Probabilistic Graphical Models & Probabilistic AI

Ben Lengerich

Lecture 23: Open Directions in Graphical Models

April 24, 2025

Reading: See course homepage

Today

- Semester Review
- Open Directions in Graphical Models
  - Context-Adaptive Models
  - Connecting LLMs to Graphical Models

Graphical Models

Why GMs?

What’s the point of GMs in the AI era?

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

The Fundamental Questions

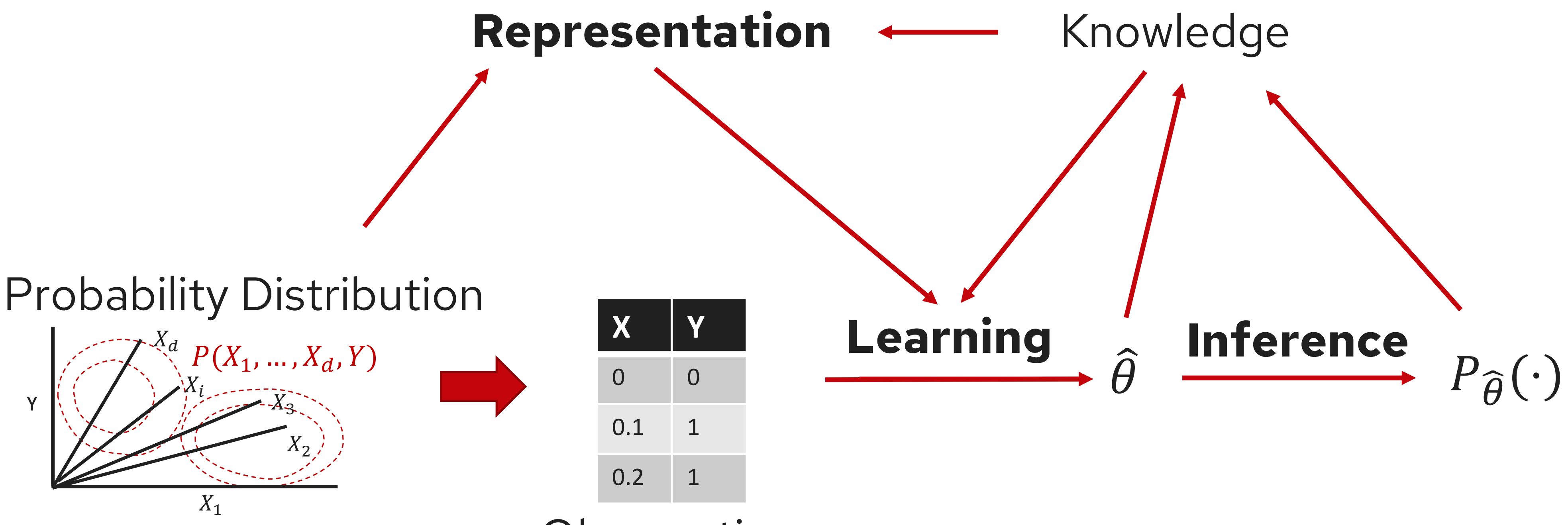
- **Representation**
  - How to encode our domain knowledge/assumptions/constraints?
  - How to capture/model uncertainties in possible worlds?
- **Inference**
  - How do I answer questions/queries according to my model and/or based on observed data?

$$\text{e.g. } P(X_i|D)$$

- **Learning**
  - What model is "right" for my data?

$$\text{e.g. } M = argmax_{M\in\mathcal{H}} F(D;M)$$

A Simplified View of our Roadmap



X	Y
0	0
0.1	1
0.2	1

Observations

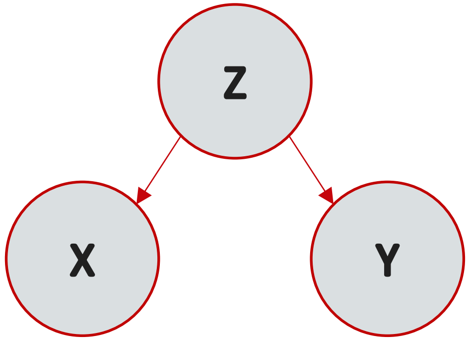
PGMs allow us to understand and structure data

- GM = Multivariate Objective Function + Structure
- PGM = Multivariate Statistics + Structure

- Formally: A PGM is a **family of distributions** on a set of random variables that are compatible with all the probabilistic independence propositions encoded by a **graph** that connects these variables.

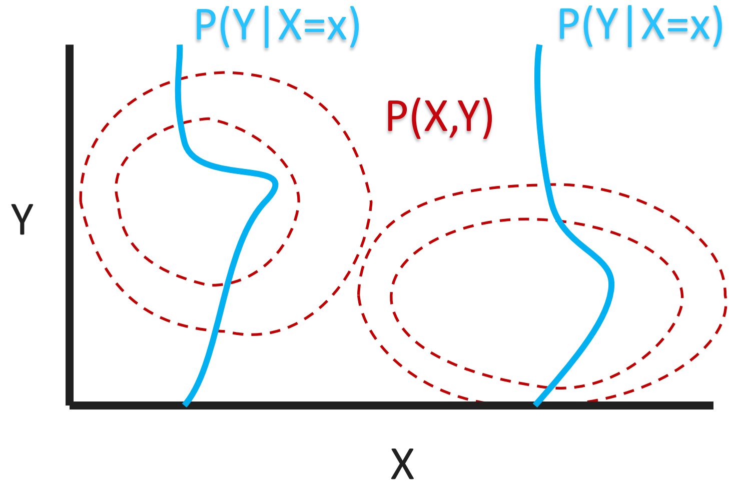
Conditional Independence

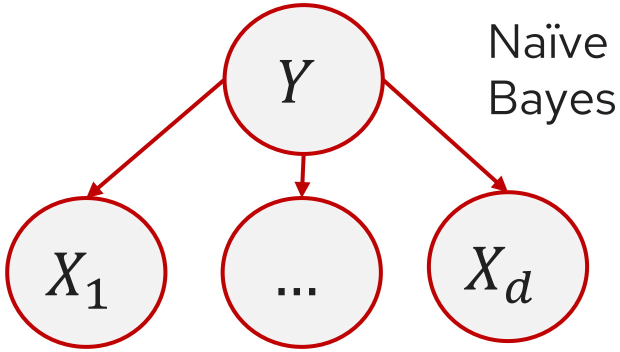
- Variables X and Y are **independent** if:
$$P(X,Y) = P(X)P(Y)$$
  - Notation: $X \perp Y$
- Variables X and Y are **conditionally independent given Z** if:
$$P(X,Y|Z) = P(X|Z)P(Y|Z)$$
  - Equivalently: $P(X|Y,Z) = P(X,Z)$
  - Notation: $X \perp Y \mid Z$



Structure Encodes Assumptions

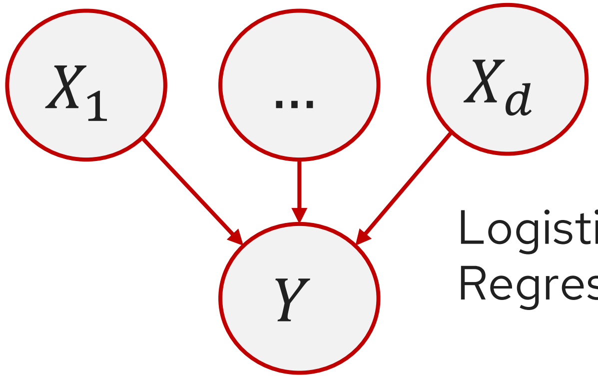
- **Generative**:
  - Models the joint distribution $\mathrm{P}(X, Y)$.
- **Discriminative**:
  - Models the conditional distribution $\mathrm{P}(Y|X)$.





Naïve Bayes

$$P(X,Y) = P(Y)\prod_i P(X_i \mid Y)$$



Logistic Regression

$$P(X,Y) = P(Y \mid X)\prod_i P(X_i)$$

Bayesian Networks (BN)

- A BN is a **directed acyclic graph** whose nodes represent the random variables and whose edges represent direct influence of one variable on another
- Provides the skeleton for representing a joint distribution compactly in a **factorized** way
- Compact representation of a set of **conditional independence** assumptions
- We can view the graph as encoding a **generative sampling process** executed by nature.

Markov Random Fields (MRFs)

- An ***undirected graphical model*** represents a distribution $P(X)$ defined by an undirected graph H and a set of positive ***potential functions*** ψ associated with the cliques of H such that:

$$\mathrm{P}(\mathrm{X}_1, \dots, \mathrm{X}_\mathrm{n}) = \frac{1}{\mathrm{Z}} \prod_{c} \psi_c(X_C)$$

where Z represents the **partition function**: $Z = \sum_X \prod_c \psi_c(X_c)$.

- The potential function can be understood as a "score" of the joint configuration

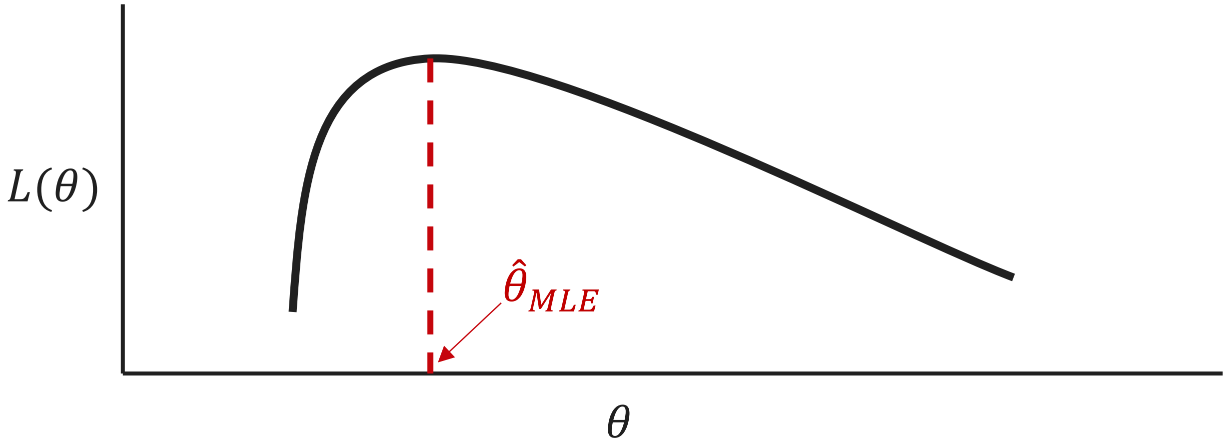
Learning

Maximum Likelihood Estimation (MLE)

- **Definition:**
  - Find $\widehat{\boldsymbol{\theta}}$ that maximizes the likelihood of observing the given data.

$$\widehat{\boldsymbol{\theta}} = \text{argmax}_{\theta} L(\theta) \text{ where } L(\theta) = P(\text{data}|\theta).$$

- **Interpretation:**
  - L(θ): Probability of the observed data given θ.
  - MLE chooses the parameter that makes the data most "likely."



