

# Explainable AI & Leverage Score Sampling

R. Teal Witter

New York University

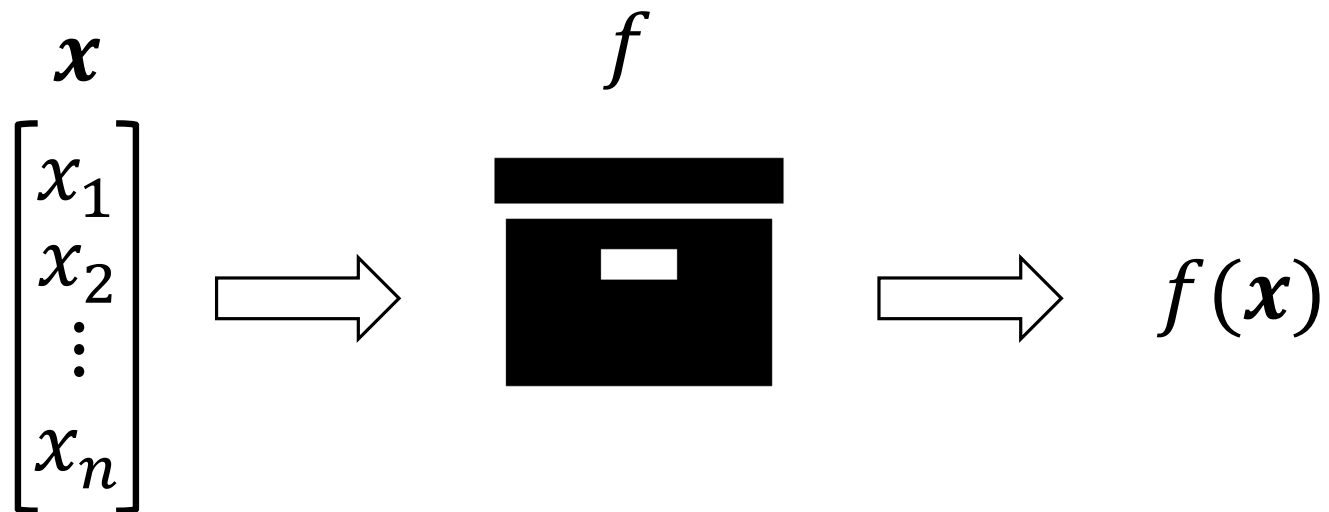
# Shapley Values & Leverage Score Sampling

Joint work with

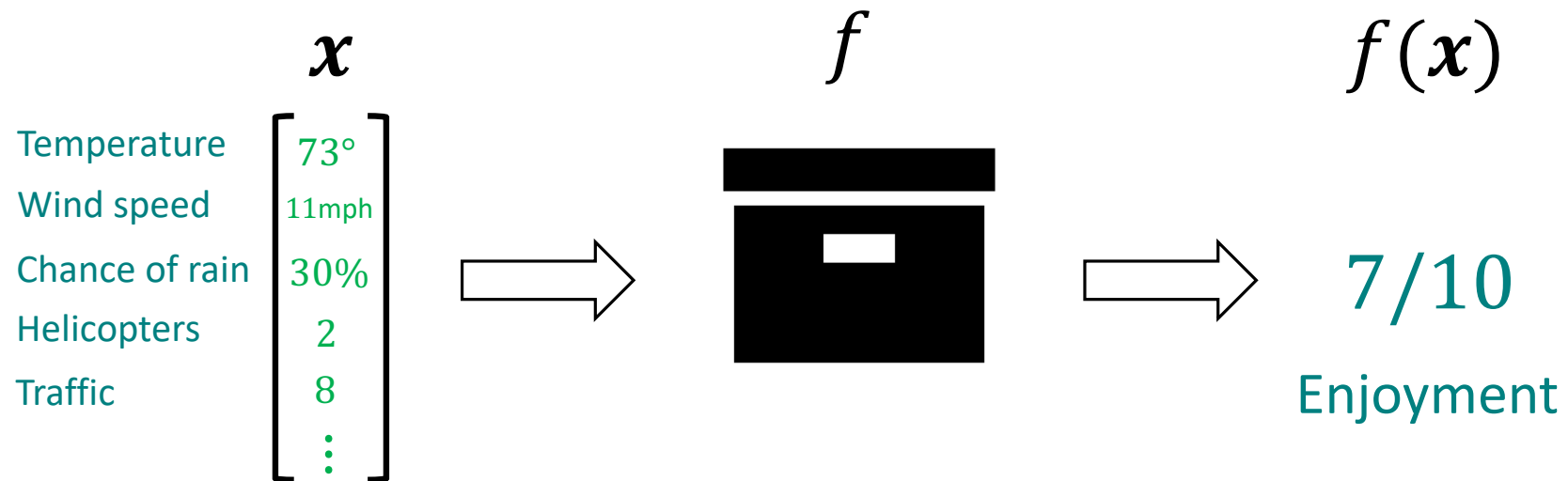


Christopher Musco  
New York University

# AI Prediction



Example: 



# Explaining Predictions

Attribute the prediction to features relative to a baseline



“Since the traffic is 8 instead of 3, the ride is 1.7 less enjoyable.”






Attribution value!

# Explaining Predictions

Attribute the prediction to features relative to a baseline



“Since the traffic is 8 instead of 3, the ride is 1.7 less enjoyable.”

								$f(x)$
Temperature	$\begin{bmatrix} 73^\circ \\ 11\text{mph} \\ 30\% \\ 2 \\ 8 \\ \vdots \end{bmatrix}$	$\begin{bmatrix} 89^\circ \\ 1\text{mph} \\ 0\% \\ 5 \\ 3 \\ \vdots \end{bmatrix}$	89°	11mph	30%	5	3	5/10
Wind speed			89°	11mph	30%	5	8	4/10
Chance of rain			73°	1mph	0%	5	3	6/10
Helicopters			73°	1mph	0%	5	8	8/10
Traffic								
	Explicand	Baseline						

# Desirable Properties

**Null Player:** *If a feature never changes the prediction, then its attribution value is 0*

**Symmetry:** *If two features always induce the same change, then their attribution values are the same*

**Additivity:** *For two predictive functions, the attribution value of a feature in the combined function is the sum of the attribution values for each function*






**Efficiency:** *The attribution values sum to the difference between the predictions on the explicand and baseline*

⇔ Shapley values!

