MLSS 2020

Machine learning for healthcare

Prof. Mihaela van der Schaar

Email: mv472@cam.ac.uk

Website: http://www.vanderschaar-lab.com/

Why is ML for healthcare different?

ML has accomplished wonders …. on well-posed problems where the notion of a "solution" is well-defined and solutions are verifiable

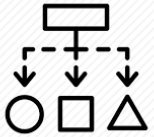
Healthcare is different – problems are not well-posed and notion of a "solution" is often not well-defined and solutions are hard to verify

This presents enormous challenges – and also enormous opportunities

Challenges

1. Augment clinical decision making

2. Support and inspire clinical discovery

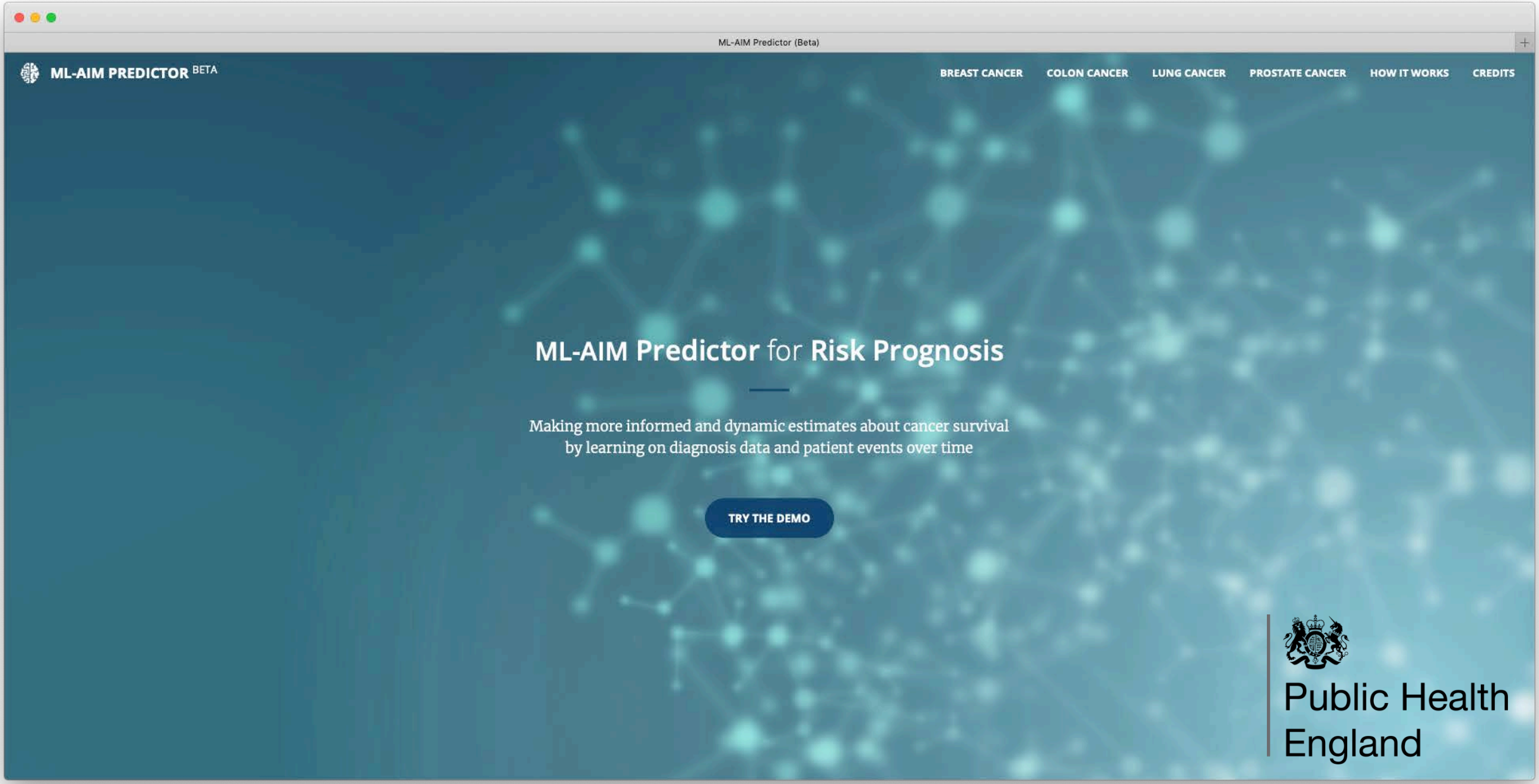


3. New problem formulations, new ML models and techniques

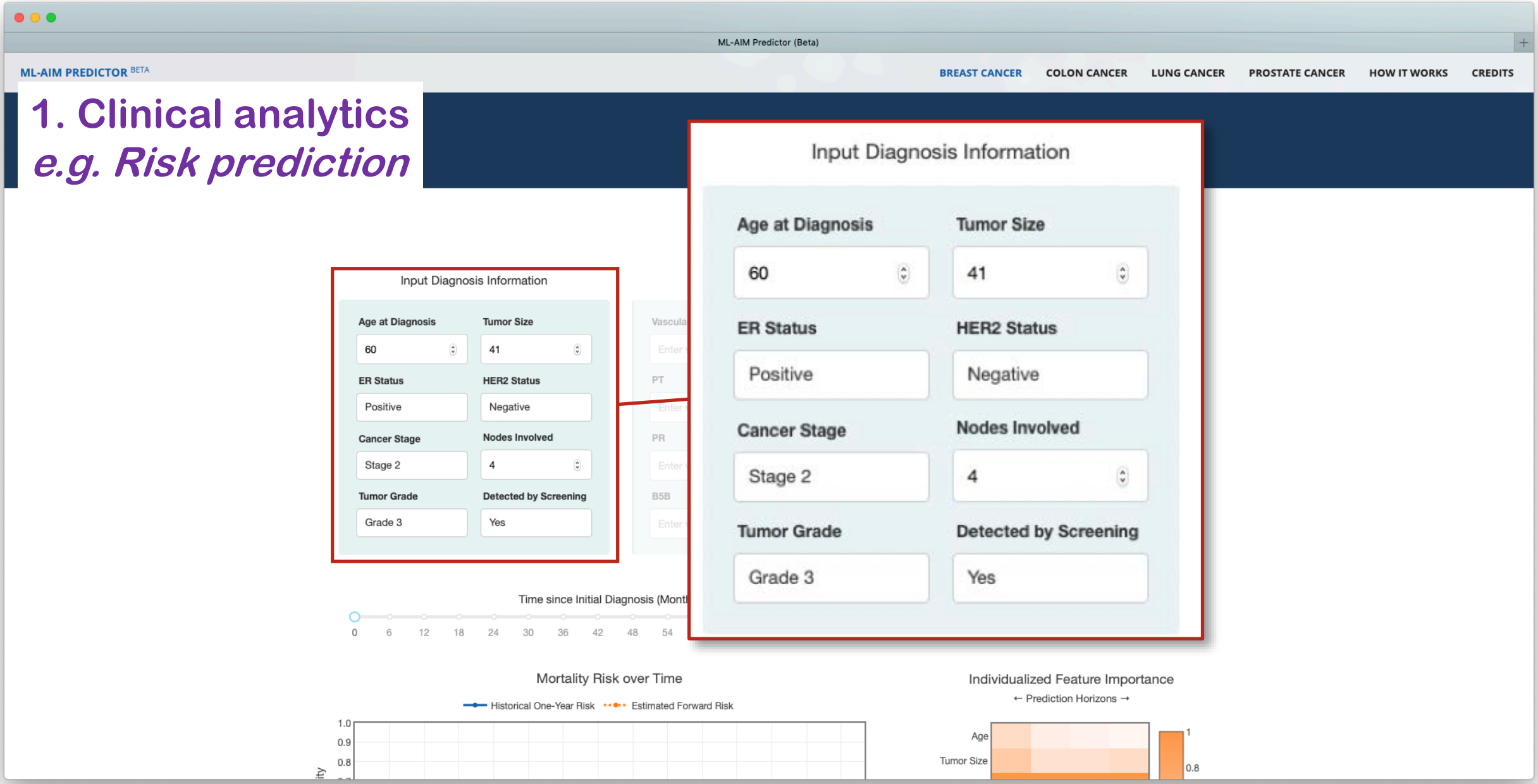
Focus of this tutorial!

Opportunities: ML can transform healthcare

1) **deliver** precision medicine at the patient-level
2) **understand** the basis and trajectories of health and disease
3) **inform and improve** clinical pathways, better utilize resources & reduce costs
4) **transform** population health and public health policy


ML-AIM Predictor (Beta)
ML-AIM PREDICTOR BETA
BREAST CANCER
COLON CANCER
LUNG CANCER
PROSTATE CANCER
HOW IT WORKS
CREDITS
ML-AIM Predictor for Risk Prognosis
Making more informed and dynamic estimates about cancer survival
by learning on diagnosis data and patient events over time
TRY THE DEMO
Public Health
England

1. Clinical analytics
e.g. Risk prediction



ML-AIM Predictor (Beta)

ML-AIM PREDICTOR BETA

BREAST CANCER | COLON CANCER | LUNG CANCER | PROSTATE CANCER | HOW IT WORKS | CREDITS

Input Diagnosis Information

Age at Diagnosis: 60
Tumor Size: 41
ER Status: Positive
HER2 Status: Negative
Cancer Stage: Stage 2
Nodes Involved: 4
Tumor Grade: Grade 3
Detected by Screening: Yes

Input Pathology Information

Vascular Invasion: N
Distance to Resection: 7
PT: X
PN: pn0
PR: N
Ki67: X
B5B: X
E-cadherin: X

or, Upload Pathology Report

marking axillary tail. At
approximately 12 o'clock position needle tract with surrounding fat
is identified. There is some fibrosis and this area extends up to
around 60mm and lies 15mm from the deep margin, 20mm from superior
and 60mm from inferior margin.
Part C - labelled ""Left sentinel node #2"".
A lymph node with surrounding fat measuring

Drag-and-Drop or Select File

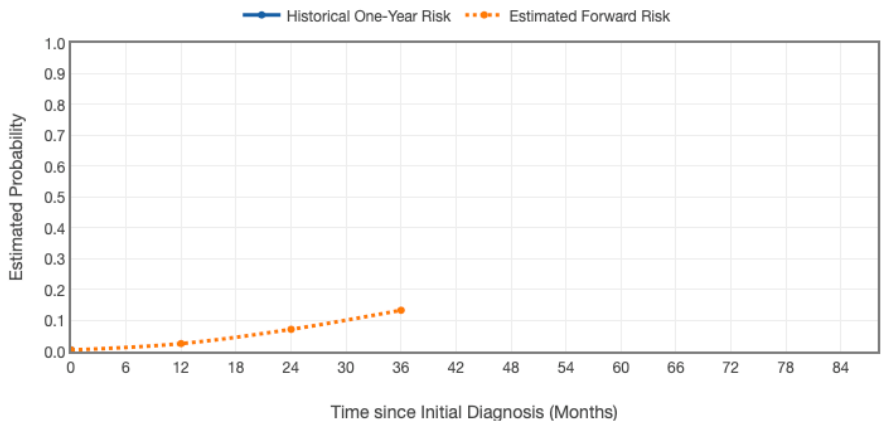
Time since Initial Diagnosis (Months)

0 6 12 18 24 30 36 42 48 54 60 66 72 78 84

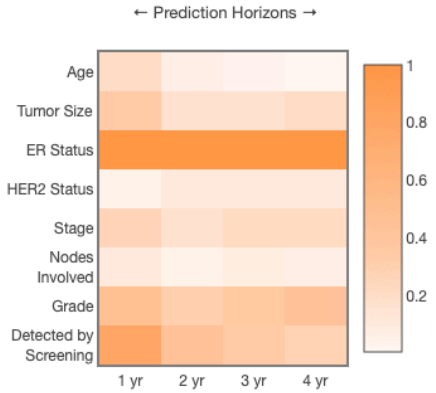
Patient 3

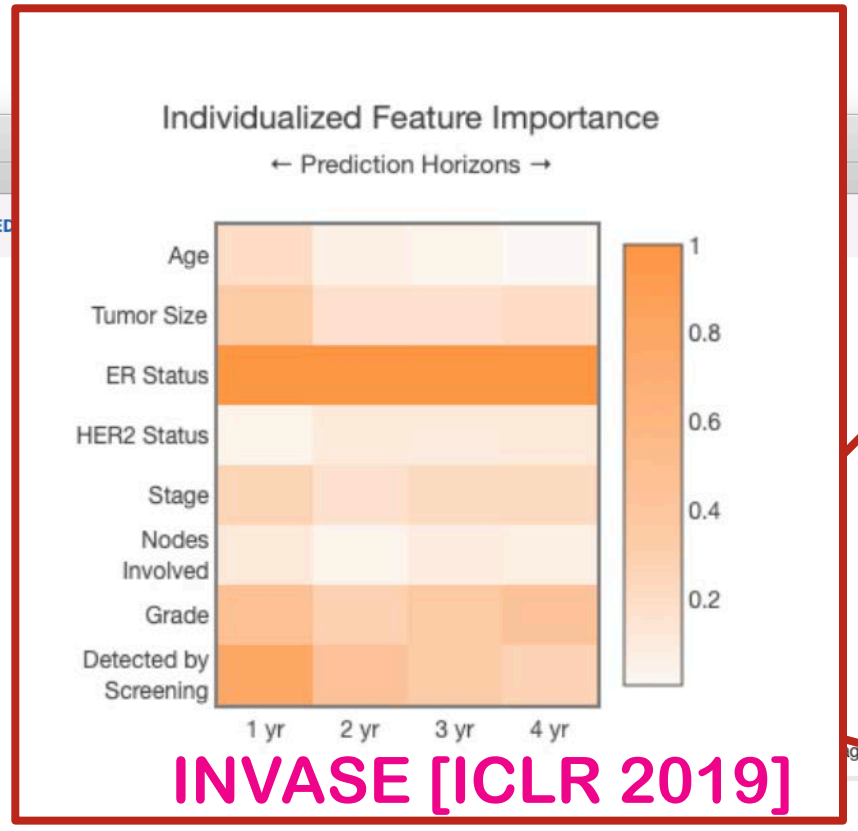
AutoPrognosis [ICML 2018]

Mortality Risk over Time

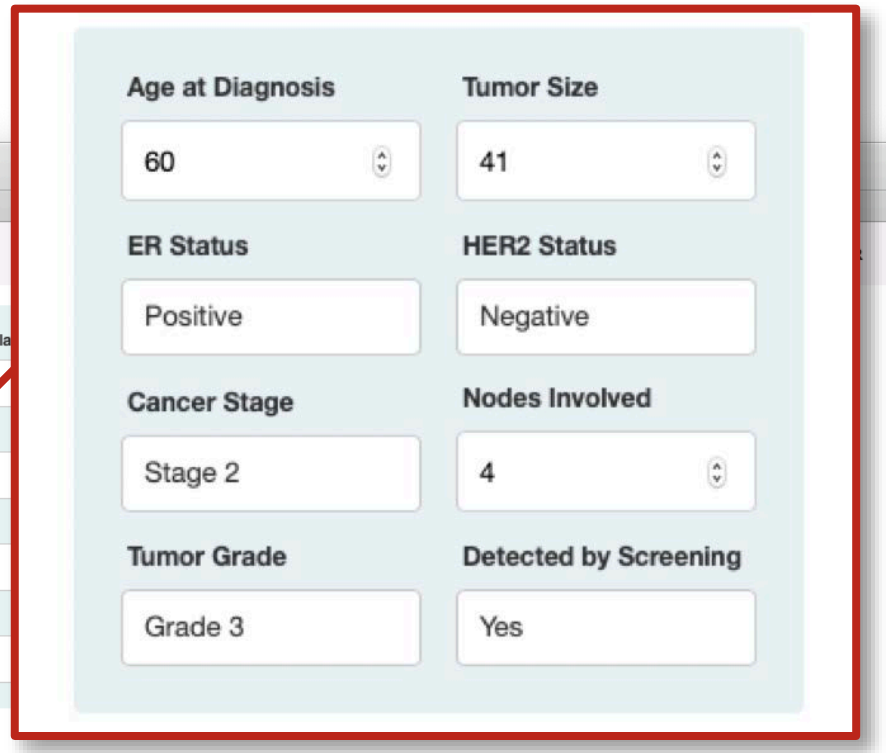


Individualized Feature Importance

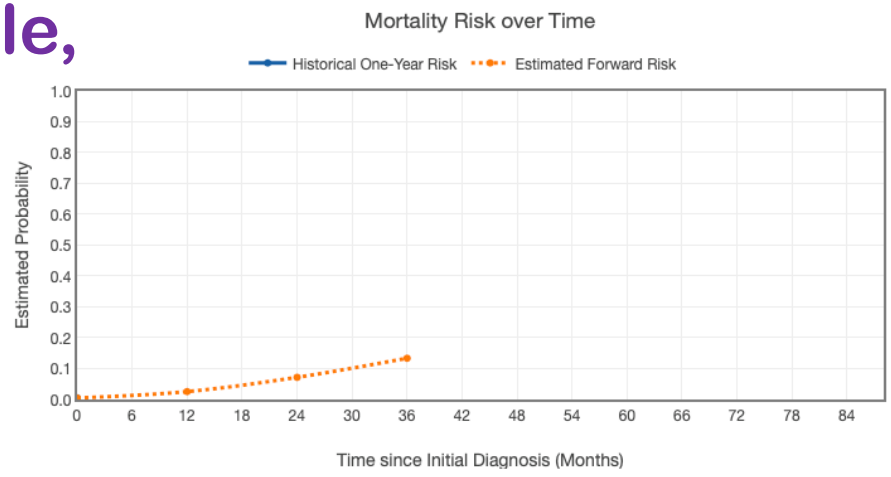


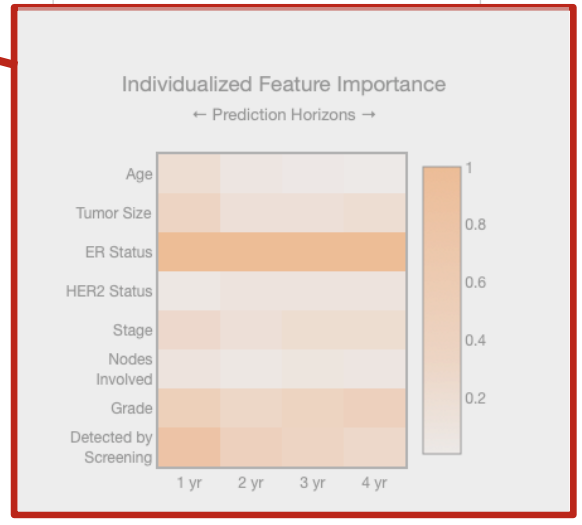


INVASE [ICLR 2019]

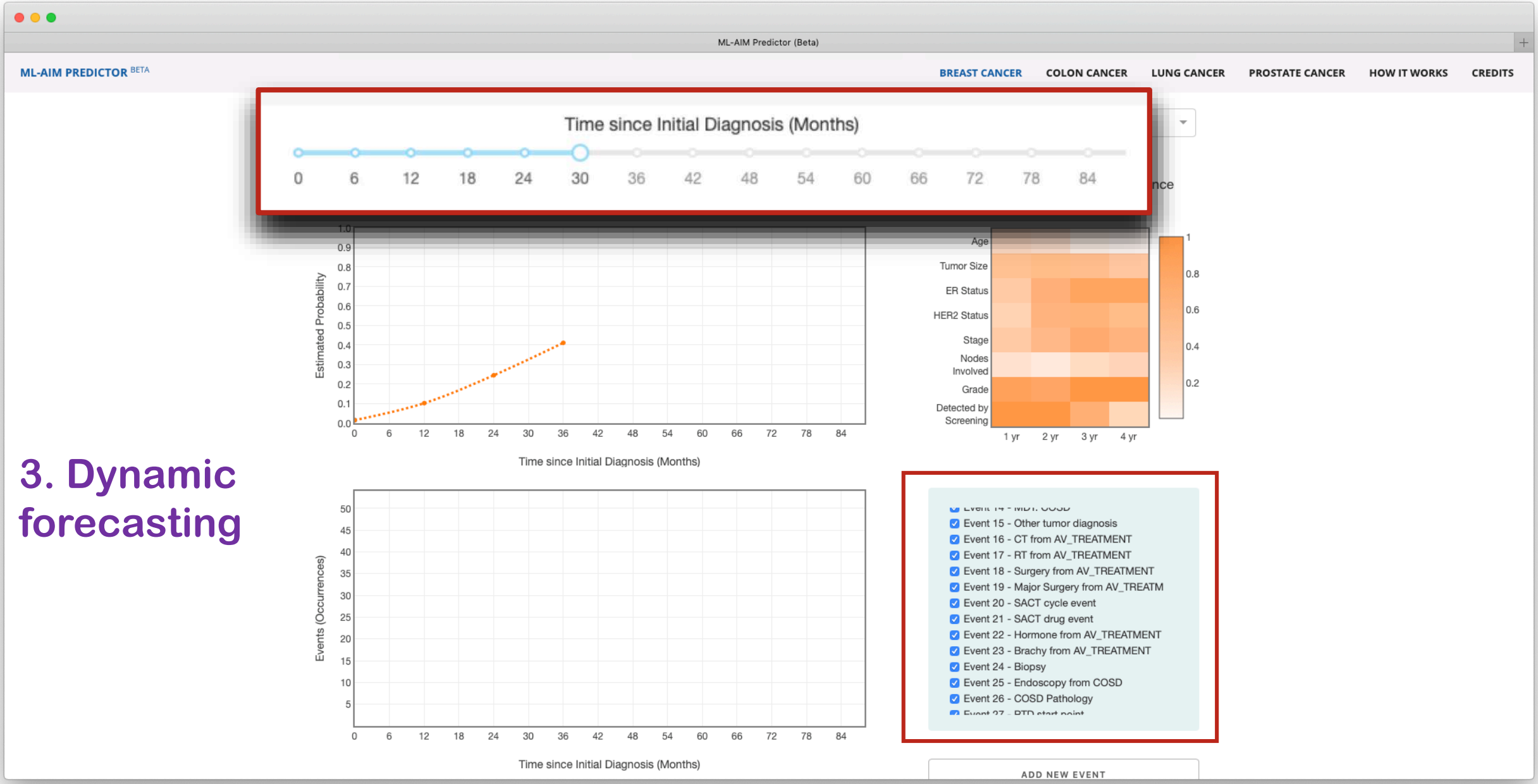


2. Interpretable, explainable & trustworthy

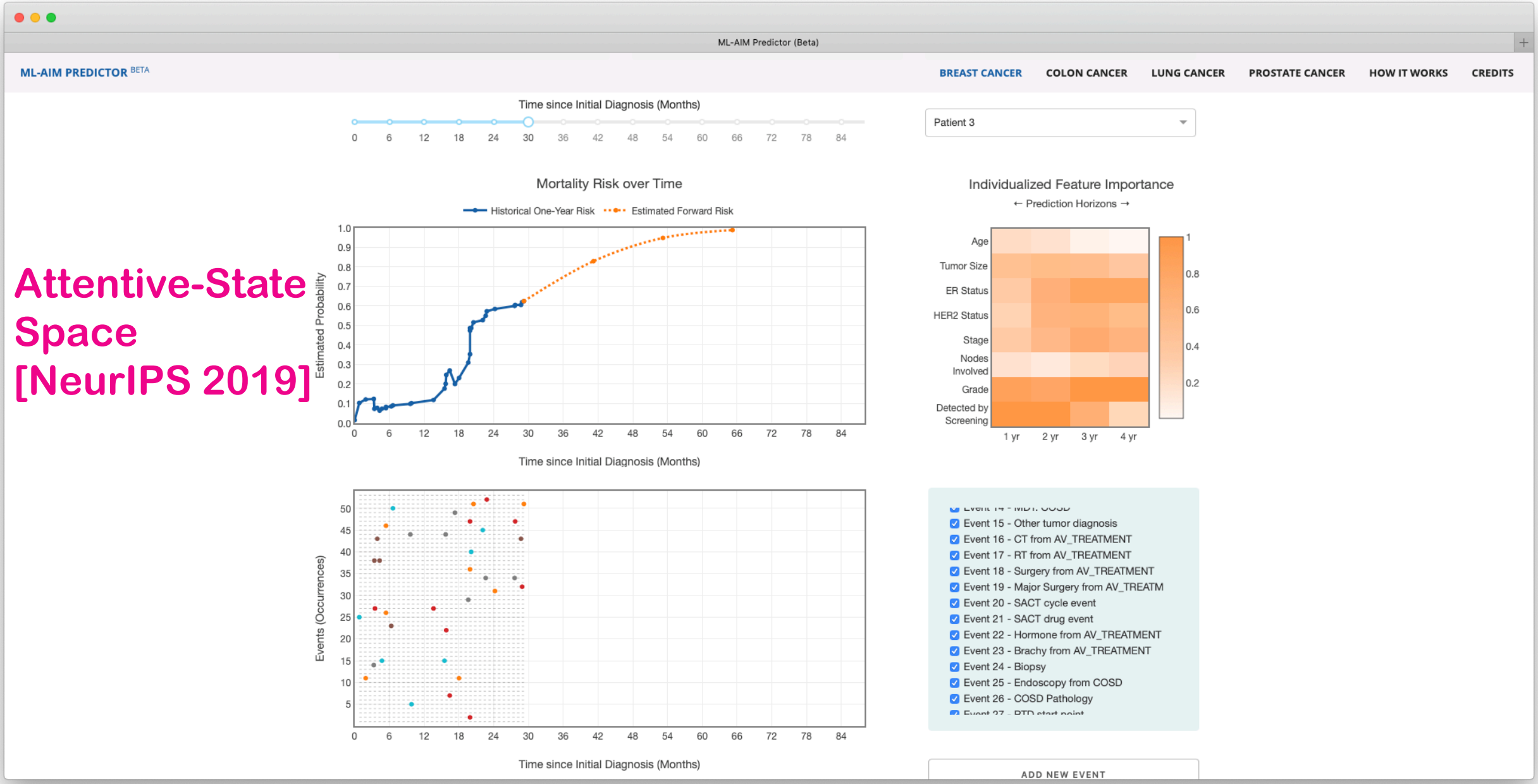




3. Dynamic forecasting



Attentive-State Space [NeurIPS 2019]