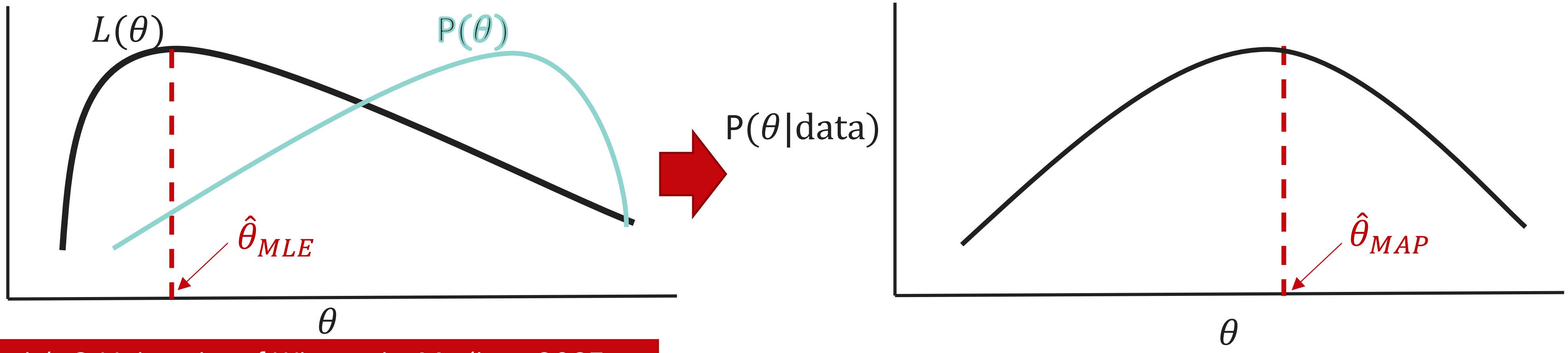
Maximum A Posteriori (MAP) Estimation

- Find

$$\hat{\theta}_{MAP} = argmax_{\theta} P(\theta|\text{data}) \propto argmax_{\theta} P(\text{data}|\theta)P(\theta)$$

- $P(\text{data}|\theta)$: Likelihood
- $P(\theta)$: Prior belief about θ

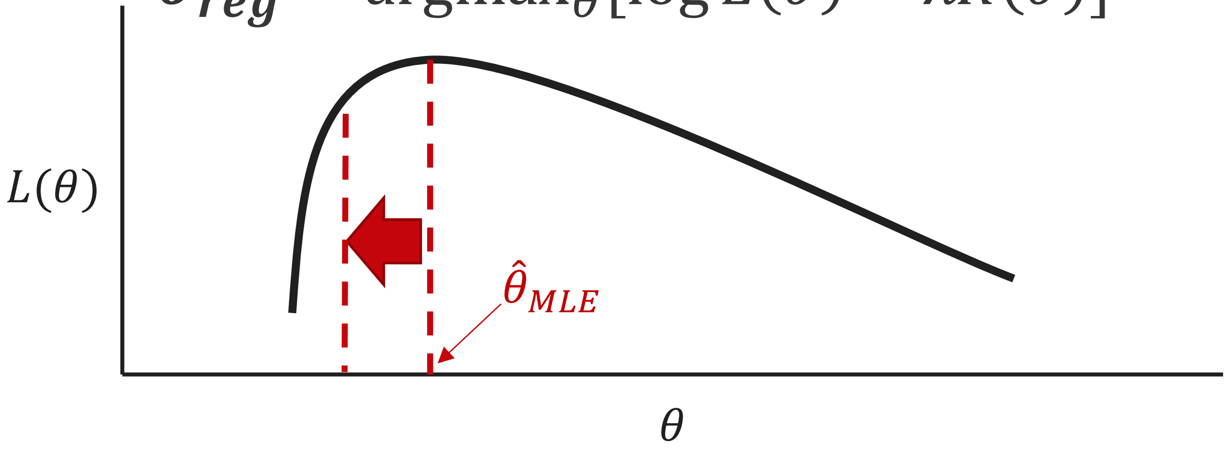
- MLE ignores $P(\theta)$
- MAP incorporates prior information.



Regularization is MAP

- **MLE with Regularization:**
  - Adds a penalty to avoid overfitting

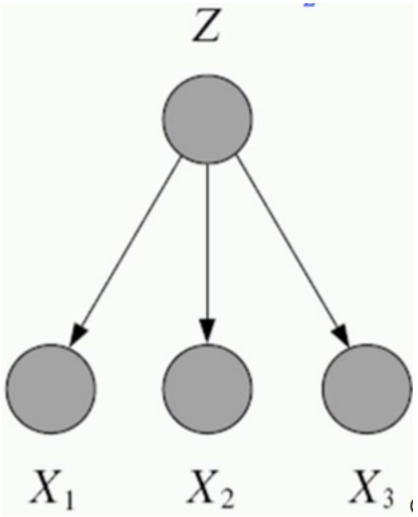
$$\widehat{\boldsymbol{\theta}_{\boldsymbol{reg}}} = \mathrm{argmax}_{\theta}[\log L(\theta) - \lambda R(\theta)]$$



- **MAP as Penalized MLE:**
  - Let $P(\theta) \propto e^{-\lambda R(\theta)}$. Then

$$\hat{\boldsymbol{\theta}}_{MAP} = argmax_{\theta}[\log L(\theta) + \log P(\theta)] = \hat{\theta}_{reg}$$

Why is learning with latent variables harder?



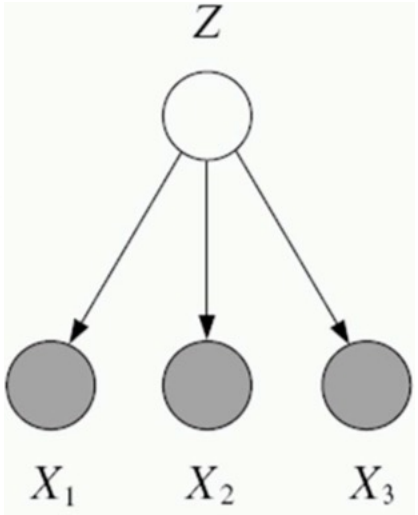
- In fully-observed IID settings, the log-likelihood decomposes into a sum of local terms:

$$\ell_c(\theta; D) = \log p(x, z \mid \theta) = \log p(z \mid \theta_z) + \log p(x \mid z, \theta_x)$$

- With latent variables, all parameters become coupled via marginalization

$$\ell_c(\theta; D) = \log \sum_z p(x, z \mid \theta) = \log \sum_z p(z \mid \theta_z) p(x \mid z, \theta_x)$$

Sum over z is inside log



Solution 1 to LV learning: Expectation-Maximization

- *"Guess a value for the LVs, then update it."*

- **E-step**:
  - Compute the expected value of the sufficient statistics of the hidden variables under current estimates of parameters

- **M-step**:
  - Using the current expected value of the hidden variables, compute the parameters that maximize the likelihood.

Solution 2 to LV learning: Variational Inference

- *"Maximize an easier lower-bound of the log-likelihood."*

$$\log p(x \mid \theta) \geq \underbrace{E_{z\sim q}[\log p(x, z \mid \theta)] + H(q)}_{\text{"ELBO": Evidence Lower Bound}} \quad +KL(q(z \mid x) \mid\mid p(z \mid x, \theta))$$

- We choose a family of variational distributions (i.e., a parameterization of a distribution of the latent variables) such that the expectations are computable.
- Then, we **maximize the ELBO** to find the parameters that gives as tight a bound as possible on the marginal probability of x.

Solution 3 to LV learning: Monte Carlo

- *"Define a distribution by drawing samples instead of a closed-form."*

- Draw random samples from desired distribution
- Yield a stochastic representation of desired distribution
  - $E_p[f(x)] = \frac{\sum_m f(X_m)}{|m|}$
- **Asymptotically exact**
- Challenges:
  - How to draw samples from desired distribution?
  - How to know we've sampled enough?

Solution 4 to LV learning: Deep Learning

- *"Define the likelihood of latent variables as delta functions."*

- Define our probabilistic model such that

$$p(z \mid x; \theta) = \delta\big(z - f(x; \theta)\big), \text{ i.e. } z = f(x; \theta),$$

- Then

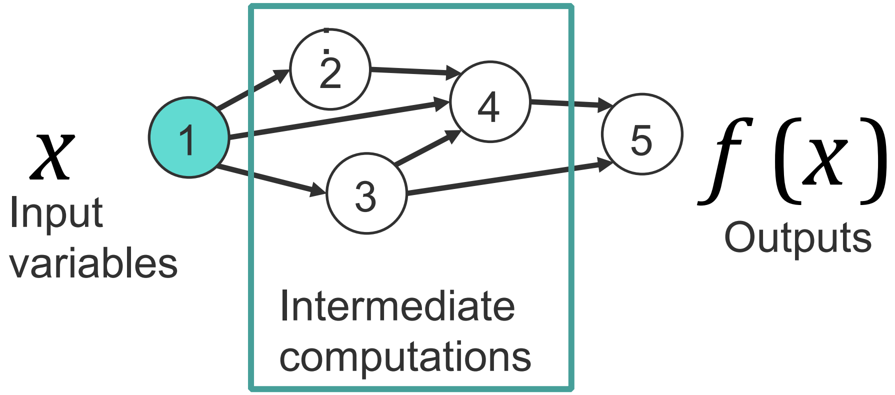
$$p(y \mid x; \theta) = p(y \mid f(x; \theta))$$

- By properly defining f with convenient activation functions (like ReLU or sigmoid), then $\hat{\theta}_{MLE}$ can be estimated by backpropagating error on y.

Deep Learning

Deep Learning via Backpropagation

- Neural networks are function compositions that can be represented as computation graphs:

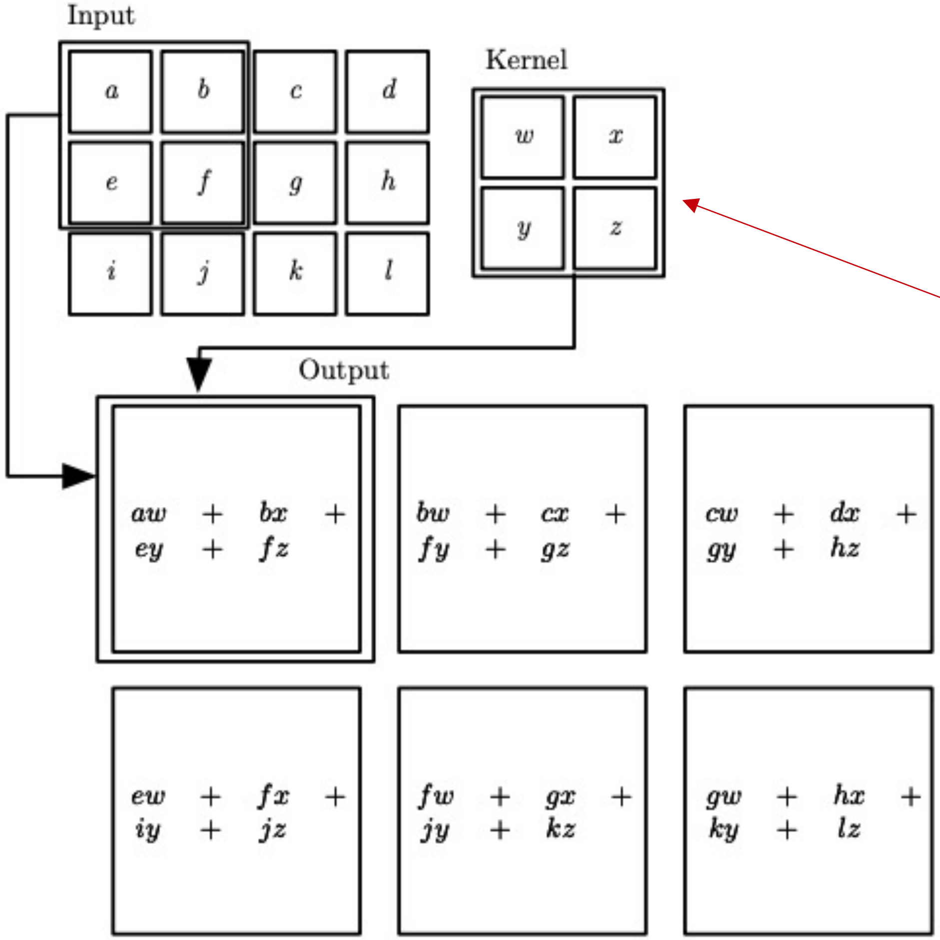


$$\frac{\partial f_n}{\partial x} =$$

- By applying the chain rule, and working in reverse order, we get:

$$\frac{\partial f_n}{\partial x} = \sum_{i_1 \in \pi(n)} \frac{\partial f_n}{\partial f_{i_1}} \frac{\partial f_{i_1}}{\partial x} = \sum_{i_1 \in \pi(n)} \frac{\partial f_n}{\partial f_{i_1}} \sum_{i_2 \in \pi(i_1)} \frac{\partial f_{i_1}}{\partial f_{i_2}} \frac{\partial f_{i_1}}{\partial x} = \dots$$

Convolutional Neural Networks [LeCun 1989]

