

# STAT 992: Foundation Models for Biomedical Data

---

Ben Lengerich

Lecture 16: Dataset Debugging

April 20, 2026



# Recall: LLM Pre-training

- **Goal:** Learn a model of language from raw text
  - **Probabilistic objective:** Max log-likelihood of observed seqs

$$\max_{\theta} \sum_i \sum_t \log P_{\theta}(X_{i,t} | X_{i,<t})$$

- **MLE Implication:** Model learns to reproduce frequency of observations in dataset, regardless of observations are “correct”

# MLE: A KL-Divergence view

$$KL(P_{data} \parallel P_{\theta}) = E_{x \sim P_{data}}[-\log P_{\theta}(x)] + \text{const.}$$

$$\operatorname{argmax}_{\theta} E_{x \sim P_{data}}[\log P_{\theta}(x)] = \operatorname{argmin}_{\theta} KL(P_{data} \parallel P_{\theta})$$

- Must spread probability mass to cover observed sequences, even if incoherent
  - Repetition and genericity ("The man said the man said...")
  - Poor calibration on out-of-distribution prompts
  - Memorization of rare patterns

# Data: The Real Bottleneck





# What can go wrong in training data?

---

- Label errors
- Duplicates/near-duplicates
- Distribution gaps
- Annotation artifacts
- Data leakage
- Confounders
- Poisoning/contamination

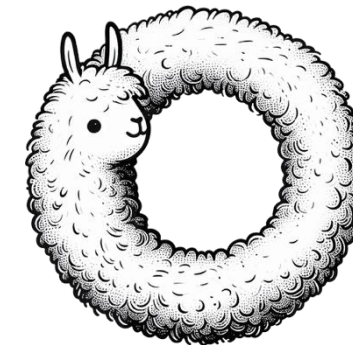
More info: <https://www.nature.com/articles/s42256-022-00516-1>

# Modern pre-training datasets

- FineWeb
- Dolma
- RedPajama
- SlimPajama
- Cosmopedia (synthetic)



TxT360



LLM360

**Dataset composition** strongly affects model behavior



# Synthetic Data

---

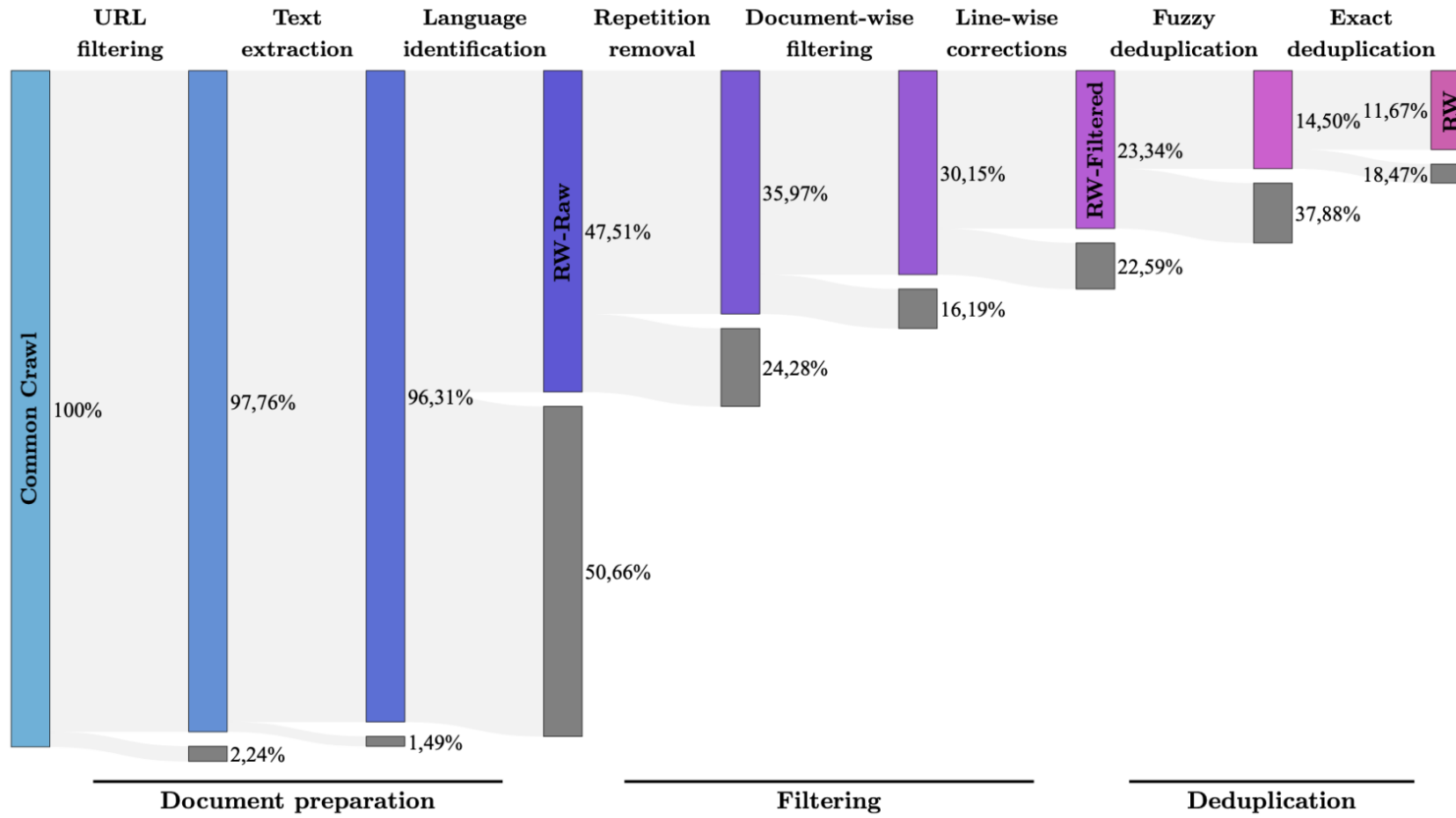
Modern LLMs train on large amounts of synthetic data.

Examples:

- reasoning datasets
- synthetic textbooks
- distillation datasets

Synthetic data expands the training distribution.

# Data filtering



Penedo, et al. (2023) The Refined Web dataset for Falcon LLM

# Embedding-Based Data Curation

---

- Core idea: embed all data points with a pretrained model (CLIP, BERT, etc.), then use the embedding space to detect problems
  - Near-duplicates cluster together → deduplicate
  - Outliers sit far from clusters → flag or remove
  - Cross-modal inconsistencies (image-text mismatch) → filter
- SemDeDup [[Abbas et al 2024](#)]
  - Embed with a foundation model, cluster, remove semantic duplicates within clusters
  - Removes 50% of data with minimal performance loss on downstream tasks
  - Key insight: exact dedup misses paraphrases and reformulations
- DataComp [[Gadre et al 2023](#)]
  - Benchmark for dataset filtering strategies on a Common Crawl image-text pool
  - Best strategy: CLIP-score filtering (keep only samples where image and text embeddings are aligned)
  - Smaller, filtered datasets outperform larger unfiltered ones



# Quality Classifiers & Learned Filters

- Core idea: we can use models to score data for other models
- FineWeb [[Penedo et al 2024](#)]
  - Train a classifier to predict "educational quality" scores on web text
  - Filter Common Crawl to keep only high-scoring documents
  - Result: 1.3T tokens that outperform 15T unfiltered tokens on benchmarks
- DSIR — Data Selection with Importance Resampling [[Xie et al., 2023](#)]
  - Given a target domain, reweight/resample a large general corpus to match the target distribution
  - Uses n-gram features for fast importance weight estimation
  - Works without any foundation model — a useful baseline
- Ask-LLM [[Sachdeva et al., 2024](#)]
  - Directly prompt an LLM: "Is this a high-quality training example?"
  - Surprisingly competitive with learned classifiers
  - Raises the question: what does "quality" even mean?

