STAT 453: Introduction to Deep Learning and Generative Models

Ben Lengerich

Lecture 25: Alignment, Explainability, and Open Directions

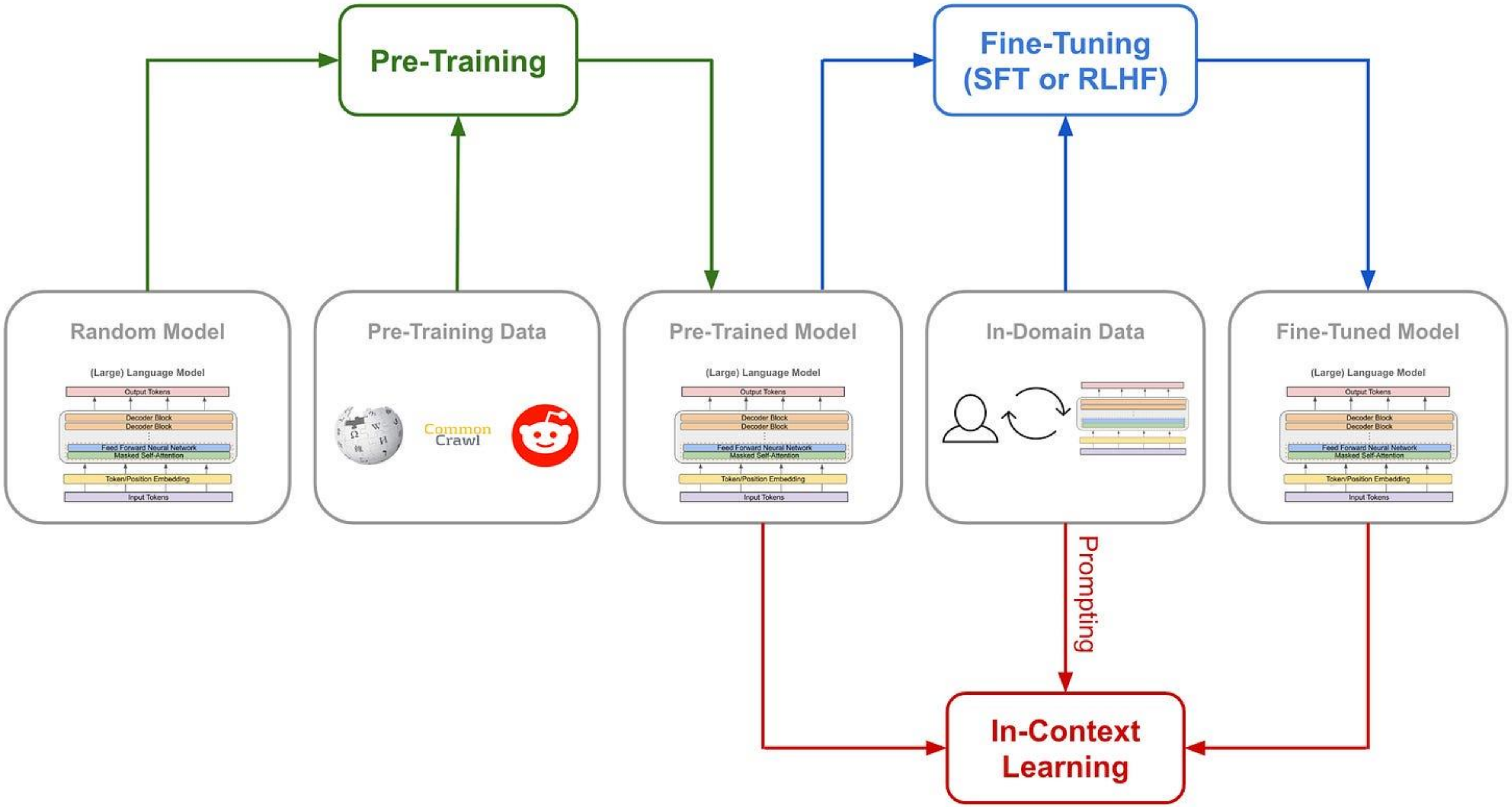
December 1, 2025

Reading: See course homepage

Today

- Optional HW5
- Project Presentation Sign-up
  - **4 minute presentations!**
- Project Final Report
  - Due Friday December 12th
  - Submit PDF via Canvas
- Final Exam
  - December 17th, 5:05-7:05PM
  - **Science 180**
  - **Study Guide Released**

Last time



https://cameronrwolfe.substack.com/p/understanding-and-using-supervised

Explainability, Alignment

2016: Setting the stage

Lipton (2016) - Interpretability is invoked when **metrics ≠ objectives**

arXiv:1606.03490v3 [cs.LG] 6 Mar 2017

The Mythos of Model Interpretability

Zachary C. Lipton [1]

Abstract

Supervised machine learning models boast remarkable predictive capabilities. But can you trust your model? Will it work in deployment? What else can it tell you about the world? We want models to be not only good, but interpretable. And yet the task of *interpretation* appears underspecified. Papers provide diverse and sometimes non-overlapping motivations for interpretability, and offer myriad notions of what attributes render models interpretable. Despite this ambiguity, many papers proclaim interpretability axiomatically, absent further explanation. In this paper, we seek to refine the discourse on interpretability. First, we examine the motivations underlying interest in interpretability, finding them to be diverse and occasionally discordant. Then, we address model properties and techniques thought to confer interpretability, identifying transparency to humans and post-hoc explanations as competing notions. Throughout, we discuss the feasibility and desirability of different notions, and question the oft-made assertions that linear models are interpretable and that deep neural networks are not.

1. Introduction

As machine learning models penetrate critical areas like medicine, the criminal justice system, and financial markets, the inability of humans to understand these models seems problematic (Caruana et al., 2015; Kim, 2015). Some suggest *model interpretability* as a remedy, but few articulate precisely *what* interpretability means or *why* it is important. Despite the absence of a definition, papers frequently make claims about the interpretability of various models. From this, we might conclude that either: (i) the definition of interpretability is universally agreed upon, but no one has managed to set it in writing, or (ii) the term interpretability is ill-defined, and thus claims regarding interpretability of various models may exhibit a quasi-scientific character. Our investigation of the literature suggests the latter to be the case. Both the motives for interpretability and the technical descriptions of interpretable models are diverse and occasionally discordant, suggesting that interpretability refers to more than one concept. In this paper, we seek to clarify both, suggesting that *interpretability* is not a monolithic concept, but in fact reflects several distinct ideas. We hope, through this critical analysis, to bring focus to the dialogue.

Here, we mainly consider supervised learning and not other machine learning paradigms, such as reinforcement learning and interactive learning. This scope derives from our original interest in the oft-made claim that linear models are preferable to deep neural networks on account of their interpretability (Lou et al., 2012). To gain conceptual clarity, we ask the refining questions: *What is interpretability and why is it important?* Broadening the scope of discussion seems counterproductive with respect to our aims. For research investigating interpretability in the context of reinforcement learning, we point to (Dragan et al., 2013) which studies the human interpretability of robot actions. By the same reasoning, we do not delve as much as other papers might into Bayesian methods, however try to draw these connections where appropriate.

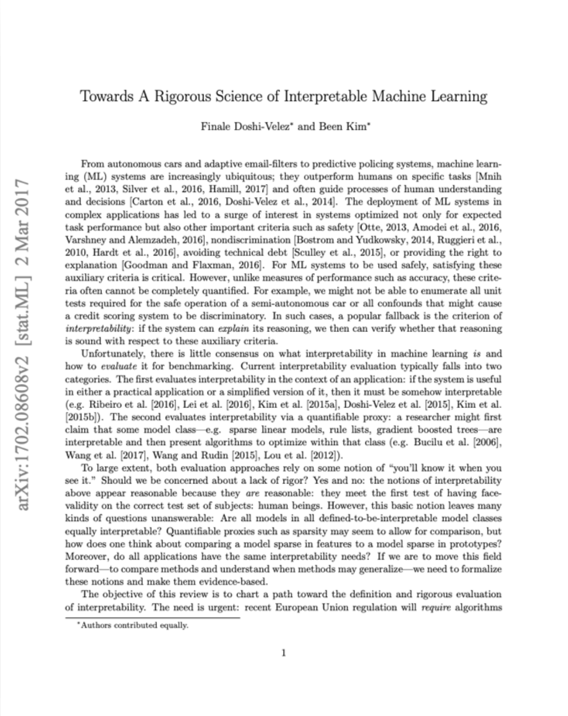
To ground any discussion of what might constitute interpretability, we first consider the various desiderata put forth in work addressing the topic (expanded in §2). Many papers propose interpretability as a means to engender trust (Kim, 2015; Ridgeway et al., 1998). But what is trust? Does it refer to faith in a model's performance (Ribeiro et al., 2016), robustness, or to some other property of the decisions it makes? Does interpretability simply mean a low-level mechanistic understanding of our models? If so does it apply to the features, parameters, models, or training algorithms? Other papers suggest a connection between an interpretable model and one which uncovers causal structure in data (Athey & Imbens, 2015). The legal notion of a *right to explanation* offers yet another lens on interpretability.

Often, our machine learning problem formulations are im-

[1] University of California, San Diego. Correspondence to: Zachary C. Lipton <zlipton@cs.ucsd.edu>.

2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016), New York, NY, USA. Copyright by the author(s).

Doshi-Velez & Kim (2017) - Three modes of evaluation: **application**-grounded, **human**-grounded, **functionally**-grounded.



arXiv:1702.08608v2 [stat.ML] 2 Mar 2017

Towards A Rigorous Science of Interpretable Machine Learning

Finale Doshi-Velez* and Been Kim*

From autonomous cars and adaptive email-filters to predictive policing systems, machine learning (ML) systems are increasingly ubiquitous; they outperform humans on specific tasks [Mnih et al., 2013, Silver et al., 2016, Hamill, 2017] and often guide processes of human understanding and decisions [Carton et al., 2016, Doshi-Velez et al., 2014]. The deployment of ML systems in complex applications has led to a surge of interest in systems optimized not only for expected task performance but also other important criteria such as safety [Otte, 2013, Amodei et al., 2016, Varshney and Alemzadeh, 2016], nondiscrimination [Bostrom and Yudkowsky, 2014, Ruggieri et al., 2010, Hardt et al., 2016], avoiding technical debt [Sculley et al., 2015], or providing the right to explanation [Goodman and Flaxman, 2016]. For ML systems to be used safely, satisfying these auxiliary criteria is critical. However, unlike measures of performance such as accuracy, these criteria often cannot be completely quantified. For example, we might not be able to enumerate all unit tests required for the safe operation of a semi-autonomous car or all confounds that might cause a credit scoring system to be discriminatory. In such cases, a popular fallback is the criterion of *interpretability*: if the system can *explain* its reasoning, we then can verify whether that reasoning is sound with respect to these auxiliary criteria.