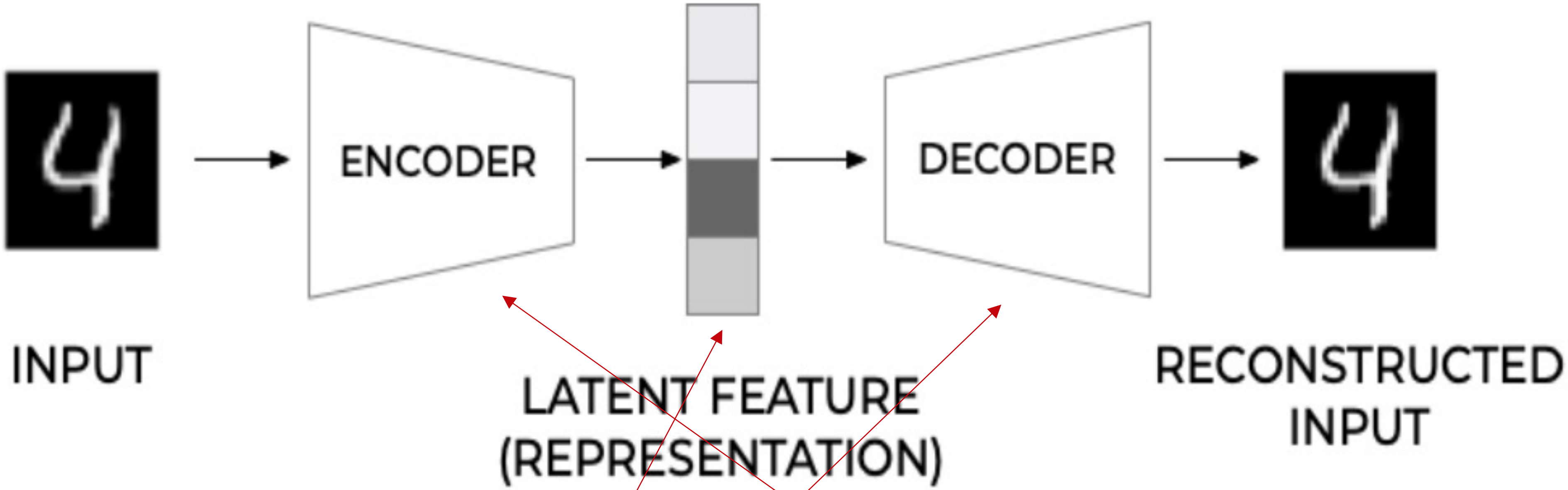


Sliding filters (kernels)

Reused weights (small)!

Fig. Goodfellow et al. 2016

Autoencoders

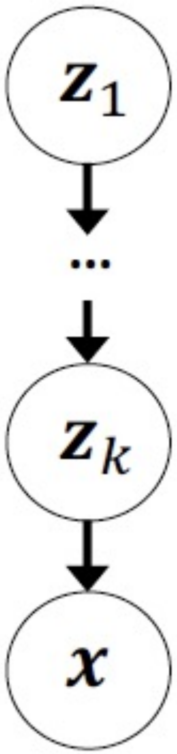


$$\tilde{x} = f(h) = f\big(g(x)\big)$$

[Michelucci 2022]

Deep Generative Models

- Define probabilistic distributions overs a set of variables
- “Deep” means multiple layers of hidden variables!
- Many forms:
  - Variational Autoencoders
  - GANs
  - Diffusion Models



The "Transformer"

- **Original Transformer** (Vaswani et al., 2017):
  - Encoder-decoder architecture for sequence-to-sequence tasks
  - Parallelizable self-attention instead of recurrence
  - Positional encodings enable order sensitivity
- **Encoder**: Processes input sequence
- **Decoder**: Generates output sequence using masked attention + encoder output
- Inspired by machine translation (observe full input sequence, predict full output sequence)

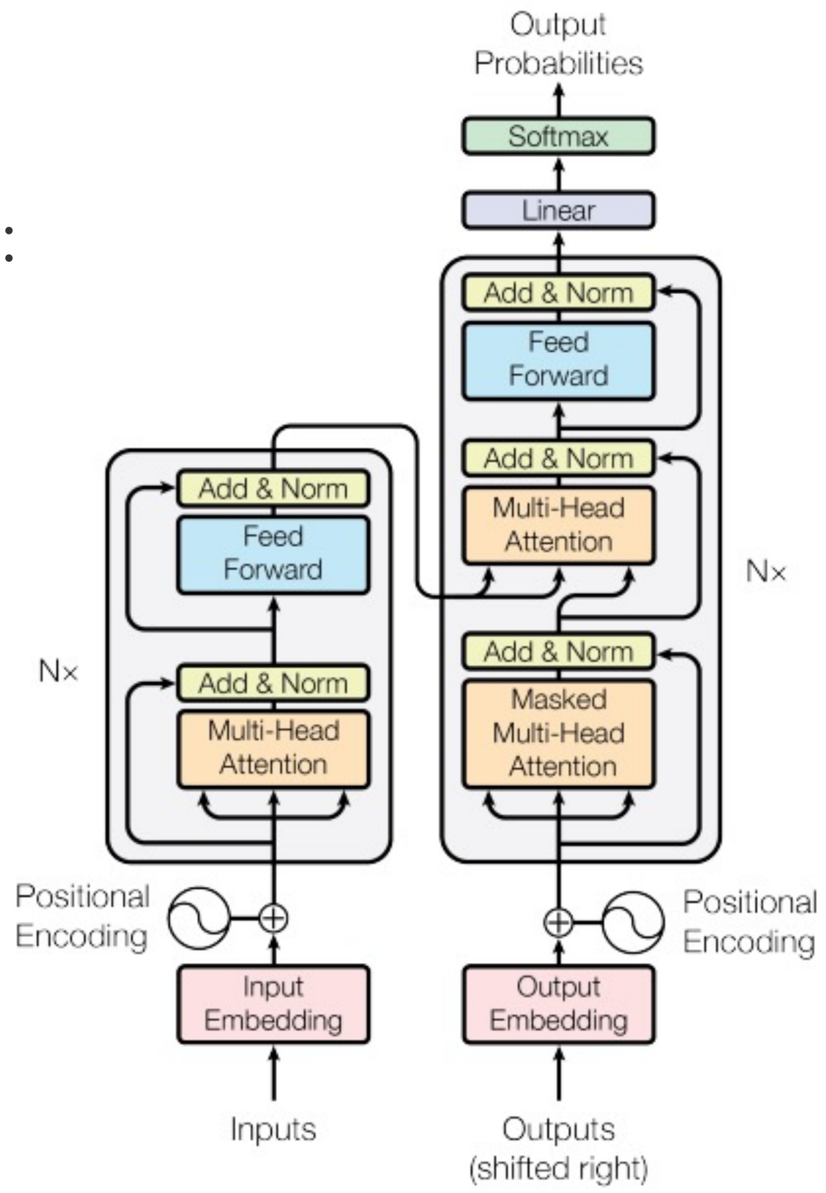


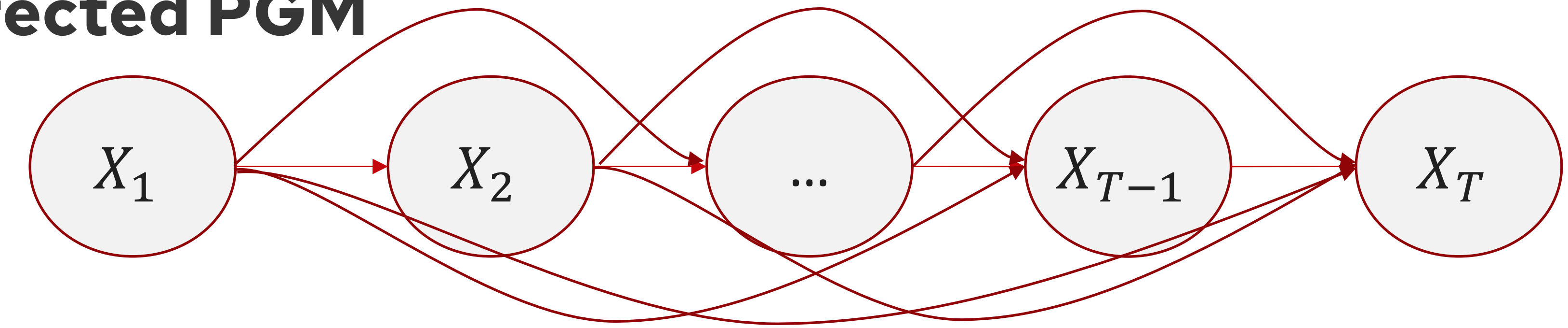
Figure 1: The Transformer - model architecture.

GPT: From Seq. Transduction to Seq. Modeling

- **Original Transformer** (Vaswani et al., 2017):
$$P(Y \mid X) = \prod_t P(Y_t \mid Y_{<t}, X)$$
  - **Conditional** sequence model for tasks like translation (input → output)
- **Generative Pretrained Transformer (GPT)** Models:
$$P(X) = \prod_t P(X_t \mid X_{<t})$$
  - **Unconditional** generative model over raw text
- Architectural consequence: **no encoder**, only a decoder with causal structure

LLMs: The definition of Generative Models

- **Directed PGM**



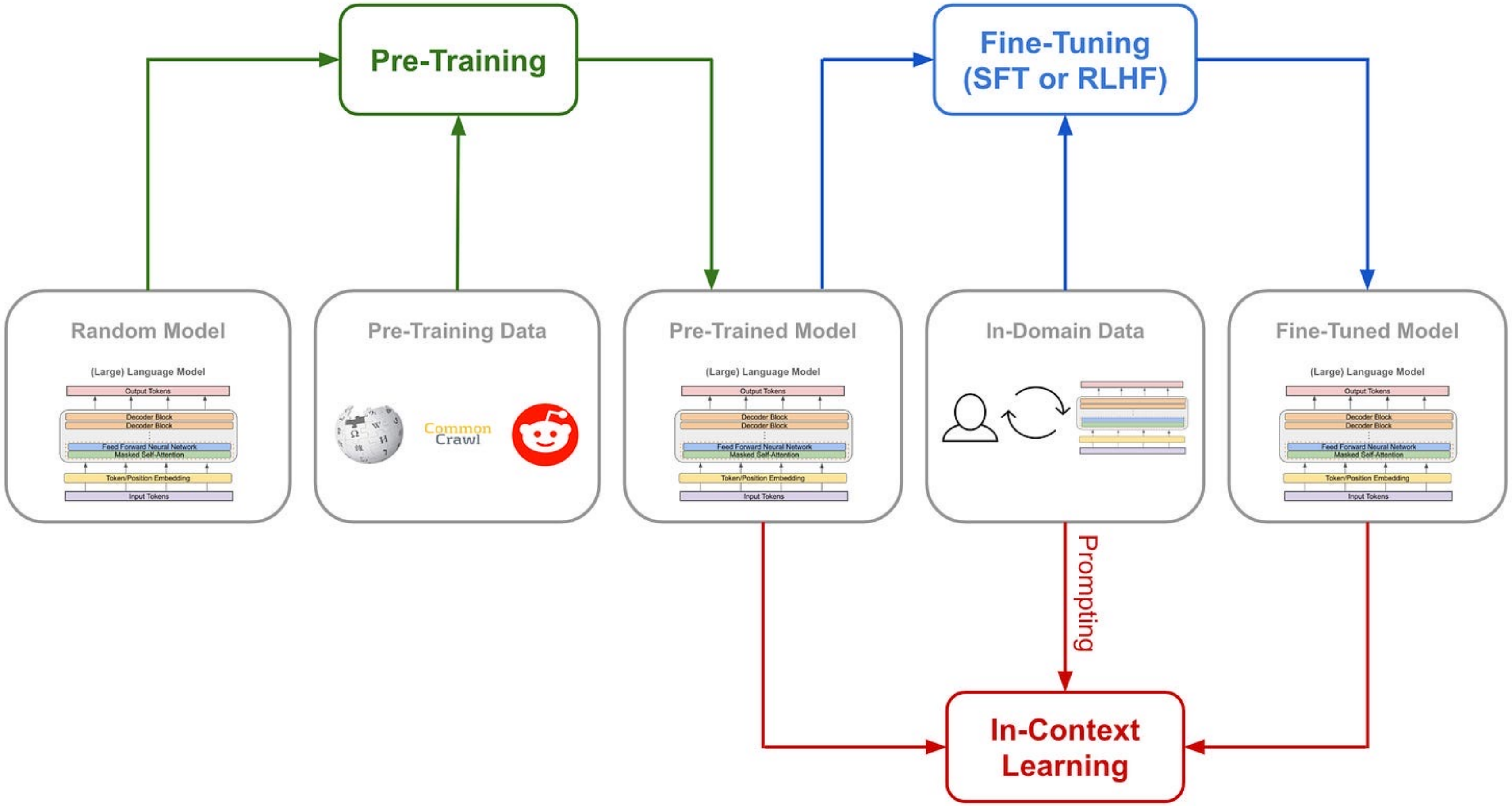
$$P_\theta(X) = \prod_i \prod_t P_\theta(X_{i,t} \mid X_{i,<t})$$

- **Probabilistic objective:** Max log-likelihood of observed seqs

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

[Radford et al., Improving Language Understanding by Generative Pre-Training]

LLM Training: Unsupervised → Supervised



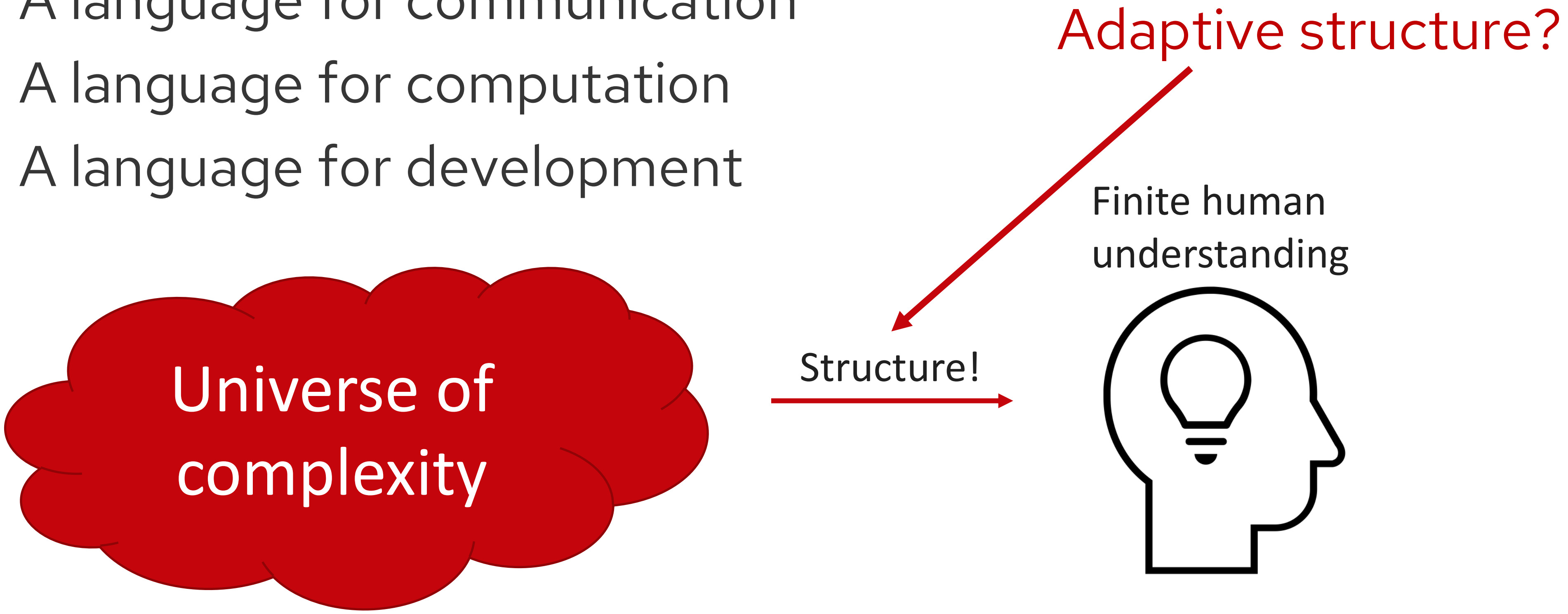
https://cameronrwolfe.substack.com/p/understanding-and-using-supervised

Open Directions in Graphical Models

Why GMs?

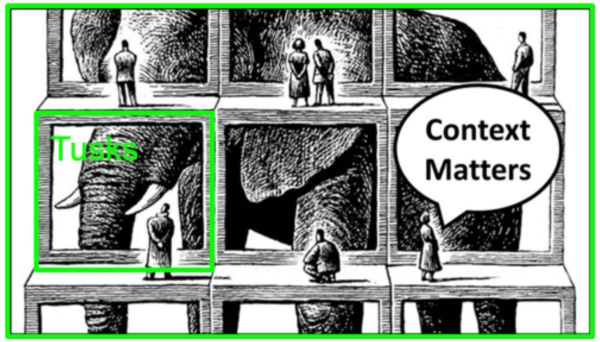
What's the point of GMs in the AI era?

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



Context-Adaptive GMs

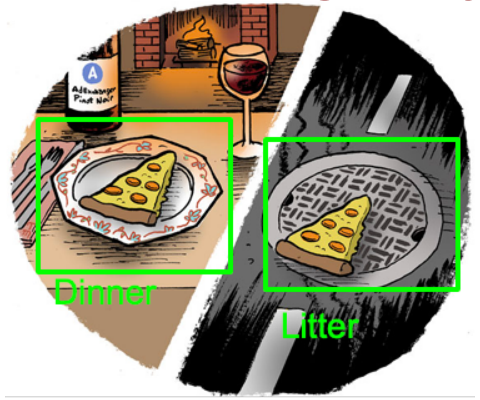
Interpreting complex systems



Elephant

- Zooming in for **personalization**
- Zooming out for **inclusion**

Latent **heterogeneity**



- Disease **subtypes**
- **Multiple-hit** mechanisms
- Prior **exposures**

Multi-modal effects

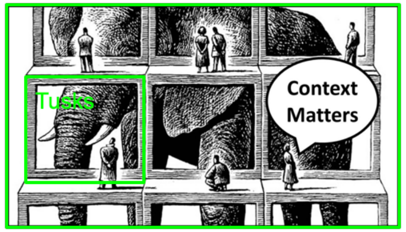
Dog

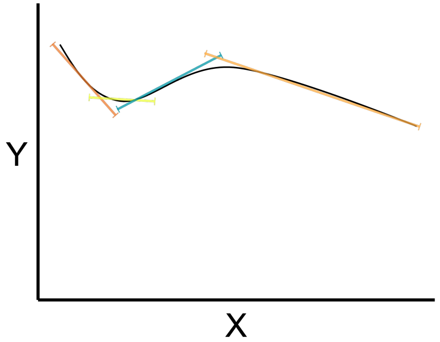
- Identifying and eliminating **biases**
- Connecting statistics to **foundation models**

Context-Adaptive GMs

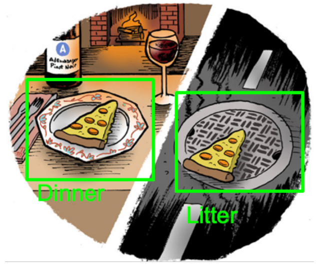
$$Y = X\hat{\beta}_{\Phi}(C) + \hat{\mu}(C) + \epsilon$$

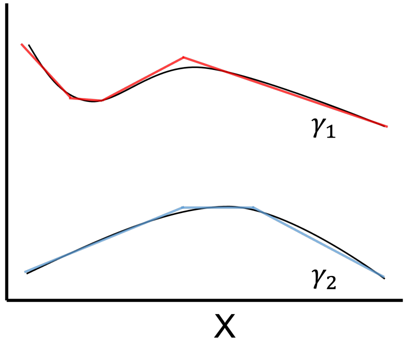
Interpreting complex systems



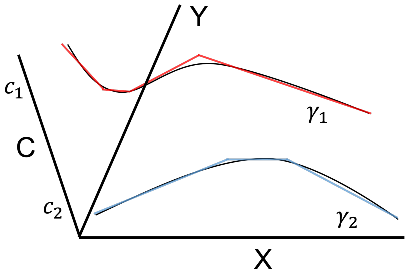


Latent **heterogeneity**





Multi-modal effects



Varying-Coefficients Regression

From

$$Y = \beta_0 + \beta_1 X_1 + \ \ldots + \beta_p X_p + \epsilon$$

To

$$Y = \beta_0(C) + \beta_1(C) X_1 + \ \ldots + \beta_p(C) X_p + \epsilon$$

Parameter-generating functions, each $R^m \rightarrow R$

Linear [Hastie & Tibshirani 1993]

Splines [Lu et al 2015]

Trees [Deshpande et al 2023]

Varying-Coefficients Regression

From

$$Y = \beta_0 + \beta_1 X_1 + \ \dots + \beta_p X_p + \epsilon$$

To

$$Y = \beta_0(C) + \beta_1(C) X_1 + \ \dots + \beta_p(C) X_p + \epsilon$$

Parameter-generating functions, each $R^m \to R$

Can these be neural networks?

Contextualized learning: A Recipe

1. Define a differentiable objective for your model of interest

$$\hat{\theta} = argmin_{\theta} \sum_{i=1}^{n} \ell(X_i, \theta)$$

$X \in \mathbb{R}^{n \times p}$

2. Replace model parameters with a differentiable context encoder

$$\widehat{\Phi} = argmin_{\Phi} \sum_{i}^{n} \ell(X_i, \Phi(C_i))$$

$C \in \mathbb{R}^{n \times c}$

$\Phi(c): \mathbb{R}^{c} \rightarrow \mathbb{R}^{|\theta|}$

3. (Optional) Re-parameterize the context encoder to constrain the solution space

$$\Phi(c; \phi, A) := \sum_{k=1}^{K} \phi(c)_k A_k$$

$K \ll |\theta|$

$A \in \mathbb{R}^{K \times |\theta|}$

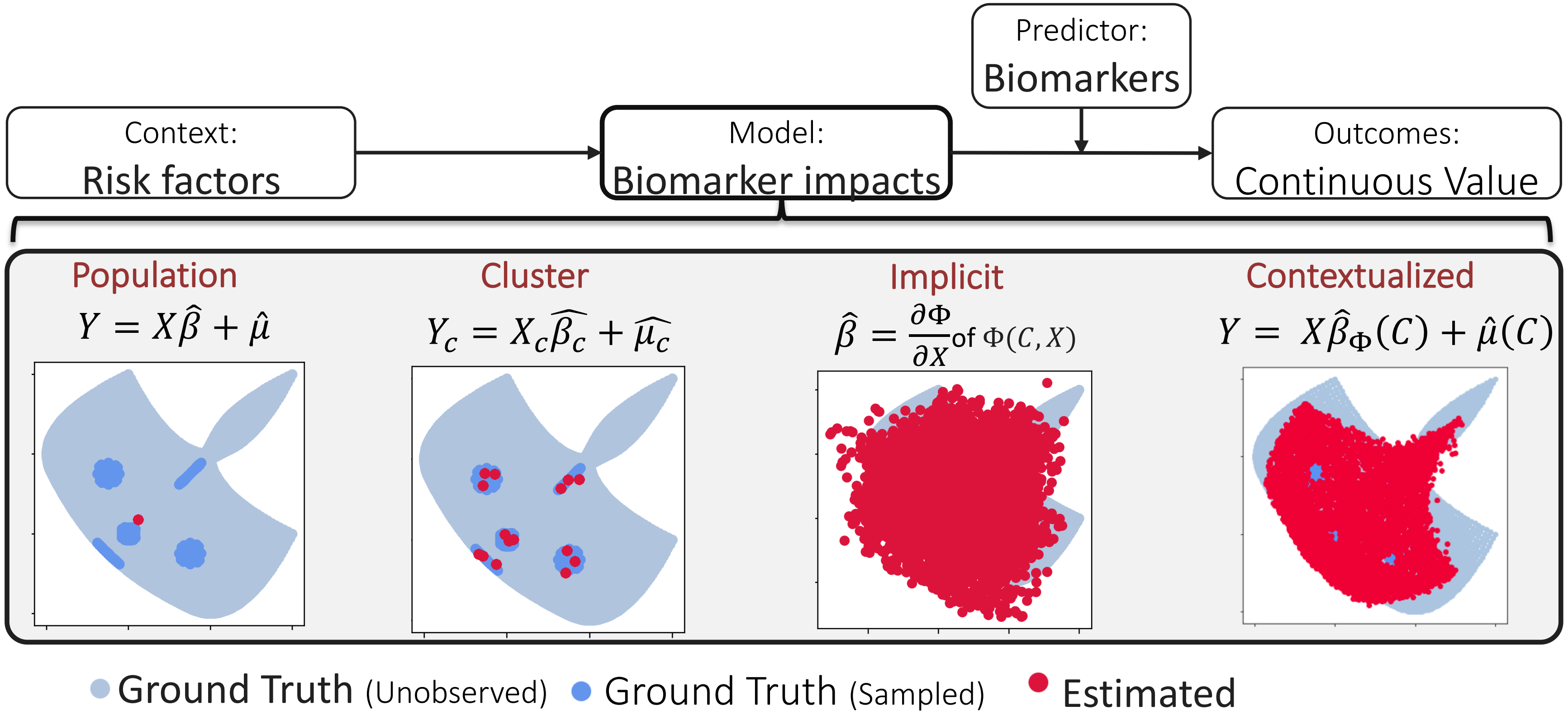
$\phi(c): \mathbb{R}^{c} \rightarrow \mathbb{R}^{K}$

