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
John, a bank customer

model

55%

chance John will have repayment problems
John, a bank customer

model

55%

chance John will have repayment problems

No loan
John, a bank customer

Why?!

Model

55% chance John will have repayment problems

No loan
John, a bank customer

Why?!

model

55%

chance John will have repayment problems

No loan
John, a bank customer

Why?! 

Why?! 

model

55%

chance John will have repayment problems

No loan
Why?!

John, a bank customer

Why?!

AI magic!

model

55%

chance John will have repayment problems

No loan
Interpretable Accurate

Complex model

Simple model
Interpretable  Accurate

Complex model

Simple model
Interpretable or accurate: *choose one.*
Interpretable or accurate: choose one.
Interpretable or accurate: choose one.
Interpretable or accurate: choose one.
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

\[
\sum X_i w_i \rightarrow Y
\]
Linear model

\[ \sum \]

\[ \begin{align*}
X_1 & \rightarrow X_1w_1 \\
X_2 & \rightarrow X_2w_2 \\
\vdots & \\
X_M & \rightarrow X_Mw_M \\
\end{align*} \]

Output: \( Y \)

Credit attributed to feature \( X_M \)
Linear model

X: Features    Y: Outcome

Complex model \( f(\cdot) \)

Black Box

\( X_1, X_2, \ldots, X_M \) → \( Y \)
**Linear model**

\[ \sum \]

\[ X_1, X_2, \ldots, X_M \rightarrow Y \]

**Complex model** \( f(.) \)

**Additive feature attribution**

For a particular prediction

\[ \sum \]

\[ \phi_1(f, x), \phi_2(x), \ldots, \phi_M(f, x) \]

\[ X_1, X_2, \ldots, X_M \rightarrow Y \]
**Linear model**

\[ X: \text{Features} \quad Y: \text{Outcome} \]

\[ X_1, X_2, \ldots, X_M \]

\[ X_i w_i, X_i w_M \]

\[ \Sigma \rightarrow Y \]

---

**Complex model** \( f(.) \)

**Black Box**

\[ X_1, X_2, \ldots, X_M \]

\[ ? \rightarrow Y \]

---

**Additive feature attribution**

For a particular prediction

\[ \varphi_M(f, x) \]

Credit attributed to feature \( X_M \)

\[ \varphi_1(f, x) \]
LIME
Ribeiro et al. 2016

Shapley reg. values
Lipovetsky et al. 2001

QII
Datta et al. 2016

DeepLIFT
Shrikumar et al. 2016

Relevance prop.
Bach et al. 2015

Shapley sampling
Štrumbelj et al. 2011

Saabas
Saabas 2014
LIME
Ribeiro et al. 2016

Shapley reg. values
Lipovetsky et al. 2001

DeepLIFT
Shrikumar et al. 2016

Relevance prop.
Bach et al. 2015

QII
Datta et al. 2016

Shapley sampling
Štrumbelj et al. 2011

Saabas
Saabas 2014
LIME
Ribeiro et al. 2016

Shapley reg. values
Lipovetsky et al. 2001

QII
Datta et al. 2016

DeepLIFT
Shrikumar et al. 2016

Relevance prop.
Bach et al. 2015

Shapley sampling
Štrumbelj et al. 2011

Saabas
Saabas 2014
LIME
Ribeiro et al. 2016

Shapley reg. values
Lipovetsky et al. 2001

QII
Datta et al. 2016

DeepLIFT
Shrikumar et al. 2016

Relevance prop.
Bach et al. 2015

Shapley sampling
Štrumbelj et al. 2011

Saabas
Saabas 2014
LIME
Ribeiro et al. 2016

Shapley reg. values
Lipovetsky et al. 2001

DeepLIFT
Shrikumar et al. 2016

Relevance prop.
Bach et al. 2015

QII
Datta et al. 2016

Shapley sampling
Štrumbelj et al. 2011

Saabas
Saabas 2014
Additive feature attribution methods

- **LIME**
  - Ribeiro et al. 2016

- **DeepLIFT**
  - Shrikumar et al. 2016

- **Shapley reg. values**
  - Lipovetsky et al. 2001

- **Relevance prop.**
  - Bach et al. 2015

- **QII**
  - Datta et al. 2016

- **Shapley sampling**
  - Štrumbelj et al. 2011

- **Saabas**
  - Saabas 2014
Additive feature attribution methods

LIME

Shapley reg. values

QII

Shapley sampling

DeepLIFT

Relevance prop.

