Contextualized Machine Learning for Applied Data Science

Feb 1, 2023
Accelerating Physics with ML

Ben Lengerich, blengeri@mit.edu
Manolis Kellis, manoli@mit.edu
Computational Biology Lab, CSAIL, MIT

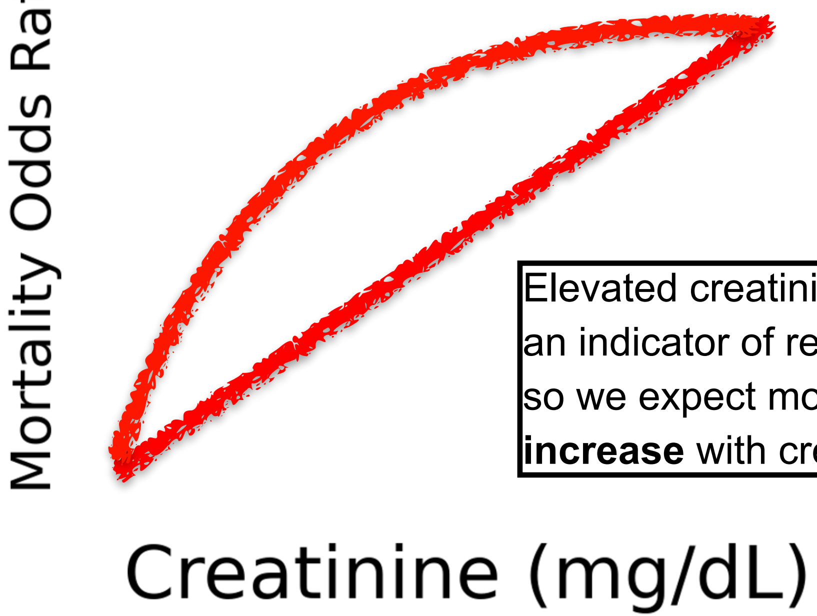
Massachusetts Institute of Technology
Computer Science and Artificial Intelligence Lab
CSAIL
BROAD INSTITUTE

Challenges with Real-World Data

- Time-varying effects
  - Noise models change
- Distribution shifts (instrument / upstream model changes)
- Outliers / non-Gaussian glitches
- Model retraining - online?
- Interpretability - What are we really learning?

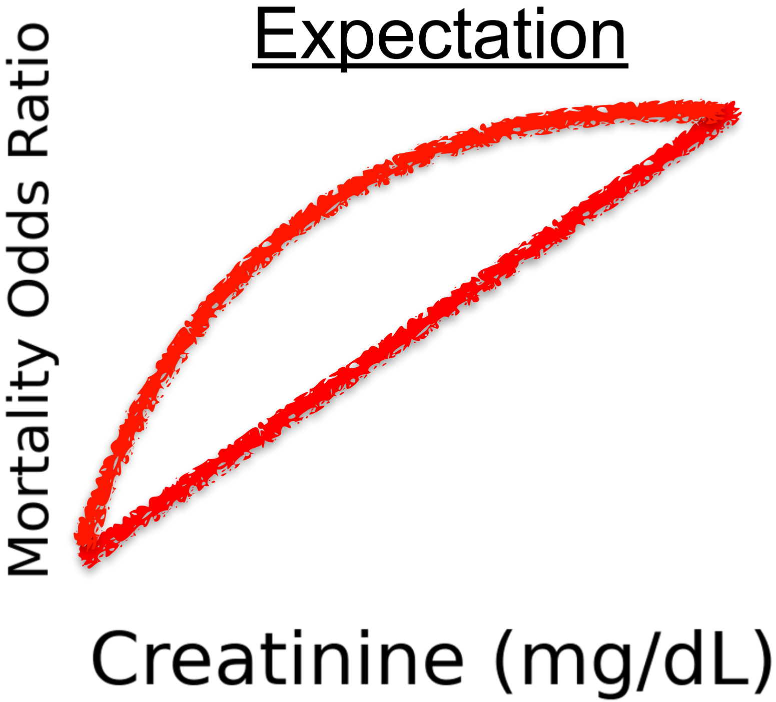
Observational data don't always match our expectations

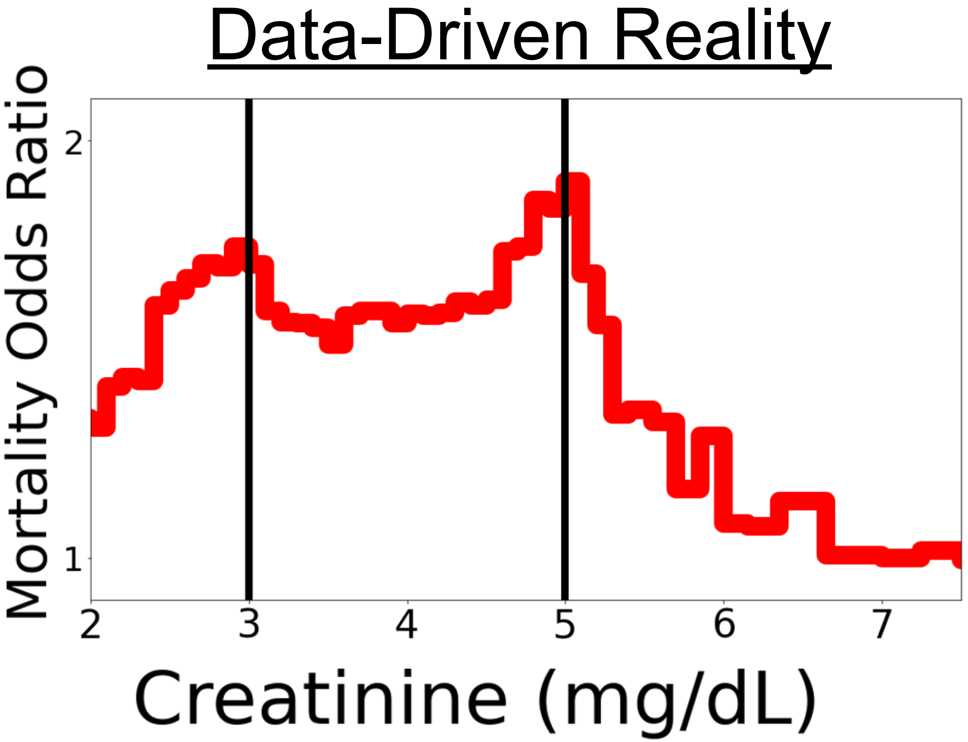
Expectation



Elevated creatinine levels are an indicator of renal failure, so we expect mortality risk to **increase** with creatinine.