Unfortunately, there is little consensus on what interpretability in machine learning *is* and how to *evaluate* it for benchmarking. Current interpretability evaluation typically falls into two categories. The first evaluates interpretability in the context of an application: if the system is useful in either a practical application or a simplified version of it, then it must be somehow interpretable (e.g. Ribeiro et al. [2016], Lei et al. [2016], Kim et al. [2015a], Doshi-Velez et al. [2015], Kim et al. [2015b]). The second evaluates interpretability via a quantifiable proxy: a researcher might first claim that some model class—e.g. sparse linear models, rule lists, gradient boosted trees—are interpretable and then present algorithms to optimize within that class (e.g. Bucilu et al. [2006], Wang et al. [2017], Wang and Rudin [2015], Lou et al. [2012]).

To large extent, both evaluation approaches rely on some notion of "you'll know it when you see it." Should we be concerned about a lack of rigor? Yes and no: the notions of interpretability above appear reasonable because they *are* reasonable: they meet the first test of having face-validity on the correct test set of subjects: human beings. However, this basic notion leaves many kinds of questions unanswerable: Are all models in all defined-to-be-interpretable model classes equally interpretable? Quantifiable proxies such as sparsity may seem to allow for comparison, but how does one think about comparing a model sparse in features to a model sparse in prototypes? Moreover, do all applications have the same interpretability needs? If we are to move this field forward—to compare methods and understand when methods may generalize—we need to formalize these notions and make them evidence-based.

The objective of this review is to chart a path toward the definition and rigorous evaluation of interpretability. The need is urgent: recent European Union regulation will *require* algorithms

*Authors contributed equally.

1

2017-2020: Fragmented Approaches

Post-Hoc	Transparency	Mechanistic
• LIME, SHAP, IG • Became industry standard	• GAMs, Monotonic Nets • Niche in healthcare/tabular	• Circuits, probing, feature geometry in LLMs • Technically deep, but rarely user-facing

Cracks in Post-Hoc Explanations

- Popular tools often look convincing but don't guarantee fidelity.
- **Adebayo et al. (2018)**: Saliency maps can be insensitive to model weights.
- **Slack et al. (2020)**: Easy to fool LIME and SHAP.
- **Jacovi & Goldberg (2020)**: Faithfulness vs plausibility gap.
- **Rudin (2019)**: Call to abandon post-hoc in high-stakes settings.

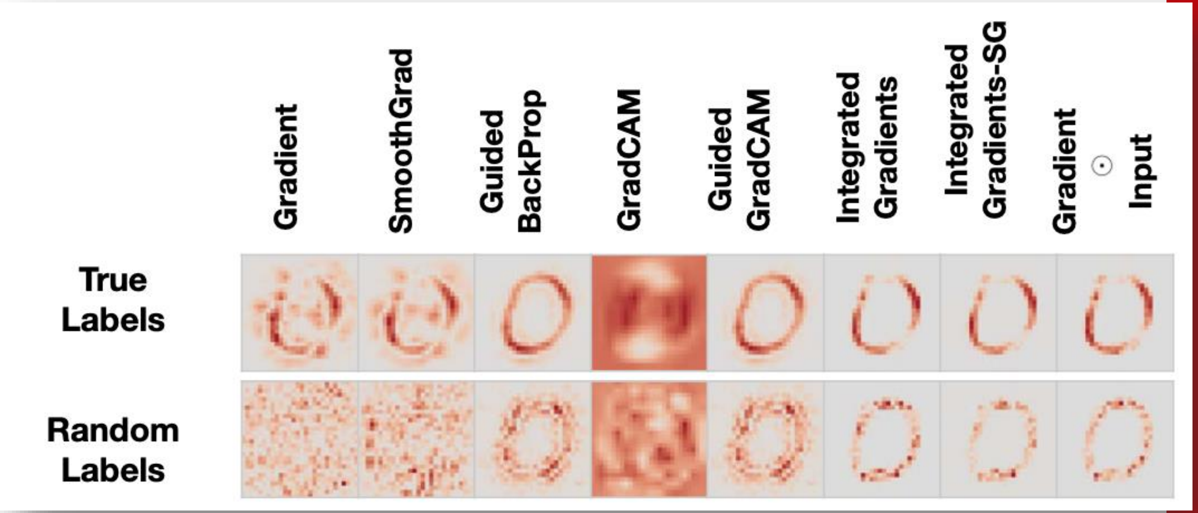


Figure 6 from *Adebayo et al. 2018*

Foundation Models take the field

2017 - Transformer

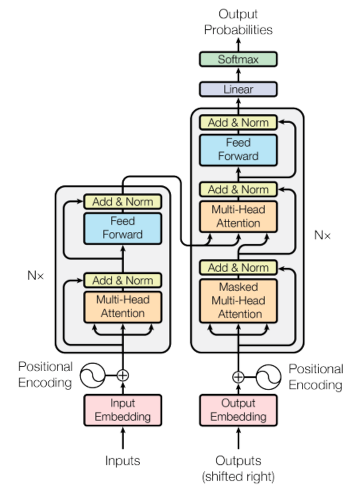


Figure 1: The Transformer - model architecture.

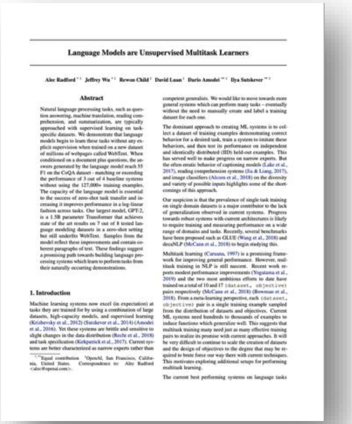
2018 - GPT

117M params

Improving Language Understanding by Generative Pre-Training

2019 – GPT2

1.5B params



Language Models are Unsupervised Multitask Learners

2020 – GPT3

175B params

Language Models are Few-Shot Learners

2024 – GPT4

>1T params

GPT-4 Technical Report

2024 – GPT5

>?? params

“Scale is all you need”?

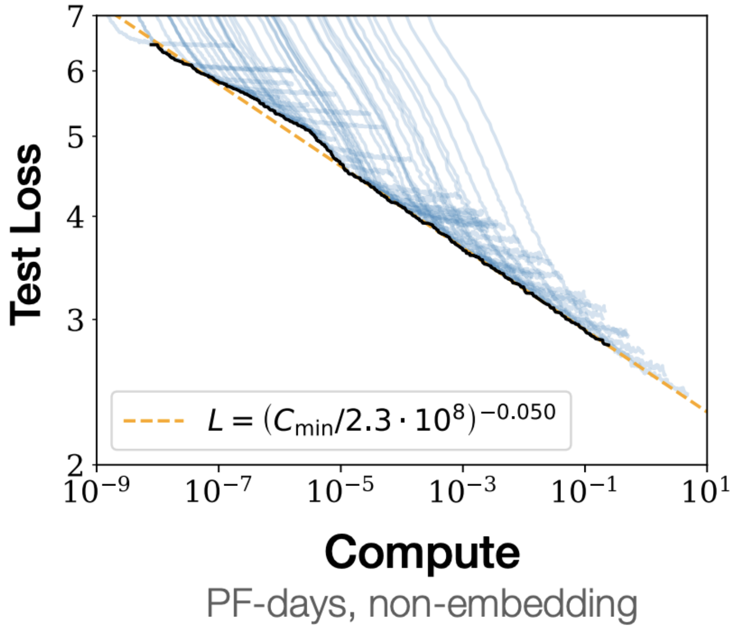


Figure 1 Language modeling performance improves smoothly as we increase the model size, datasetset size, and amount of compute[2] used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.

“Scaling Laws for Neural Language Models”. Kaplan et al 2021

“Scale is all you need”?

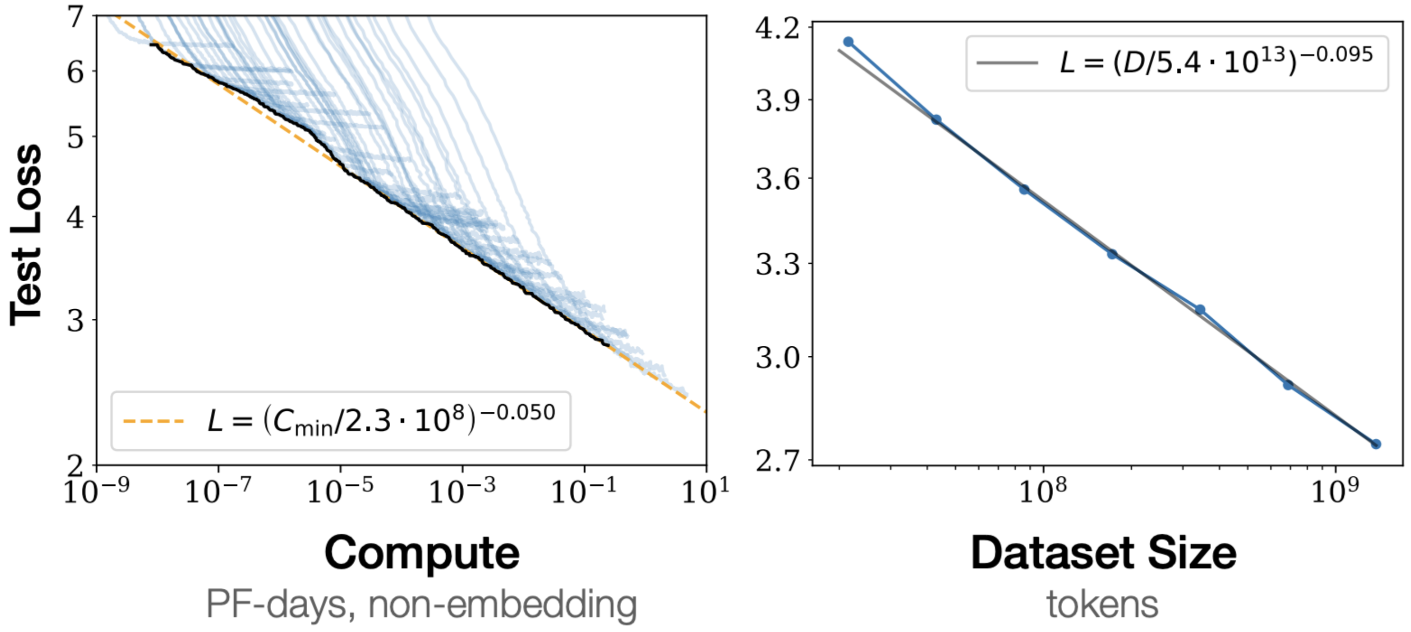


Figure 1 Language modeling performance improves smoothly as we increase the model size, datasetset size, and amount of compute[2] used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.