ML-AIM Predictor (Beta)

ML-AIM PREDICTOR BETA

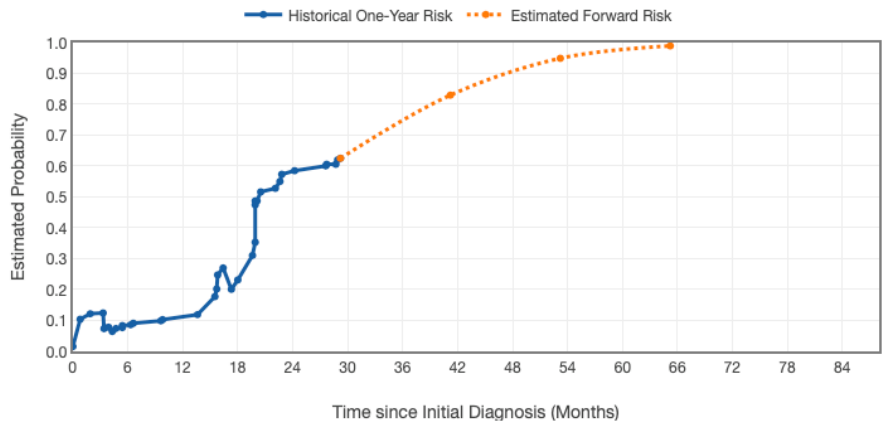
BREAST CANCER | COLON CANCER | LUNG CANCER | PROSTATE CANCER | HOW IT WORKS | CREDITS

Time since Initial Diagnosis (Months)

0 6 12 18 24 30 36 42 48 54 60 66 72 78 84

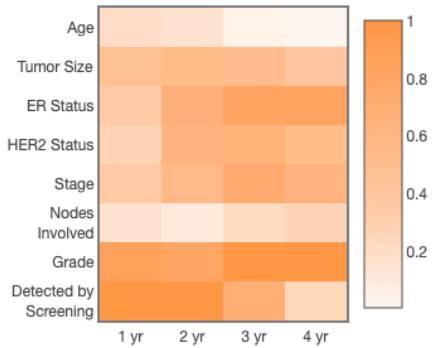
Patient 3

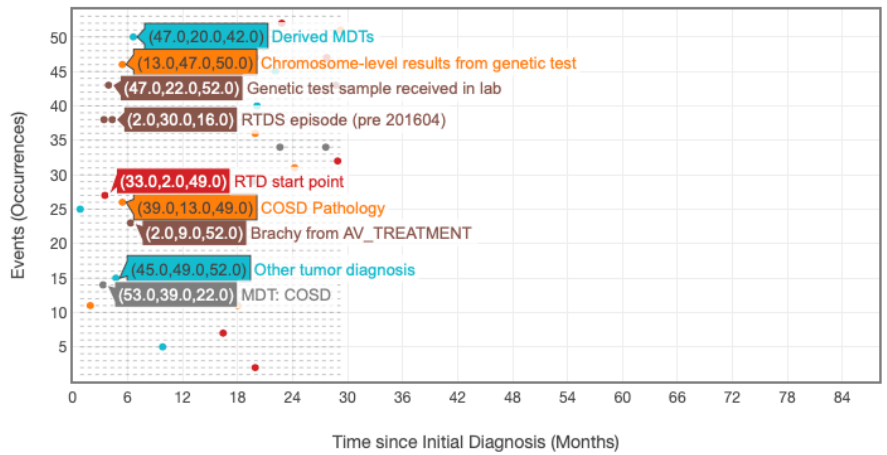
Mortality Risk over Time



Individualized Feature Importance

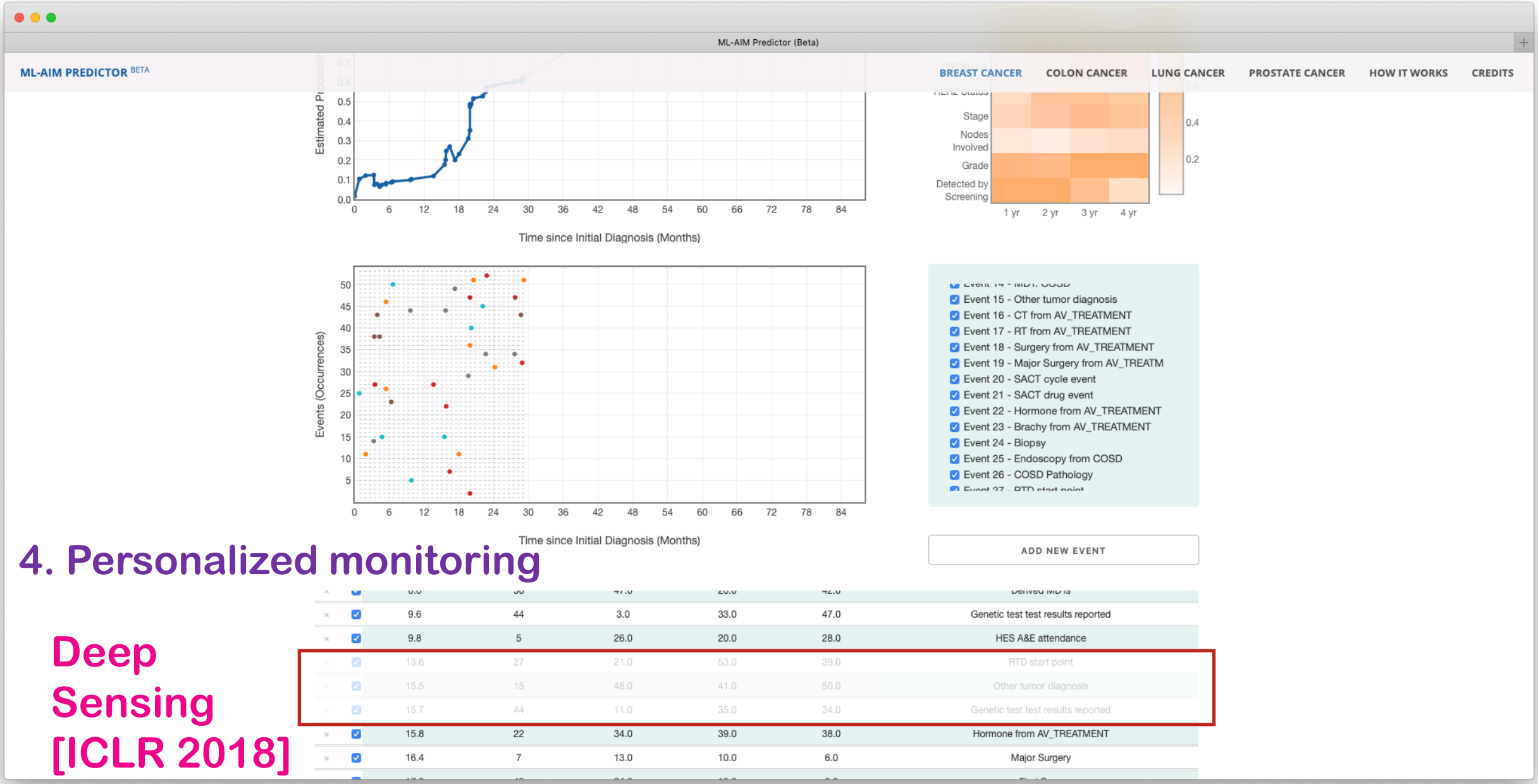
← Prediction Horizons →





- Event 15 - Other tumor diagnosis
- Event 16 - CT from AV_TREATMENT
- Event 17 - RT from AV_TREATMENT
- Event 18 - Surgery from AV_TREATMENT
- Event 19 - Major Surgery from AV_TREATM
- Event 20 - SACT cycle event
- Event 21 - SACT drug event
- Event 22 - Hormone from AV_TREATMENT
- Event 23 - Brachy from AV_TREATMENT
- Event 24 - Biopsy
- Event 25 - Endoscopy from COSD
- Event 26 - COSD Pathology

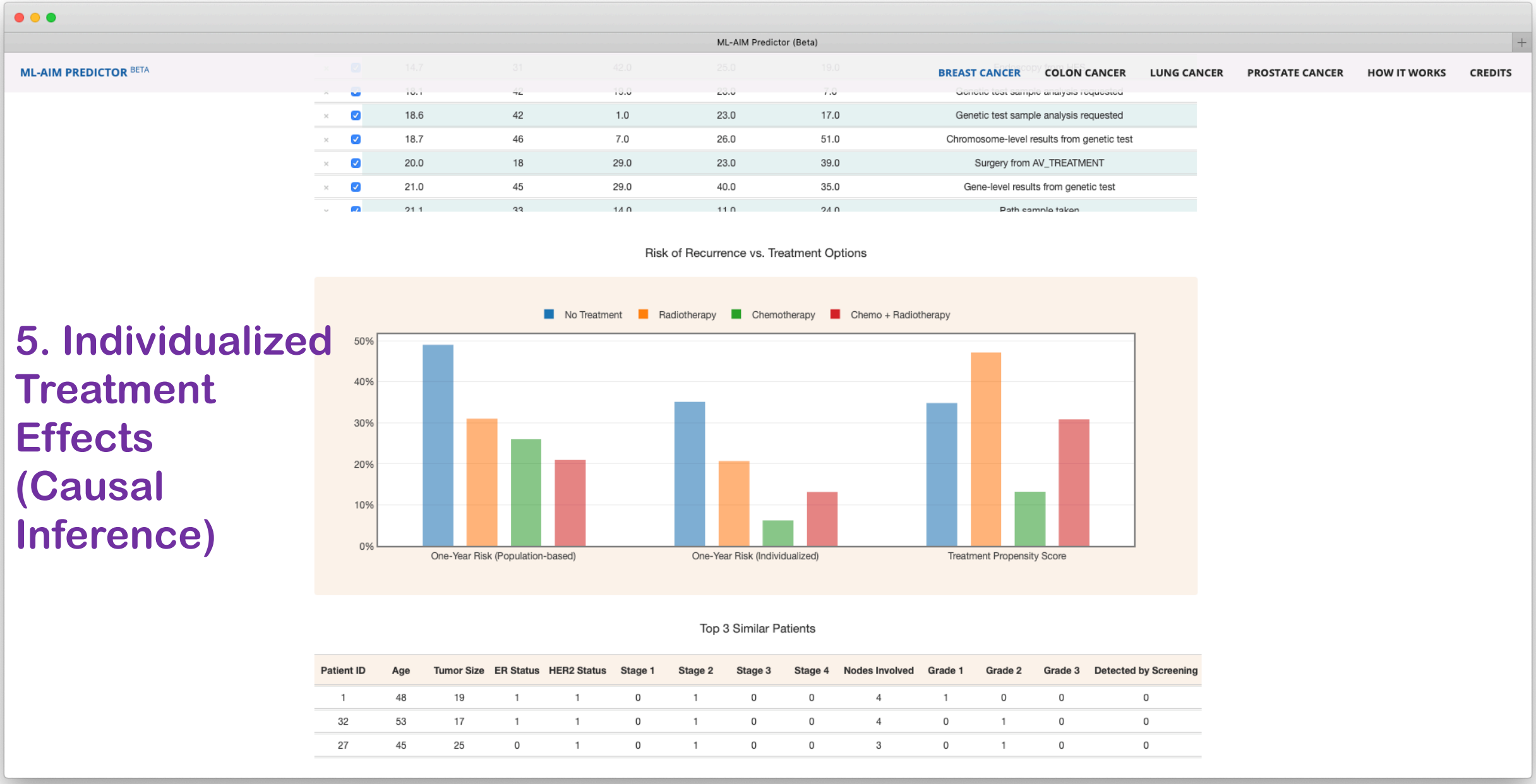
ADD NEW EVENT



4. Personalized monitoring

Deep Sensing [ICLR 2018]

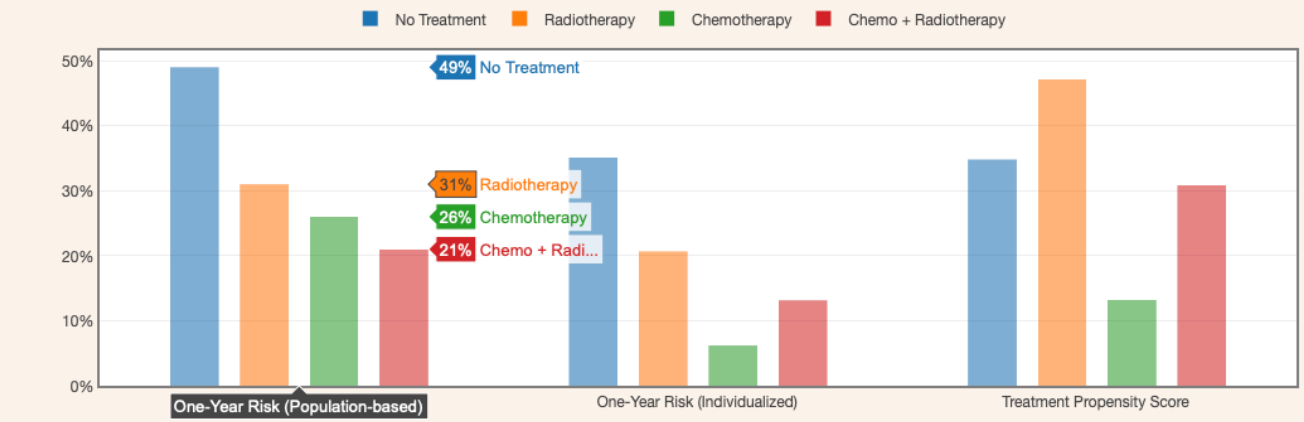
5. Individualized Treatment Effects (Causal Inference)



Patient ID	Age	Tumor Size	ER Status	HER2 Status	Stage 1	Stage 2	Stage 3	Stage 4	Nodes Involved	Grade 1	Grade 2	Grade 3	Detected by Screening
1	48	19	1	1	0	1	0	0	4	1	0	0	0
32	53	17	1	1	0	1	0	0	4	0	1	0	0
27	45	25	0	1	0	1	0	0	3	0	1	0	0

×	☑	18.6	42	1.0	23.0	17.0	Genetic test sample analysis requested
×	☑	18.7	46	7.0	26.0	51.0	Chromosome-level results from genetic test
×	☑	20.0	18	29.0	23.0	39.0	Surgery from AV_TREATMENT
×	☑	21.0	45	29.0	40.0	35.0	Gene-level results from genetic test

Risk of Recurrence vs. Treatment Options

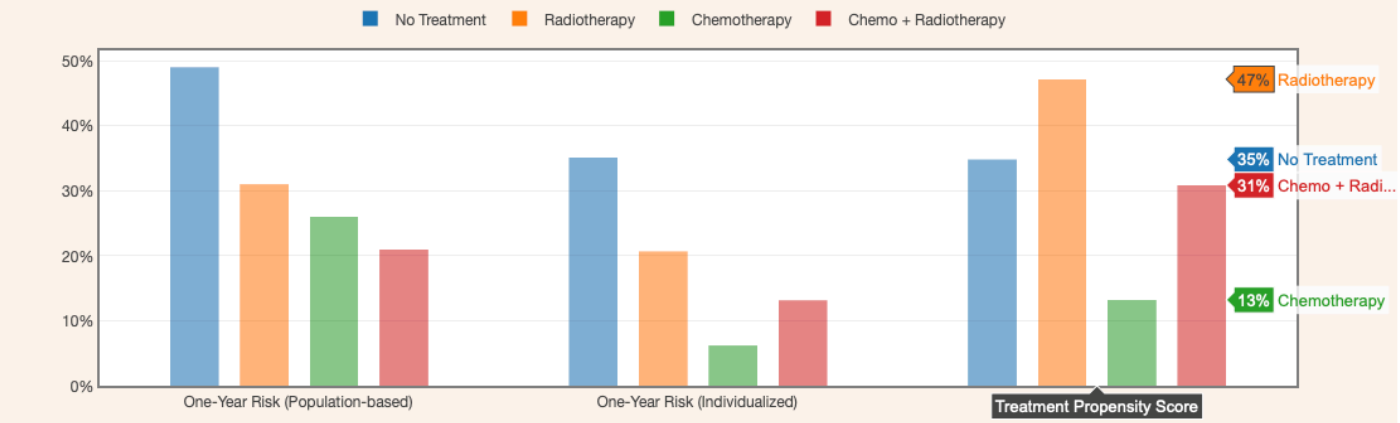


Top 3 Similar Patients

Patient ID	Age	Tumor Size	ER Status	HER2 Status	Stage 1	Stage 2	Stage 3	Stage 4	Nodes Involved	Grade 1	Grade 2	Grade 3	Detected by Screening
1	48	19	1	1	0	1	0	0	4	1	0	0	0
32	53	17	1	1	0	1	0	0	4	0	1	0	0
27	45	25	0	1	0	1	0	0	3	0	1	0	0

×	☑	18.6	42	1.0	23.0	17.0	Genetic test sample analysis requested
×	☑	18.7	46	7.0	26.0	51.0	Chromosome-level results from genetic test
×	☑	20.0	18	29.0	23.0	39.0	Surgery from AV_TREATMENT
×	☑	21.0	45	29.0	40.0	35.0	Gene-level results from genetic test

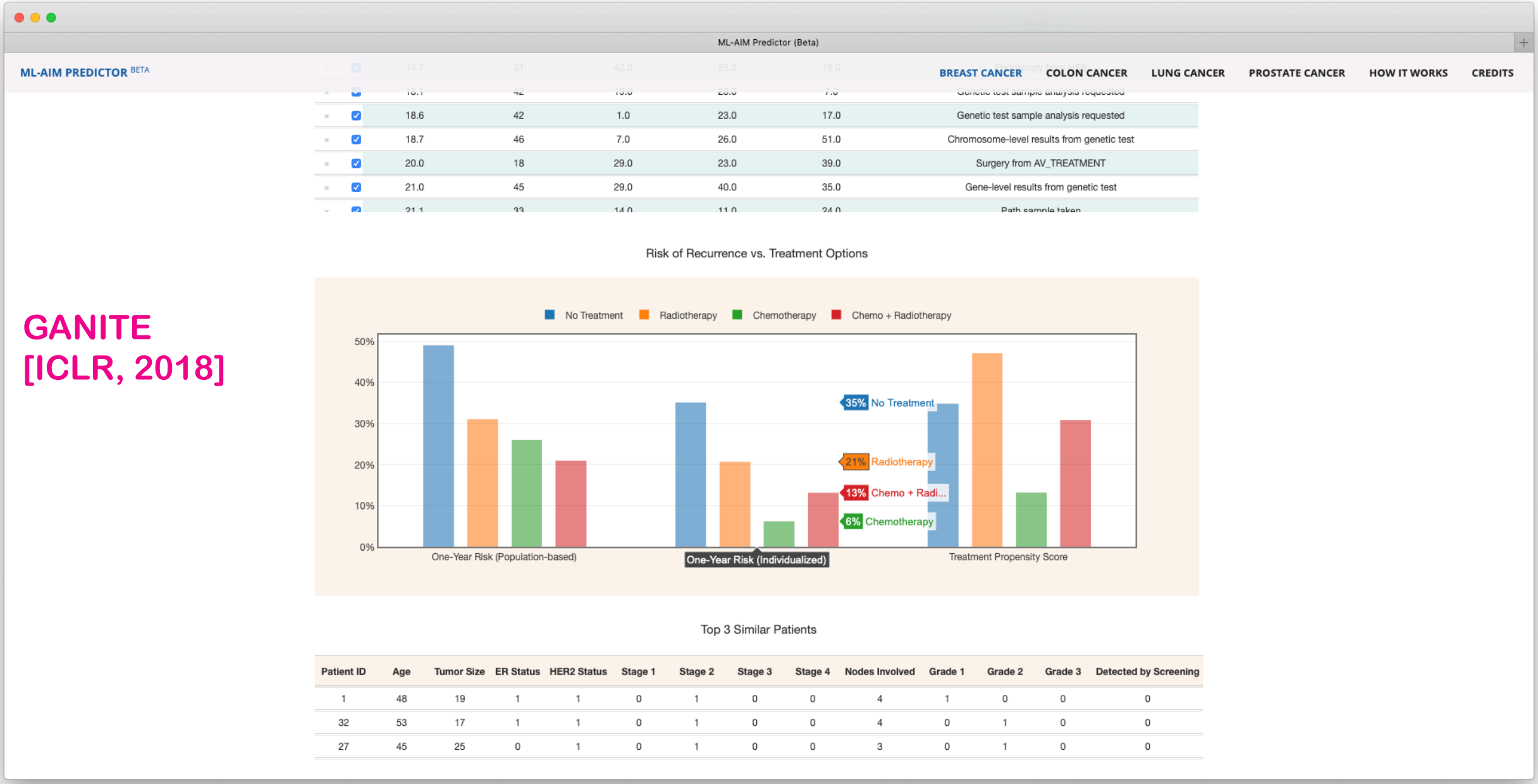
Risk of Recurrence vs. Treatment Options



Top 3 Similar Patients

Patient ID	Age	Tumor Size	ER Status	HER2 Status	Stage 1	Stage 2	Stage 3	Stage 4	Nodes Involved	Grade 1	Grade 2	Grade 3	Detected by Screening
1	48	19	1	1	0	1	0	0	4	1	0	0	0
32	53	17	1	1	0	1	0	0	4	0	1	0	0
27	45	25	0	1	0	1	0	0	3	0	1	0	0

GANITE
[ICLR, 2018]



Patient ID	Age	Tumor Size	ER Status	HER2 Status	Stage 1	Stage 2	Stage 3	Stage 4	Nodes Involved	Grade 1	Grade 2	Grade 3	Detected by Screening
1	48	19	1	1	0	1	0	0	4	1	0	0	0
32	53	17	1	1	0	1	0	0	4	0	1	0	0
27	45	25	0	1	0	1	0	0	3	0	1	0	0

Lecture plan

1. Clinical analytics at scale: AutoML
2. Interpretable, explainable and trustworthy ML
3. Dynamic forecasting
4. Personalized monitoring and screening
5. Individualized treatment effects

6. How can we get data? Synthetic data generation

Part 1: Clinical analytics *at scale*

How can we automate the process of building analytics?