4. Optimize end-to-end

$$\hat{\phi}, \hat{A} = argmin_{\phi, A} \sum_{i}^{n} \ell(X_i, \Phi(C_i; \phi, A))$$

Lengerich et al 2023

Toy Example: Linear Regression

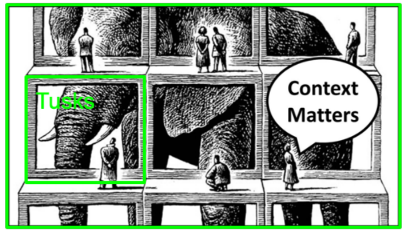


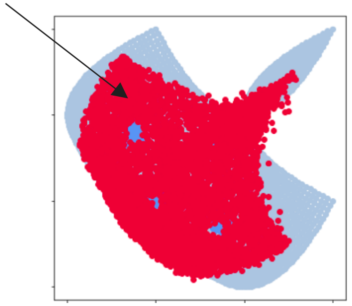
[Lengerich et al 2023](#)

Toy Example: Linear Regression

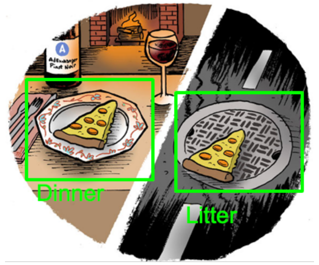
$$Y = X\hat{\beta}_{\Phi}(C) + \hat{\mu}(C)$$

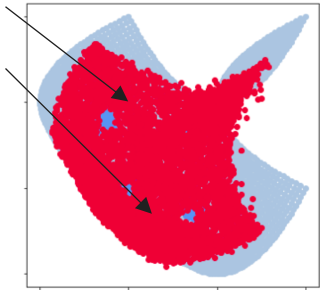
Interpreting complex systems



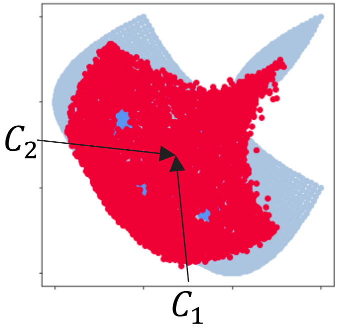


Latent **heterogeneity**



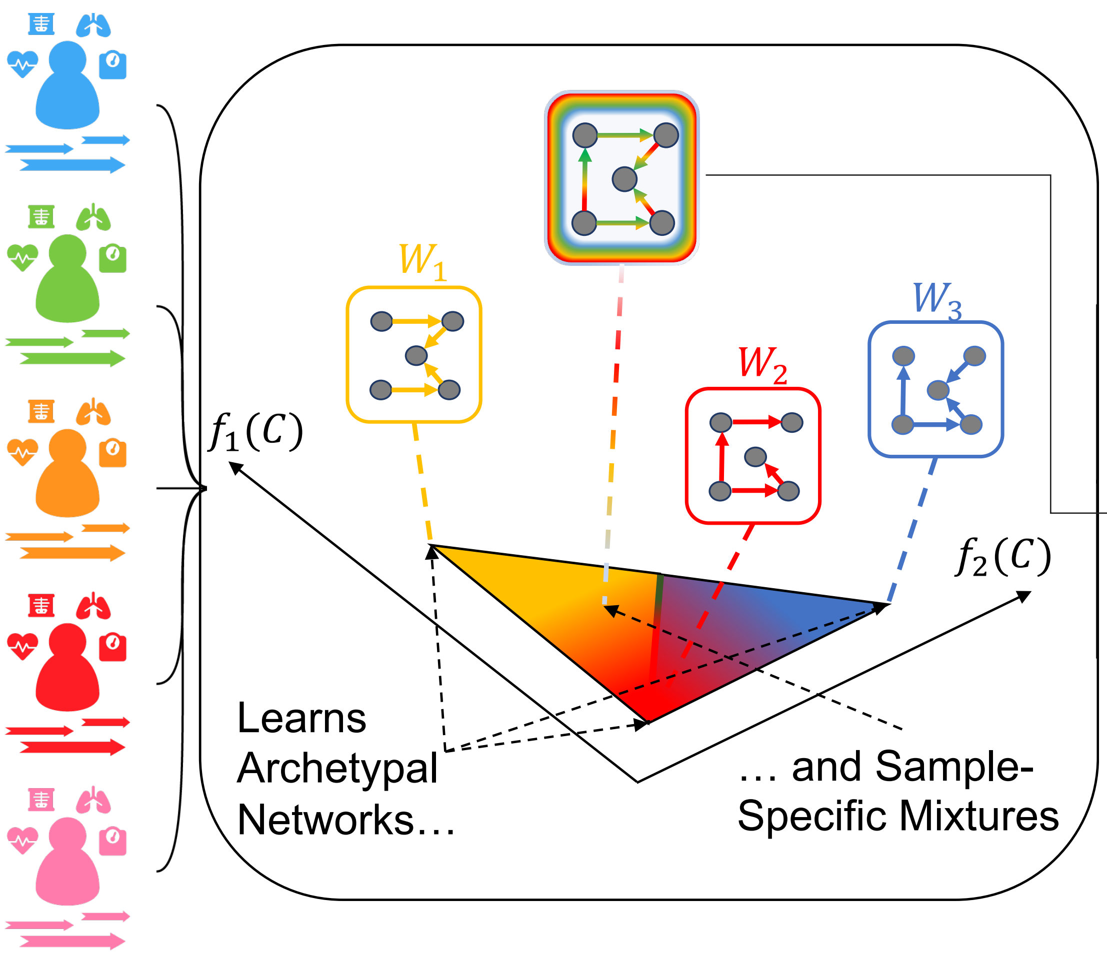


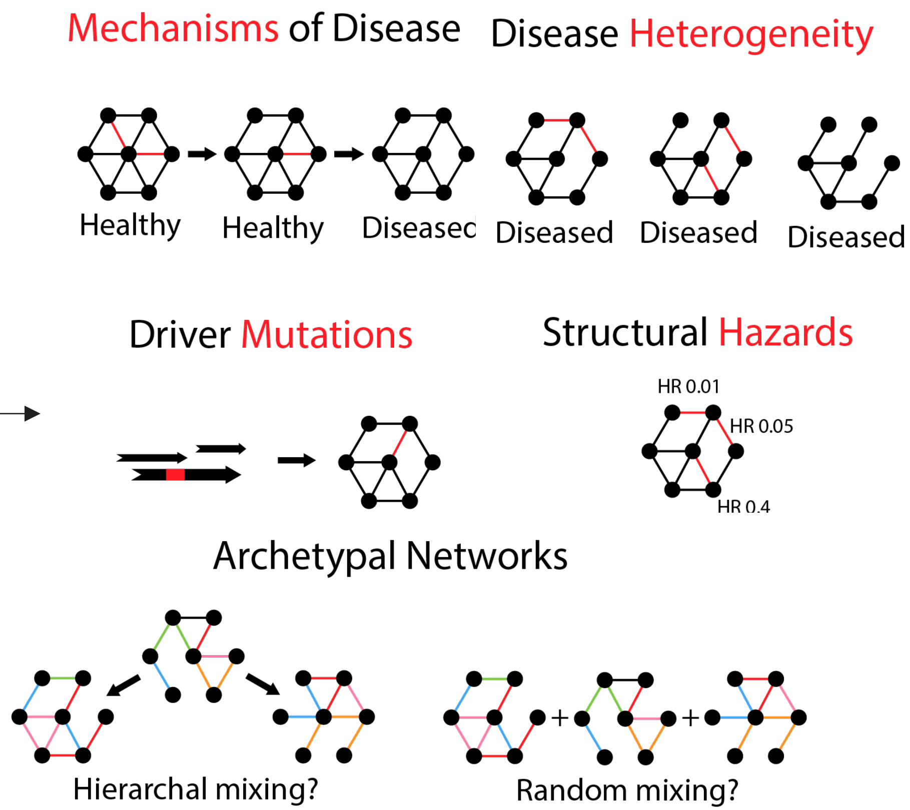
Multi-modal effects



Lengerich et al 2023

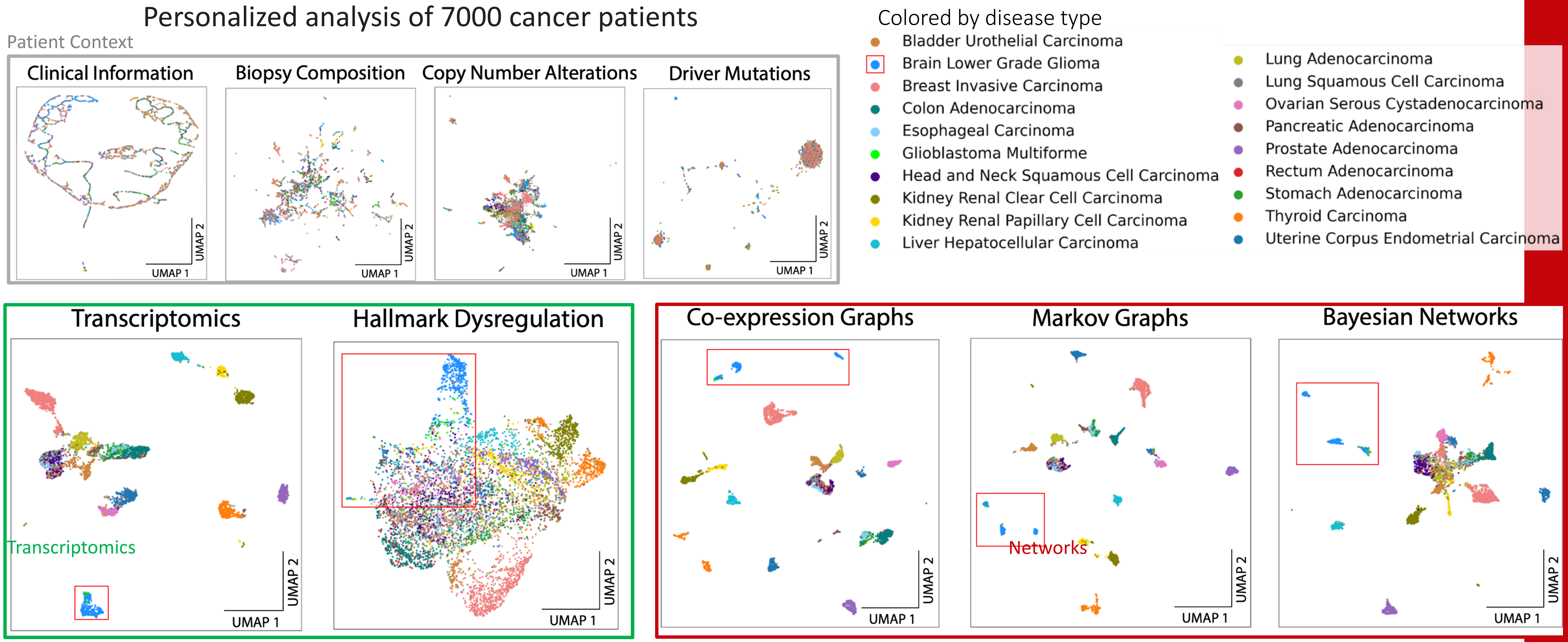
Contextualized GMs enable new studies of biology





Ellington et al. PNAS 2025 (to appear)