“Scaling Laws for Neural Language Models”. Kaplan et al 2021

“Scale is all you need”?

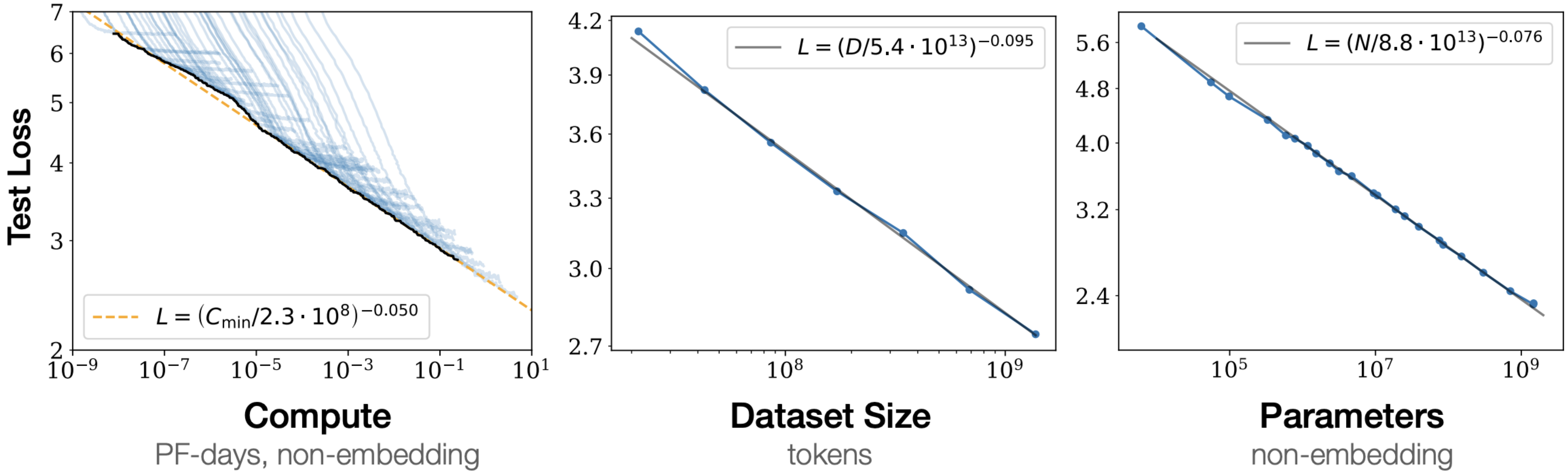


Figure 1 Language modeling performance improves smoothly as we increase the model size, datasetset size, and amount of compute[2] used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.

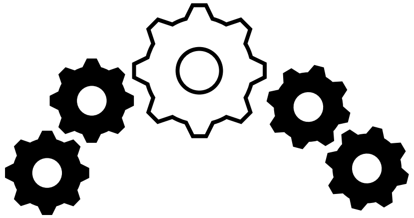
“Scaling Laws for Neural Language Models”. Kaplan et al 2021

System Design view of interpretability

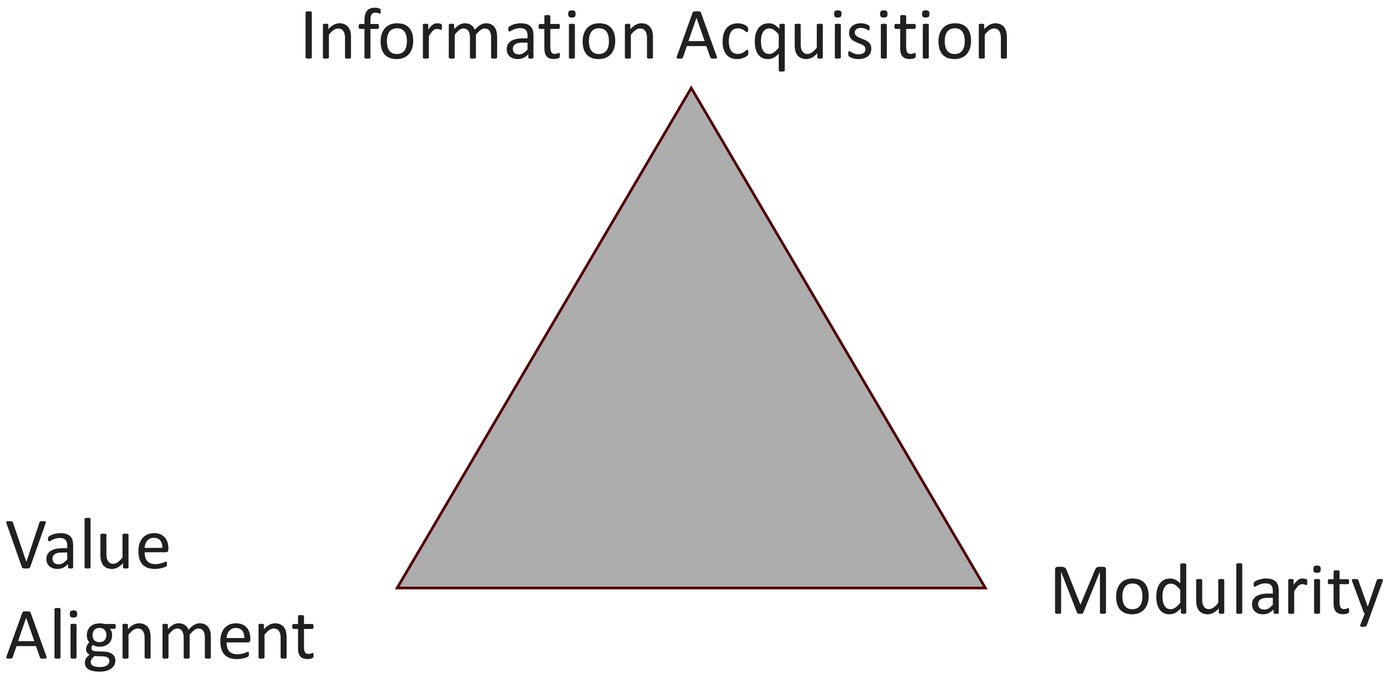
Individual vs System-Level Stats

Example: Basketball

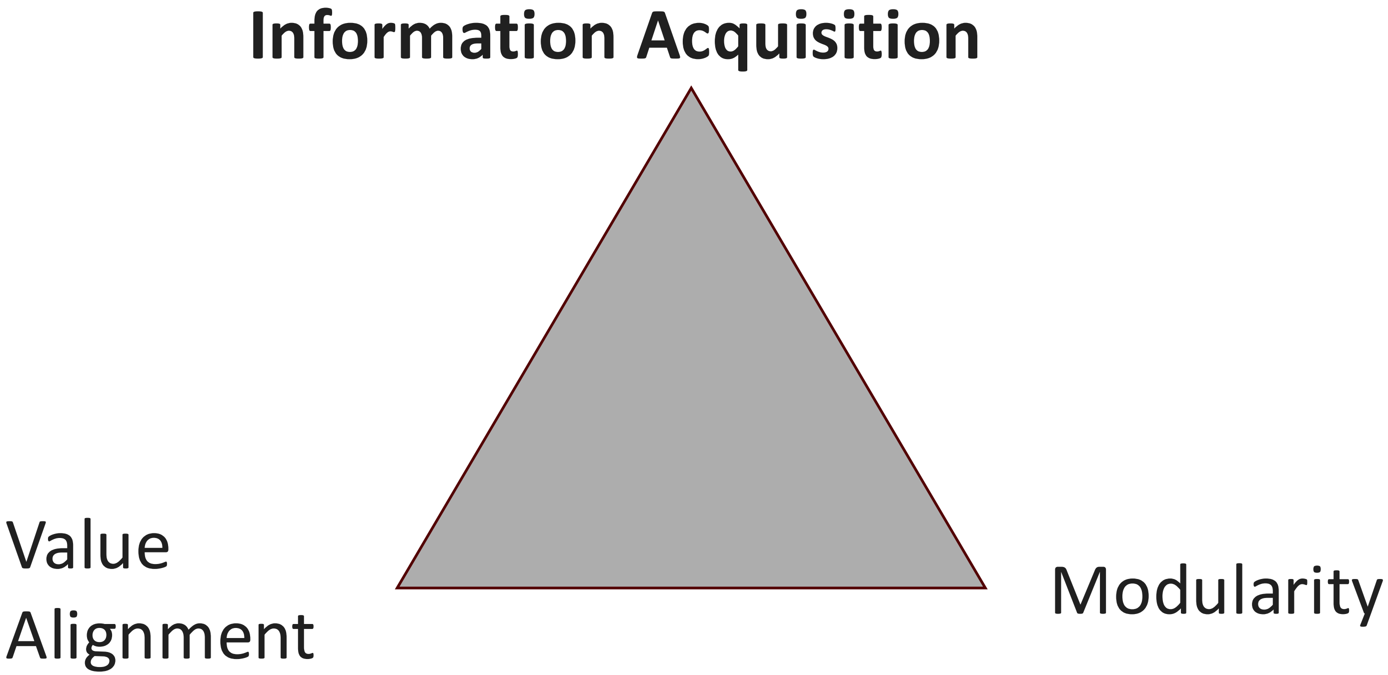
Individual Stats	System Stats
• PPG, APG, PER • Russell Westbrook 2016–17: 31.6 PPG, 10.4 APG, PER 30.6. • Historic individual success	• Net Rating = Offensive – Defensive Rating • Measured on lineups, not individuals • Correlates better with team wins