Cardiovascular disease
- Risk of CVD events
- Mortality risk after heart-failure
- Mortality risk – Cardiac transplantation

Cancer: Breast, Prostate, Colon

Cystic Fibrosis

Asthma

Hospital care

Alzheimer's disease

ML for Risk Prediction in Healthcare

+ High predictive accuracy (for some datasets)
+ Data-driven, few assumptions
- Many algorithms: Which one to choose?
- Many hyper-parameters: Need expertise in data science

AUROC	MAGGIC	UK Biobank	UNOS-I	UNOS-II
Best predictor	0.80 ± 0.004	0.76 ± 0.002	0.78 ± 0.002	0.65 ± 0.001
	NN	GradientBoost	ToPs	ToPs
Best Clinical Score	0.70 ± 0.007	0.70 ± 0.003	0.62 ± 0.001	0.56 ± 0.001
Cox PH	0.75 ± 0.005	0.74 ± 0.002	0.70 ± 0.001	0.59 ± 0.001

ML for Risk Prediction in Healthcare

+ High predictive accuracy (for some datasets)
+ Data-driven, few assumptions
- Many algorithms: Which one to choose?
- Many hyper-parameters: Need expertise in data science

AUROC	MAGGIC	UK Biobank	UNOS-I	UNOS-II
Best predictor	0.80 ± 0.004	0.76 ± 0.002	0.78 ± 0.002	0.65 ± 0.001
	NN	GradientBoost	ToPs	ToPs
Best Clinical Score	0.70 ± 0.007	0.70 ± 0.003	0.62 ± 0.001	0.56 ± 0.001
Cox PH	0.75 ± 0.005	0.74 ± 0.002	0.70 ± 0.001	0.59 ± 0.001

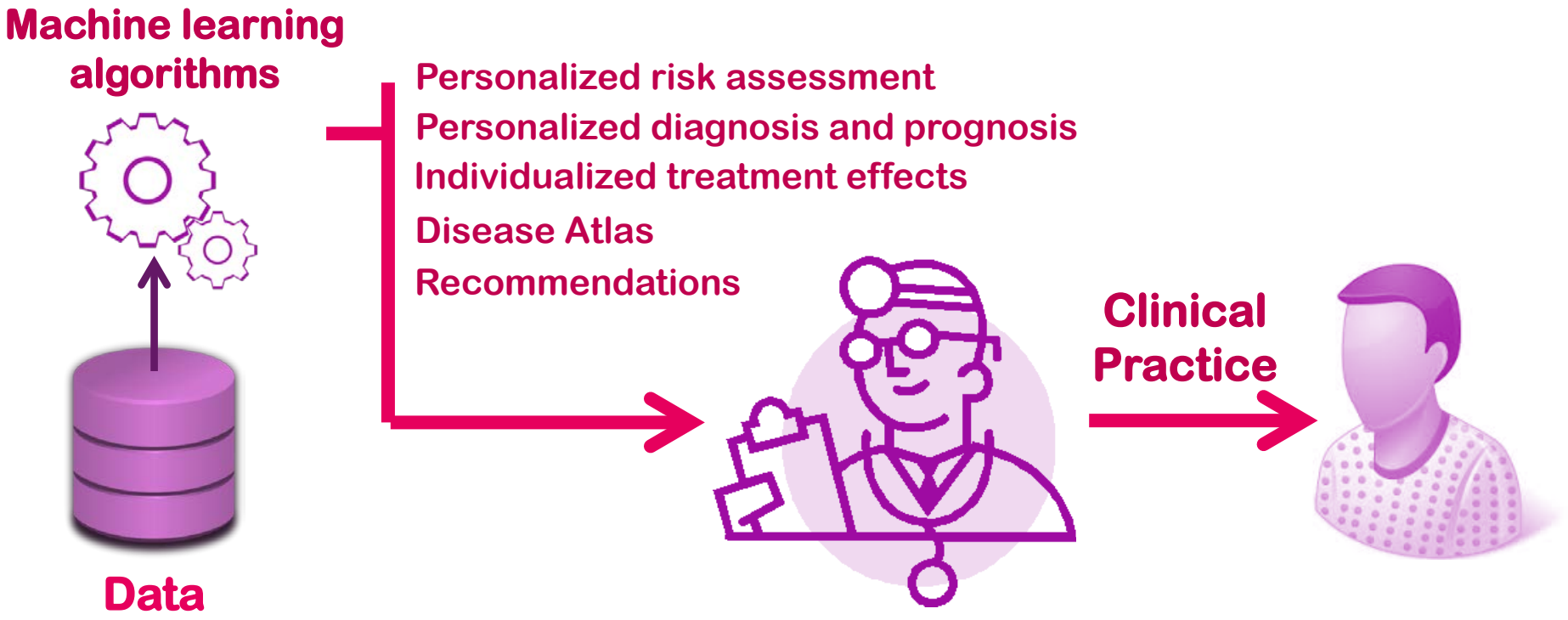
- Can we predict in advance which method is best?
- Can we do better?
- Many metrics of performance (AUROC, AUPRC, C-index, quality of well-being)

The "Augmented" MD

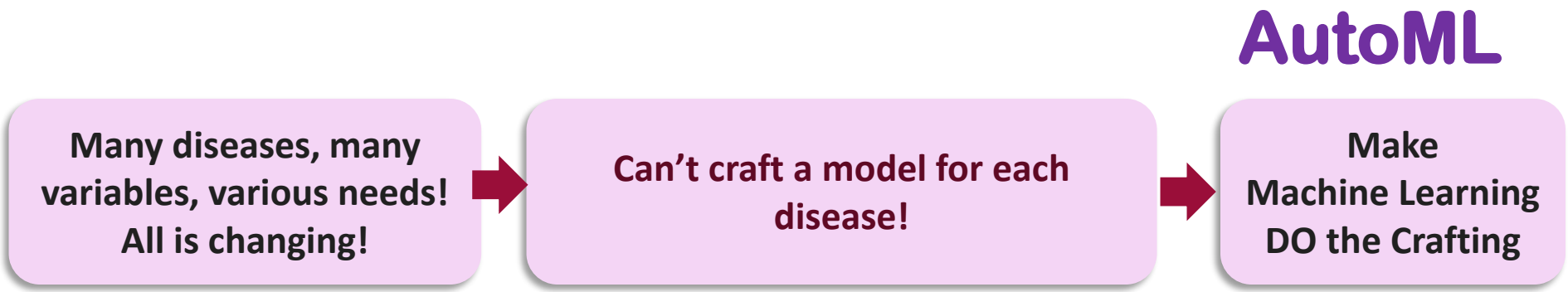
- **Machine learning**

  …***can't*** do medicine!

  …***can*** provide doctors with actionable information!



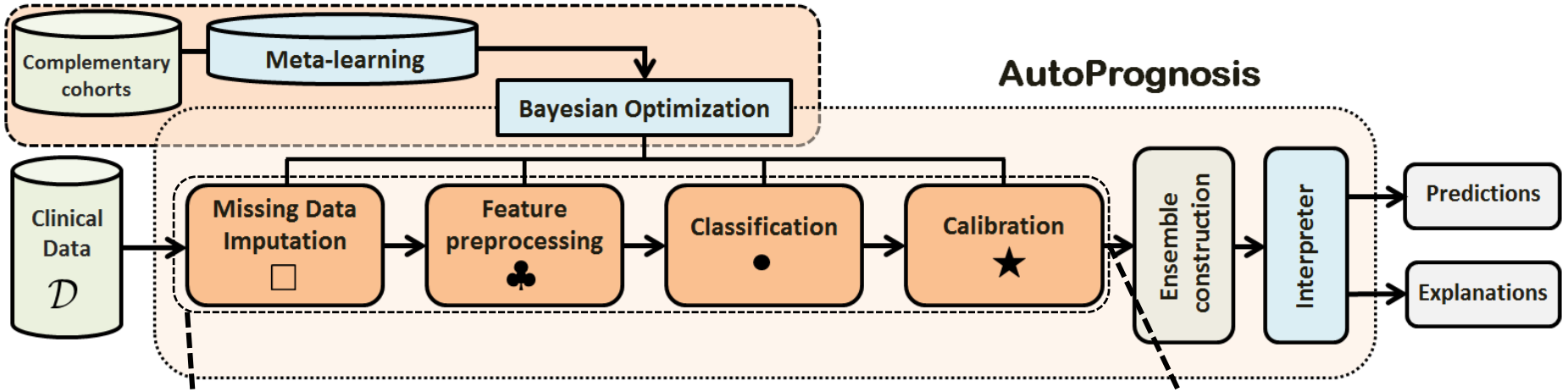
How?

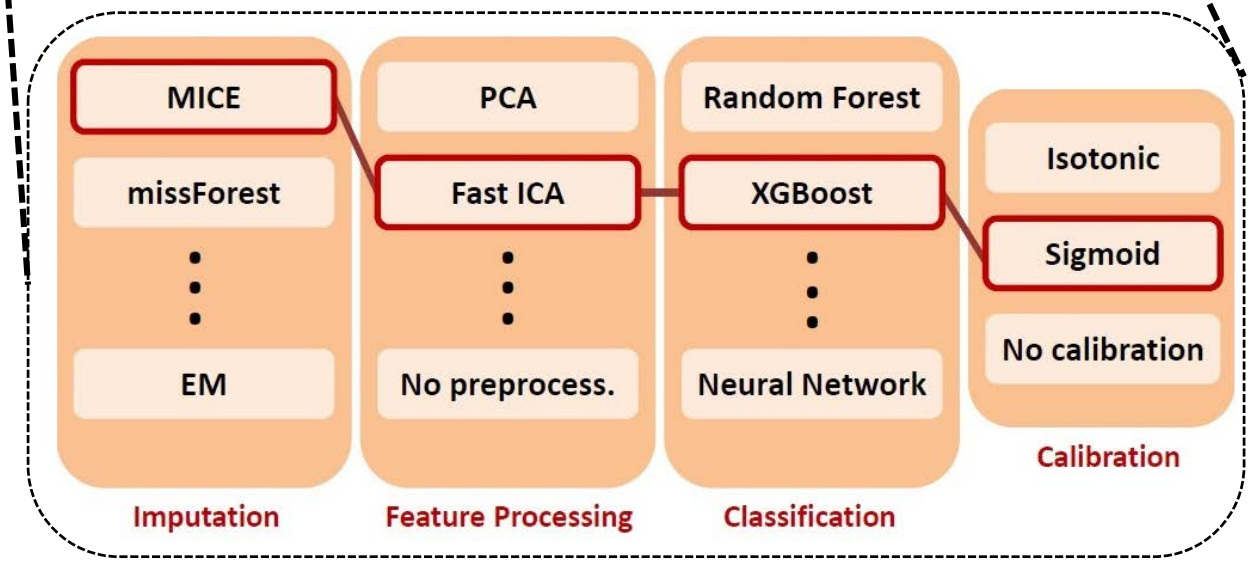


- **Previous AutoML? Auto-WEKA and Auto-Sklearn**
  - Limited performance gains
  - Meta-learning
  - Simplistic handling of missing data
  - Do not capture uncertainty
  - Limited to classification problems (survival, competing risks, time-series etc.)

AutoPrognosis [Alaa & vdS, ICML 2018]:

A tool for crafting Clinical Scores





We need an entire pipeline!

Each pipeline is a path of algorithms

- 8 imputation algorithms, 10 feature preprocessing algorithms, 20 classifiers, 3 calibration methods
- MANY hyperparameters in each algorithm
- Total number of hyperparameters = 110

Pipeline Stage	Algorithms				
□ Data Imputation	□ missForest (2) □ Matrix completion (2)	□ Median (0) □ MICE (1)	□ Most-frequent (0) □ GAIN	□ Mean (0) □ None (0)	□ EM (1)
♣ Feature process.	♣ Feature agglo. (4) ♣ R. kitchen sinks (2)	♣ Kernel PCA (5) ♣ Nystroem (5)	♣ Polynomial (3) ♣ Linear SVM (3)	♣ Fast ICA (4) ♣ Select Rates (3)	♣ PCA (2) ♣ None (0)
• Prediction	• Bernoulli NB (2) • Gaussian NB (0) • Multinomial NB (2) • Ridge Class. (1) • DMGP	• AdaBoost (4) • XGBoost (5) • R. Forest (5) • Bagging (4) • CMGP	• Decision Tree (4) • Extr. R. Trees (5) • Neural Net. (5) • k-NN (1) • DeepHit	• Grad. Boost. (6) • Light GBM (5) • Log. Reg. (0) • Surv. Forest (5) • HBM	• LDA (4) • L. SVM (4) • GP (3) • Cox Reg. (0) • TOPs
★ Calibration	★ Sigmoid (0)	★ Isotonic (0)	★ None (0)		

- Find the best paths and tune parameters:
  A hard optimization problem??

Automated Pipeline Configuration

- Imputation algorithms $\mathcal{A}_d$
  Hyperparameters Θ_d
- Feature process. algorithms $\mathcal{A}_f$
  Hyperparameters Θ_f
- Classification algorithms $\mathcal{A}_c$
  Hyperparameters Θ_c
- Calibration algorithms $\mathcal{A}_a$
  Hyperparameters Θ_a

- Set of all pipelines $\mathcal{P} = \mathcal{A}_d \times \mathcal{A}_f \times \mathcal{A}_c \times \mathcal{A}_a$
- Set of all hyperparameters $\Theta = \Theta_d \times \Theta_f \times \Theta_c \times \Theta_a$
- Set of all pipeline configurations $\mathcal{P}_\Theta$

- Combined Pipeline Selection and Hyperparameter (CPSH) optimization

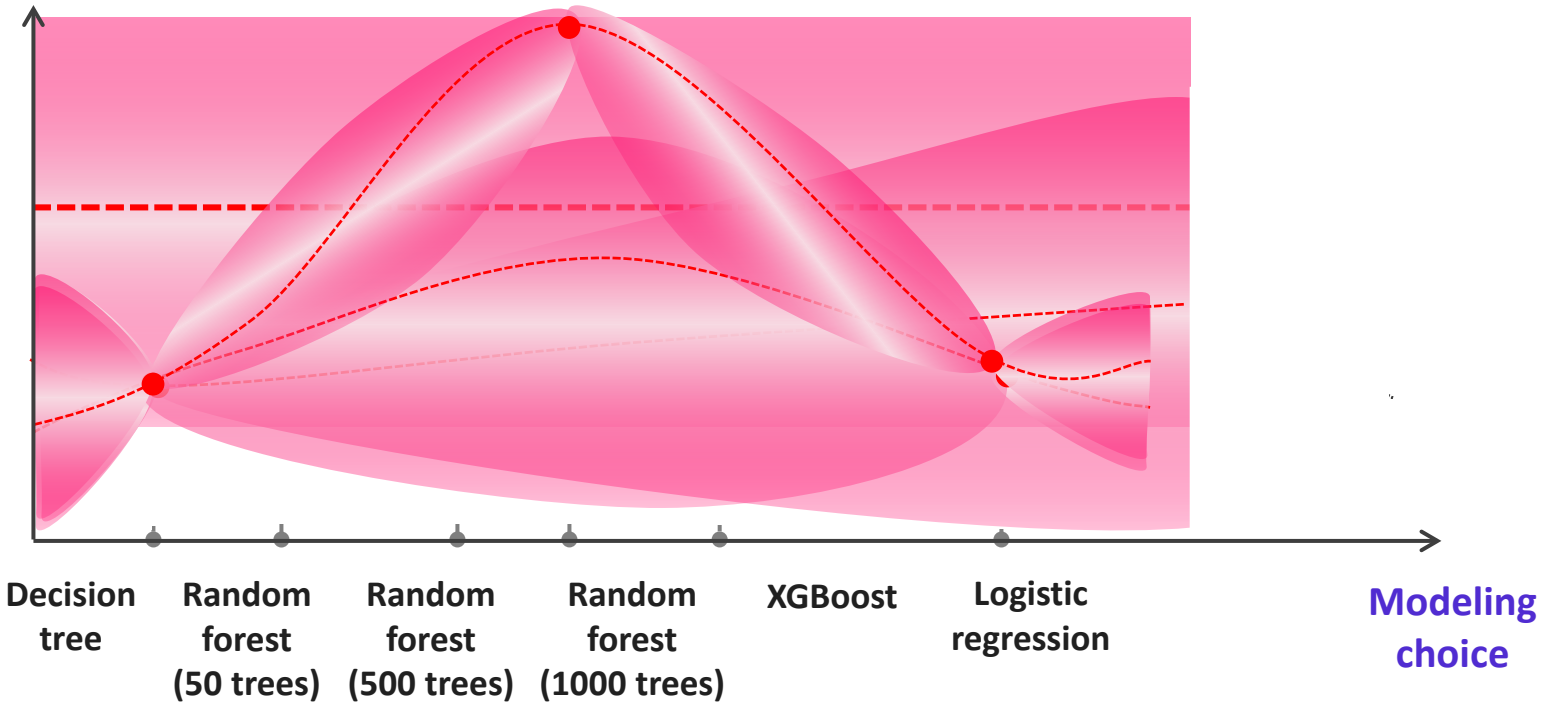
$$P^*_{\theta^*} \in \arg\max_{P_\theta \in \mathcal{P}_\Theta} \frac{1}{K} \sum_{i=1}^{K} \mathcal{L}(P_\theta; \mathcal{D}^{(i)}_{\text{train}}, \mathcal{D}^{(i)}_{\text{valid}})$$

A hard *learning and optimization* problem

Bayesian optimization

- Gaussian process prior, balances exploration and exploitation.

Performance function $P^*_{\theta^*} \in \arg\max_{P_\theta \in \mathcal{P}_\Theta} \frac{1}{K} \sum_{i=1}^{K} \mathcal{L}(P_\theta; \mathcal{D}^{(i)}_{\text{train}}, \mathcal{D}^{(i)}_{\text{valid}})$



$$\mathcal{P} = \mathcal{A}_d \times \mathcal{A}_f \times \mathcal{A}_c \times \mathcal{A}_a$$

Curse of dimensionality

- **The CPSH problem** $\arg\max_{P_\theta \in \mathcal{P}_\Theta} \frac{1}{K} \sum_{i=1}^{K} \mathcal{L}(P_\theta; \mathcal{D}_{\text{train}}^{(i)}, \mathcal{D}_{\text{valid}}^{(i)})$
- **Bayesian optimization**

Gaussian process prior

$$f \sim \mathcal{GP}(\mu(P_\theta), k(P_\theta, P'_\theta))$$

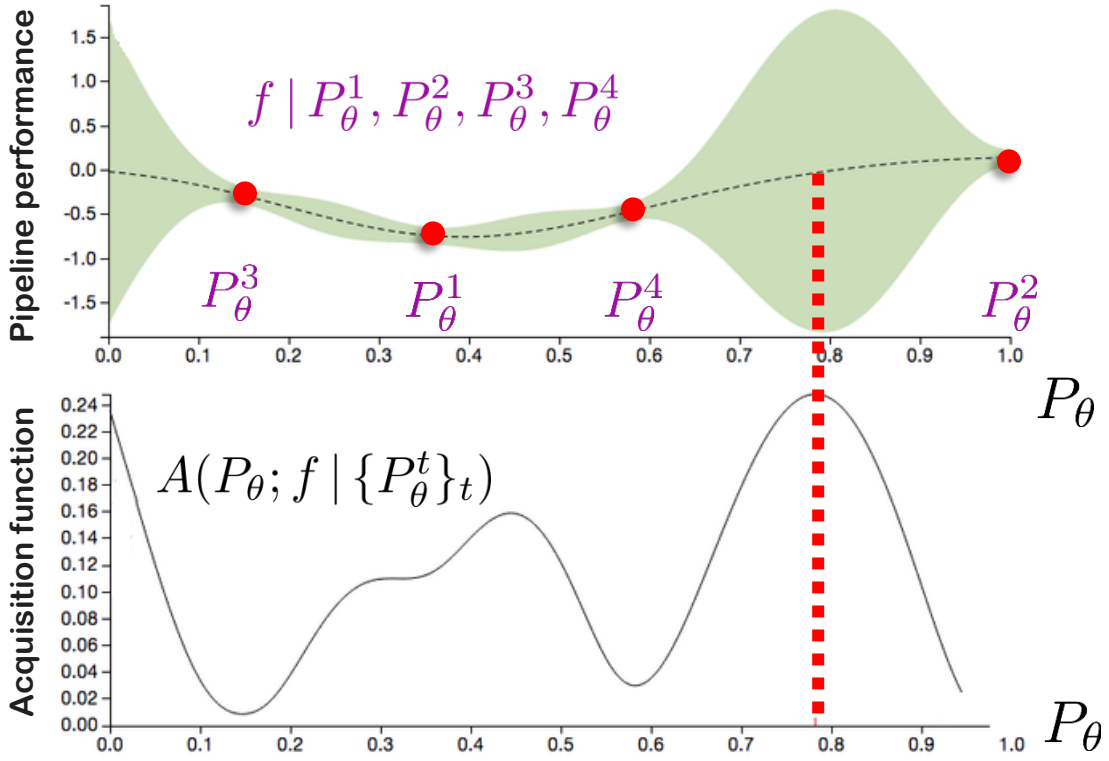
Gaussian process posterior

$$f \mid \{P_\theta^t\}_t$$

Select new pipeline via acquisition function

$$P_\theta^{t+1} = \arg\max_{P_\theta} A(P_\theta; f \mid \{P_\theta^t\}_t)$$

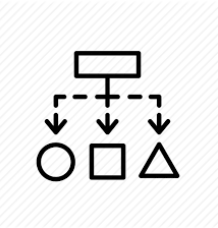
$$f(P_\theta) = \frac{1}{K} \sum_{i=1}^{K} \mathcal{L}(P_\theta; \mathcal{D}_{\text{train}}^{(i)}, \mathcal{D}_{\text{valid}}^{(i)}) + \varepsilon$$



Bayesian Optimization does not work well for $D > 10$ [Wang, 2013]

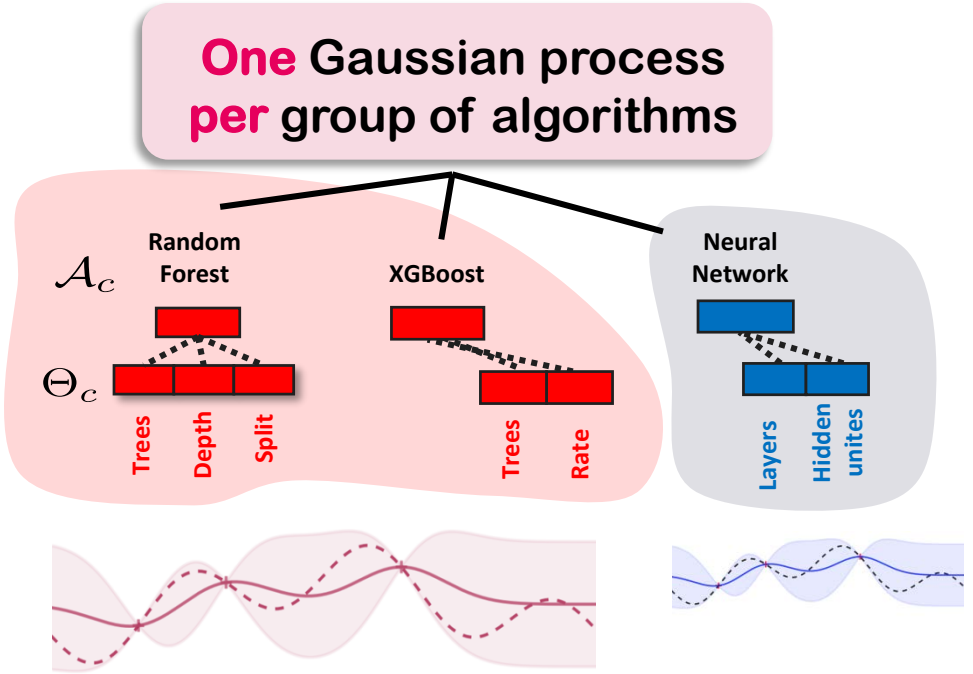
Bayesian Optimization with Structured Kernel Learning

[Alaa & vdS, ICML 2018]



- **Main idea:** Some algorithms are "correlated" and some are not => Correlated algorithms should be made to share information
- Correlation is not known in advance, so must be learned
- Learn a structured kernel that groups correlated algorithms:

  - Low dimensionality for every group
  - Relevant information sharing within a group



Sparse additive BO [Alaa & vdS, ICML 2018]

- **Main idea:** Not **ALL** algorithms have correlated performance!
- **Example:** XGBoost could be correlated with Random forest, but not with neural networks!

Improvement in learning rate

Full BO

$$\Theta\left(t^{-\frac{\alpha}{2\alpha+d}}\right)$$

Sparse BO

$$\Theta\left(m \cdot t^{-\frac{\alpha}{2\alpha+d/m}}\right)$$

…but the structure of the kernel is unknown!