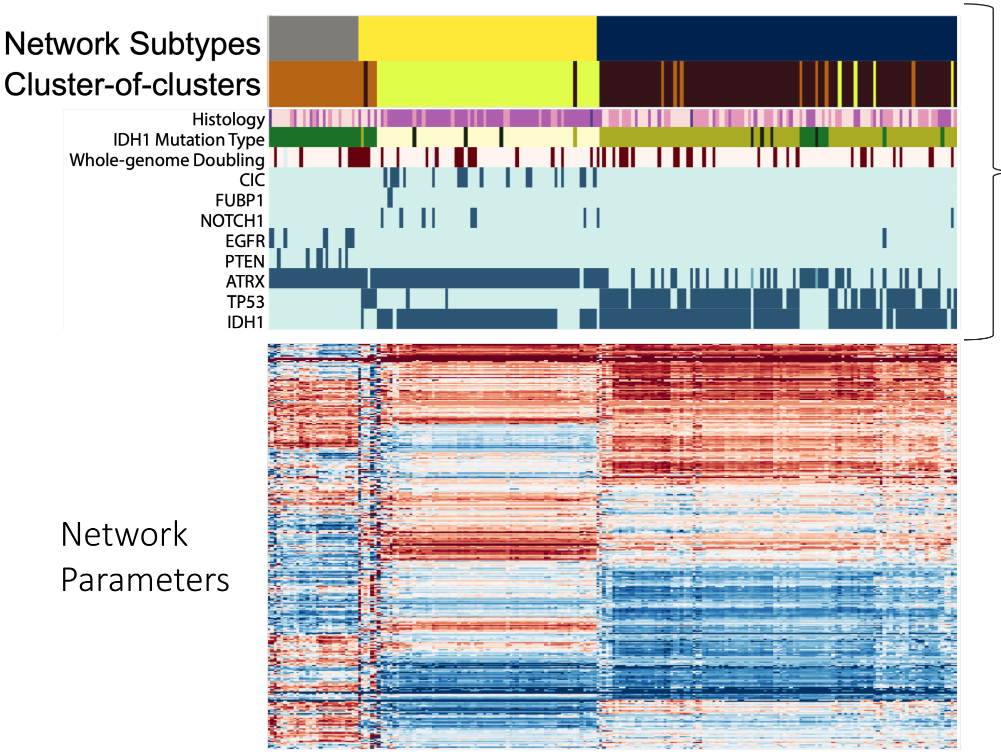
Contextualized GMs enable new studies of biology

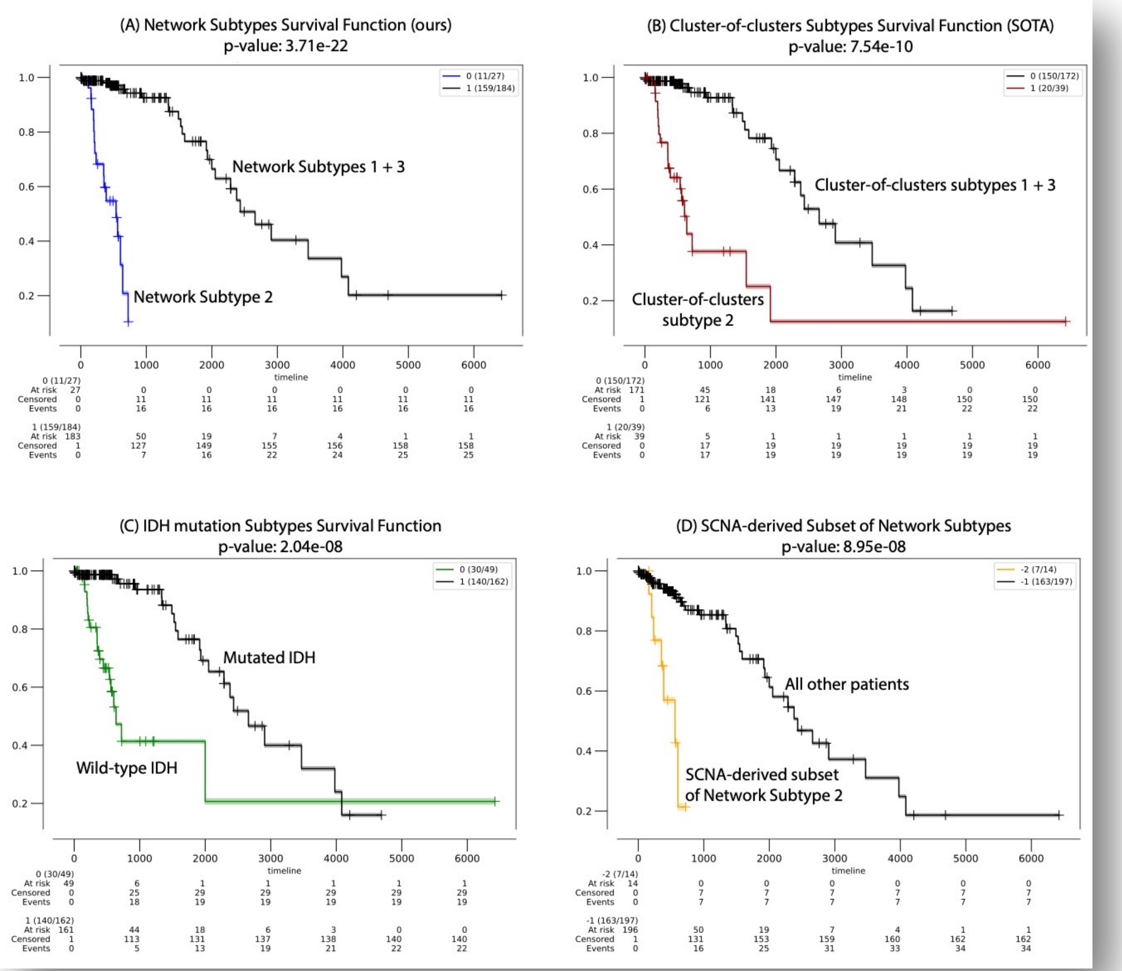


Ellington et al. PNAS 2025 (to appear)

Contextualized GMs enable new studies of biology

Personalized analysis of 7000 cancer patients

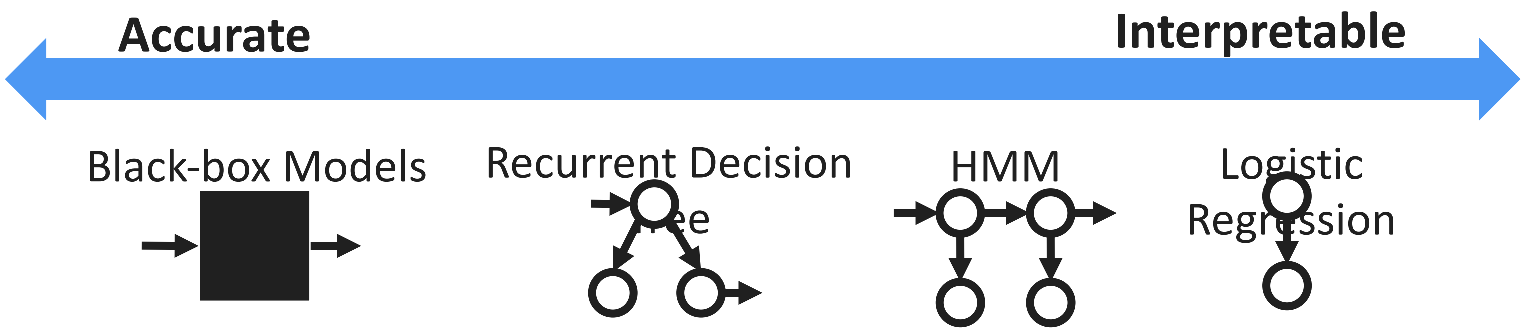




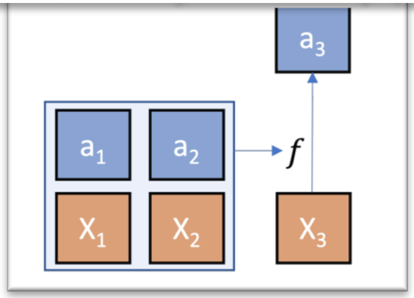
Ellington et al. PNAS 2025 (to appear)

Contextualized GMs work within RL too

Want to model recurrent processes of medical decisions as RL policies



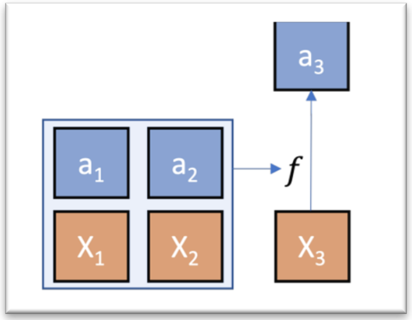
Contextualized Policy Recovery (CPR)

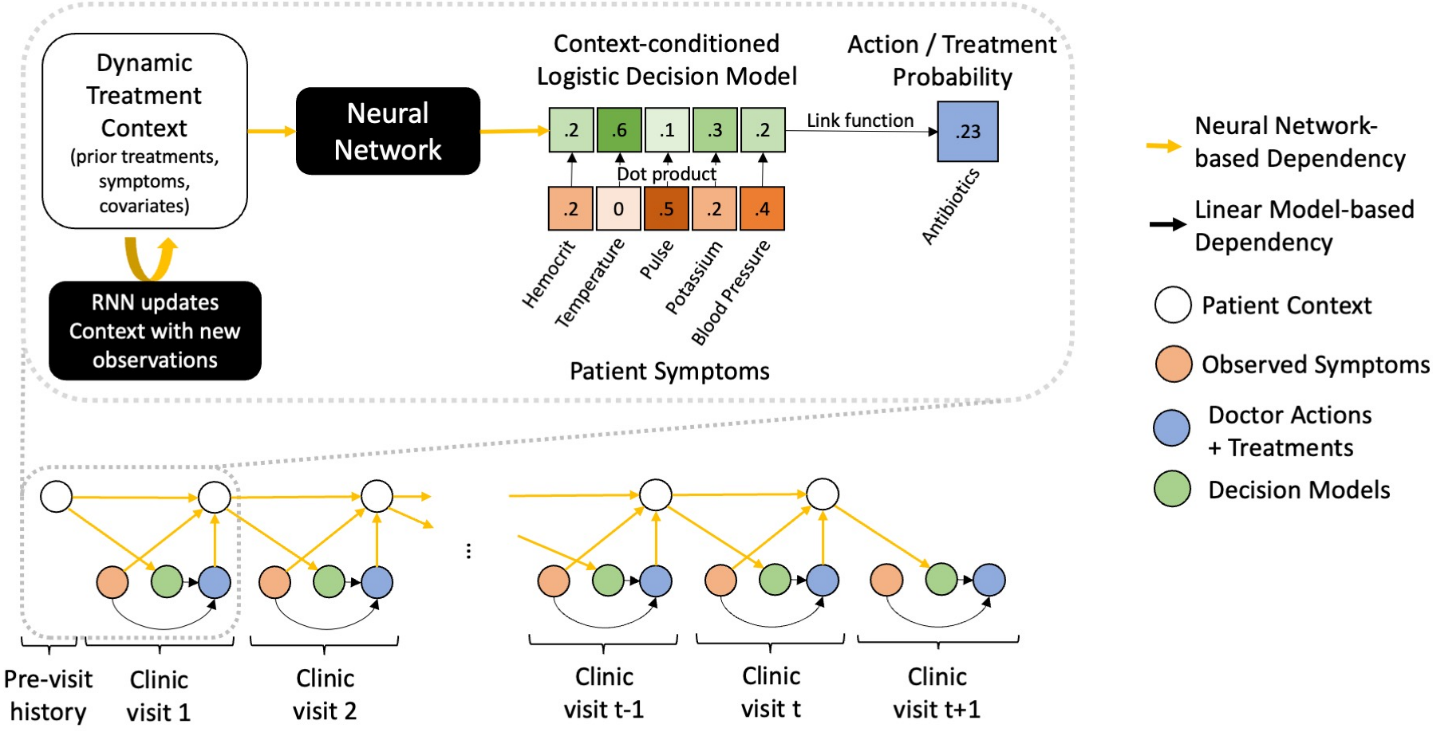


Deuschel et al. ICML 2023

Contextualized GMs work within RL too

Contextualized Policy Recovery (CPR)