System Design benefits of interpretability

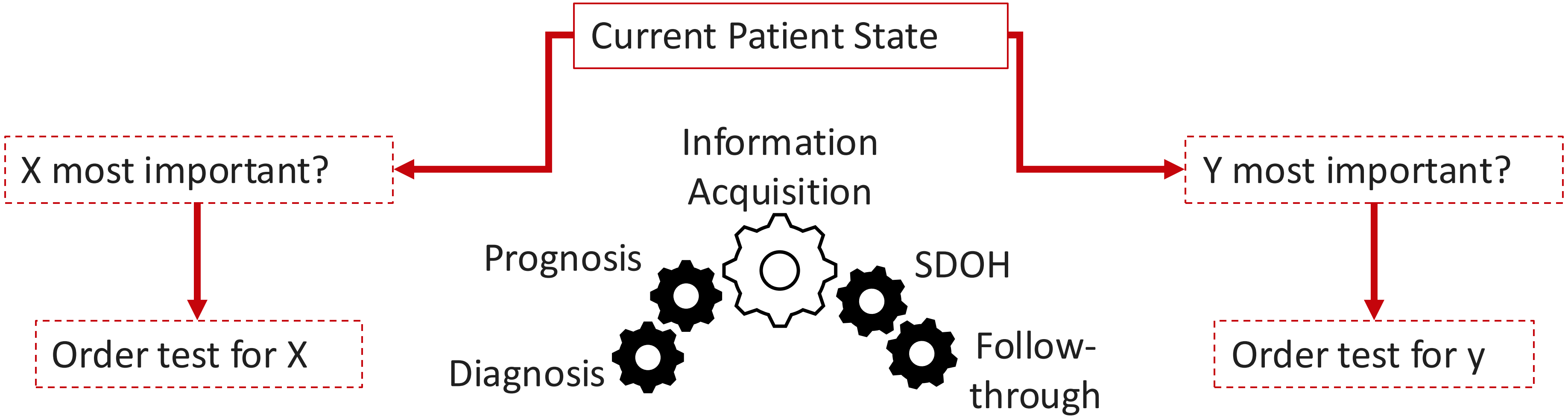


System Design benefits of interpretability



Information Acquisition: What should we measure?

- Predictive models often take measurements as fixed.
- In practice, measurement is **active** and costly.
  - Especially true in biomedicine
- Interpretability can highlight *missing but valuable* information.



Case study: Severe Maternal Morbidity (SMM)

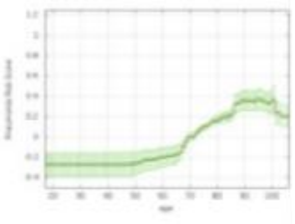
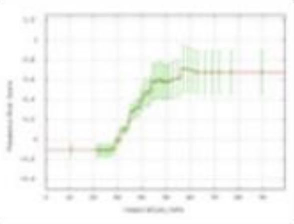
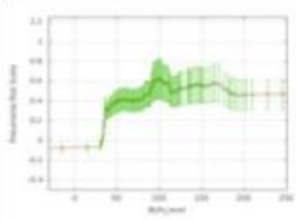
- Predict SMM via Generalized Additive Model (GAM)

[Hastie and Tibshirani (1993)]

Decompose complex outcomes into a sum of univariate functions

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

Components can be individually visualized:

$F(x) = y = \beta_0 +$  $+$  $+ \ldots +$ 

Case study: Severe Maternal Morbidity (SMM)

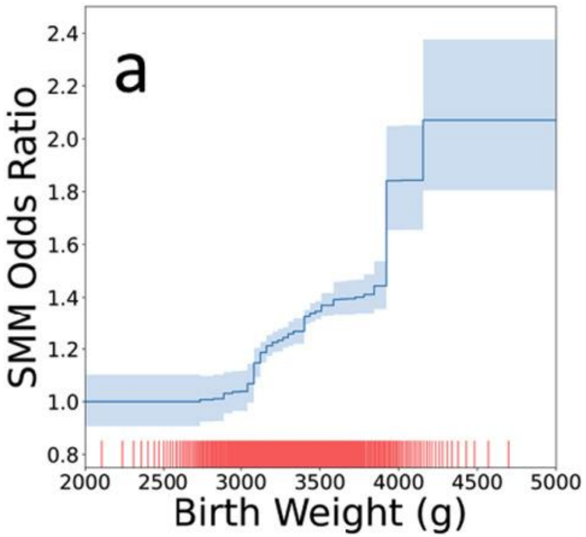


Figure 1 Generalized additive model (GAM) plots showing odds of SMM for (a) baby birthweight

Lengerich et al. *Insights into severe maternal morbidity in the NTSV population.* AJOG 2021

Case study: Severe Maternal Morbidity (SMM)

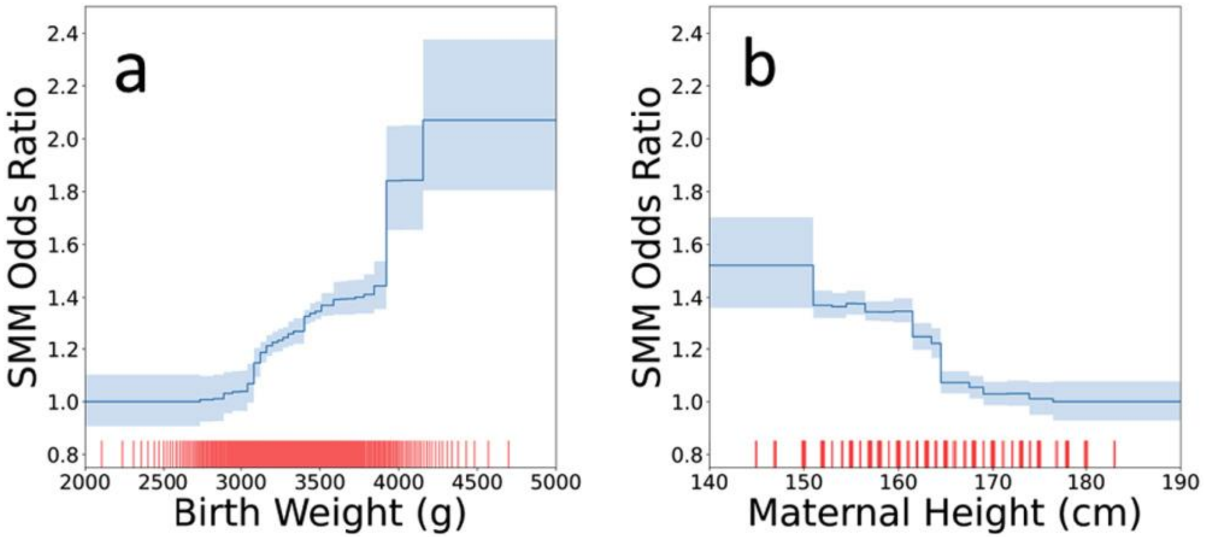


Figure 1 Generalized additive model (GAM) plots showing odds of SMM for (a) baby birthweight and (b) maternal height.

Lengerich et al. *Insights into severe maternal morbidity in the NTSV population.* AJOG 2021

Case study: Severe Maternal Morbidity (SMM)

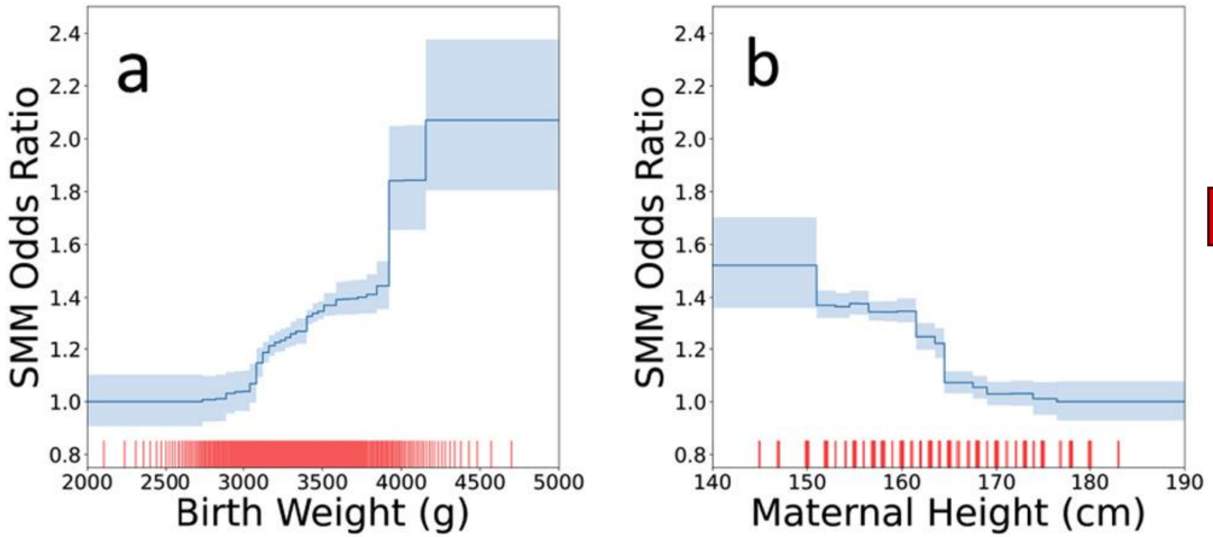
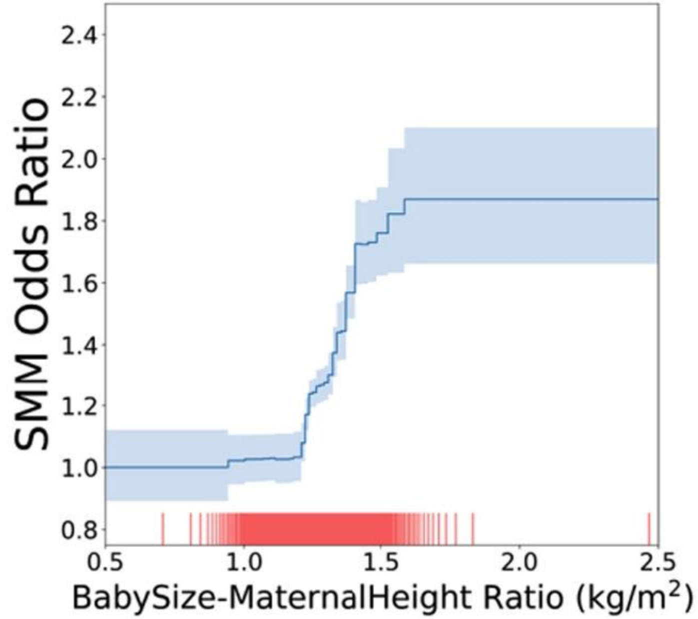


Figure 1 Generalized additive model (GAM) plots showing odds of SMM for (a) baby birthweight and (b) maternal height.