*Lloyd Shapley received the Nobel Prize in Economics for formulating Shapley values.*

# Shapley Values for Feature Attribution

Let  $S \subseteq [n]$  and define  $v(S) = f(\mathbf{x}^S)$  where






$S$						$f(\mathbf{x}^S)$
$\{2,3\}$	89°	11mph	30%	5	3	5/10
$\{2,3,5\}$	89°	11mph	30%	5	8	4/10

*The SHAP paper (one of the first to use Shapley values in explainable AI) has 25k+ citations.*



# Shapley Values for Feature Attribution

Let  $S \subseteq [n]$  and define  $v(S) = f(\mathbf{x}^S)$  where

$S$						$f(\mathbf{x}^S)$
$\{2,3\}$	89°	11mph	30%	5	3	5/10
$\{2,3,5\}$	89°	11mph	30%	5	8	4/10

For a set function  $v: 2^{[n]} \rightarrow \mathbb{R}$ , the  $i$ th Shapley value is

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

# Estimating Shapley values

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

**Monte Carlo Sampling:** Sample  $S, S \cup \{i\}$  to use  $v(S \cup \{i\}) - v(S)$

... but samples only used for one Shapley value

**Maximum Reuse Sampling:** Sample  $S$  to either add/subtract  $v(S)$  for all  $i$

... but magnitude of  $v(S)$  is much larger than magnitude of  $v(S \cup \{i\}) - v(S)$

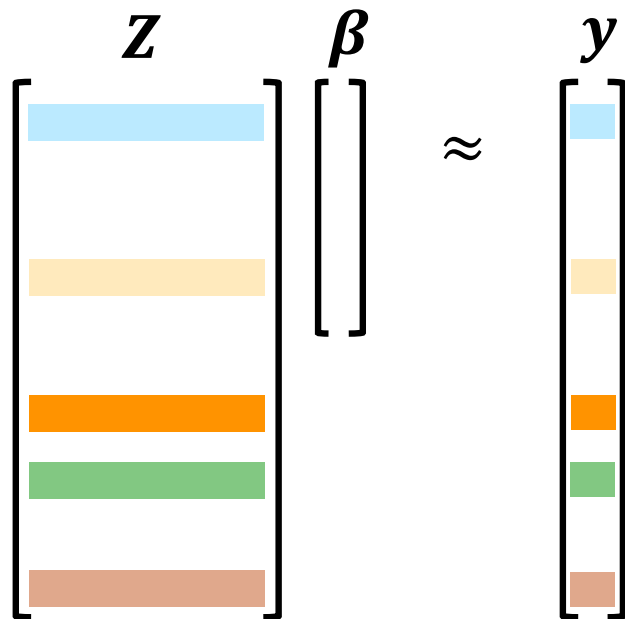
**Permutation Sampling:** Sample  $S_1 \subset S_2 \subset \dots \subset S_n$  to use  $v(S_{\ell+1}) - v(S_\ell)$

... but only 2x reuse

# Regression Formulation

Constraint to satisfy *efficiency* property

$$\phi = \arg \min_{\beta: \langle \beta, 1 \rangle = v([n]) - v(\emptyset)} ||\mathbf{Z}\beta - \mathbf{y}||_2$$



Connection known since 80's [CGKR 1988]

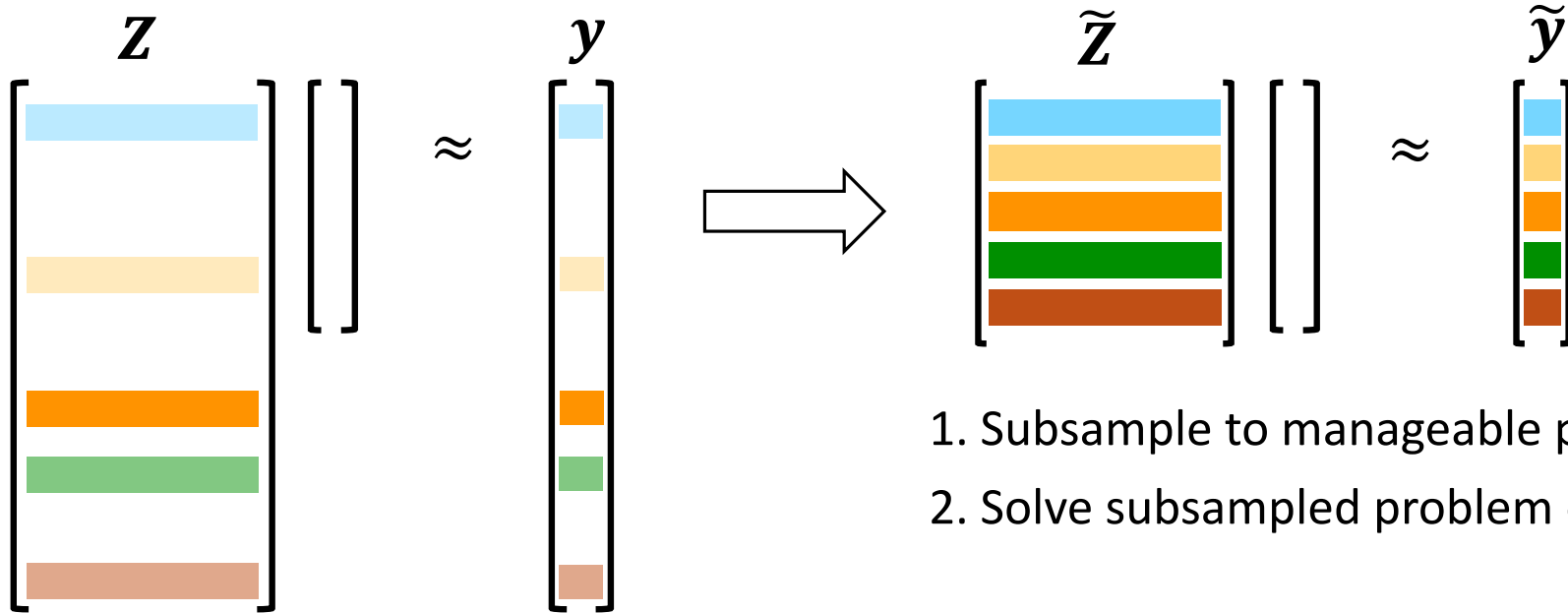
Each row/entry corresponds to binary vector

$$\mathbf{z} \in \{0,1\}^n : 0 < ||\mathbf{z}||_1 < n$$

Weighted by  $w(||\mathbf{z}||_1) = \frac{1}{\binom{n}{||\mathbf{z}||_1} (n - ||\mathbf{z}||_1) ||\mathbf{z}||_1}$