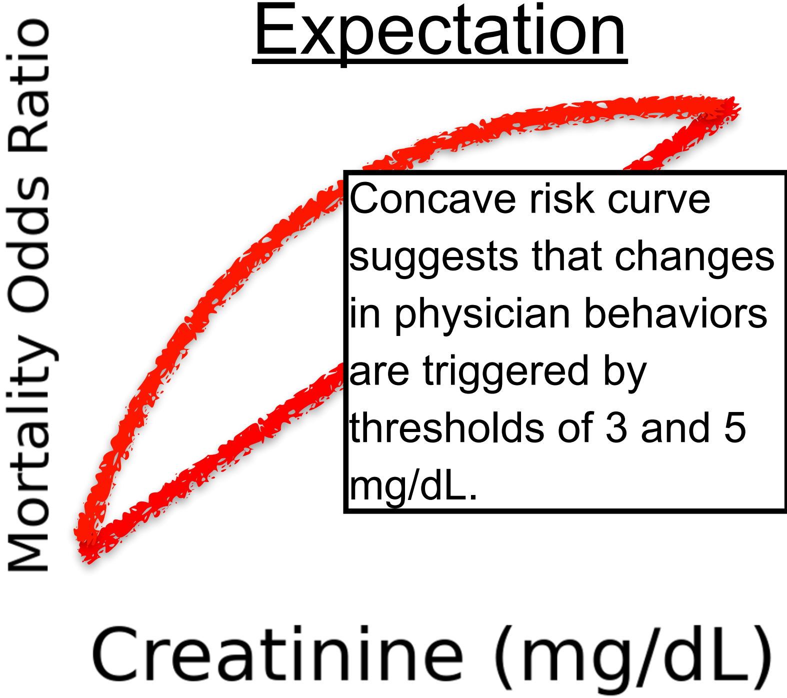
Observational data don't always match our expectations





Observational data don't always match our expectations

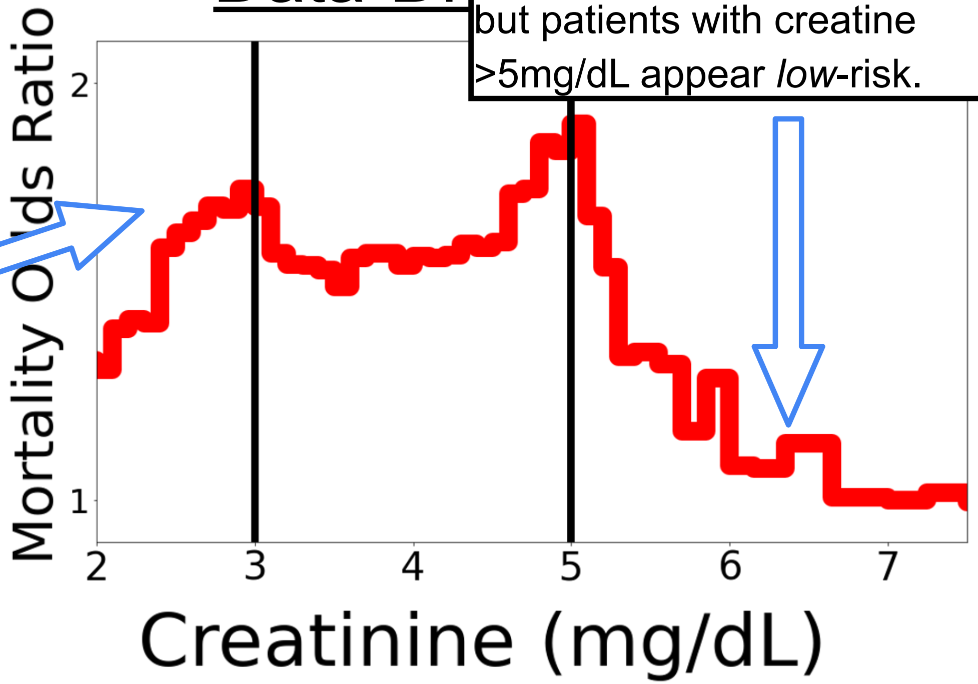
Expectation



Concave risk curve suggests that changes in physician behaviors are triggered by thresholds of 3 and 5 mg/dL.

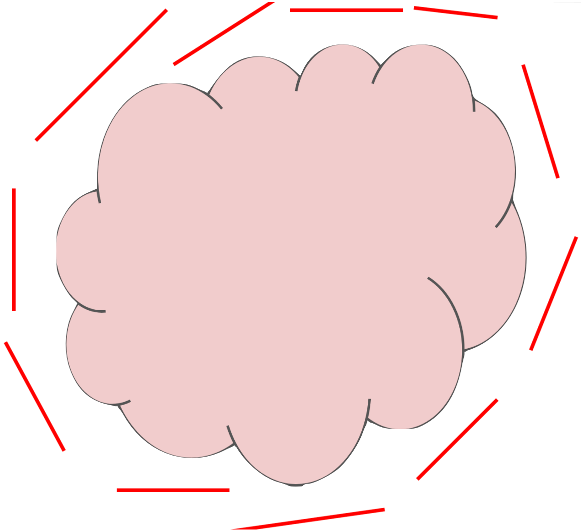
Data-Dr

Elevated creatinine levels are an indicator of renal failure, but patients with creatine >5mg/dL appear *low*-risk.



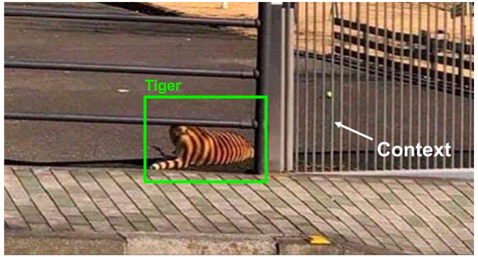
One Solution: Locally-Interpretable Models

- Local models can be both interpretable and accurate
- Universal approximators

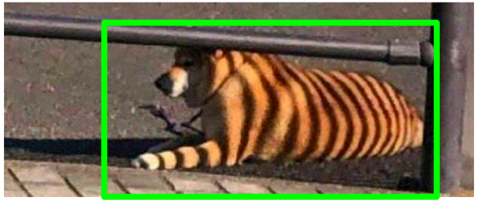


Local models: 3 Philosophies

1 Local models are incorrect, obscured by context factors

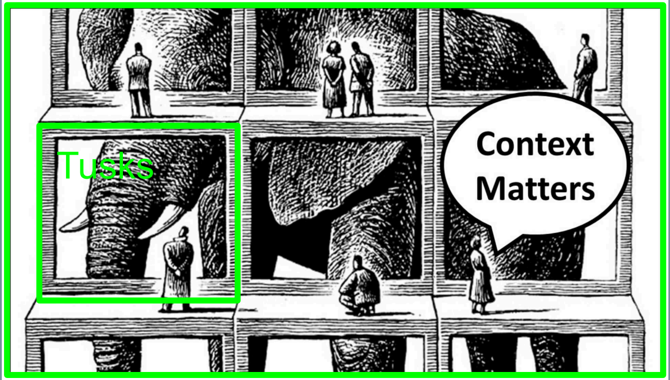


Tiger or Dog?



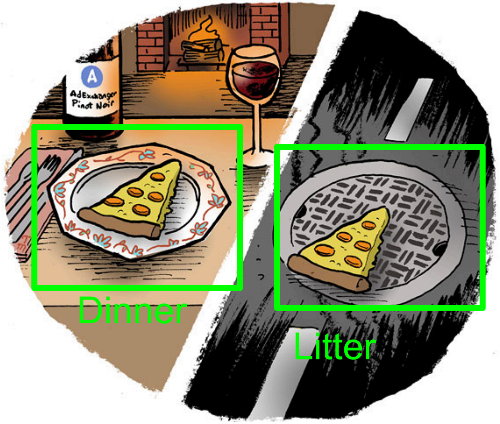
Solution: Subtract influence of unseen context factors to estimate universal effects