Lengerich et al. *Insights into severe maternal morbidity in the NTSV population.* AJOG 2021

BabySize-MaternalHeight Ratio (BMSR)
= birth weight (kg) / maternal height^2



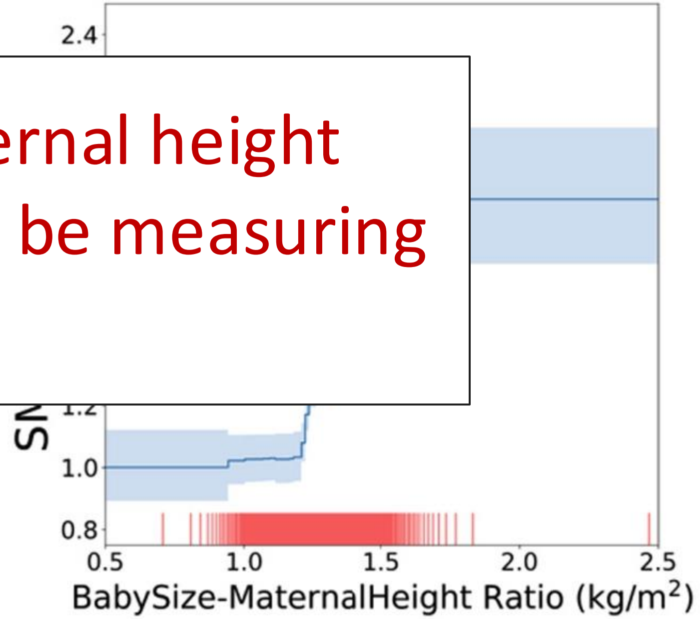
#1 Feature Importance: more than preeclampsia, etc.

Case study: Severe Maternal Morbidity (SMM)

Figure 1 Generalized additive model (GAM) plots showing odds of SMM for (a) baby birthweight and (b) maternal height.

Lengerich et al. *Insights into severe maternal morbidity in the NTSV population.* AJOG 2021

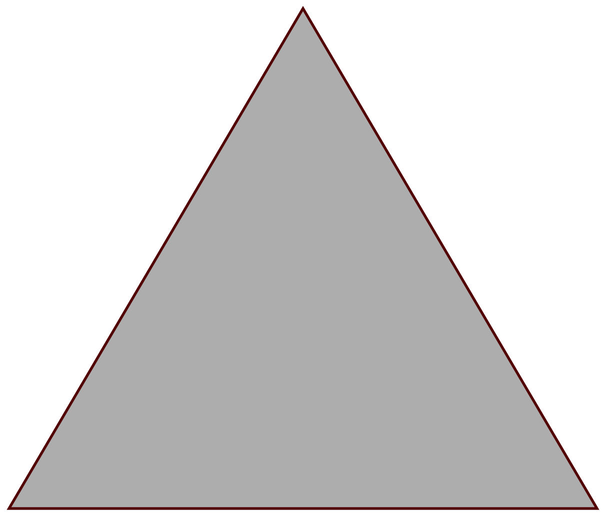
BabySize-MaternalHeight Ratio (BMSR)
= birth weight (kg) / maternal height^2



We just happened to have maternal height available; what should we actually be measuring clinically?

#1 Feature Importance: more than preeclampsia, etc.

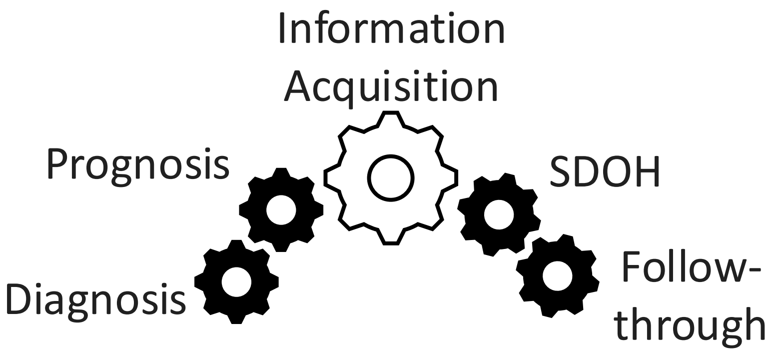
System Design benefits of interpretability



Modularity: Swappable, testable Components

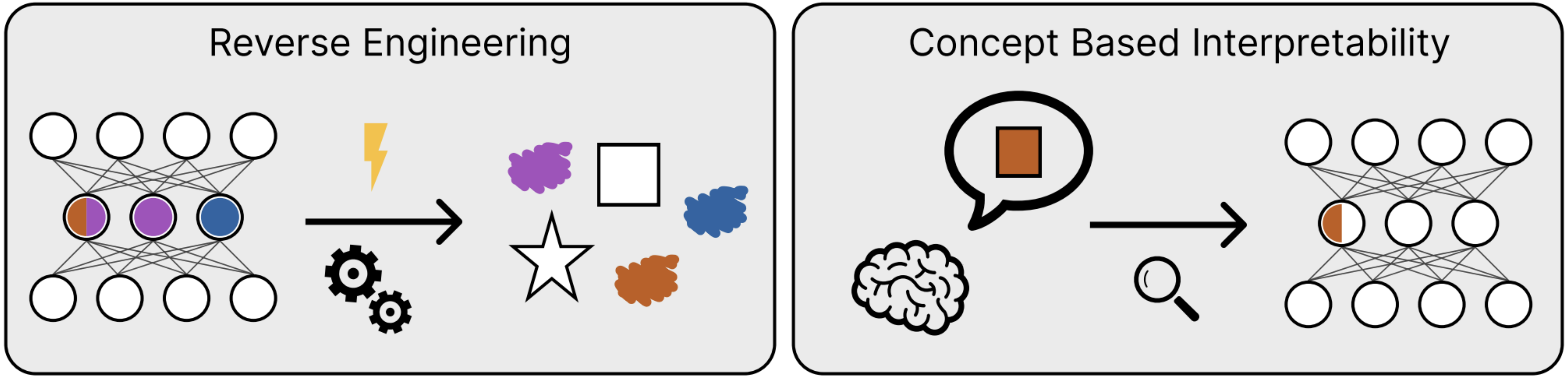
- Interpretability *connects component-level performance to system-level performance*
- Each component has a ”job”

  → new versions can be tested and adopted



Modularity: Swappable, testable Components

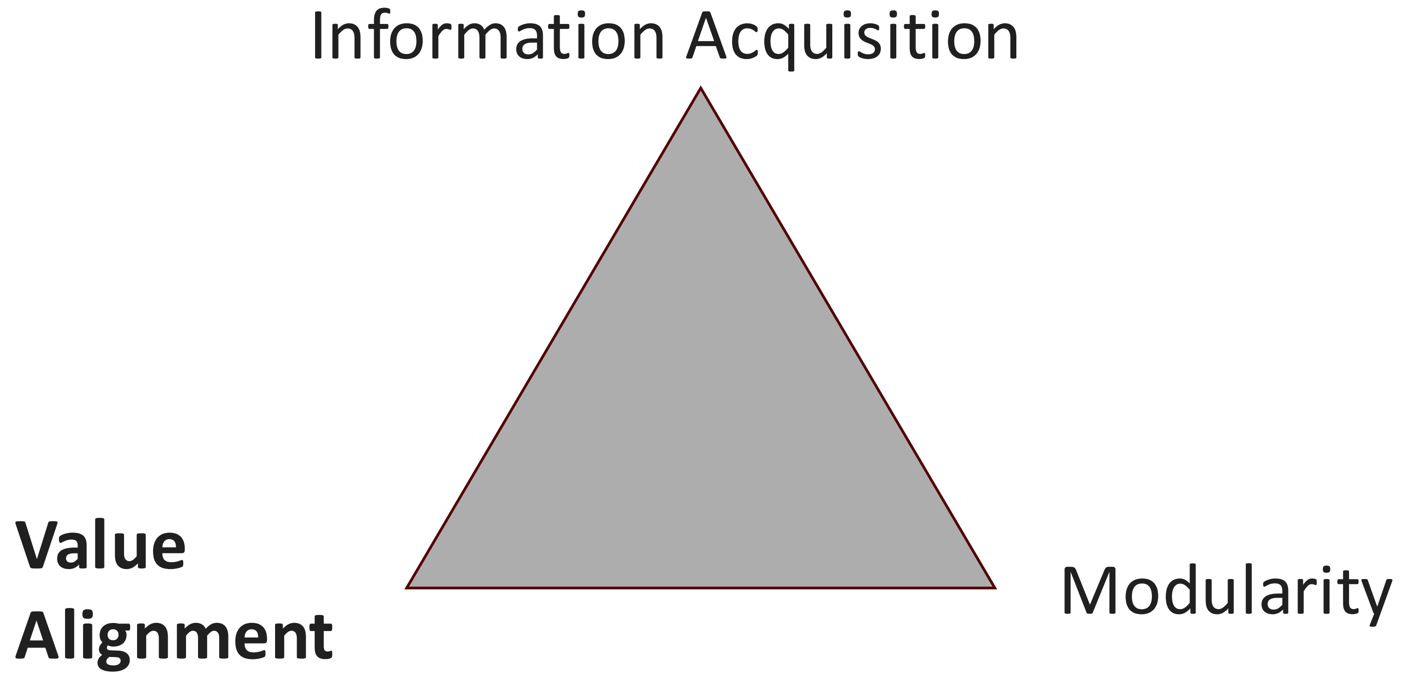
- Currently: extract components from trained models



https://arxiv.org/pdf/2501.16496

- Could we incorporate these components from the start of model training?