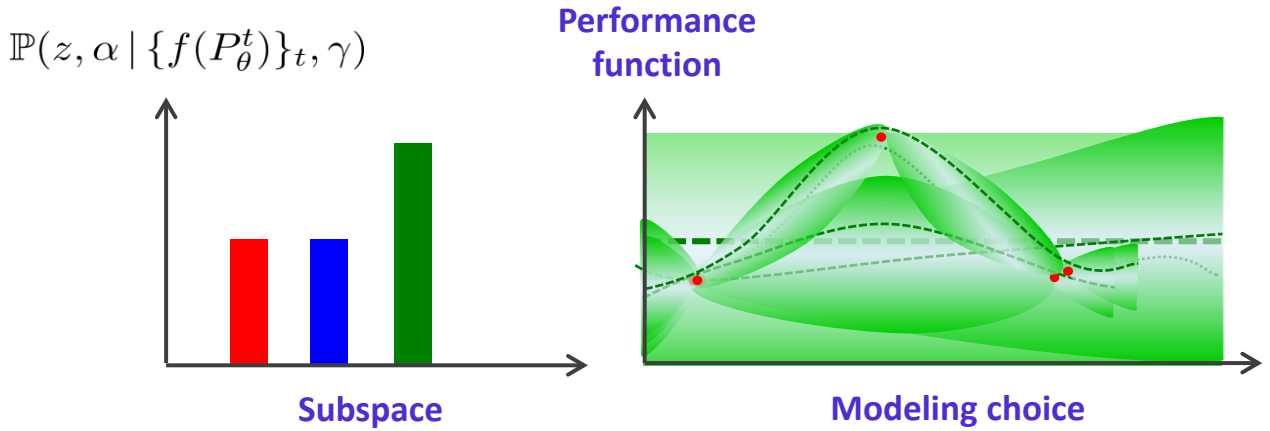
- **Define variable** $z_i \in \{1, \ldots, m\}$: indicator for subspace allocation for modeling choice i
- **Hierarchical Bayesian Prior**

Prior on decompositions

$$\alpha \sim \text{Dirichlet}(M, \gamma)$$
$$z_i \sim \text{Multinomial}(\alpha)$$

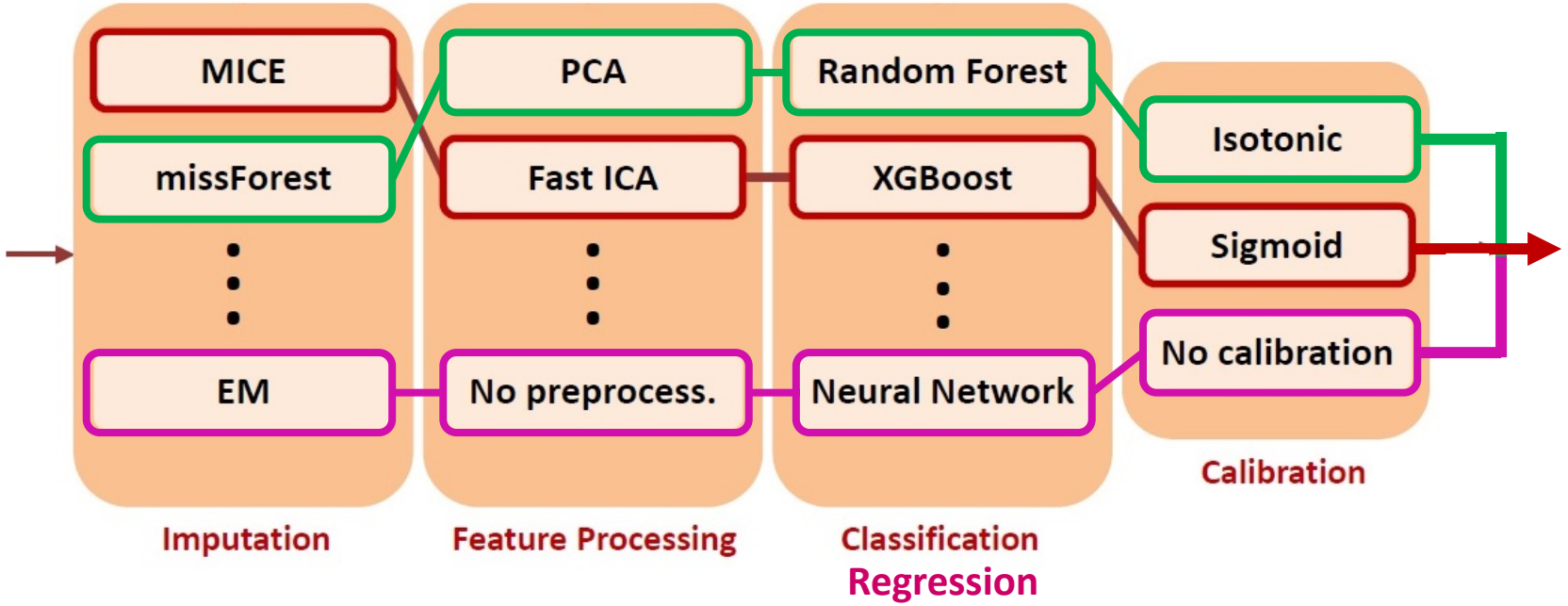
Posterior decompositions

$$\mathbb{P}(z, \alpha \mid \{f(P_\theta^t)\}_t, \gamma) \propto \mathbb{P}(\{f(P_\theta^t)\}_t \mid z)\, \mathbb{P}(z \mid \alpha)\, \mathbb{P}(\alpha, \gamma)$$



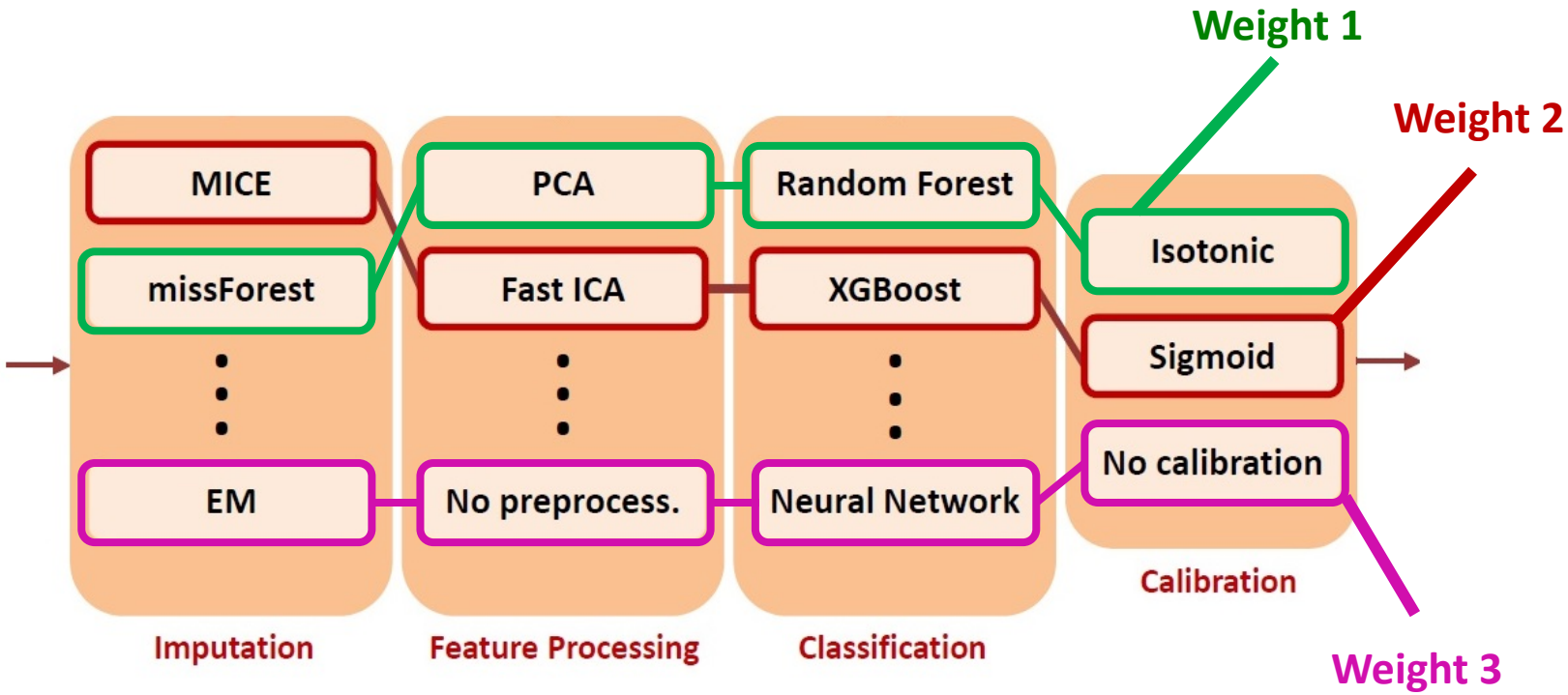
Ensembles

- **Instead of the single best pipeline we use an ensemble**
- **Why?**
  - ***Uncertainty*:** finite data set to learn from, so we are not sure which pipeline is “best”
  - ***Information loss:*** using a single pipeline discards useful information from other pipelines



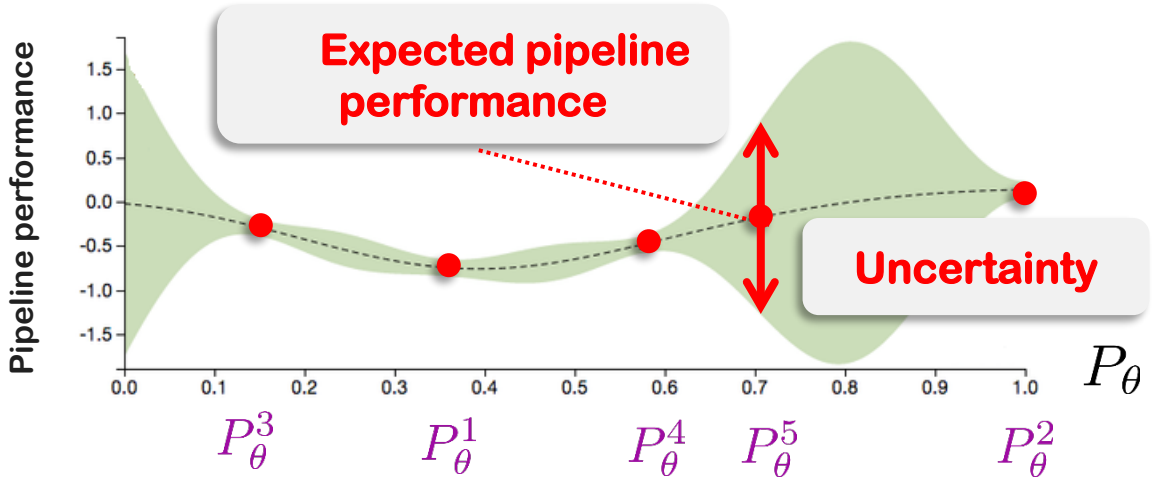
Building ensembles…

- Soft selection of algorithms: augment a weight assigned to each model in the BO domain.



Post-hoc Ensemble Construction

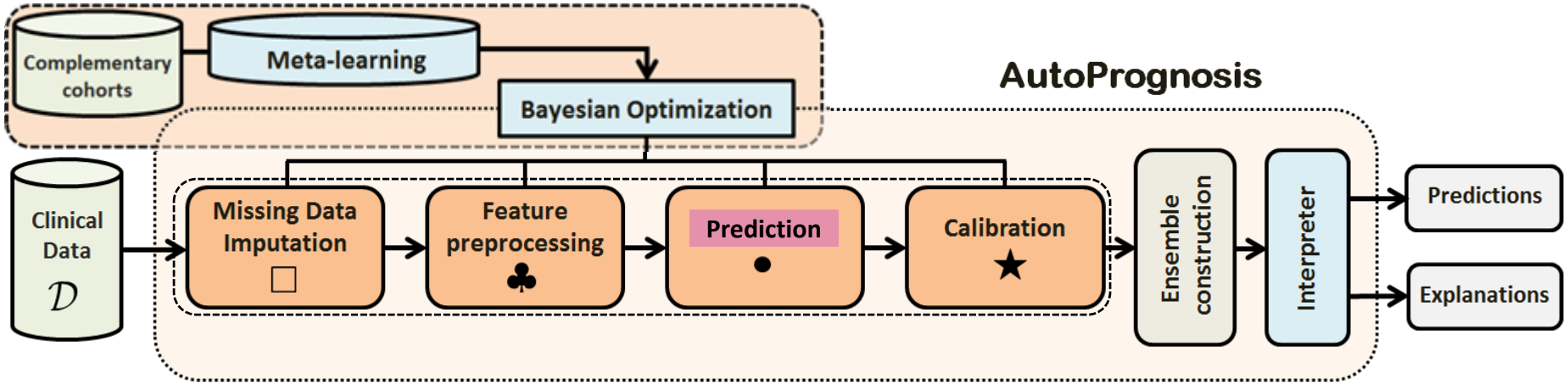
- Bayesian model averaging



- Create an ensemble using the posterior distribution of performances
- Create a linear combination of pipelines $\sum_i w_i P_\theta^i$
- Weight of every pipeline = empirical probability of it being the best!

$$
\begin{aligned}
w_i &= \mathbb{P}(P_\theta^{i^*} = P_\theta^i \mid \mathcal{H}_t) \\
&= \prod_{j \neq i} \Phi\left((\mu_i - \mu_j) \cdot (\sigma_i^2 + \sigma_j^2)^{-\frac{1}{2}}\right),
\end{aligned}
$$

AutoPrognosis in practice



Cardiovascular - ICML 2018
Cystic Fibrosis - Scientific Reports - 2018
UK Biobank - Plos One 2018
Breast Cancer – 2019
Covid - 2020

Adjutorium: AutoPrognosis for Covid-19

Our goal: Provide evidence that reliably assists the difficult decisions clinicians and managers have to make to save lives

Use depersonalized data

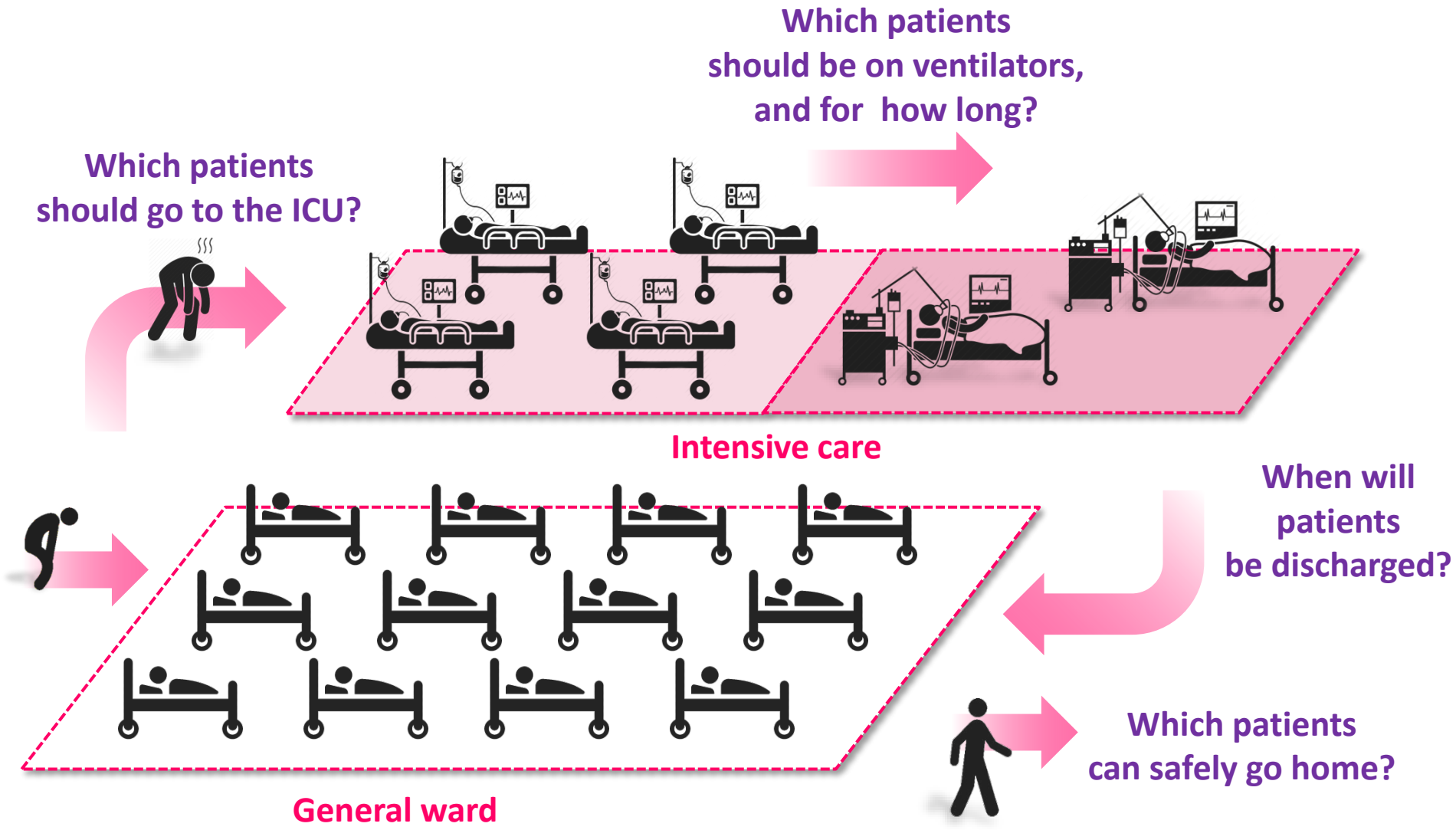
- demographic info, comorbidities, hospitalization details, outcomes

to:

- **forecast personalized risk for each patient**
- **forecast personalized patient benefit from resources**
- **forecast which treatments are needed by each patient and when**
- **forecast which resources are needed by each patient and when**
- **forecast future resource requirements at the hospital level**

www.vanderschaar-lab.com/covid-19/

Decisions that healthcare professionals need to make



AutoPrognosis at work
Covid-19: Who needs ventilation?

- **AUC-ROC accuracy for predicting whether a patient will need ventilation based on info available at hospital admission**

Model	AUC-ROC
AP: all features	**0.771 ± 0.002**
AP: age + specific comorbidities	0.761 ± 0.001
AP: age + no. of comorbidities	0.720 ± 0.003
Cox Regression: all features	0.690 ± 0.002
Charlson Comorbidity Index	0.618 ± 0.002

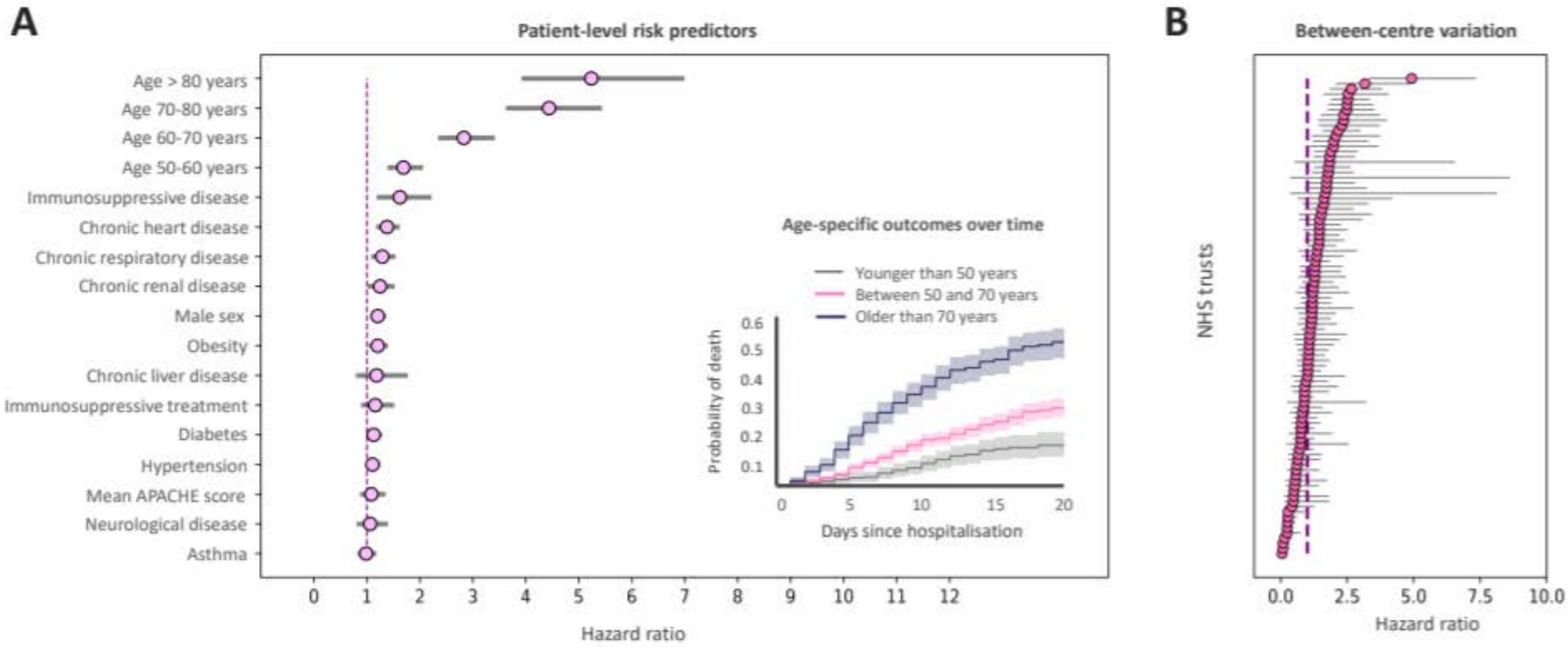
News

Trials begin of machine learning system to help hospitals plan and manage COVID-19 treatment resources developed by NHS Digital and University of Cambridge

Trials have begun of a system that will use machine learning to help predict the upcoming demand for intensive care (ICU) beds and ventilators needed to treat patients with COVID-19 at individual hospitals and across regions in England.

Date: 20 April 2020

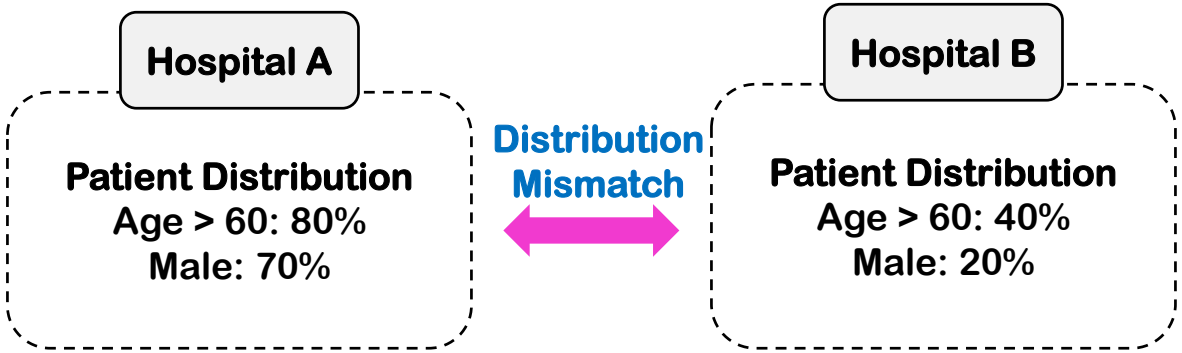
Between-centre differences for COVID-19 ICU mortality from early data in England [Intensive Care Medicine, May 2020]



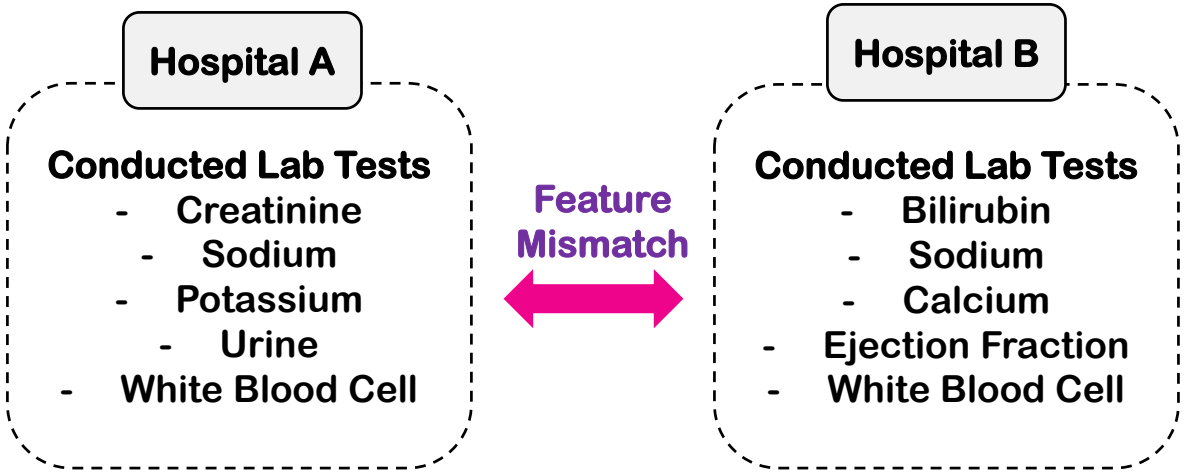
How can we transfer what we have learned to other jurisdictions?