# Kernel SHAP

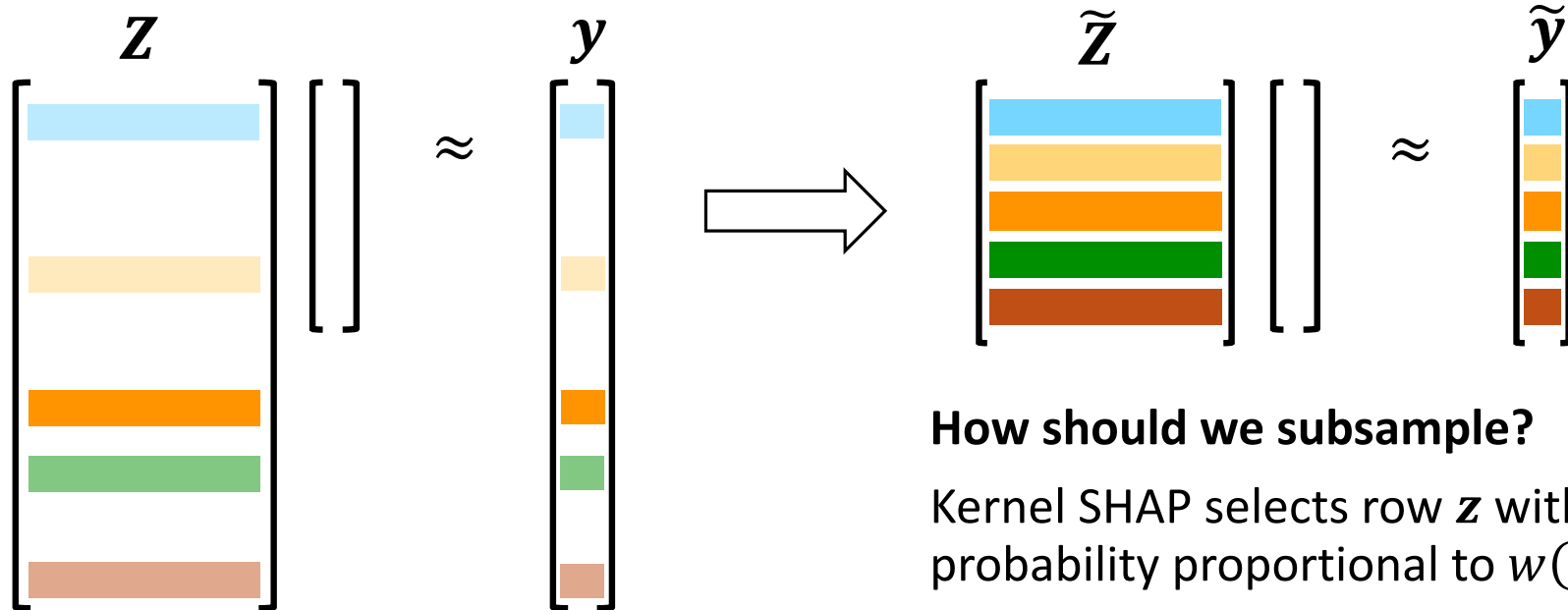
$$\tilde{\phi} = \arg \min_{\beta: \langle \beta, 1 \rangle = v([n]) - v(\emptyset)} ||\tilde{\mathbf{Z}}\beta - \tilde{\mathbf{y}}||_2$$



1. Subsample to manageable problem
2. Solve subsampled problem exactly

# Kernel SHAP

$$\tilde{\phi} = \arg \min_{\beta: \langle \beta, 1 \rangle = v([n]) - v(\emptyset)} ||\tilde{\mathbf{Z}}\beta - \tilde{\mathbf{y}}||_2$$



# Constrained to Unconstrained Regression

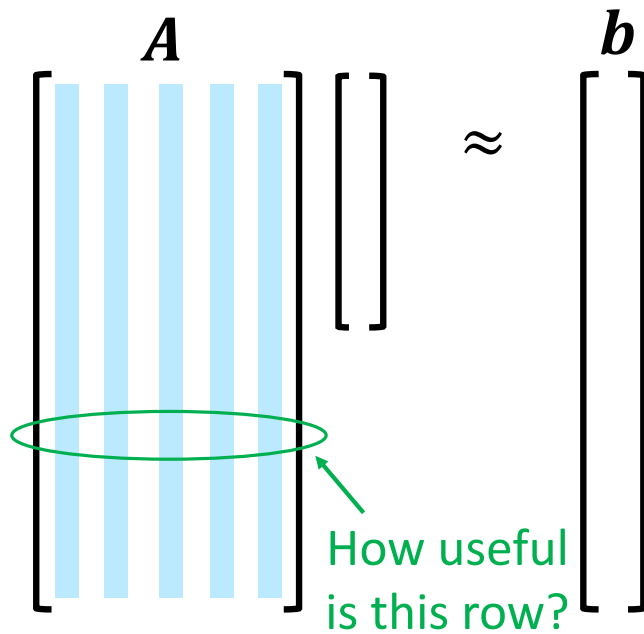
$$\begin{aligned}\phi &= \arg \min_{\beta: \langle \beta, \mathbf{1} \rangle = v([n]) - v(\emptyset)} ||\mathbf{Z}\beta - \mathbf{y}||_2 \\ &= \arg \min_{\beta} ||\mathbf{A}\beta - \mathbf{b}||_2 + \mathbf{1} \frac{v([n]) - v(\emptyset)}{n}\end{aligned}$$

- By constraint, we know the component in the  $\mathbf{1}$  direction
- Only optimize to residual target in space orthogonal to  $\mathbf{1}$

Formulate as unconstrained problem so we can apply our favorite tools!



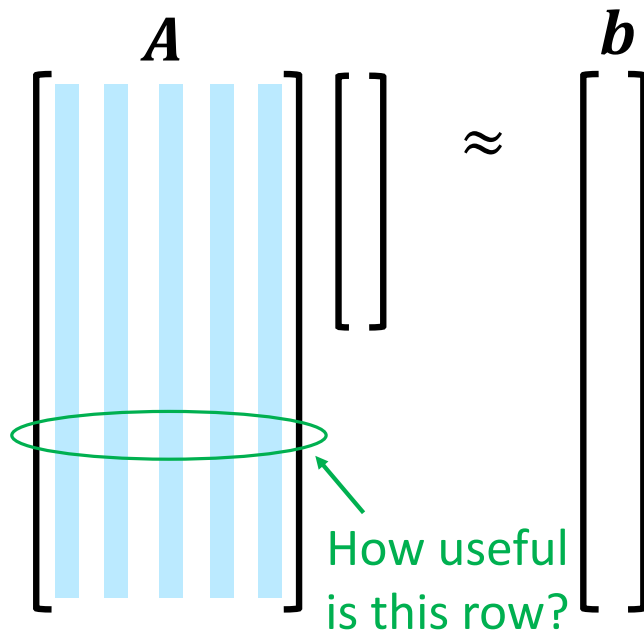
# Leverage Scores



Row  $\mathbf{z}$  has “leverage”:

$$\ell_{\mathbf{z}} = \mathbf{A}_{\mathbf{z}}(\mathbf{A}_{\mathbf{z}}^{\top} \mathbf{A}_{\mathbf{z}})^+ \mathbf{A}_{\mathbf{z}}^{\top}$$

# Leverage Scores and Shapley values



Row  $z$  has “leverage”:

$$\ell_z = A_z(A_z^\top A_z)^+ A_z^\top$$

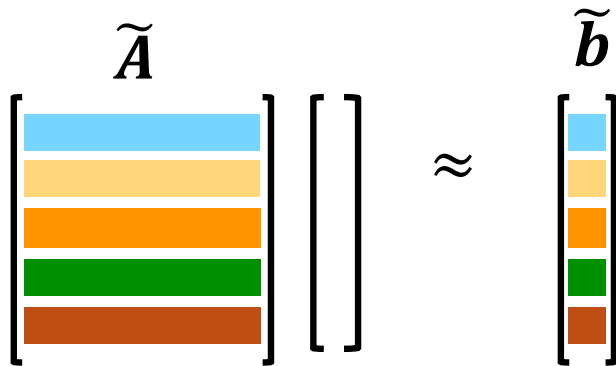
$$\ell_z = \left( \frac{n}{||z||} \right)^{-1}$$

Very similar to weighting in Shapley value definition!



# Leverage SHAP

$$\tilde{\phi} = \arg \min_{\beta} ||\tilde{A}\beta - \tilde{b}||_2 + \mathbf{1} \frac{v([n]) - v(\emptyset)}{n}$$



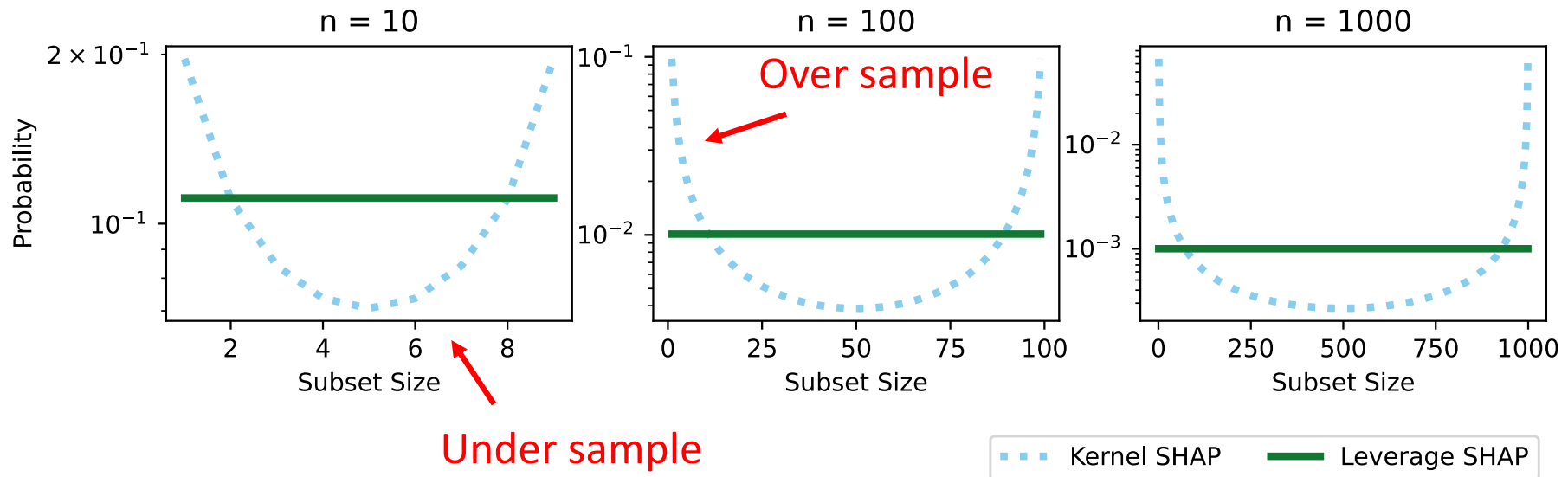
Leverage SHAP selects row  $\mathbf{z}$  with probability proportional to leverage score!

+ Paired Sampling

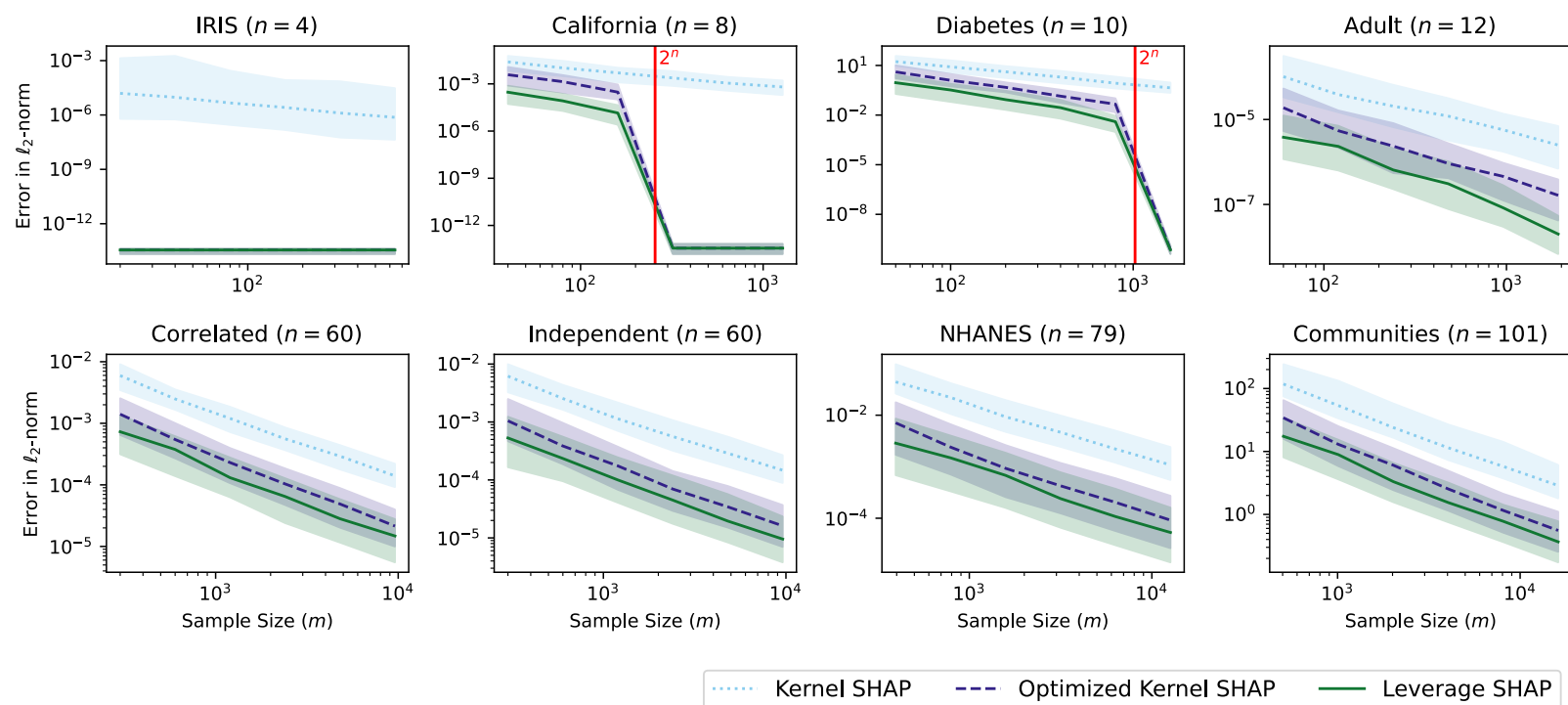
+ Bernoulli Sampling

# Leverage SHAP vs Kernel SHAP Probabilities

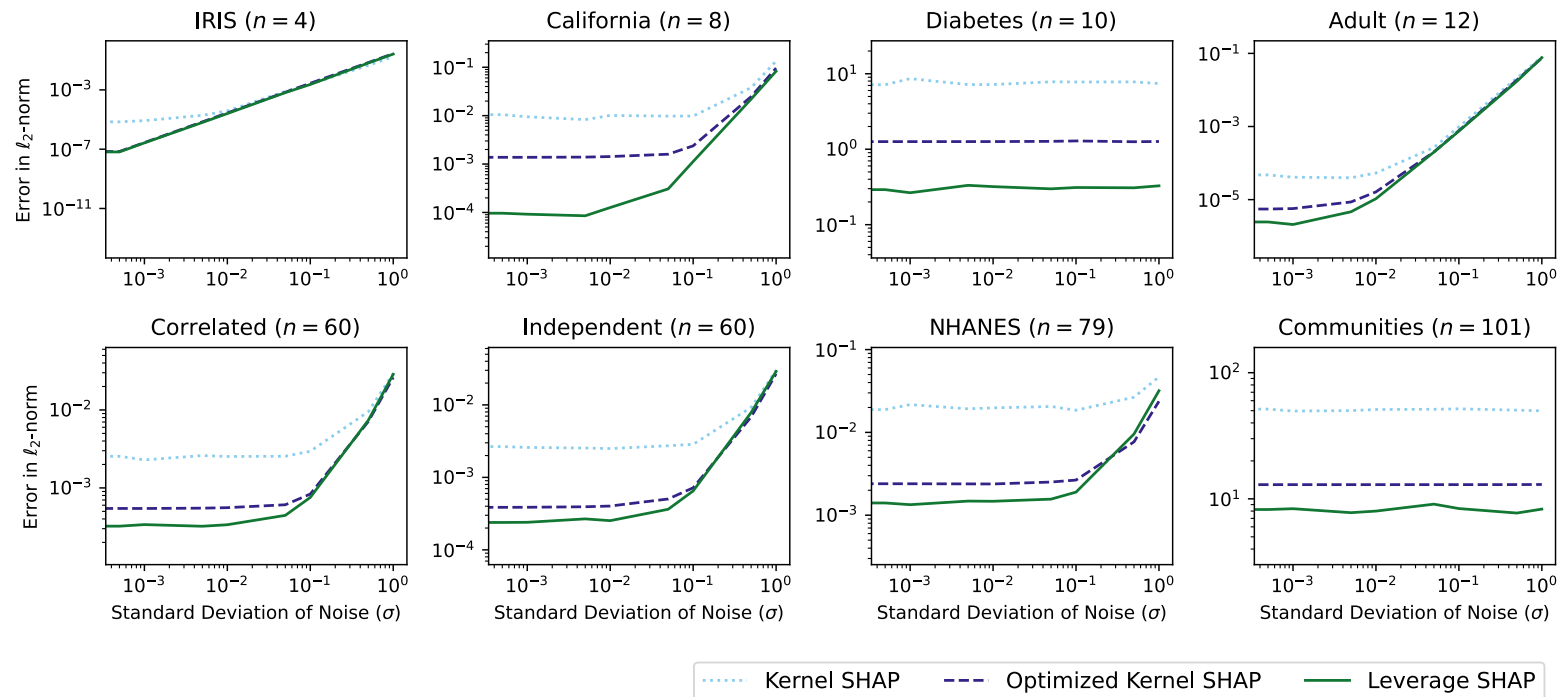
Kernel SHAP and Leverage SHAP Probability Distributions



# Accuracy by Sample Size



# Accuracy by Noise



Robustness is useful, e.g.,  $v(S) = \mathbb{E}_{x^S}[f(x^S)]$

# Theoretical Guarantee

As long as  $m = \tilde{O}\left(\frac{n}{\epsilon}\right)$ , the Leverage SHAP solution  $\tilde{\boldsymbol{\phi}}$  satisfies

$$||\mathbf{A}\tilde{\boldsymbol{\phi}} - \mathbf{b}||_2^2 \leq (1 + \epsilon)||\mathbf{A}\boldsymbol{\phi} - \mathbf{b}||_2^2$$

with probability 9/10.

Guarantee similar to standard leverage analysis but proof requires

- Modifications for paired sampling
- Modifications for sampling without replacement

# Interpretable Corollary

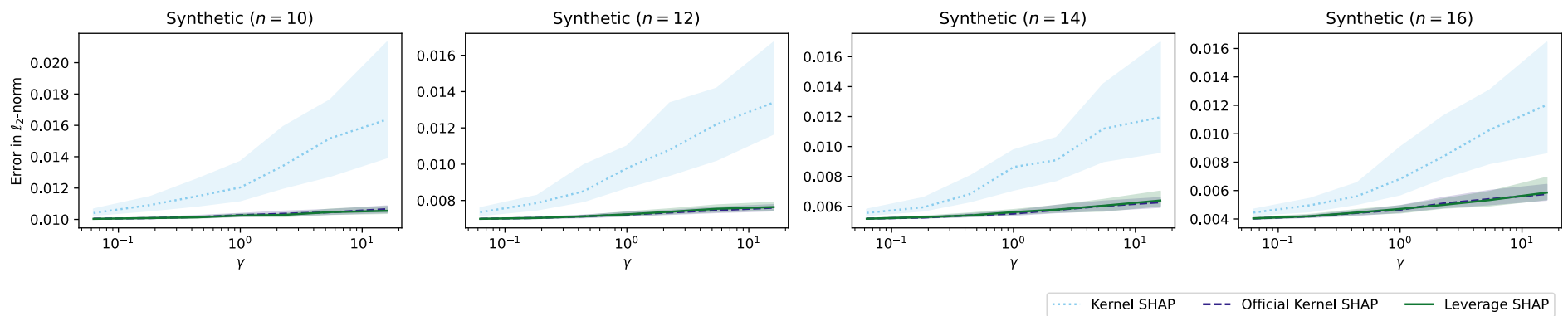
As long as  $m = \tilde{O}\left(\frac{n}{\epsilon}\right)$ , the Leverage SHAP solution  $\tilde{\phi}$  satisfies

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

with probability 9/10 where  $\gamma = \frac{\|A\phi - b\|_2^2}{\|A\phi\|_2^2} \in [0, \infty)$

**Intuition:** We can find  $\tilde{\phi}$  close to the optimal in objective value but, when optimal solution is bad,  $\tilde{\phi}$  will be far from  $\phi$

# $\gamma$ in Practice\*

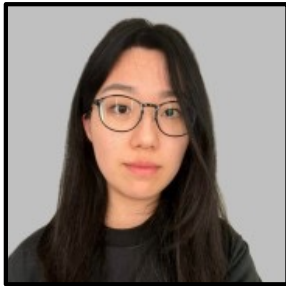


**Takeaway:**  $\gamma$  is a parameter of regression (not artifact of analysis)

\*Computing  $\gamma$  requires exponential time so experiments are small

# Banzhaf Values & Leverage Score Sampling

Joint work  
with



Yurong Liu  
NYU



Filip Korn  
Google



Tarfah Alrashed  
Google



Dimitris Paparas  
Google



Juliana Freire  
NYU



# Desirable Properties

**Null Player:** *If a feature never changes the prediction, then its attribution value is 0*

**Symmetry:** *If two features always induce the same change, then their attribution values are the same*

**Additivity:** *For two predictive functions, the attribution value of a feature in the combined function is the sum of the attribution values for each function*

**- Efficiency:** *The attribution values sum to the difference between the predictions on the explicand and baseline*

**+ 2-Efficiency:** *If two features are combined, their combined attribution value is the sum of the features' individual attribution values*



**Banzhaf values!**

*John Banzhaf is an activist lawyer, he used Banzhaf values to argue a Nassau County voting system was unfair.*

# Banzhaf Values

For a set function  $v: 2^{[n]} \rightarrow \mathbb{R}$ , the  $i$ th Banzhaf value is

$$\phi_i = \frac{1}{2^{n-1}} \sum_{S \in [n] \setminus \{i\}} v(S \cup \{i\}) - v(S)$$

Banzhaf values are

- Simpler
- Empirically easier to approximate

## Estimating Banzhaf values

$$\phi_i = \frac{1}{2^{n-1}} \sum_{S \in [n] \setminus \{i\}} v(S \cup \{i\}) - v(S)$$

**Monte Carlo (MC):** Sample  $S, S \cup \{i\}$  to use  $v(S \cup \{i\}) - v(S)$

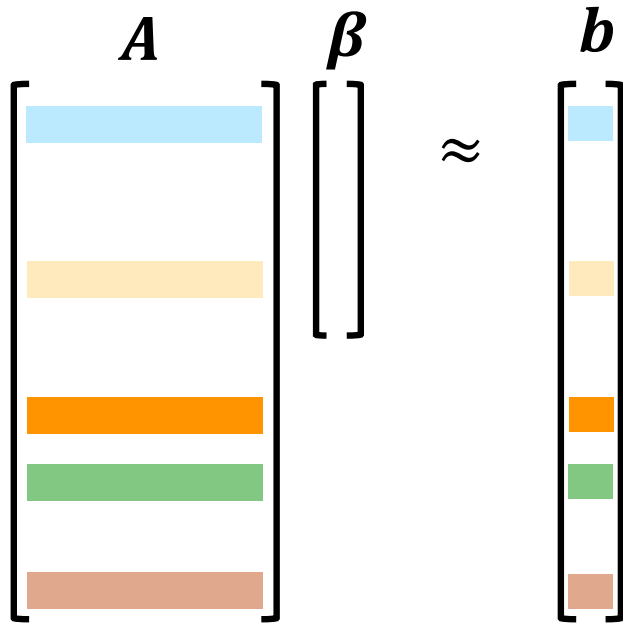
... but samples only used for one Shapley value

**Maximum Sampling Reuse (MSR):** Sample  $S$  to either add/subtract  $v(S)$  for all  $i$

... but magnitude of  $v(S)$  is much larger than magnitude of  $v(S \cup \{i\}) - v(S)$

# Regression Formulation

$$\phi = \arg \min_{\beta} ||A\beta - b||_2$$

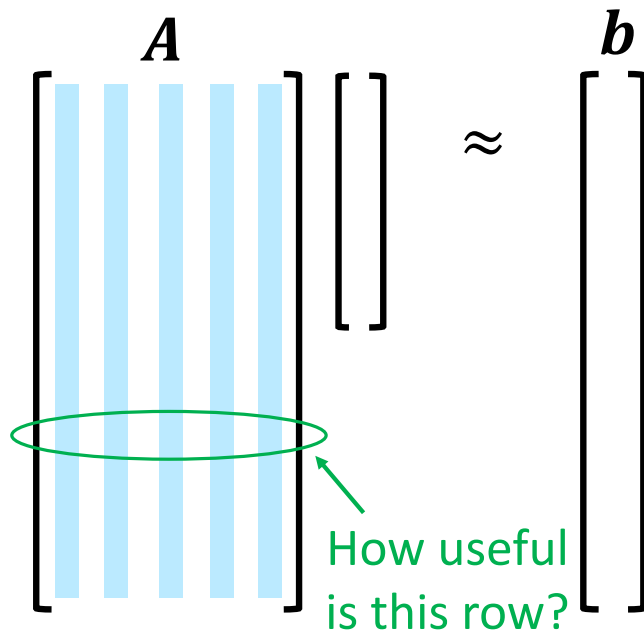


Special case known since 90's [HH 1992]

Each row/entry corresponds to binary vector

$$\mathbf{z} \in \left\{ -\frac{1}{2}, \frac{1}{2} \right\}^n$$

# Leverage Scores and Banzhaf values



Row  $z$  has “leverage”:

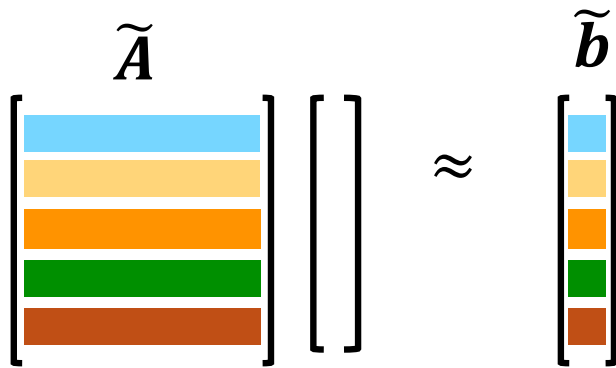
$$\ell_z = A_z(A_z^\top A_z)^{-1} A_z^\top$$

$$\ell_z = \frac{n}{2n}$$

Very similar to weighting in Banzhaf value definition!