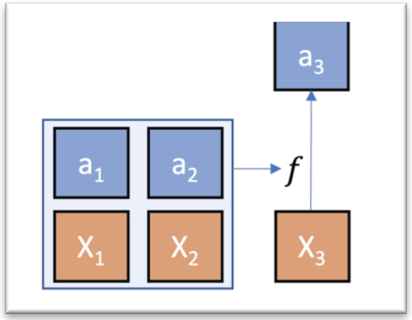




Deuschel et al. ICML 2023

Contextualized GMs work within RL too

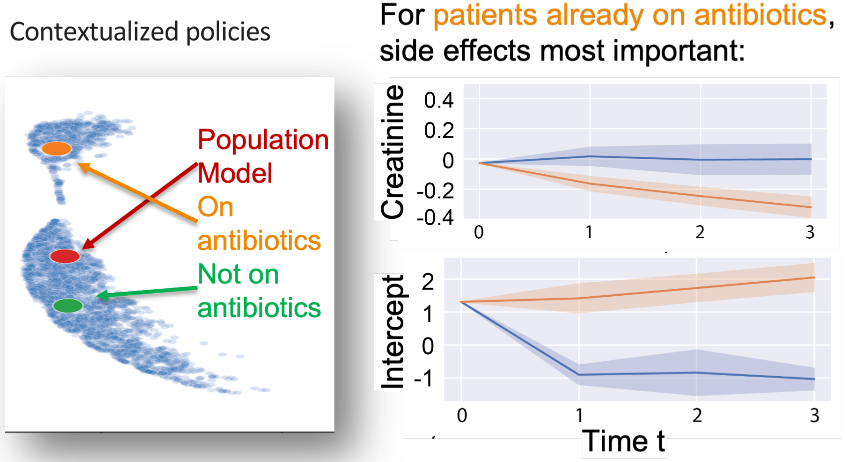
Contextualized Policy Recovery (CPR)



SOTA Accuracy

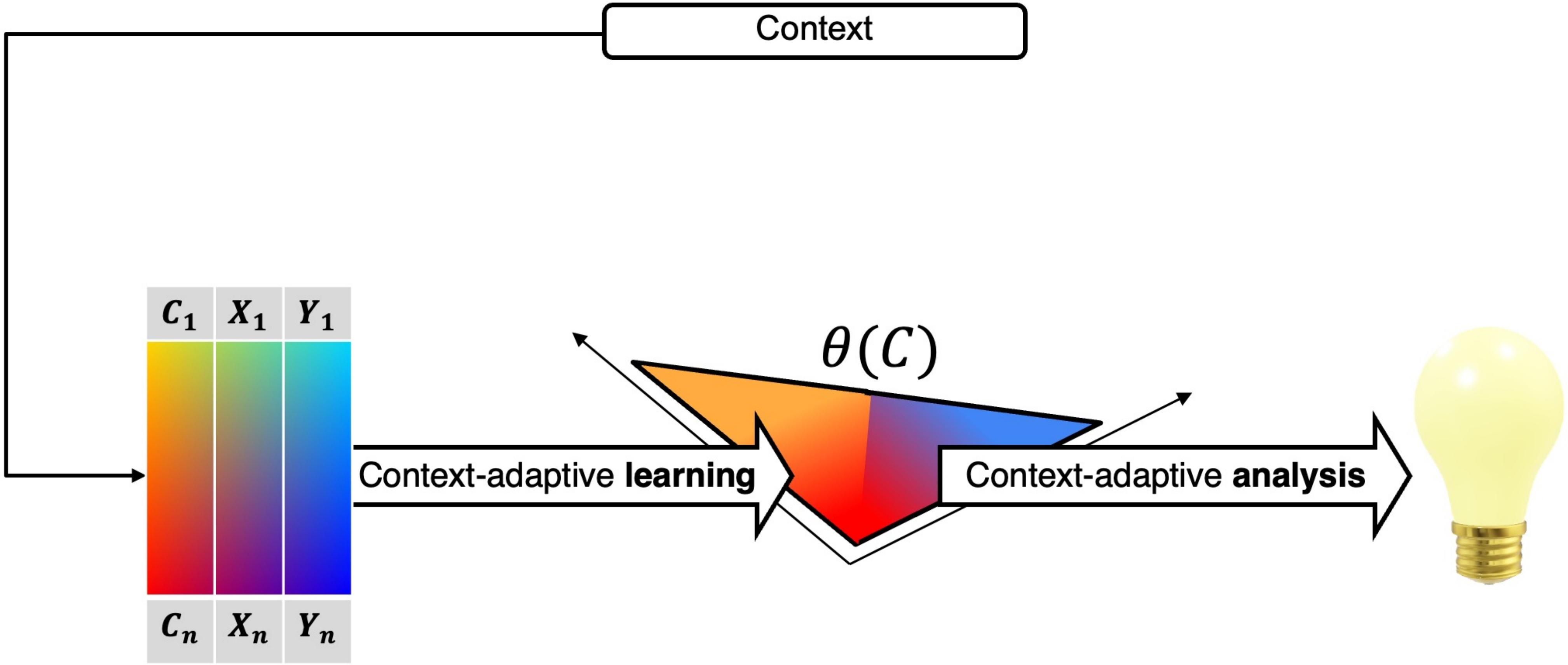
Model	AUROC	
	ADNI MRI Scans	MIMIC Antibiotics
stic regression	0.66 ± 0.01	0.57 ± 0.01
INTERPOLE	0.60 ± 0.04	NR
POETREE	0.62 ± 0.01	0.65 ± 0.04
CPR-RNN	**0.72 ± 0.01**	**0.82 ± 0.00**
CPR-LSTM	**0.72 ± 0.01**	**0.82 ± 0.00**
RNN	0.72 ± 0.01	0.83 ± 0.00
LSTM	0.71 ± 0.01	0.84 ± 0.00

Contexualized Understanding

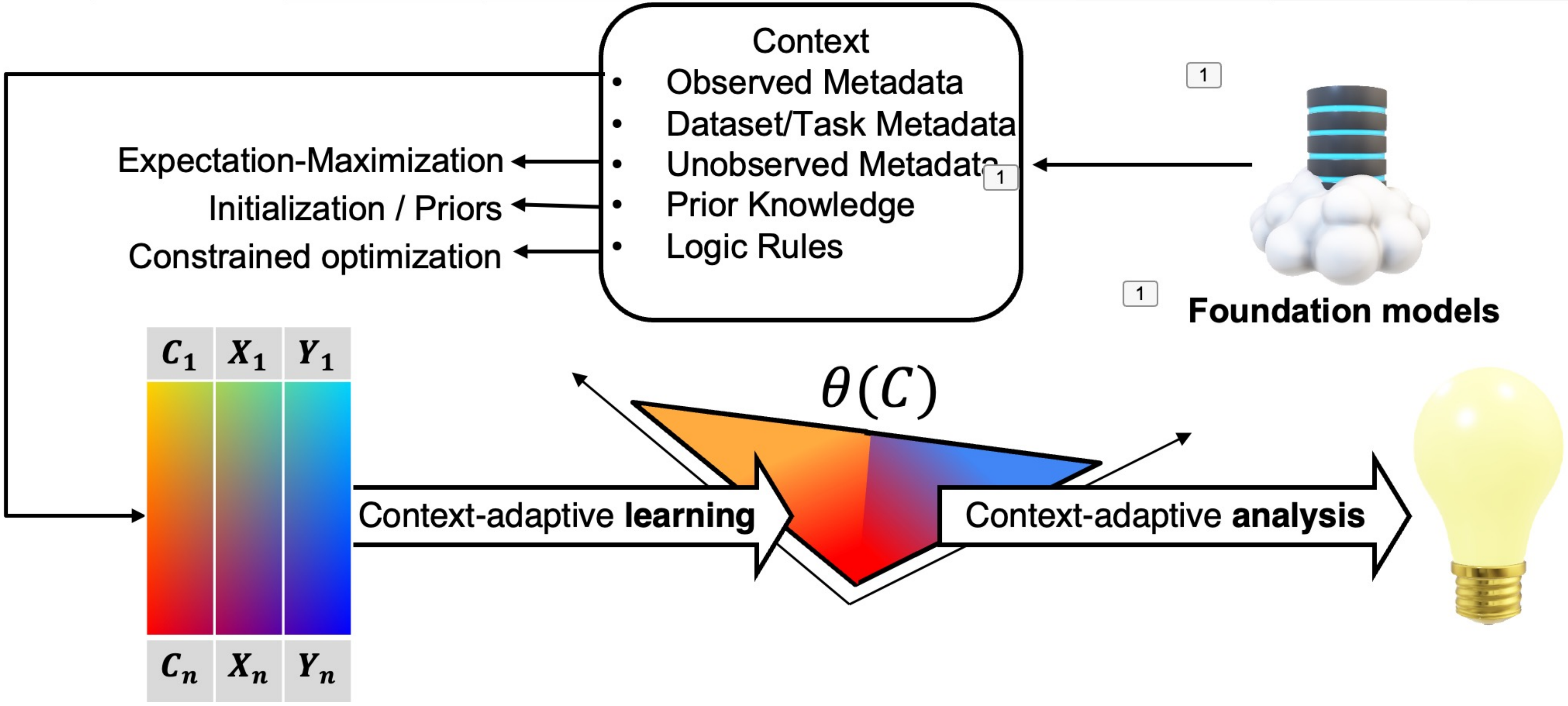


Deuschel et al. ICML 2023

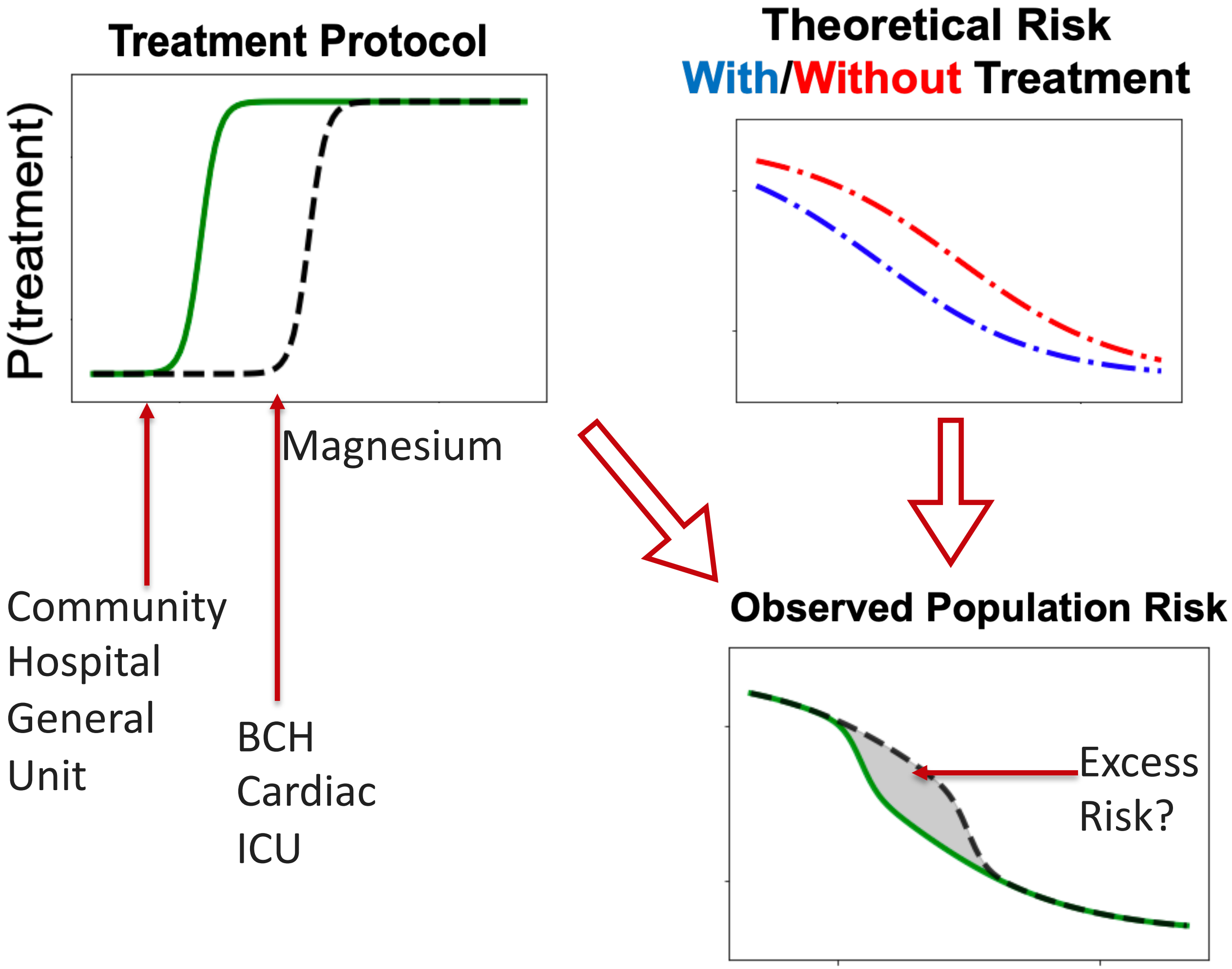
Context connects statistical ML to persistent knowledge