Challenges

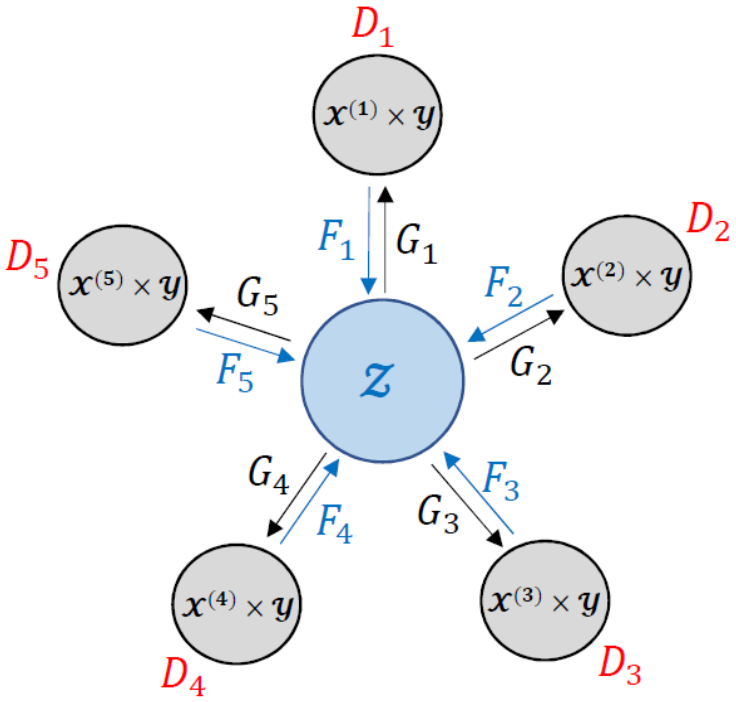
- **Distribution Mismatch**: Auxiliary data does not come from the **same distribution as target data**



- **Feature Mismatch**: Different data collectors collect **different pieces of information** for each sample

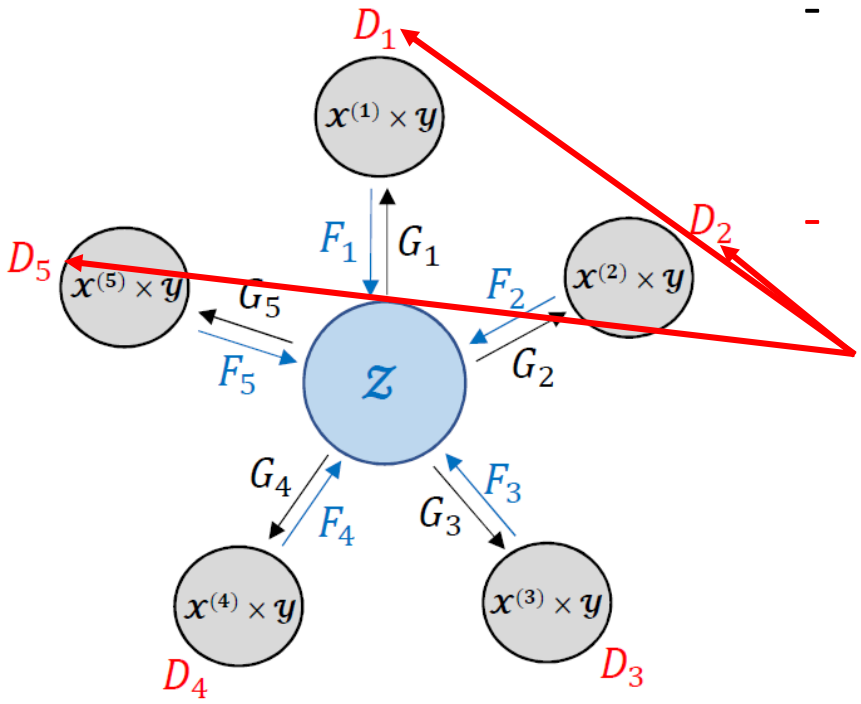


Radial GAN [Joon, Jordon, vdS, ICML 2018]



- Use **multiple GAN architectures** to "**translate**" the data from one dataset to another

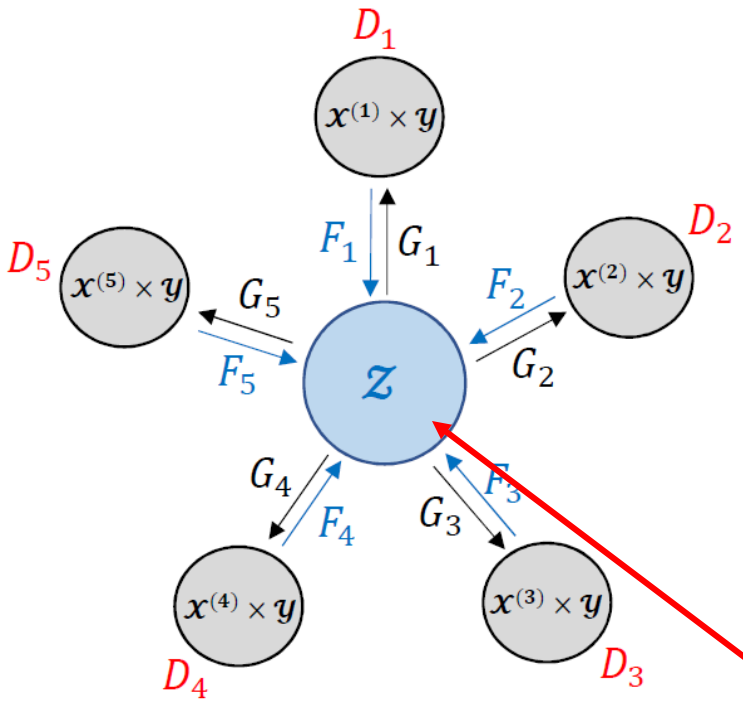
Radial GAN [Joon, Jordon, vdS, ICML 2018]



- Use **multiple GAN architectures** to "**translate**" the data from one dataset to another

- **Distribution Mismatch** is dealt with by the **adversarial framework** which ensures that "translation" respects the **target distribution**

Radial GAN [Joon, Jordon, vdS, ICML 2018]



- Use **multiple GAN architectures** to "**translate**" the data from one dataset to another
- **Distribution Mismatch** is dealt with by the **adversarial framework** which ensures that "translation" respects the **target distribution**
- **Feature mismatch** is dealt with by introducing a **latent space** through which all samples are mapped

Does it work?

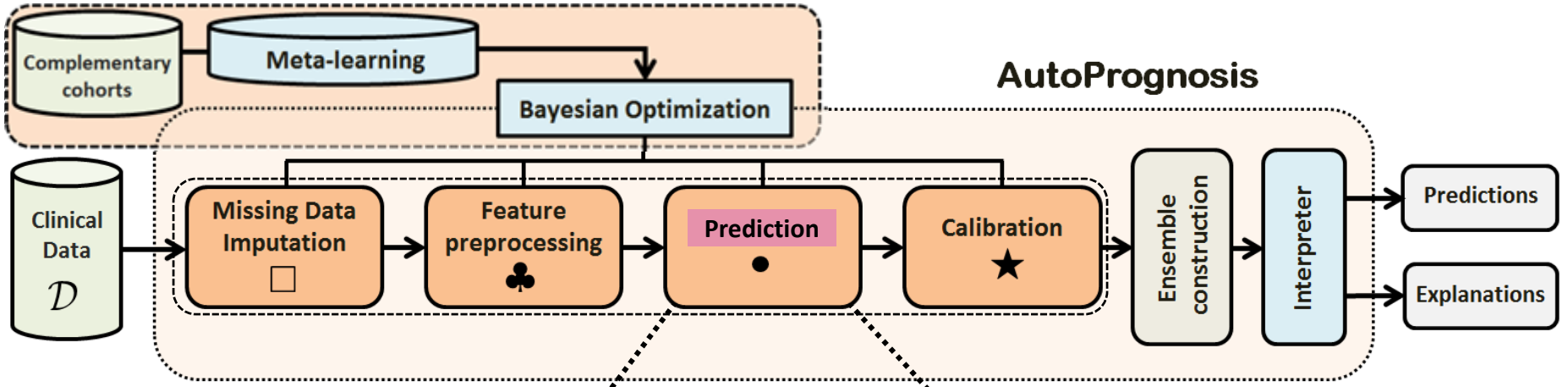
MAGGIC Dataset – Heart Failure

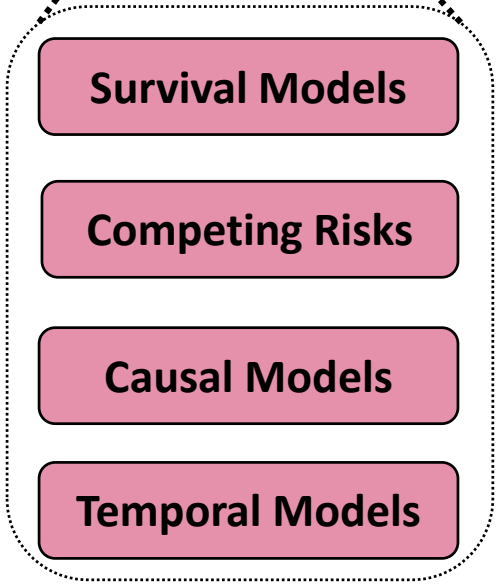
- A collection of **different medical studies**
- Label is set as **1-year all-cause mortality**

- Total number of **features** across all studies **is 216**
  (1) Average number of features in each study: **66**
  (2) Average number of **shared** features: **35 (53.8%)**

Algorithm	$M = 3$		$M = 5$		$M = 7$	
	AUC	APR	AUC	APR	AUC	APR
RadialGAN	.0154±.0091	.0243±.0096	.0292±.0009	.0310±.0096	.0297±.0071	.0287±.0073
Simple-combine	.0124±.0020	.0110±.0016	.0132±.0020	.0118±.0026	.0135±.0017	.0156±.0025
Co-GAN	.0058±.0028	.0085±.0026	.0094±.0018	.0139±.0036	-.0009±.0015	-.0013±.0027
StarGAN	.0119±.0015	.0150±.0013	.0150±.0025	.0191±.0013	.0121±.0020	.0160±.0021
Cycle-GAN	-.0228±.0112	-.0306±.0085	-.0177±.0082	-.0196±.0085	-.0076±.0022	-.0168±.0030
(Wiens et al., 2014)	-.0314±.0075	-.0445±.0125	-.0276±.0057	-.0421±.0052	-.0292±.0054	-.0411±.0063

Automated ML for clinical analytics (beyond predictions)





Clairvoyance

Lee, Alaa, Zame, vdS, AISTATS 2019

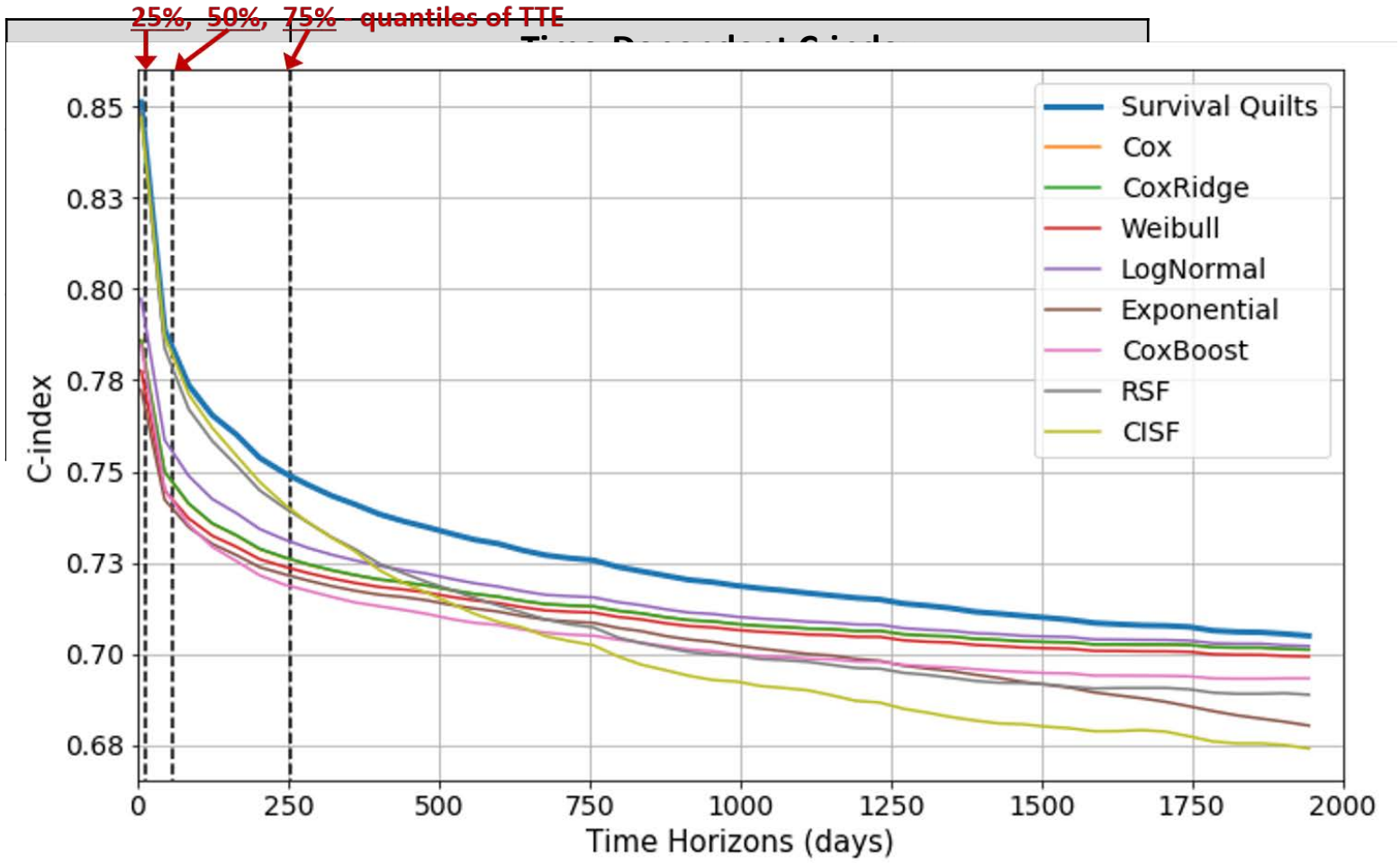
Alaa, vdS, NIPS 2017
Lee, Yoon, vdS, TBME 2019

Alaa, vdS, ICML 2019

Zhang, Jarett, vdS, AISTATS 2020

Survival Analysis – What is the challenge?

- There is no "best" survival model for all survival horizons
  - depends on time-horizon



Survival Quilts [Lee, Zame, Alaa, vdS, AISTATS 2019]

- **Main Idea:** autoML for survival analysis
  - ensemble over different time horizons
  - both discrimination among predicted risks and calibration

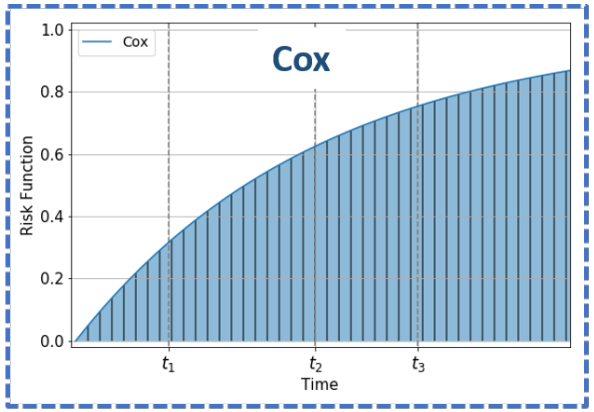
- **Our goal:** estimate risk function $R(t|\mathbf{x}) = 1 - S(t|\mathbf{x}) = \mathbb{P}(T \leq t|\mathbf{x})$ maximizing following optimization problem:

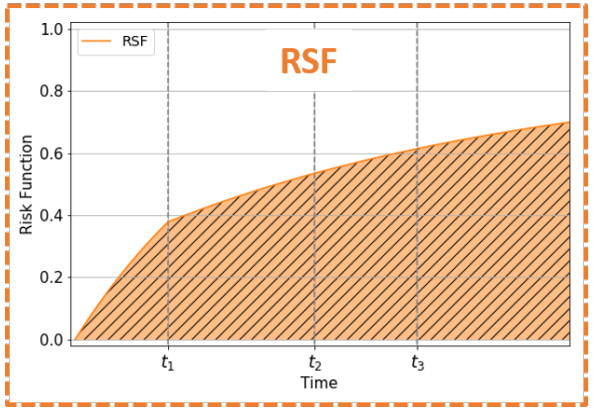
$$\max_{R \in \mathcal{R}} f(R)$$
$$\text{s.t. } g(R) \leq c,$$

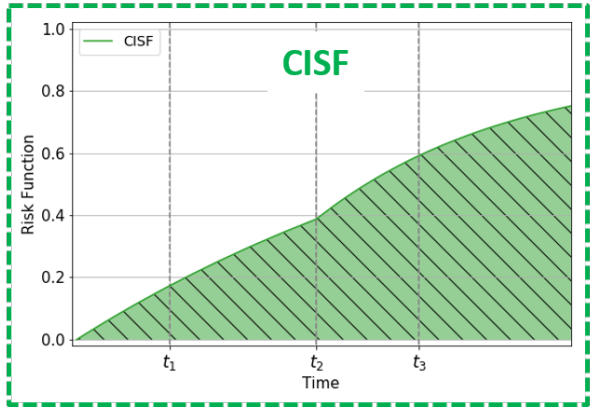
Constrained BO
Quilting Pattern
Composition Problem

  - $f(\cdot)$: Time-dependent C-Index (T. A. Gerds et al., 2013)
  - $g(\cdot)$: Time-dependent Brier Score (U. B. Mogensen et al., 2013)

Survival Quilts: A Schematic Overview

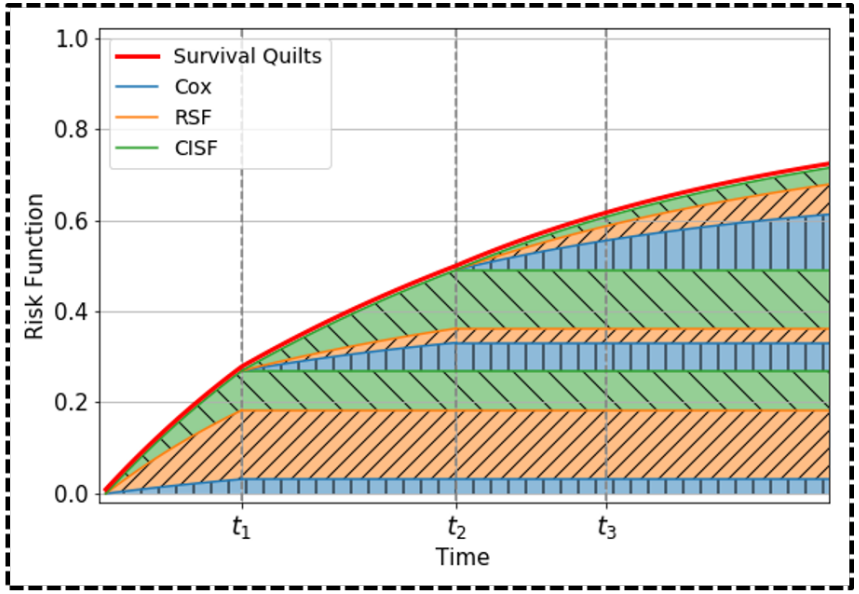






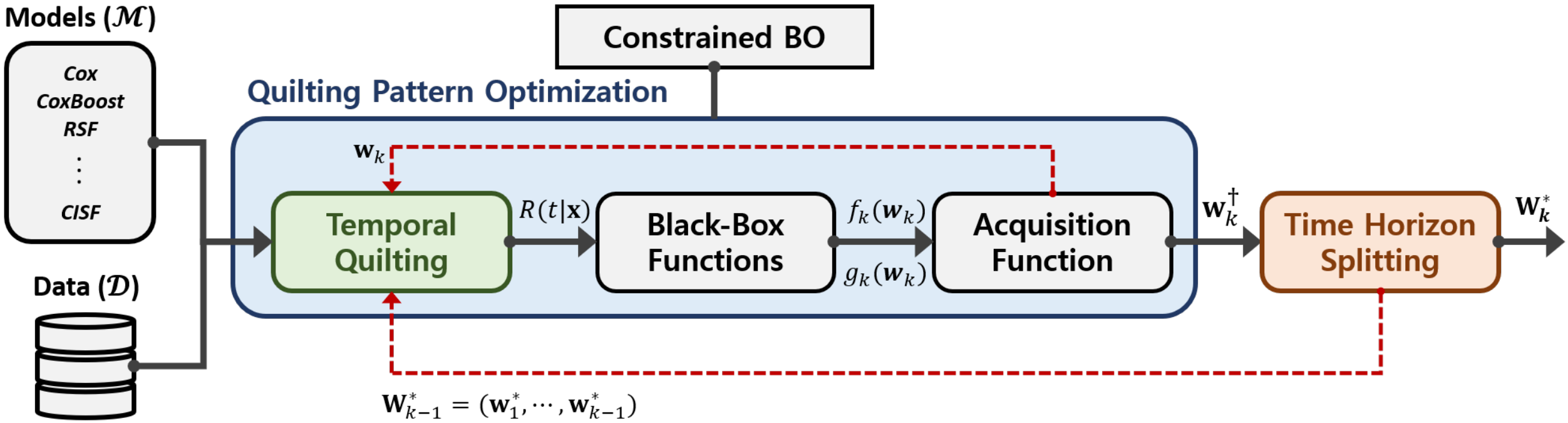
Temporal Quilting

Quilting Pattern



$$w_1 = \begin{bmatrix} 0.1 \\ 0.4 \\ 0.5 \end{bmatrix} \quad w_2 = \begin{bmatrix} 0.2 \\ 0.2 \\ 0.6 \end{bmatrix} \quad w_3 = \begin{bmatrix} 0.5 \\ 0.4 \\ 0.1 \end{bmatrix}$$

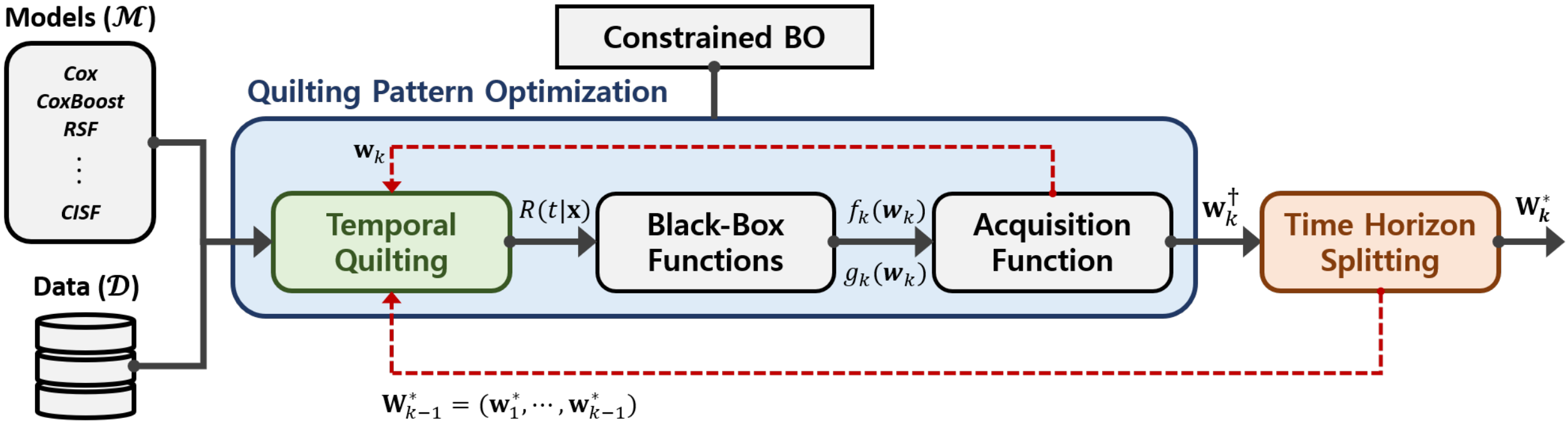
Survival Quilts [Lee, Zame, Alaa, vdS, AISTATS 2019]