# Real-world data





# Label Error Detection in Benchmarks

## Method:

- Train a model, collect predicted class probabilities for each example
- Estimate the joint distribution of noisy labels  $\times$  true labels:  $p(\tilde{y}, y^*)$
- Identify examples where the given label is likely wrong (high confidence for a different class)
- Rank by confidence gap  $\rightarrow$  prioritize for re-labeling

## Results:

- Applied to 10 major benchmarks (ImageNet, CIFAR-10/100, MNIST, QuickDraw, Amazon Reviews, etc.)
- Found label error rates of  $\sim 3.4\%$  on average
- ImageNet:  $\sim 6\%$  of validation labels are wrong
- Removing these errors and retraining improves accuracy

## PERVASIVE LABEL ERRORS IN TEST SETS DESTABILIZE MACHINE LEARNING BENCHMARKS

**Curtis G. Northcutt**  
ChipBrain, MIT  
Boston, MA, USA  
curtis@chipbrain.com  
cgn@mit.edu

**Anish Athalye**  
Dept. of EECS  
MIT  
Cambridge, MA, USA  
aathalye@mit.edu

**Jonas Mueller**  
Amazon Web Services  
East Palo Alto, CA, USA  
jonaswmueller@gmail.com



# LLMs as Data Annotators & Auditors

## LLMs as annotators:

- Gilardi et al. (2023): ChatGPT outperforms crowd-workers on text annotation tasks (relevance, stance, topic, frame detection)
- He et al., AnnoLLM (2024): LLM-based annotation with chain-of-thought explanations
- Cost comparison: ~\$0.002/label (LLM) vs. ~\$0.50/label (crowd-worker) — 250× cheaper

## LLMs as auditors:

- Use LLMs to re-annotate existing datasets, flag disagreements with human labels
- Disagreements are candidates for label errors, ambiguity, or annotation guideline failures
- "LLM-as-judge" paradigm: use one model to score the outputs or labels of another

## Limitations:

- LLMs inherit biases from pretraining (Western, English-centric, majority viewpoints)
- Struggle with specialized domain knowledge (medical coding, rare diseases)
- Confidently wrong — high fluency masks errors



# Specific Challenges of Biomedical Data

- **Challenge 1: Label noise from clinical workflows**
  - ICD codes assigned for billing, not research — systematic inaccuracies
  - "Diagnosis present on admission" flags are often unreliable
  - Lab values have instrument-specific artifacts (different hospitals, different ranges)
- **Challenge 2: Non-stationarity from treatment policies**
  - Treatment guidelines change over time → the data distribution shifts
  - Goodhart's Law in action (callback to slide 10): biomarker thresholds drive treatment, treatment changes outcomes
  - A model trained on 2015 data may learn associations that are reversed by 2025
- **Challenge 3: Survivorship & selection bias**
  - EHR data only contains patients who showed up at the hospital
  - ICU datasets are conditioned on being sick enough to be admitted
- **Challenge 4: Multi-site distribution shift**
  - Same disease, different documentation practices across hospitals
  - MIMIC (Boston) vs. eICU (multi-center) vs. UK Biobank
  - Foundation models pretrained on one site may not transfer

# Real implications of real-world dataset debugging?



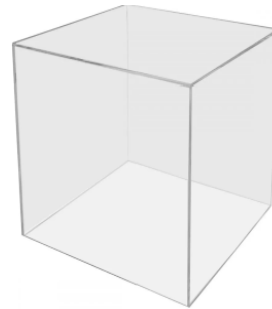
# Real-world data are confusing!



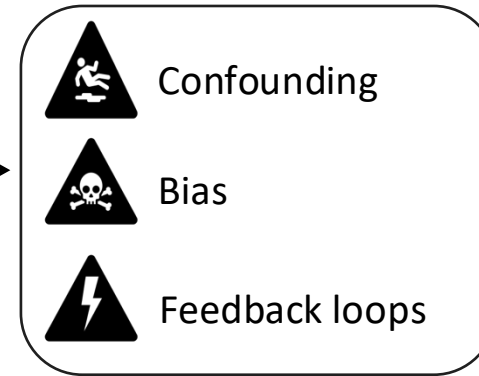
Real-world complexity



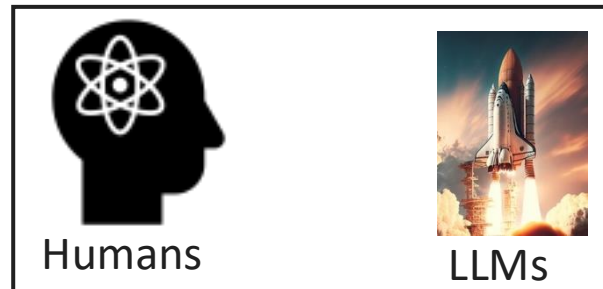
Dataset  
(here be dragons!)



Interpretable  
Data Science



Insight



Reasoning with prior knowledge

# Goodheart's Law

---

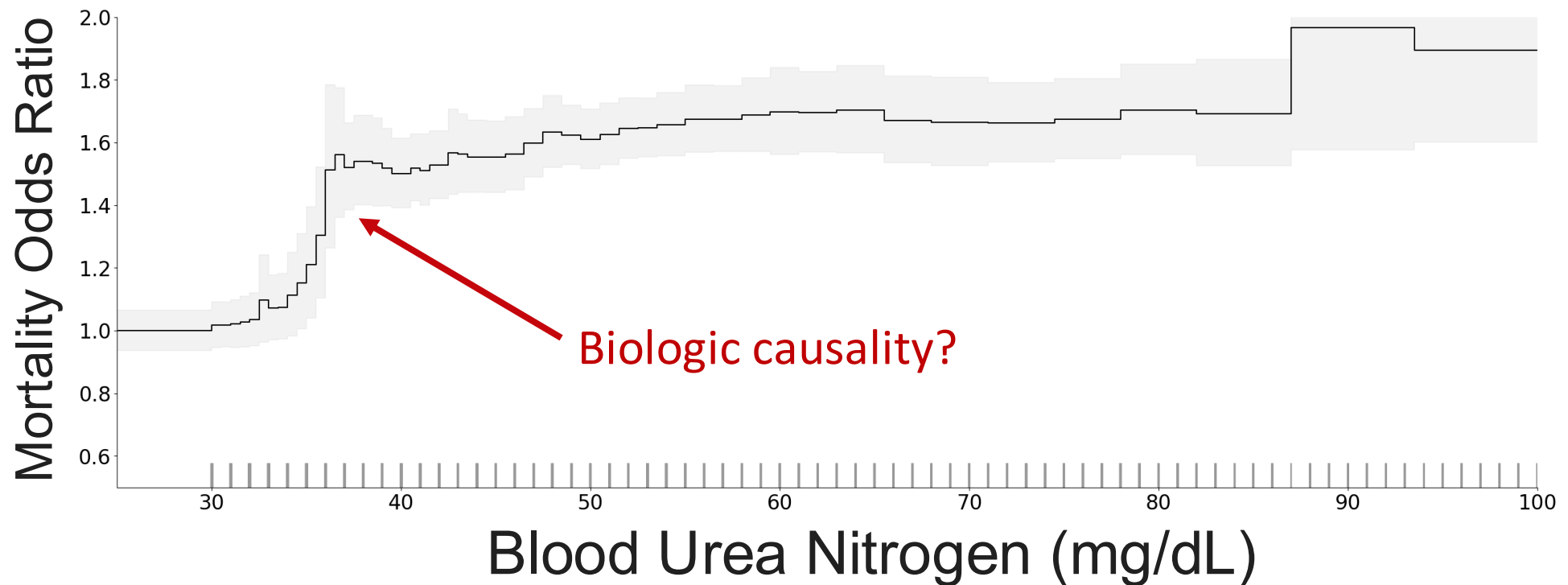
When a measure becomes a target, it ceases to be a good measure.

## A form of **Goodheart's Law** for biomarkers

When a biomarker is used to guide treatment decisions, it ceases to predict outcomes.