Saabas
Additive feature attribution methods

LIME

DeepLIFT

Shapley reg. values

Relevance prop.

QII

Shapley sampling

Saabas
Additive feature attribution methods

LIME

Shapley reg. values

QII

Shapley sampling

DeepLIFT

Relevance prop.

Saabas
Additive feature attribution methods

LIME

DeepLIFT

Shapley reg. values

Relevance prop.

QII

Shapley sampling

Saabas
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%

\[ \mathbb{E}[f(x)] \]
Base rate
20%

$E[f(x)]$

Prediction for John
55%

$f(x)$
Base rate

20%

$E[f(x)]$

Prediction for John

55%

$f(x)$

How did we get here?
35%

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

$\phi_0$  $\phi_1$

Age = 20
$E[f(x)]$  

$\phi_0 \rightarrow \phi_1 \rightarrow \phi_2 \rightarrow \phi_3$  

$E[f(x) \mid x_1, x_2, x_3]$  

Open accounts = 1
The order matters!

\[ E[f(x)] \]

\[ f(x) \]

\[ \phi_0 \rightarrow \phi_1 \rightarrow \phi_2 \rightarrow \phi_3 \rightarrow \phi_4 \]

Age = 20

Day trader
The order matters!

$0 \to E[f(x)] \to f(x) \to \phi_0 \to \phi_2 \to \phi_1 \to \phi_3 \to \phi_4$

Day trader

Age = 20
35%

$E[f(x)]$  $E[f(x) \mid x_1]$

$\phi_0$  $\phi_1$

Age = 20
$E[f(x)]$
$35\%$

$E[f(x)]$

$E[f(x) \mid x_1]$

$\phi_0$

$\phi_1$

Age = 20
\[ E[f(x) \mid x_1, x_2] \]
$E[f(x)] \quad \phi_0 \quad \phi_1 \quad \phi_2 \quad \phi_3$

$E[f(x) \mid x_1, x_2, x_3]$

90%

Open accounts = 1
The order matters!

$E[f(x)]$ 

$\phi_0$  $\phi_1$ 

Age = 20 

Day trader 

$\phi_2$  $\phi_3$  $\phi_4$
The order matters!

\[ C \xrightarrow{E[f(x)]} \phi_0 \xrightarrow{\phi_2} \phi_3 \]

Day trader

\[ \text{Age} = 20 \]
The order matters!

Lloyd Shapley

Nobel Prize in 2012

$E[f(x)] \rightarrow f(x) \rightarrow \phi_3$

$\phi_0 \rightarrow \phi_2 \rightarrow \phi_1 \rightarrow \phi_4$

Day trader

Age = 20
Shapley properties

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

```
\[ E[f(x)] \quad \phi_0 \quad \phi_1 \quad \phi_2 \quad \phi_3 \quad \phi_4 \]

Age = 20

Day trader
```
Shapley properties

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

**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.

\[
\begin{align*}
\phi_0 & \quad E[f(x)] \\
\phi_1 & \quad \text{Age} = 20 \\
\phi_2 & \quad \text{Day trader} \\
\phi_3 & \\
\phi_4 &
\end{align*}
\]
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.

Violating consistency means you can’t trust feature orderings based on your attributions. ...even within the same model!
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 Objective

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

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

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

The loss \( L \), regularizer \( \Omega \), and local kernel \( \pi_{x'} \) were all chosen heuristically...
L, Ω, and $\pi_x$, are forced under local accuracy and consistency!
$L, \Omega, \text{ and } \pi_{x'}$ are forced under local accuracy and consistency!

$$L(f, g, \square_{x^0}) = \sum_{z^0 \in Z} f(h_x^{-1}(z^0)) - g(z^0) - \square_{x^0}(z^0)$$

$\boxtimes(g) = 0$

$$\Box_{x^0}(z^0) = \frac{(M - 1)}{(M \choose |z^0|)(M - |z^0|)}$$
$L, \Omega, \text{ and } \pi_x,$ are forced under local accuracy and consistency!

