Interpretability and Shapley Values

R. Teal Witter

Machine Learning: Supervised Tasks

What is the risk of heart disease?

Input:

Age	35 yr
Cholesterol	240 mg/dL
Smoking	Former
Family History	No
Exercise	Moderate
Glucose	115 mg/dL

Output: 6 / 10 risk

Machine Learning: Text Generation

In three sentences of Shakespearean English, please explain why coding is so fun.

Verily, coding doth transport the mind to realms of boundless creation, where one doth conjure wondrous contrivances from naught but mere symbols and letters. With each keystroke, the coder weaveth intricate spells that command the very essence of machines to dance to their will, solving puzzles with elegance and grace. In this noble craft, the thrill of discovery and the joy of crafting artful algorithms doth bring forth a delight that knoweth no bounds.

Source: ChatGPT

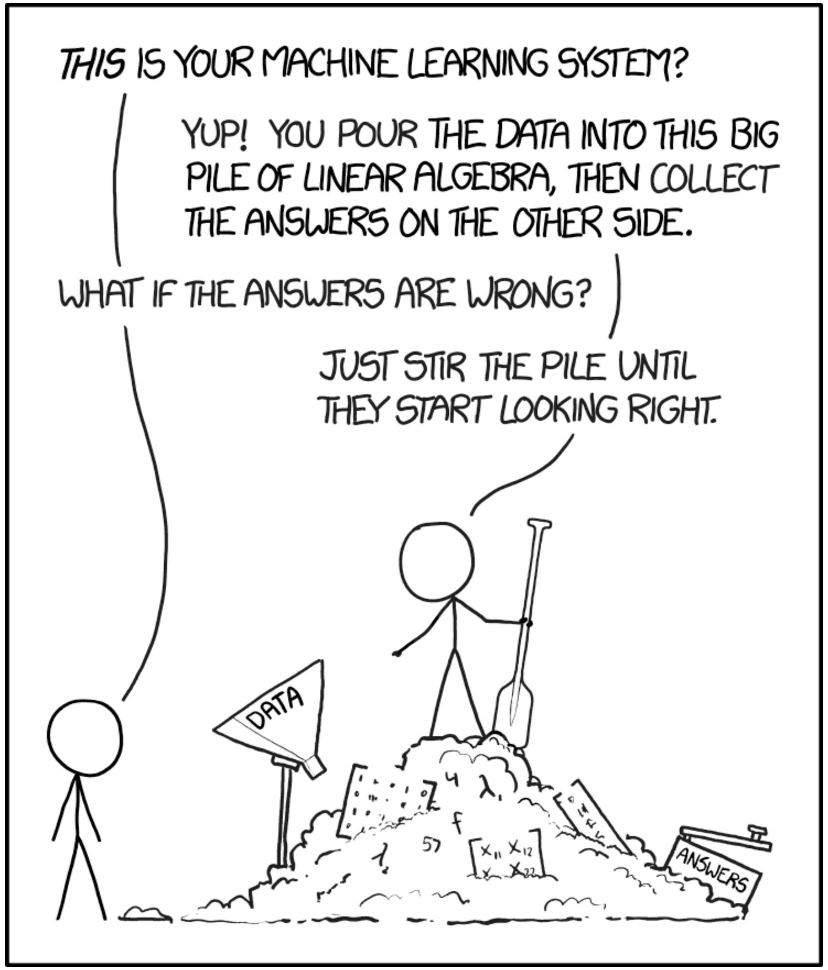
Machine Learning: Image Generation

Source: Microsoft

Source: ChatGPT

“An underwater library filled with ancient books and jellyfish swimming around.”

Motivation



Machine learning works *really* well so let's use it!

If we use ML to affect people's lives in domains like law, healthcare, and finance, **how can we *justify* ML decisions?**

Machine Learning

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \Rightarrow f \Rightarrow f(\boldsymbol{x})$$

Typically, f is a neural network

Example:

$\boldsymbol{x}$ $\quad$ f $\quad$ $f(\boldsymbol{x})$

Age	35 yr
Cholesterol	240 mg/dL
Smoking	Former
Family History	No
Exercise	Moderate
Glucose	115 mg/dL

6/10 risk

Explaining Predictions

Attribution value

Attribute the prediction to features

"Since **cholesterol** is 240 mg/dL, the risk is 3 higher than baseline."

$$\begin{array}{cccccccc}
 & \{1,\mathbf{2},3,4,5,6\} & & \{1,\mathbf{2},4,5\} & \{1,4,5\} & & \{\mathbf{2},3,6\} & \{3,6\} \\
\textbf{Cholesterol} & \begin{bmatrix} \text{35 yr} \\ \textbf{240 mg/dL} \\ \text{Former} \\ \text{No} \\ \text{Moderate} \\ \text{115 mg/dL} \end{bmatrix} & \cdots & \begin{bmatrix} \text{35 yr} \\ \textbf{240 mg/dL} \\ * \\ \text{No} \\ \text{Moderate} \\ * \end{bmatrix} & \begin{bmatrix} \text{35 yr} \\ * \\ * \\ \text{No} \\ \text{Moderate} \\ * \end{bmatrix} & \cdots & \begin{bmatrix} * \\ \textbf{240 mg/dL} \\ \text{Former} \\ * \\ * \\ \text{115 mg/dL} \end{bmatrix} & \begin{bmatrix} * \\ * \\ \text{Former} \\ * \\ * \\ \text{115 mg/dL} \end{bmatrix} \\
f & 6 & & 5 & 3 & & 7 & 6
\end{array}$$