# Kernel Banzhaf

$$\tilde{\phi} = \arg \min_{\beta} ||\tilde{A}\beta - \tilde{b}||_2$$

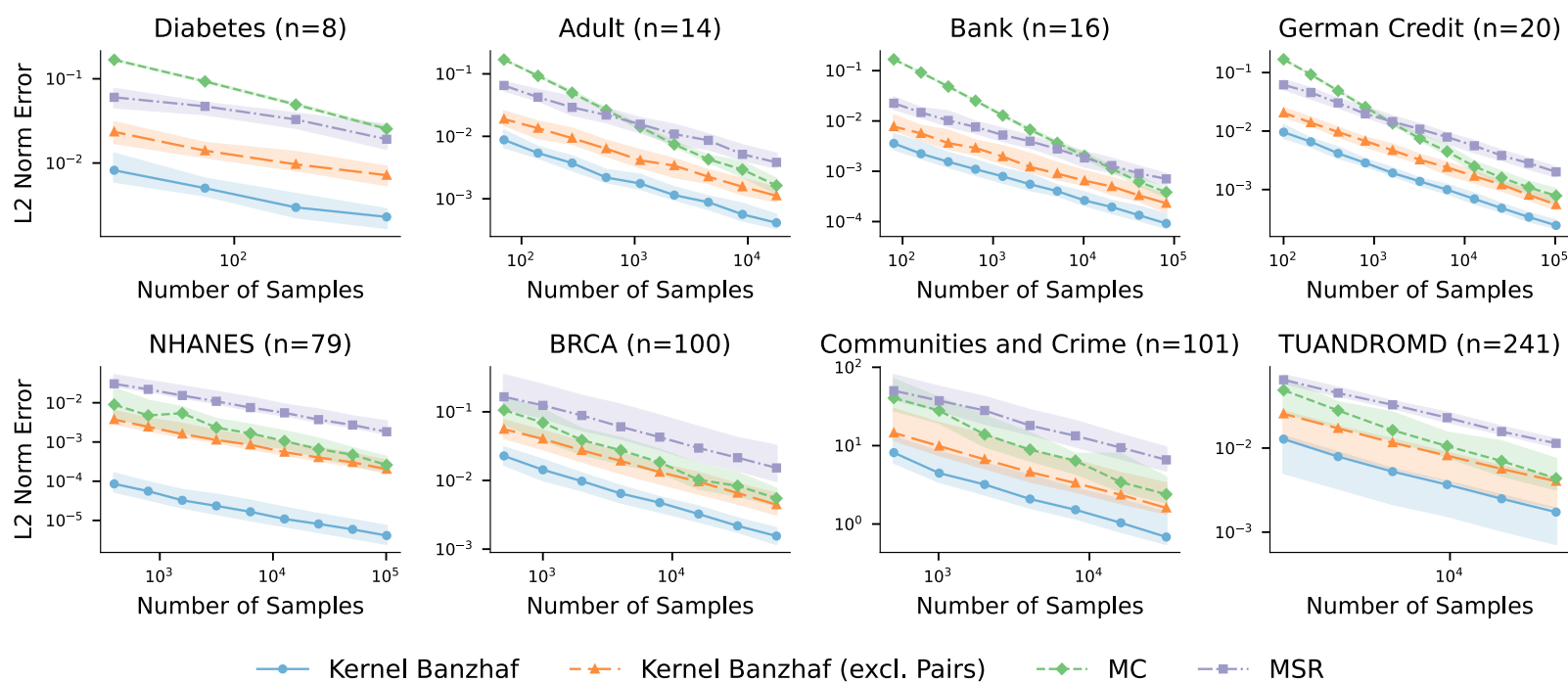


Kernel Banzhaf selects row  $z$  with probability proportional to leverage score!

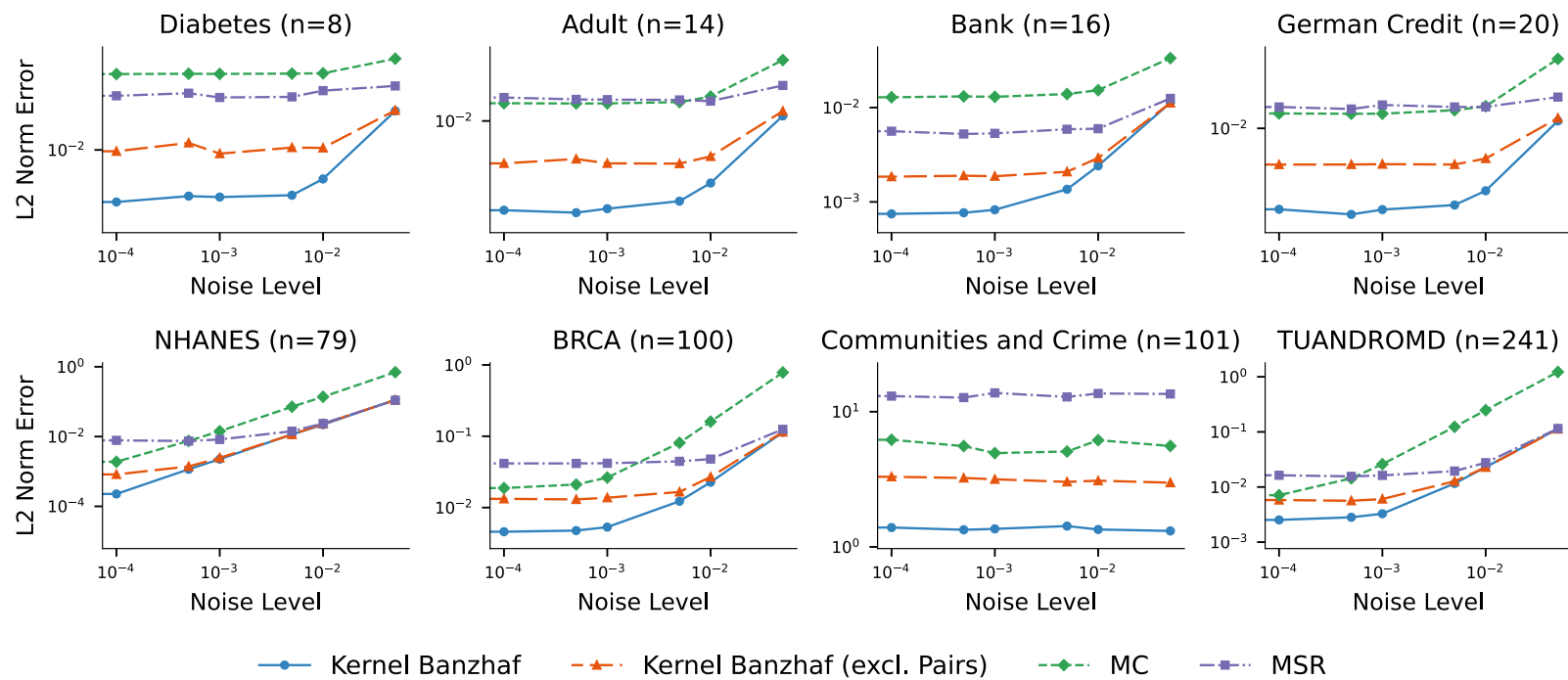
**+ Paired Sampling**

(But not Bernoulli Sampling)

# Accuracy by Number of Samples



# Accuracy by Noise





# Theoretical Guarantees

As long as  $m = \tilde{O}\left(\frac{n}{\epsilon}\right)$ , the Kernel Banzhaf solution  $\tilde{\boldsymbol{\phi}}$  satisfies

$$\|\mathbf{A}\tilde{\boldsymbol{\phi}} - \mathbf{b}\|_2^2 \leq (1 + \epsilon)\|\mathbf{A}\boldsymbol{\phi} - \mathbf{b}\|_2^2$$

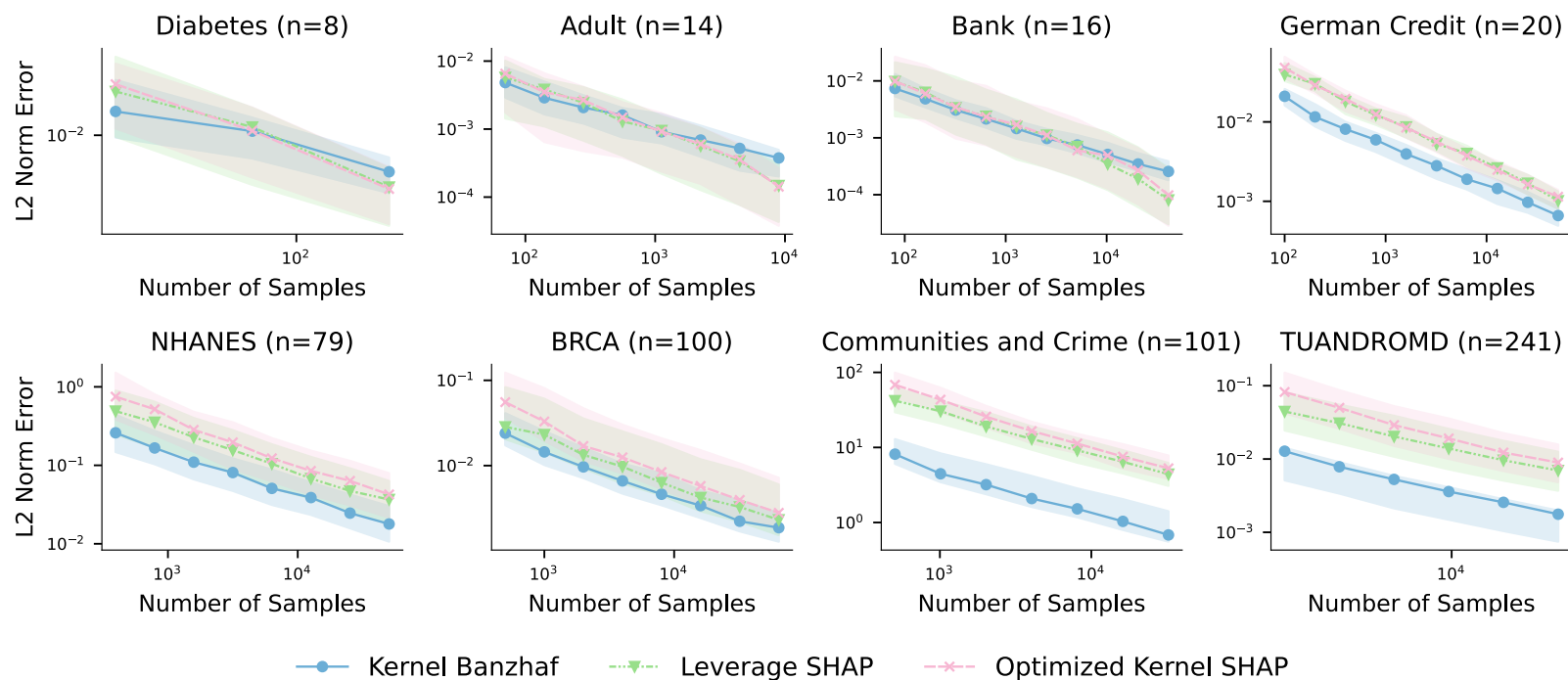
and, for  $\gamma = \frac{\|\mathbf{A}\boldsymbol{\phi} - \mathbf{b}\|_2^2}{\|\mathbf{A}\boldsymbol{\phi}\|_2^2} \in [0, \infty)$ , ↕ Equivalent for Banzhaf values

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

with probability 9/10

Guarantee similar to standard leverage analysis but proof requires modifications for paired sampling

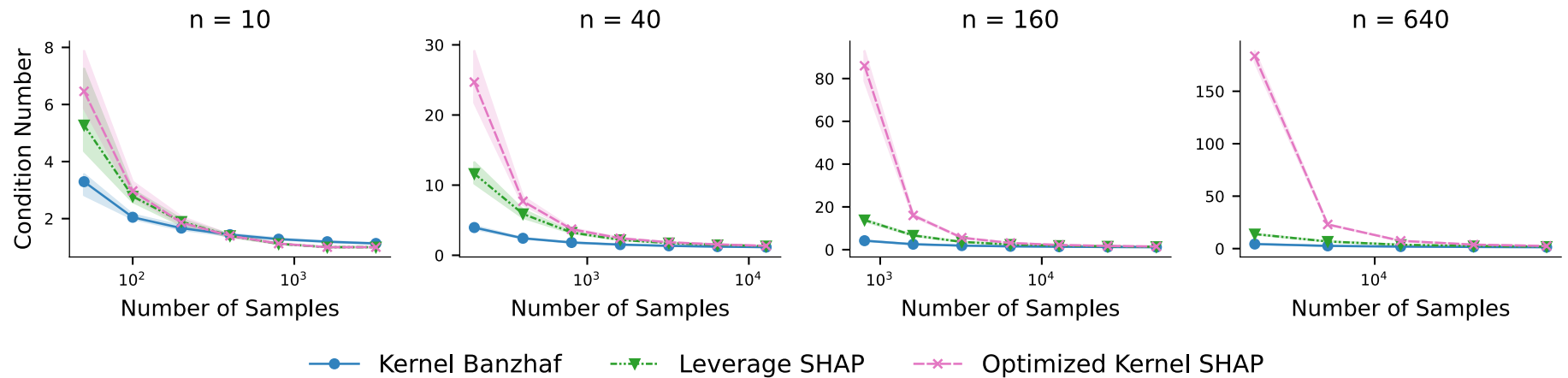
# Shapley vs Banzhaf Estimators



Why do Leverage SHAP and Kernel Banzhaf perform differently?

# Condition Number of $\tilde{\mathbf{A}}$

Subsampled Banzhaf problem is more well-conditioned.



Since the full problem is well-conditioned, a subsampled problem with large condition number means weights won't generalize.

# Thank you!

**Any questions or comments?**

*Let me know! I'm submitting both papers tonight!*

**Any ideas for future work?**

*At 7am ET tomorrow, I'll start looking for new projects 😊*

**Any feedback on presentation?**

*I plan to present this as my job talk!*