# Real-world effects are surprising and may not be causal

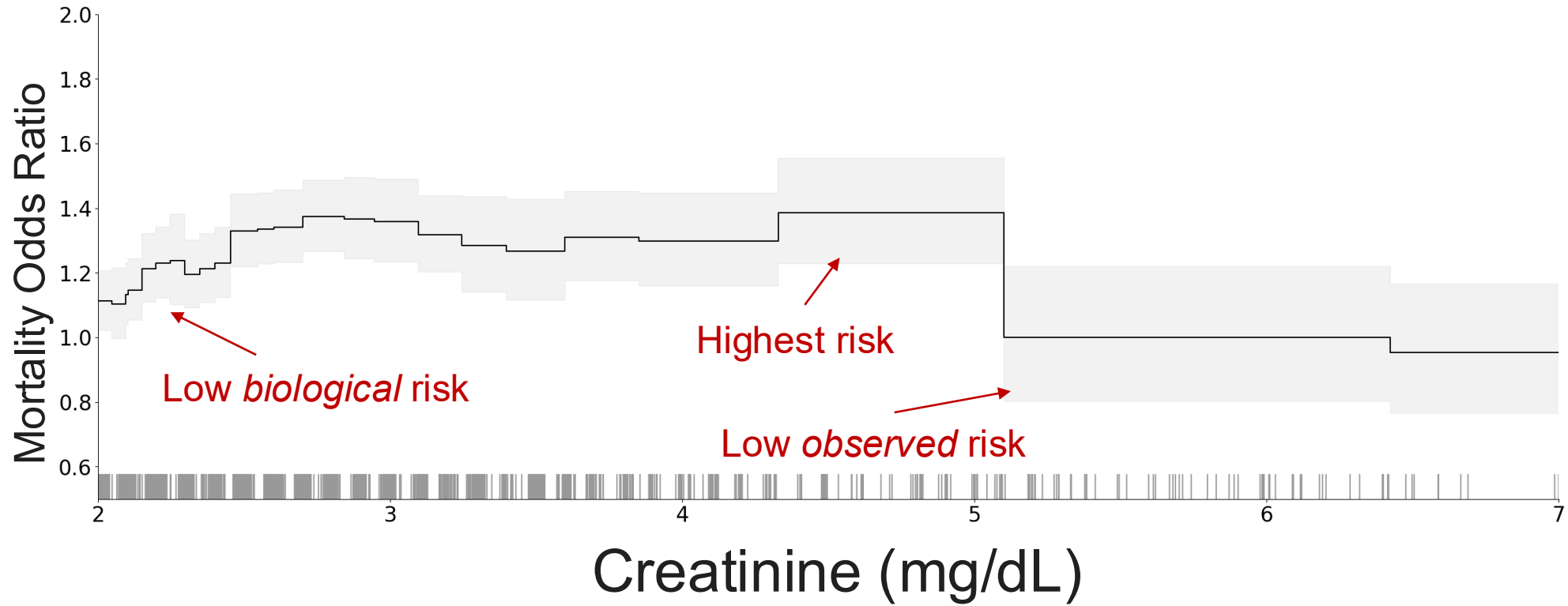
In-hospital mortality risk for hospitalized patients with pneumonia:



Lengerich et al. *NPJ Digital Medicine* 2026 (to appear)

# Real-world effects are surprising and may not be causal

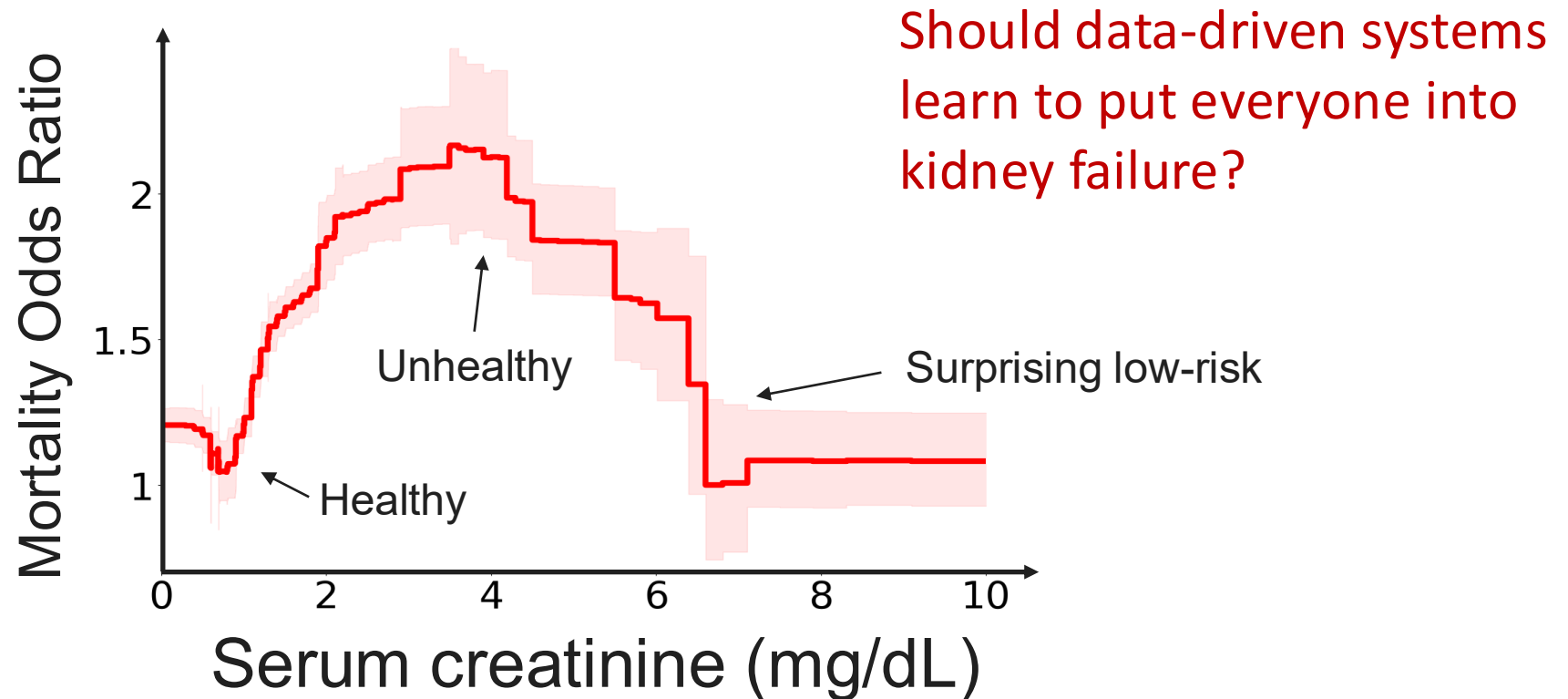
In-hospital mortality risk for hospitalized patients with pneumonia:



Lengerich et al. *NPJ Digital Medicine* 2026 (to appear)

# Real-world effects are surprising and may not be causal

MIMIC-IV mortality risk for hospitalized patients:

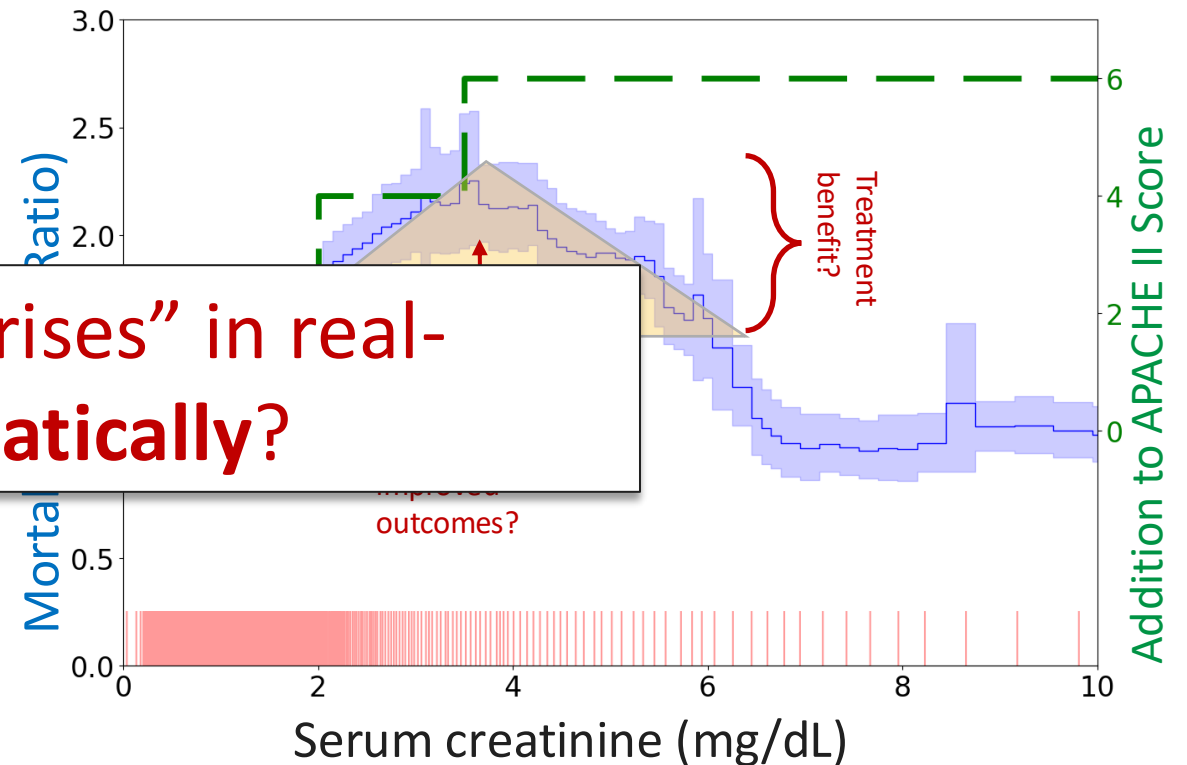


Lengerich et al. *NPJ Digital Medicine* 2026 (to appear)

# “Surprises” in real-world data have varied implications

- Hidden complexity
- Interface between natural cause + human behavior
- Data problem?
- Experimental design problem?
- **Opportunity?**

Can we find “surprises” in real-world data **automatically**?





# What would it take to find surprises automatically?

---

- An expectation of “normal”
- Scalable anomaly-detection procedure
- Uncertainty quantification

## Today:

1. Statistical Tests
2. LLMs

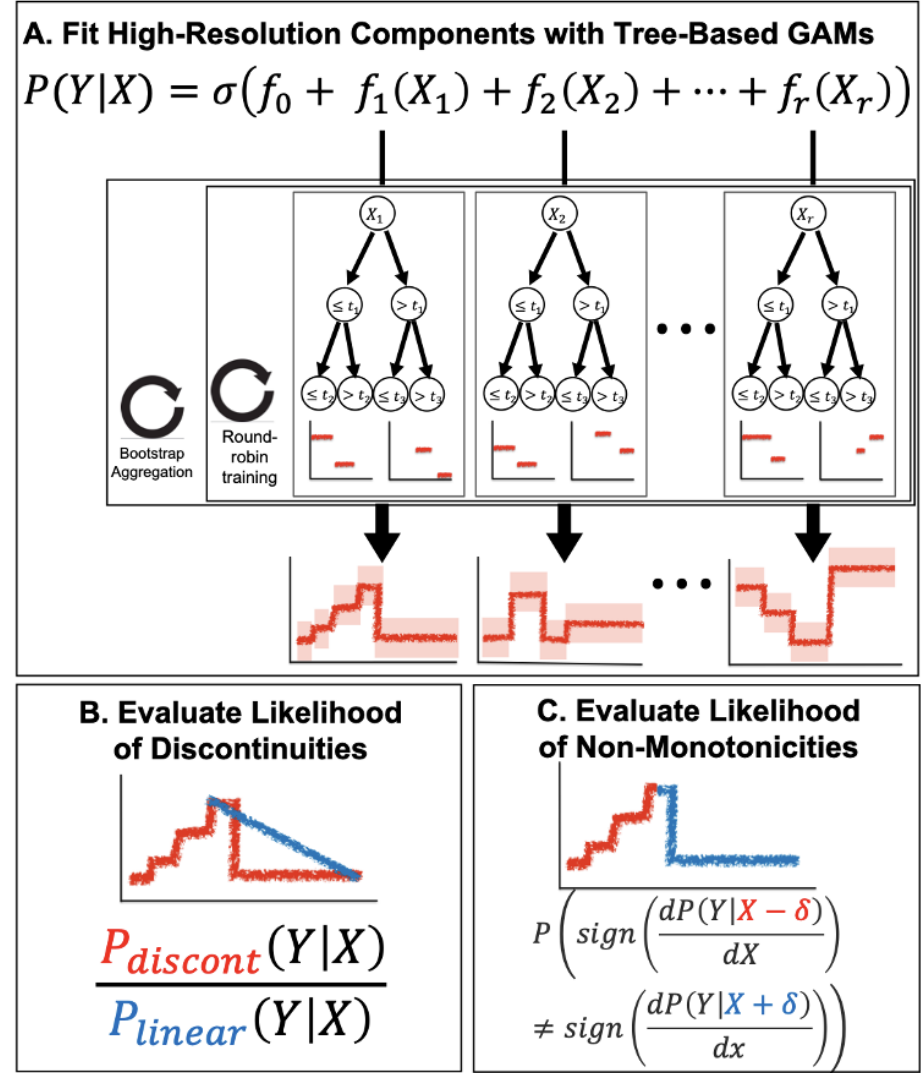
# Systematically identifying these artifacts

1. Estimate an interpretable ML model.

- Decomposes outcome into univariate components
- Corrects for correlations, etc.



2. Test each component for surprising shapes.



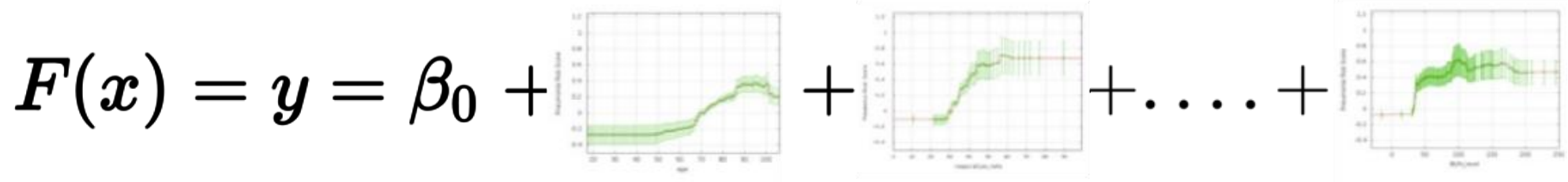
# Generalized Additive Models (GAMs)

[Hastie and Tibshirani (1993)]

Decompose complex outcomes into a sum of univariate functions

$$F(x) = y = \beta_0 + f_1(x_1) + f_2(x_2) + \dots + f_r(x_r)$$

Components can be individually visualized:

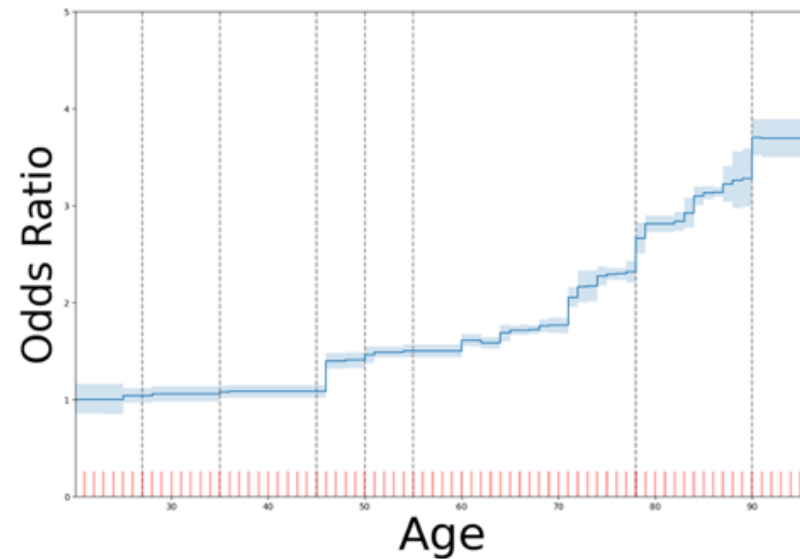


# Automatically finding discontinuous effects

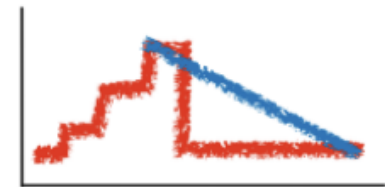
```
# Automatically find discontinuous effects.
from ebm_utils.analysis.changepoints import find_and_plot_discontinuities
discontinuities = find_and_plot_discontinuities(X, Y, min_samples=200, min_effect_size=1.2)
```



Feature	Value	# Samples	Effect Size	P-Ratio
Glucose	106.000000	5268.0	1.205830	986.0
Glucose	99.000000	7978.0	1.111032	840.0
Age	50.000000	4503.0	1.199735	820.0
Age	78.000000	5824.0	1.146653	797.0
Glucose	122.000000	3851.0	1.182639	646.0
Age	90.000000	3702.0	1.188723	640.0



## B. Evaluate Likelihood of Discontinuities



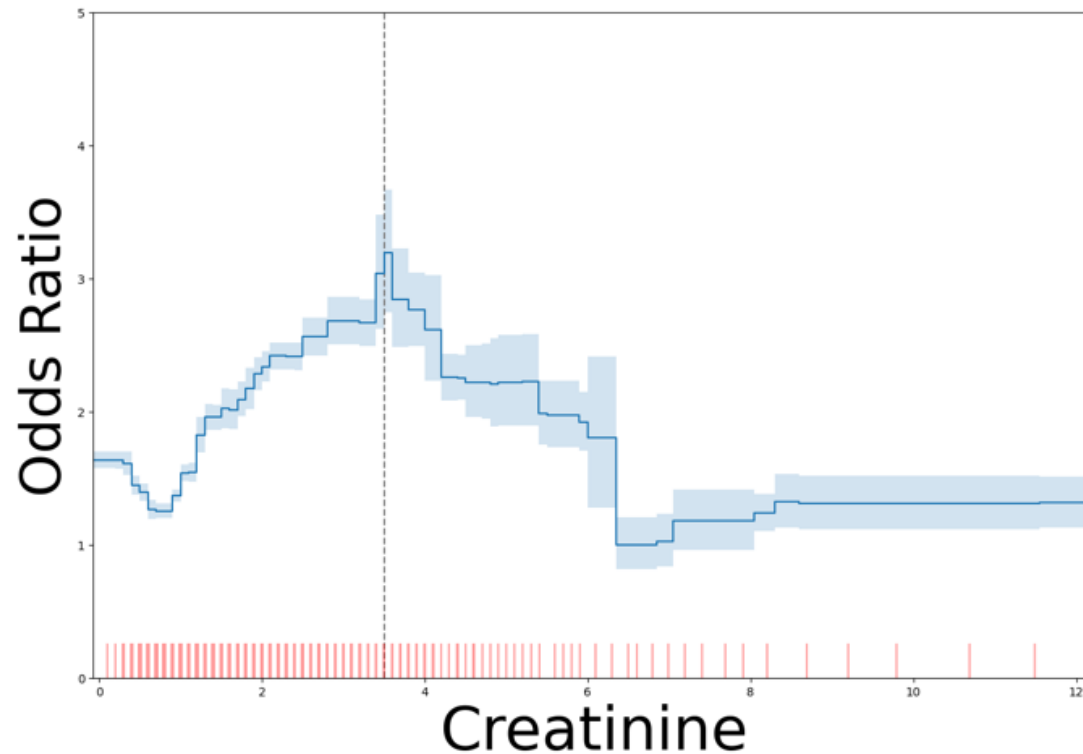
$$\frac{P_{discont}(Y|X)}{P_{linear}(Y|X)}$$

# Automatically finding non-monotonic effects

```
find_and_plot_non_monotonicities(X, Y, prob_threshold=0.1, counter_causal_only=True)
```



Feature	Value
SBP	81.0
Chloride	115.0
Creatinine	3.5
Magnesium	2.6
Platelet	16.0
Potassium	6.7
Temp	35.0



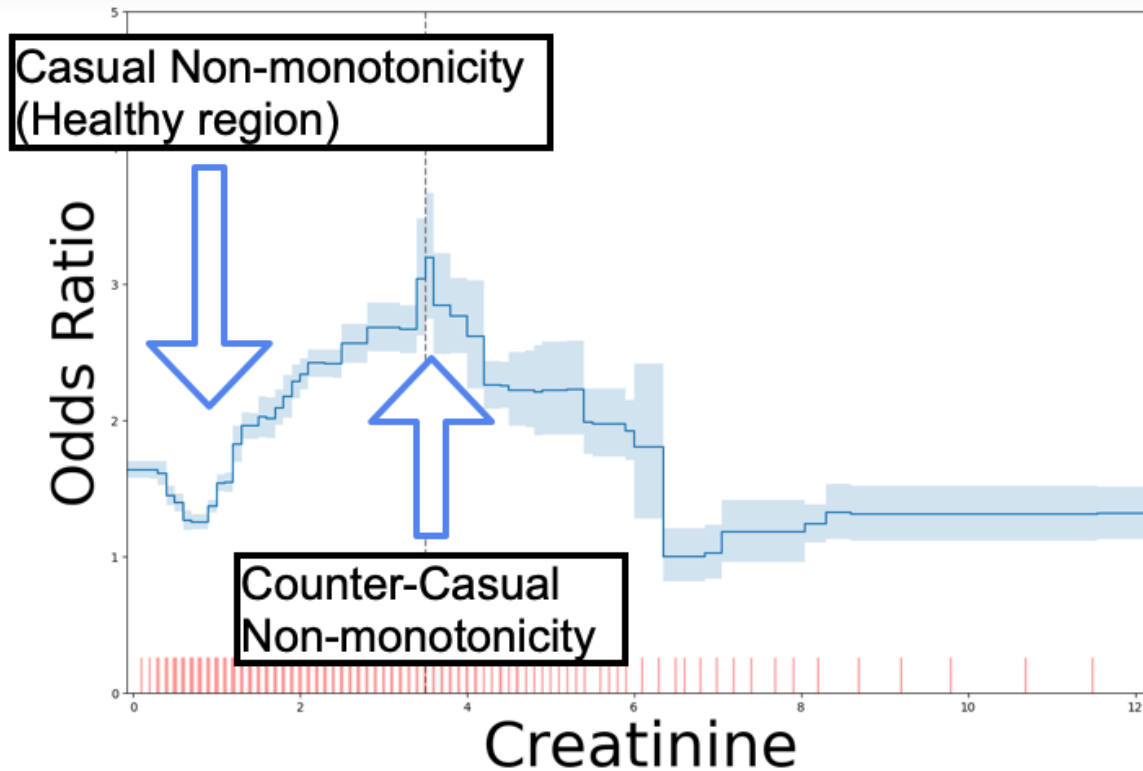
### C. Evaluate Likelihood of Non-Monotonicities

$$P \left( \text{sign} \left( \frac{dP(Y|X - \delta)}{dX} \right) \neq \text{sign} \left( \frac{dP(Y|X + \delta)}{dx} \right) \right)$$

Lengerich et al. *NPJ Digital Medicine* 2026 (to appear)

# Automatically finding non-monotonic effects

```
find_and_plot_non_monotonicities(X, Y, prob_threshold=0.1, counter_causal_only=True)
```



**C. Evaluate Likelihood of Non-Monotonicities**

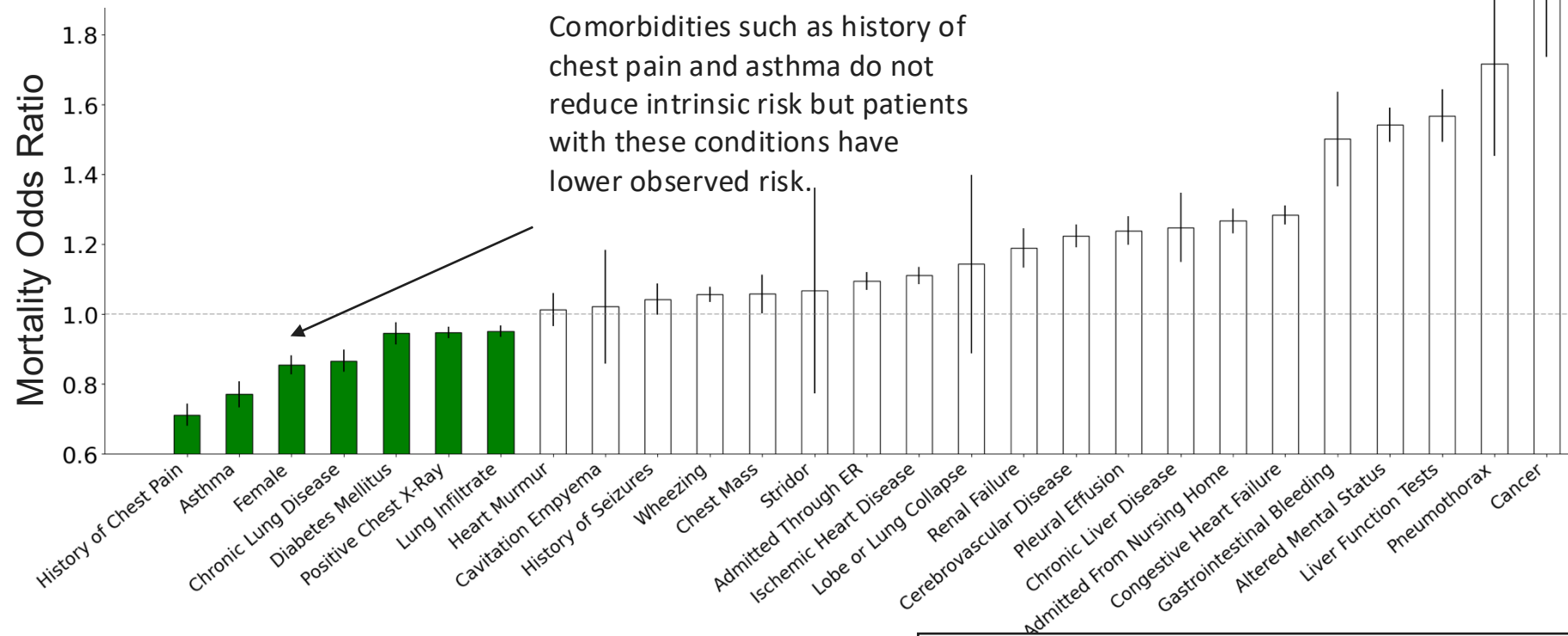
$$P\left(\text{sign}\left(\frac{dP(Y|X - \delta)}{dX}\right) \neq \text{sign}\left(\frac{dP(Y|X + \delta)}{dx}\right)\right)$$

Lengerich et al. *NPJ Digital Medicine* 2026 (to appear)

# Can we expand surprise-finding to all prior knowledge?

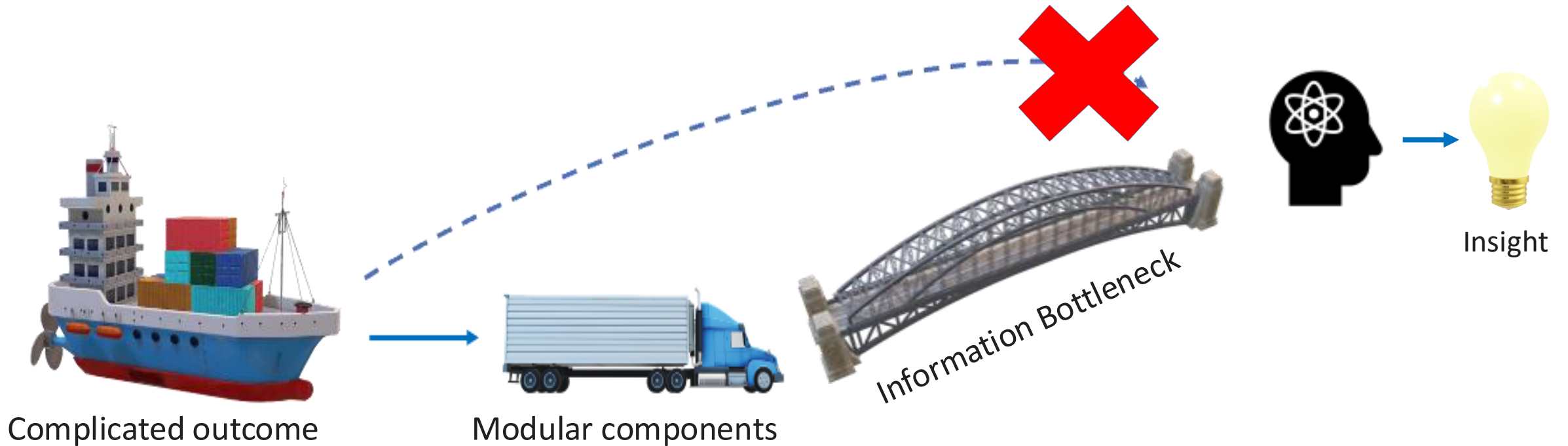
So far: Surprises contradict biological causality of **smooth, monotone** effects

What about surprises that contradict prior knowledge?



# The core idea of GAMs: Modularity → Interpretability

An **information bottleneck** limits human understanding of complicated ideas...

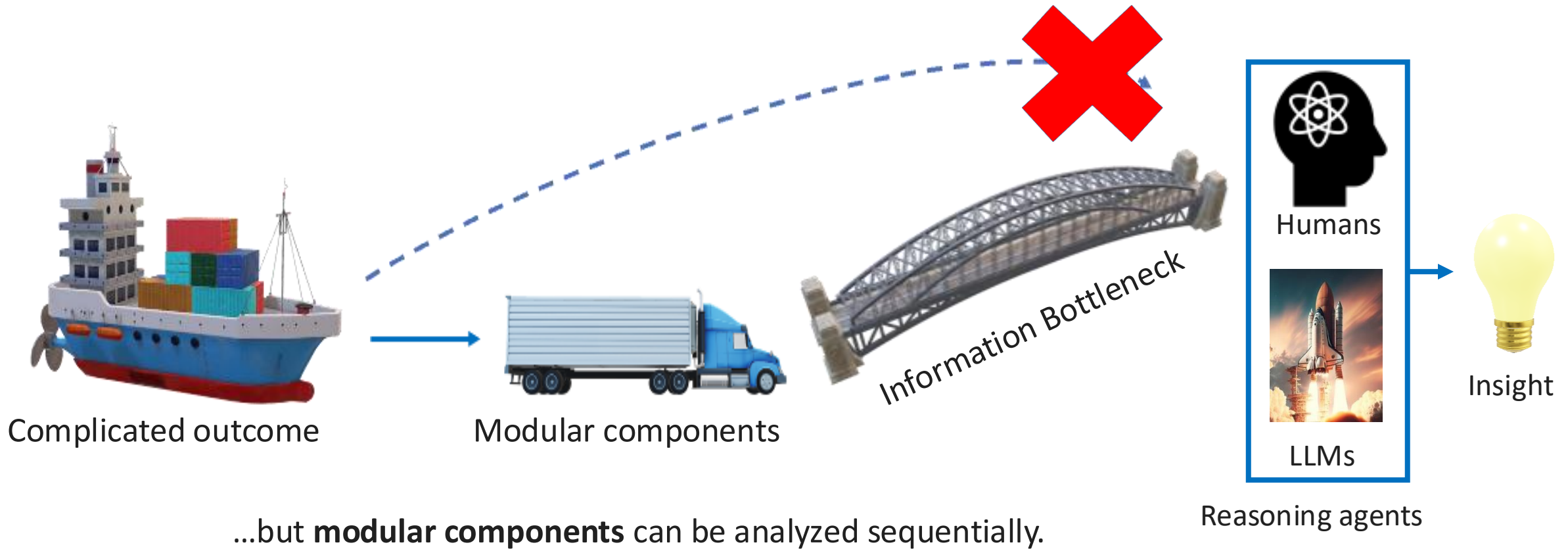


...but **modular components** can be analyzed sequentially.

*"LLMs Understand Glass-Box Models, Discover Surprises, and Suggest Repairs".*  
Lengerich et al. Arxiv 2023

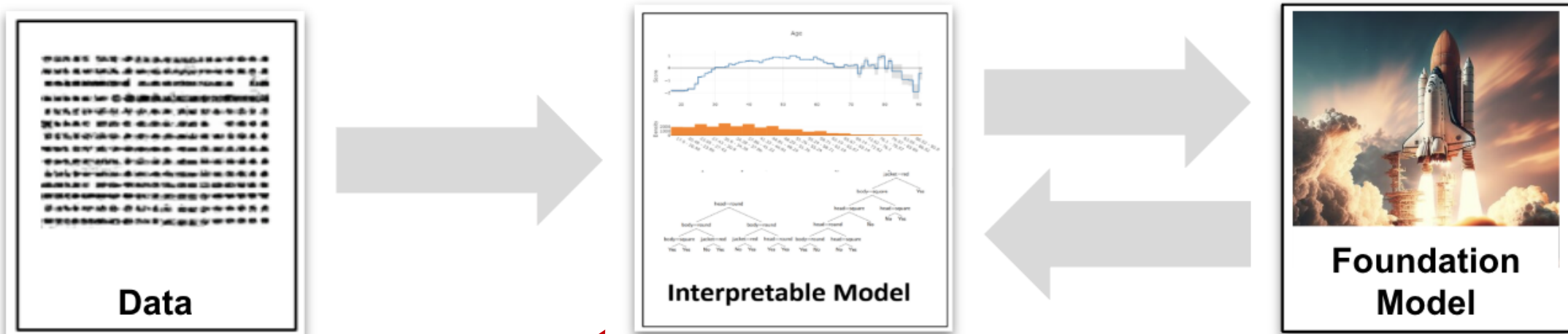
# GAMs + LLMs: Modularity → Automated Interpretability

An **information bottleneck** limits human understanding of complicated ideas...



*"LLMs Understand Glass-Box Models, Discover Surprises, and Suggest Repairs".*  
Lengerich et al. Arxiv 2023

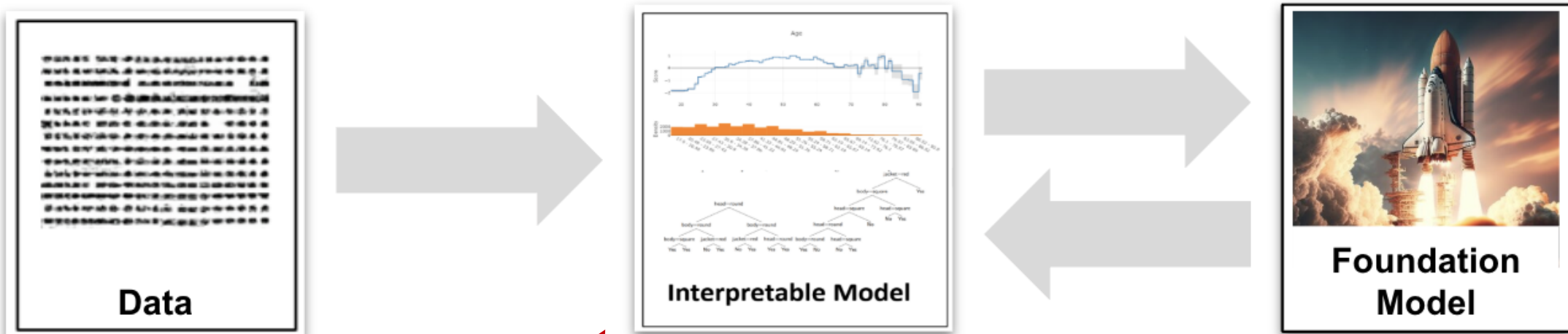
# Statistical ML + LLMs → Automated Data Science?



Privacy-preserving **data compression**

*"LLMs Understand Glass-Box Models, Discover Surprises, and Suggest Repairs".*  
Lengerich et al. Arxiv 2023

# Statistical ML + LLMs → Automated Data Science?



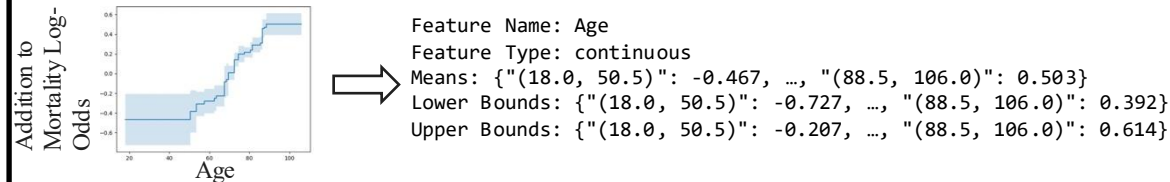
Privacy-preserving **data compression**

*"LLMs Understand Glass-Box Models, Discover Surprises, and Suggest Repairs".*  
Lengerich et al. Arxiv 2023

# Statistical ML + LLMs → Automated Data Science?

## Univariable Graphs

Univariable Graphs can be encoded as text...



...and understood by LLMs...

**GPT-4:** "[...] The general pattern of the graph indicates that as the age increases, the log-odds contribution also increases, meaning that older age is associated with a higher risk of death due to pneumonia.

In the age range from about 18 to 67.5 years, the contribution to the log-odds is negative, which means the model predicts a lower risk of death from pneumonia in this age group. Starting from around 69.5 years, the contribution becomes positive, indicating a higher risk of death from pneumonia as the patient's age increases.

It's important to note that the exact turning point where the risk increases can vary, as indicated by the lower and upper bounds for each age range. These bounds represent the 95% confidence interval, which is a measure of the uncertainty around the estimate. However, the overall trend remains the same: older age is associated with higher risk of death from pneumonia.

...for conversational data science.

### Summarization

**Q:** Does the risk increase monotonically with age?  
**GPT-4:** "Yes, according to the provided data [...]"

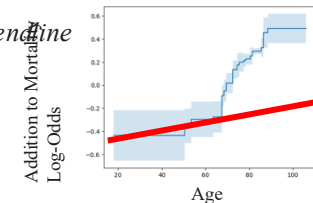
### Application

**Q:** What would be the average risk of an 82-year old patient?  
**GPT-4:** "[...] according to the model, an 82-year old patient would have a 57.2% probability of dying from pneumonia on average. ..."

### Simplification

**Q:** What is the best trendline? Express your answer as a linear equation of the form  $y=mx+b$ .

**GPT-4:** "[...] Therefore, the best trendline (linear equation) for the graph is:  
 The graph is:  
 $y \approx -0.558006 + 0.003896x$ "



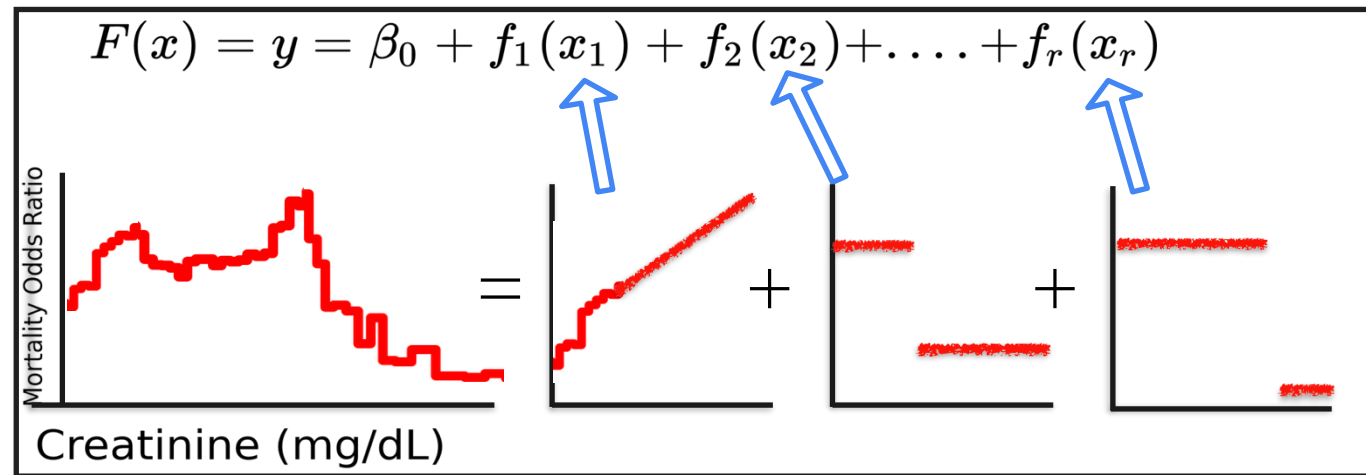
	GPT-3.5	GPT-4
Reading a Value from a Graph	3/75	64/75
Deciding Monotonicity	2/12	11/12
Finding the Largest Jump	9/31	27/31

"LLMs Understand Glass-Box Models, Discover Surprises, and Suggest Repairs".  
 Lengerich et al. Arxiv 2023

# Statistical ML + LLMs → Automated Data Science?

## Multivariable Graphs

**Hierarchical** reasoning through separable component functions



*"LLMs Understand Glass-Box Models, Discover Surprises, and Suggest Repairs".*  
Lengerich et al. Arxiv 2023

# Surprise-Finding: LLMs vs Human experts

Benchmarked in a **blinded study** against doctors

1. GPT and 4 Doctors independently evaluate effects from a GAM.
2. Doctors grade other responses. Tell them it's doctors rating doctor explanations. Secretly, LLM explanations were mixed in.

Anomaly Detector	# of Anomalies per Feature	% Ratings of >2 ("Agree")	
		Anomaly identification	Anomaly explanation
Self (Doctor)	0.64(0.55,0.73)	98.9(95.8,100.0)	92.2(70.2,100.0)
Other Doctor	0.64(0.55,0.73)	92.0(85.6,98.4)	82.0(71.4,92.6)
GPT-4	1.0(0.93,1.07)	66.7(54.2,79.2)	63.0(53.6,72.4)

↑  
But more exhaustive

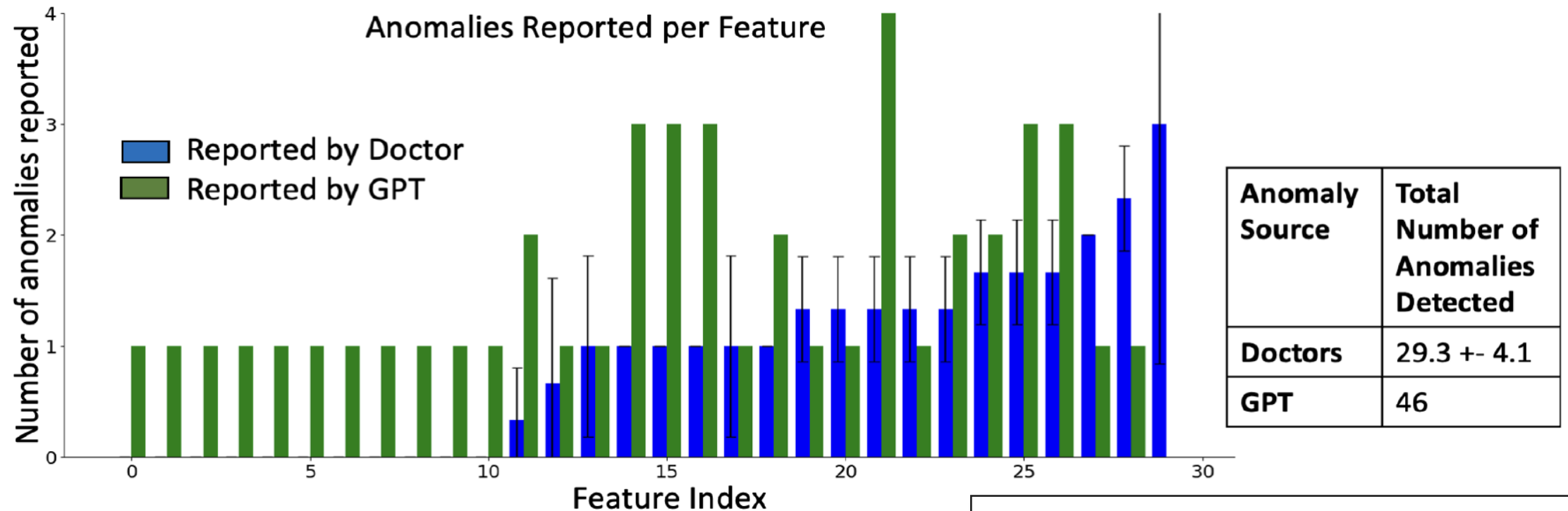
↙ ↘  
GPT-4 not as good as doctors

Lengerich et al. *JAMIA Open* 2026 (to appear)

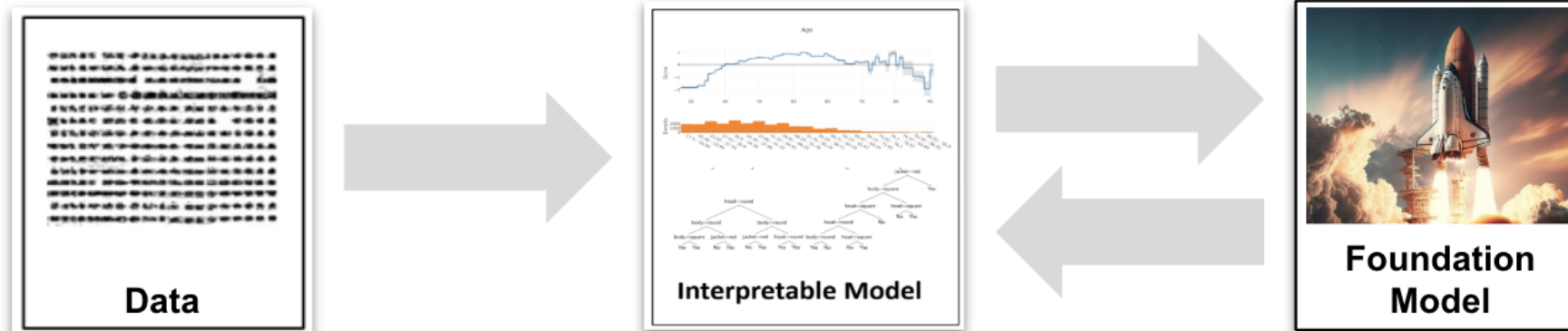
# Surprise-Finding: LLMs vs Human experts

Benchmarked in a **blinded study** against doctors

1. GPT and 4 Doctors independently evaluate effects from a GAM.
2. Doctors grade other responses. Tell them it's doctors rating doctor explanations. Secretly, LLM explanations were mixed in.



# This is just the start of automated data science





# The big picture: Foundation Models for Dataset Debugging

- **Before Training**

- Embedding-based curation (DataComp, SemDeDup)
- Quality classifiers (FineWeb-Edu, Ask-LLM)
- Deduplication & filtering pipelines (RefinedWeb)

- **After Training**

- Label error detection (Confident Learning)
- Training dynamics analysis (Dataset Cartography)
- Data attribution (TRAK, Datamodels, Influence Functions)
- Memorization auditing (Carlini et al.)

- **Human-in-the-Loop / LLM-in-the-Loop**

- GAMs + statistical tests for surprise-finding, LLMs as automated auditors
- Conversational data science for iterative debugging



# Open Problems & Future Directions

---

- **Problem 1:** Debugging at trillion-token scale — current methods don't scale to pretraining corpora
- **Problem 2:** Multimodal consistency — how do you debug alignment between images, text, genomics, EHR codes simultaneously?
- **Problem 3:** Synthetic data feedback loops — models generating training data for the next generation of models; model collapse risk (Shumailov et al., 2023)
- **Problem 4:** Evaluating the debuggers — how do we know if a "surprise" is a real artifact vs. a false alarm? Ground truth is expensive
- **Problem 5:** Causal reasoning about artifacts — knowing *that* an artifact exists vs. knowing *why* it exists (requires causal models + domain knowledge + LLM reasoning)
- **Problem 6:** Closing the loop — from debugging to repair. Can we automatically fix datasets, not just flag problems?

Questions?