$\boxed{v(S \cup \{\mathbf{2}\}) - v(S)}$

$\boxed{v(S \cup \{\mathbf{2}\}) - v(S)}$

Shapley Values

The Shapley value for feature i:

$$\phi_i = \frac{1}{n} \sum_{S \subseteq [n] \setminus \{i\}} \frac{\boxed{v(S \cup \{i\}) - v(S)}}{\binom{n-1}{|S|}}$$

+ Based on axioms and studied by game theorists
+ The de facto explainable AI method (25k citations and 20k repos)
- Not necessarily the “right” answer (several limitations)

Shapley Values

The Shapley value for feature i:

$$\phi_i = \frac{1}{n} \sum_{S \subseteq [n] \setminus \{i\}} \frac{v(S \cup \{i\}) - v(S)}{\binom{n-1}{|S|}}$$

$$\phi_i = \underbrace{\frac{1}{n} \sum_{k=0}^{n-1} \underbrace{\frac{1}{\binom{n-1}{k}} \sum_{S \subseteq [n] \setminus \{i\} : |S| = k} v(S \cup \{i\}) - v(S)}_{\text{Average over sets of size } k}}_{\text{Average over all sizes } k}$$

Shapley Values

The Shapley value for feature i:

$$\phi_i = \frac{1}{n} \sum_{S \subseteq [n] \setminus \{i\}} \frac{v(S \cup \{i\}) - v(S)}{\binom{n-1}{|S|}}$$

Question: How do we compute Shapley values?

Shapley Values

The Shapley value for feature i:

$$\phi_i = \frac{1}{n} \sum_{S \subseteq [n] \setminus \{i\}} \frac{v(S \cup \{i\}) - v(S)}{\binom{n-1}{|S|}}$$

When $i = 1$ and $n = 4$:

$$\phi_1 = \frac{1}{4}\Bigg(\frac{v(\{1\}) - v(\emptyset)}{\binom{3}{0}} + \frac{v(\{1,2\}) - v(\{2\})}{\binom{3}{1}} + \frac{v(\{1,3\}) - v(\{3\})}{\binom{3}{1}} + \frac{v(\{1,4\}) - v(\{4\})}{\binom{3}{1}} + \frac{v(\{1,2,3\}) - v(\{2,3\})}{\binom{3}{2}} + \frac{v(\{1,2,4\}) - v(\{2,4\})}{\binom{3}{2}} + \frac{v(\{1,3,4\}) - v(\{3,4\})}{\binom{3}{2}} + \frac{v(\{1,2,3,4\}) - v(\{2,3,4\})}{\binom{3}{3}} \Bigg)$$

Shapley Values

The Shapley value for feature i:

$$\phi_i = \frac{1}{n} \sum_{S \subseteq [n] \setminus \{i\}} \frac{v(S \cup \{i\}) - v(S)}{\binom{n-1}{|S|}}$$

When $i = 1$ and $n = 5$:

$$\begin{aligned}\phi_1 = \frac{1}{5}\Bigg(&\frac{v(\{1\}) - v(\emptyset)}{\binom{4}{0}} + \frac{v(\{1,2\}) - v(\{2\})}{\binom{4}{1}} + \frac{v(\{1,3\}) - v(\{3\})}{\binom{4}{1}} + \frac{v(\{1,4\}) - v(\{4\})}{\binom{4}{1}} + \frac{v(\{1,5\}) - v(\{5\})}{\binom{4}{1}} + \frac{v(\{1,2,3\}) - v(\{2,3\})}{\binom{4}{2}} \\ &+ \frac{v(\{1,2,4\}) - v(\{2,4\})}{\binom{4}{2}} + \frac{v(\{1,2,5\}) - v(\{2,5\})}{\binom{4}{2}} + \frac{v(\{1,3,4\}) - v(\{3,4\})}{\binom{4}{2}} + \frac{v(\{1,3,5\}) - v(\{3,5\})}{\binom{4}{2}} \\ &+ \frac{v(\{1,4,5\}) - v(\{4,5\})}{\binom{4}{2}} + \frac{v(\{1,2,3,4\}) - v(\{2,3,4\})}{\binom{4}{3}} + \frac{v(\{1,2,3,5\}) - v(\{2,3,5\})}{\binom{4}{3}} + \frac{v(\{1,2,4,5\}) - v(\{2,4,5\})}{\binom{4}{3}} \\ &+ \frac{v(\{1,3,4,5\}) - v(\{3,4,5\})}{\binom{4}{3}} + \frac{v(\{1,2,3,4,5\}) - v(\{2,3,4,5\})}{\binom{4}{4}}\Bigg)\end{aligned}$$

Shapley Values

The Shapley value for feature i:

$$\phi_i = \frac{1}{n} \sum_{S \subseteq [n] \setminus \{i\}} \frac{v(S \cup \{i\}) - v(S)}{\binom{n-1}{|S|}}$$