2 Local models are context-specific views of a universal phenomena



Solution: Context-specific models —> reconstruct global model

3 Local models are accurate views of context-specific phenomena



Solution: Context-specific models —> context-specific effects

Contextualized
Heterogeneous Modeling Toolbox

Massachusetts Institute of Technology

CSAIL Computer Science and Artificial Intelligence Lab

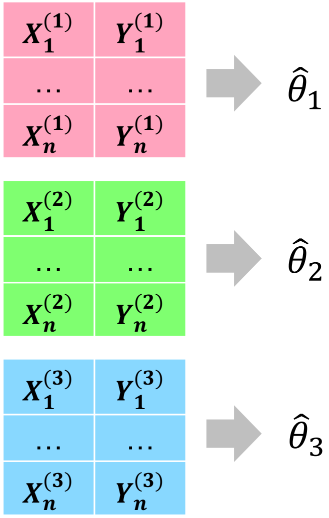
BROAD INSTITUTE

Our Solution:
Contextualized Machine Learning

Contextualized Machine Learning

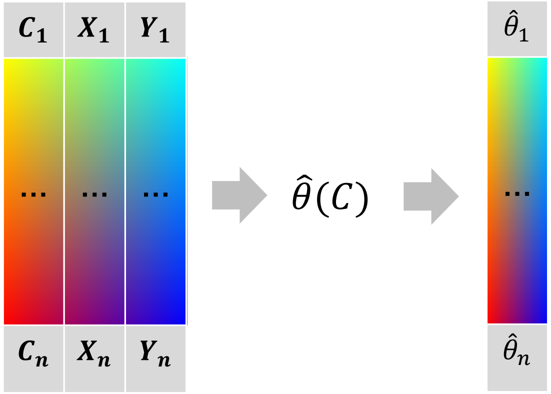
Cohort Modeling

statistics of discrete partitions

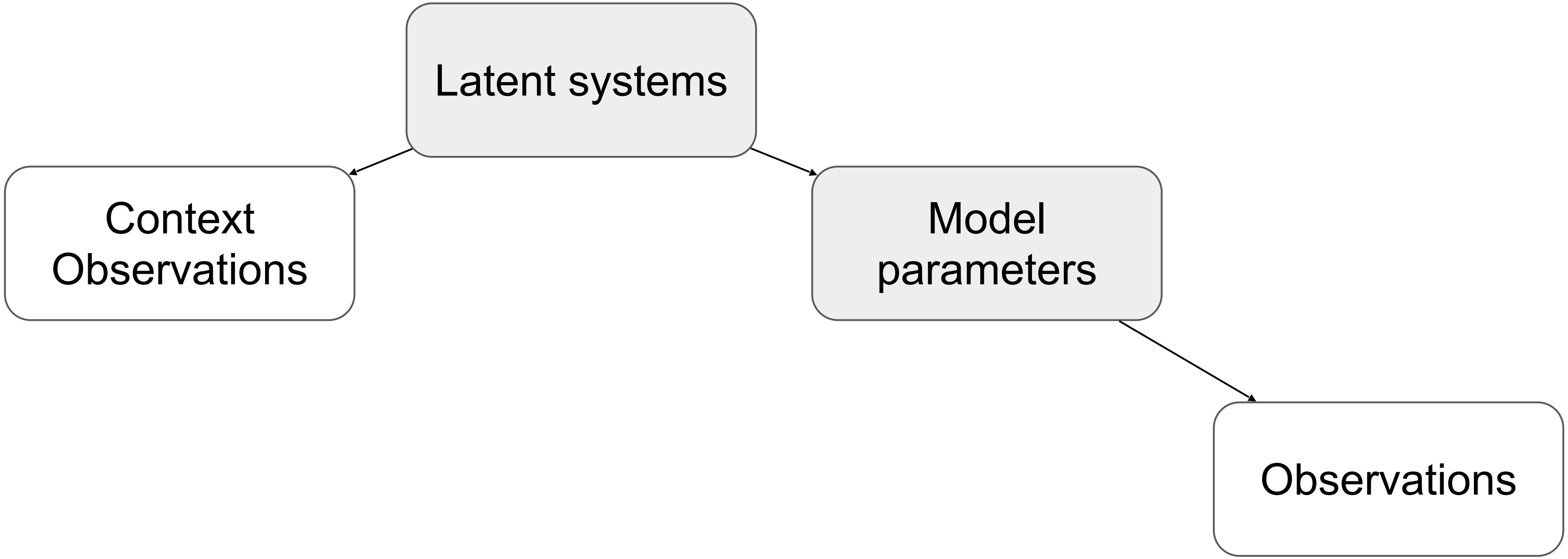


Contextualized Modeling

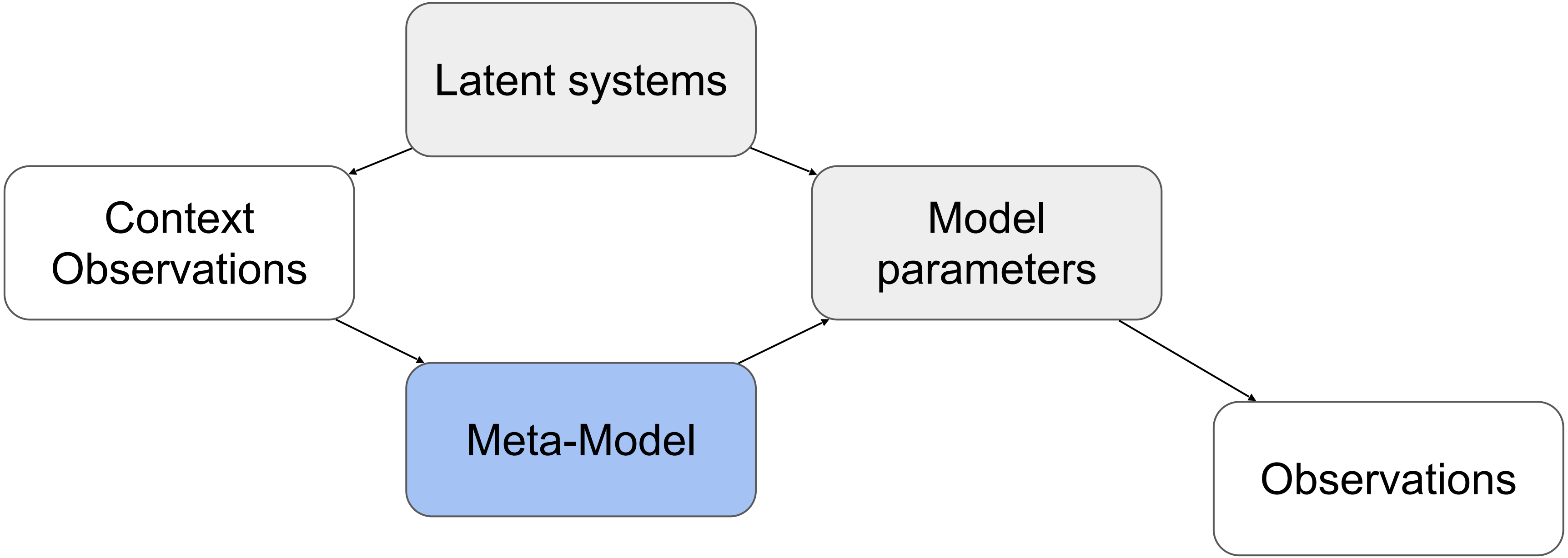
parameters as functions of context



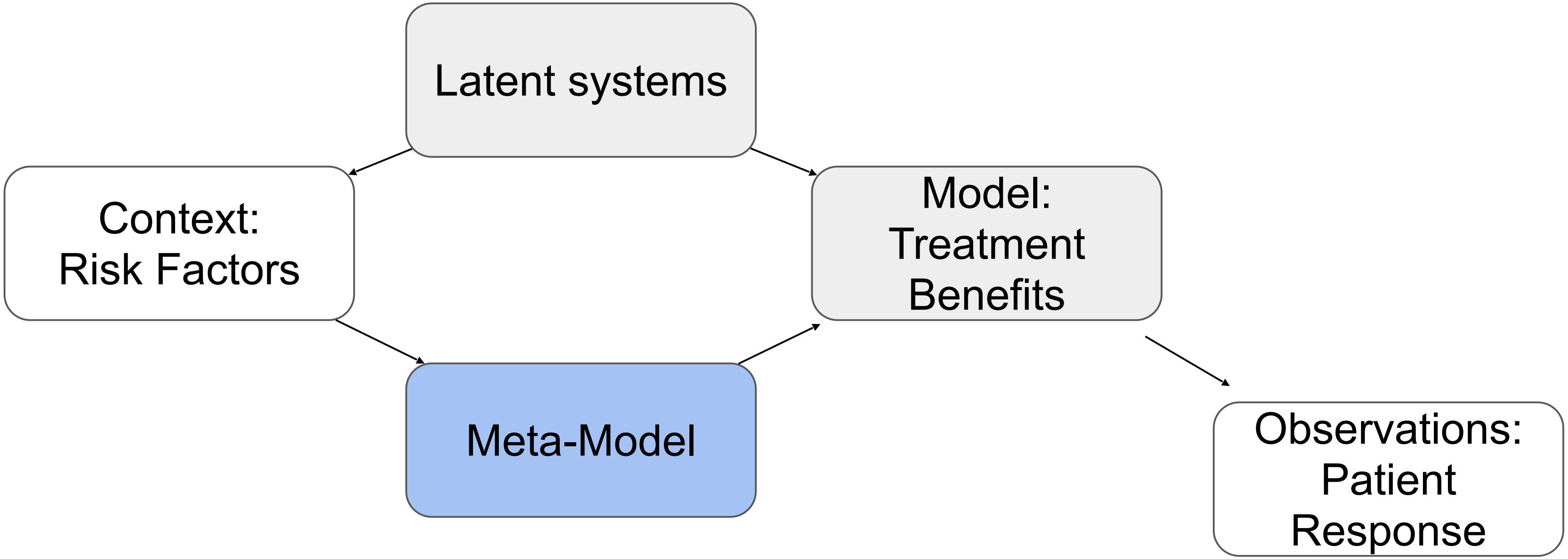
Contextualized Machine Learning



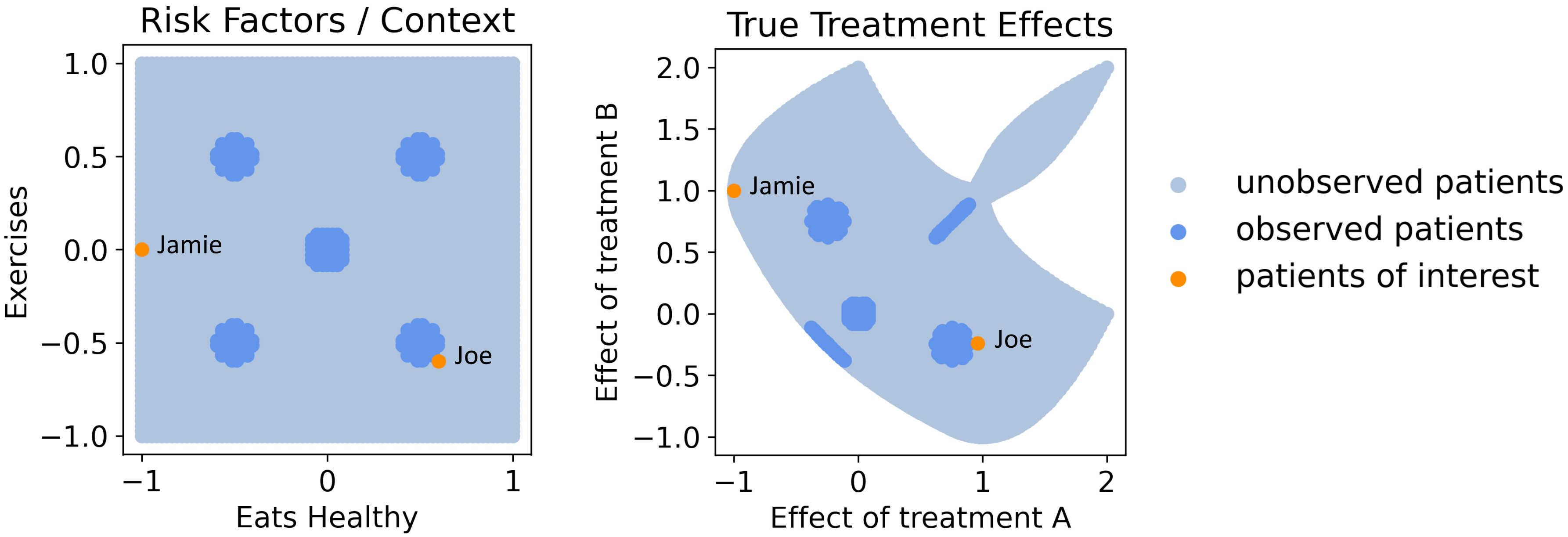
Contextualized Machine Learning



Toy Example: Heterogeneous Treatment Effects

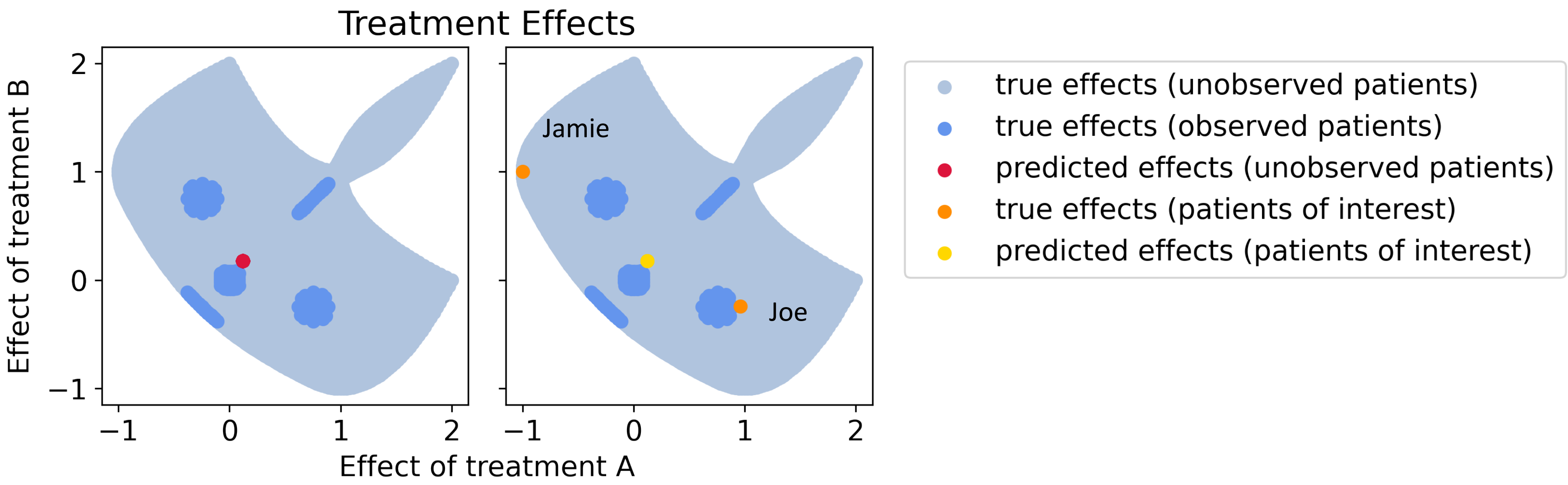


Toy Example: Heterogeneous Treatment Effects



Population Model: No Heterogeneity

Learn a single (population) model by solving $Y = X\hat{\beta} + \hat{\mu}$



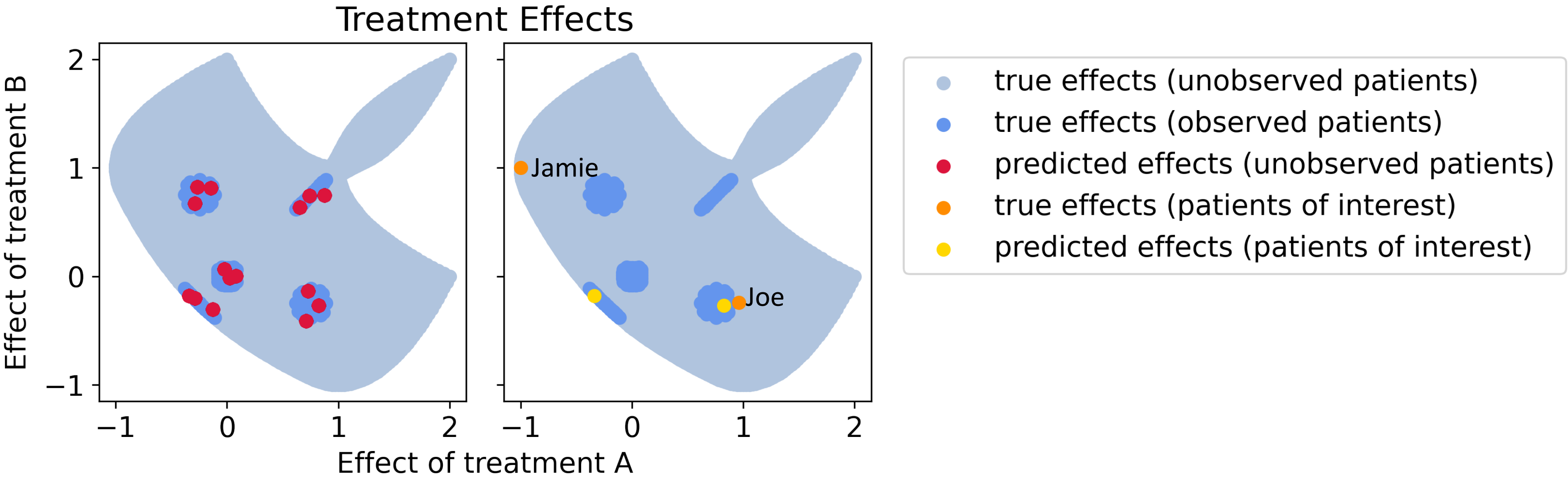
Massachusetts Institute of Technology

Computer Science and Artificial Intelligence Lab CSAIL

BROAD INSTITUTE

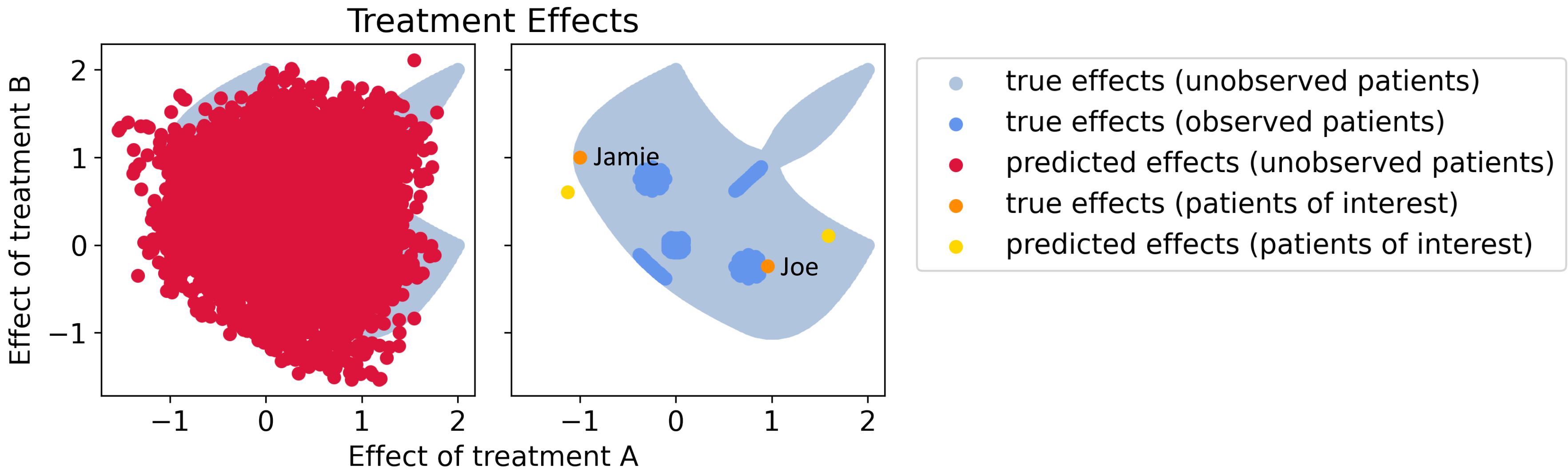
Cluster-Based Models: Limited Heterogeneity