Context connects statistical ML to persistent knowledge

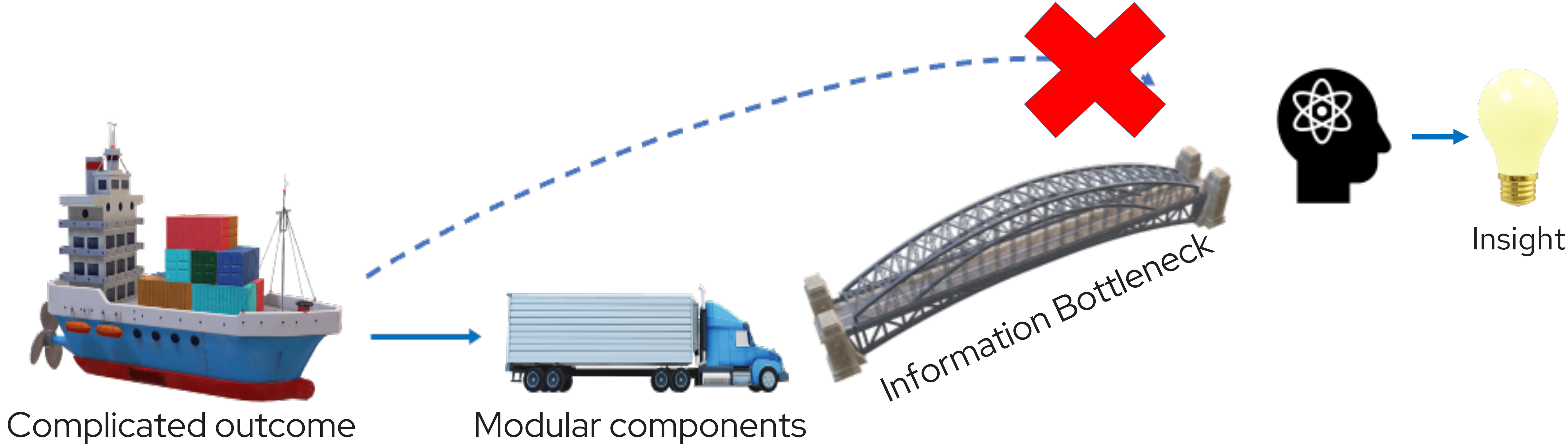


A recent personal story



A core idea of GMs: Modularity → interpretability

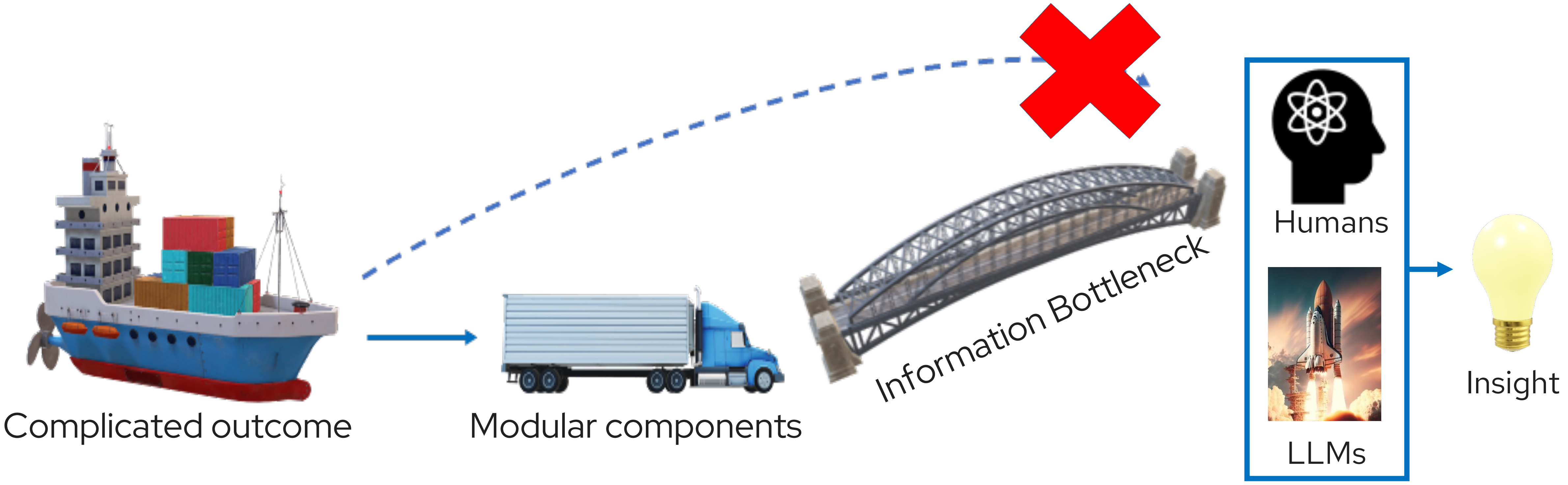
An **information bottleneck** limits human understanding of complicated ideas…



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

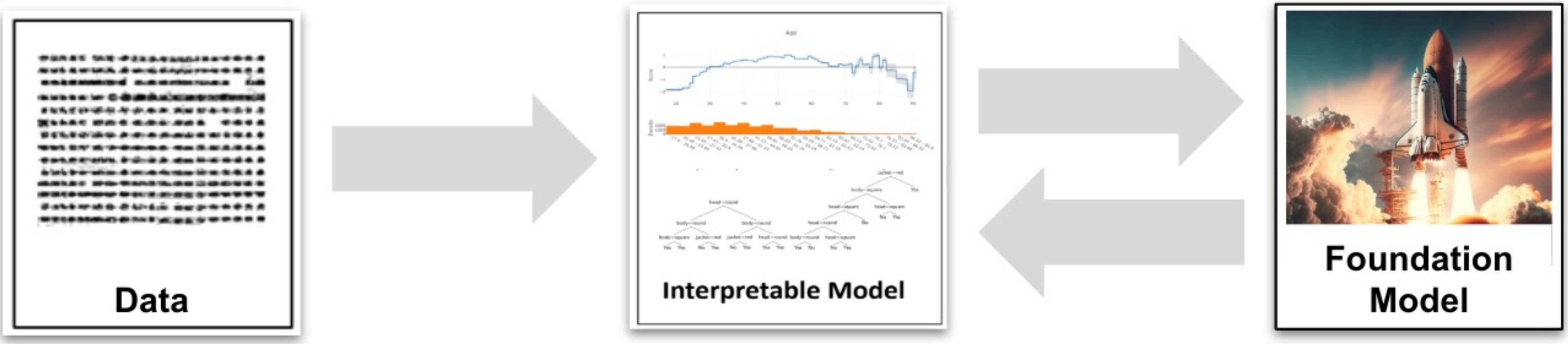
GMs + LLMs: Modularity → Automated Interpretability

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



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

GMs + LLMs → Tremendous potential



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

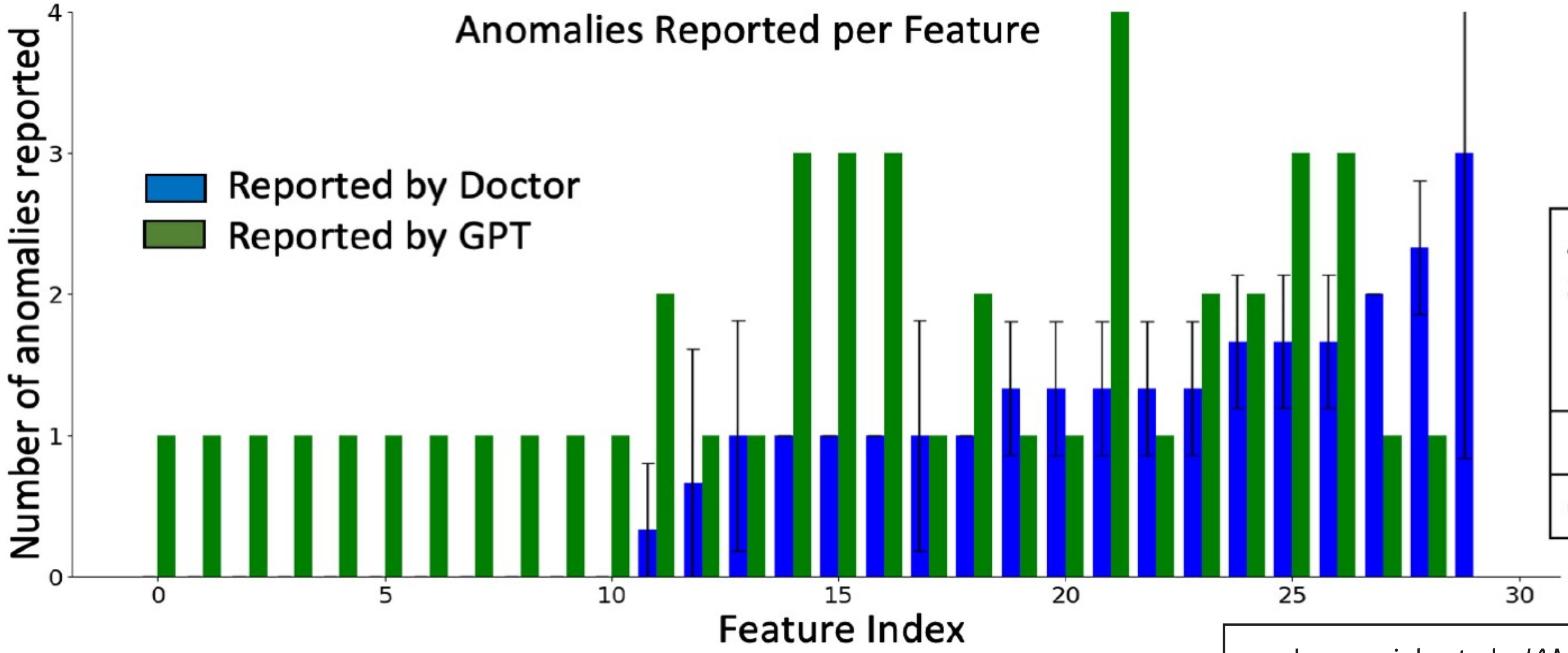
GPT-4 not as good as doctors

Lengerich et al. *JAMIA Open* 2025 (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.



Anomaly Source	Total Number of Anomalies Detected
Doctors	29.3 +- 4.1
GPT	46

Lengerich et al. *JAMIA Open* 2025 (to appear)

Many open problems and opportunities

- **Scalability of Contextualized Learning:** Systems for storing, accessing, and generating context-specific models
- **Integration with Emerging Biomedical Technologies:** More views of personal context (wearables) and fine-grained interventions (CRISPR, Perturb-seq)
- **Combining Episodic and Semantic Memory:** Beyond Archetypes
- **Ethical and Privacy Considerations:** Which features should be used to personalize risk models? Which should be invariant?
- **Robust Local Interpretations:** Can we guarantee robustness of local interpretations via smoothness, (adversarial) robustness, or other properties?
- **Federated learning and data sharing:** How can local models be pooled into meta-models with only minimal access to original data?
- **Communication protocols:** Should all communication be routed through the meta-model?
- **Resource efficiency and accessibility:** When can the meta-model be ignored?
- **Longitudinal studies and real-world impact:** What kinds of personalized interventions really make a difference?

What will you do with the language of complexity?

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON