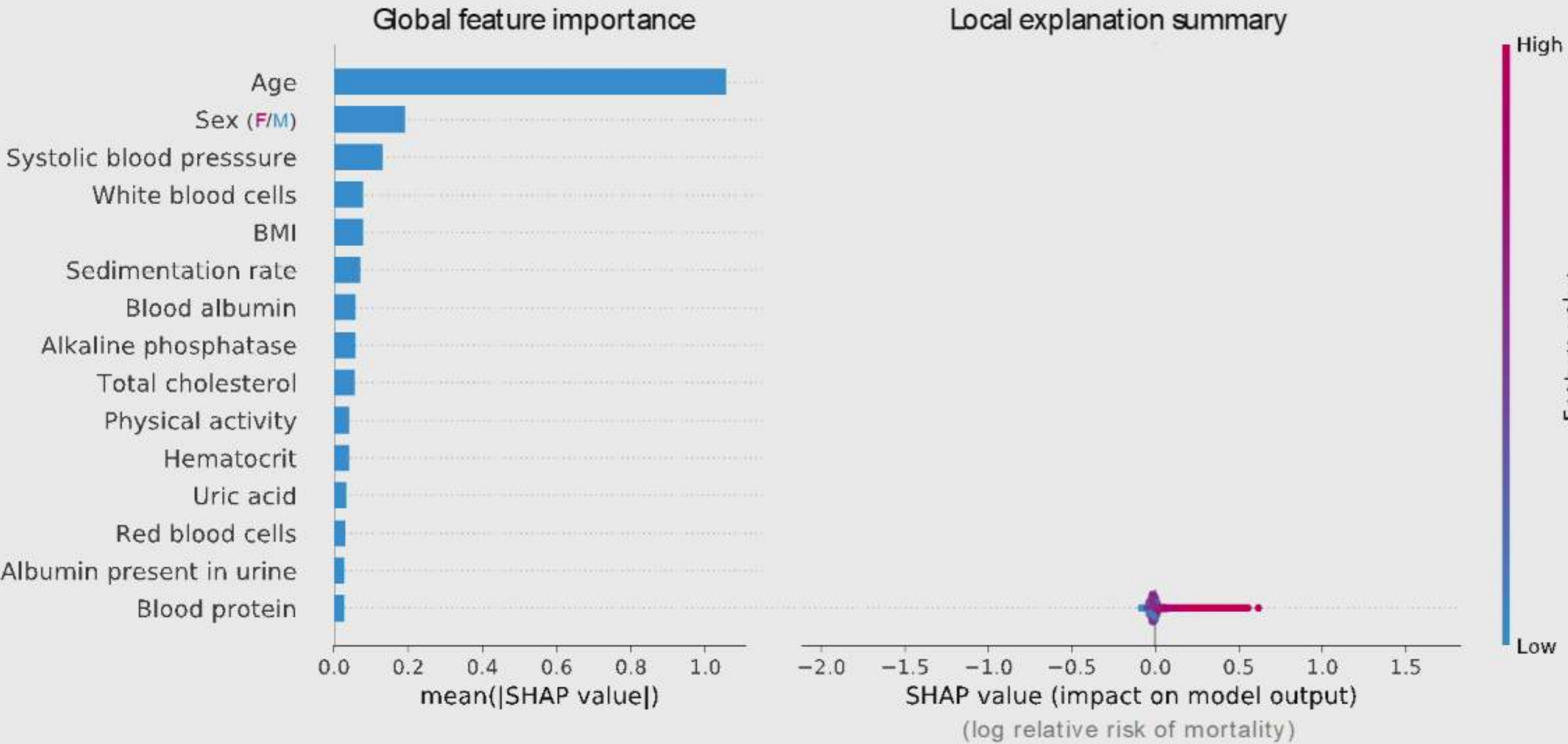
Reveal rare high-magnitude mortality effects



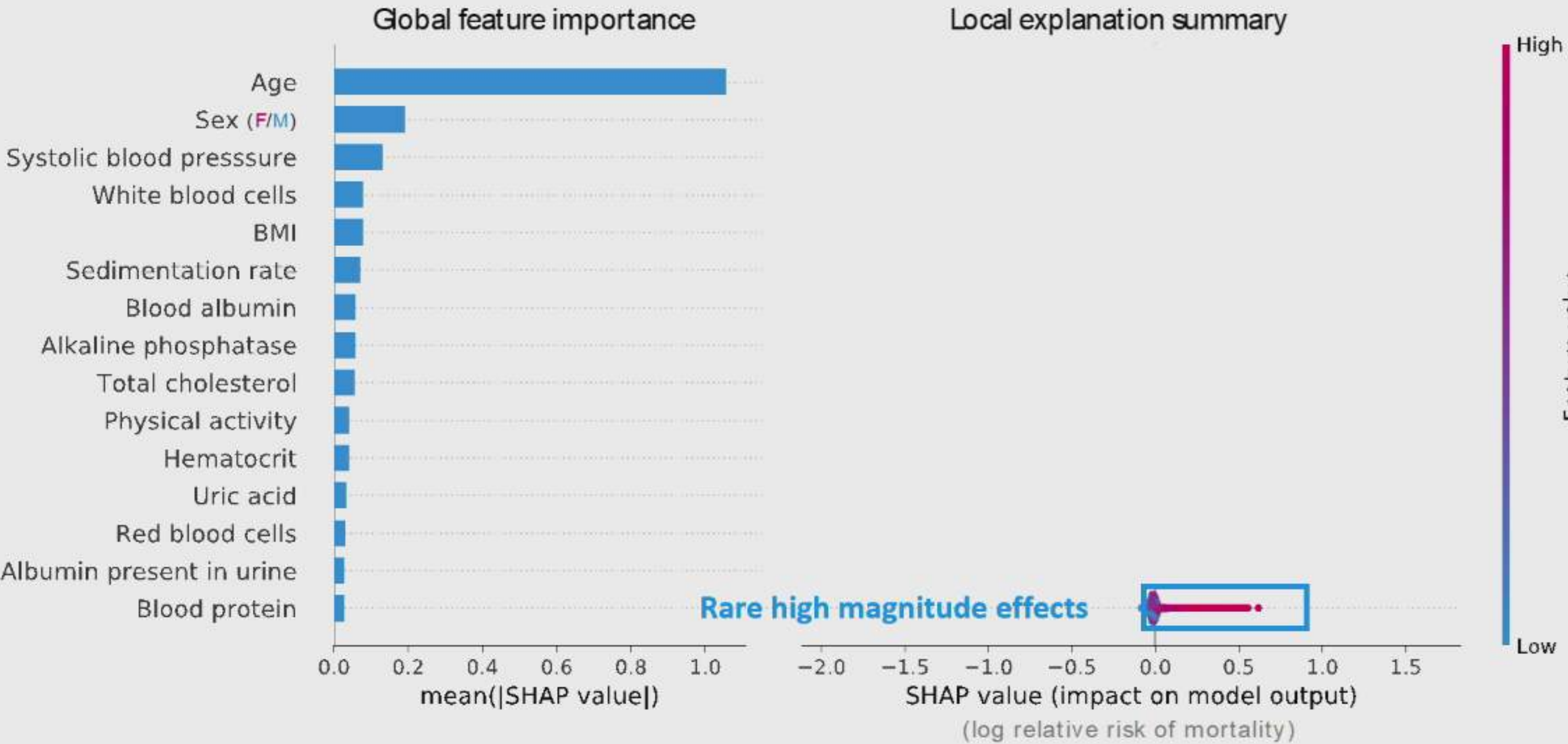
Reveal rare high-magnitude mortality effects



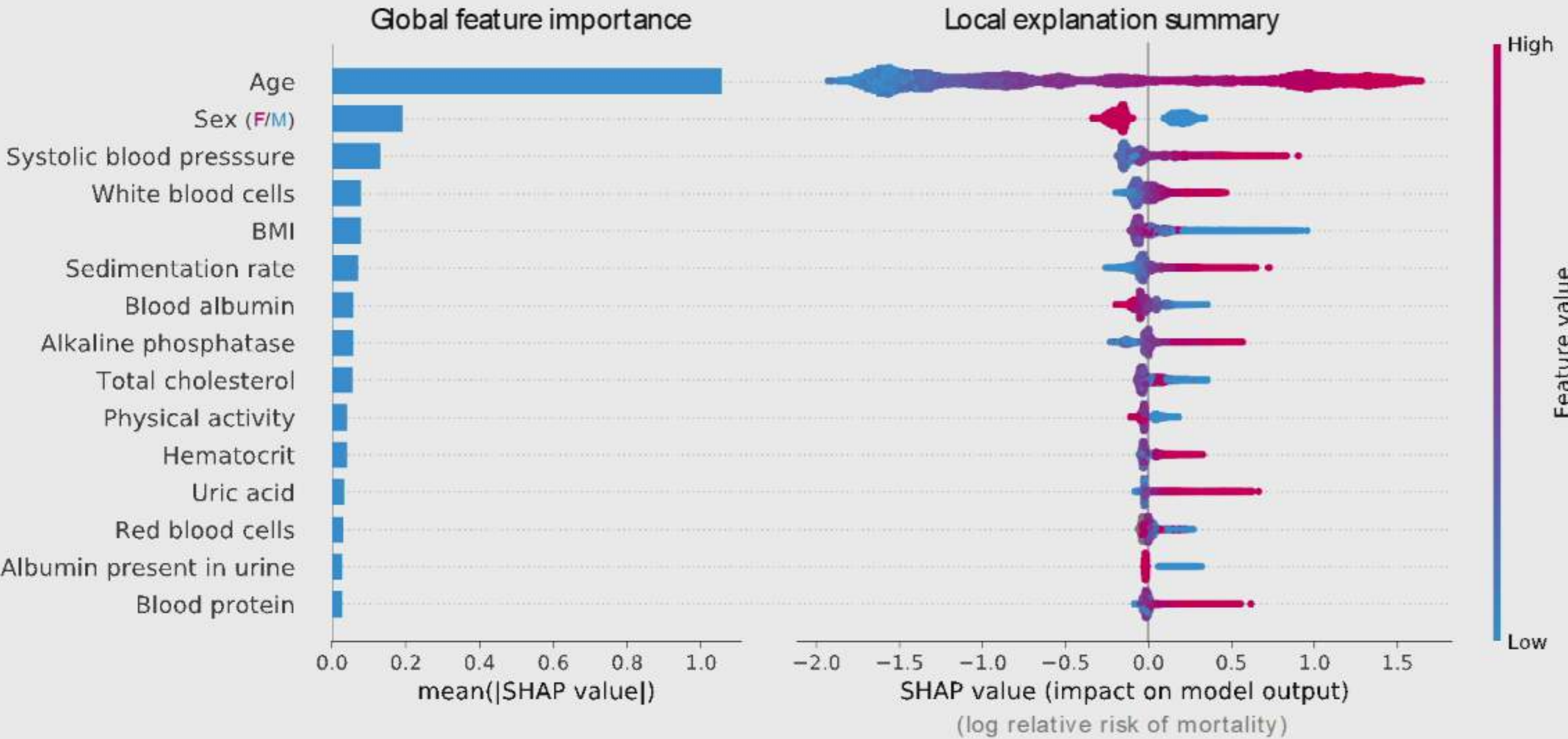
Reveal rare high-magnitude mortality effects



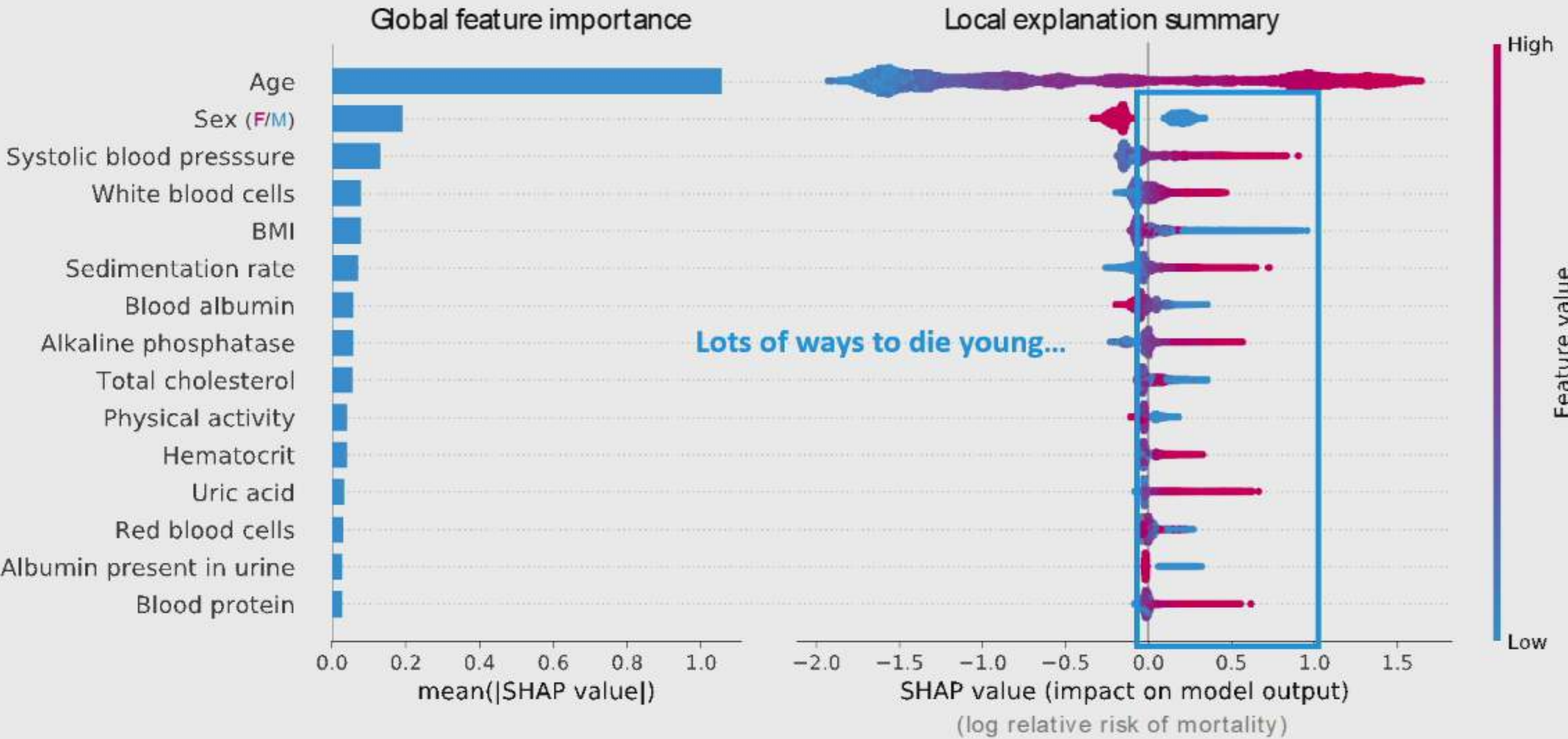
Reveal rare high-magnitude mortality effects



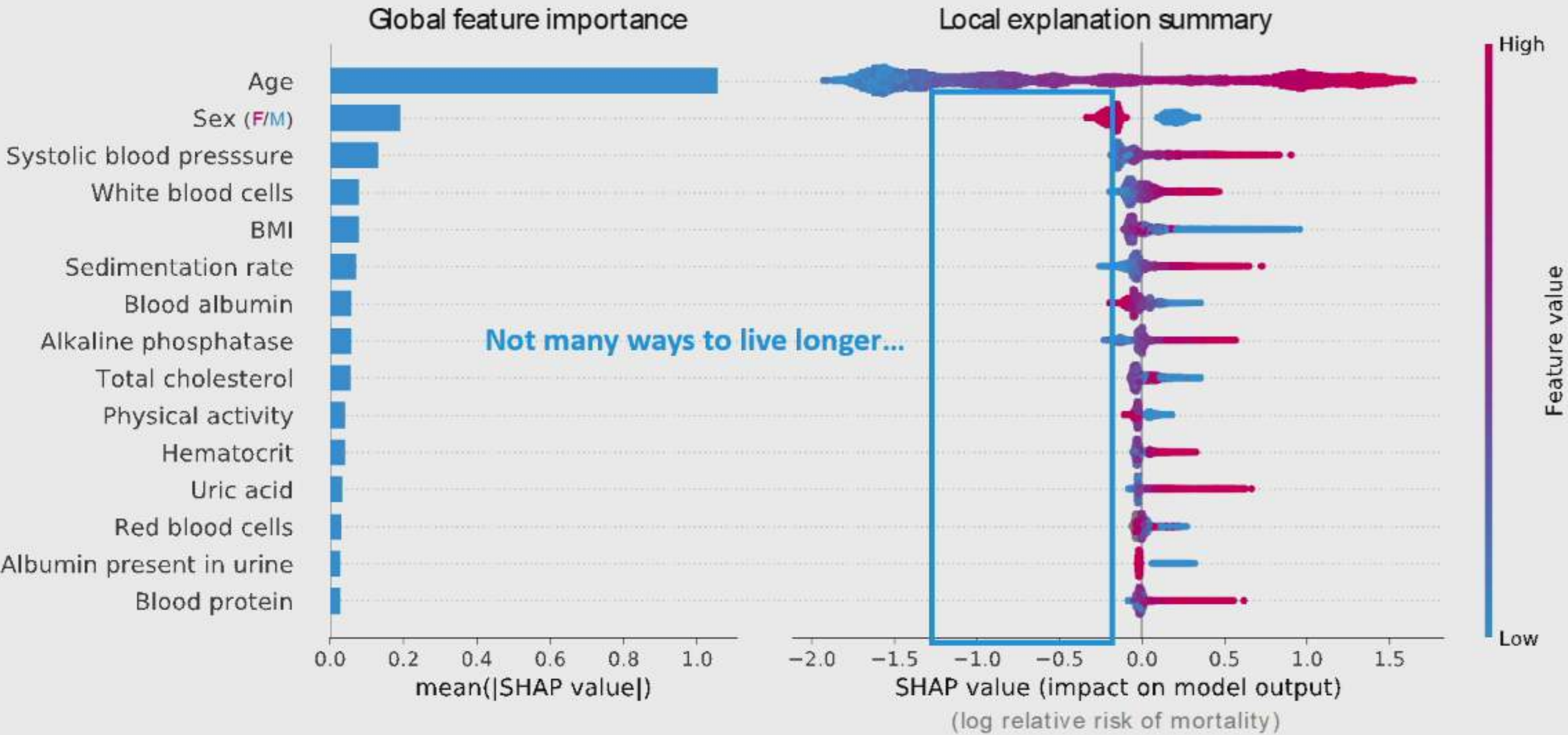
Reveal rare high-magnitude mortality effects



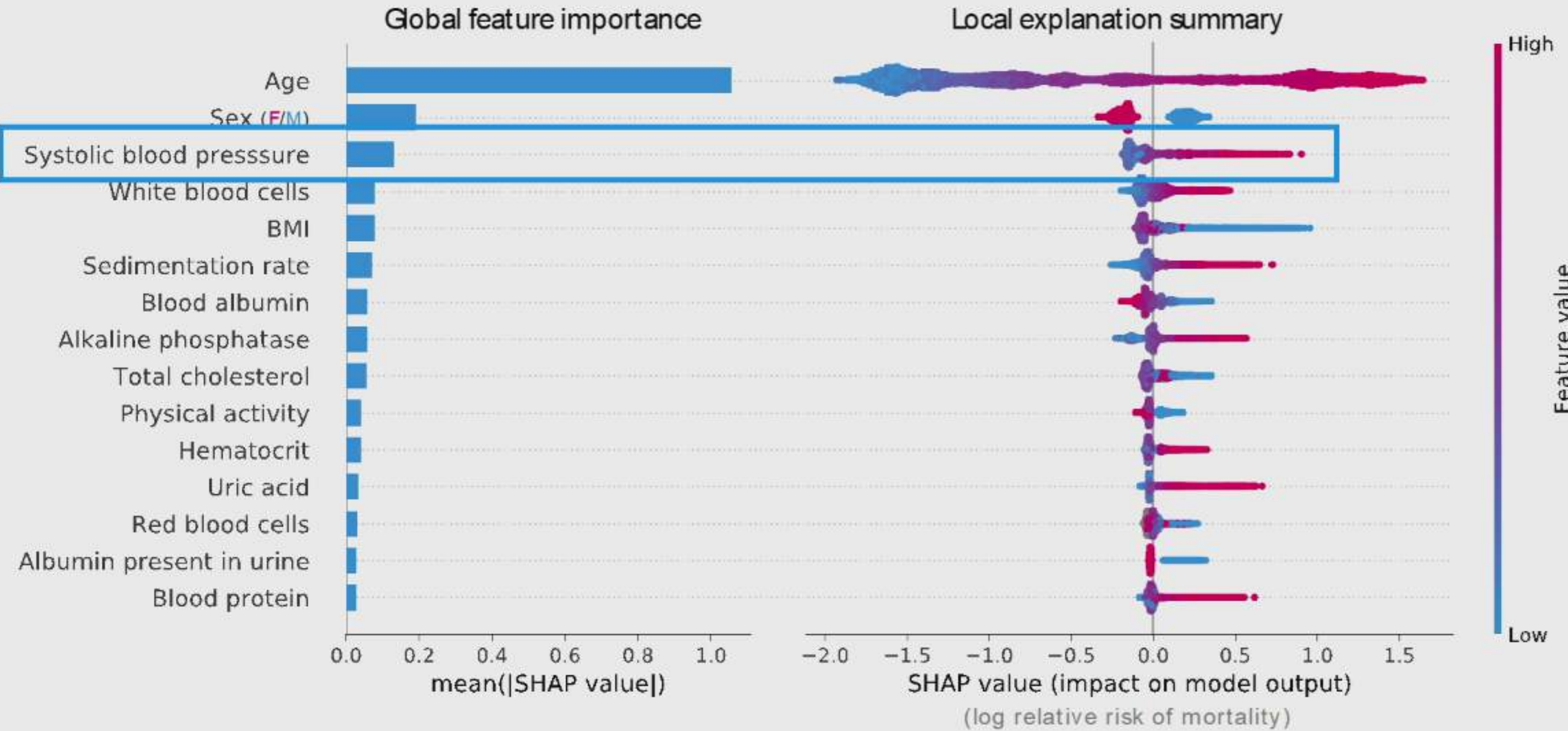
Reveal rare high-magnitude mortality effects



Reveal rare high-magnitude mortality effects

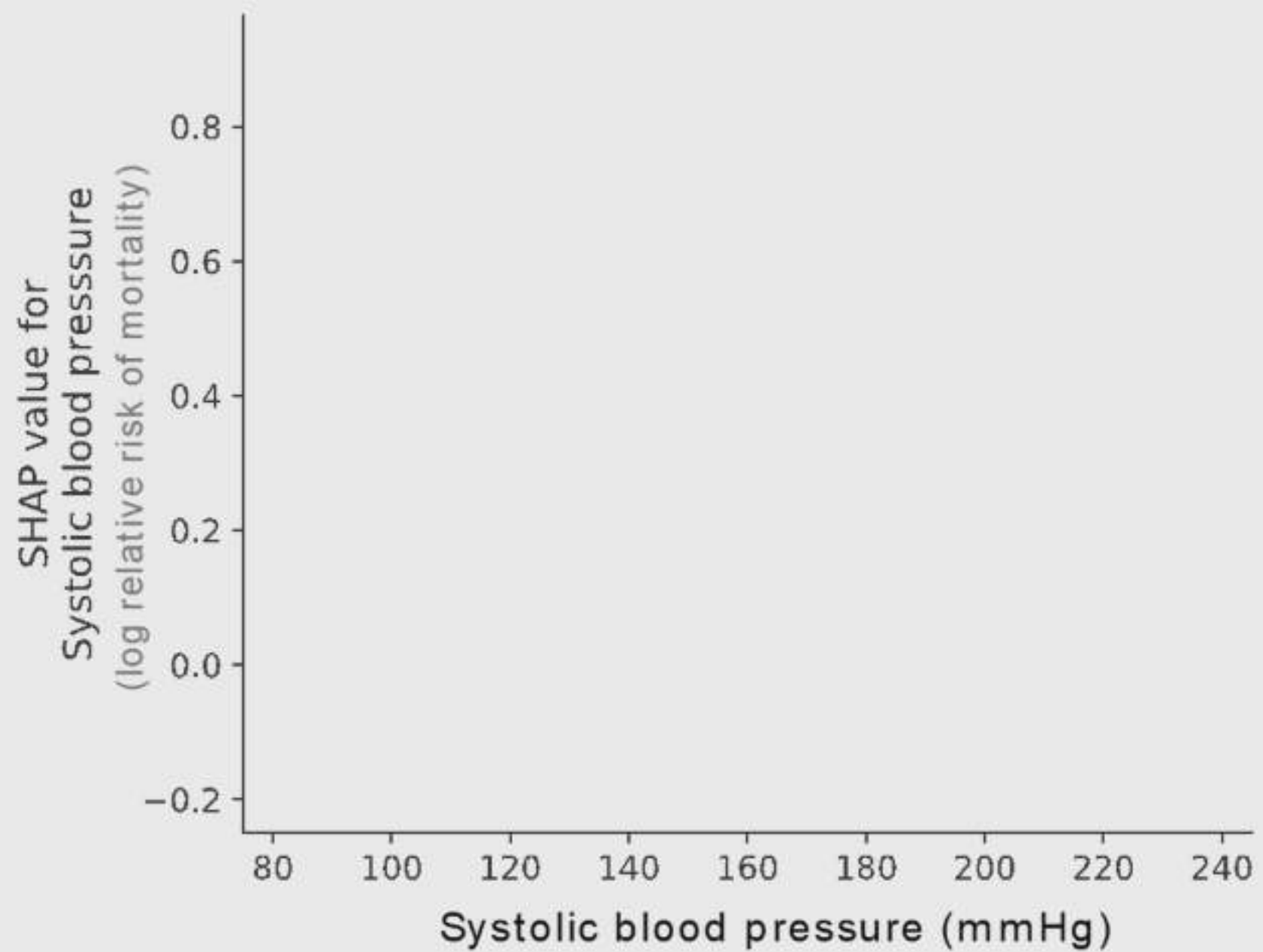


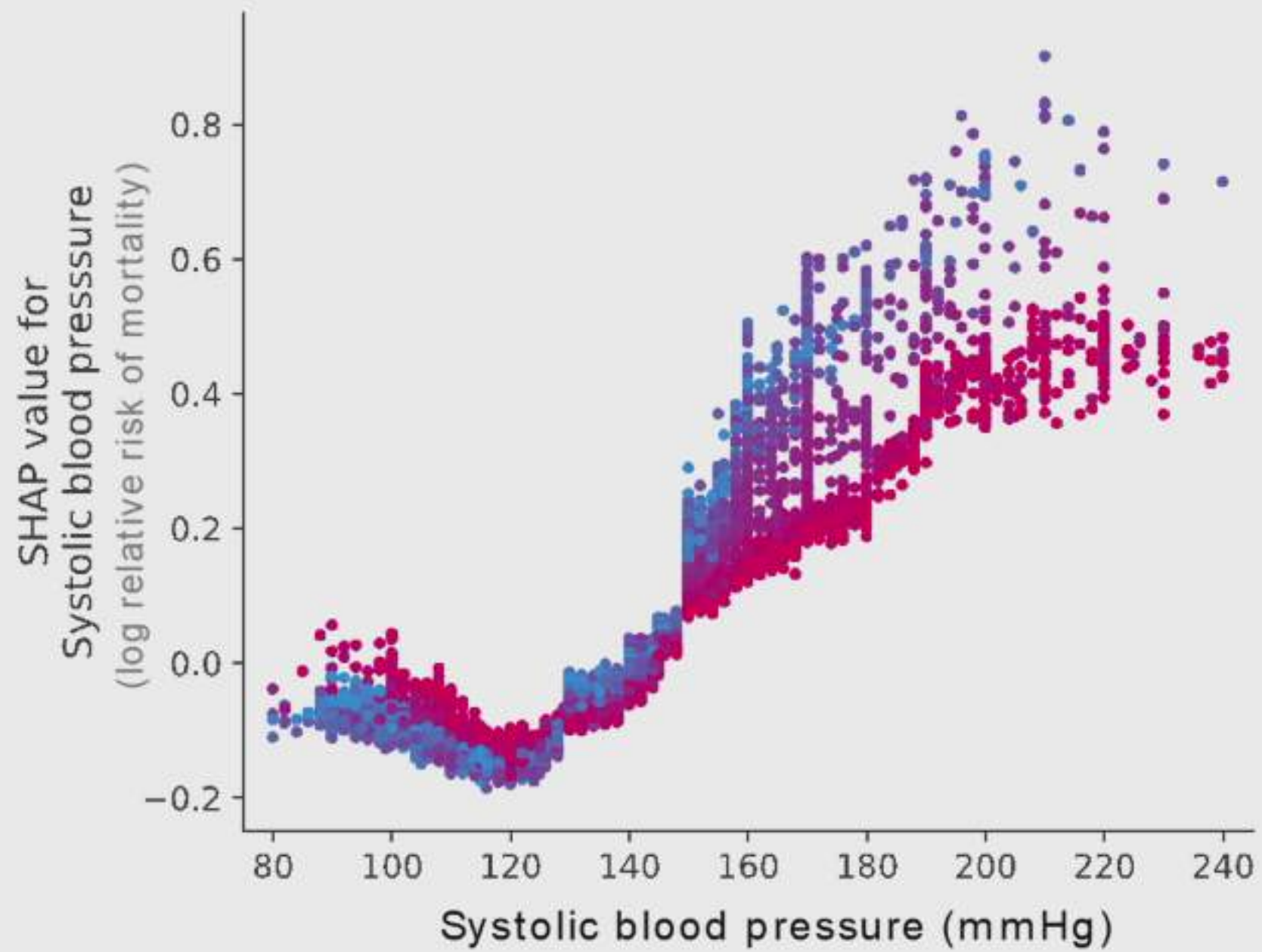
Reveal rare high-magnitude mortality effects

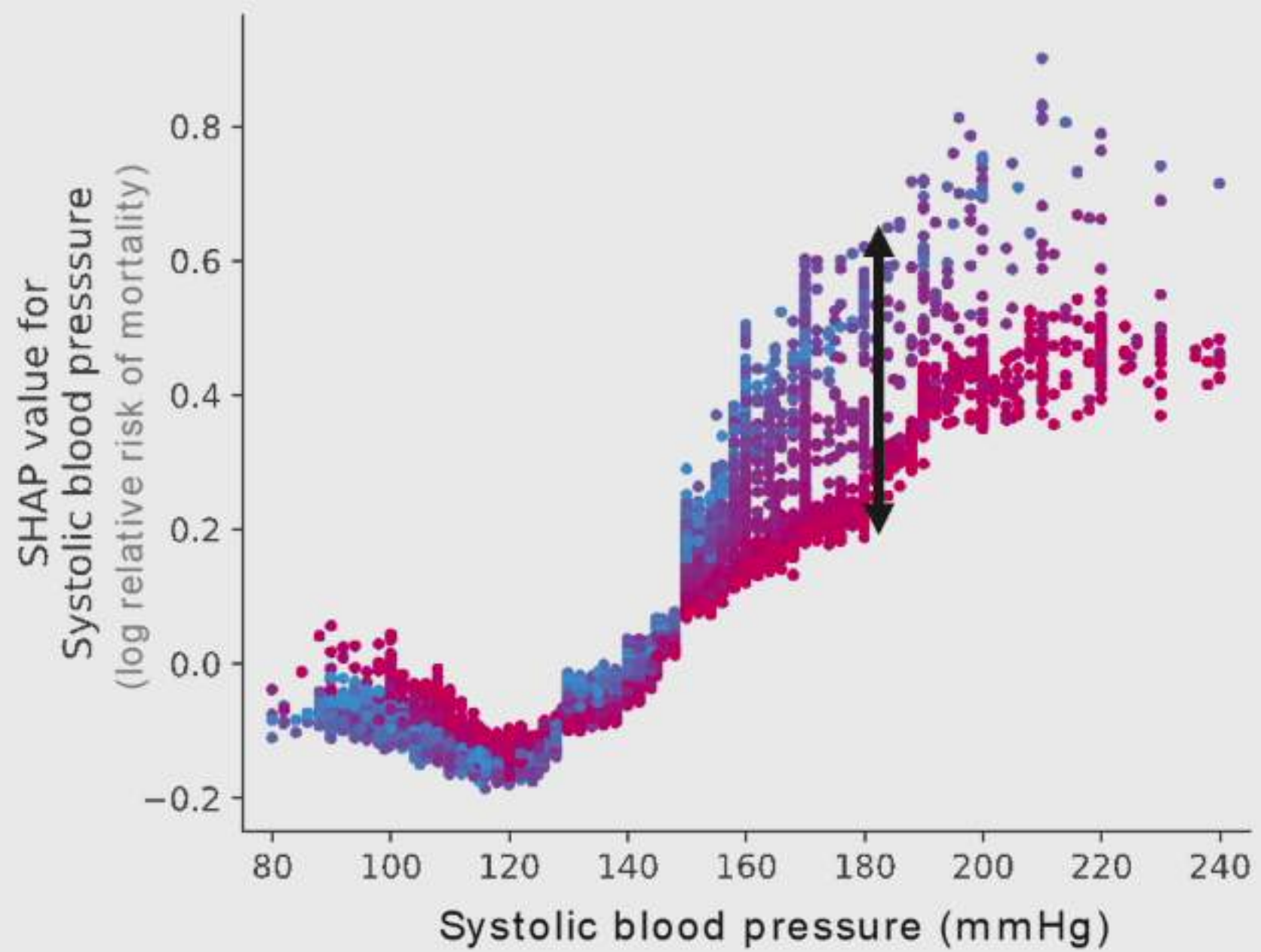


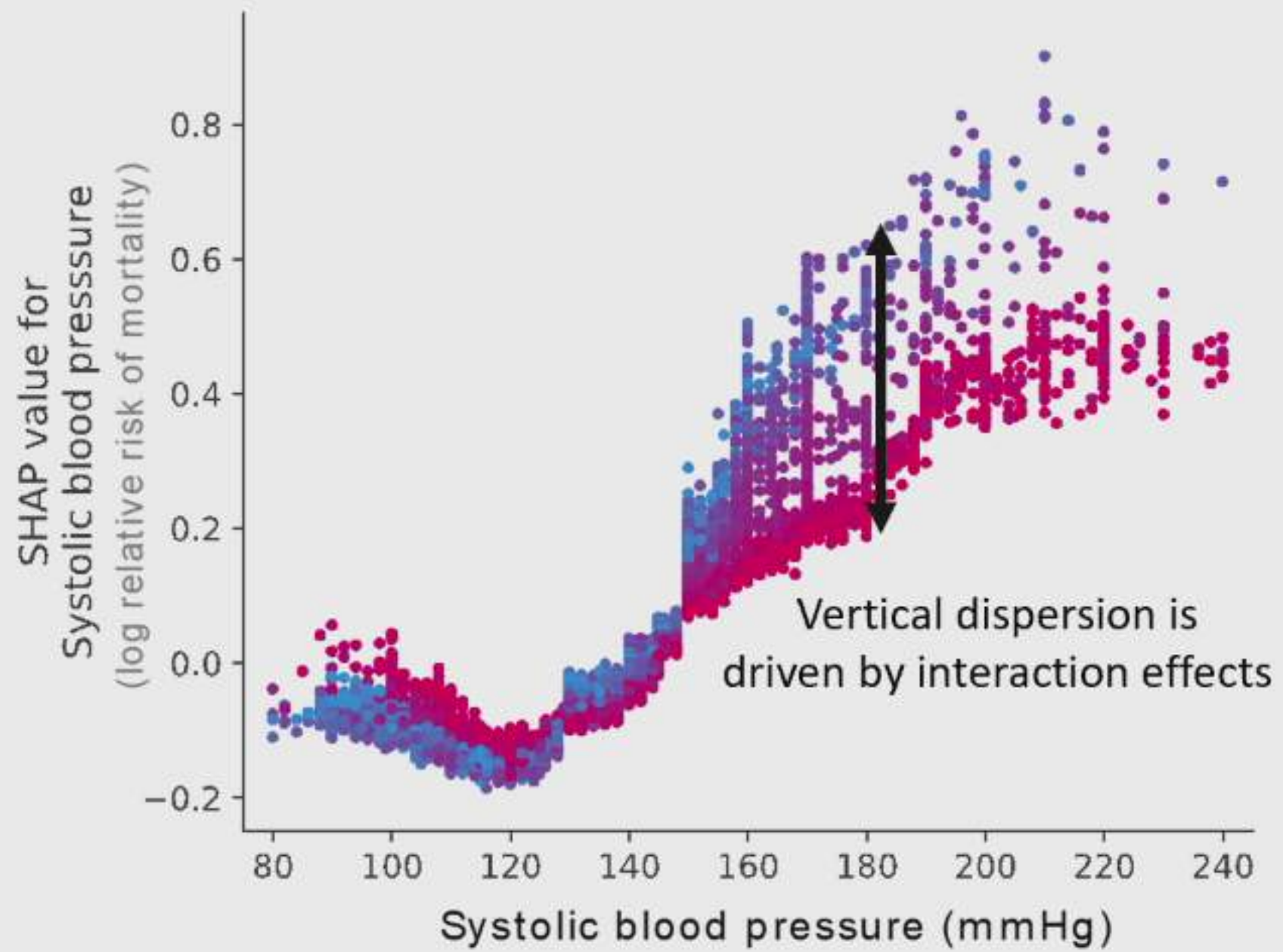
80 100 120 140 160 180 200 220 240

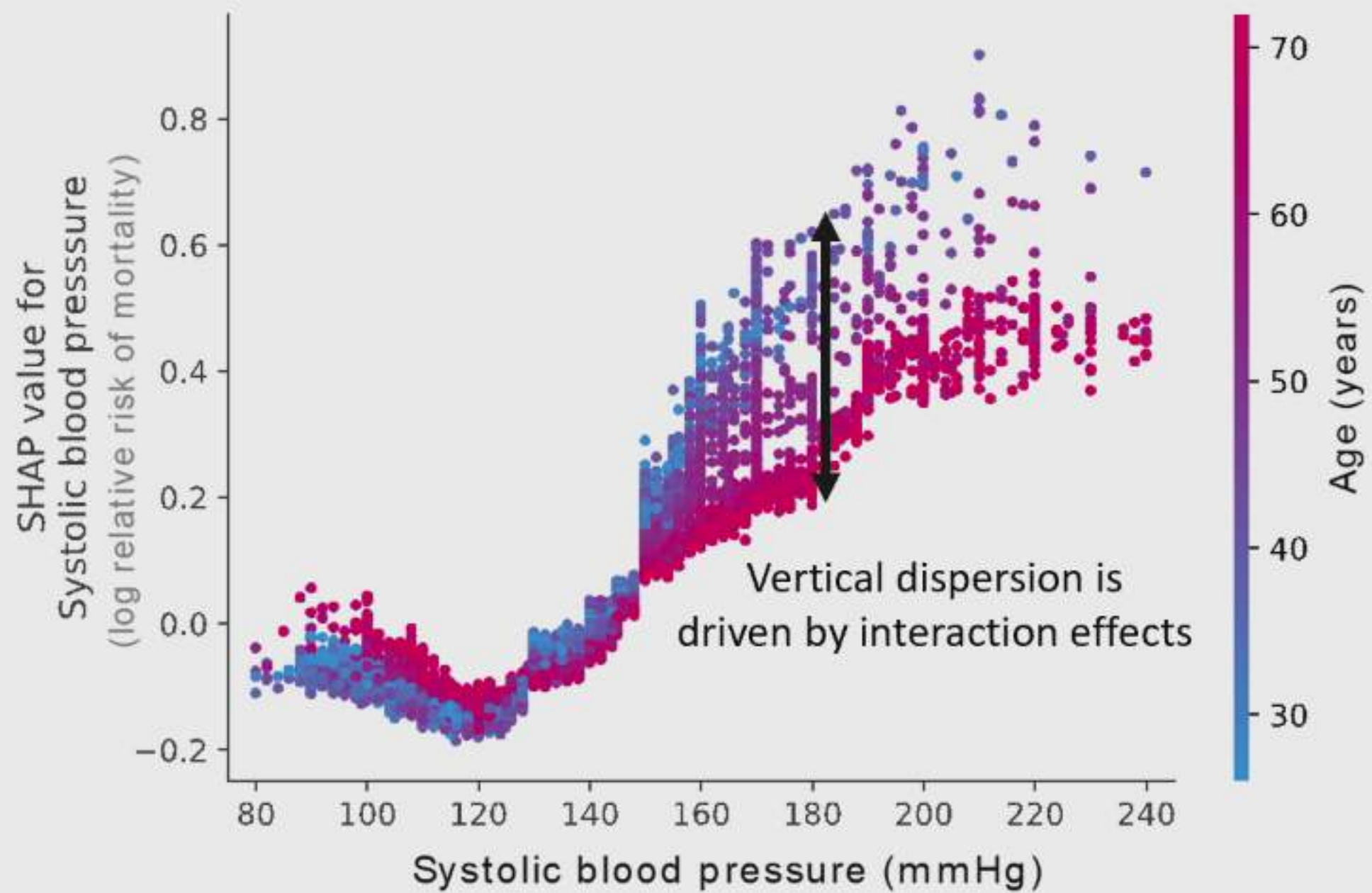
Systolic blood pressure (mmHg)



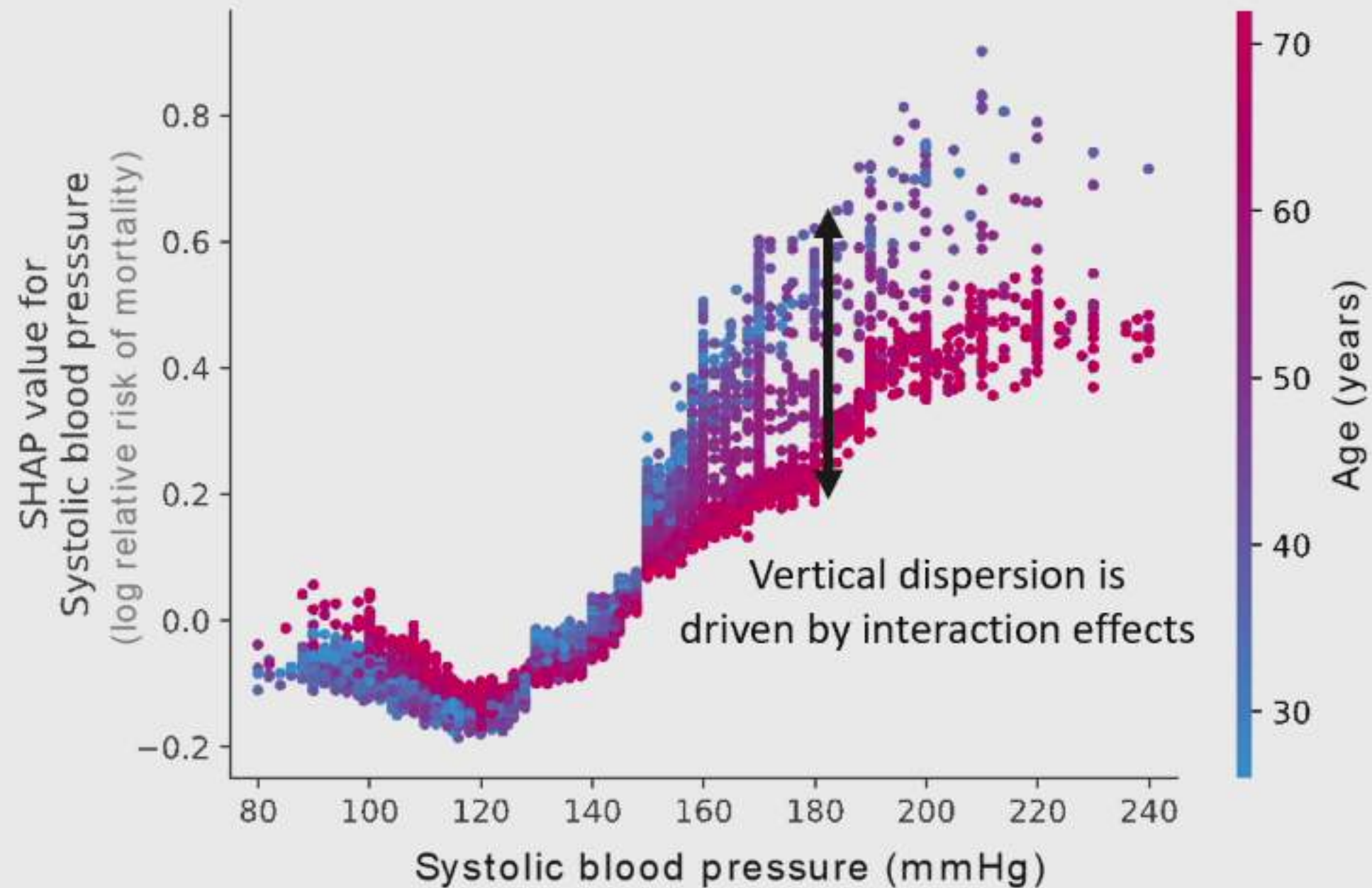




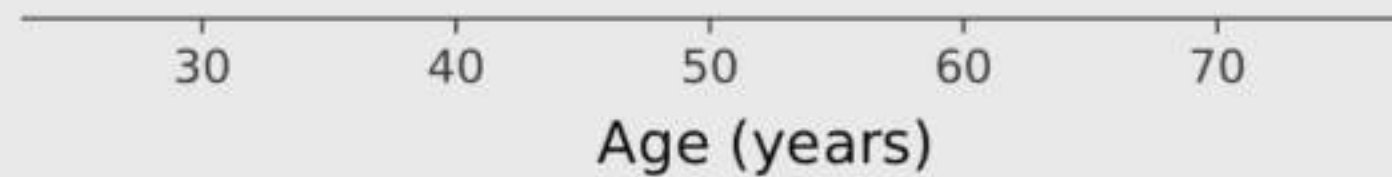




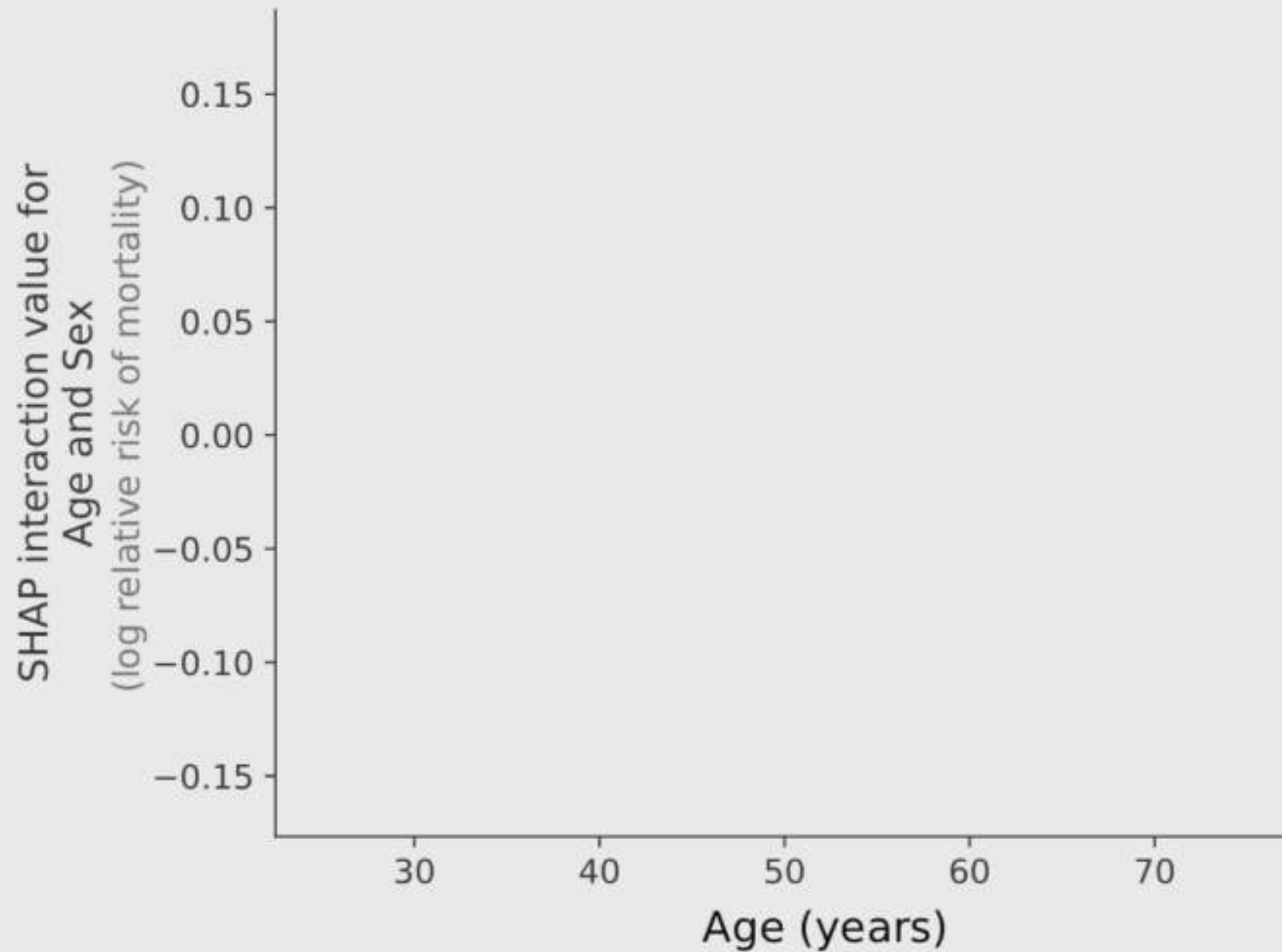
Dependence plots reveal the increased danger of early onset high blood pressure



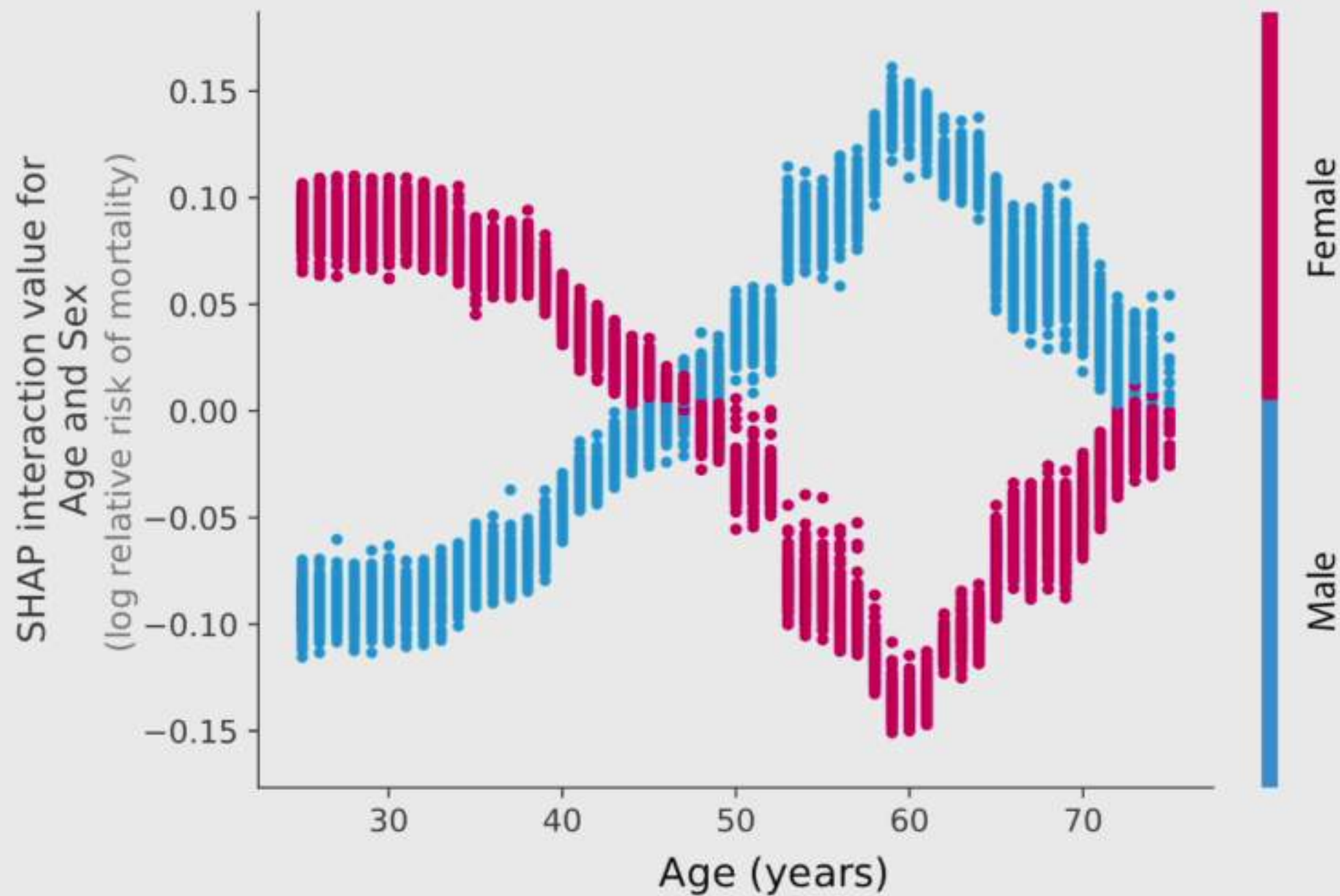
The varying risk of sex over a lifetime



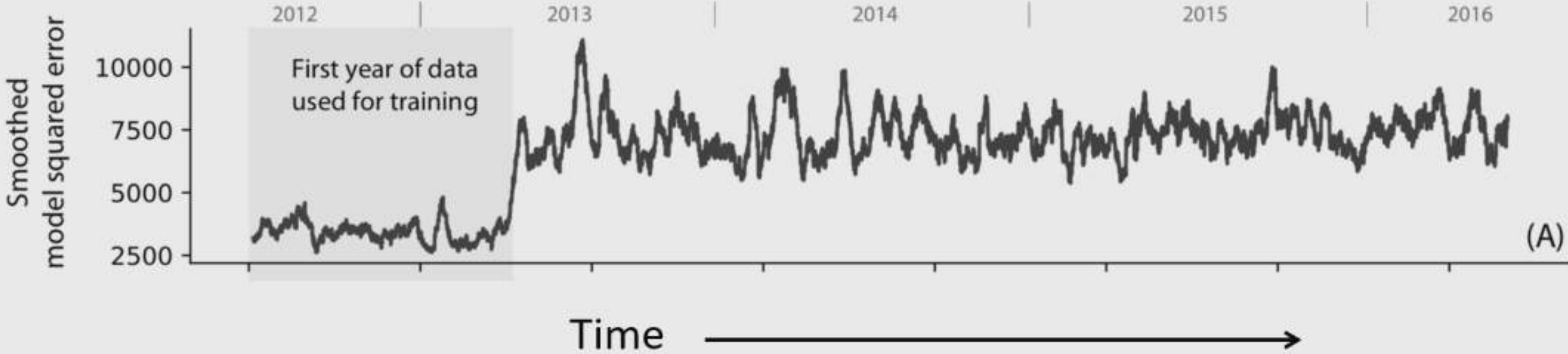
The varying risk of sex over a lifetime



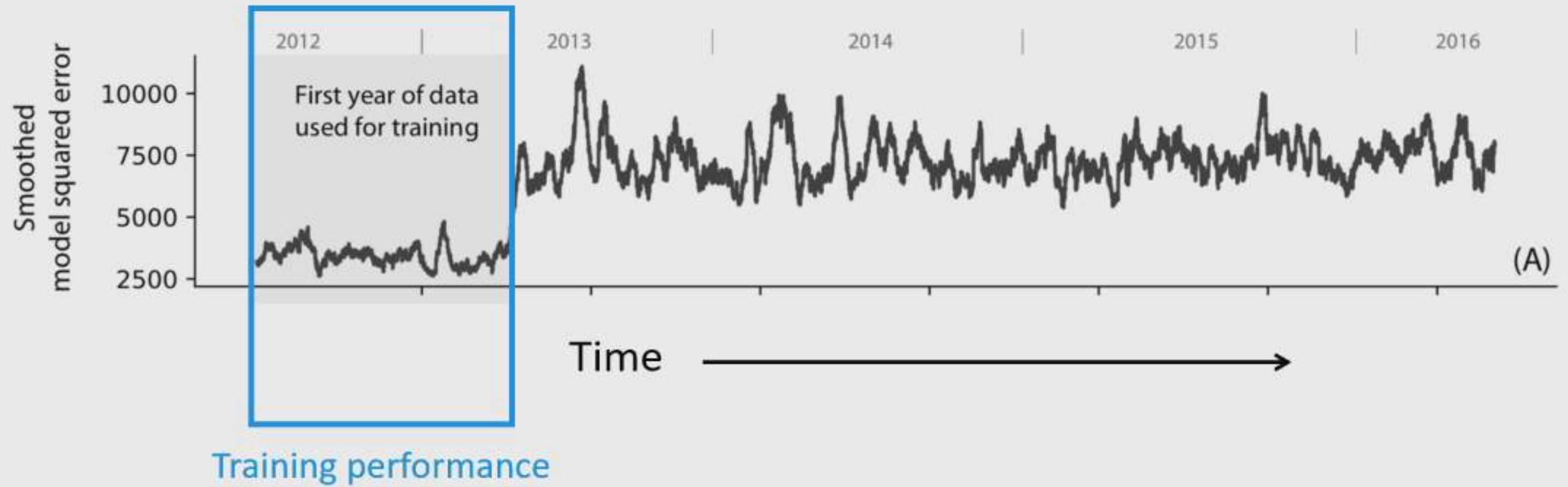
The varying risk of sex over a lifetime



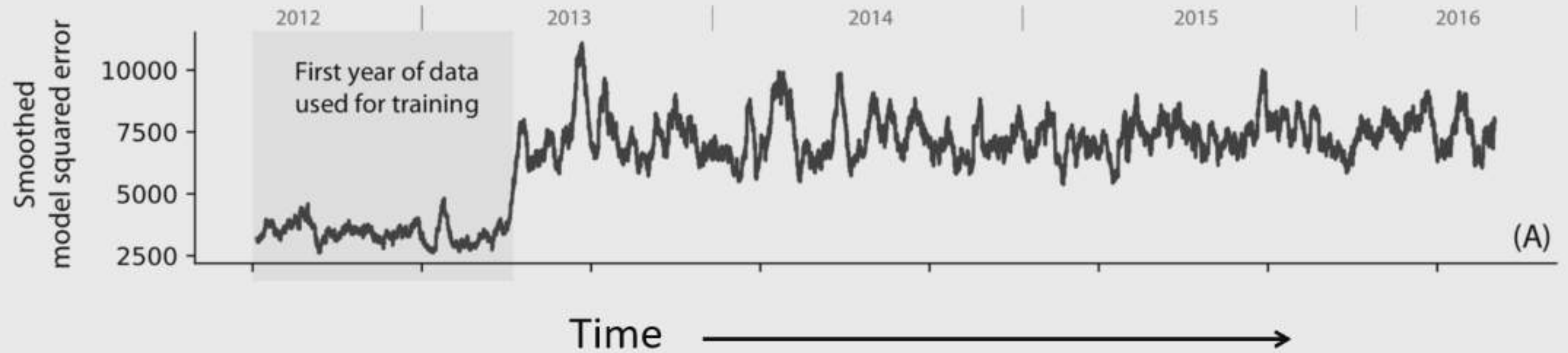
Model monitoring



Model monitoring

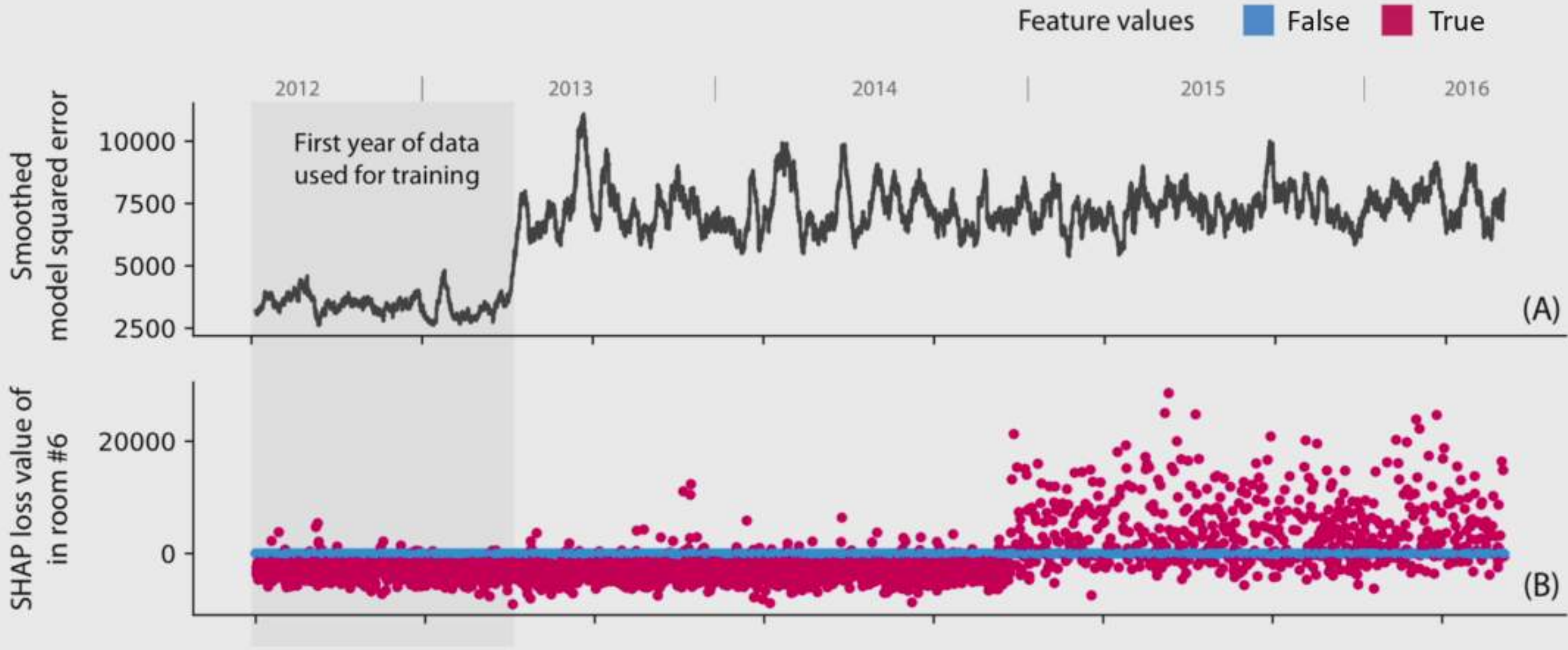


Model monitoring

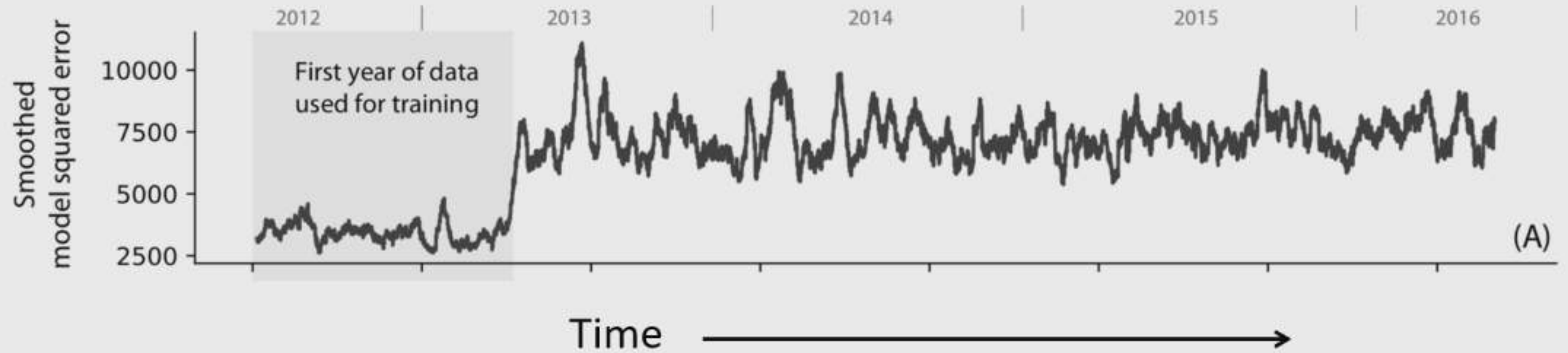


Can you find where we introduced the bug?

Model monitoring

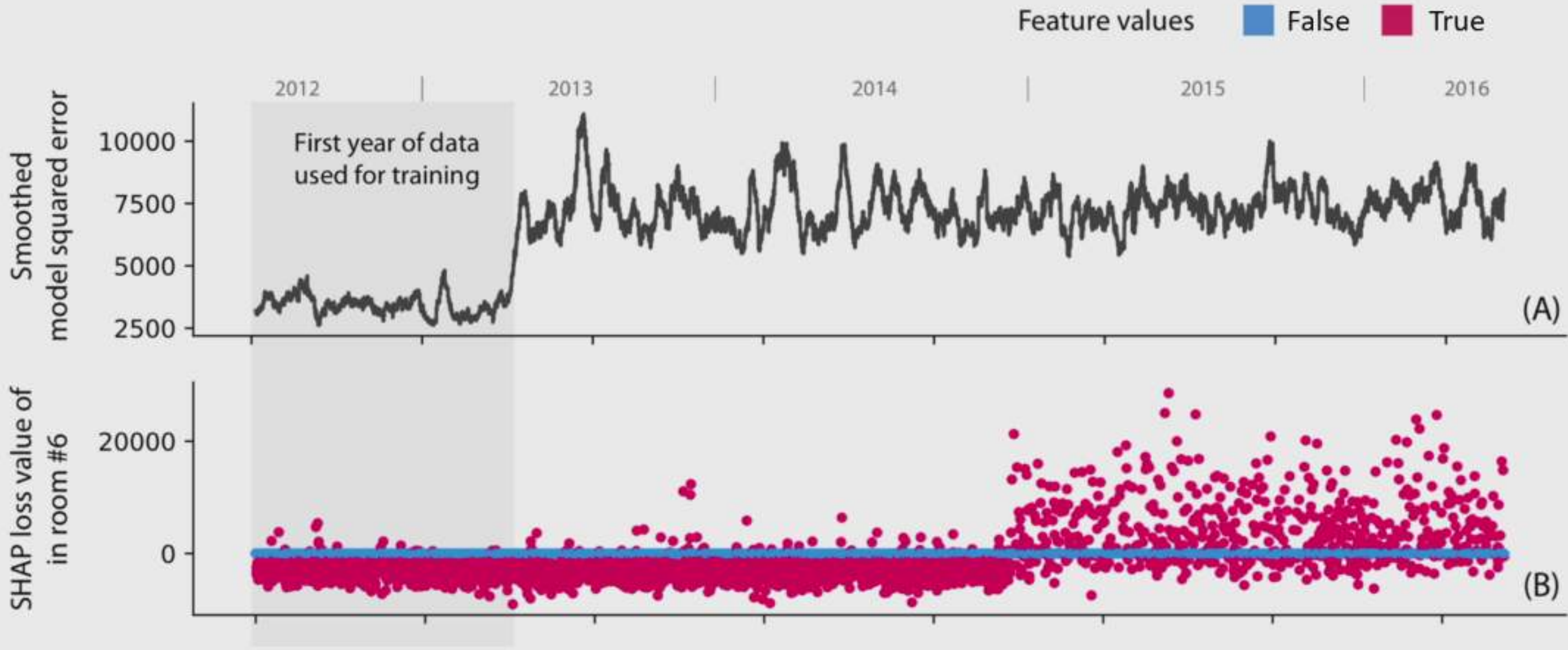


Model monitoring

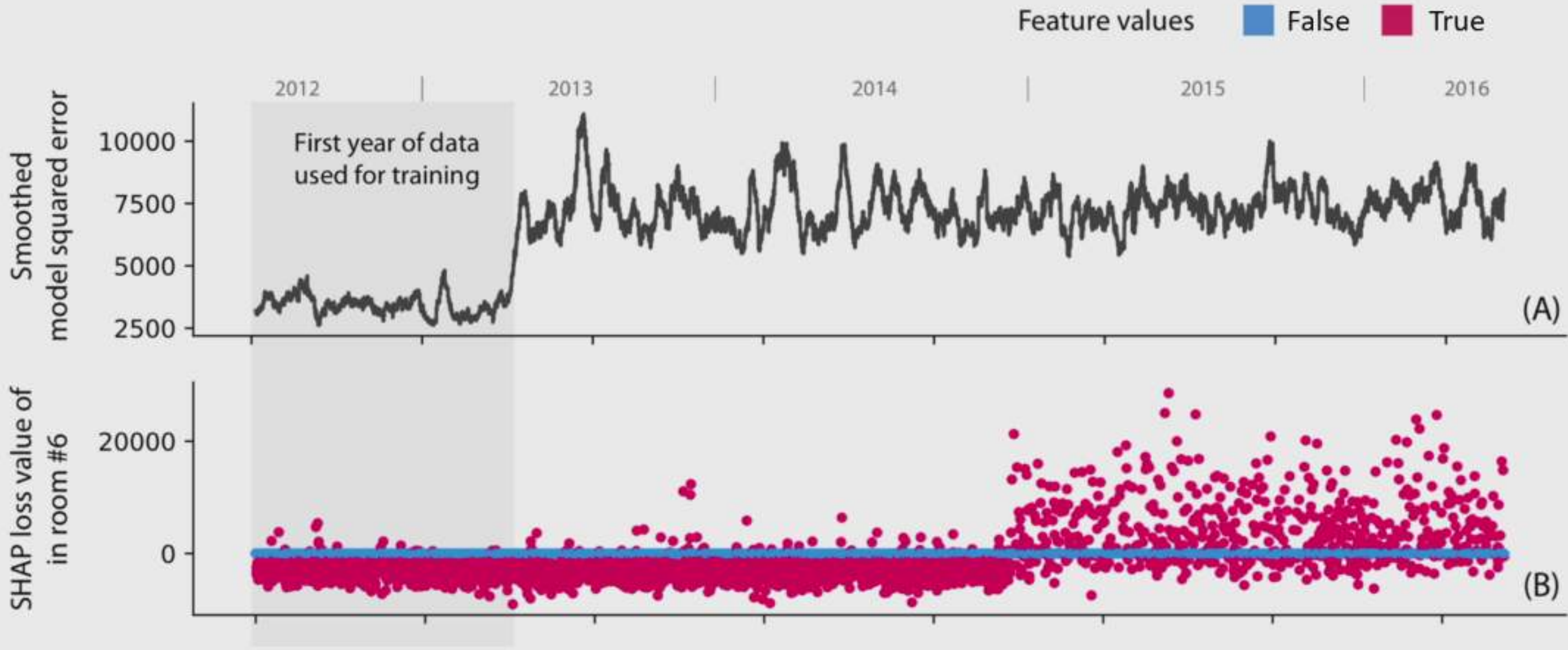


Can you find where we introduced the bug?

Model monitoring

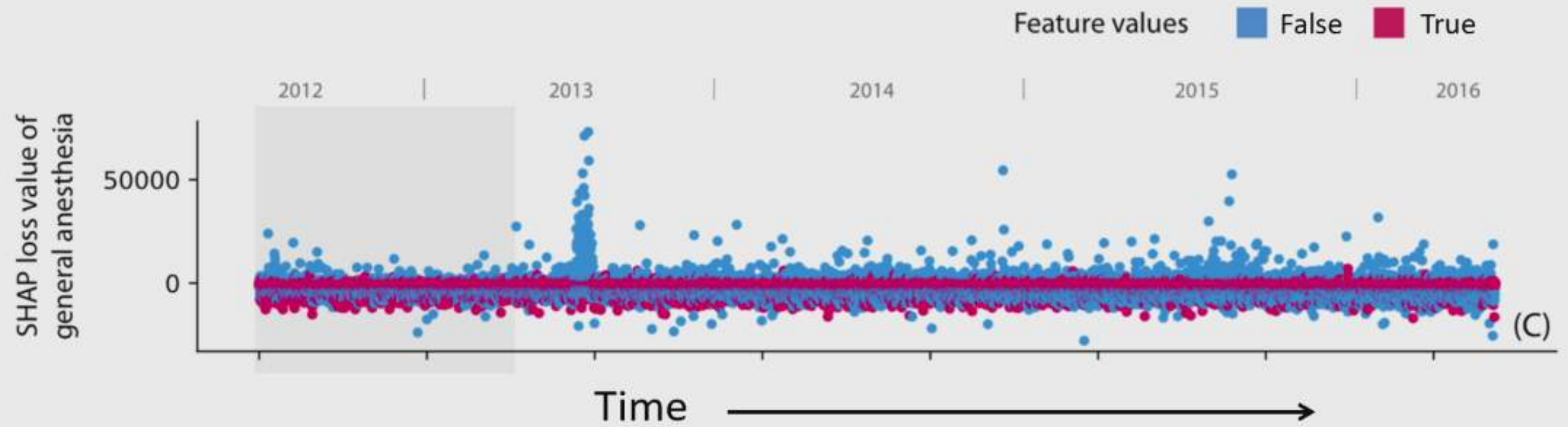


Model monitoring

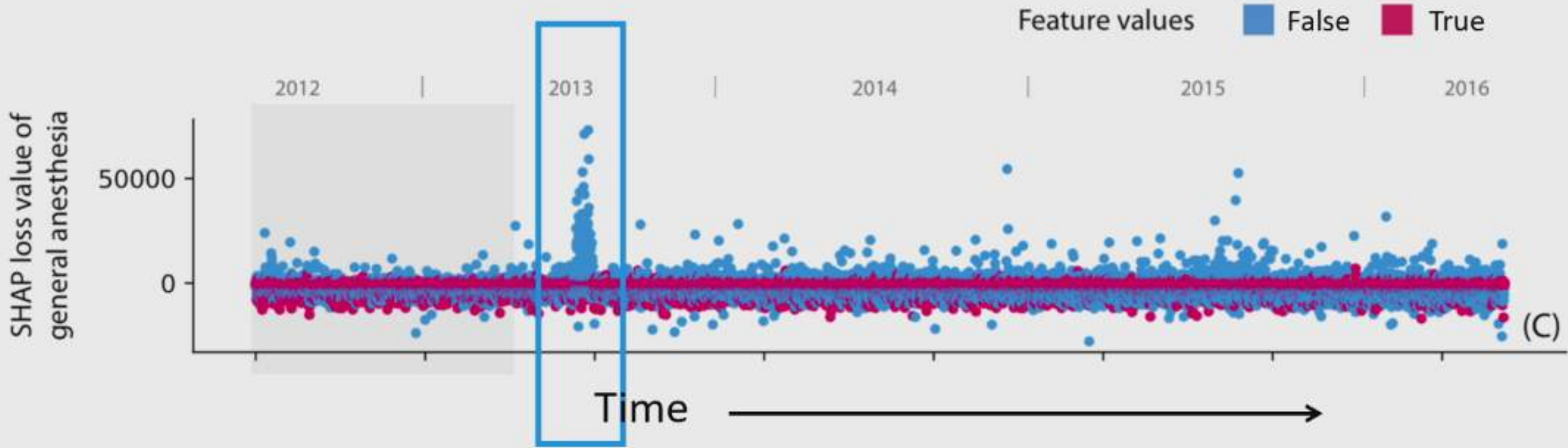


Now can you find where we introduced the bug?

Model monitoring

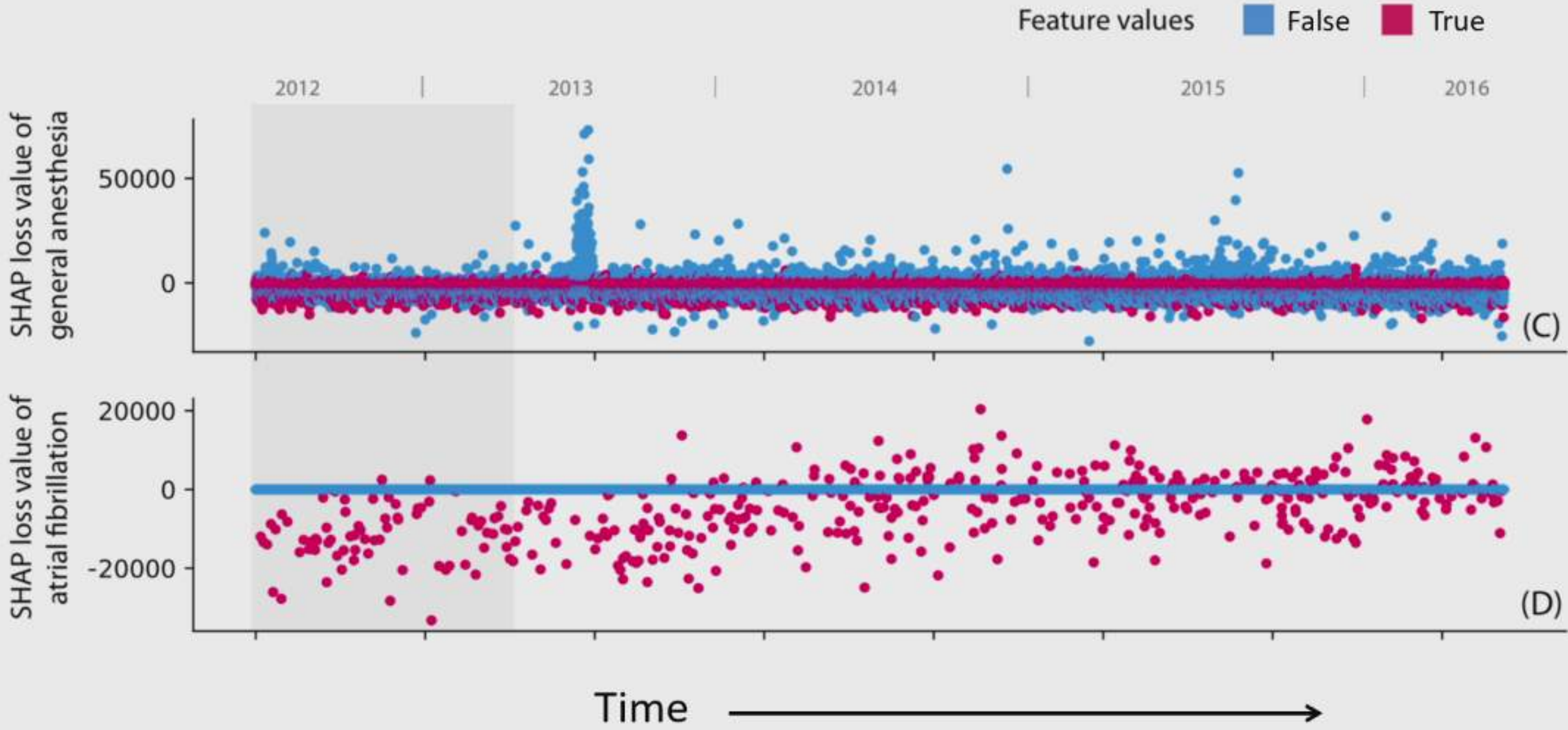


Model monitoring

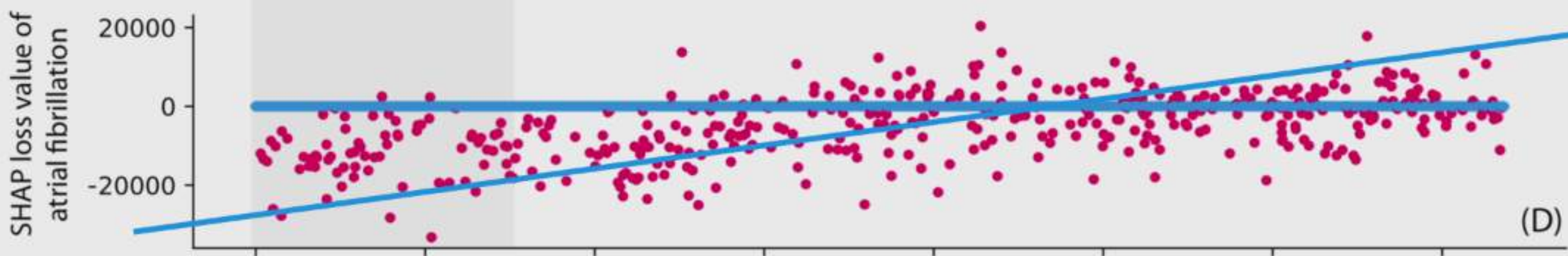
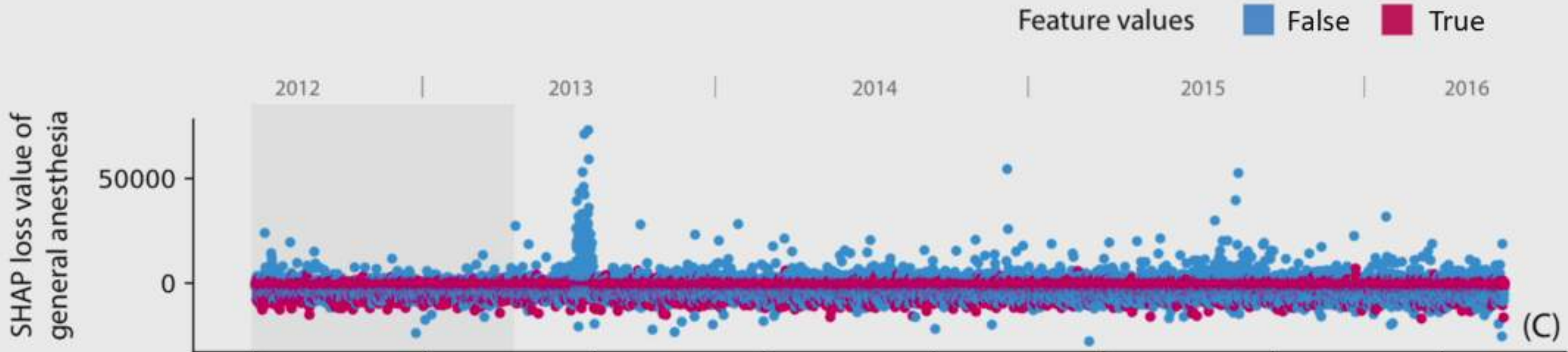


Transient electronic medical record

Model monitoring



Model monitoring



Gradual change in atrial fibrillation ablation procedure durations

Time →

Explainable AI for Science and Medicine

Theory



Unification of explanation methods



Strong uniqueness results

Practice



New estimation methods for the classic Shapley values



Explainable AI tools

Application



Anesthesia safety



Mortality risk
Hospital scheduling

Don't take my word for it, try it yourself 😊

github.com/slundberg/shap

Don't take my word for it, try it yourself 😊



```
import lightgbm as lgb
```



github.com/slundberg/shap

Don't take my word for it, try it yourself 😊

dmlc
XGBoost

```
import lightgbm as lgb
```

 CatBoost

 scikit
learn

github.com/slundberg/shap


TensorFlow

 Keras

 PyTorch

Don't take my word for it, try it yourself 😊

github.com/slundberg/shap

Don't take my word for it, try it yourself 😊



github.com/slundberg/shap

Don't take my word for it, try it yourself 😊



github.com/slundberg/shap

Don't take my word for it, try it yourself 😊



github.com/slundberg/shap

Don't take my word for it, try it yourself 😊



github.com/slundberg/shap

Don't take my word for it, try it yourself 😊



github.com/slundberg/shap



Don't take my word for it, try it yourself 😊



github.com/slundberg/shap



Don't take my word for it, try it yourself 😊



github.com/slundberg/shap



BANK OF ENGLAND

Don't take my word for it, try it yourself 😊



github.com/slundberg/shap



BANK OF ENGLAND



Future Work

Theory

Practice

Application

Future Work

Theory

Exploring fundamental interpretability tradeoffs in the presence of correlated features

Practice

Application

Future Work

Theory

Exploring fundamental interpretability tradeoffs in the presence of correlated features

Using explanation constraints to guide model training

Practice

Application

Future Work

Theory

Exploring fundamental interpretability tradeoffs in the presence of correlated features

Using explanation constraints to guide model training

Practice

Efficient and general model monitoring tools

Application

Future Work

Theory

Exploring fundamental interpretability tradeoffs in the presence of correlated features

Using explanation constraints to guide model training

Practice

Efficient and general model monitoring tools

Integrating causal modeling assumptions to enhance the interpretability of feature attributions

Application

Future Work

Theory

Exploring fundamental interpretability tradeoffs in the presence of correlated features

Using explanation constraints to guide model training

Practice

Efficient and general model monitoring tools

Integrating causal modeling assumptions to enhance the interpretability of feature attributions

Application

In-the-loop high-stakes decision making

Future Work

Theory

Exploring fundamental interpretability tradeoffs in the presence of correlated features

Using explanation constraints to guide model training

Practice

Efficient and general model monitoring tools

Integrating causal modeling assumptions to enhance the interpretability of feature attributions

Application

In-the-loop high-stakes decision making

Understanding adverse drug interactions /genomics/proteins

Future Work

Theory

Exploring fundamental interpretability tradeoffs in the presence of correlated features

Using explanation constraints to guide model training

Practice

Efficient and general model monitoring tools

Integrating causal modeling assumptions to enhance the interpretability of feature attributions

Application

In-the-loop high-stakes decision making

Understanding adverse drug interactions /genomics/proteins

Augmented Intelligence for Finance



Su-In Lee

Collaborations



Hugh Chen



Pascal Sturmfels



Su-In Lee



Alex Okeson



Nao Hiranuma

Collaborations



Hugh Chen



Pascal Sturmfels



Su-In Lee



Ruqian Chen

UW Math



Alex Okeson



Nao Hiranuma

Collaborations



Hugh Chen



Pascal Sturmfels



Su-In Lee



Alex Okeson



Nao Hiranuma



Joe Janizek



Gabe Erion



Alex DeGrave



Ruqian Chen

UW Math

Collaborations



Hugh Chen



Pascal Sturmfels



Su-In Lee



Ruqian Chen



Alex Okeson



Nao Hiranuma



Joe Janizek



Gabe Erion



Alex DeGrave

Collaborations

Hiranuma, Lundberg, and Lee. AIControl: Replacing matched control experiments with machine learning improves ChIP-seq peak identification. Nucleic Acids Research, 2019


PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING


 Hugh Chen



 Pascal Sturmfels



 Su-In Lee


UW MSTP
Medical Scientist Training Program




Collaborations

Chen, Lundberg, Lee. Hybrid Gradient Boosting Trees and Neural Networks for Forecasting Operating Room Data. NeurIPS Workshop ML4H: Machine Learning for Health, 2017.


 Ruqian Chen
UW Math


 Alex Okeson


 Nao Hiranuma




 Joe Janizek Gabe Erion Alex DeGrave

Collaborations



Ruqian Chen



Alex Okeson



Nao Hiranuma



Joe Janizek



Gabe Erion



Alex DeGrave



Hugh Chen



Pascal Sturmfels



Su-In Lee

UW MSTP

Erion, Chen, Lundberg, Lee. Anesthesiologist-level forecasting of hypoxemia with only SpO2 data using deep learning. NeurIPS Workshop ML4H: Machine Learning for Health, 2017.

Collaborations



Hugh Chen



Pascal Sturmfels



Su-In Lee



Ruqian Chen



Alex Okeson



Nao Hiranuma



Joe Janizek



Gabe Erion



Alex DeGrave

Estimating drug-drug interaction effects, Manuscript in preparation.



Hugh Chen



Pascal Sturmfels



Su-In Lee



Ruqian Chen



Alex Okeson



Nao Hiranuma



Joe Janizek



Gabe Erion



Alex DeGrave

Collaborations

Manuscript in under review.



Safiye Celik

Su-In Lee

Collaborations

Lee, et al. A machine learning approach to integrate big data for precision medicine in acute myeloid leukemia. Nature communications, 2018.



Ruqian Chen

UW Math



Alex Okeson



Nao Hiranuma



Joe Janizek



Gabe Erion



Alex DeGrave



Hugh Chen



Pascal Sturmfels



Su-In Lee



Alex Okeson



Nao Hiranuma



Joe Janizek



Gabe Erion



Alex DeGrave



Ruqian Chen

UW Math

Collaborations



Hugh Chen



Pascal Sturmfels



Su-In Lee



Ruqian Chen

UW Math



Alex Okeson



Nao Hiranuma



Joe Janizek



Gabe Erion



Alex DeGrave

Collaborations

Anesthesiology & Pain Medicine



Monica
Vavilala



Bala Nair



Jerry Kim



Hugh Chen



Pascal Sturmfels



Su-In Lee



Ruqian Chen

UW Math



Alex Okeson



Nao Hiranuma



Joe Janizek



Gabe Erion



Alex DeGrave

Collaborations

Anesthesiology & Pain Medicine



Monica Vavilala



Bala Nair



Jerry Kim

Kidney Research Institute



Jonathan Himmelfarb



Nisha Bansal



Ronit Katz

Collaborations

Cardiology



Jordan Prutkin



Ruqian Chen

UW Math

W PAUL G. ALLEN SCHOOL OF COMPUTER SCIENCE & ENGINEERING



Hugh Chen



Pascal Sturmfels



Su-In Lee



Nao Hiranuma



Joe Janizek



Gabe Erion



Alex DeGrave

Anesthesiology & Pain Medicine



Monica Vavilala



Bala Nair



Jerry Kim

Kidney Research Institute



Jonathan Himmelfarb



Nisha Bansal



Ronit Katz

Cardiology



Jordan Prutkin



Ruqian Chen

UW Math

W PAUL G. ALLEN SCHOOL OF COMPUTER SCIENCE & ENGINEERING



Hugh Chen



Pascal Sturmfels



Su-In Lee



Nao Hiranuma



Joe Janizek



Gabe Erion



Alex DeGrave



Alex Okeson

Anesthesiology & Pain Medicine



Monica Vavilala



Bala Nair



Jerry Kim

Kidney Research Institute



Jonathan Himmelfarb



Nisha Bansal



Ronit Katz

Collaborations

University of Toronto



Michael Hoffman



Linda Penn



William Tu



Brian Raught

Cardiology



Jordan Prutkin

W PAUL G. ALLEN SCHOOL OF COMPUTER SCIENCE & ENGINEERING



Hugh Chen



Pascal Sturmfels



Su-In Lee



Ruqian Chen



Alex Okeson



Nao Hiranuma



Joe Janizek



Gabe Erion



Alex DeGrave

Collaborations

University of Toronto



Michael Hoffman



Linda Penn



William Tu



Brian Raught

Lundberg et al. ChromNet: Learning the human chromatin network from all ENCODE ChIP-seq data. Genome Biology, 2016. (F1000Prime recommended)



Monica Vavilala



Bala Nair



Jerry Kim



Jonathan Himmelfarb



Nisha Bansal



Ronit Katz

Thanks!