Step 1: Temporal quilting - **construct valid risk functions for a given set of weights (a quilting pattern) for survival models over time horizons**

- Risk predictions at past time horizons are carried forward to future time horizons to provide a consistent risk function

Survival Quilts [Lee, Zame, Alaa, vdS, AISTATS 2019]



Step 2: Quilting pattern is optimized using Constrained BO

Step 3: Quilting pattern is made robust through Time Horizon Splitting (Endogenous time horizon splitting)

Step 2: Construct the Constrained BO

- Model the ***black-box optimization*** problem as

$$\mathbf{W}^* = \arg\max_{\mathbf{W}} \frac{1}{J}\sum_{j=1}^{J} \mathcal{L}_f(\mathbf{W}; \mathcal{D}_{\text{tr}}^{(j)}, \mathcal{D}_{\text{va}}^{(j)}) \quad \text{s.t.} \quad \frac{1}{J}\sum_{j=1}^{J} \mathcal{L}_g(\mathbf{W}; \mathcal{D}_{\text{tr}}^{(j)}, \mathcal{D}_{\text{va}}^{(j)}) \leq c$$

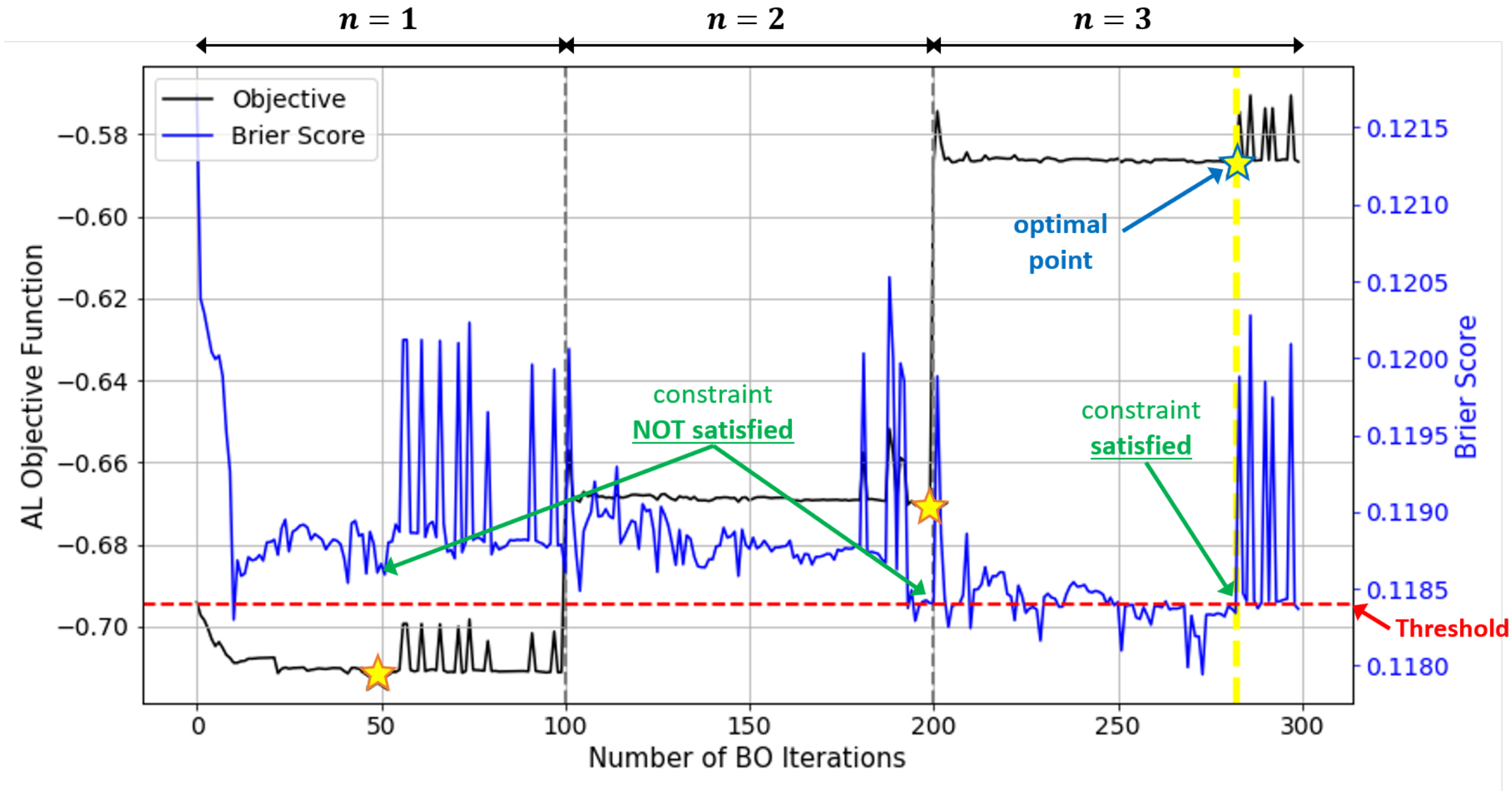
where $\mathbf{W} = [\mathbf{w_1}, \cdots, \mathbf{w_K}]$ and $\mathbf{w_k}$ is vector comprising all weights at time t_k

- Construct BO with GP priors:

$$f \sim \mathcal{GP}(\mu_f(\mathbf{W}), \kappa_f(\mathbf{W}, \mathbf{W}')) \qquad g \sim \mathcal{GP}(\mu_g(\mathbf{W}), \kappa_g(\mathbf{W}, \mathbf{W}'))$$

- Augmented Lagrangian BO with GP priors:

$$L(\mathbf{w}_k; \lambda, \rho) = -f_k(\mathbf{w}_k) + \lambda \cdot (g_k(\mathbf{w}_k) - c) + \frac{1}{\rho}\max(0, g_k(\mathbf{w}_k) - c)^2$$


$n = 1$
$n = 2$
$n = 3$
Objective
Brier Score
optimal point
constraint NOT satisfied
constraint satisfied
Threshold
AL Objective Function
Brier Score
Number of BO Iterations
−0.58
−0.60
−0.62
−0.64
−0.66
−0.68
−0.70
0.1215
0.1210
0.1205
0.1200
0.1195
0.1190
0.1185
0.1180
0
50
100
150
200
250
300

Results

- **Time-dependent C-index**

$$C(t) = \mathbb{P}(\hat{R}(t|\mathbf{x}_i) > \hat{R}(t|\mathbf{x}_j)|\Delta_i = 1, T_i \leq t, T_i < T_j)$$

- **Time-dependent Brier Score**

$$BS(t) = \mathbb{E}\left[(\mathbf{1}(T_i \leq t) - \hat{R}(t|\mathbf{x}_i))^2\right]$$

Average performance over different time-horizons

Models	METABRIC		UNOS		SUPPORT	
	C-index	Brier Score	C-index	Brier Score	C-index	Brier Score
Cox	0.645±0.03	0.191±0.01	0.620±0.02	0.198±0.02	0.734±0.01	0.173±0.00
CoxRidge	0.648±0.03	0.188±0.01	0.622±0.02	0.196±0.02	0.734±0.01	0.173±0.00
SurvReg	0.644±0.03	0.191±0.01	0.619±0.02	0.199±0.02	0.731±0.01	0.175±0.00
CoxBoost	0.653±0.03	0.186±0.01	0.621±0.02	0.197±0.02	0.733±0.01	0.174±0.00
gbmSurv	0.630±0.04	0.207±0.01	0.612±0.04	0.208±0.01	0.736±0.01	0.185±0.00
RSF	0.698±0.01	0.183±0.01	0.632±0.03	0.200±0.02	0.748±0.01	0.167±0.00
cForest	0.703±0.01	0.183±0.01	0.631±0.04	0.195±0.02	0.748±0.01	0.167±0.00
Survival Quilts	0.711±0.01	0.182±0.01	0.636±0.03	0.196±0.02	0.759±0.01	0.169±0.00

Part 2: Interpretable, explainable and trustworthy ML

How to turn ML models into actionable intelligence?

We need

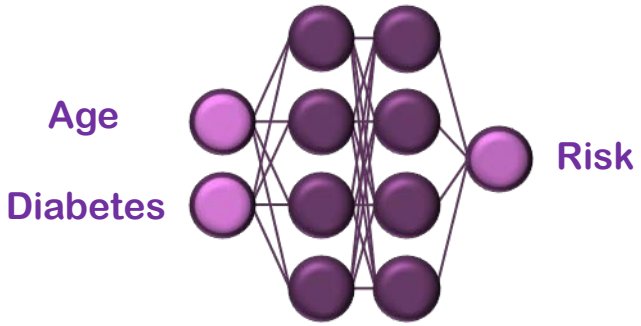
- ***Transparency***: **Users need to comprehend how the model makes predictions**
- ***Risk understanding***: **Users need to understand, quantify and manage risk**
- ***Avoid implicit bias***: **Users need to be able to check whether the model does not learn biases**
- ***Discovery***: **Users need to distil insights and new knowledge from the learned model**
- ***Know what we do not know:*** **Users need to have a quantification of the model's prediction uncertainty**

Interpretability, explainability and trustworthiness

Understand

why a prediction is made by the model

Interpretability



Interpretation 1

$$\boldsymbol{Risk} \approx \beta_o\, Age + \beta_1\, Diabetes$$

Interpretation 2

Feature importance: β_o, β_1

what can we **learn** from the model

Explainability

All possible interpretations

User context

Explanation module

Interpretation 2

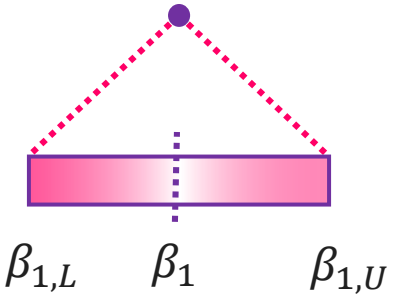
Feature importance: β_o, β_1

how **trustworthy** is the model's prediction

Trustworthiness

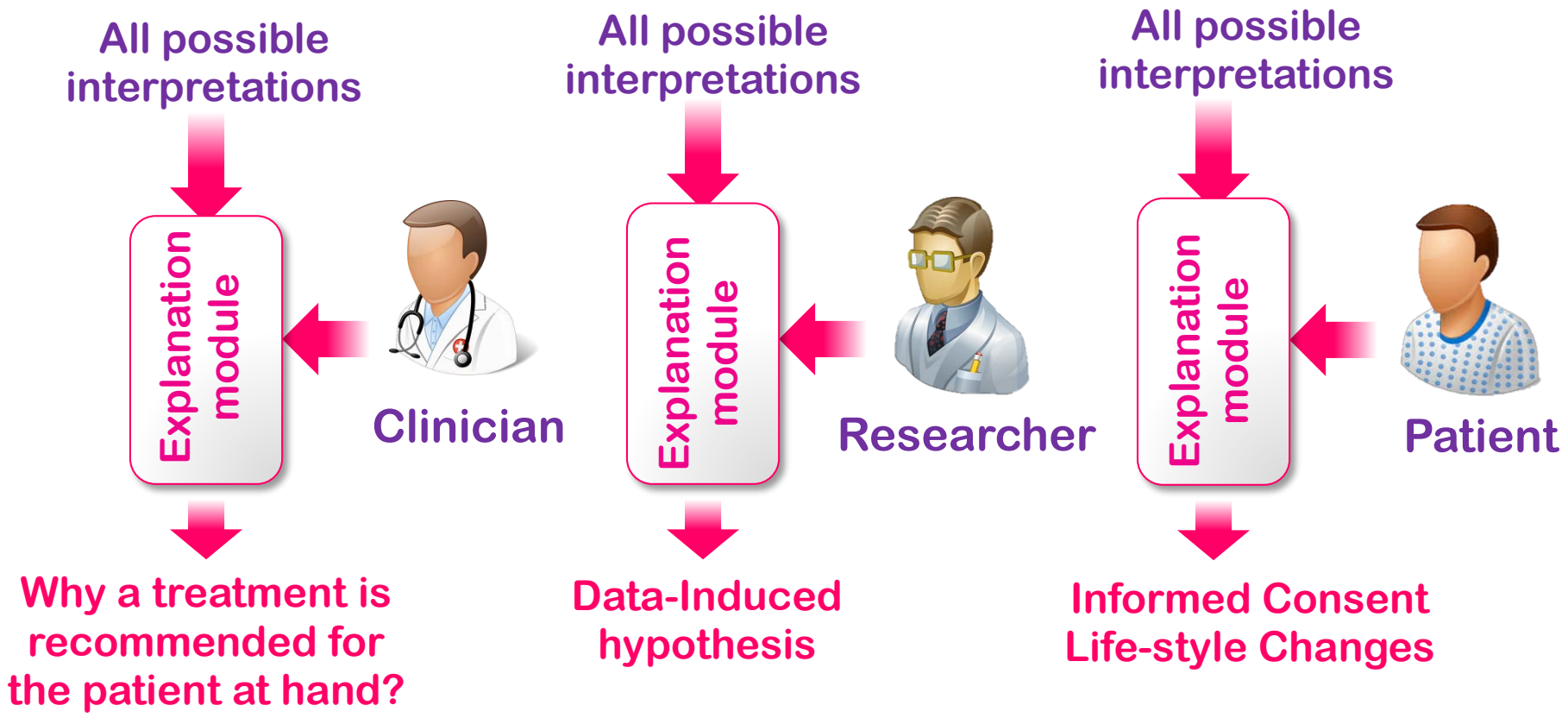
$$\boldsymbol{Risk} \approx \beta_o\, Age + \beta_1\, Diabetes$$

Confidence interval



Explainability = Tailored interpretability

- Different users seek different forms of “understanding”…



Many kinds of interpretations exist…

- Current methods are tailored to *one* type of interpretation
  Uncovering one of the following
  - What features are globally important, i.e. for the entire population?
  - What features are locally important, i.e. for this patient?
  - Features interaction
  - Model non-linearity

- Desiderata
  - Model-independent: general, not tailored to specific models
  - Post-hoc: should not interfere with model training, which may introduce bias and compromise accuracy

Which features of an individual are relevant for a prediction?

Mary

Age, Gender, Diabetes, Hypertension, SBP, ….

Black Box Predictive Model

1-year mortality due to Cardiovascular Disease: 0.78

Interpretations

Age, Diabetes, Hypertension, SBP

Various interpretation methods

Method	Feature importance	Individualized feature importance	Model-independent	Identifying the set of relevant features for each instance
LASSO [Tibshirani, 1996]	✓		✓	
Knock-off [Candes et al, 2016]	✓		✓	
L2X [Chen et al, 2018]	✓	✓	✓	
LIME [Ribeiro et al, 2016]	✓	✓	✓	
SHAP [Lundberg et al, 2017]	✓	✓	✓	
DeepLIFT [Shrikumar et al, 2017]	✓	✓		
Saliency [Simonyan et al, 2013]	✓	✓		
TreeSHAP [Lundberg et al, 2018]	✓	✓		
Pixel-wise [Batch et al, 2015]	✓	✓		
INVASE [Yoon, Jordon and van der Schaar, 2019]	✓	✓	✓	✓

INVASE discovers *the number* of relevant features for each instance

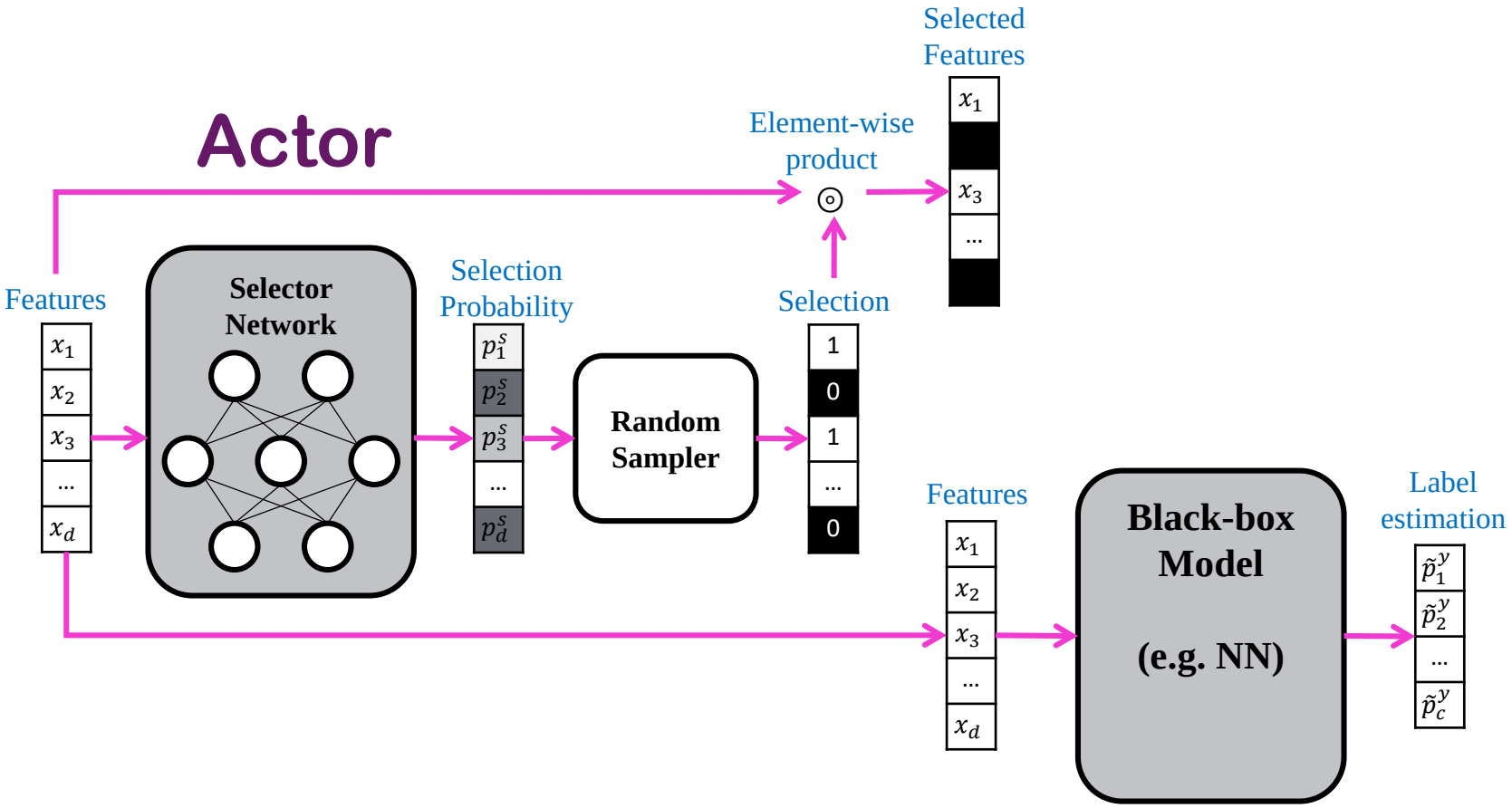
INVASE [Yoon, Jordon, vdS, ICLR 2019]

- How can we learn individualized feature importance?
- Key idea: **Use Reinforcement Learning (RL)**
  - Make observations
  - Select "actions" on the basis of these observations
  - Determine "rewards" for these actions

  - Ultimately learn a policy which selects the best actions
    - i.e. actions that maximize rewards given observations

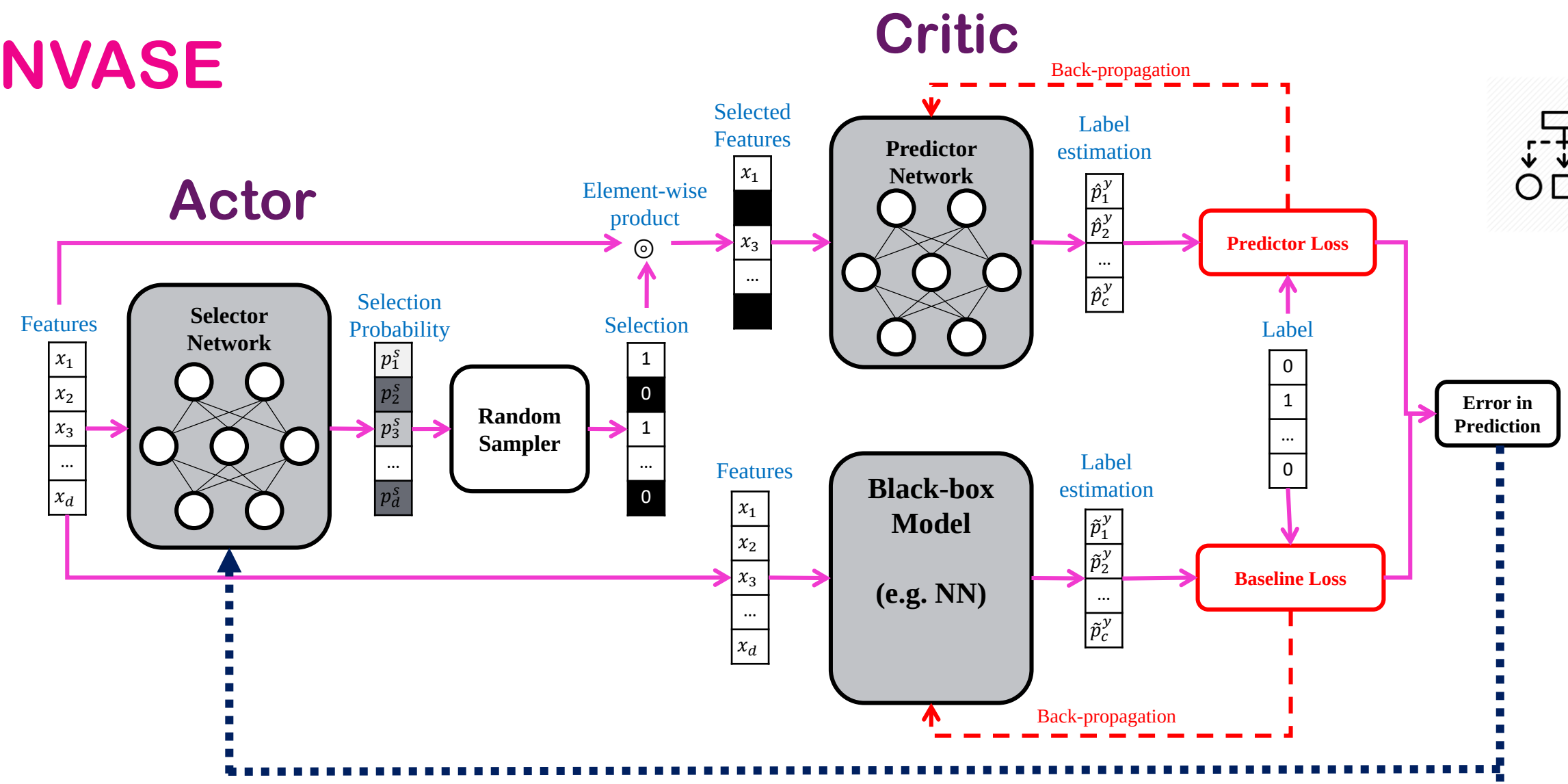
- We use the Actor-Critic approach to RL

INVASE



- **Selector network (actor)** takes instances and outputs vector of selection probabilities.

INVASE



- **Predictor network (critic)** receives the selected features, makes predictions and provides **feedback to the actor**.

INVASE: Instance-wise feature importance for prediction

Find selector function S that minimizes features selected $S(x)$ while satisfying equality constraints on the conditional distribution of the predictions.

- **Objective: minimize $S(\boldsymbol{x})$**
- **Constraints:**

$$\left(Y \middle| \boldsymbol{X}^{(S(\boldsymbol{x}))} = \boldsymbol{x}^{(S(\boldsymbol{x}))}\right) \overset{\mathrm{d}}{=} (Y|\boldsymbol{X} = \boldsymbol{x})$$

- $\boldsymbol{x}$: Features for *a given realization*
- $S: \mathcal{X} \to \{0,1\}^d$: Selector function, $S(\boldsymbol{x})$: Selected features
- Y: Predictions made by black-box model

INVASE: Instance-wise feature importance for prediction

Find selector function S that minimizes features selected $S(x)$ while satisfying equality constraints on the conditional distribution of the predictions.

- **Objective: minimize** $S(\boldsymbol{x})$
- **Constraints:**

$$\left(Y \middle| \boldsymbol{X}^{(S(x))} = \boldsymbol{x}^{(S(x))}\right) \overset{\mathrm{d}}{=} (Y|\boldsymbol{X} = \boldsymbol{x})$$

- **Lagrangian optimization:**

$$\mathcal{L}(S) = \mathbb{E}[KL\left(Y \middle| \boldsymbol{X}^{(S(x))} = x^{(S(x))}\right)||(Y|\boldsymbol{X} = x) + \lambda||S(x)||]$$

- **Challenging problem:**
  - Output space of the selector function is large - its size increases exponentially with the dimension of the feature space!
  - We do not have access to the densities required – need to be learned

[Yoon, Jordon, vdS, ICLR 2019]

Are we done?

- **NO!**
- **Need to ALSO understand what the model discovered: feature/statistical interactions, model non-linearity, etc.**

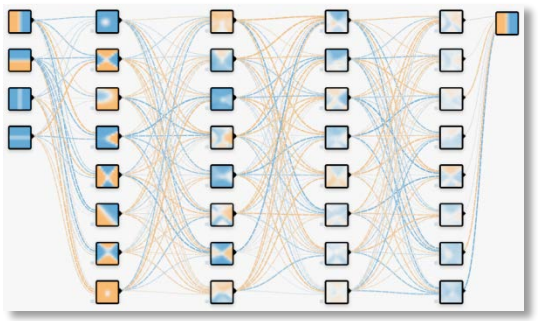
Method	Feature importance	Individualized feature importance	Feature interaction	Model-independent	Post-hoc
LIME [Ribeiro et al, 2016]	✓			✓	✓
SHAP [Lundberg et al, 2017]	✓		✓	✓	
DeepLIFT [Shrikumar et al, 2017]	✓				
INVASE [Yoon, Jordon and van der Schaar, 2019]	✓	✓		✓	✓
L2X [Chen et al, 2018]	✓	✓			
GAM [Lou et al, 2013]	✓		✓		
NIT [Tsang et al, 2018]	✓		✓		

What we are aiming for?

- Understand what the model discovered:
  feature importance, instance-wise feature importance, feature/statistical interactions, model non-linearity, etc.
- Produce a transparent risk equation describing the model for approval in practice guidelines (e.g. American Joint Committee on Cancer)
- Enable model explainability, not only interpretability

Demystify black-box models using symbolic metamodels
[A. Alaa & vdS, NeurIPS 2019]

Black-box ML model



$f(\mathbf{x})$

Symbolic Metamodeling

White-box (transparent) model

$$\alpha_1 X_1 + \alpha_2 X_2^2 + \alpha_3 X_1 X_2$$
$$\alpha_4 X_3^3 + \alpha_5 \log(X_4)$$

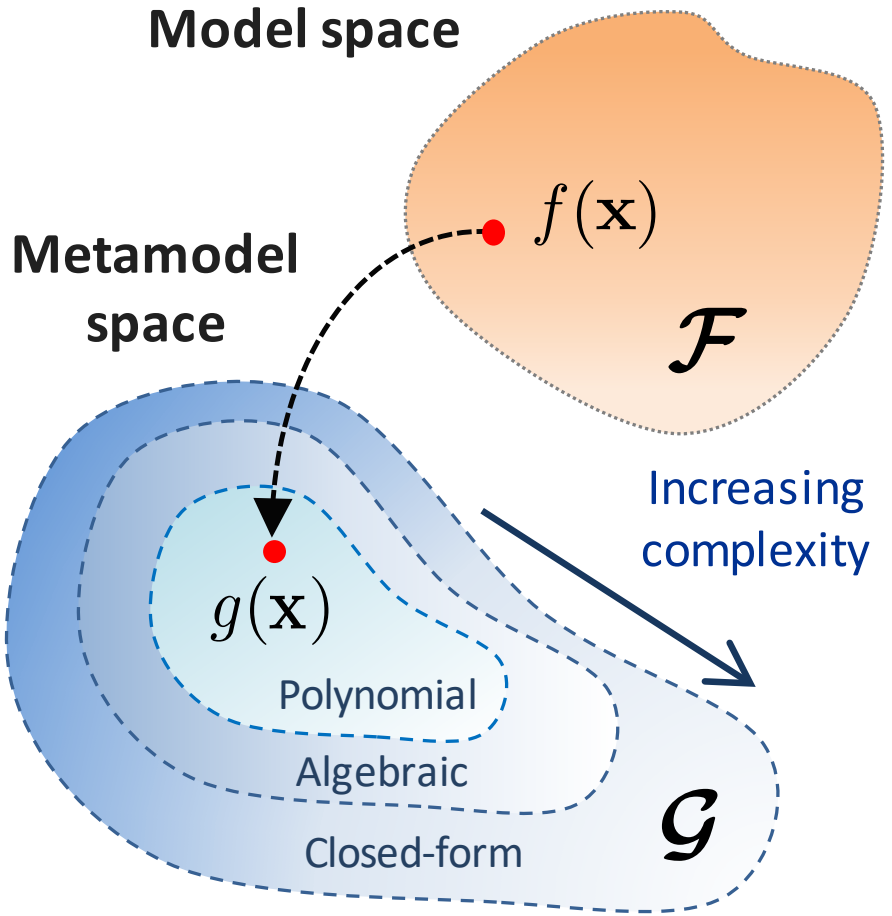
$g(\mathbf{x})$

- **Metamodel =** a model of a model.
- **A symbolic metamodel** outputs a **transparent function** ***describing*** the predictions of the black box model
- **Metamodeling** needs only **query access** to **trained black-box model**.

Symbolic metamodeling

Black-box ML model

Model space (uninterpretable)



White-box model

Metamodel space

Metamodel space can be *chosen* by the user!

How are we going to achieve this?

- **Kolmogorov-Arnold Theorem** [Kolmogorov et al, 1961]

  Every multivariate continuous function can be written as a finite composition of **univariate** continuous functions

$$g(\mathbf{x}) = \sum_{q=0}^{r} g_q \left(\sum_{p=1}^{n} g_{q,p}(x_p) \right)$$

- **The symbolic metamodeling problem**

Metamodel representation

$$g(\mathbf{x};\theta) = \sum_{q=0}^{2n} G \left(\sum_{p=1}^{n} G(x_p;\theta_{q,p});\theta_q \right)$$

Metamodel optimization

$$\theta^* = \arg\min_{\theta\in\Theta} \ell(f(\mathbf{x}), g(\mathbf{x};\theta))$$

What basic functions?

- **Meijer G-functions** [C. S. Meijer, 1936]

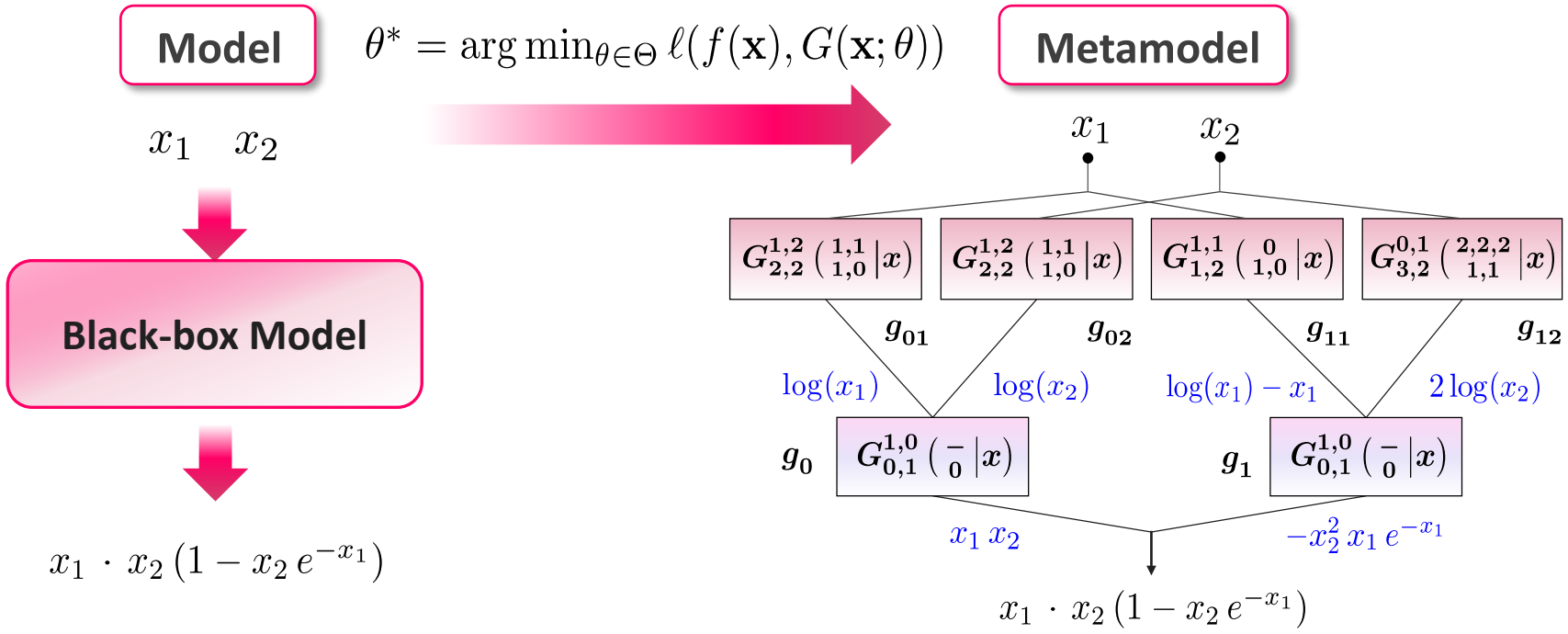
$$G_{p,q}^{m,n}\left(\begin{matrix}a_1,\ldots,a_p\\ b_1,\ldots,b_q\end{matrix}\middle|x\right)=\frac{1}{2\pi i}\int_L \frac{\prod_{j=1}^{m}\Gamma(b_j-s)\prod_{j=1}^{n}\Gamma(1-a_j+s)}{\prod_{j=m+1}^{q}\Gamma(1-b_j+s)\prod_{j=n+1}^{p}\Gamma(a_j-s)}\,x^s\,ds$$

- **Very general class of functions**
- **Parameter selection yields many familiar functions**

G-function	Equivalent function	G-function	Equivalent function
$G_{0,1}^{1,0}\left(\begin{matrix}-\\ 0\end{matrix}\middle|-x\right)$	e^x	$G_{2,2}^{1,2}\left(\begin{matrix}\frac{1}{2},1\\ \frac{1}{2},0\end{matrix}\middle|x^2\right)$	$2\arctan(x)$
$G_{2,2}^{1,2}\left(\begin{matrix}1,1\\ 1,0\end{matrix}\middle|x\right)$	$\log(1+x)$	$G_{1,2}^{2,0}\left(\begin{matrix}1\\ \alpha,0\end{matrix}\middle|x\right)$	$\Gamma(\alpha,x)$
$G_{0,2}^{1,0}\left(\begin{matrix}-\\ 0,\frac{1}{2}\end{matrix}\middle|\frac{x^2}{4}\right)$	$\frac{1}{\sqrt{\pi}}\cos(x)$	$G_{1,2}^{2,0}\left(\begin{matrix}1\\ 0,\frac{1}{2}\end{matrix}\middle|x^2\right)$	$\sqrt{\pi}\,\mathrm{erfc}(x)$
$G_{0,2}^{1,0}\left(\begin{matrix}-\\ \frac{1}{2},0\end{matrix}\middle|\frac{x^2}{4}\right)$	$\frac{1}{\sqrt{\pi}}\sin(x)$	$G_{0,2}^{1,0}\left(\begin{matrix}-\\ \frac{a}{2},\frac{-a}{2}\end{matrix}\middle|\frac{x^2}{4}\right)$	$J_a(x)$

Building a symbolic metamodel

- **Metamodel construction is "analogous" to a 2-layer neural network**

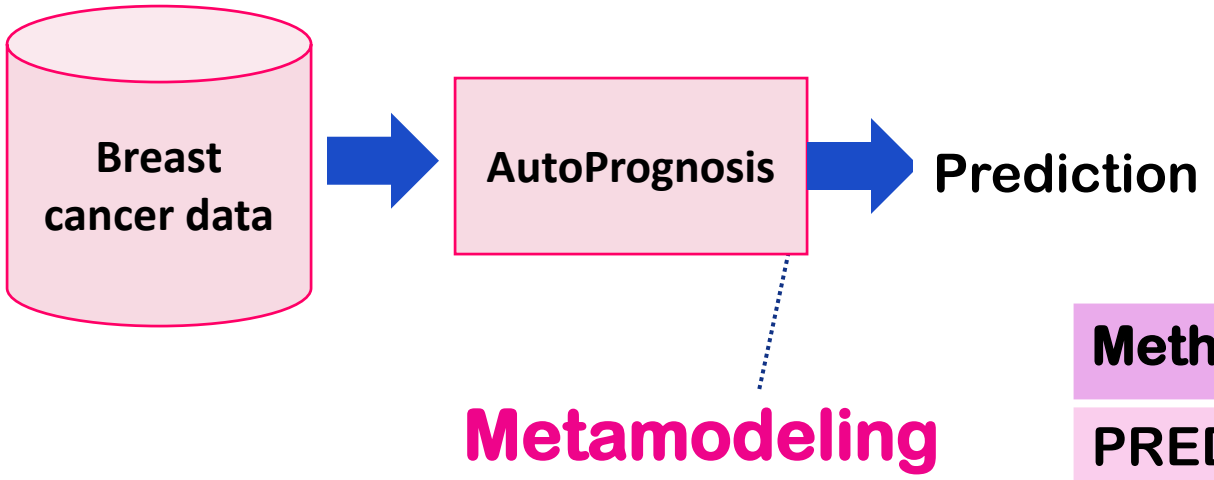


Parameters of a Meijer-G function can be learned by gradient descent!
This can be done very fast!

Interpretability using symbolic metamodeling in practice

[A. Alaa & vdS, NeurIPS 2019]

Example: Predicting breast cancer risk survival (5 years)

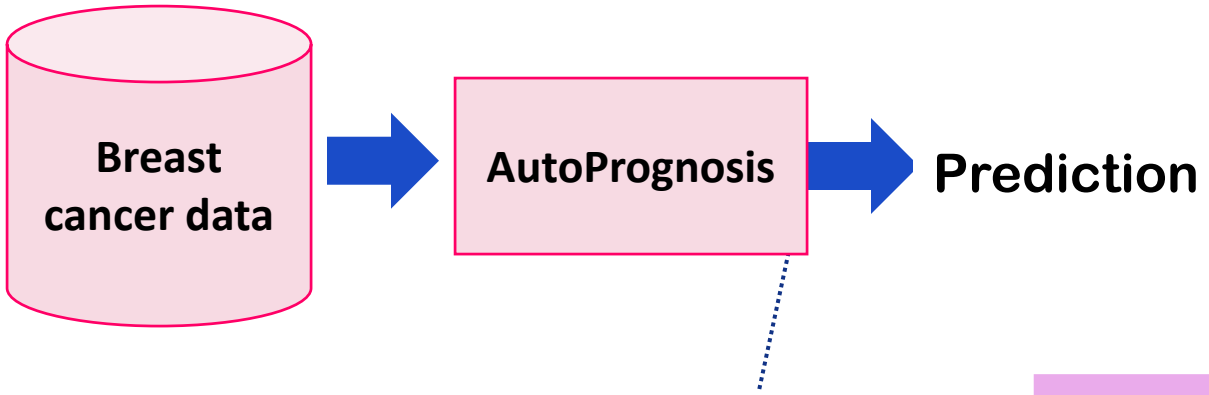


Method	AUC-ROC
PREDICT	0.75 ± 0.0033
AutoPrognosis	0.84 ± 0.0032

Interpretability using symbolic metamodeling in practice

[A. Alaa & vdS, NeurIPS 2019]

Example: Predicting breast cancer risk survival (5 years)



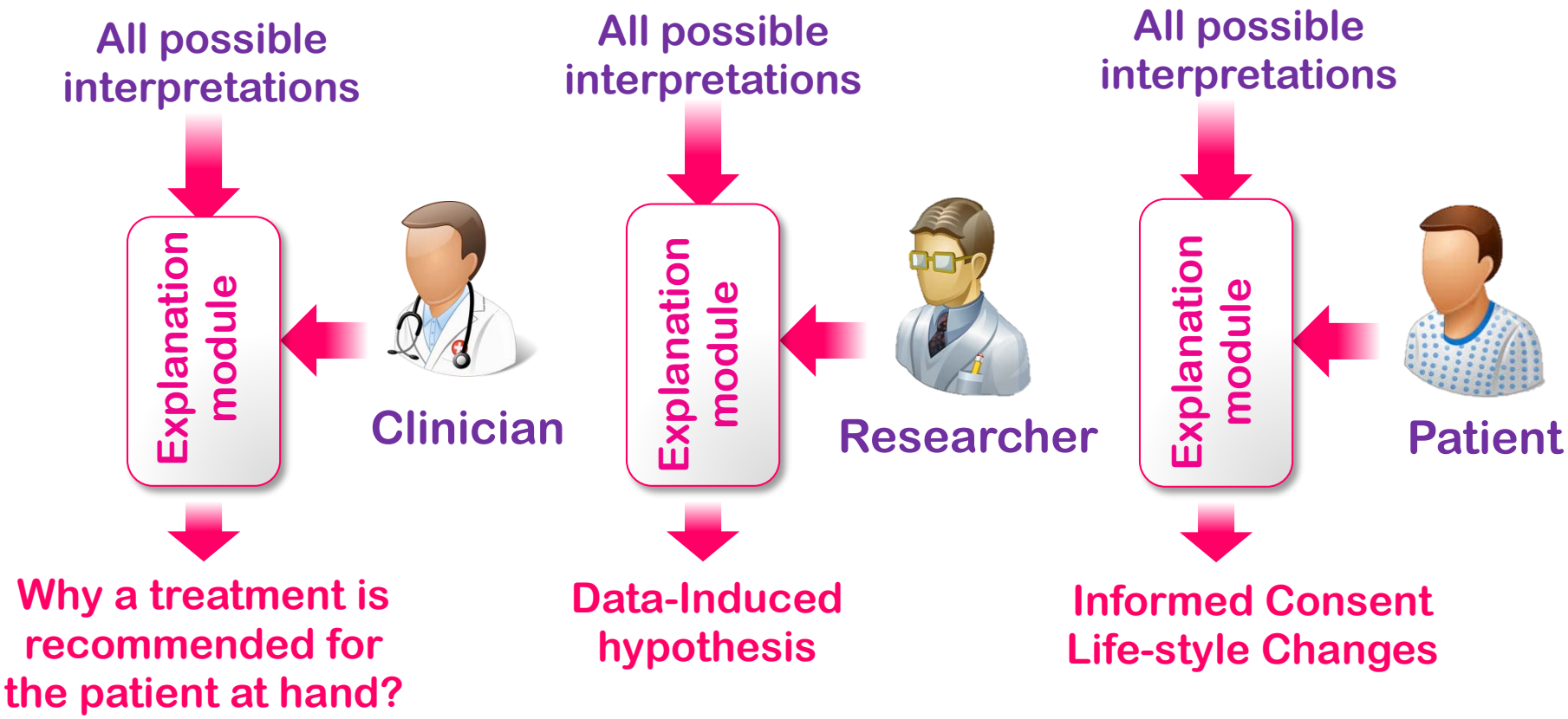
$$f(Age, ER, HER2, Tumor\ size, Grade, Nodes, Screening)$$

$$\exp\left(\frac{\text{Age}}{5} - \log\left(\frac{\text{Tumor size}}{100}\right) + \frac{1}{10}\log(\text{Nodes})\right) \times \exp\left(\frac{\text{ER} \cdot \text{Nodes}}{20} + \frac{\text{ER} \cdot \text{Tumor size}}{23}\right)$$

Method	AUC-ROC
PREDICT	0.75 ± 0.0033
AutoPrognosis	0.84 ± 0.0032
Metamodel	0.83 ± 0.0020

Explainability = Tailored interpretability

- Different users seek different forms of “understanding”…



Metamodels: How to use them

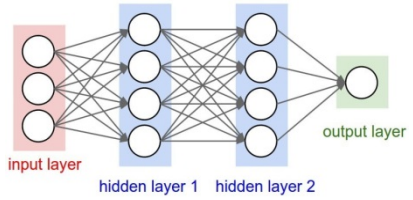
- Different forms of interpretations can be extracted from a Metamodel's forward and backward views!

Forward use

Input= features
Output=risk

- Treatment justification
- Hypothesis induction
- Variable interactions
- Variable importance

Black-box model



Symbolic Metamodel

Metamodel

g(Age, ER, HER2, Tumor size, Nodes)

Backward use

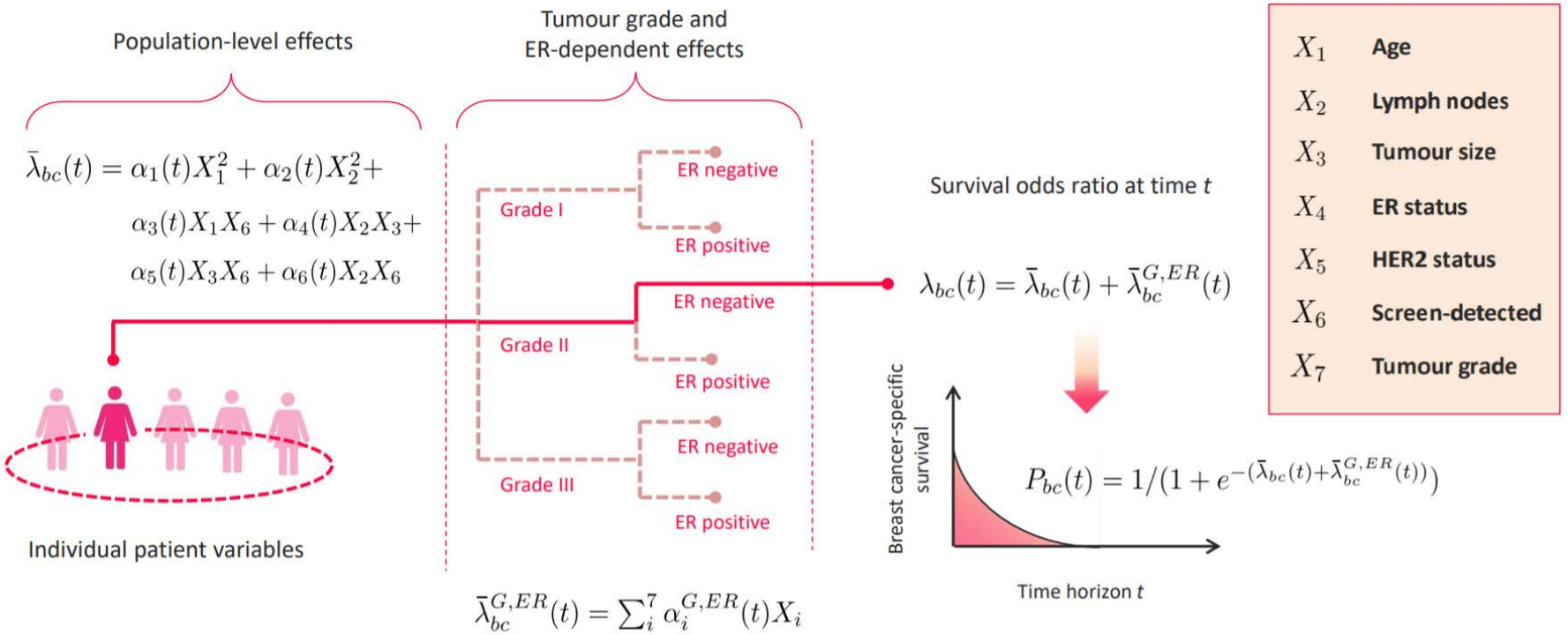
Input= reduced risk
Output=features

- Modifiable variables
- Dosage recommendation
- Policy design

Metamodels for researchers

- Discovery! Data-induced hypothesis

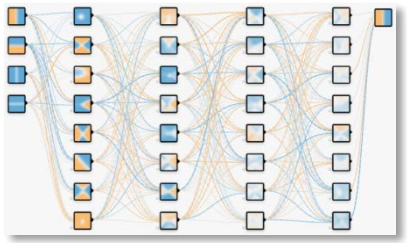
- **Example:** (a) Breast cancer subtype definition can be refined through cancer grades, (b) Risk grows quadratically with Lymph nodes



Metamodels support robust discovery

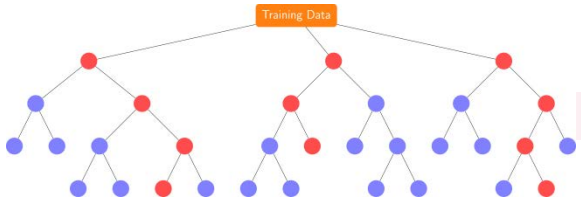
- **Regardless of the model f(x), g(x) is always a symbolic expression**
- **Unified format for many different types of black-box models: identify common discoveries by comparing their Metamodels**

Neural network



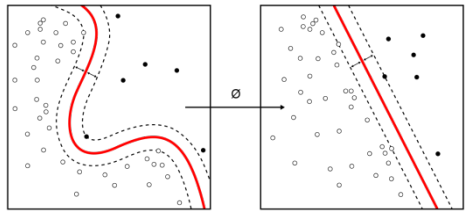
$$\alpha_1 X_1 + \alpha_2 X_2^2 + \alpha_3 X_1 X_2$$
$$\alpha_4 X_3^3 + \alpha_5 \log(X_4)$$

Random forest



$$\alpha_1 X_1 + \alpha_2 X_2^2 + \alpha_3 X_1 X_2$$
$$\alpha_4 X_3^3 + \alpha_5 \log(X_4)$$

SVM

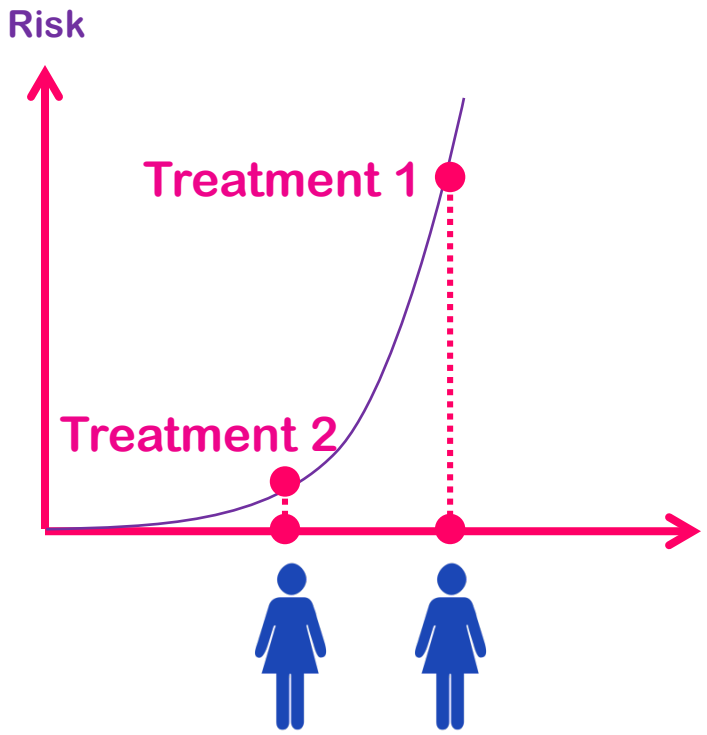


$$\alpha_1 X_1 + \alpha_2 X_2^2 + \alpha_3 X_1 X_2$$
$$\alpha_4 X_3^3 + \alpha_5 \log(X_4)$$

Report common discoveries

Metamodels for clinicians

- Understand why how predictions or treatment recommendations are being made by the ML-model
- Example: Two patients with apparently similar features get different treatment recommendations!



Explanation:

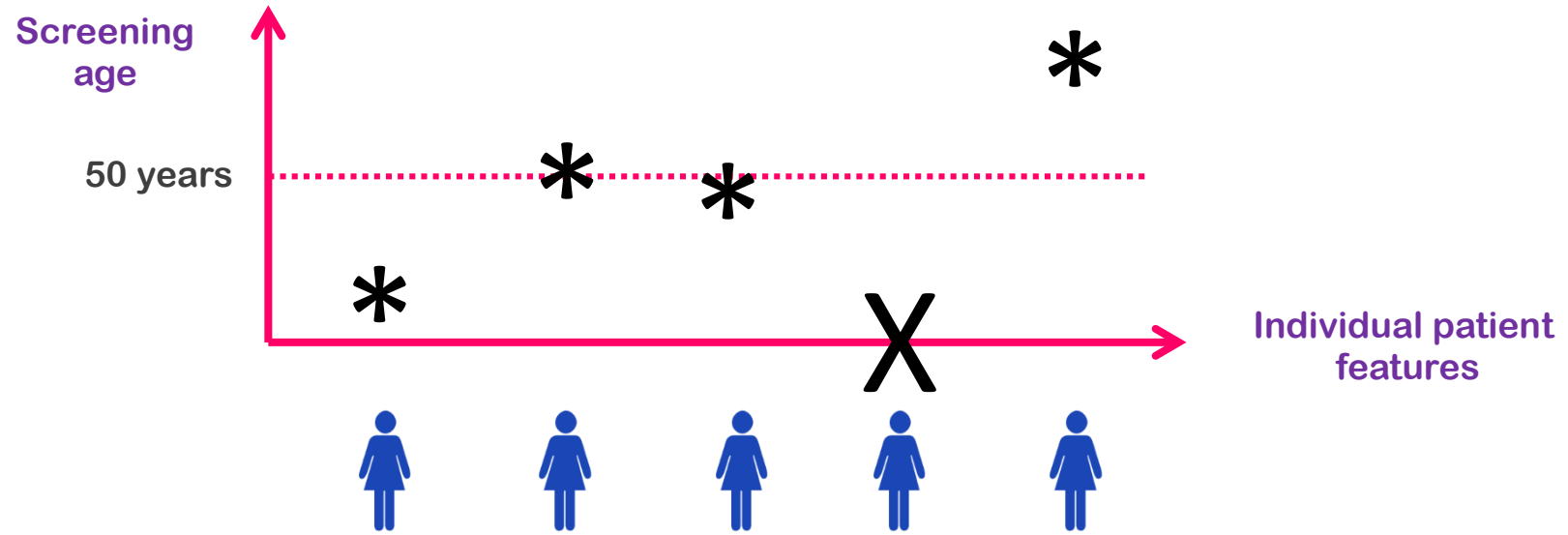
Small changes in number of Lymph node causes a quadratic increase in risk

Metamodels for policy-makers

- Policy makers can be informed how to design more efficient, individualized screening programs (e.g. best age to screen)

Can be set through the inverse Metamodel equation

$$Screening\ age = g^{-1}(Family\ history, Genetics, BMI \mid Risk = X\ \%)$$



Metamodels for patients

- Patients can be informed how to alter behavior to lower risk.

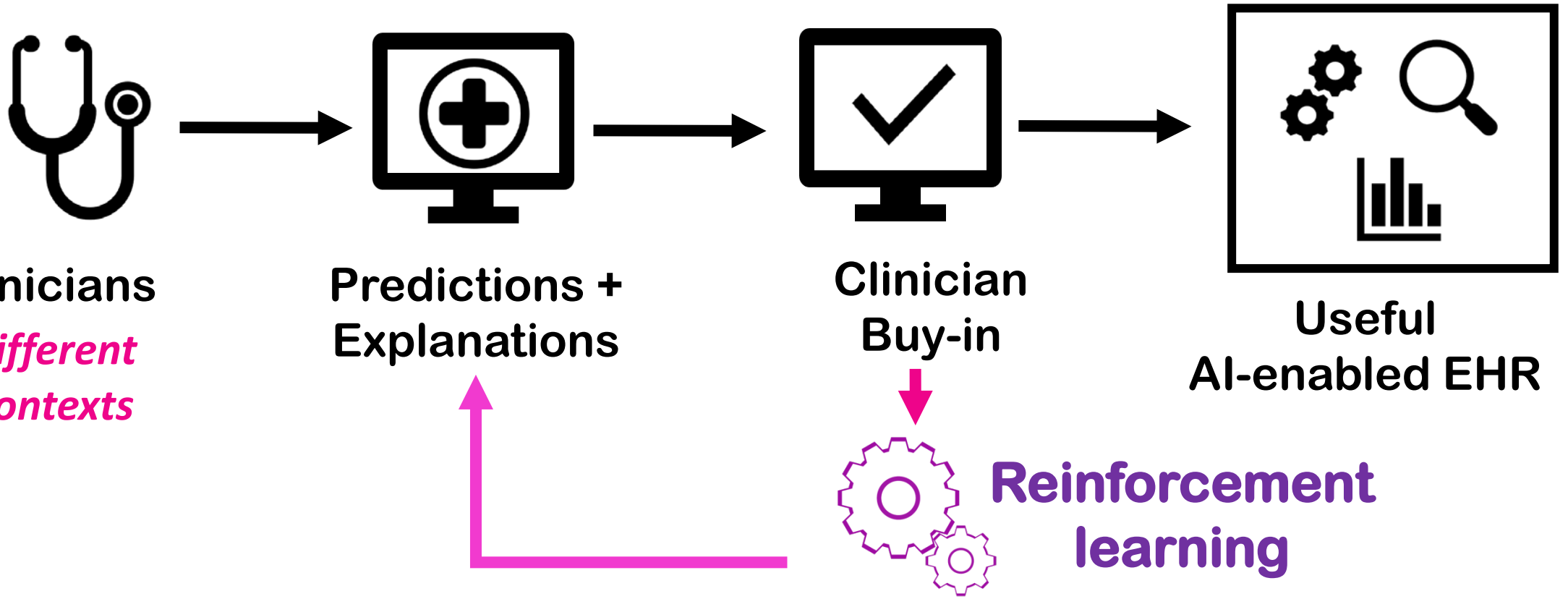
Can be set through the inverse Metamodel equation

$$\text{BMI Reduction} = g^{-1}(\textit{Family history}, \textit{Genetics}, \textit{Diabetes} \mid \textit{Risk} = X\ \%)$$

How to enable clinicians to collaborate with ML-enabled healthcare systems?

We build ‘recommender systems’

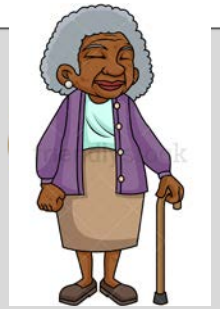
[Lahav, Mastronarde, vdS, NeurIPS ML4Health, 2018]



See more info on
Recommender systems & Personalized education

www.vanderschaar-lab.com/EduAdvance.html

Clinicians and AI working together



- **Patient Betty** is an 86-year-old non-Caucasian female suffering from heart failure
- Betty has a BMI of 21.6
- Betty exhibits rales and shortness of breath at rest

- Our model predicts the probability of Betty dying within 1 year is **83.5%**

Patients most similar to Betty

Patients like Betty who survived

Clinicians and AI working together

- 19 cardiologists
  - Different sub-specialties
  - Different types of patients
  - 4 different countries (USA, UK, Italy, Netherlands)

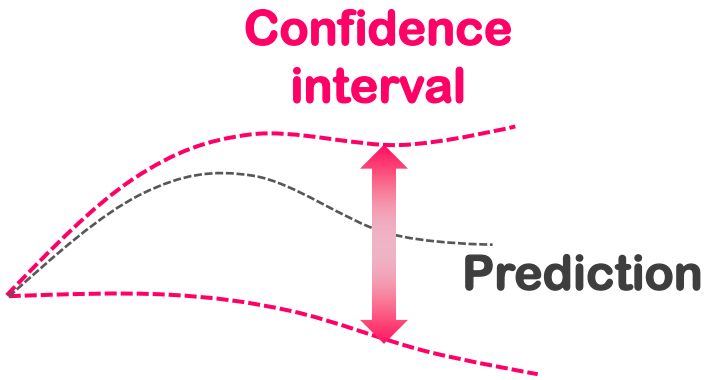
  - Different age groups (38 to 75)
  - Different familiarity with AI

  - Trustworthiness is key!

Building trust into predictions

- ML/deep learning methods achieve high accuracy, but…
  - they fall short in quantifying the uncertainty in their predictions!

- Uncertainty quantification is crucial



High-confidence prediction

Low-confidence prediction

ML informs individualized decision making!

- Model is not equally confident in every prediction it makes!
- Need to know what we do know and what we do not know!

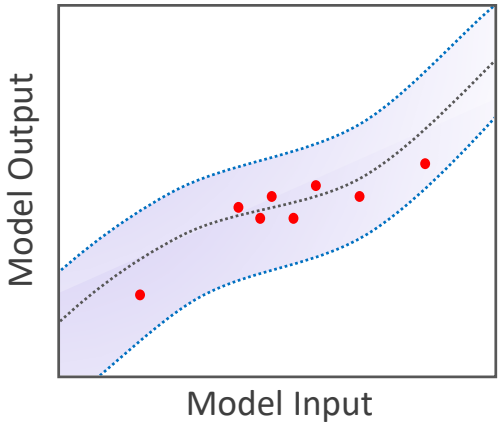
Quantifying uncertainty

- Need estimates of uncertainty that satisfy two key criteria

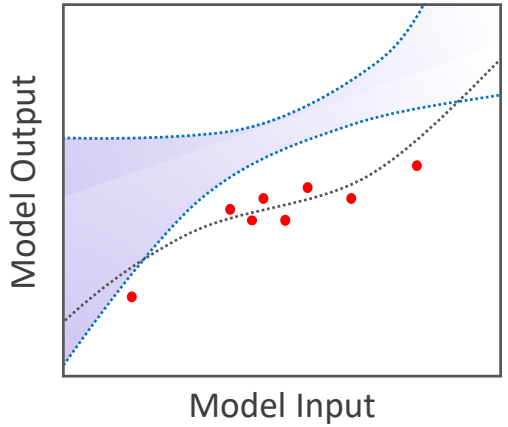
Coverage

Uncertainty intervals contain the true outcome with high probability

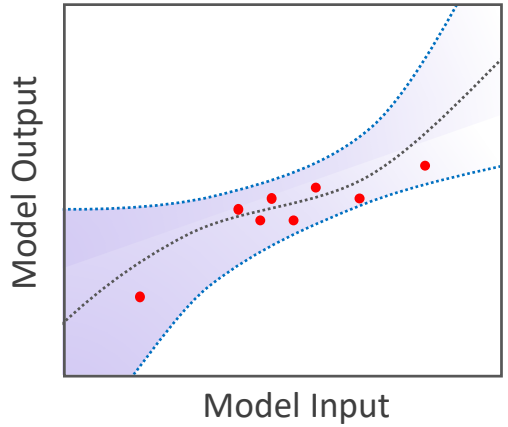
Accurate coverage



No coverage



Accurate coverage

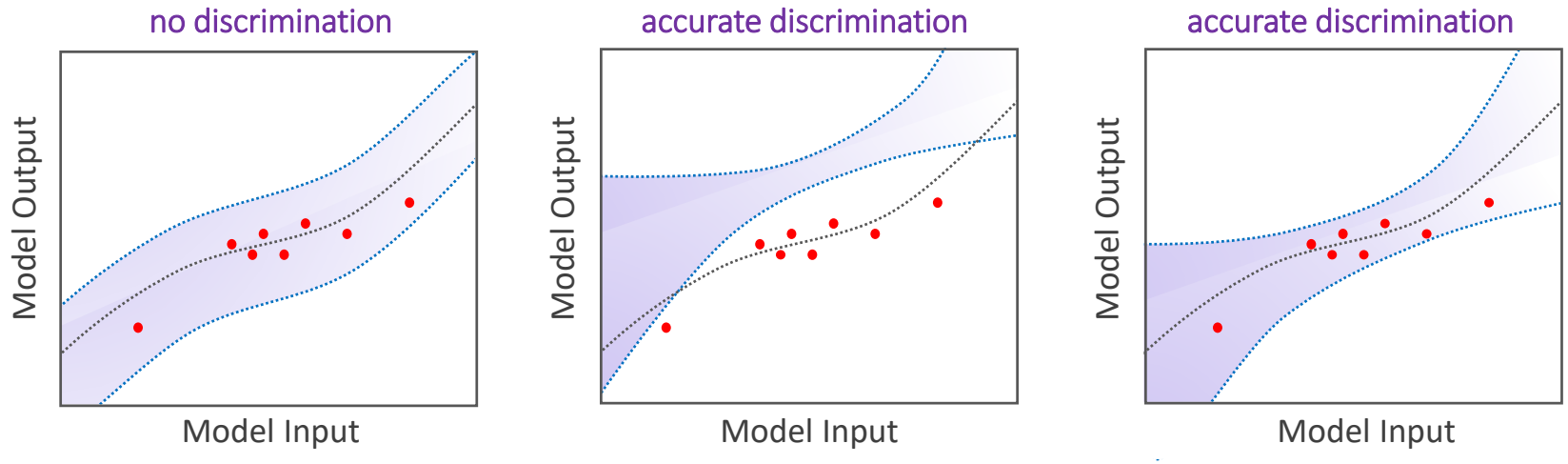


Quantifying uncertainty

- Need estimates of uncertainty that satisfy two key criteria

Discrimination

Width of uncertainty interval reflects level of confidence



Quantifying uncertainty

- Need estimates of uncertainty that satisfy two key criteria

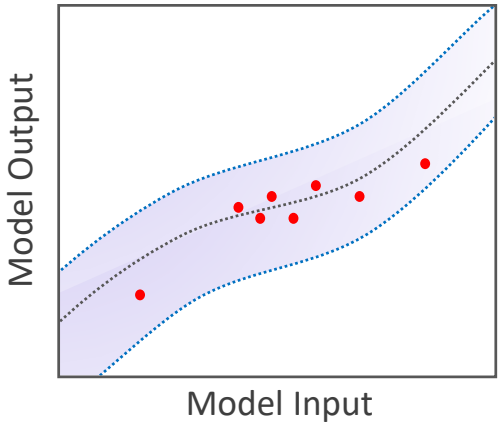
Coverage

Uncertainty intervals contain the true outcome with high probability

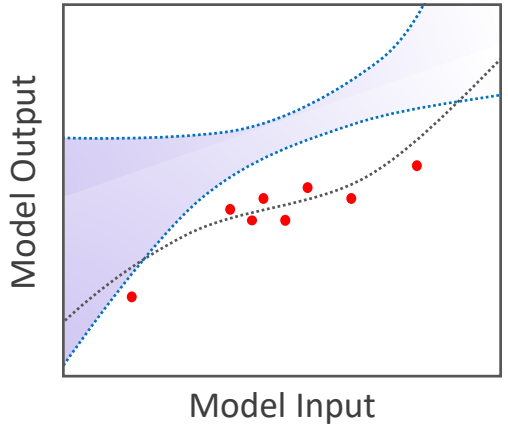
Discrimination

Width of uncertainty interval reflects level of confidence

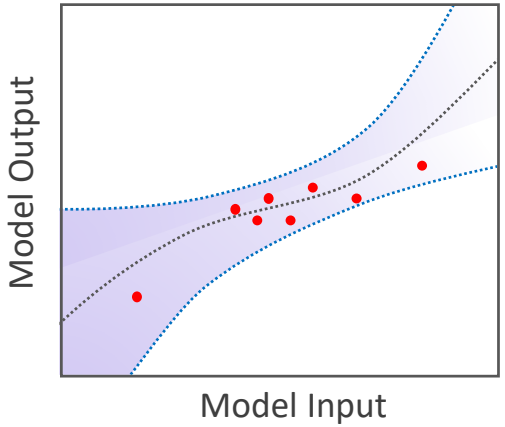
Accurate coverage, no discrimination



No coverage, accurate discrimination

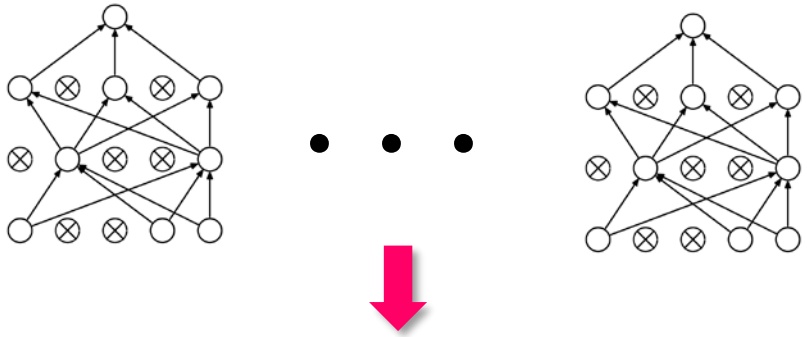


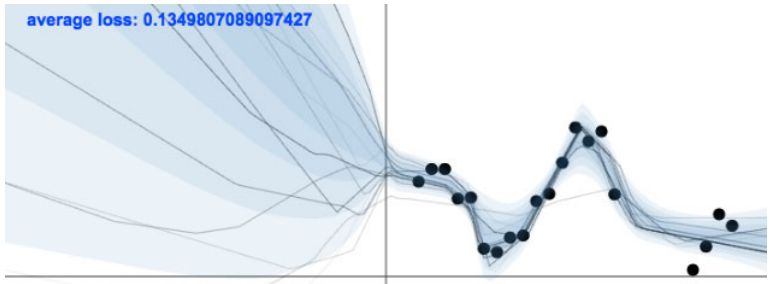
Accurate coverage and accurate discrimination



State-of-the-art approaches

- Bayesian approach
  - Monte Carlo dropout = Variational inference [Gal & Ghahramani, 2015]

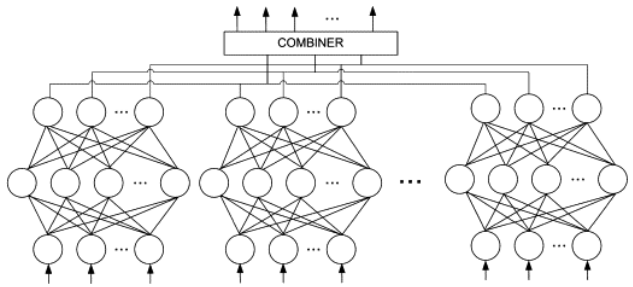




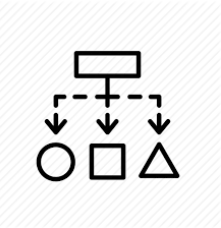
Not post-hoc

No coverage guarantees…

- Deep ensembles
  - Explicitly train many networks and assume the prediction is their average.



Our approach: Post-hoc methodology with frequentist coverage guarantees



	Method	Post-hoc vs Built-in	Coverage
	Bayesian neural nets (Ritter et al., 2018)	Built-in	No guarantees
	Probabilistic backprop. (Blundell et al., 2015)	Built-in	No guarantees
	Monte Carlo dropout (Gal & Ghahramani, 2016)	Built-in	No guarantees
	Deep Ensembles (Lakshminarayanan et al., 2017)	Built-in	No guarantees
*	**Discriminative Jackknife (Alaa and vdS,ICML2020)**	**Post-hoc**	**1-α**

- Does not interfere with model training or compromise accuracy!

Alaa and vdS,ICML2020 - Frequentist Uncertainty in Recurrent Neural Networks via *Blockwise* Influence Functions

Acknowledgements

Dr. Ahmed Alaa
Dr. Hyunsuk Lee
Jinsung Yoon
Alexis Bellot
Changhee Lee
Yao Zhang
Ioana Bica
Zhaozhi Qian
Trent Kyono
James Jordon
Dan Jarrett
Alihan Huyuk

Email: mv472@cam.ac.uk
Website: http://www.vanderschaar-lab.com/

http://www.vanderschaar-lab.com