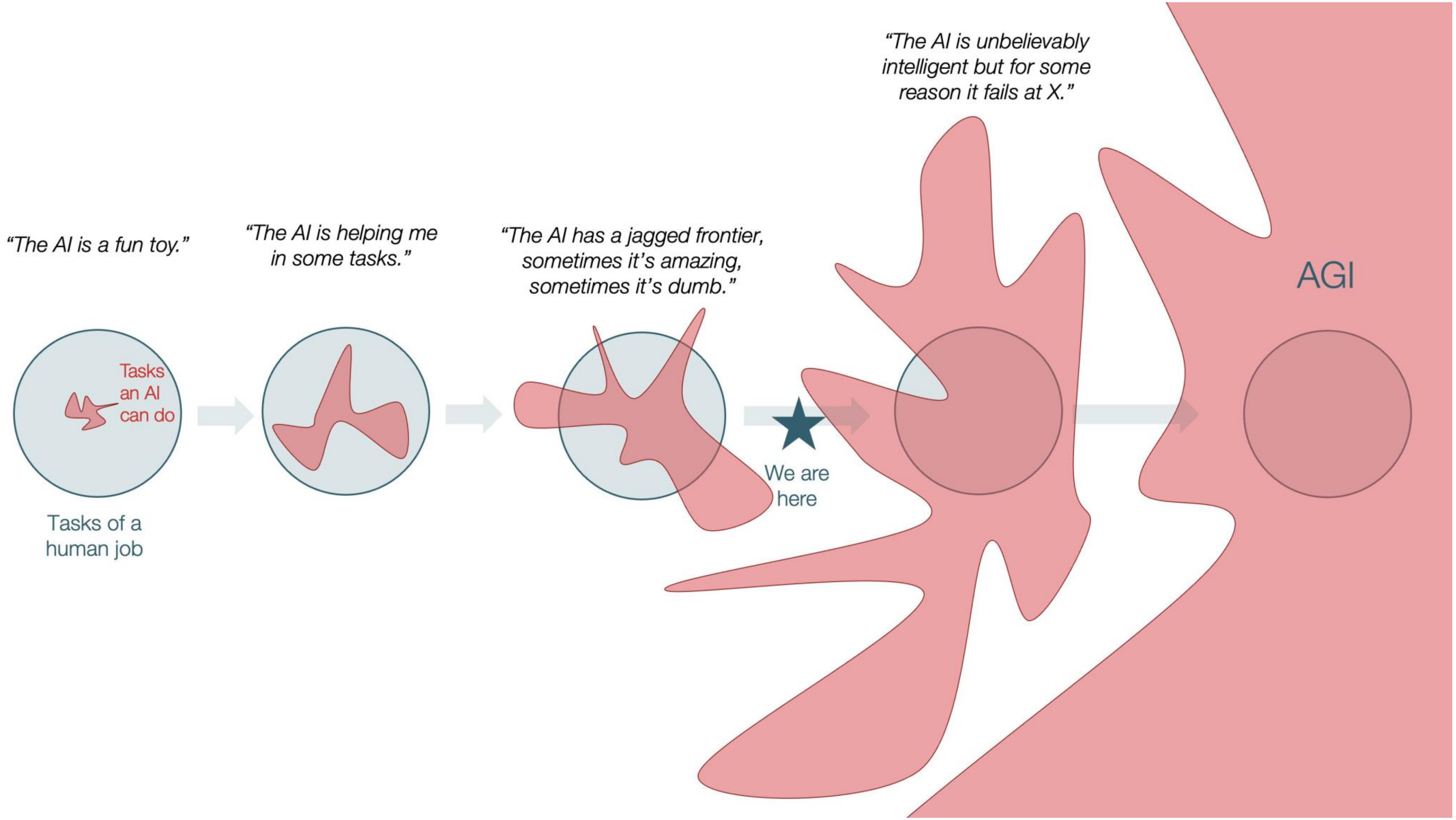
System Design benefits of interpretability



Alignment: What did the model learn to optimize?

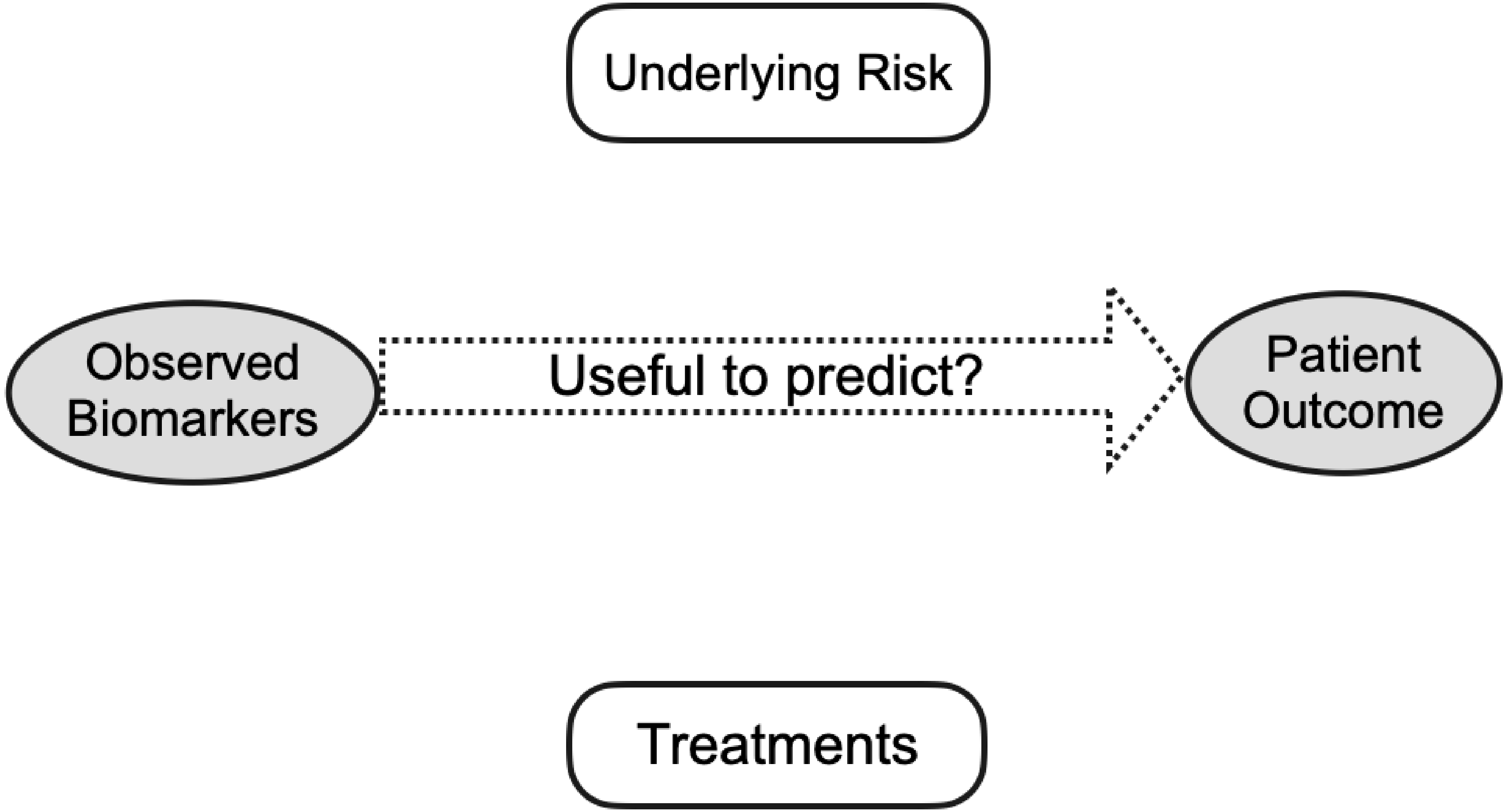
- Connect probabilistic objectives to value-based objectives
- Outer vs inner alignment:
  - **Outer alignment:** Is the loss function we train on actually aligned with human goals?
  - **Inner alignment:** Given that loss, does the trained model's internal representation faithfully implement that goal, even off-distribution?

Jagged Performance = Mis-alignment?