This means we can now estimate the Shapley values using linear regression!
$L$, $\Omega$, and $\pi_{\chi'}$ 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

[Graphs showing comparison between SHAP, Shapley sampling, LIME, and True Shapley value for dense and sparse original models.]
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

Inputs

Outputs

Preusience
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

Potential desat. region

45 dynamic features

43 other patient time series features

Odds ratio (current odds/typical odds)

-20 min  -15 min  -10 min  -5 min  (now)  +5 min

Anaesthesia ready

Prediction window

Prediction

Inputs

Outputs

\( \mathcal{L} \) prescience

Procedure history  
Procedure future

28
Prescience predicts hypoxemia and explains why
Prescience predicts hypoxemia and explains why
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

Potential desat. region

45 dynamic features

43 other patient time series features

\[ \text{Odds ratio} = 2.4 \]

5 min. odds ratio

Explanation

Why?

Inputs

Outputs

\[ \frac{\text{Prescience}}{\text{Prediction window}} \]
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.
An interpretability vs. accuracy tradeoff

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

Graph:
- TPR (% of desats correctly predicted)
- FPR (% of non-desats incorrectly predicted)
- GBM trees (AUC 0.90)
- Linear lasso (AUC 0.86)
- Random (AUC 0.5)

Diagram:
- Complex model $f(.)$
- Generalized linear model

Black Box
- $X_1, X_2, \ldots, X_p$
- $Y$
- $X$: Features
- $Y$: Outcome
- $\Sigma$
- $Y$

30
An interpretability vs. accuracy tradeoff

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

When FPR=0.1, TPR changes +15%

- GBM trees (AUC 0.90)
- Linear lasso (AUC 0.86)
- Random (AUC 0.5)

Complex model $f(.)$

Generalized linear model

Black Box

X: Features
Y: Outcome

$X_1, X_2, \ldots, X_p$
Using SHAP values in the operating room
Using SHAP values in the operating room

<table>
<thead>
<tr>
<th>base value</th>
<th>hypoxemia fold risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.125</td>
<td>2.4</td>
</tr>
<tr>
<td>0.25</td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
</tr>
</tbody>
</table>
Using SHAP values in the operating room

Red features push the risk higher

Succinylcholine | Peak pressure | SpO2 | Tidal volume | Height/weight

base value

hypoxemia fold risk

2.4
Using SHAP values in the operating room

Red features push the risk higher

Succinylcholine  Peak pressure  SpO2  Tidal volume  Height/weight

Feature impact

base value

hypoxemia fold risk

2.4
Using SHAP values in the operating room

Red features push the risk higher

Succinylcholine | Peak pressure | SpO2 | Tidal volume | Height/weight

base value

hypoxemia fold risk 2.4

Green features push the risk lower

Pulse | Sevoflurane | Respiration rate | Ablation in proc. text

Feature impact

31
Using SHAP values in the operating room

Red features push the risk higher

base value

hypoxemia fold risk 2.4

Green features push the risk lower

Feature impact

B

hypo2 fold risk

SpO2

16


NSR

Anesthesia ready

Now
Using SHAP values in the operating room

- Red features push the risk higher
- Green features push the risk lower

- Features: Succinylcholine, Peak pressure, SpO2, Tidal volume, Height/weight, Pulse, Sevoflurane, Respiration rate, ablation in proc text

- Hypoxemia fold risk: 2.4
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

Succinylcholine, Peak pressure, SpO2, Tidal volume, Height/weight

Green features push the risk lower

Pulse, Sevoflurane, Respiration rate, ablation in proc text

Feature impact

B

Now

hypoxia fold risk

16

8

4

2

1

0.5

0.25

0.125

SpO2

Tidal volume

Height/weight


NSR Anesthesia ready NSR NSR NSR Procedure
Using SHAP values in the operating room

Red features push the risk higher

base value

hypoxemia fold risk 2.4

Green features push the risk lower

Feature impact

Succinylcholine  Peak pressure  SpO2  Tidal volume  Height/weight  Pulse  Sevoflurane  Respiration rate  ablation in proc text

B

16


hypoxia fold risk

SpO2


Anesthesia ready

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.

![Diagram showing patient attributes and hypoxemia prediction](image-url)
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

![Graph showing explanation runtime vs. number of features in the model.](image)
Room for improvement: Model agnostic approaches can be slow and variable
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

<table>
<thead>
<tr>
<th>Model</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Regression</td>
<td>74.6%</td>
</tr>
<tr>
<td>Random Forests</td>
<td>65.0%</td>
</tr>
<tr>
<td>Decision Trees</td>
<td>61.5%</td>
</tr>
<tr>
<td>Ensemble Methods</td>
<td>44.3%</td>
</tr>
<tr>
<td>Gradient Boosted Machines</td>
<td>39.6%</td>
</tr>
<tr>
<td>Neural Networks</td>
<td>38.6%</td>
</tr>
<tr>
<td>Bayesian Techniques</td>
<td>31.8%</td>
</tr>
<tr>
<td>SVMs</td>
<td>30.8%</td>
</tr>
<tr>
<td>CNNs</td>
<td>16.6%</td>
</tr>
<tr>
<td>RNNs</td>
<td>12.9%</td>
</tr>
</tbody>
</table>

Kaggle 2017 survey of Data Scientists
Tree-based models are the most popular complex models used in industry

- Logistic Regression: 74.6%
- Random Forests: 65.0%
- Decision Trees: 61.5%
- Ensemble Methods: 44.3%
- Gradient Boosted Machines: 39.6%
- Neural Networks: 38.6%
- Bayesian Techniques: 31.8%
- SVMs: 30.8%
- CNNs: 16.6%
- RNNs: 12.9%

*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

\[ O(TL2^MN) \] Exponential

The solution depends on an exponential number of expected values!
SHAP values for trees

Direct Solution

$O(TLM! N)$ \text{ Factorial}

$O(TL2^M N)$ \text{ Exponential}

to

Tree SHAP

$O(TLD^2)$ \text{ Polynomial}
Tree SHAP is fast and exact
Tree SHAP is fast and exact

Explanation runtime (simulated data)

- Model Agnostic Lower Bound
- Tree SHAP

Minutes of runtime (explaining 10k predictions)

Explanation variability (simulated data)

- IME
- Kernel SHAP
- Tree SHAP

Std. deviation as % of magnitude
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 (example)
- Keep Negative (mask)
- Keep Negative (example)
- Keep Absolute (mask)
- Keep Absolute (example)
- Remove Positive (mask)
- Remove Positive (example)
- Remove Negative (mask)
- Remove Negative (example)
- Remove Absolute (mask)
- Remove Absolute (example)
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
Different evaluation metrics

<table>
<thead>
<tr>
<th>Decision Tree</th>
<th>Random Forest</th>
</tr>
</thead>
<tbody>
<tr>
<td>TreeExplainer (independent)</td>
<td>TreeExplainer (independent)</td>
</tr>
<tr>
<td>TreeExplainer</td>
<td>TreeExplainer</td>
</tr>
<tr>
<td>Saabas</td>
<td>Saabas</td>
</tr>
<tr>
<td>Kernel SHAP 1000 mean ref.</td>
<td>Kernel SHAP 1000 mean ref.</td>
</tr>
<tr>
<td>IME 1000</td>
<td>IME 1000</td>
</tr>
<tr>
<td>mean([TreeExplainer])</td>
<td>mean([TreeExplainer])</td>
</tr>
<tr>
<td>Gain/Gini importance</td>
<td>Gain/Gini importance</td>
</tr>
<tr>
<td>Random</td>
<td>Random</td>
</tr>
</tbody>
</table>

Higher score

Lower score
Improved feature selection power

Single decision tree (minimum interactions)

- Gain
- Permutation
- mean(|SHAP value|)
- mean(SHAP value of loss)

Fraction of support recovered

Size of support (# of features in the model)

(out of 200 total features)
Improved feature selection power

Single decision tree
(minimum interactions)

Random forest with 10 trees
(minimum interactions)

Gain
Permutation
mean(|SHAP value|)
mean(SHAP value of loss)

Fraction of support recovered

Size of support (# of features in the model)

(out of 200 total features)
Consistency with human intuition

AND (true/true) - Decision Tree

*Human consensus*

<table>
<thead>
<tr>
<th>Feature</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fever</td>
<td>8</td>
</tr>
<tr>
<td>Cough</td>
<td>8</td>
</tr>
</tbody>
</table>

Features in the model:
- Fever
- Cough
- Headache
Consistency with human intuition

AND (true/true) - Decision Tree

Human consensus
SHAP values

Feature attribution value

Fever
Cough
Headache
Features in the model
Consistency with human intuition

**AND (true/true) - Decision Tree**

<table>
<thead>
<tr>
<th>Feature</th>
<th>Feature attribution value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fever</td>
<td>4</td>
</tr>
<tr>
<td>Cough</td>
<td>2</td>
</tr>
<tr>
<td>Headache</td>
<td>0</td>
</tr>
</tbody>
</table>

- **Human consensus**
- **SHAP values**
- **Heuristic values (Saabas)**
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

mean(|SHAP value|)
Mortality risk model

Global feature importance

Age

mean(|SHAP value|)
Mortality risk model

Global feature importance

- Age
- Sex (F/M)
- Systolic blood pressure

mean(|SHAP value|)
Mortality risk model

Global feature importance

- Age
- Sex (F/M)
- Systolic blood pressure
- White blood cells

mean(|SHAP value|)
Mortality risk model

Global feature importance

- Age
- Sex (F/M)
- Systolic blood pressure
- White blood cells
- BMI
- Sedimentation rate
- Blood albumin
- Alkaline phosphatase
- Total cholesterol
- Physical activity
- Hematocrit
- Uric acid
- Red blood cells
- Albumin present in urine
- Blood protein

mean(|SHAP value|)
Mortality risk model

Conflates the prevalence of an effect with the magnitude of an effect
Reveal rare high-magnitude mortality effects

Global feature importance

- Age
- Sex (F/M)
- Systolic blood pressure
- White blood cells
- BMI
- Sedimentation rate
- Blood albumin
- Alkaline phosphatase
- Total cholesterol
- Physical activity
- Hematocrit
- Uric acid
- Red blood cells
- Albumin present in urine
- Blood protein

Conflates the prevalence of an effect with the magnitude of an effect.
Reveal rare high-magnitude mortality effects

Global feature importance

Local explanation summary

- Age
- Sex (F/M)
- Systolic blood pressure
- White blood cells
- BMI
- Sedimentation rate
- Blood albumin
- Alkaline phosphatase
- Total cholesterol
- Physical activity
- Hematocrit
- Uric acid
- Red blood cells
- Albumin present in urine
- Blood protein

mean(|SHAP value|)

SHAP value (impact on model output) (log relative risk of mortality)
Reveal rare high-magnitude mortality effects

Global feature importance

- Age
- Sex (F/M)
- Systolic blood pressure
- White blood cells
- BMI
- Sedimentation rate
- Blood albumin
- Alkaline phosphatase
- Total cholesterol
- Physical activity
- Hematocrit
- Uric acid
- Red blood cells
- Albumin present in urine
- Blood protein

Local explanation summary

SHAP value (impact on model output)
(log relative risk of mortality)
Reveal rare high-magnitude mortality effects

Global feature importance

- Age
- Sex (F/M)
- Systolic blood pressure
- White blood cells
- BMI
- Sedimentation rate
- Blood albumin
- Alkaline phosphatase
- Total cholesterol
- Physical activity
- Hematocrit
- Uric acid
- Red blood cells
- Albumin present in urine
- Blood protein

mean(|SHAP value|)

Local explanation summary

SHAP value (impact on model output)
(log relative risk of mortality)
Reveal rare high-magnitude mortality effects

Global feature importance

- Age
- Sex (F/M)
- Systolic blood pressure
- White blood cells
- BMI
- Sedimentation rate
- Blood albumin
- Alkaline phosphatase
- Total cholesterol
- Physical activity
- Hematocrit
- Uric acid
- Red blood cells
- Albumin present in urine
- Blood protein

Local explanation summary

SHAP value (impact on model output)
(log relative risk of mortality)
Reveal rare high-magnitude mortality effects

Global feature importance

<table>
<thead>
<tr>
<th>Feature</th>
<th>SHAP value (impact on model output)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td></td>
</tr>
<tr>
<td>Sex (F/M)</td>
<td></td>
</tr>
<tr>
<td>Systolic blood pressure</td>
<td></td>
</tr>
<tr>
<td>White blood cells</td>
<td></td>
</tr>
<tr>
<td>BMI</td>
<td></td>
</tr>
<tr>
<td>Sedimentation rate</td>
<td></td>
</tr>
<tr>
<td>Blood albumin</td>
<td></td>
</tr>
<tr>
<td>Alkaline phosphatase</td>
<td></td>
</tr>
<tr>
<td>Total cholesterol</td>
<td></td>
</tr>
<tr>
<td>Physical activity</td>
<td></td>
</tr>
<tr>
<td>Hematocrit</td>
<td></td>
</tr>
<tr>
<td>Uric acid</td>
<td></td>
</tr>
<tr>
<td>Red blood cells</td>
<td></td>
</tr>
<tr>
<td>Albumin present in urine</td>
<td></td>
</tr>
<tr>
<td>Blood protein</td>
<td></td>
</tr>
</tbody>
</table>

(mean(|SHAP value|))
Reveal rare high-magnitude mortality effects

Global feature importance

- Age
- Sex (F/M)
- Systolic blood pressure
- White blood cells
- BMI
- Sedimentation rate
- Blood albumin
- Alkaline phosphatase
- Total cholesterol
- Physical activity
- Hematocrit
- Uric acid
- Red blood cells
- Albumin present in urine
- Blood protein

Local explanation summary

Lots of ways to die young...

mean(|SHAP value|) vs. SHAP value (impact on model output)

(log relative risk of mortality)
Reveal rare high-magnitude mortality effects

Global feature importance

- Age
- Sex (M/F)
- Systolic blood pressure
- White blood cells
- BMI
- Sedimentation rate
- Blood albumin
- Alkaline phosphatase
- Total cholesterol
- Physical activity
- Hematocrit
- Uric acid
- Red blood cells
- Albumin present in urine
- Blood protein

Local explanation summary

- SHAP value (impact on model output)
  (log relative risk of mortality)

Not many ways to live longer...
Reveal rare high-magnitude mortality effects

Global feature importance:
- Age
- Sex (M/F)
- Systolic blood pressure
- White blood cells
- BMI
- Sedimentation rate
- Blood albumin
- Alkaline phosphatase
- Total cholesterol
- Physical activity
- Hematocrit
- Uric acid
- Red blood cells
- Albumin present in urine
- Blood protein

Local explanation summary:

mean(|SHAP value|) vs. SHAP value (impact on model output)
(log relative risk of mortality)
Vertical dispersion is driven by interaction effects.
Vertical dispersion is driven by interaction effects.
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

![Graph showing smoothed model squared error over time, with a shaded area indicating the first year of data used for training.](Image)
Model monitoring

Training performance

First year of data used for training

Smoothed model squared error

Time

(A)
Model monitoring

Can you find where we introduced the bug?
Model monitoring

(A) Smoothed model squared error over time from 2012 to 2016. The first year of data used for training is highlighted.

(B) SHAP loss value of in room #6 plotted over time with feature values indicated by color (False = blue, True = red).
Model monitoring

Can you find where we introduced the bug?
Model monitoring

Feature values
- False
- True

(A) Smoothed model squared error
- First year of data used for training

(B) SHAP loss value of in room #6
Model monitoring

Now can you find where we introduced the bug?
Model monitoring

SHAP loss value of general anesthesia

Feature values
- False
- True

Time


(C)
Model monitoring

Transient electronic medical record
Model monitoring

SHAP loss value of general anesthesia

SHAP loss value of atrial fibrillation

Feature values
- False
- True

Time
Model monitoring

Gradual change in atrial fibrillation ablation procedure durations
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 😊

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
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

ROLLS ROYCE

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
Collaborations

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

Manuscript in under review.
Collaborations
Lundberg et al. ChromNet: Learning the human chromatin network from all ENCODE ChIP-seq data. Genome Biology, 2016. (F1000Prime recommended)
Thanks!