Scalable Attribution and Selection for Machine Learning via Spectral Methods

Landon Butler
Ph.D. Student, Berkeley EECS

Collaborators

Justin Kang

Efe Erginbas

Kannan Ramchandran

Bin Yu

Abhineet Agarwal

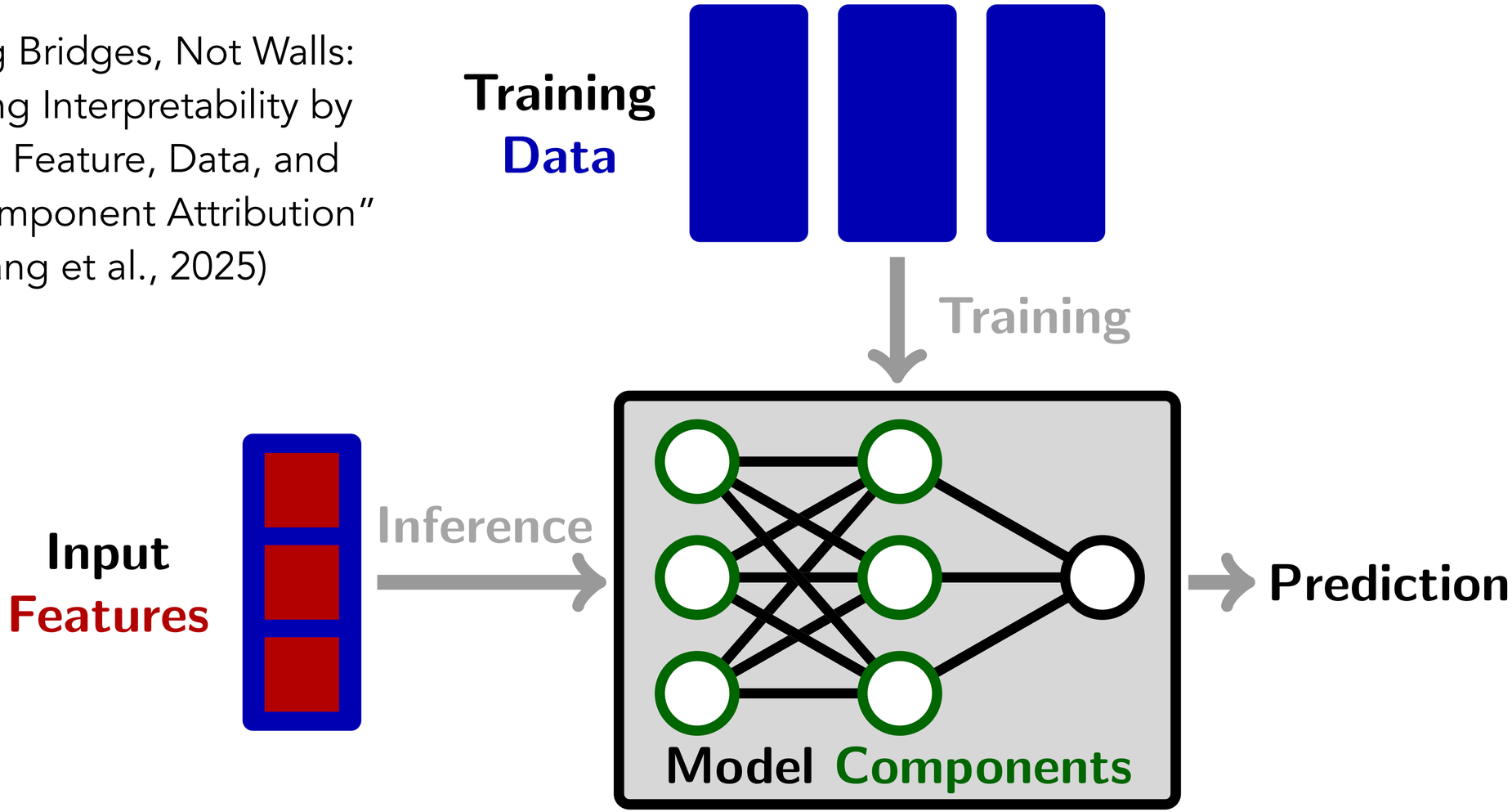
Ramtin Pedarsani

Fabian Fumagalli

R. Teal Witter

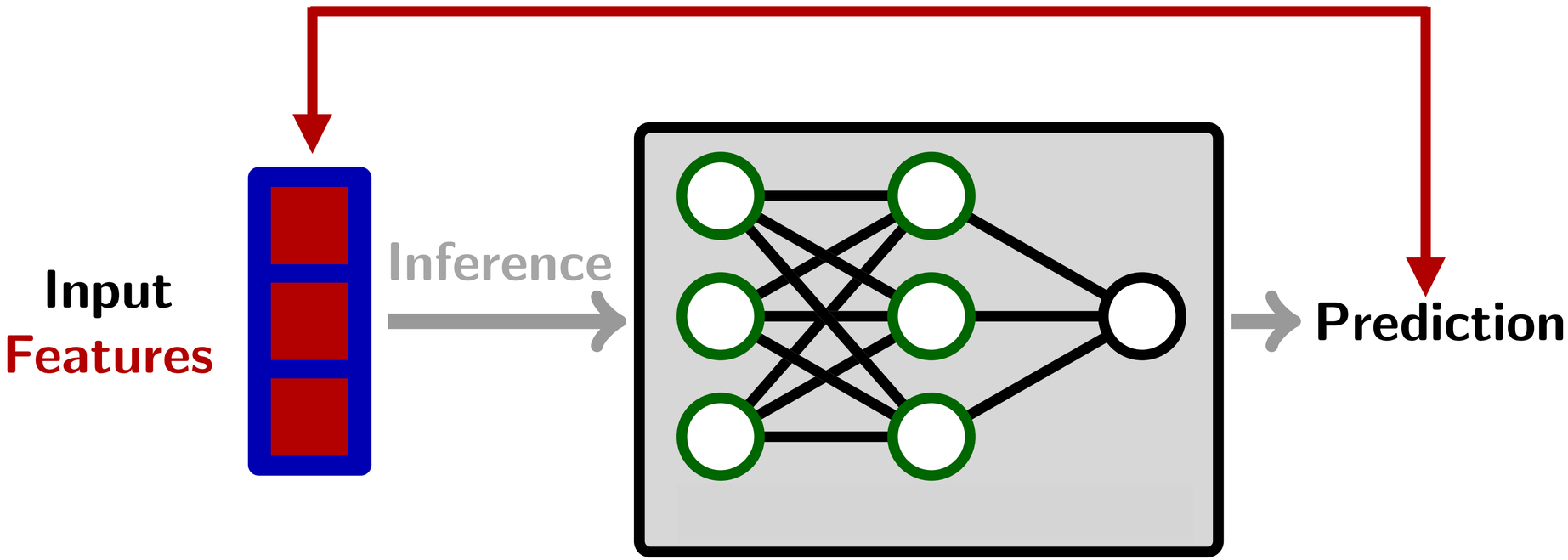
Machine Learning Systems

"Building Bridges, Not Walls: Advancing Interpretability by Unifying Feature, Data, and Model Component Attribution" (Zhang et al., 2025)



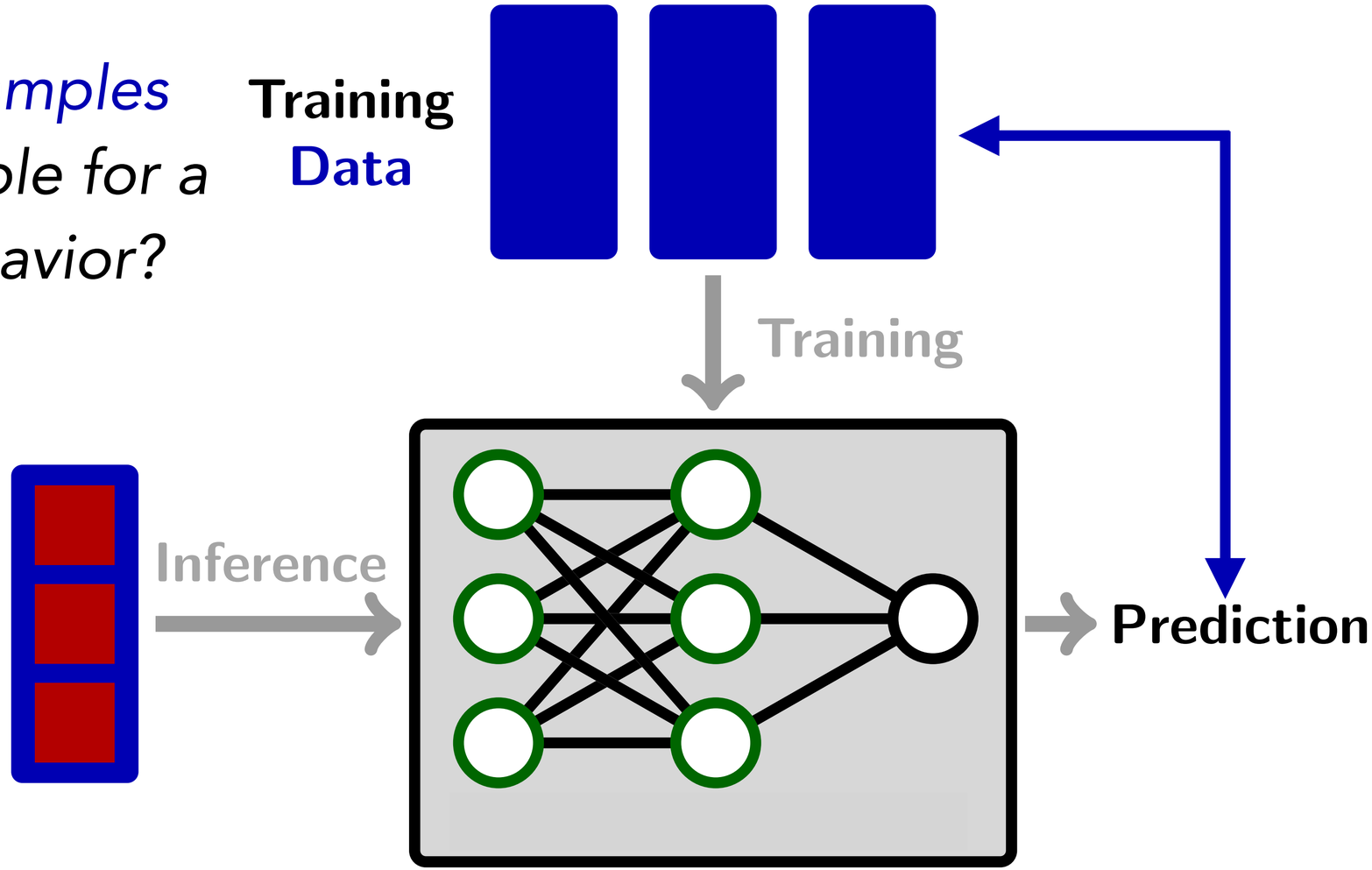
Feature Attribution

Which input features are most responsible for a given model behavior?



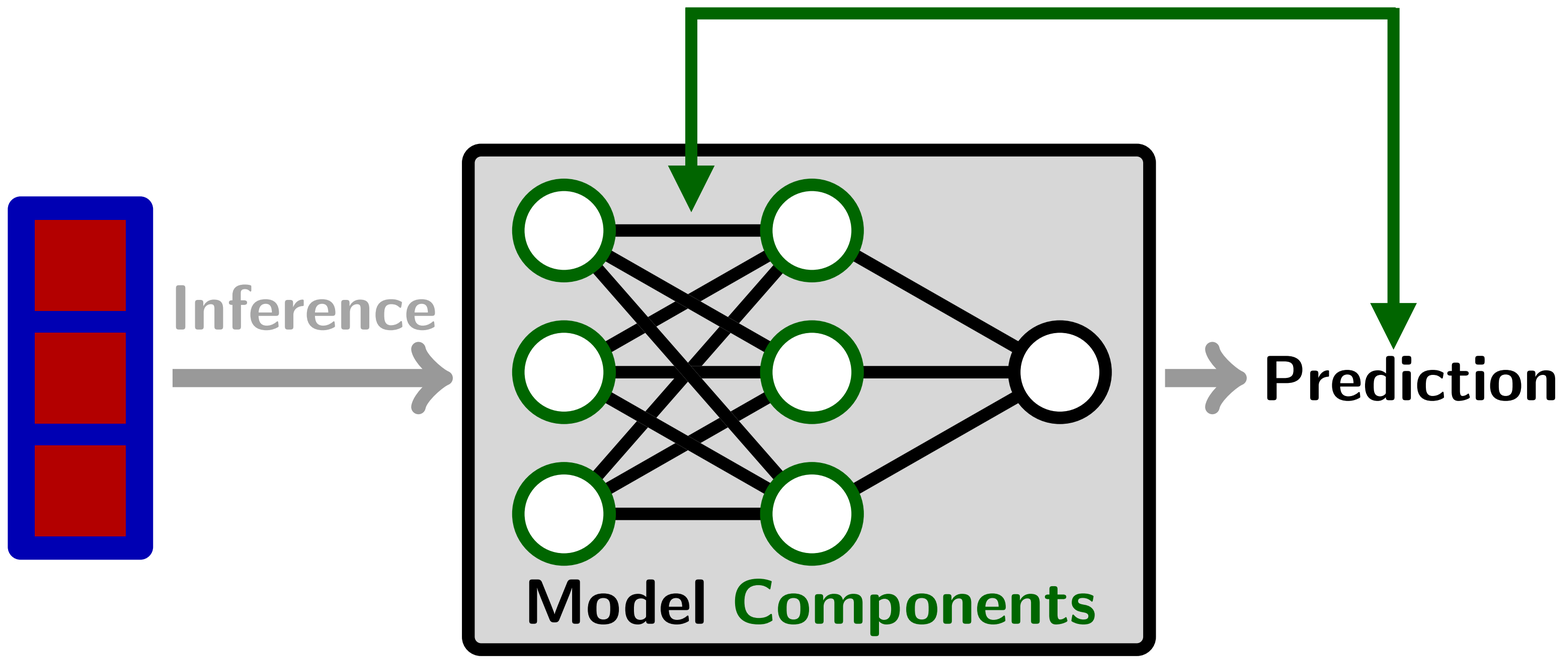
Data Attribution

Which training samples are most responsible for a given model behavior?



Model Component Attribution

Which model components are most responsible for a given model behavior?



History of Attribution

Feature

Causal Inference: Do-calculus (Pearl, 1985), Potential outcomes (Neyman, 1923; Rubin, 1974)

Interpretable By-Design: Decision Trees (e.g. Breiman, 1984), GAMs (Hastie & Tibshirani, 1986)

NN Grad-Based Methods: Saliency maps (Simonyan et al., 2013), Int. Gradients (Sundararajan et al., 2017)

Black-Box Model Agnostic: SHAP (Lundberg & Lee, 2017), LIME (Ribeiro et al., 2016)

Data

Data Attribution: Influence Functions (Koh & Liang, 2017), DataModels (Ilyas et al., 2022)

Comp.

Component Attribution: Linear Probes (Alain & Bengio, 2016), Circuit Discovery (Wang et al., 2022)

+ many applications

Attribution and Selection

Techniques for ***attribution*** can also inform strategies for ***selection***

e.g. top-k selection

LLM-Related Selection Applications:

Feature Selection: Prompt Compression (Li et al., 2025), RAG (Lewis et al., 2020)

Data Selection: Data Curation (Raffel et al., 2020), Curriculum Learning (Bengio et al., 2009)

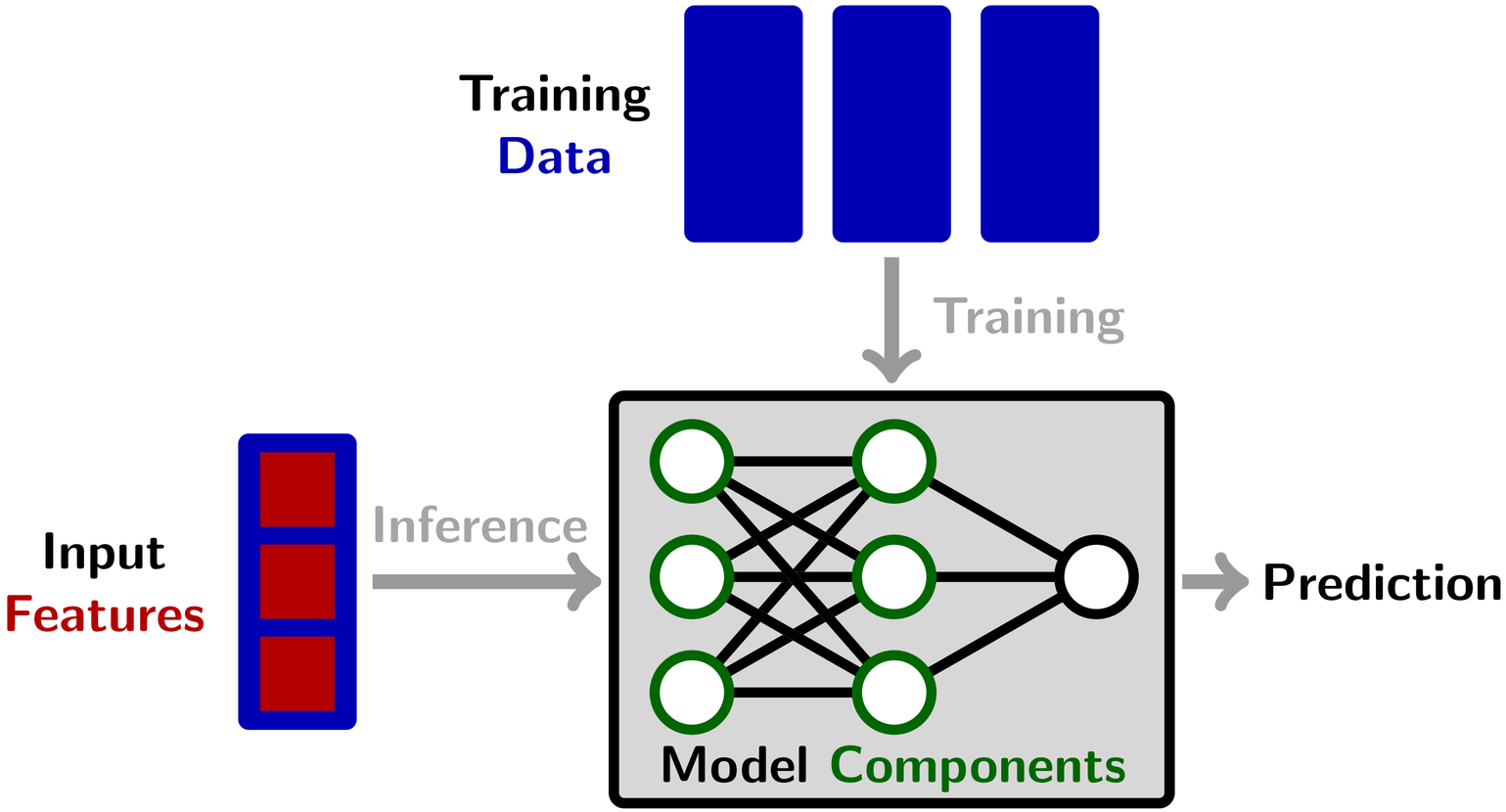
Component Selection: Model Pruning (Ma et al., 2023), Steering (Turner et al., 2023)