Cluster C, then for each cluster solve $Y_c = X_c\,\hat{\beta}_c + \hat{\mu}_c$



Implicit Models: Unorganized

$$Y = \Phi(C, X) \rightarrow \hat{\beta} = \frac{\partial \Phi}{\partial X}$$



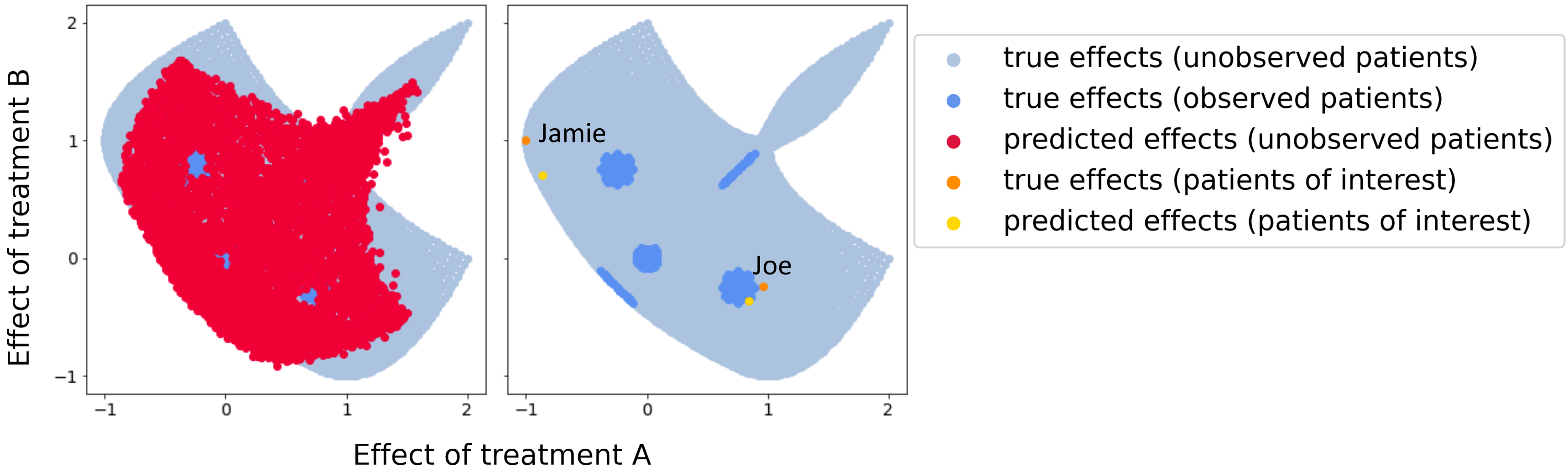
Massachusetts Institute of Technology

Computer Science and Artificial Intelligence Lab
CSAIL

BROAD INSTITUTE

Contextualized Models: Generalizability by Learning Latent Structure

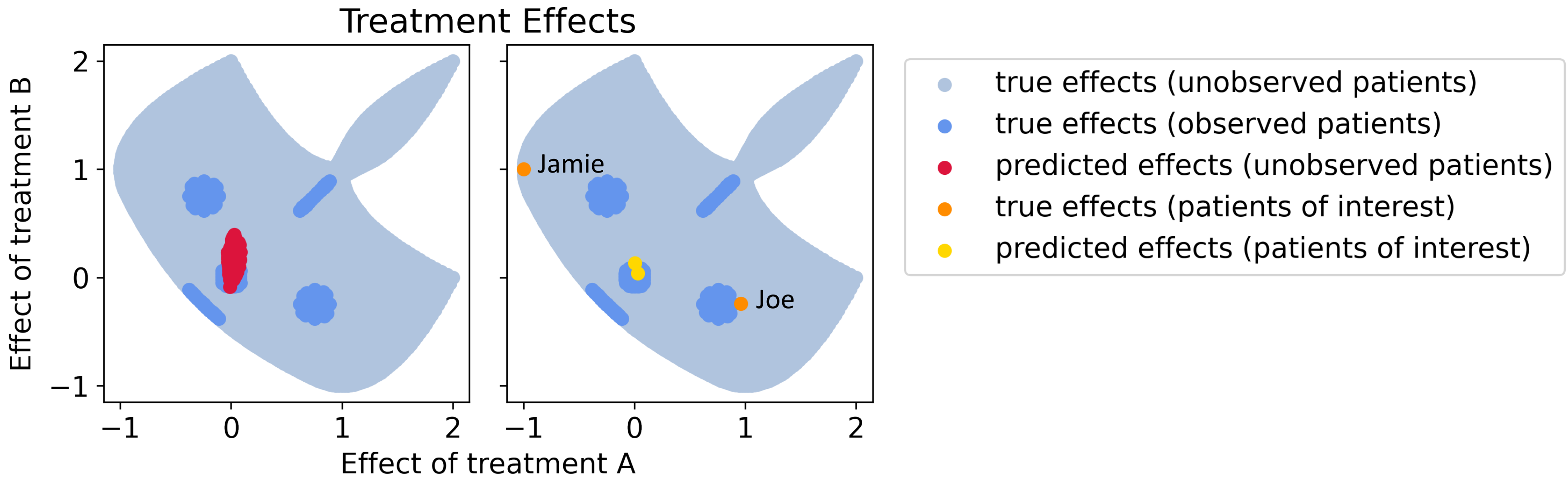
$$Y = X\ \beta_{\Phi}(C,\ \epsilon) + \mu_{\Phi}(C,\ \epsilon)$$



Massachusetts Institute of Technology
Computer Science and Artificial Intelligence Lab
CSAIL
BROAD INSTITUTE

In the worst case, context encoders recapitulate the population model

$$Y = X\,\beta_\Phi(\epsilon) + \mu_\Phi(\epsilon) \rightarrow Y = X\hat{\beta} + \hat{\mu}$$



Massachusetts Institute of Technology
Computer Science and Artificial Intelligence Lab
CSAIL
BROAD INSTITUTE

Contextualized

Heterogeneous Modeling Toolbox

contextualized.ml

```
pip install contextualized-ml
```

```
from contextualized.easy import ContextaulizedRegressor
model = ContextualizedRegressor()
model.fit(C, X, Y)
```

PyTorch

How to: Contextualizing Models with Deep Learning

1. Define a differentiable objective for your model of interest

$$\hat{\theta} = argmin_{\theta}\ loss(X, \theta)$$

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

2. Replace model parameters with a differentiable context encoder

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

$$[\hat{\theta}_1, \ \dots, \hat{\theta}_n] = \hat{\Phi}(C)$$

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

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

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

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

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

$K \ll |\theta|$

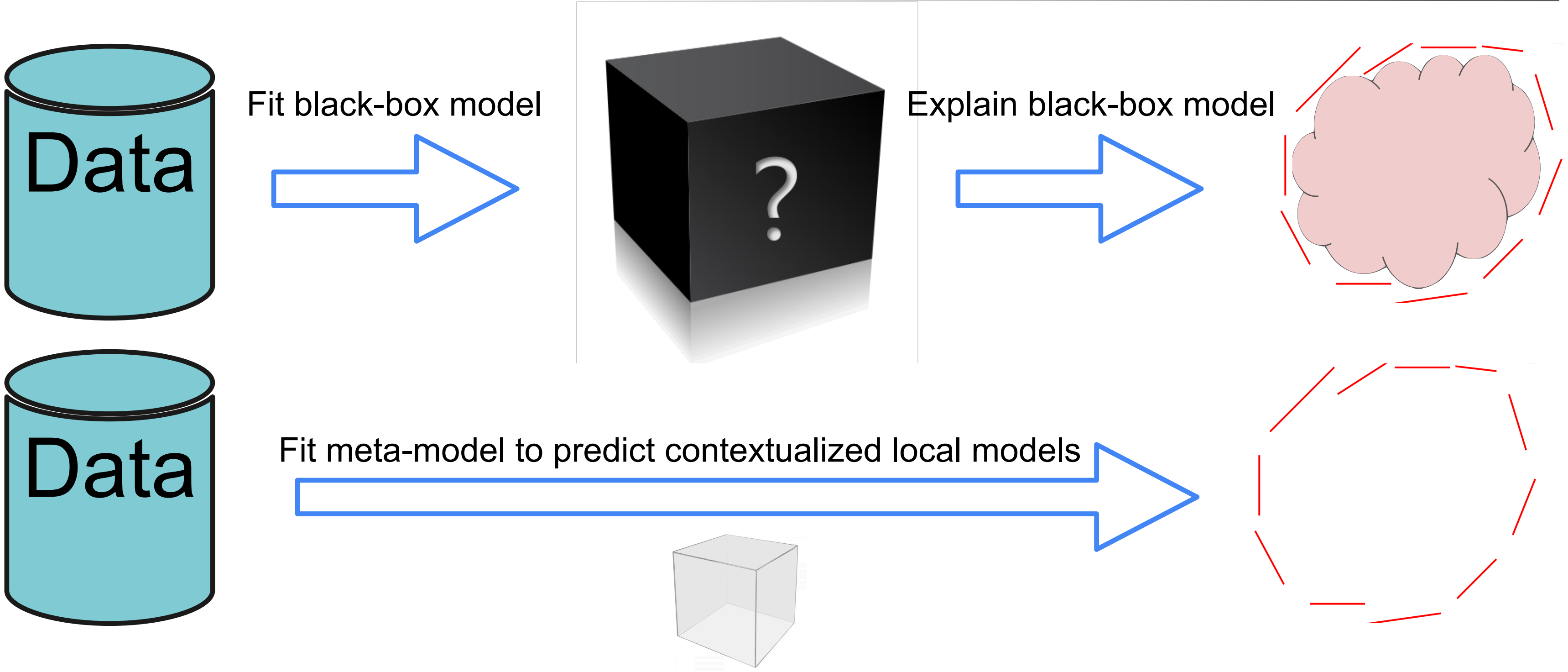
$A \in \mathbb{R}^{K \times |\theta|}$

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

4. Learn with your favorite auto-differentiation library

Massachusetts Institute of Technology | CSAIL Computer Science and Artificial Intelligence Lab | BROAD INSTITUTE

Contextual Meta-Models Generate Interpretable Local Models

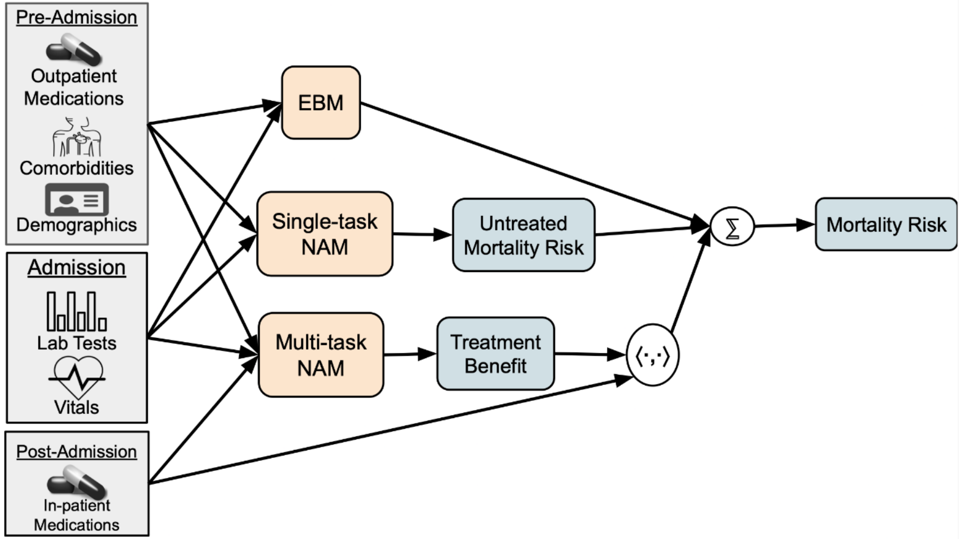


Vignettes of Contextualized Machine Learning

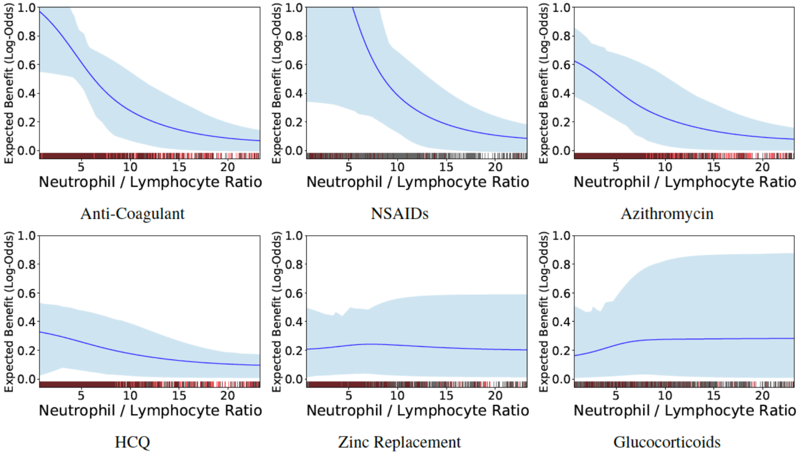
Contextualized Treatment Benefits in Covid-19

Tree-based EBMs are **great** at modeling healthcare data, but not differentiable. Can we combine EBM benefits with contextualized treatment estimation?

Solution:



Reveals that treatment effectiveness changes based on inflammation and thrombosis factors:

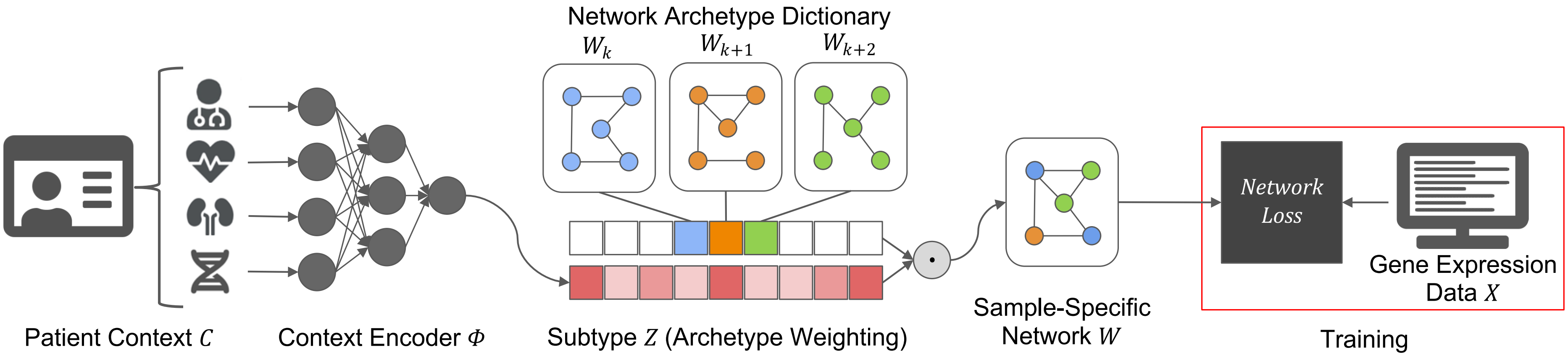


[Lengerich JBI 2022]

Massachusetts Institute of Technology
Computer Science and Artificial Intelligence Lab CSAIL
BROAD INSTITUTE

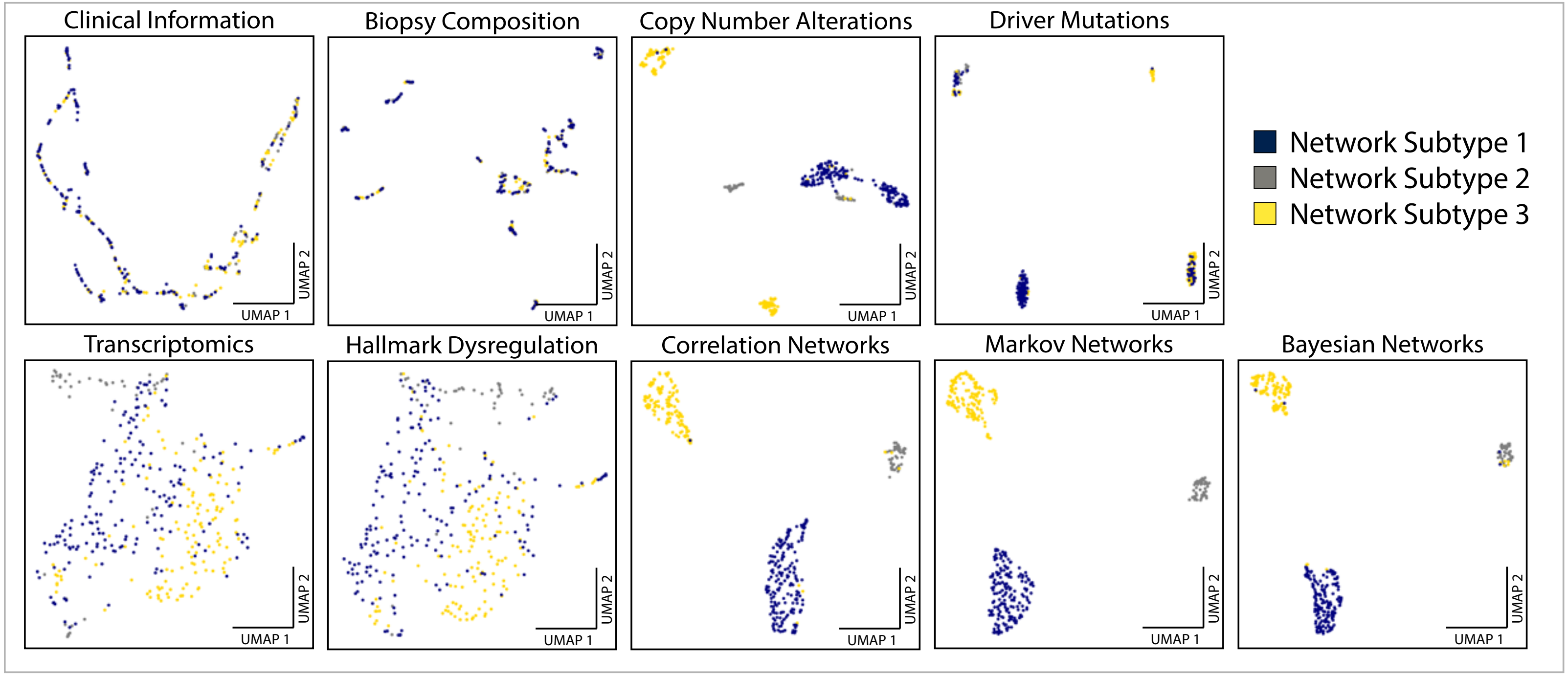
Contextualized Gene Regulatory Networks

Learning Sample-Specific Contextualized Graphical Models



Contextualized Gene Regulatory Networks in Cancer

Brain Glioma Embeddings Reveal Network-Based Subtypes

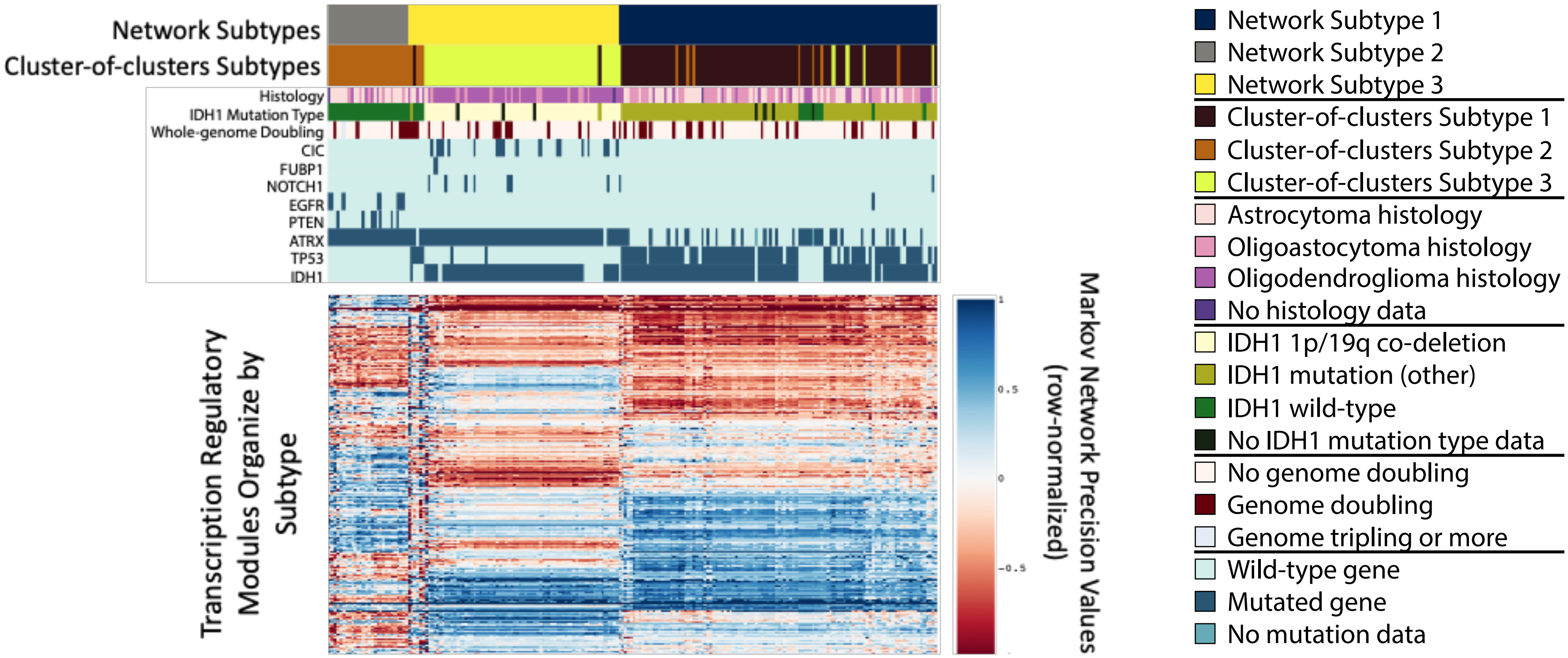


Massachusetts Institute of Technology

Computer Science and Artificial Intelligence Lab CSAIL

BROAD INSTITUTE

Networks Organize Into Subtypes in Cancer



Massachusetts Institute of Technology

Computer Science and Artificial Intelligence Lab CSAIL

BROAD INSTITUTE

Key Takeaways

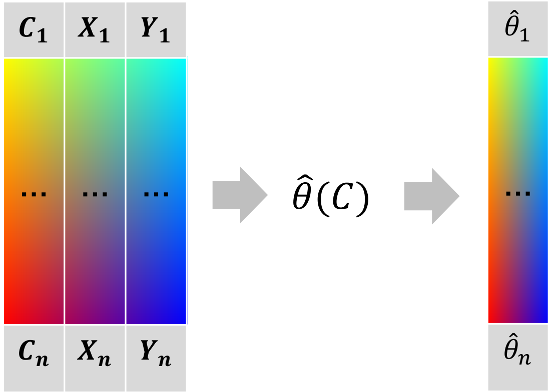
- Contextualized Machine Learning learns *meta*-models that generate parameters from context.
- All differentiable machine learning models can be expressed as contextualized models.
- Available in PyTorch sklearn-like API:

contextualized.ml

Contextualized
Heterogeneous Modeling Toolbox

Contextualized Modeling

parameters as functions of context



Massachusetts Institute of Technology
Computer Science and Artificial Intelligence Lab
CSAIL
BROAD INSTITUTE

Thank you!

- Ben Lengerich - blengeri@mit.edu
- Manolis Kellis - manoli@mit.edu
- Kellis Lab @MIT CSAIL

Contextualized
Heterogeneous Modeling Toolbox

contextualized.ml

Demos at contextualized.ml/docs

Massachusetts Institute of Technology

Computer Science and Artificial Intelligence Lab
CSAIL

BROAD INSTITUTE