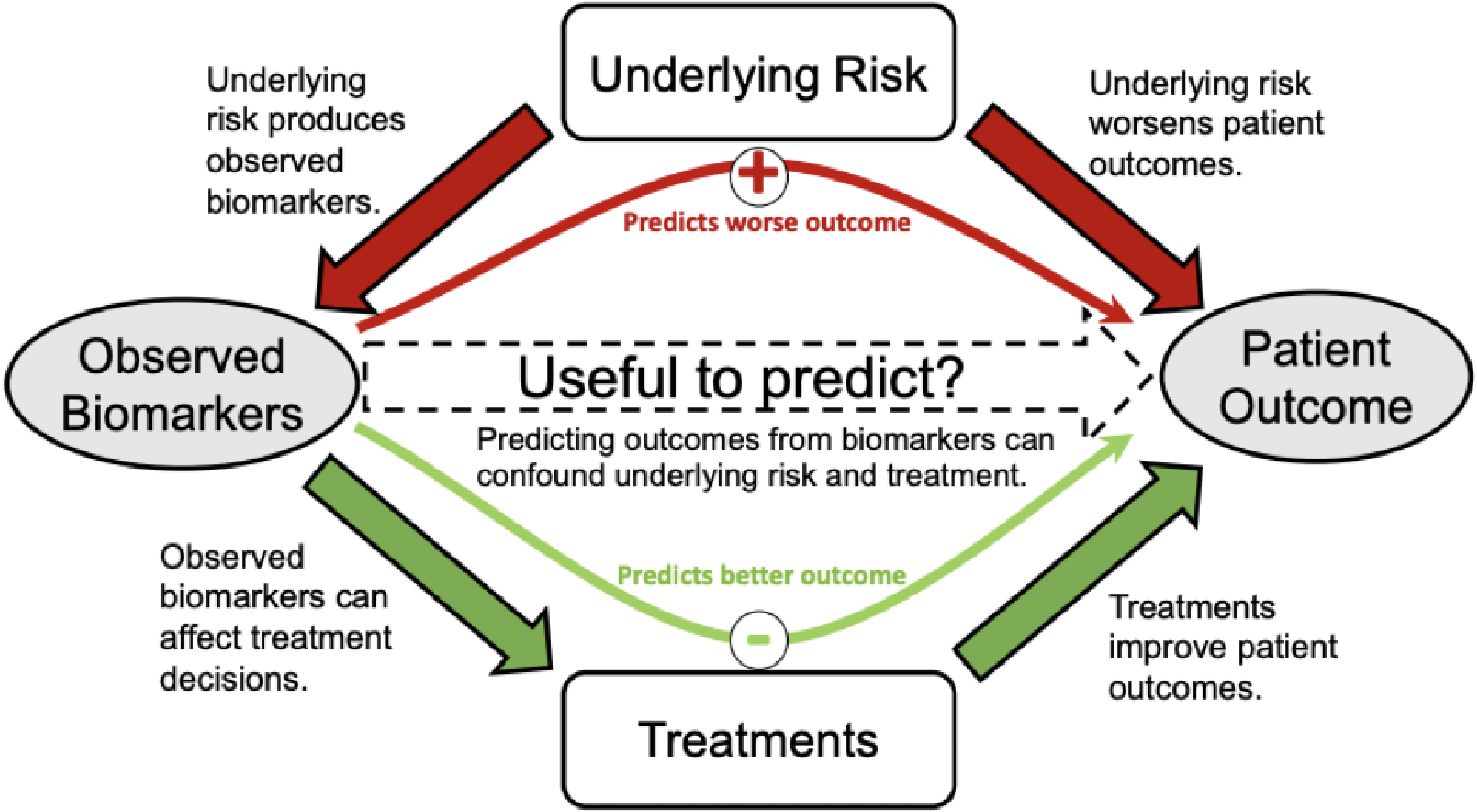


https://x.com/tomaspueyo/status/1993360931267473662

Example: Medicine

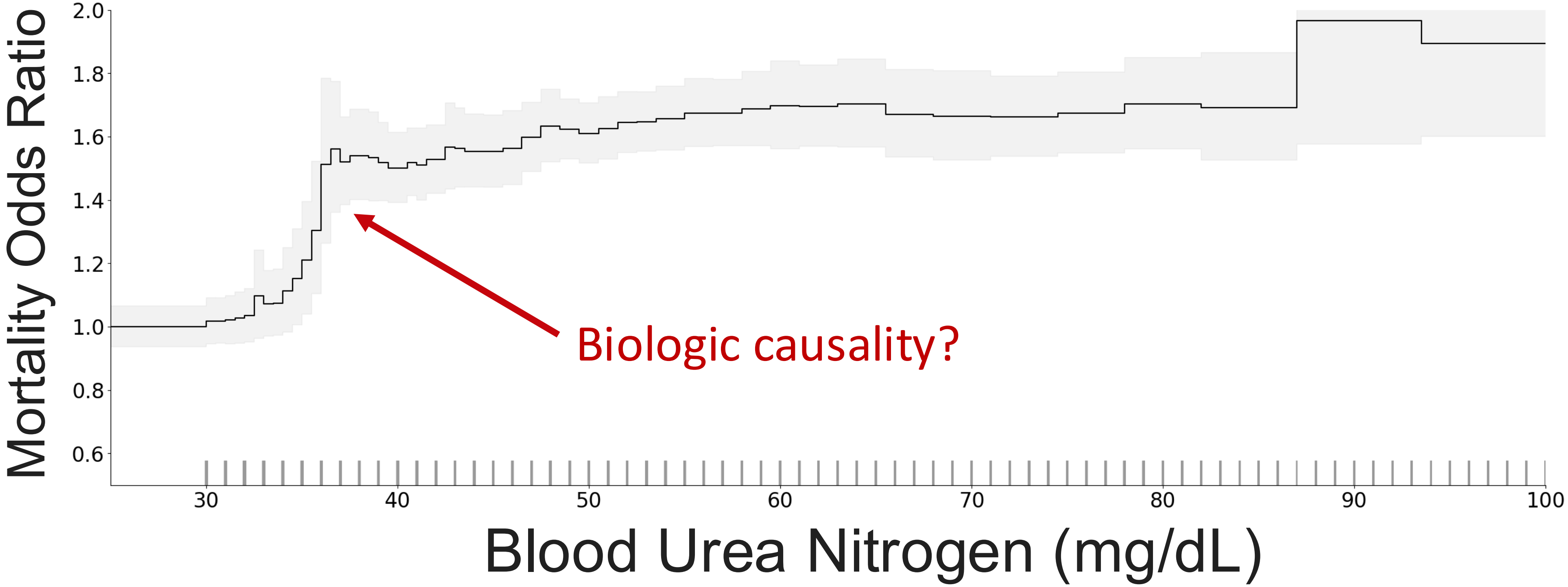


Example: Medicine



Real-world effects are surprising and may not be causal

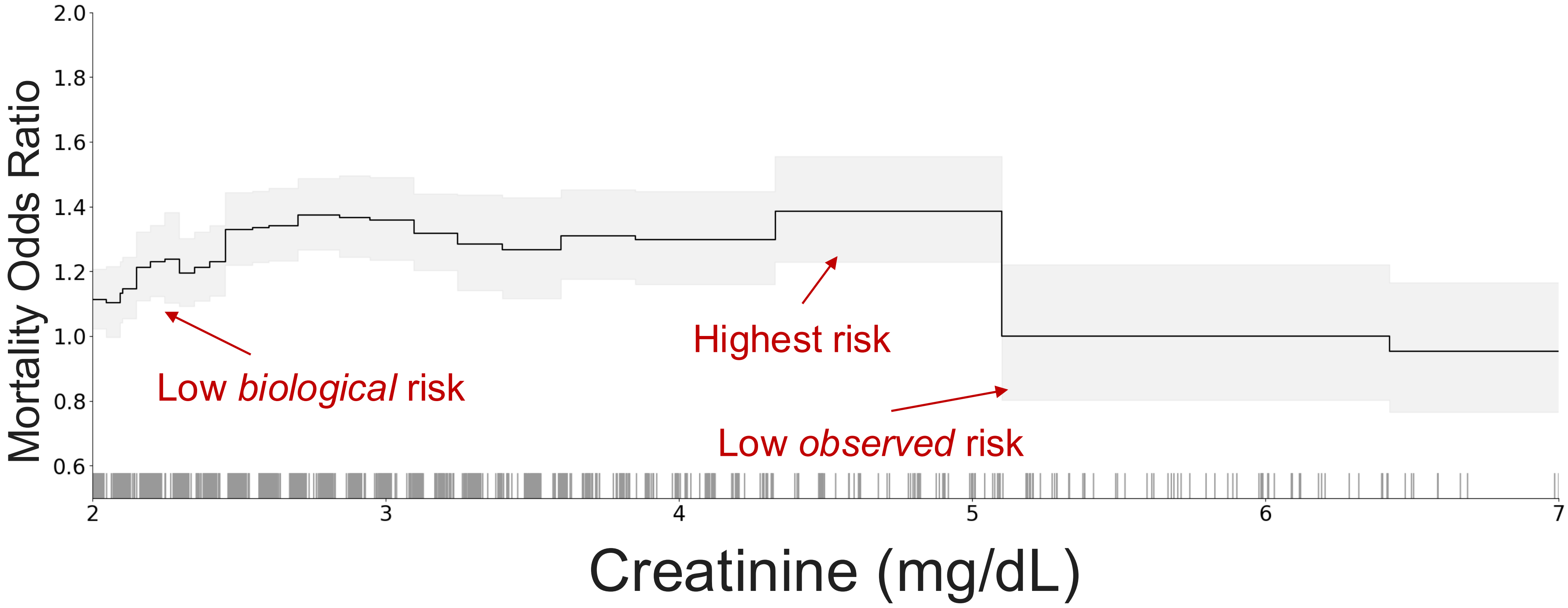
In-hospital mortality risk for hospitalized patients with pneumonia:



Lengerich et al. *The Hidden Cost of Round Numbers and Sharp Thresholds in Clinical Practice.* NPJ Digital Medicine 2025