Design and apply *scalable* algorithms for attribution and selection across these three lenses

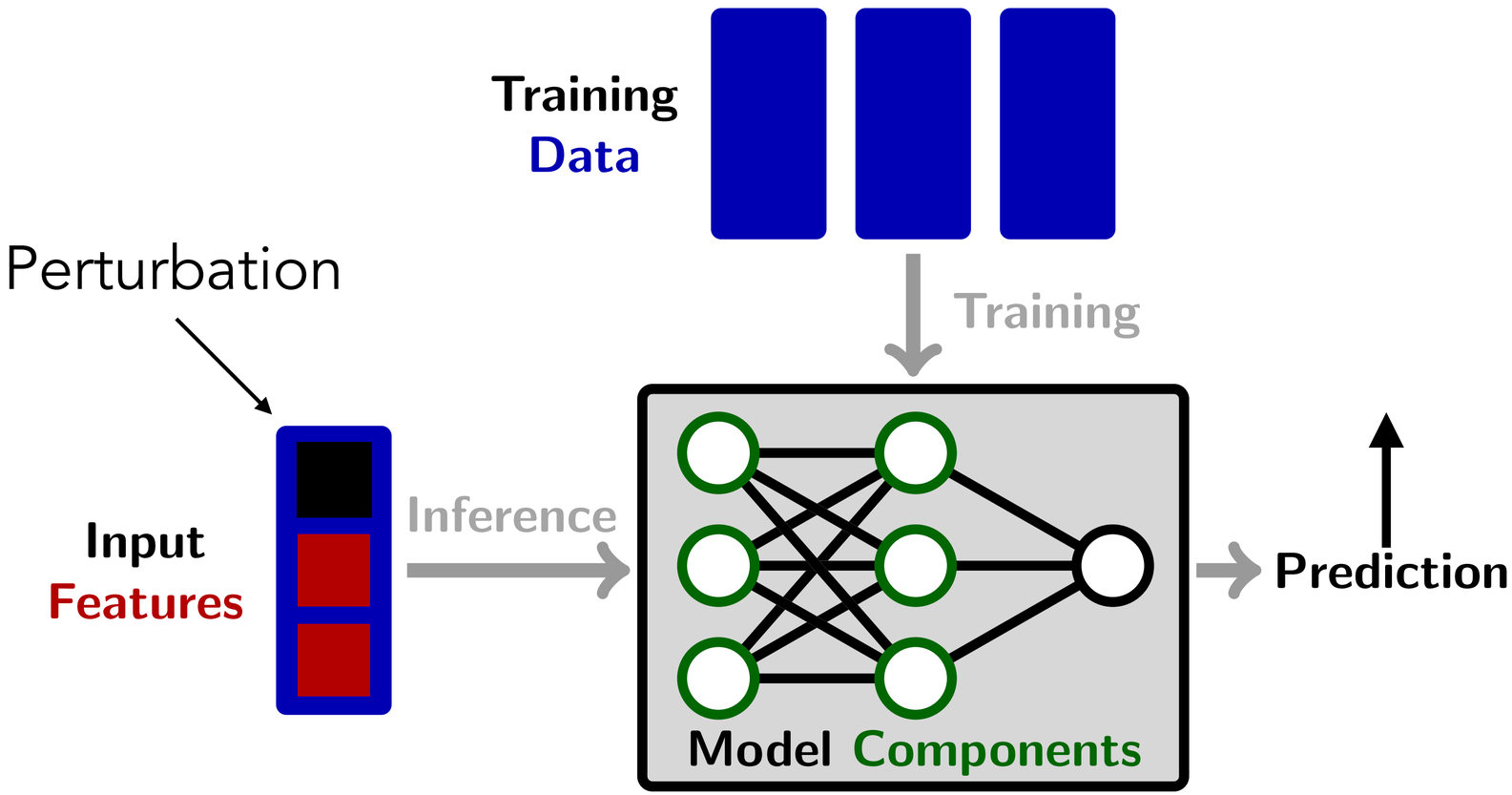
(black-box treatment until component lense)

Framework

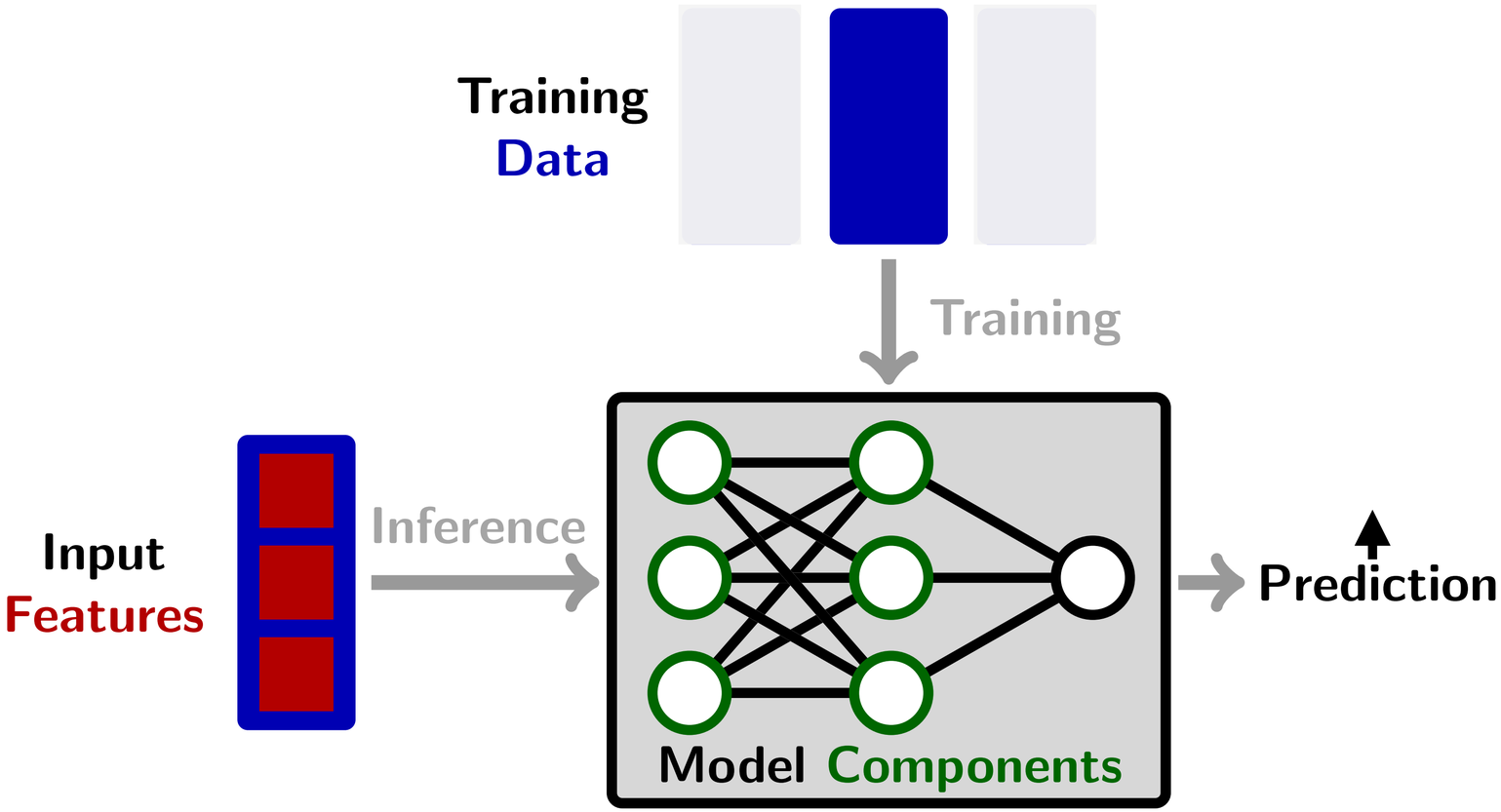
Perturbation



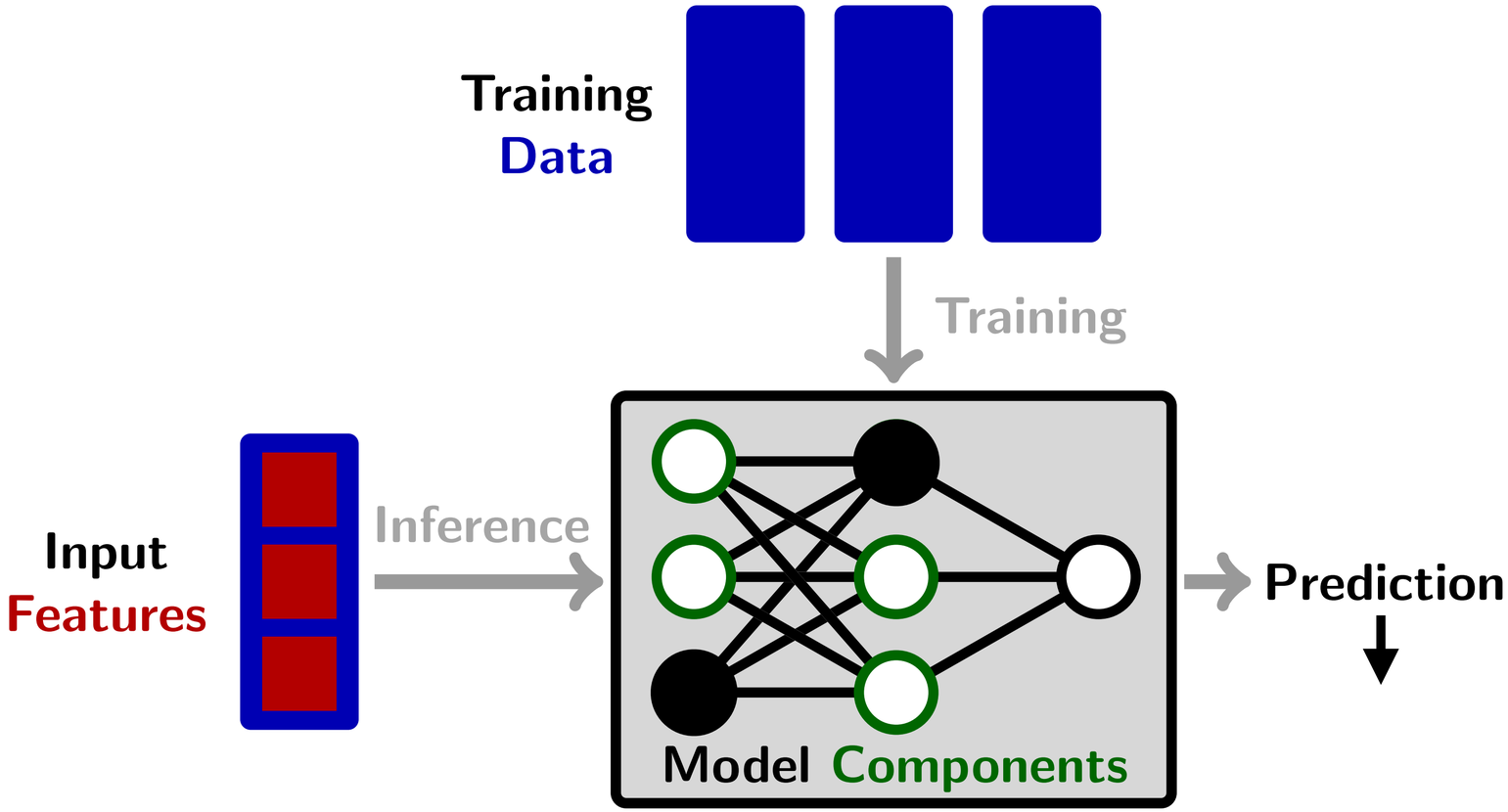
Feature Perturbation



Data Perturbation

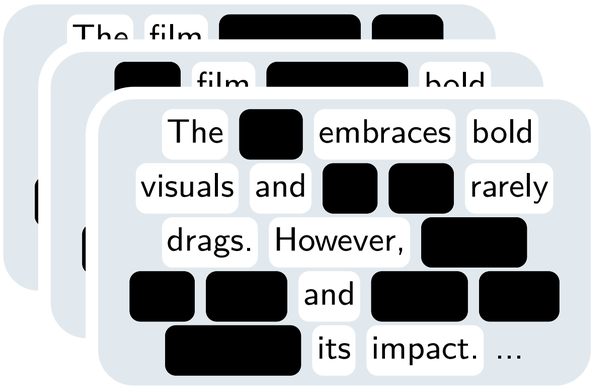


Component Perturbation

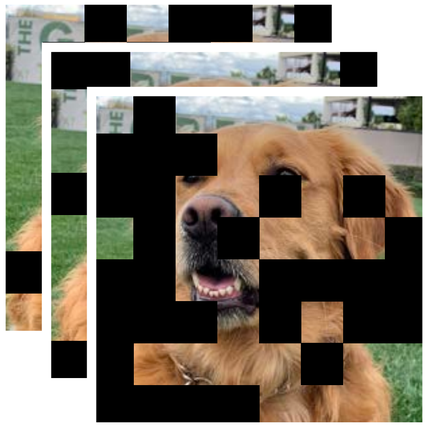


Perturbation

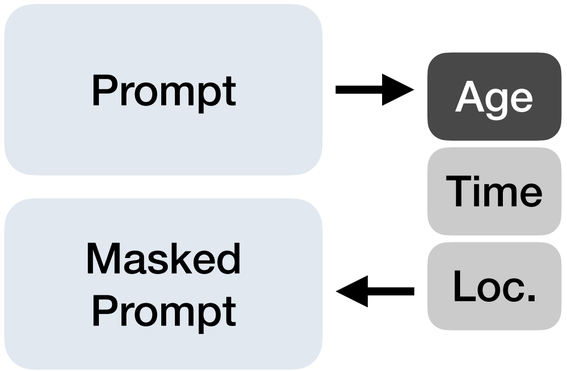
Feature



Text



Vision

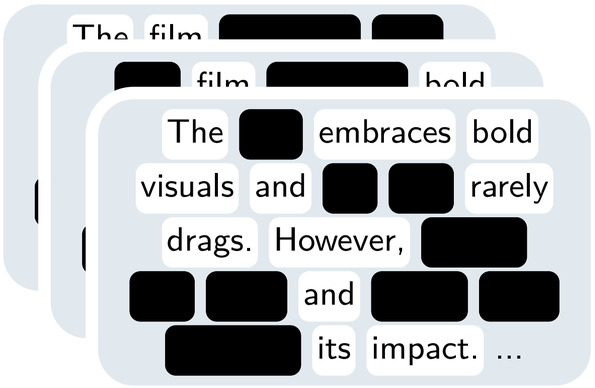


Concept

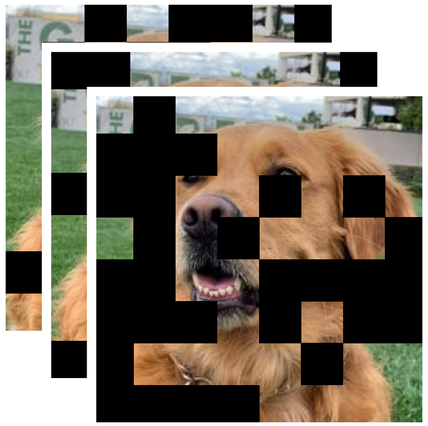
Key Challenge:
In-distribution perturbations

Perturbation

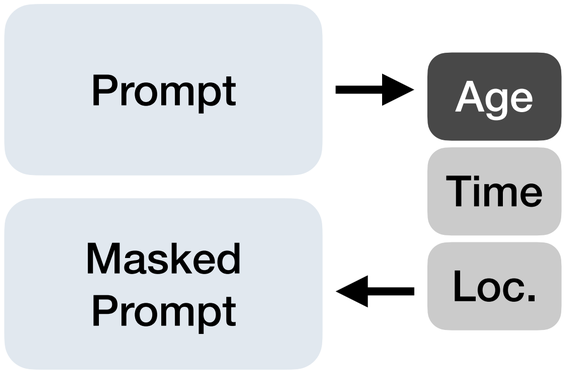
Feature



Text



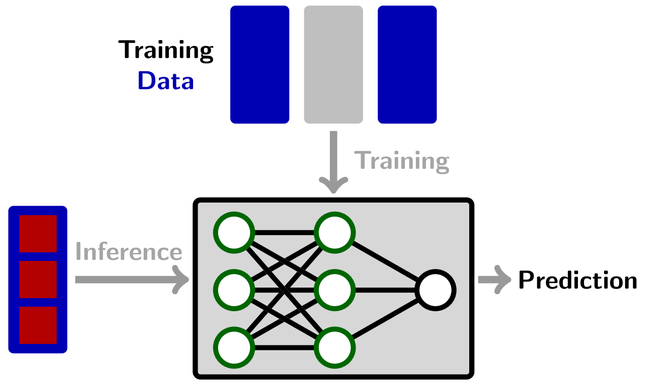
Vision



Concept

Key Challenge:
In-distribution perturbations

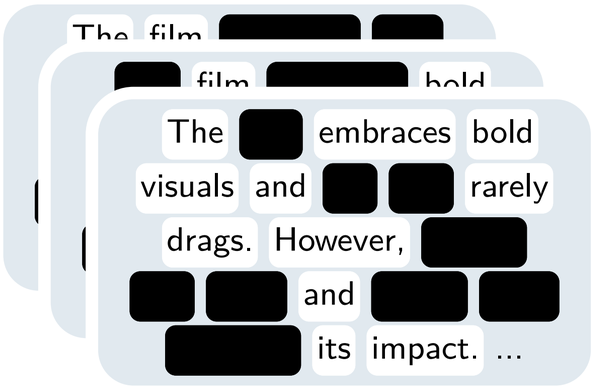
Data



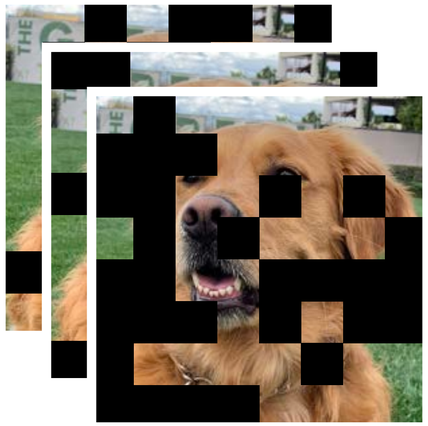
Key Challenge:
Retraining Cost

Perturbation

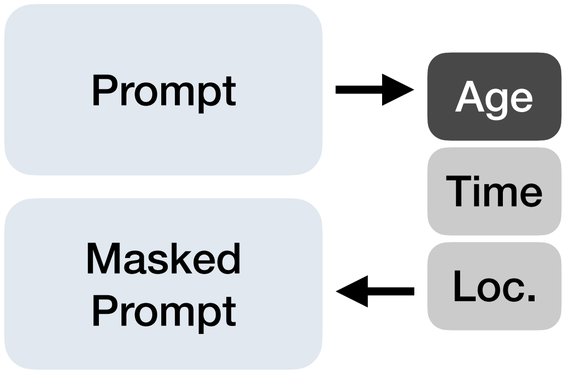
Feature



Text



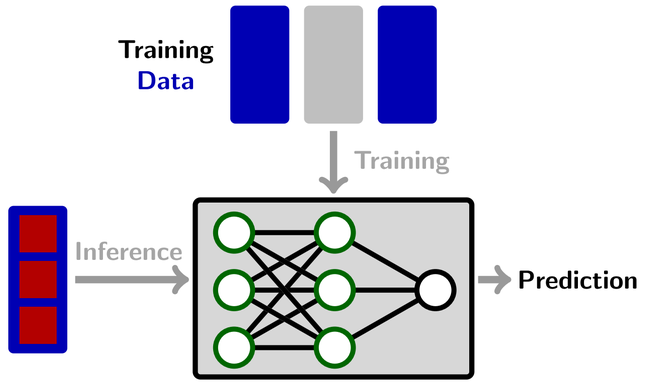
Vision



Concept

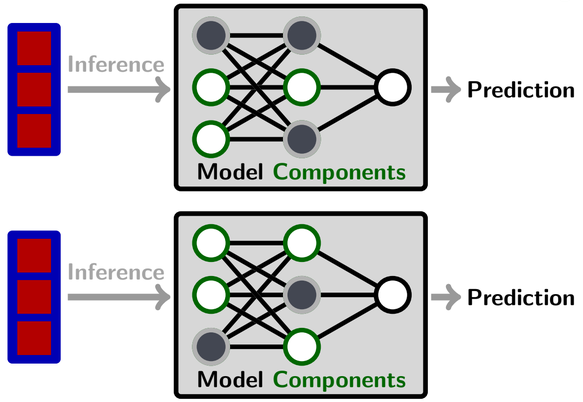
Key Challenge:
In-distribution perturbations

Data



Key Challenge:
Retraining Cost

Model Component



Key Challenge:
In-distribution perturbations

Value Function

Let $S \subseteq [n]$ be the set of features / data points / components unperturbed

We use $f : 2^{[n]} \to \mathbb{R}$ to measure the effect on model behavior

Examples

Classification

Logit of predicted class

LLM

Perplexity w.r.t. original output

Accuracy

Performance on validation set

From Value Functions to Selection

$S \subseteq [n]$ **is the set of features / data points / components unperturbed**

Value function $f : 2^{[n]} \to \mathbb{R}$

$$\max_{\ell \leq |S| \leq r} f(S)$$

Examples

Prompt compression

Data selection

Model pruning

Submodular set functions (Nemhauser et al., 1978)
Greedy solutions (Nemhauser et al., 1978)
Ellipsoid Method (Grötschel et al., 1981)

From Value Functions to Attribution

$S \subseteq [n]$ **is the set of features / data points / components unperturbed**

Value function $f : 2^{[n]} \to \mathbb{R}$

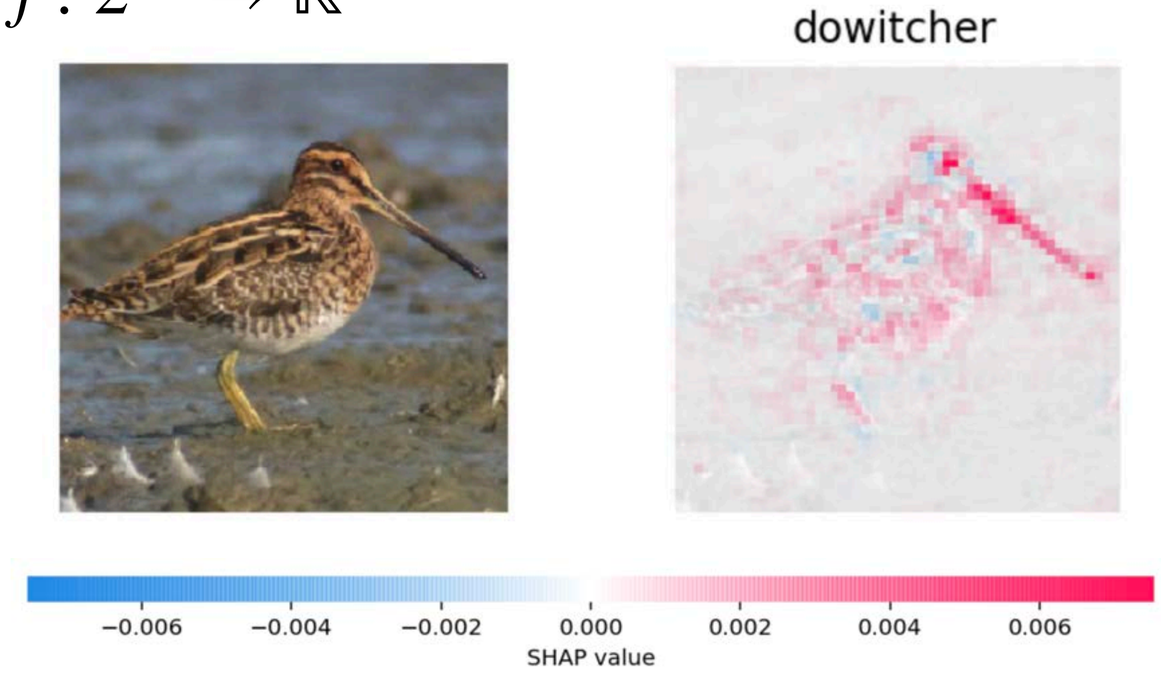
How much value does $i \in [n]$ ***contribute?***

$$\Delta_i(S) = f(S \cup \{i\}) - f(S)$$

$$Attr_i = \sum_{S \subseteq [n] \setminus \{i\}} w(S)\Delta_i(S)$$

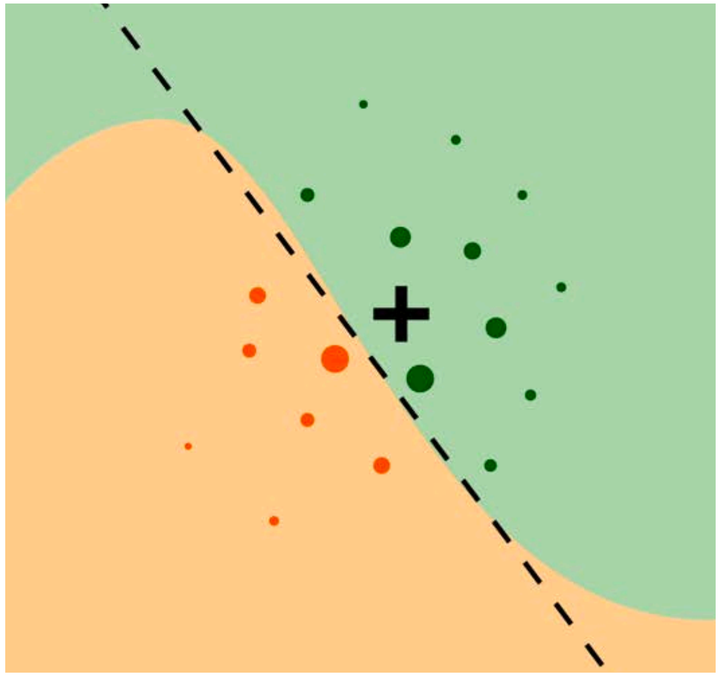
Shapley value uses $w(S) = \frac{1}{n}\binom{n-1}{|S|}^{-1}$

Shapley Value (Shapley, 1953)
Banzhaf Power Index (Banzhaf, 1965)
SHAP (Lundberg & Lee, 2017)



Similar definitions extend to interactions $\{i, j\}$

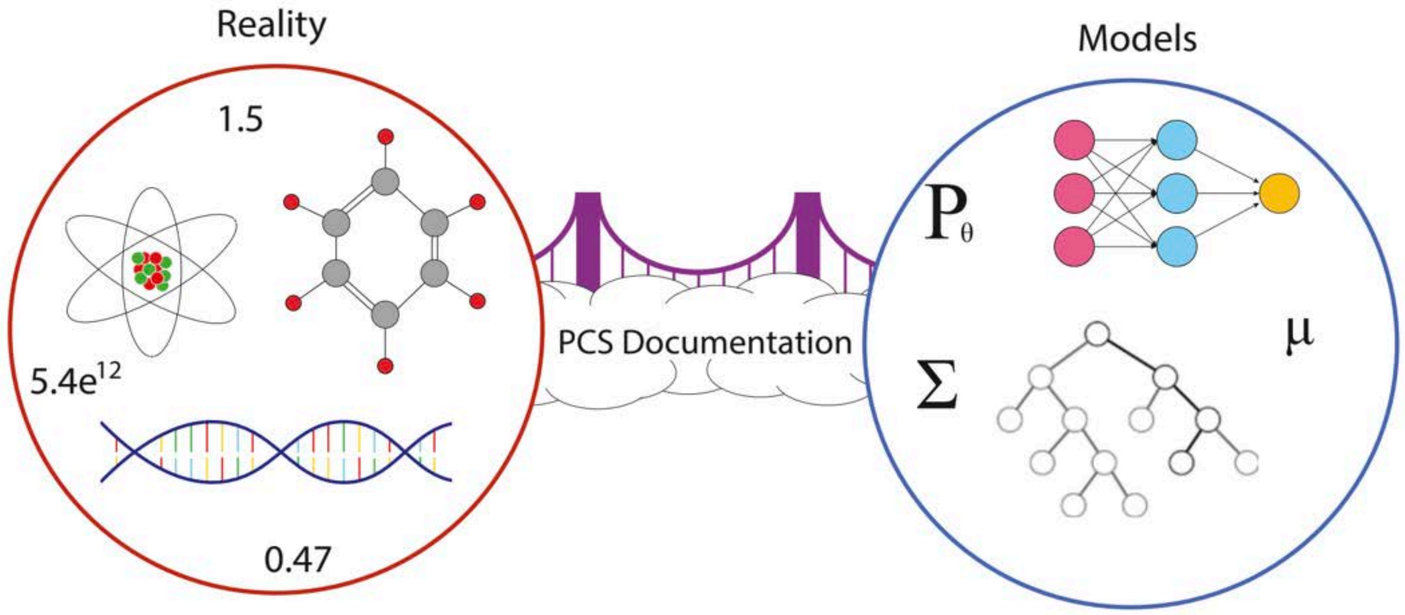
Limitations: Locality and Reality-Checking



Framework for Local Analysis

(Reference input, dataset,
or model configuration)

Veridical Data Science
(Yu & Kumbier, 2019)



External Reality-Checking Relies on Human Validation

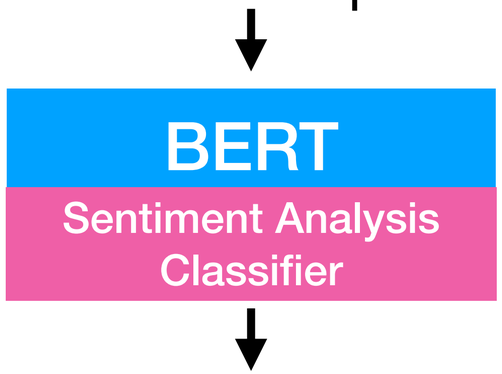
(Discovering new physical, chemical,
or biological laws)

Algorithms

Sentiment Analysis Example

Review:

A visual masterpiece with a great lead, yet undone by a hollow script.



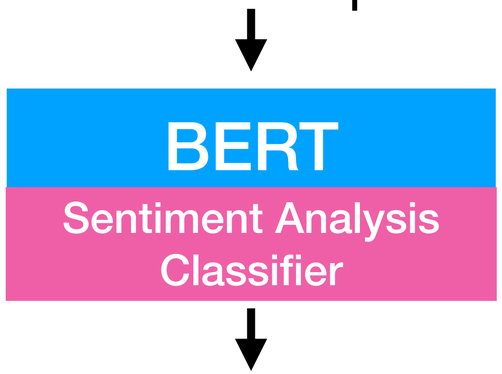
Sentiment Score

-0.95

Sentiment Analysis Example

Review:

A visual masterpiece with a great lead, yet undone by a hollow script.

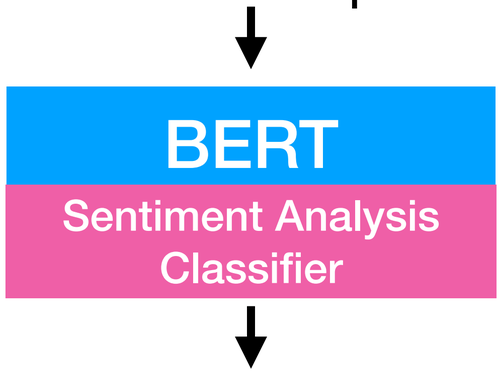


Sentiment Score

-0.95

Review:

A visual masterpiece with a great lead, yet undone by a hollow script.

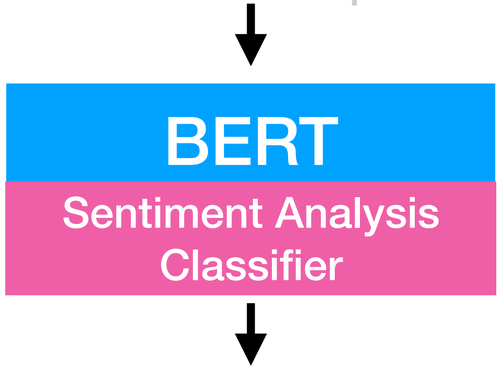


Sentiment Score

-0.03

Review:

A visual masterpiece with a great lead, yet undone by a hollow script.



Sentiment Score

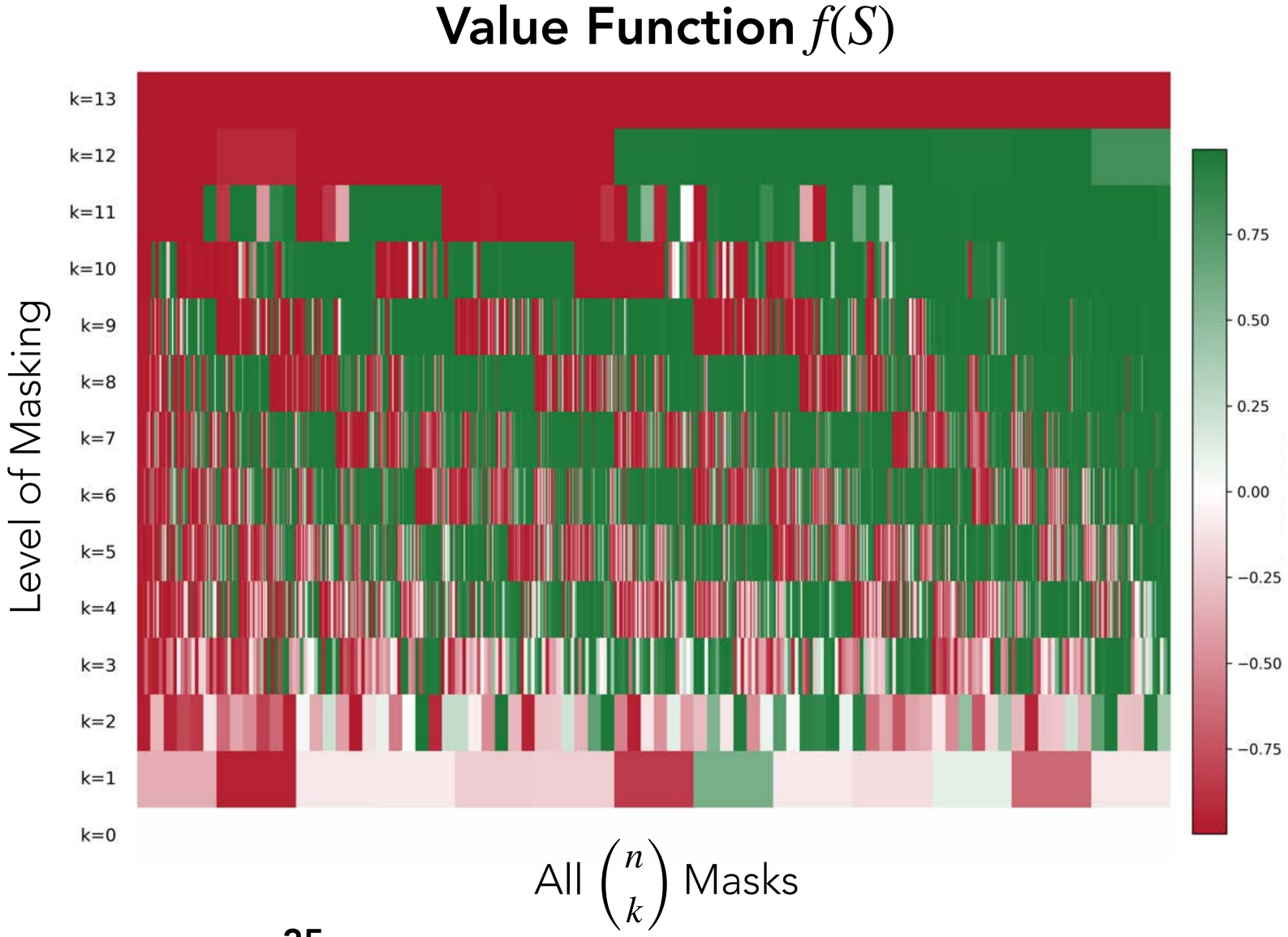
+0.73

Sentiment Analysis Example

A visual masterpiece with a great lead, yet undone by a hollow script.

A visual masterpiece with a great lead, yet undone by a hollow script.

A visual masterpiece with a great lead, yet undone by a hollow script.



Fourier Decomposition

Every value function $f(S)$ admits a unique decomposition onto a orthonormal parity (XOR) function basis:

$$f(S) = \sum_{T \subseteq [n]} F(T) \prod_{i \in T} x_i$$

The film embraces bold visuals and the plot rarely drags. However, glaring plot holes and forced jokes undercut its impact. ...

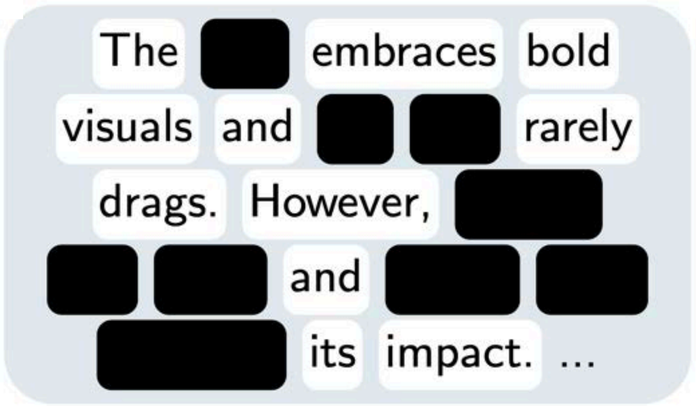
$$f : 2^{[n]} \to \mathbb{R}$$

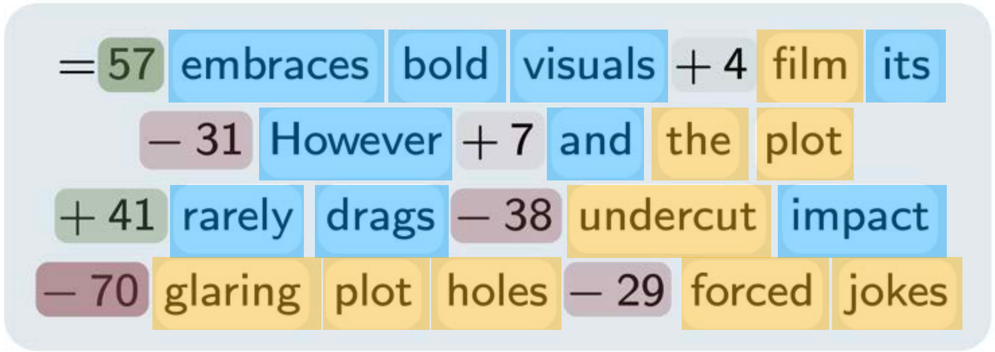
$=57$ embraces bold visuals $+4$ film its -31 However $+7$ and the plot $+41$ rarely drags -38 undercut impact -70 glaring plot holes -29 forced jokes

Fourier Decomposition

Every value function $f(S)$ admits a unique decomposition onto a orthonormal parity (XOR) function basis:

$$f(S) = \sum_{T \subseteq [n]} F(T) \prod_{i \in T} x_i$$



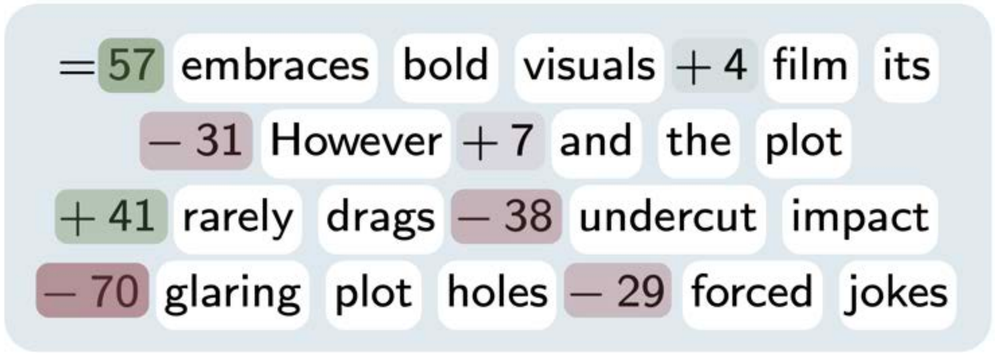


-1 +1

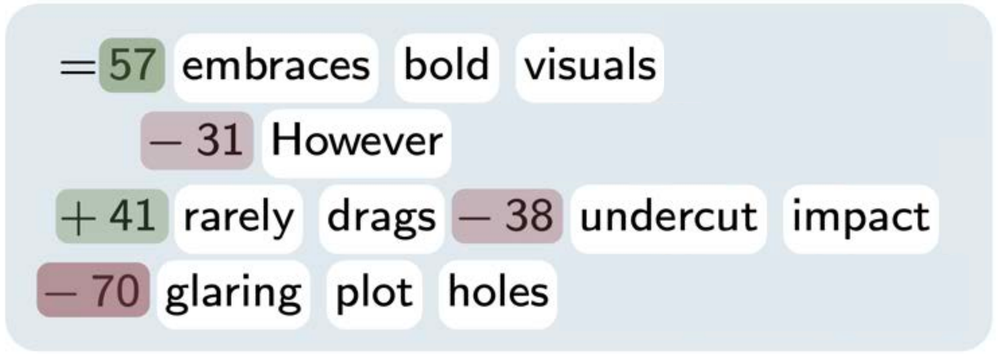
Fourier Decomposition

Possesses many key properties for learning, including Parseval's Theorem:

$$\mathbb{E}_S\left[\left(f(S)-\hat{f}(S)\right)^2\right]=\sum_{T\subseteq[n]}\left(F(T)-\hat{F}(T)\right)^2$$



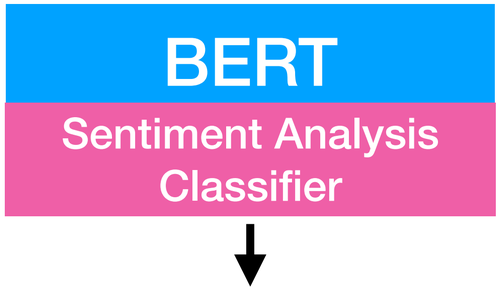
$$f: 2^{[n]} \to \mathbb{R}$$



$$\hat{f}: 2^{[n]} \to \mathbb{R}$$

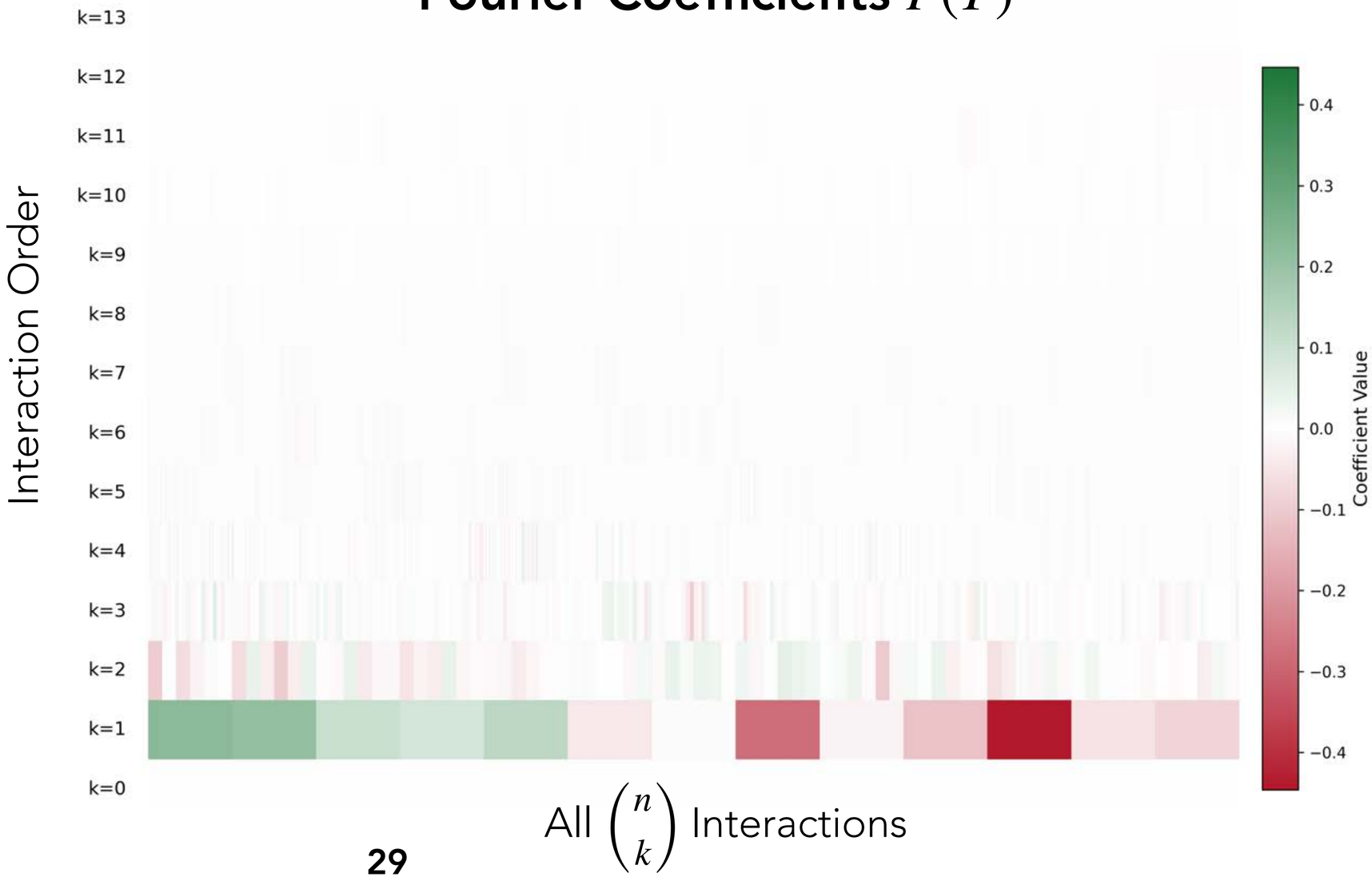
Structural Observations of Fourier Transform

Review:
A visual masterpiece with a great lead, yet undone by a hollow script.



Sentiment Score

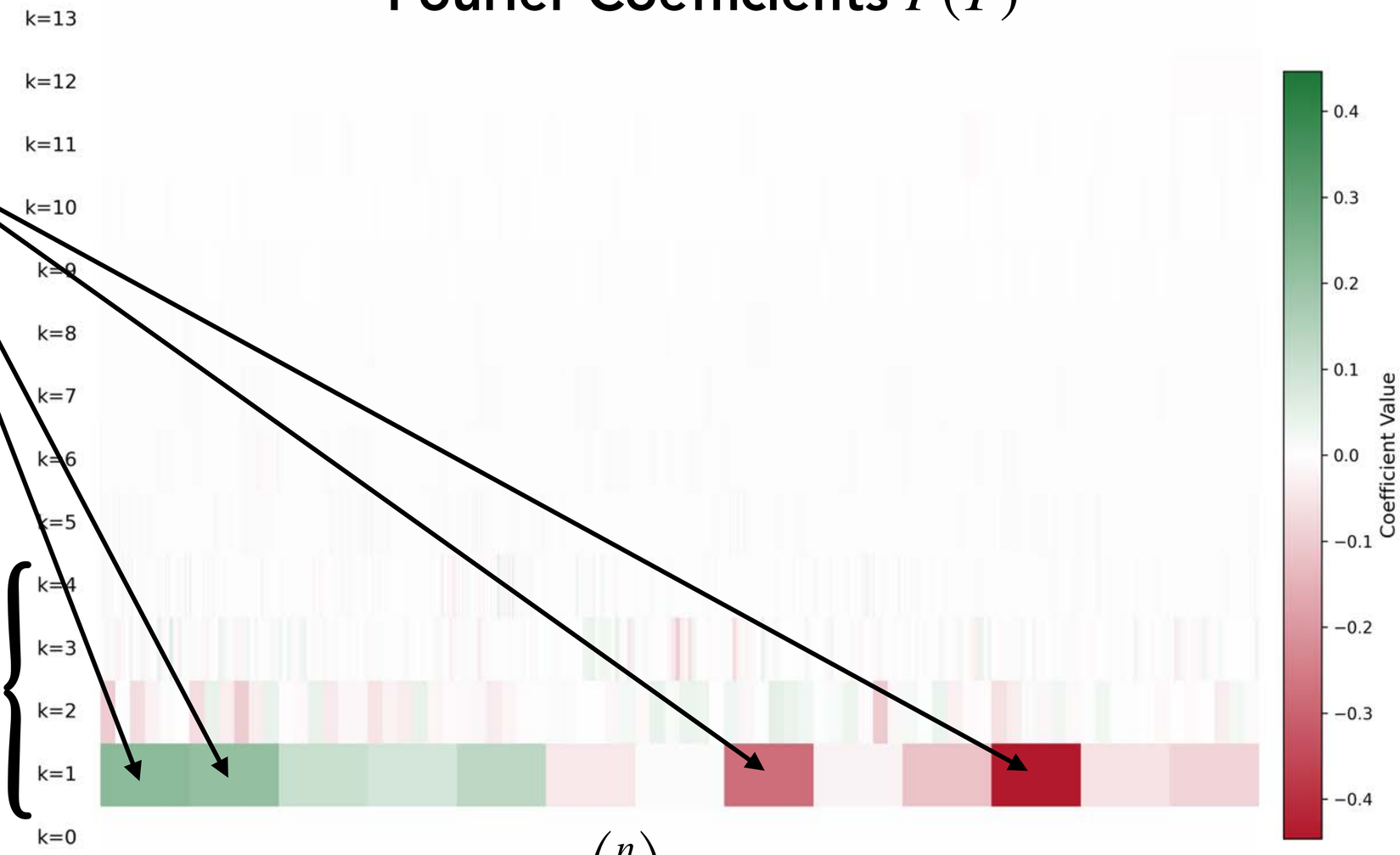
Fourier Coefficients $F(T)$



Structural Observations of Fourier Transform

Fourier Coefficients $F(T)$

A small number of Fourier coefficients dominate (Sparsity)

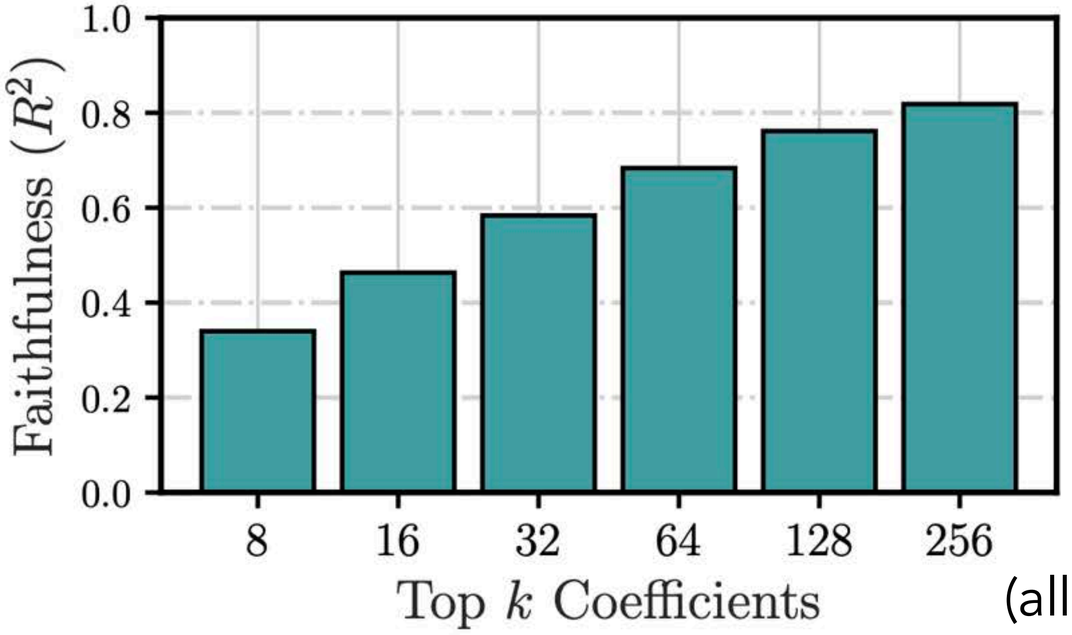


Influential Fourier coefficients tend to be low-degree

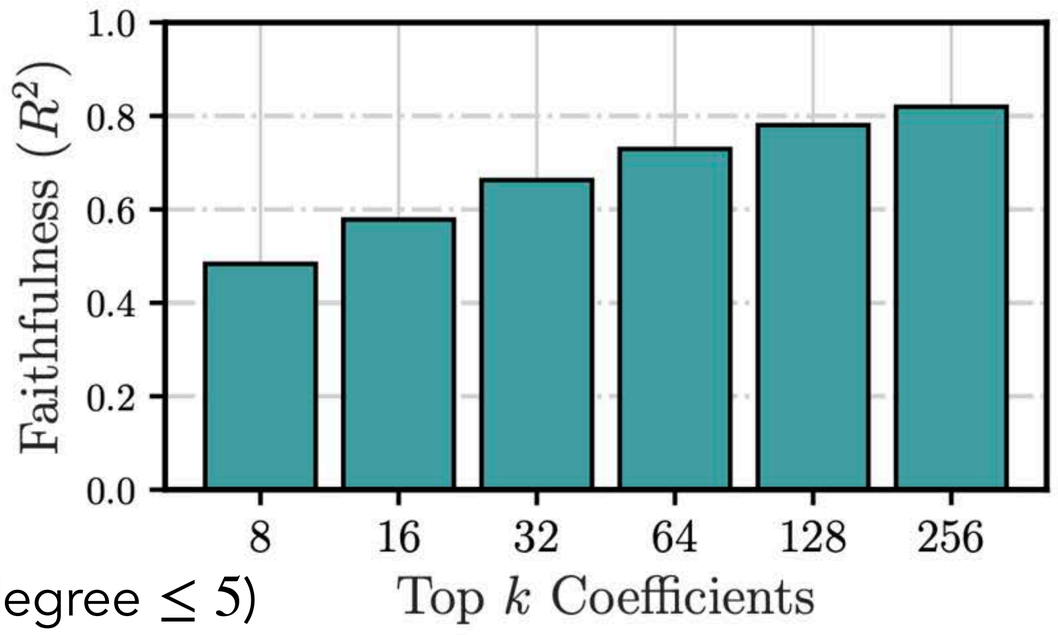
All $\binom{n}{k}$ Interactions

Empirical Sparsity / Low-Degree

Long Sentiment Analysis (300 features)



CLIP Image-Caption (80 features)

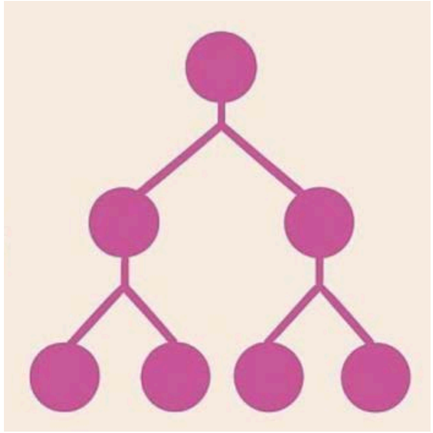


(all with degree ≤ 5)

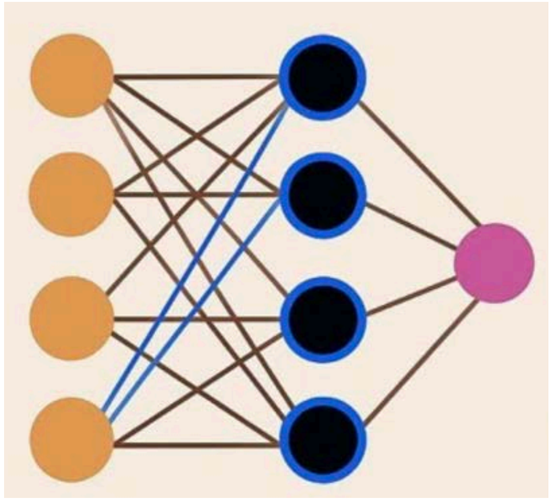
Similar results with value functions for

CIFAR-10 data attribution / Llama-3 Attention Head Pruning

Theoretical Sparsity / Low-Degree



Tree Ensembles with T trees of depth d are d-degree and $O(T4^d)$ sparse



"Spectral Bias"

For (infinite-width, randomly initialized) neural networks, ***sparse*** functions are easier to learn than ***dense*** functions.

Deep learning generalizes because the parameter-function map is biased towards simple functions (Valle Perez et al., 2019)

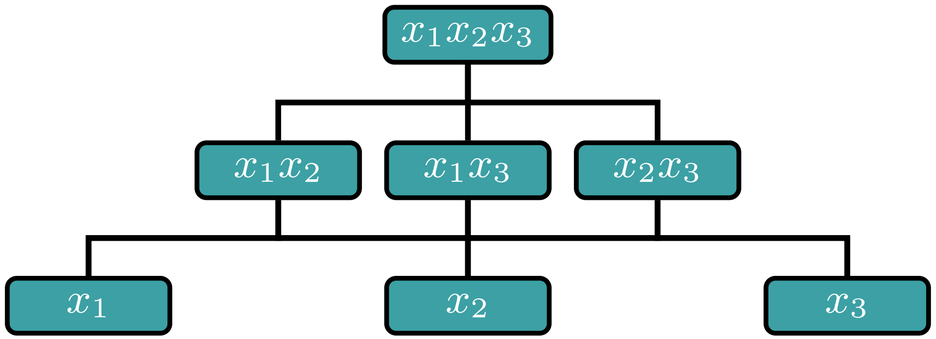
A Fine-Grained Spectral Perspective on Neural Networks (Yang & Salman, 2019)

Failure of Sparsity

Degree d AND term

$x_1 x_2 x_3$

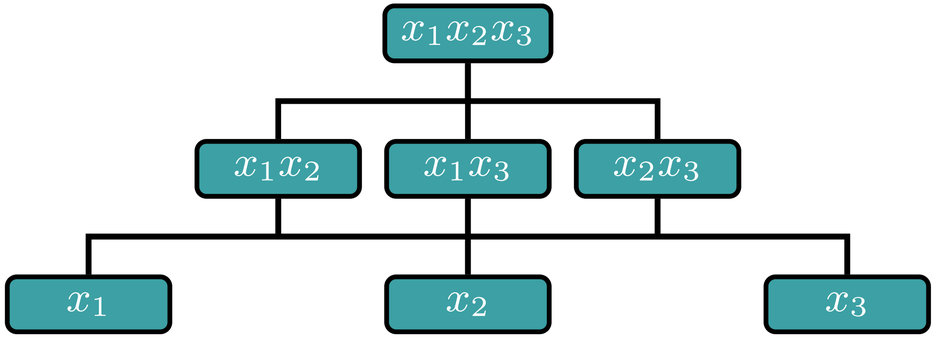
XOR w/ 2^d terms



Degree d OR term

$x_1 x_2 x_3$

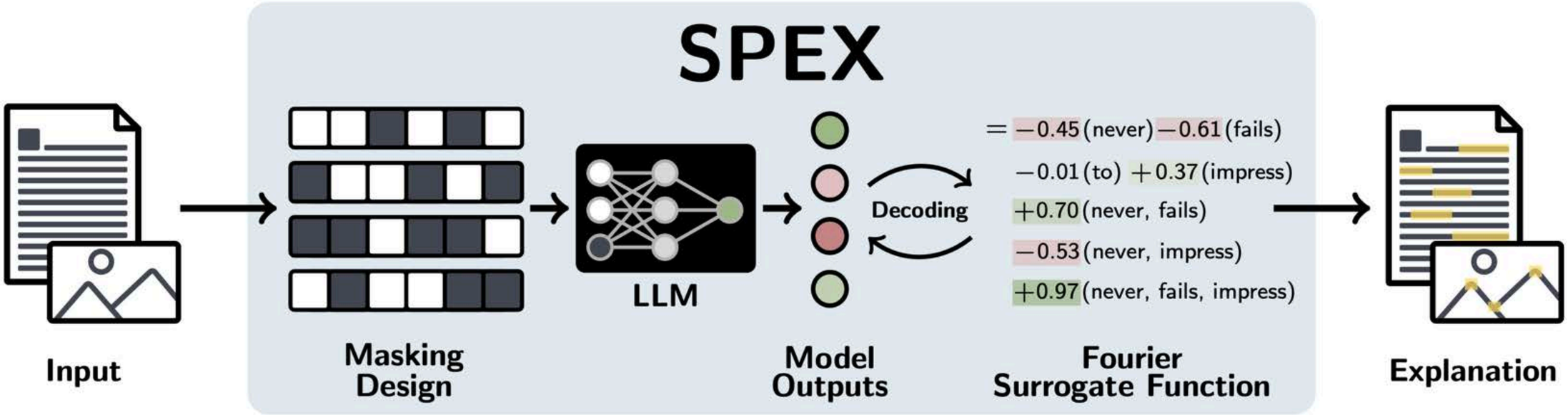
XOR w/ 2^d terms



If value function leads to high-degree ANDs / ORs, will be dense under XOR

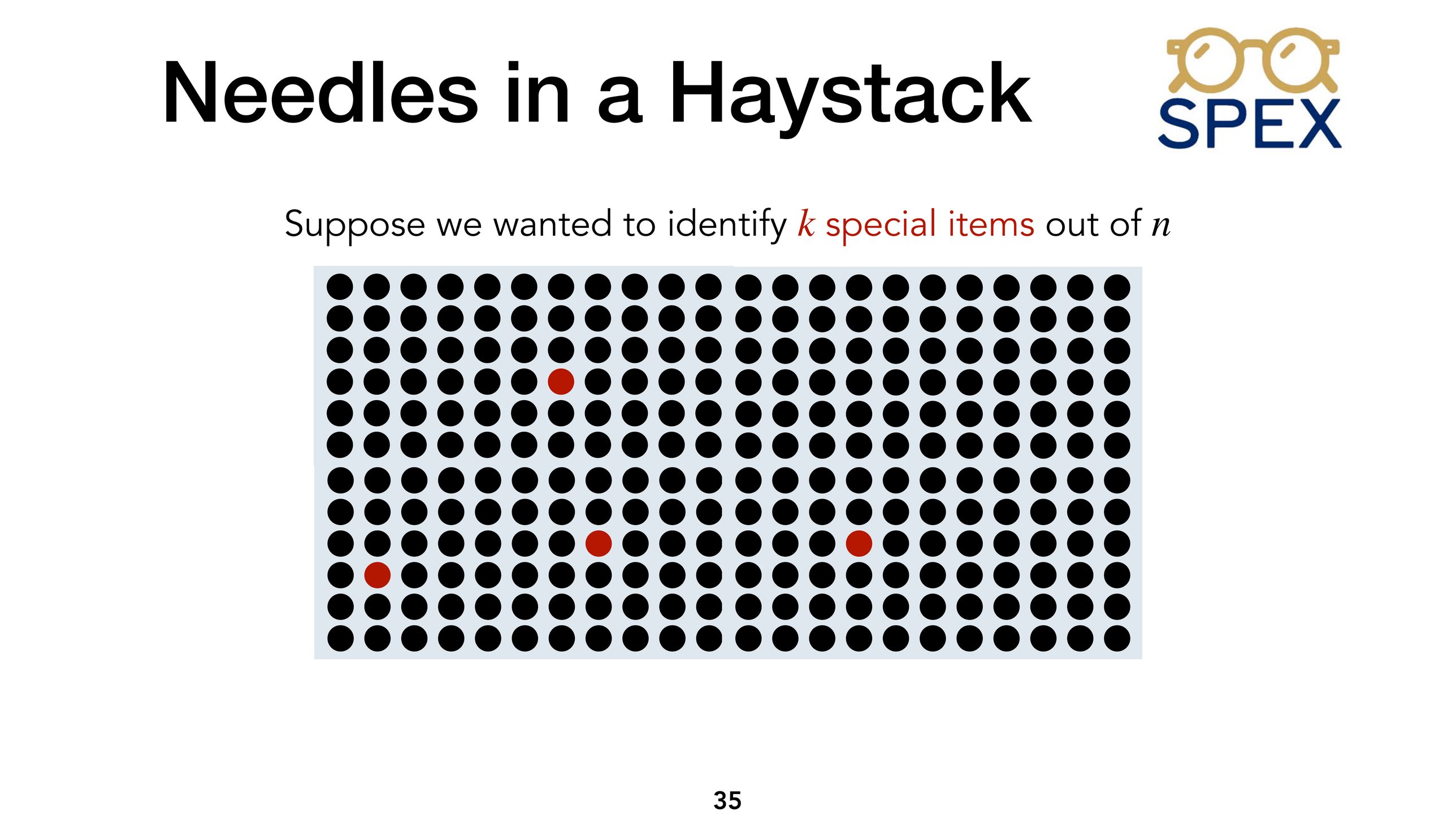
SPEX (Spectral Explainer)

Focus on (noisy) sparse support recovery



Sparse Boolean Polynomials: (Kocaoglu et al., 2014), (Negahban & Shah, 2012)
Sparse Fourier Transforms: (Stobbe & Krause, 2012), (Hassanieh et al., 2012), (Li et al., 2014)

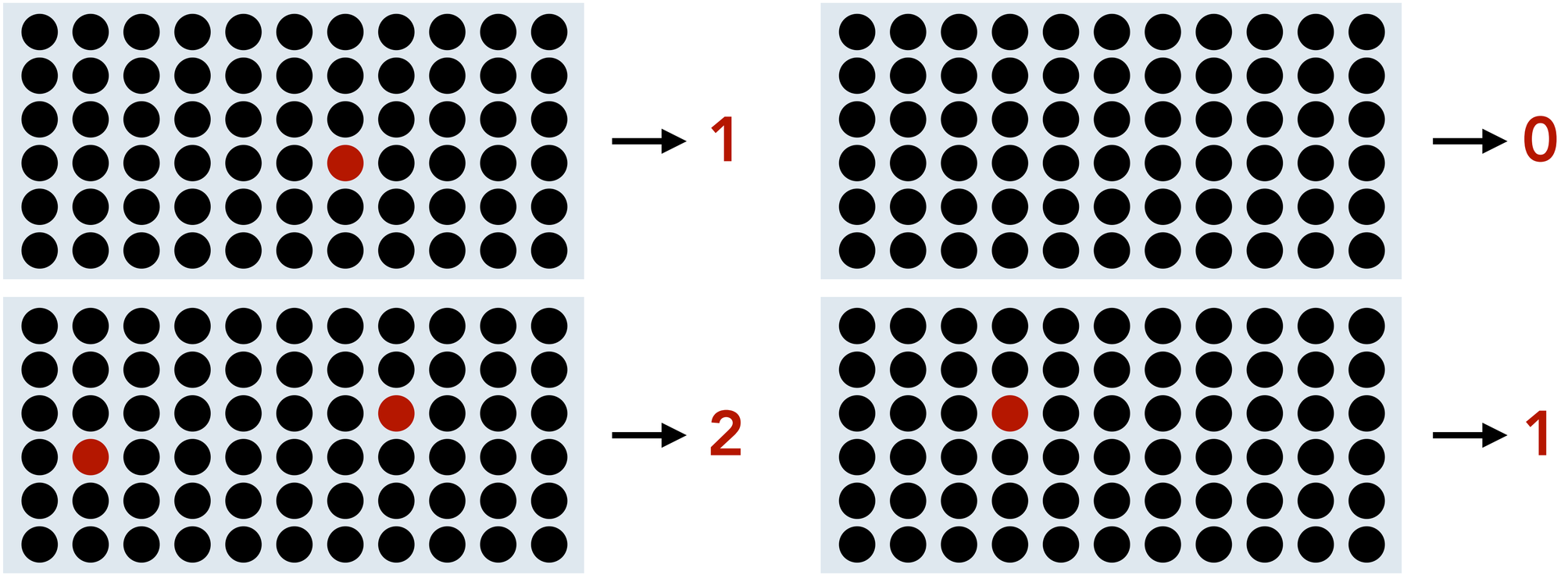
Needles in a Haystack



Suppose we wanted to identify k special items out of n

Pooled Measurements

If we are able to identify counts from pooled measurements

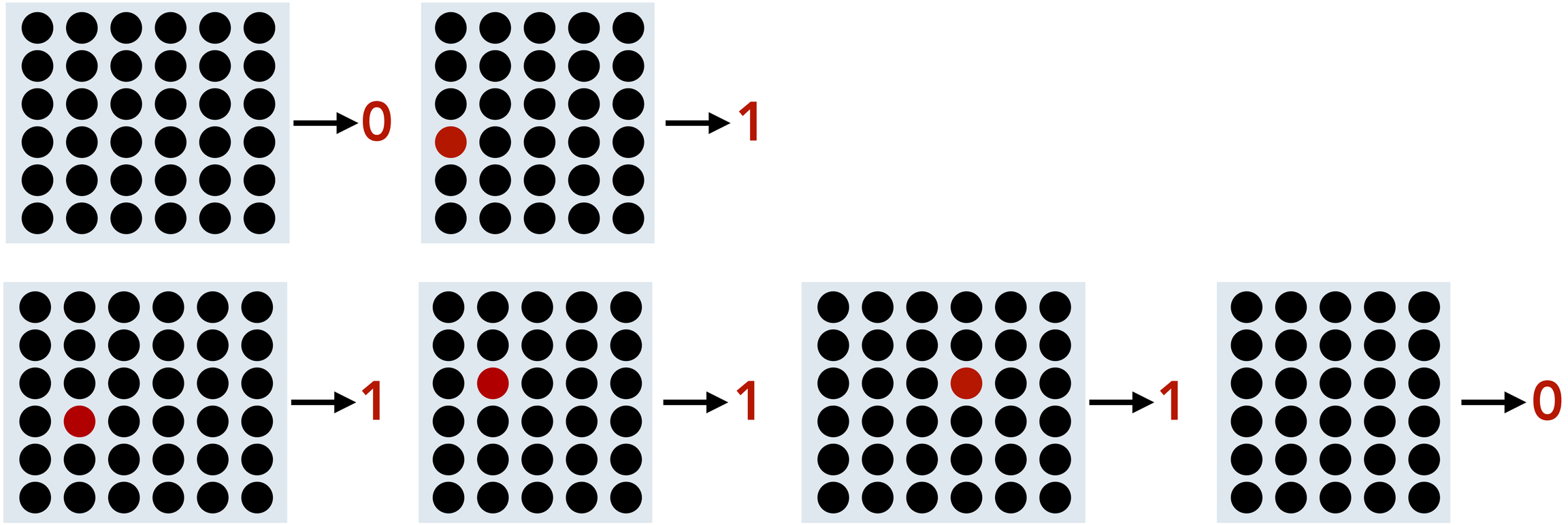


Compressed Sensing: (Candes et al., 2006), (Donoho, 2006)

Group Testing: (Dorfman, 1943), (Sobel & Groll, 1959)

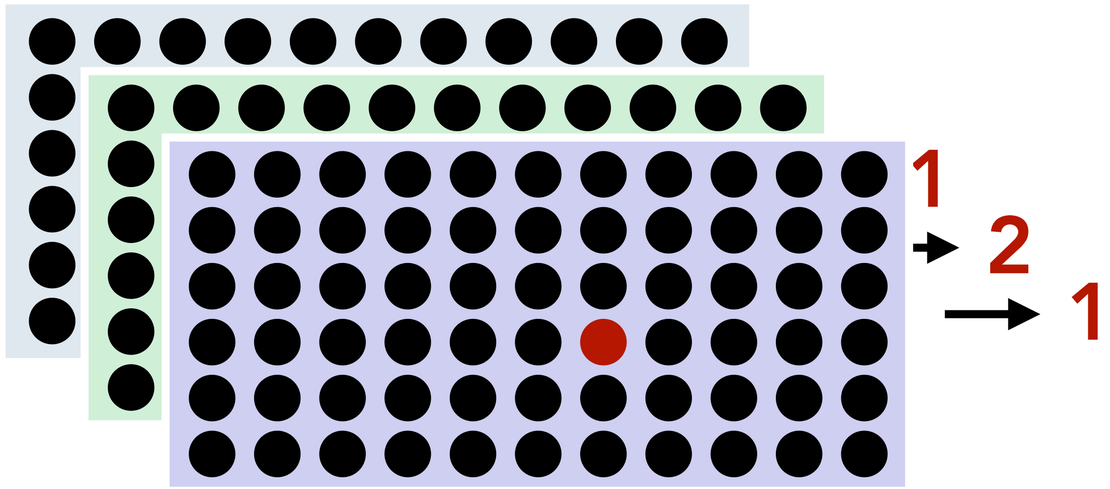
Sketching: (Alon, 1996), (Muthukrishnan, 2005)

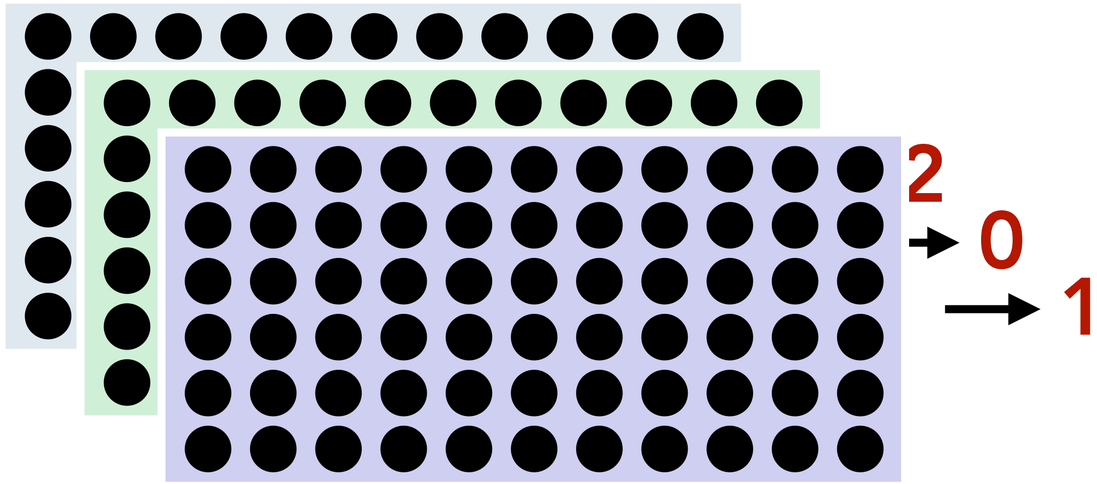
Adaptive Measurements

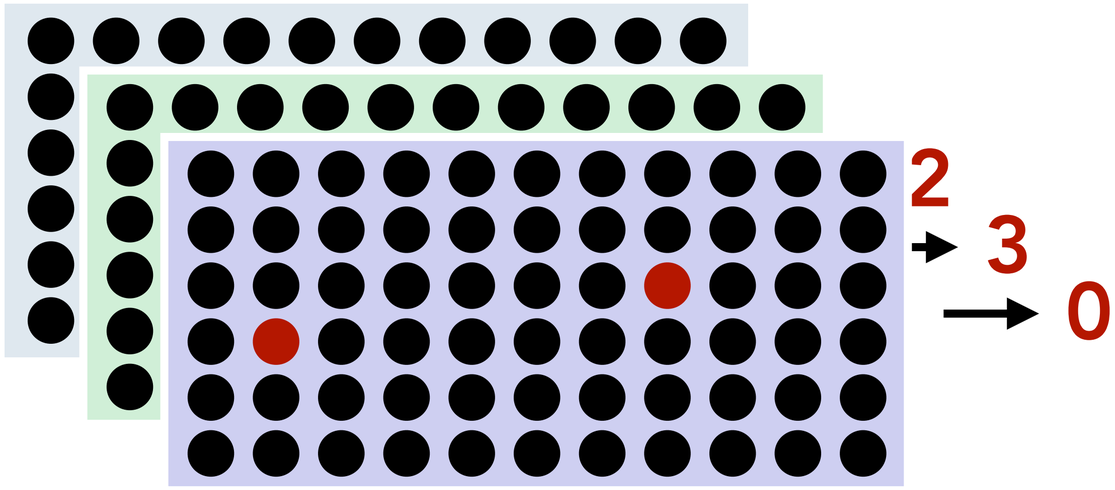


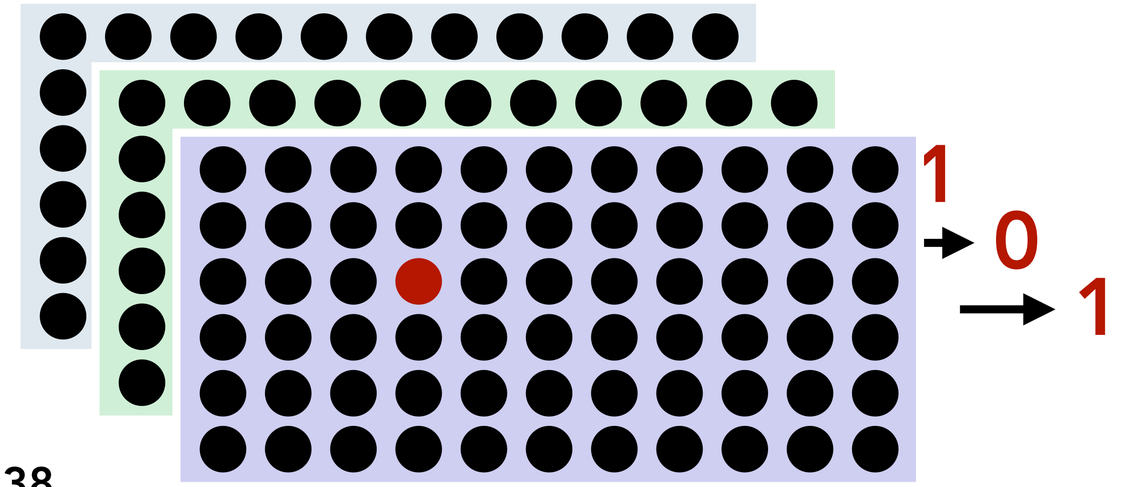
Non-Adaptive Measurements

Must select all groups beforehand (clever overlaps)





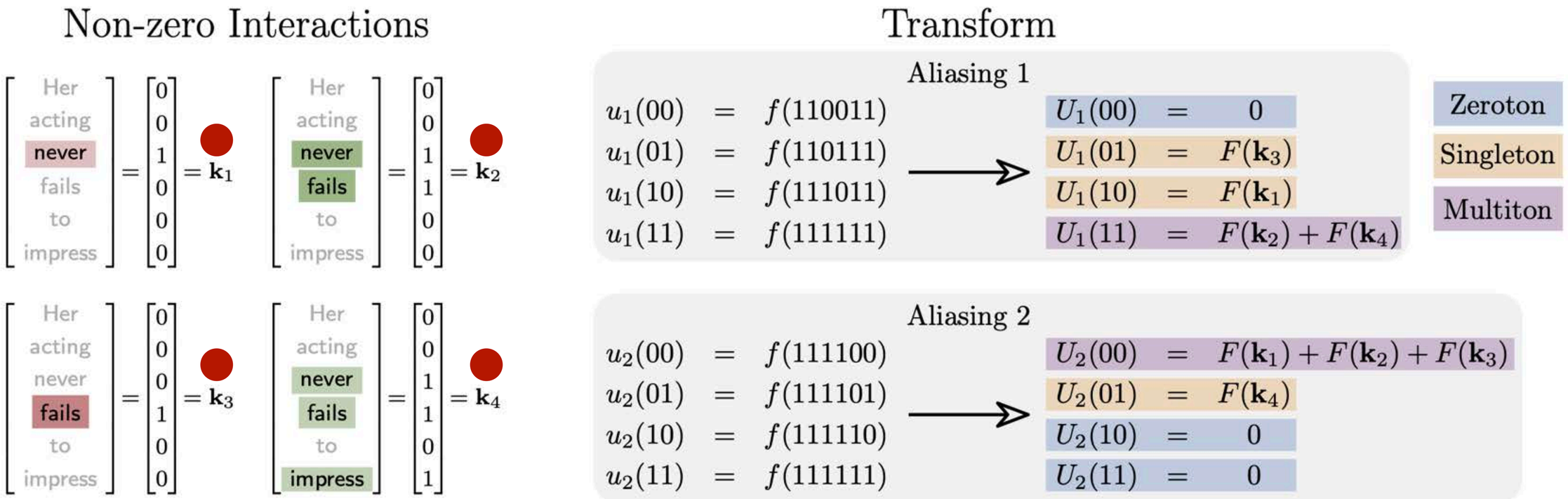




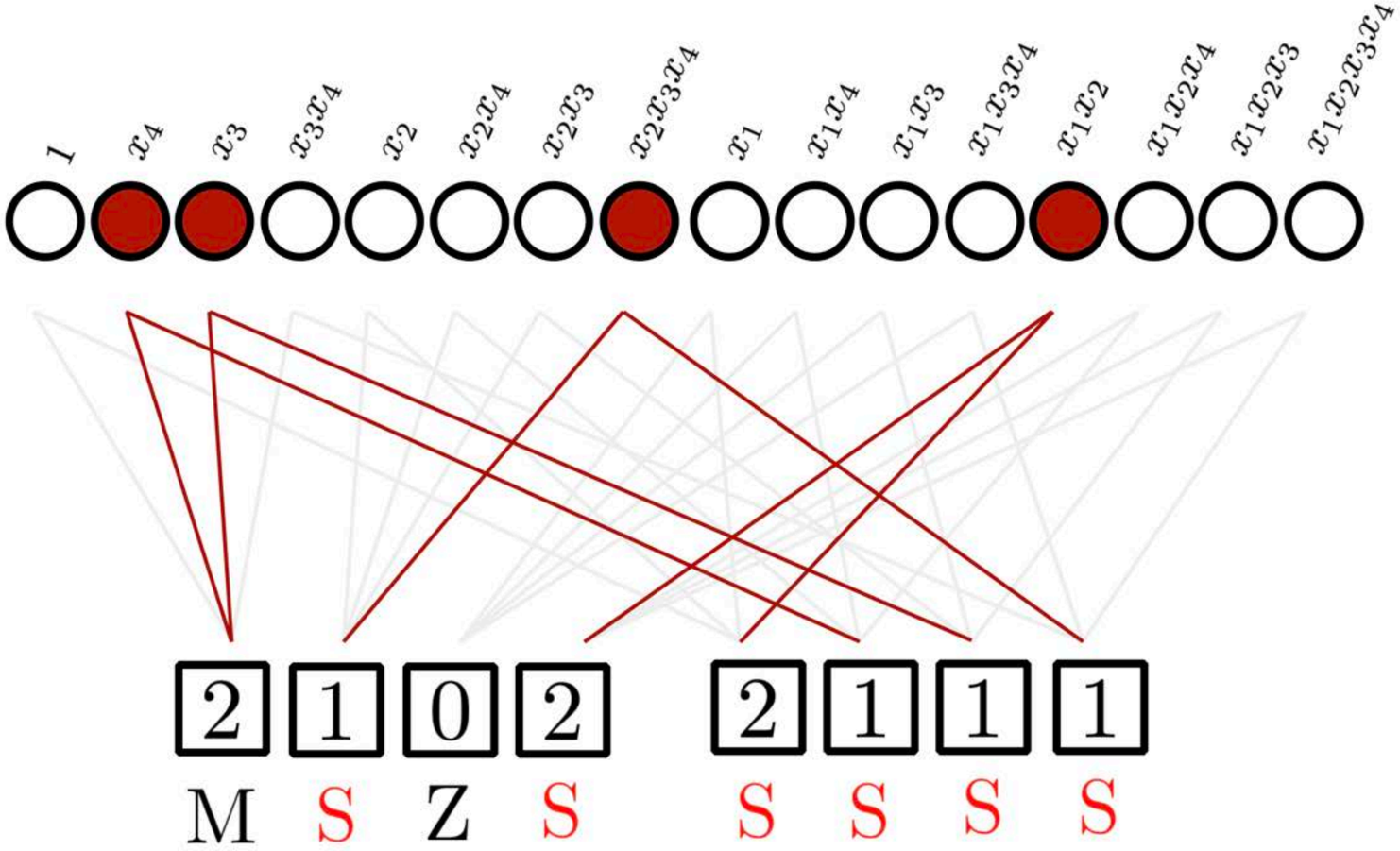
Aliasing

By selectively subsampling f, we pool many Fourier coefficients together.

Subsampled function $u : 2^{[b]} \to \mathbb{R}$ has FCs which are sums of original FCs.



Shifting



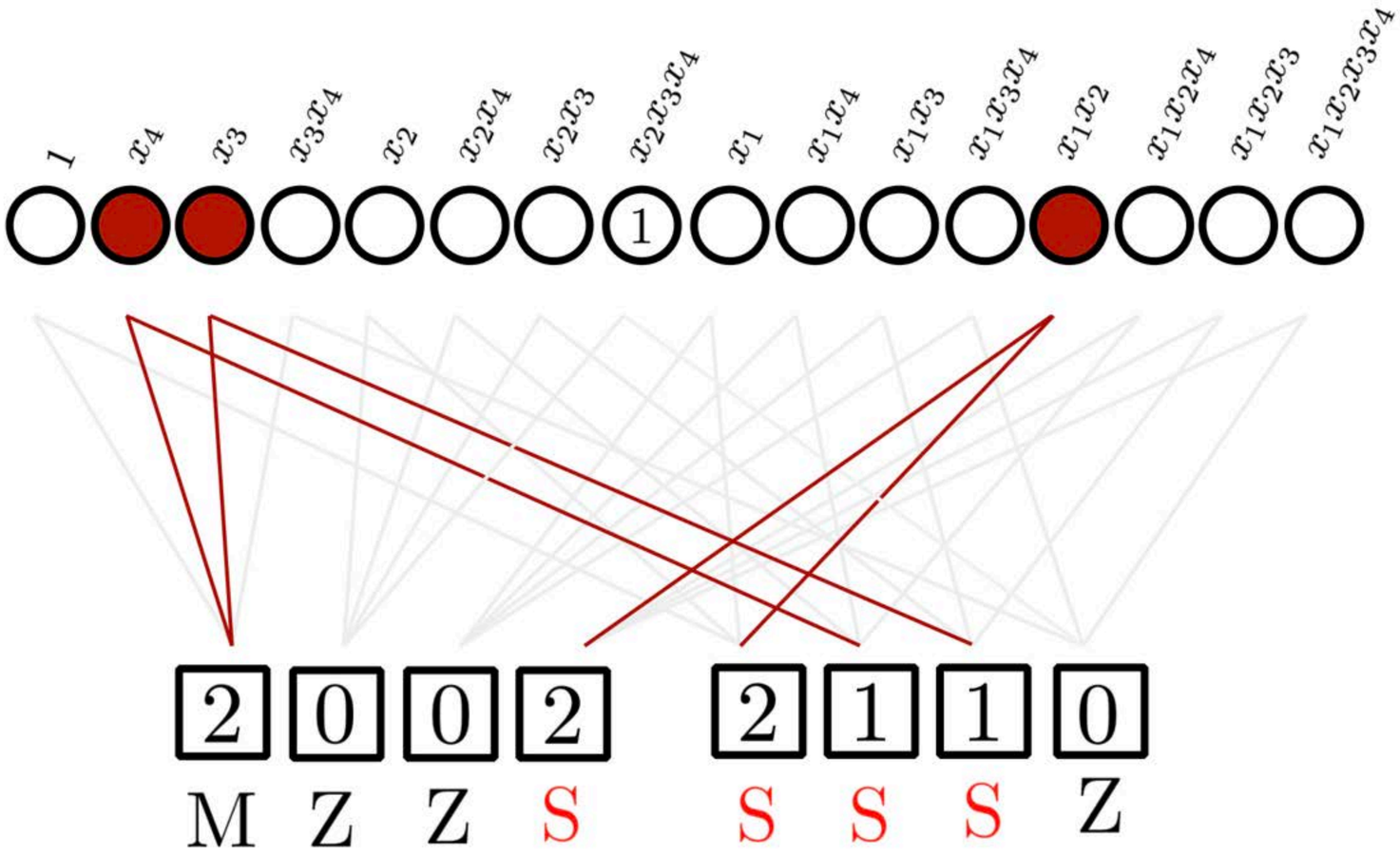
We have summed Fourier coefficients, not counts.

We can apply random shifts (negations) to Fourier coefficients to identify pooled counts.*

For singleton pools, this allows us to identify the corresponding coefficient.

*BCH Channel Codes tells us how to design non-adaptive shifts

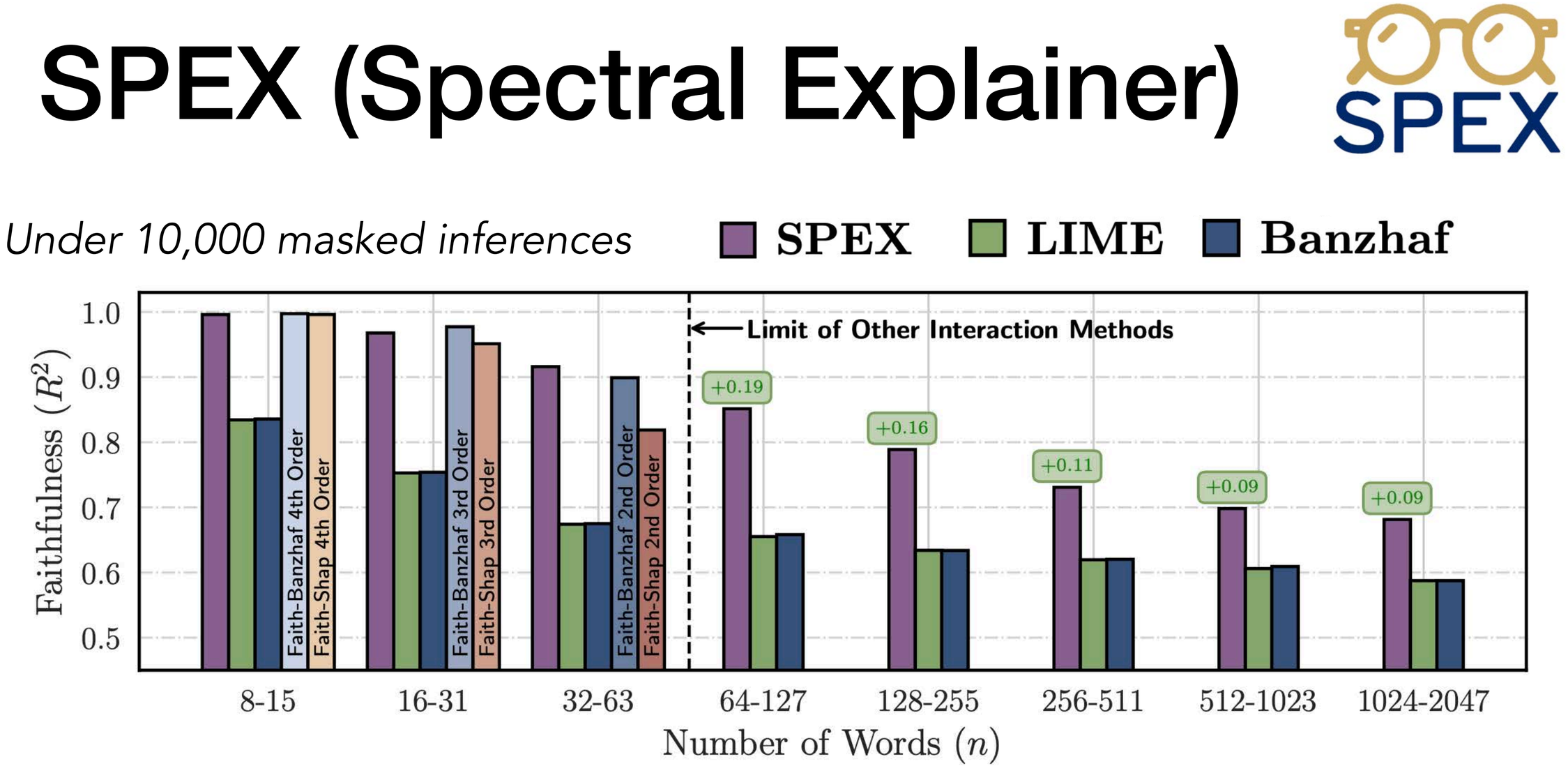
Decoding



Singletons can be identified and *peeled* from other pools

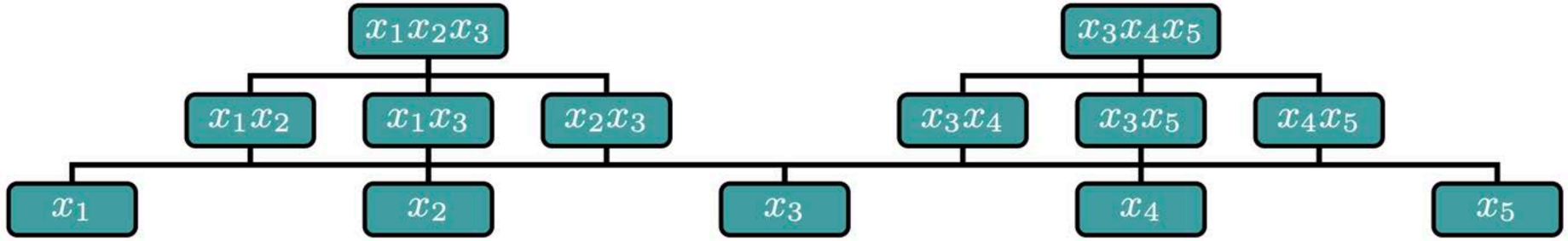
Iteratively decode until all pools are empty

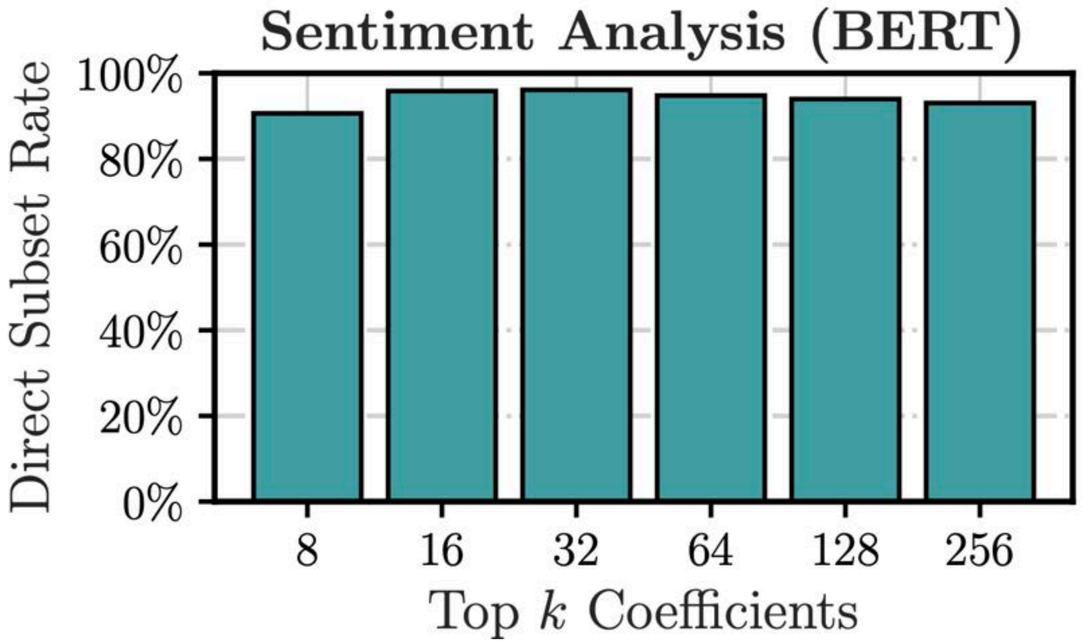
SPEX (Spectral Explainer)



"SPEX: Scaling Feature Interaction Explanations for LLMs",
KBAEPRY, ICML 2025

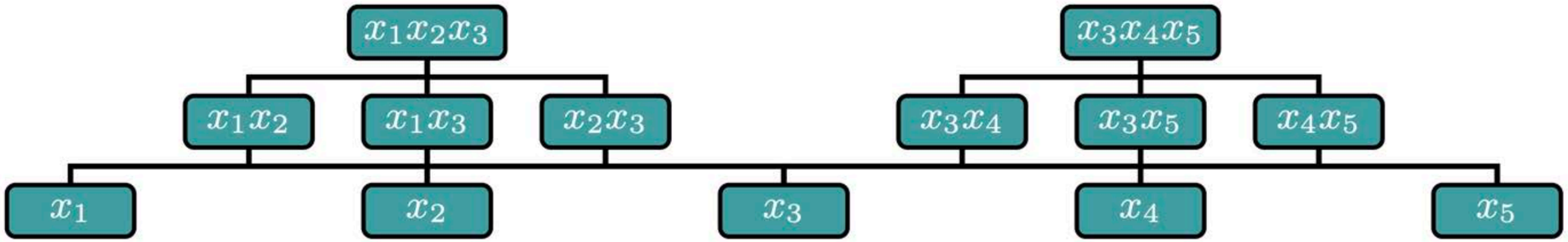
Hierarchical Structures

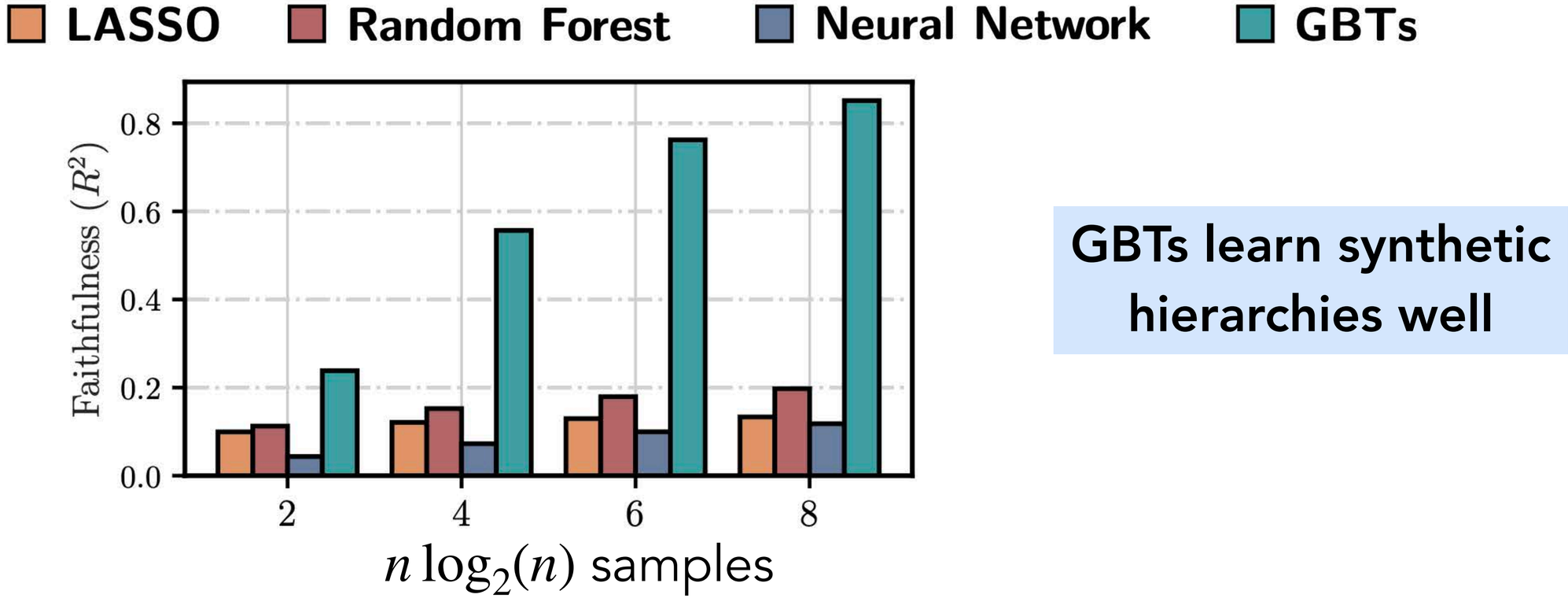




Among the top k Fourier coefficients,
>90% of their subsets are also in top k

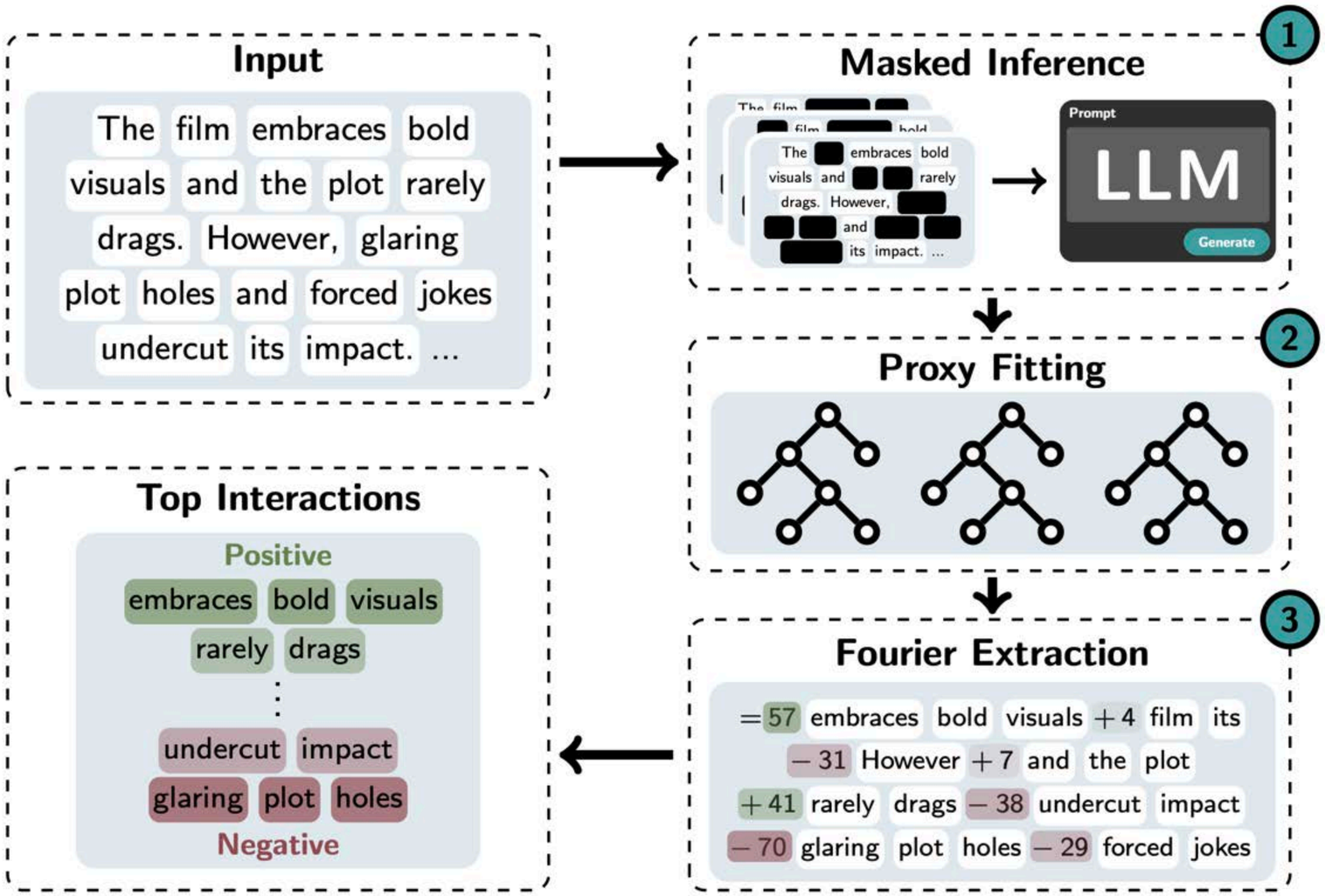
Trees and Hierarchies





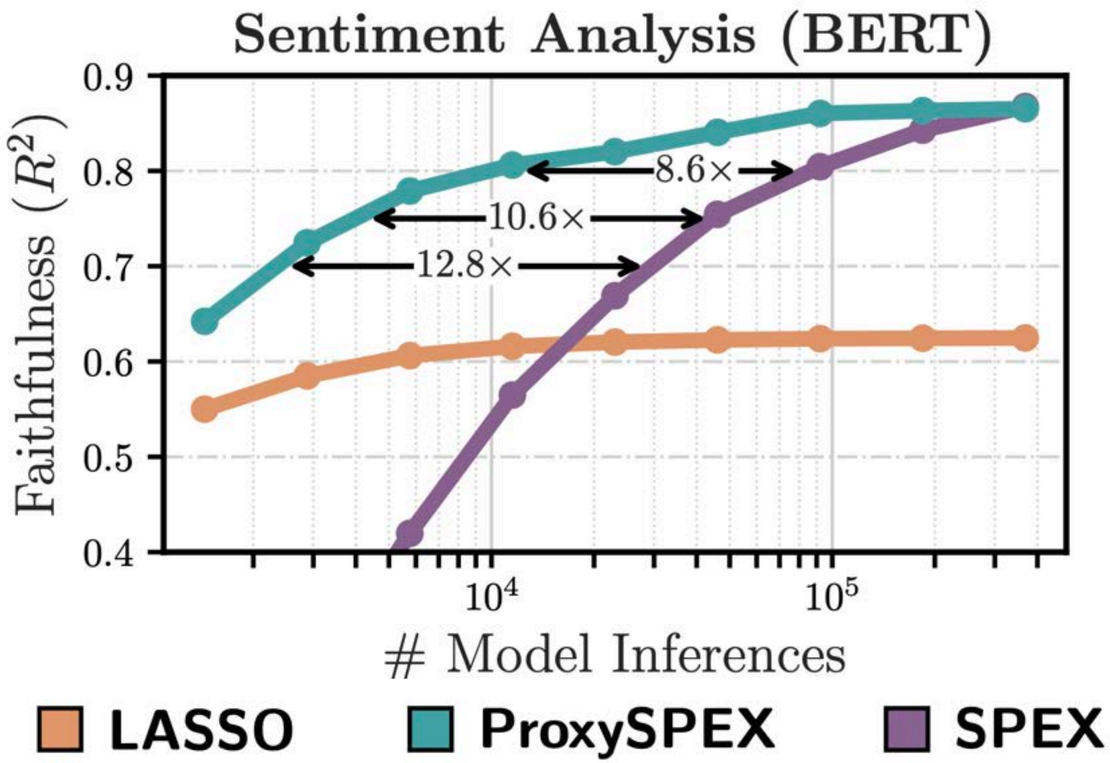
GBTs learn synthetic hierarchies well

ProxySPEX



We use GBT as **proxy** model, then extract Fourier support

ProxySPEX



~10x Inference Reduction from SPEX

"ProxySPEX: Inference-Efficient Interpretability via Sparse Feature Interactions in LLMs"

BKAEYR, NeurIPS 2025 (Spotlight)

Sparse Fourier for Selection

$$\max_{\mathbf{x}} \sum_{T \in \mathcal{K}} F(T) \prod_{i \in T} x_i$$

⇕ **Change of Variable**

Binary Integer Linear Program
(with clever linear constraints)

Solved through Branch-and-Cut algorithms provided by standard optimization libraries

Sparse Fourier for Attribution

Marginal

Banzhaf ψ_i	$-2F(\{i\})$
Shapley ϕ_i	$(-2)\sum_{\substack{S\supseteq\{i\} \\ \lvert S\rvert \text{ is odd}}} \frac{F(S)}{\lvert S\rvert}$
Influence ξ_i	$\sum_{S\ni i} F(S)^2$

Interaction

Möbius $I^{\mathrm{M}}(T)$	$(-2)^{\lvert T\rvert}\sum_{S\supseteq T} F(S)$
Or $I^{\mathrm{O}}(T)$	$\begin{cases}\sum_{S\subseteq[n]} F(S) & \text{if } T=\emptyset \\ -(-2)^{\lvert T\rvert}\sum_{S\supseteq T}(-1)^{\lvert S\rvert}F(S) & \text{if } T\neq\emptyset\end{cases}$
Banzhaf Interaction $I^{\mathrm{B}}(T)$	$-2F(T)$
Shapley Interaction $I^{\mathrm{S}}(T)$	$(-2)^{\lvert T\rvert}\sum_{S\supseteq T \text{ s.t. } (-1)^{\lvert S\rvert}=(-1)^{\lvert T\rvert}} \frac{F(S)}{\lvert S\rvert-\lvert T\rvert+1}$

Polynomial under **And / Or** basis

Shapley Feature Estimation

SHAP

50K citations

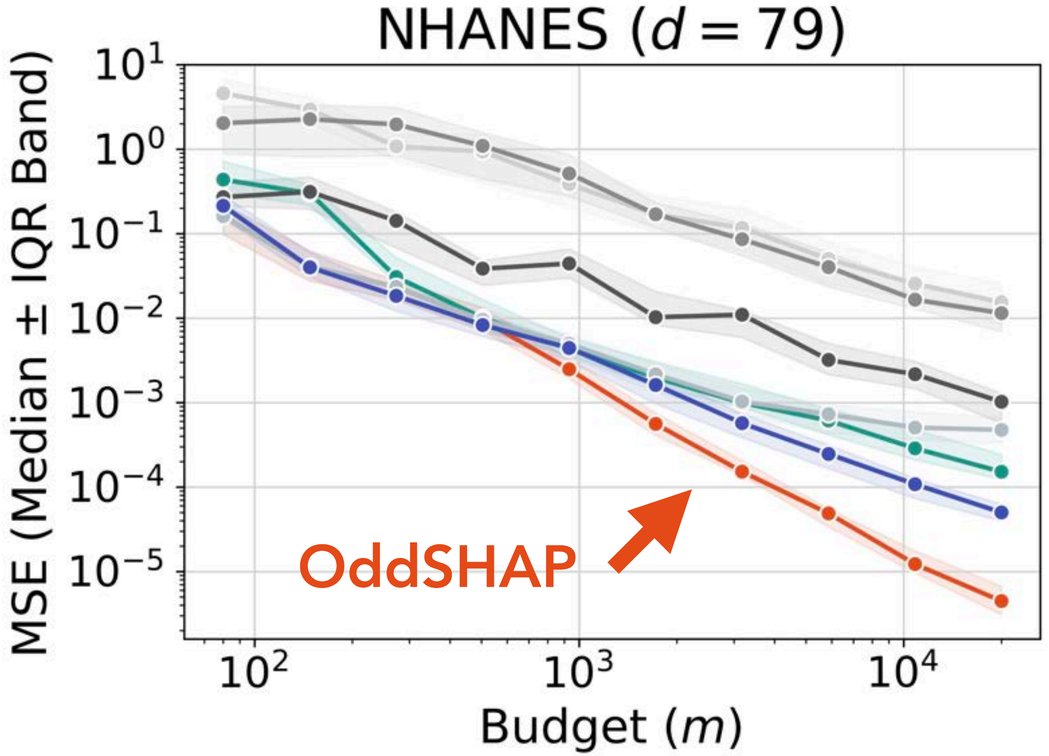
$$\text{Shapley } \phi_i \quad (-2) \sum_{\substack{S \supseteq \{i\} \\ |S| \text{ is odd}}} \frac{F(S)}{|S|}$$

Observation 3.1 (Shapley Values of Odd and Even Functions). *Let $f : 2^{[d]} \to \mathbb{R}$ be a function. Then, for all $i \in [d]$,*

$$\phi_i(f) = \phi_i(f_{\text{odd}}). \tag{7}$$

"An Odd Estimator for Shapley Values"

FBKRW, Submitted to ICML 2026



Applications

Feature Removal

Multi-Hop Question Answering

Question:
The magazine that nominated George Rainsford for their Best Actor award in 2017 comes out every week on what day of the week?

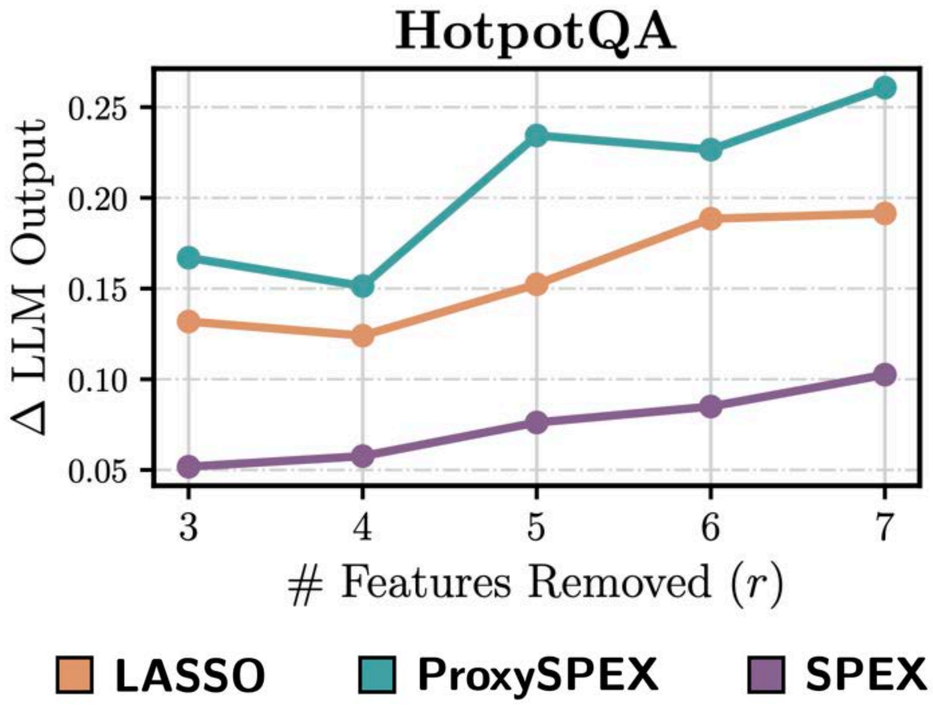
Title: George Rainsford (actor)
George Rainsford is an English actor and has been nominated for a Best Actor award in the 2017 TV Choice Awards.

Title: NFL regular season
The National Football League (NFL) regular season begins the...

⋮

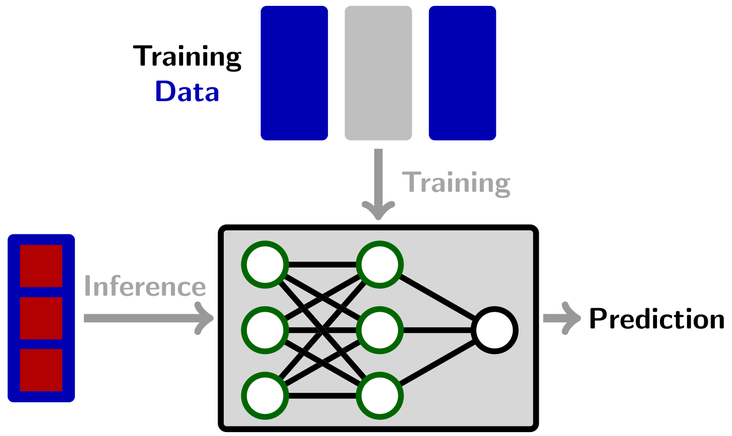
Title: TV Choice
TV Choice is a British weekly TV listings magazine published by H. Bauer Publishing. It features weekly TV broadcast programming listings and goes on sale every Tuesday.

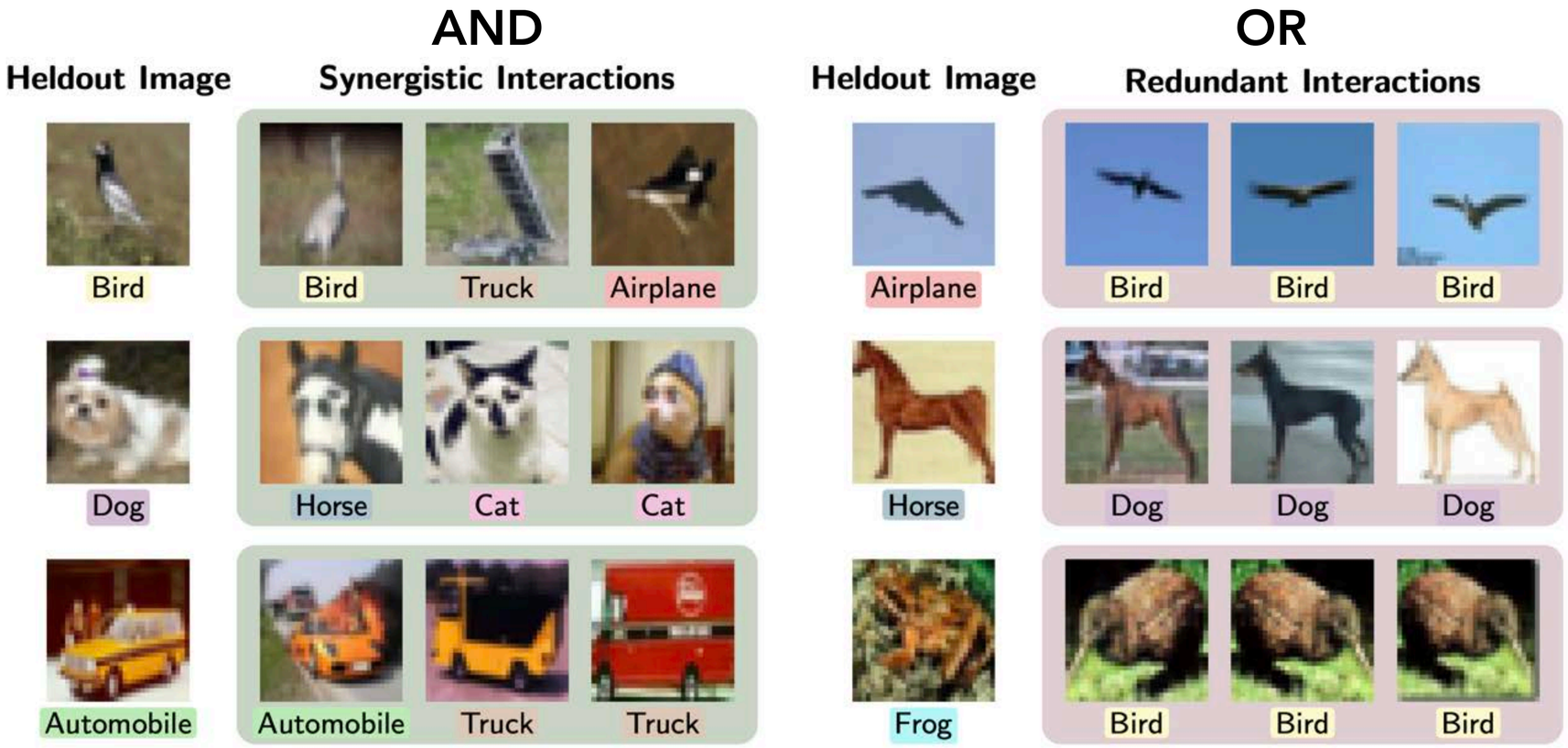
Goal: Select the sentences that cause the most significant change in the LLMs output



Not enough samples for SPEX

Data Interaction Attribution



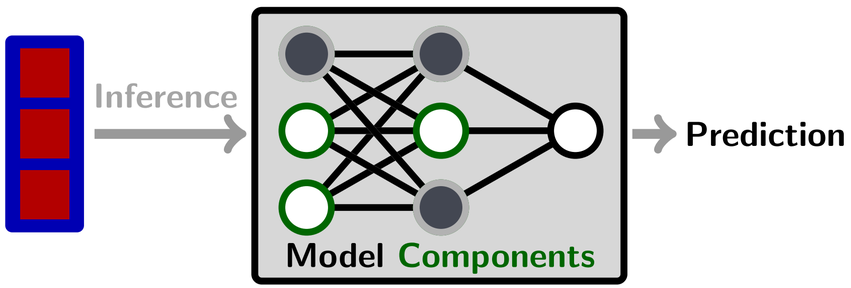


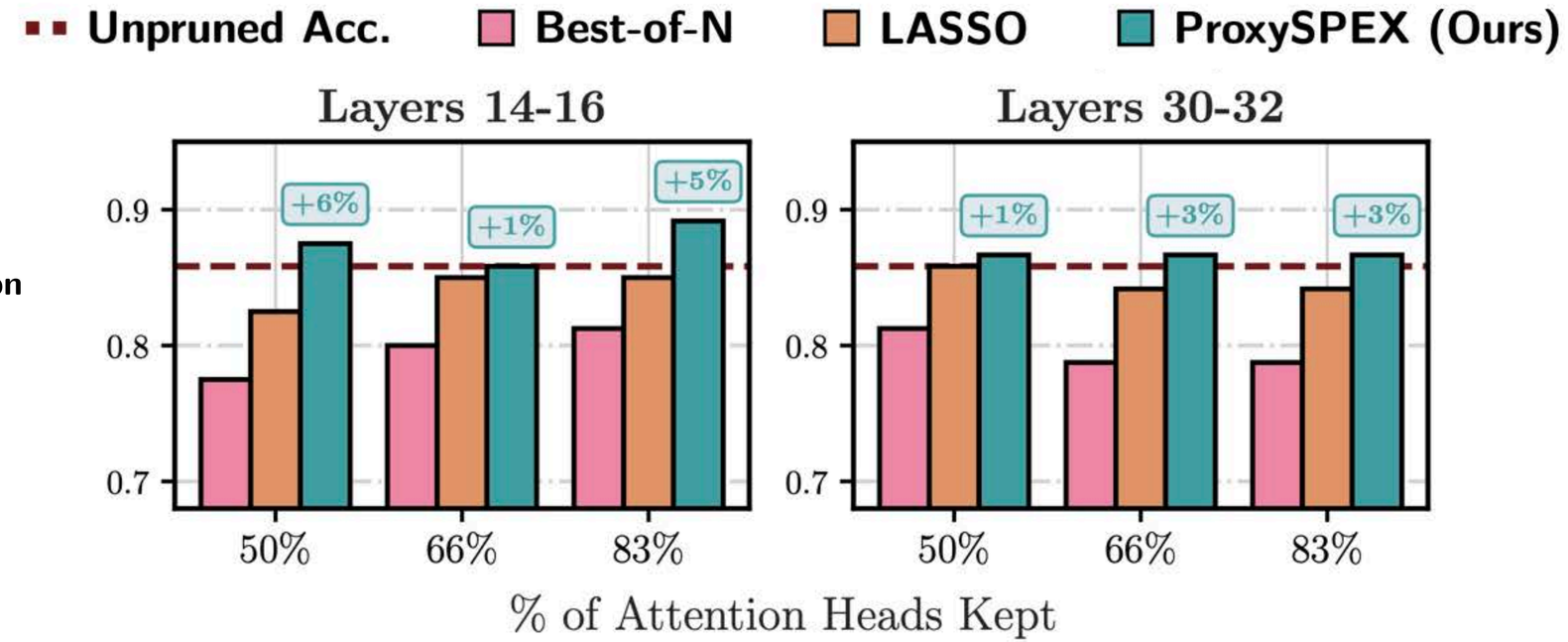
Synergistic (AND) Interactions: Data that shape decision making in a way they couldn't individually.

Redundant (OR) Interactions: Semantic duplicate for the purpose of classification.

Model Component Pruning

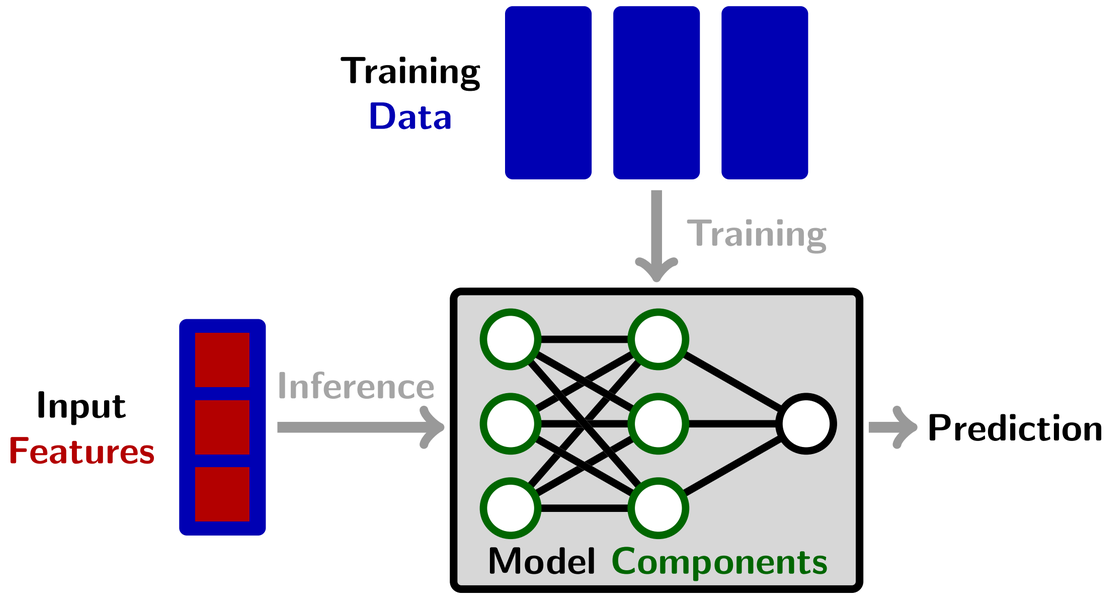
Task: Maximize accuracy on History MMLU benchmark keeping only $k\,\%$ of attention heads of Llama-3.1-8B





Strategic pruning can improve accuracy as compared to complete model!

Summary



Developed and applied *scalable* algorithms for attribution and selection for features, data, and components

Scale achieved through identifying Fourier structures

SPEX

ProxySPEX