When $i = 1$ and $n = 6$:

$$\begin{aligned}
\phi_1 = \frac{1}{6}\Big(& \frac{v(\{1\}) - v(\emptyset)}{\binom{5}{0}} + \frac{v(\{1,2\}) - v(\{2\})}{\binom{5}{1}} + \frac{v(\{1,3\}) - v(\{3\})}{\binom{5}{1}} + \frac{v(\{1,4\}) - v(\{4\})}{\binom{5}{1}} + \frac{v(\{1,5\}) - v(\{5\})}{\binom{5}{1}} + \frac{v(\{1,6\}) - v(\{6\})}{\binom{5}{1}} + \frac{v(\{1,2,3\}) - v(\{2,3\})}{\binom{5}{2}} \\
& + \frac{v(\{1,2,4\}) - v(\{2,4\})}{\binom{5}{2}} + \frac{v(\{1,2,5\}) - v(\{2,5\})}{\binom{5}{2}} + \frac{v(\{1,2,6\}) - v(\{2,6\})}{\binom{5}{2}} + \frac{v(\{1,3,4\}) - v(\{3,4\})}{\binom{5}{2}} + \frac{v(\{1,3,5\}) - v(\{3,5\})}{\binom{5}{2}} + \frac{v(\{1,3,6\}) - v(\{3,6\})}{\binom{5}{2}} \\
& + \frac{v(\{1,4,5\}) - v(\{4,5\})}{\binom{5}{2}} + \frac{v(\{1,4,6\}) - v(\{4,6\})}{\binom{5}{2}} + \frac{v(\{1,5,6\}) - v(\{5,6\})}{\binom{5}{2}} + \frac{v(\{1,2,3,4\}) - v(\{2,3,4\})}{\binom{5}{3}} + \frac{v(\{1,2,3,5\}) - v(\{2,3,5\})}{\binom{5}{3}} \\
& + \frac{v(\{1,2,3,6\}) - v(\{2,3,6\})}{\binom{5}{3}} + \frac{v(\{1,2,4,5\}) - v(\{2,4,5\})}{\binom{5}{3}} + \frac{v(\{1,2,4,6\}) - v(\{2,4,6\})}{\binom{5}{3}} + \frac{v(\{1,2,5,6\}) - v(\{2,5,6\})}{\binom{5}{3}} + \frac{v(\{1,3,4,5\}) - v(\{3,4,5\})}{\binom{5}{3}} \\
& + \frac{v(\{1,3,4,6\}) - v(\{3,4,6\})}{\binom{5}{3}} + \frac{v(\{1,3,5,6\}) - v(\{3,5,6\})}{\binom{5}{3}} + \frac{v(\{1,4,5,6\}) - v(\{4,5,6\})}{\binom{5}{3}} + \frac{v(\{1,2,3,4,5\}) - v(\{2,3,4,5\})}{\binom{5}{4}} \\
& + \frac{v(\{1,2,3,4,6\}) - v(\{2,3,4,6\})}{\binom{5}{4}} + \frac{v(\{1,2,3,5,6\}) - v(\{2,3,5,6\})}{\binom{5}{4}} + \frac{v(\{1,2,4,5,6\}) - v(\{2,4,5,6\})}{\binom{5}{4}} + \frac{v(\{1,3,4,5,6\}) - v(\{3,4,5,6\})}{\binom{5}{4}} \\
& + \frac{v(\{1,2,3,4,5,6\}) - v(\{2,3,4,5,6\})}{\binom{5}{5}} \Big)
\end{aligned}$$

Shapley Values

The Shapley value for feature i:

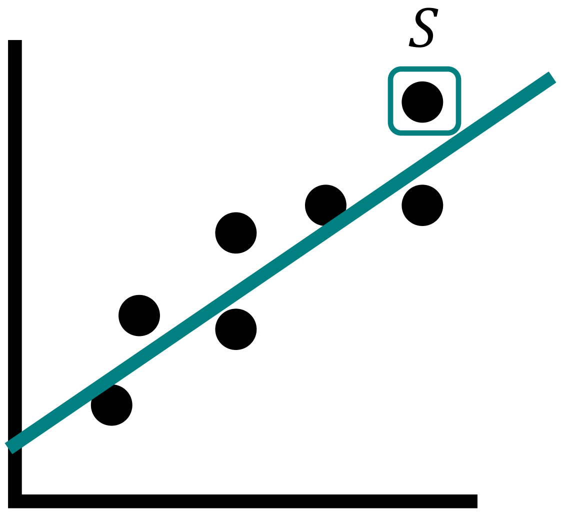
$$\phi_i = \frac{1}{n} \sum_{S \subseteq [n] \setminus \{i\}} \frac{v(S \cup \{i\}) - v(S)}{\binom{n-1}{|S|}}$$

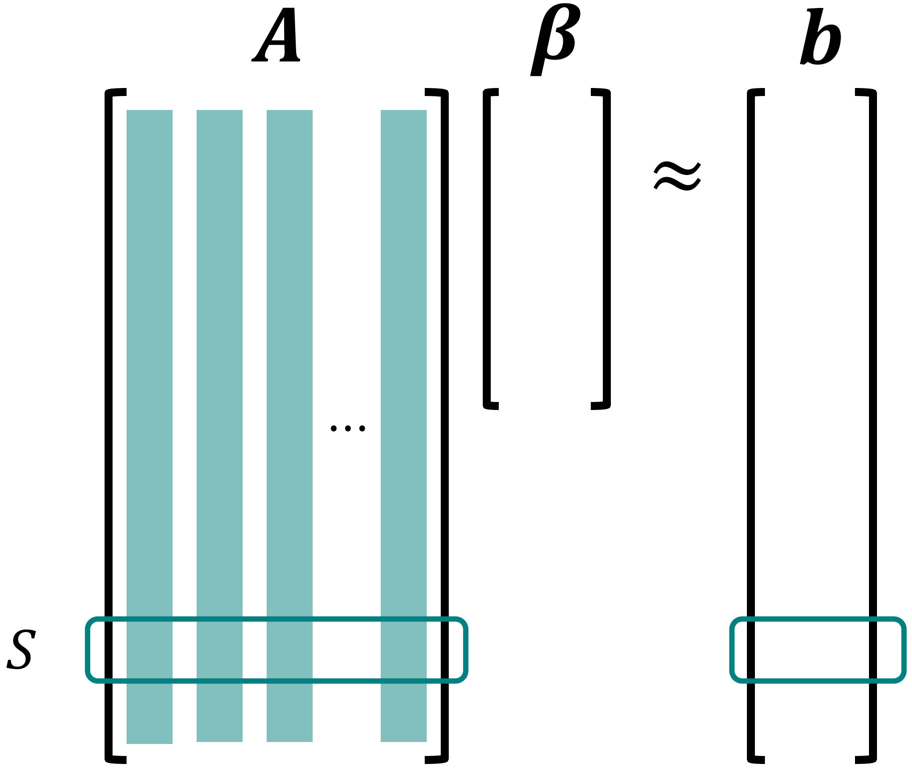
Challenge: There are $O(2^n)$ terms!

Big **Question:** How do we *efficiently* compute Shapley values?

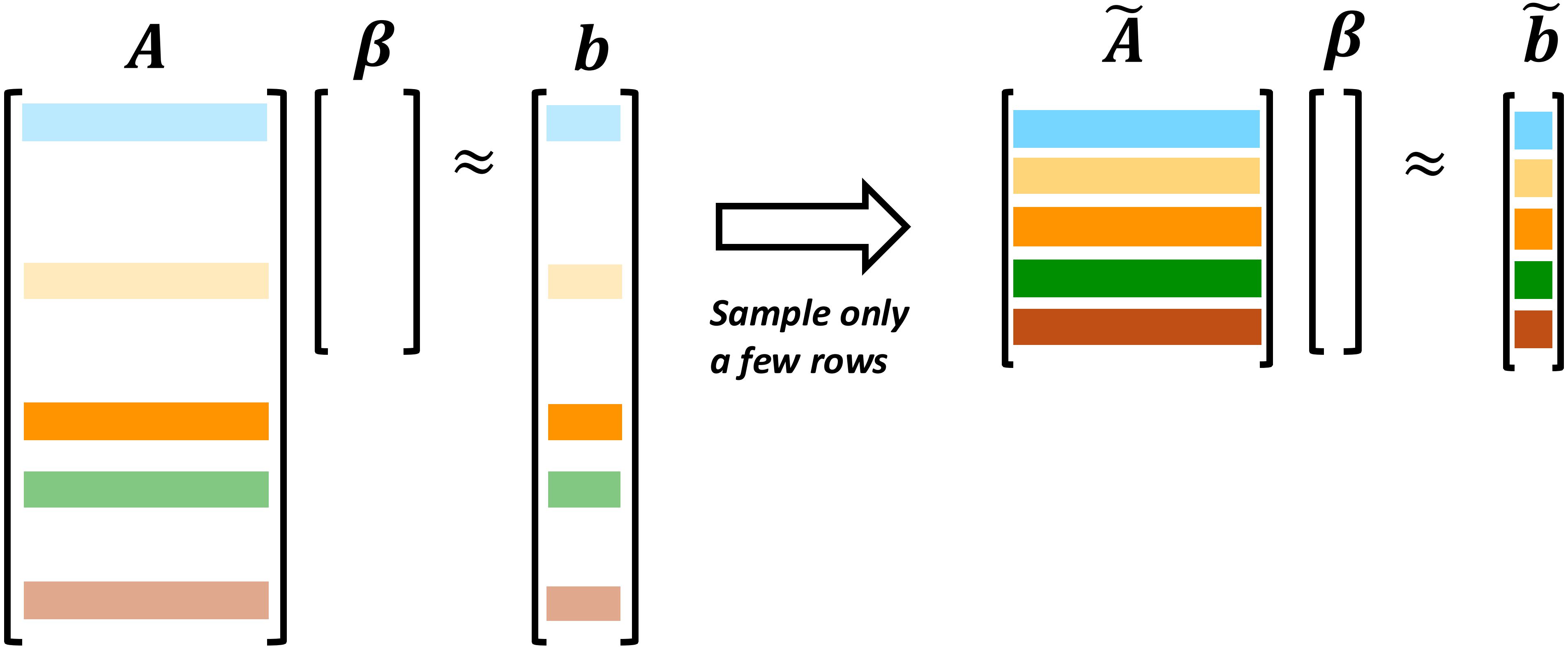
Regression Formulation

Lemma [CGKR ‘88]: We can compute Shapley values from the solution to a special linear regression problem.





Kernel SHAP


$\boldsymbol{A}$
$\boldsymbol{\beta}$
$\approx$
$\boldsymbol{b}$
Sample only
a few rows
$\tilde{\boldsymbol{A}}$
$\boldsymbol{\beta}$
$\approx$
$\tilde{\boldsymbol{b}}$

Beyond Kernel SHAP

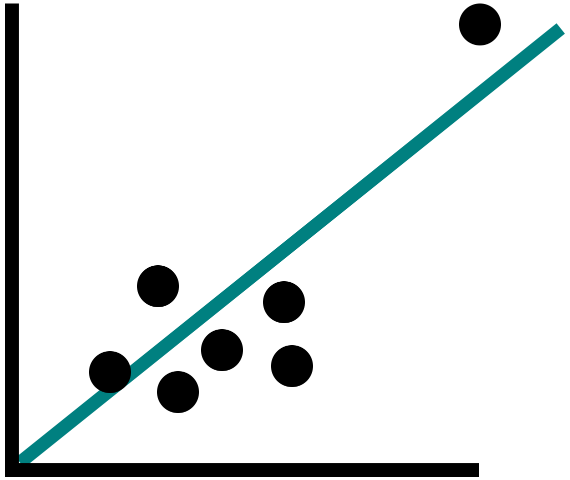
Question: How *should* we sample points?

Ideally, we want:

- Good performance
- Theoretical guarantees

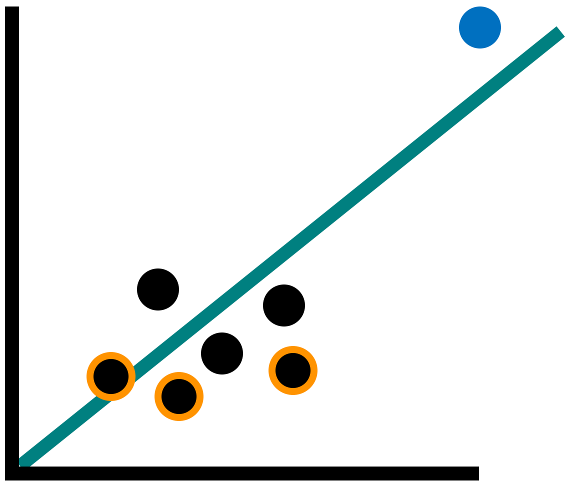
Leverage Scores

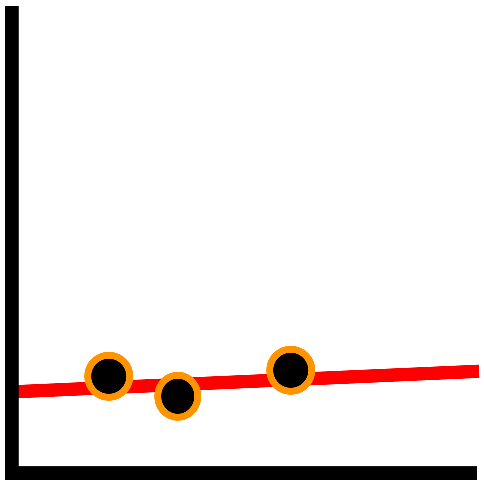
Challenge of Sampling: Which points preserve the line?



Leverage Scores

Challenge of Sampling: Which points preserve the line?





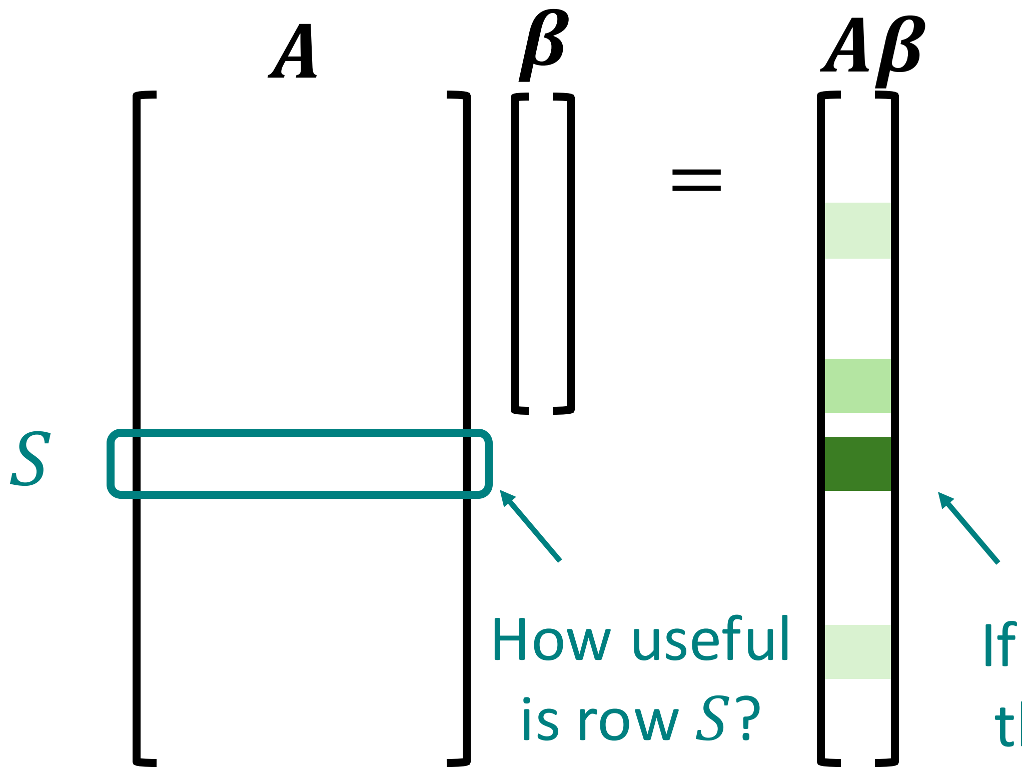
+ Without the high-leverage point, we find a *very* different line

Leverage Scores

Challenge of Sampling: Which points preserve the line?

+ With the high-leverage point, we find a close line

Leverage Scores

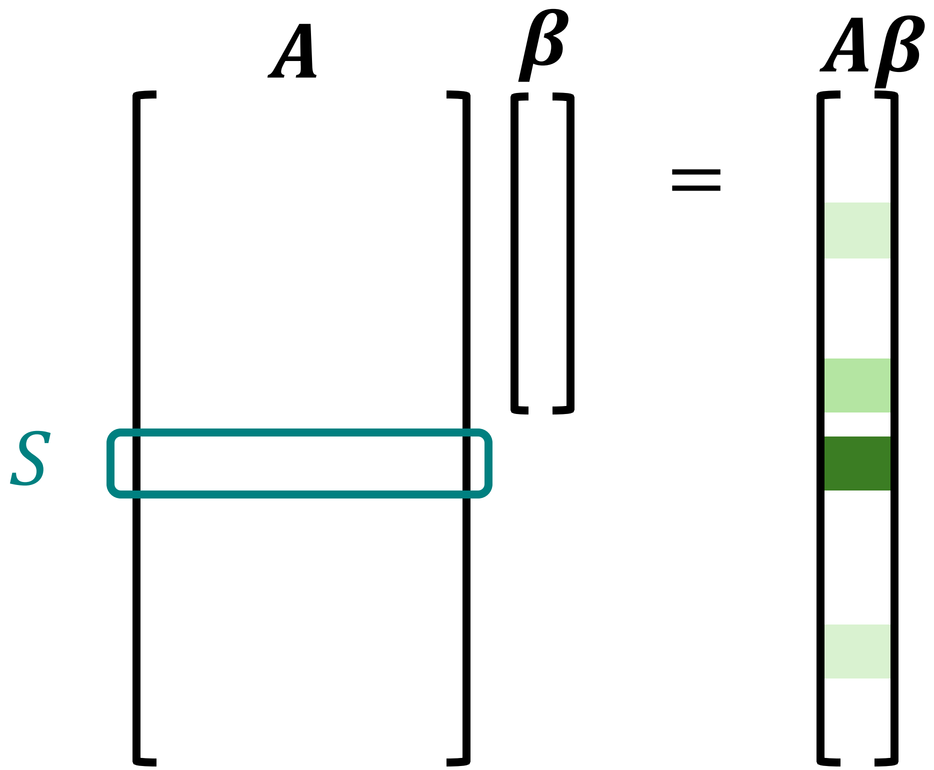


If there is an $\boldsymbol{A\beta}$ like this, then row S is "by itself".

Row S has "leverage":

$$\ell_S = \max_{\boldsymbol{\beta}} \frac{(\boldsymbol{A\beta})^2{}_S}{||\boldsymbol{A\beta}||_2^2}$$

Leverage Scores



Row S has "leverage":

$$\ell_S = \max_{\boldsymbol{\beta}} \frac{(\boldsymbol{A\beta})^2{}_S}{||\boldsymbol{A\beta}\,||_2^2}$$

$$= \frac{1}{\binom{n}{|S|}}$$

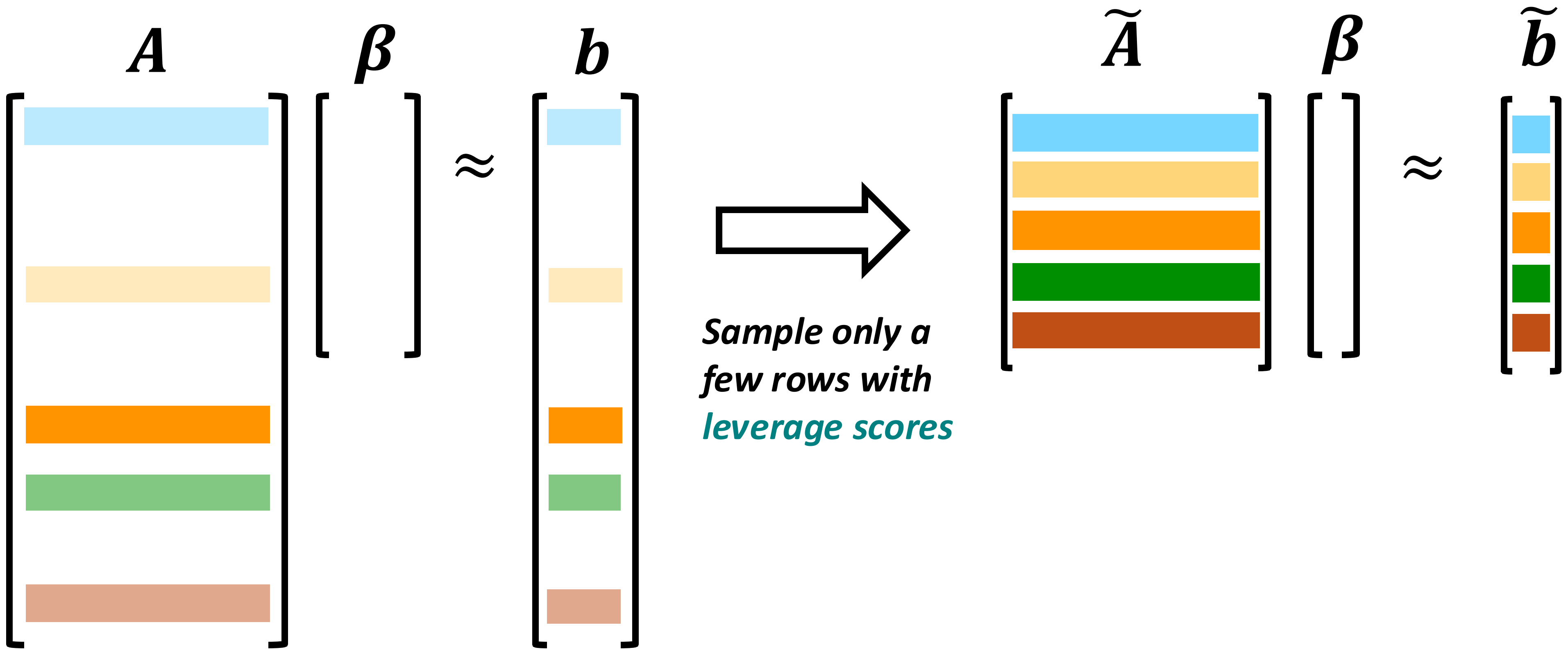
Very similar to weighting in Shapley value definition!

Shapley Values

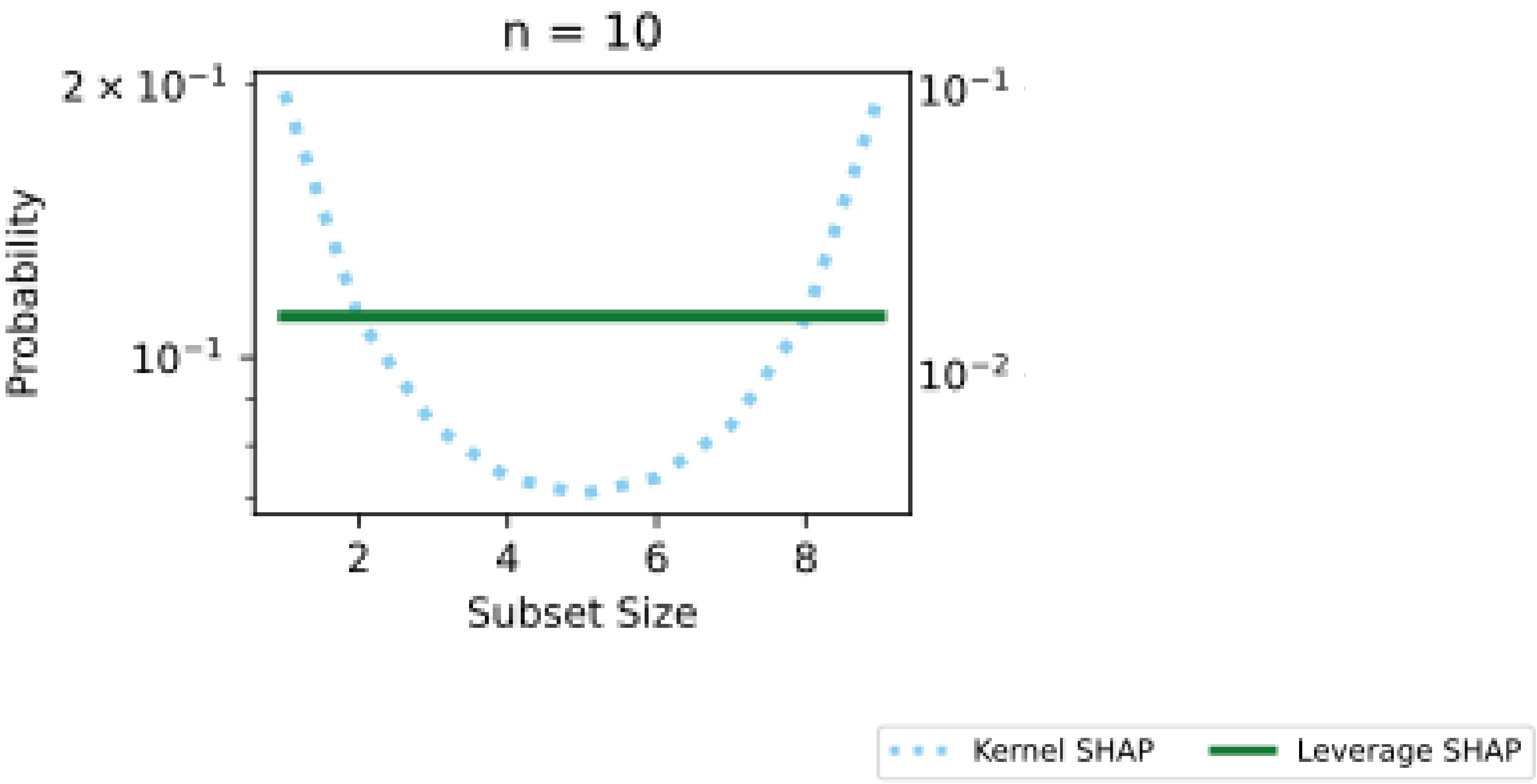
The Shapley value for feature i:

$$\phi_i = \underbrace{\frac{1}{n}\sum_{k=0}^{n-1} \underbrace{\frac{1}{\binom{n-1}{k}} \sum_{S\subseteq[n]\setminus\{i\}:|S|=k} v(S \cup \{i\}) - v(S)}_{\text{Average over sets of size } k}}_{\text{Average over all sizes } k}$$

Leverage SHAP



Leverage SHAP vs Kernel SHAP Sampling



Leverage SHAP Performance

$$\ell_2\text{-error: } ||\boldsymbol{\phi} - \tilde{\boldsymbol{\phi}}||_2^2 = \sum_{i=1}^{n} \left(\phi_i - \tilde{\phi}_i\right)^2$$

	California	Diabetes	Adult	Correlated	Independent	NHANES	Communities
Kernel SHAP							
Mean	0.0208	15.4	0.000139	0.00298	0.00324	0.0358	130.0
1st Quartile	0.0031	3.71	1.48e-05	0.00166	0.00163	0.0106	33.5
2nd Quartile	0.0103	8.19	3.86e-05	0.00249	0.00254	0.0221	53.6
3rd Quartile	0.029	20.1	0.000145	0.00354	0.00436	0.0418	132.0
Optimized Kernel SHAP							
Mean	0.00248	2.33	1.81e-05	0.000739	0.000649	0.00551	21.8
1st Quartile	0.000279	0.549	2.16e-06	0.00027	0.000187	0.000707	5.85
2nd Quartile	0.00138	1.26	5.43e-06	0.000546	0.000385	0.0024	13.0
3rd Quartile	0.0036	3.03	1.63e-05	0.00101	0.000964	0.00665	25.1
Leverage SHAP							
Mean	0.000186	0.63	5.21e-06	0.000458	0.000359	0.00385	14.7
1st Quartile	1.91e-05	0.0631	6.3e-07	0.000139	9.51e-05	0.000333	3.6
2nd Quartile	8.31e-05	0.328	2.33e-06	0.000376	0.000235	0.00149	8.9
3rd Quartile	0.000231	0.769	7.09e-06	0.000617	0.000556	0.00401	15.3

Leverage SHAP Guarantee

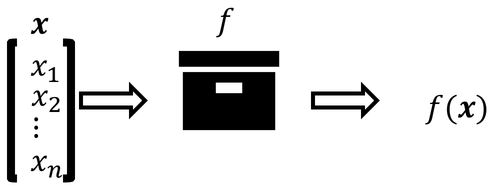
Lemma [MW '24]: Let $\gamma = \frac{||\boldsymbol{A}\,\boldsymbol{\phi} - \boldsymbol{b}||_2^2}{||\boldsymbol{A}\,\boldsymbol{\phi}||_2^2}$ and $\epsilon > 0$. With $O\left(n \log n + \frac{n}{\epsilon}\right)$ samples and with probability 99/100, the Leverage SHAP solution $\tilde{\boldsymbol{\phi}}$ satisfies

$$||\tilde{\boldsymbol{\phi}} - \boldsymbol{\phi}||_2^2 \leq \epsilon\,\gamma ||\boldsymbol{\phi}||_2^2$$

Intuition: We can accurately recover Shapley values, especially when the associated linear regression problem has a good solution

Explainable AI: Today

Goal: Attribute predictions to features

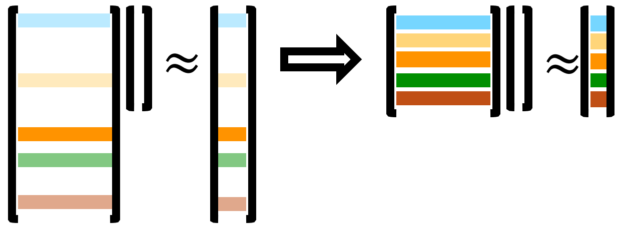


Approach: Use Shapley values

$$\phi_i = \frac{1}{n} \sum_{S \subseteq [n] \setminus \{i\}} \frac{v(S \cup \{i\}) - v(S)}{\binom{n-1}{|S|}}$$

Challenge: $O(2^n)$ terms

My Work: Apply leverage scores

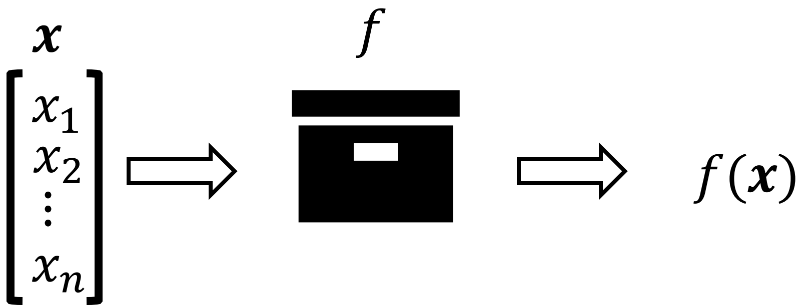


Result: Better Shapley approximation!

	California	Diabetes	Adult	Correlated	Independent	NHANES	Communities
Kernel SHAP							
Mean	0.0208	15.4	0.000139	0.00298	0.00324	0.0358	130.0
1st Quartile	0.0031	3.71	1.48e-05	0.00166	0.00163	0.0106	33.5
2nd Quartile	0.0103	8.19	3.86e-05	0.00249	0.00254	0.0221	53.6
3rd Quartile	0.029	20.1	0.000145	0.00354	0.00436	0.0418	132.0
Optimized Kernel SHAP							
Mean	0.00248	2.33	1.81e-05	0.000739	0.000649	0.00551	21.8
1st Quartile	0.000279	0.549	2.16e-06	0.00027	0.000187	0.000707	5.85
2nd Quartile	0.00138	1.26	5.43e-06	0.000546	0.000385	0.0024	13.0
3rd Quartile	0.0036	3.03	1.63e-05	0.00101	0.000964	0.00665	25.1
Leverage SHAP							
Mean	0.000186	0.63	5.21e-06	0.000458	0.000359	0.00385	14.7
1st Quartile	1.91e-05	0.0631	6.3e-07	0.000139	9.51e-05	0.000333	3.6
2nd Quartile	8.31e-05	0.328	2.33e-06	0.000376	0.000235	0.00149	8.9
3rd Quartile	0.000231	0.769	7.09e-06	0.000617	0.000556	0.00401	15.3

Explainable AI: Future

$$\phi_i = \frac{1}{n} \sum_{S \subseteq [n] \setminus \{i\}} \frac{v(S \cup \{i\}) - v(S)}{\binom{n-1}{|S|}}$$



Game Theory

Explainable AI

Game theory provides a rigorous foundation for explainable AI.

1. What is the “right” explanation technique?

2. How can we efficiently compute it?

Research Tools

Randomized algorithms

Leverage score sampling*, locality sensitive hashing, doubly robust estimators*

Classical optimization methods

Linear regression*, linear programming, semidefinite programs*

Deep optimization methods

Neural networks*, graph neural networks, diffusion, transformers*

Thank You

Please let me know if you have any questions, comments, and/or ideas for collaborations!

Email: rtealwitter@nyu.edu

Website: www.rtealwitter.com