Real-world effects are surprising and may not be causal

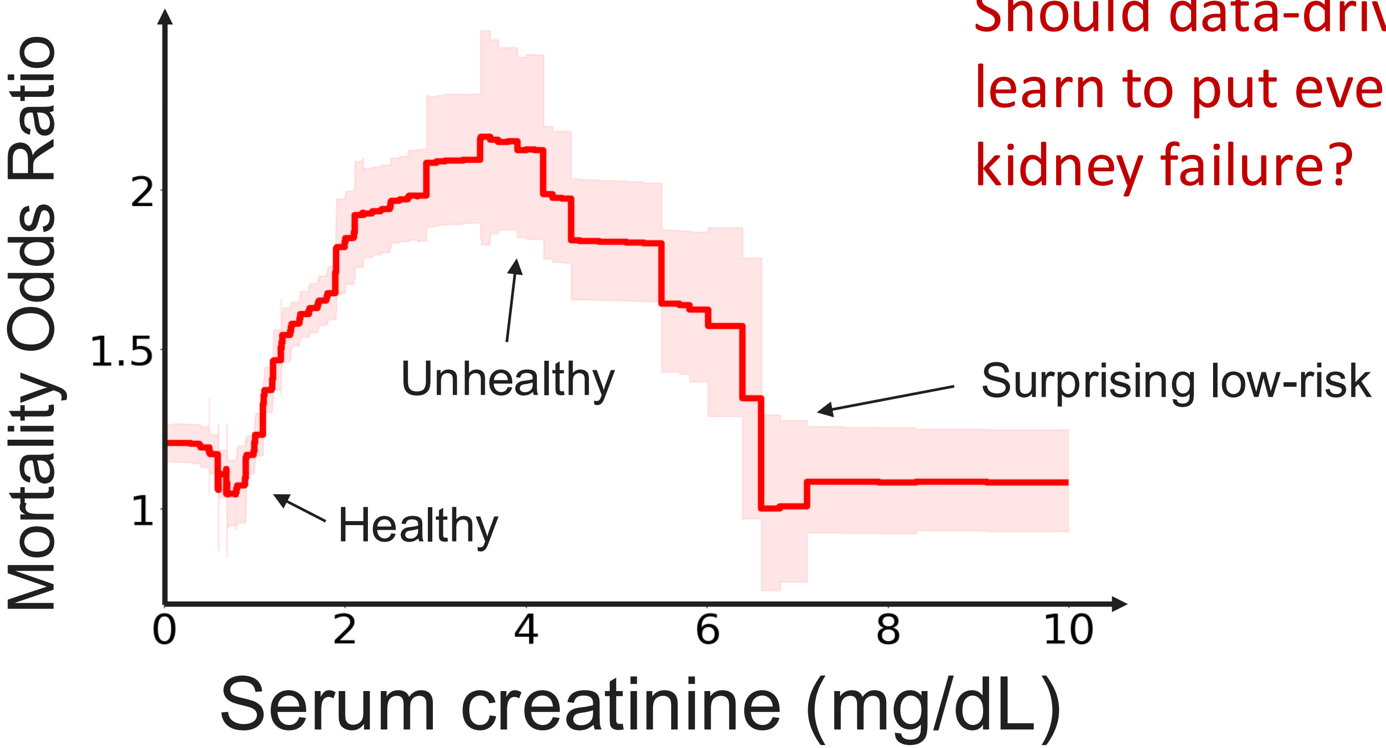
In-hospital mortality risk for hospitalized patients with pneumonia:



Lengerich et al. *The Hidden Cost of Round Numbers and Sharp Thresholds in Clinical Practice.* NPJ Digital Medicine 2025

Real-world effects are surprising and may not be causal

MIMIC-IV mortality risk for hospitalized patients:



Should data-driven systems learn to put everyone into kidney failure?

Lengerich et al. *The Hidden Cost of Round Numbers and Sharp Thresholds in Clinical Practice.* NPJ Digital Medicine 2025

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.

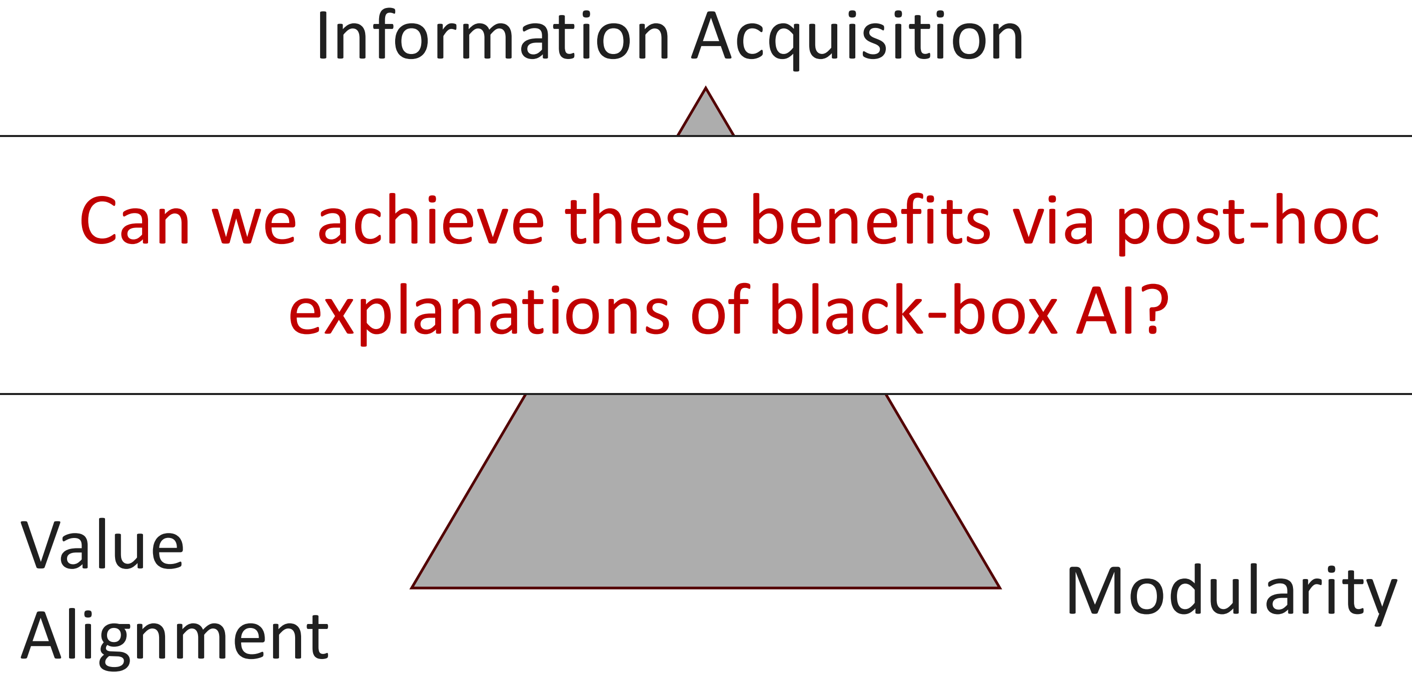
What should we do?

- Two paths:
  - Correct for the complications at training time
  - Extract and correct for the complications after training
- Key Point:
  - Ignoring complicated features does not remove their effects from the trained model.
    - Correlations and associations make all kinds of effects still show up in the trained model.

Example: Training to be invariant to “race”

- Suppose we have a dataset that contains a “race” feature and we want our trained model to be invariant to “race”. What should we do?
- ~~Remove “race” from training and assume the model ignores those effects.~~
- Train on all features including “race” and then:
  - (Maybe) Remove the learned component associated with “race”
  - (Maybe) Drop the “race” feature at test-time
  - (Maybe) Train with a modified loss function to encourage invariant predictions
  - (Maybe)

Today: System Design benefits of interpretability



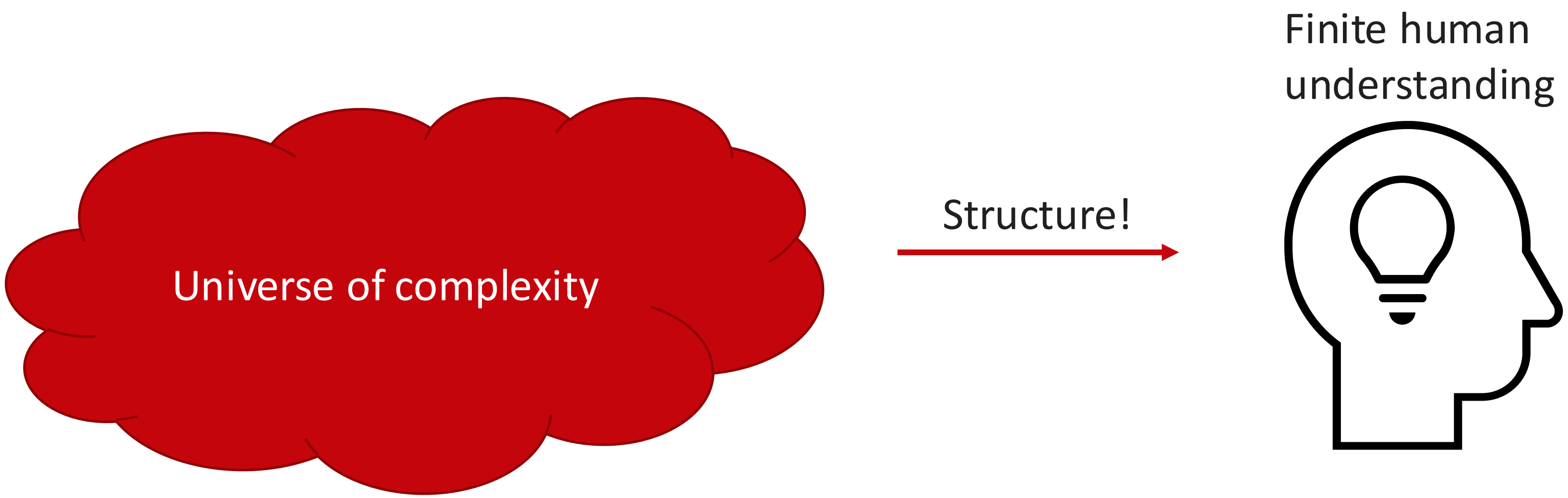
Can we achieve these benefits via post-hoc explanations of black-box AI?

Open Problems

https://www.youtube.com/watch?v=aR20FWCCjAs

What’s the point of statistics anymore?

- A language for communication
- A language for computation
- A language for development



Some open problems from Ilya

- Models show impressive eval performance but lack real-world economic impact and exhibit jaggedness, like repeating bugs in coding tasks.
- Human emotions serve as robust value functions? Current AI lacks similar mechanisms.
- Pre-training scales uniformly but hits data walls; RL consumes more compute but needs better efficiency via value functions.
- Humans generalize better than models with fewer samples and unsupervised learning.
- Alignment involves designing AI to care for sentient life, including AIs, for broader empathy over human-centric values?

Some open problems from Ilya

- Models show impressive eval performance but lack real-world economic impact and exhibit jaggedness, like repeating bugs in coding tasks.
- H
  si
- Pr
  compute but needs better efficiency via value functions.
- Humans generalize better than models with fewer samples and unsupervised learning.
- Alignment involves designing AI to care for sentient life, including AIs, for broader empathy over human-centric values?

You all now have the tools and vocabulary to discuss SOTA research that is worth billions of $.

More open problems

- RL (how to effectively train at scale with distant reward signals)
- Scaling verifiable rewards
- Combining LLMs with symbolic reasoning
- Combining LLMs with graphical models
- Continual learning
- Formal theory of alignment.
- Post-hoc interpretability of large models.
- Ante-hoc interpretable-by-design large models.
- Ethical and technical fusion: aligning not just models, but the human-model system.

Questions?

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON