Medicine 2.0:
Using Machine Learning to Transform Medical Practice and Discovery

Mihaela van der Schaar
University of Oxford
The Alan Turing Institute

"to make great leaps in data science research in order to change the world for the better."

Acknowledgements

- Research support
  - ONR
  - Alan Turing Institute
  - UK Cystic Fibrosis Thrust

- PhD students
  - Ahmed Alaa
  - Jinsung Yoon

Acknowledgements

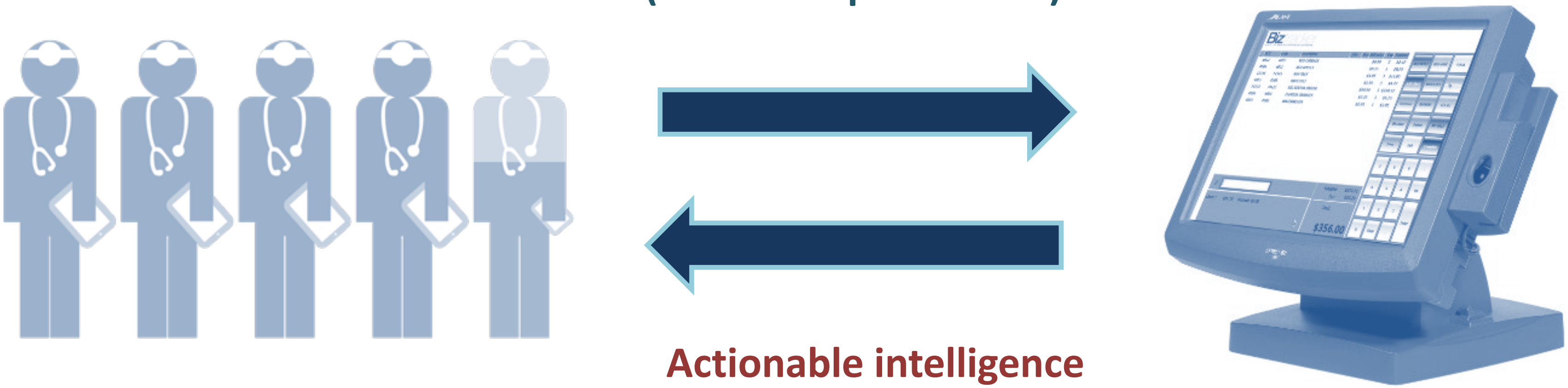
- Clinicians
  - – Dr. Amitava Banerjee (Cardiology)
  - – Dr. Raffaele Bugiardini (Cardiology)
  - – Dr. Martin Cadeiras (Cardiology)
  - – Dr. Camelia Davtyan (Internal Medicine)
  - – Dr. Steve Harris (Intensive Care)
  - – Dr. Scott Hu (Intensive Care)
  - – Dr. Dan Lasserson (Chronic Disease)
  - – Dr. Luke Macyszyn (Neurosurgery)
  - – Dr. Paolo Puddu (Cardiology)
  - – Dr. Mindy Ross (Asthma)

Machine Learning & Medicine

Vision: capitalize on increasing availability of data to extract ***actionable intelligence*** in order to improve clinical practice (saves lives, reduces costs) and advance medical discovery

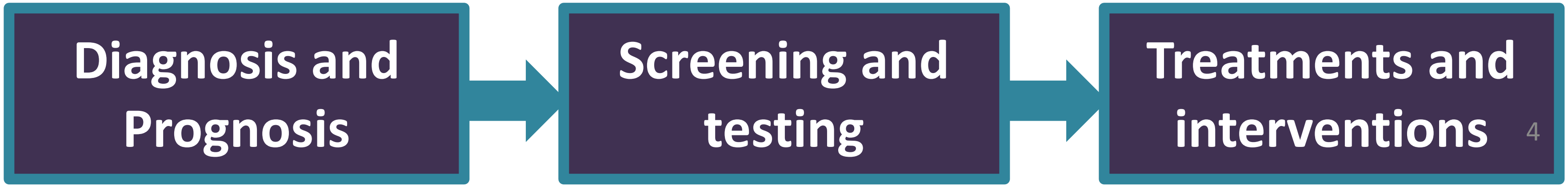
Healthcare practice = Observational data

(Natural experiments!)



Actionable intelligence

(Predictions, recommendations, practice guidelines, treatment effects, etc)



Diagnosis and Prognosis

Screening and testing

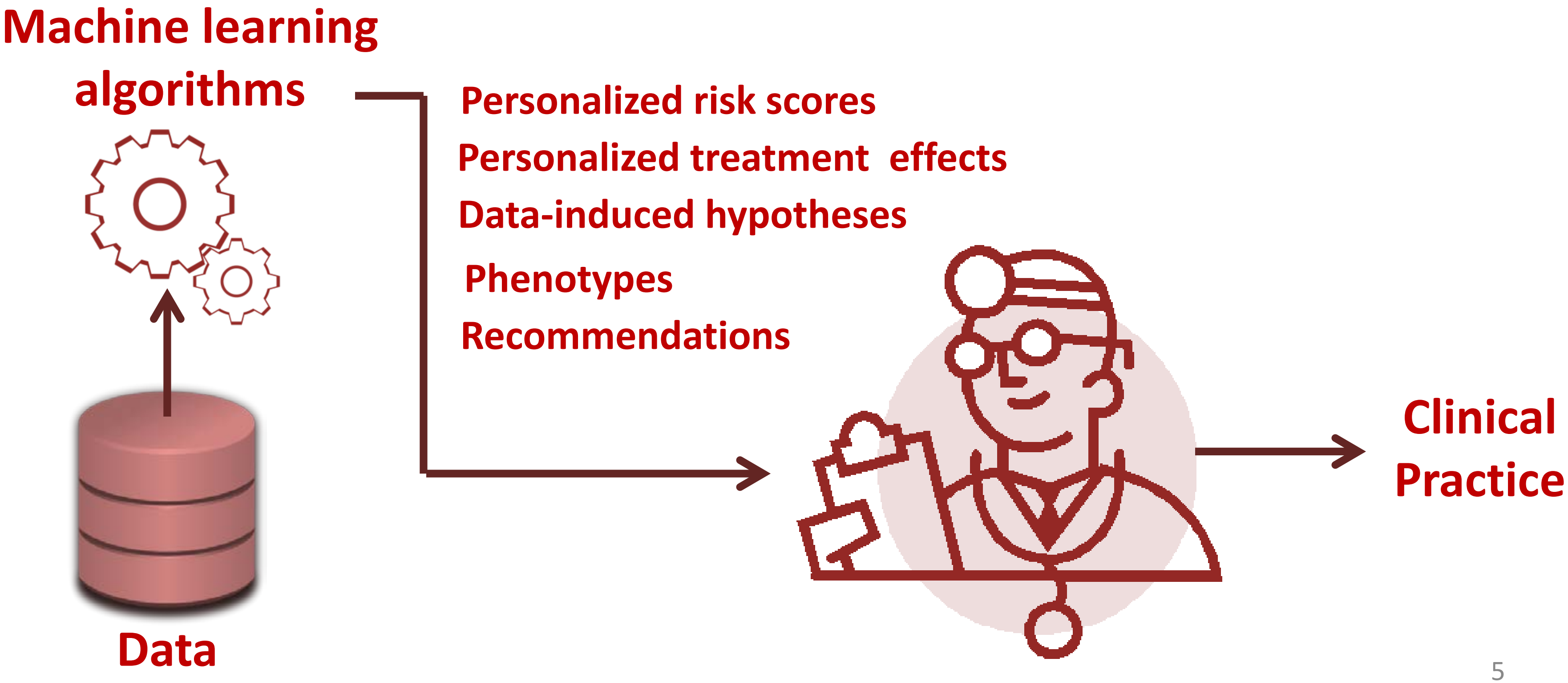
Treatments and interventions

The “Augmented” MD

- Machine learning

  …***can’t*** do medicine!

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



Machine learning algorithms

Personalized risk scores

Personalized treatment effects

Data-induced hypotheses

Phenotypes

Recommendations

Data

Clinical Practice

New Tools and Methods

- Learning/decision making
  - from time-series data
  - from many kinds of data (images, vital signs, etc.)
- Causal inference
- Graphical models
- Reinforcement learning
- Deep learning

Long Road ....
Some Steps Along the Way

- Individualized treatment effects
- Risk scoring for critical care
    - Problem and why it is important
    - Current solutions and limitations
    - New solutions and impact

Individualized Treatment Effects

- Most treatments have **different effects** for **different patients**
- Not enough to know that the treatment **works well on average**, need to know its effect on an **individual**!

Which treatment should be used for *this* patient?

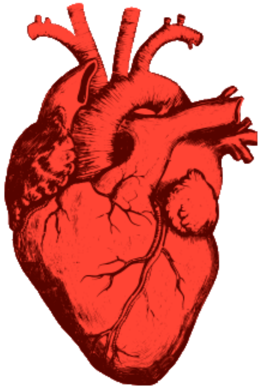
- **chemotherapy regime, medication, type of surgery ...**

Use machine learning to estimate individualized treatment effects from observational data *without* using clinical trials

- **why so important?**

Who should get a heart?

Ann



Bob

- **Factual outcome**
  - How long will Ann/Bob survive while waiting?
- **Counterfactual outcome**
  - How much will Ann/Bob benefit from this heart had she/he got it ?

Evaluation on Real-World Data

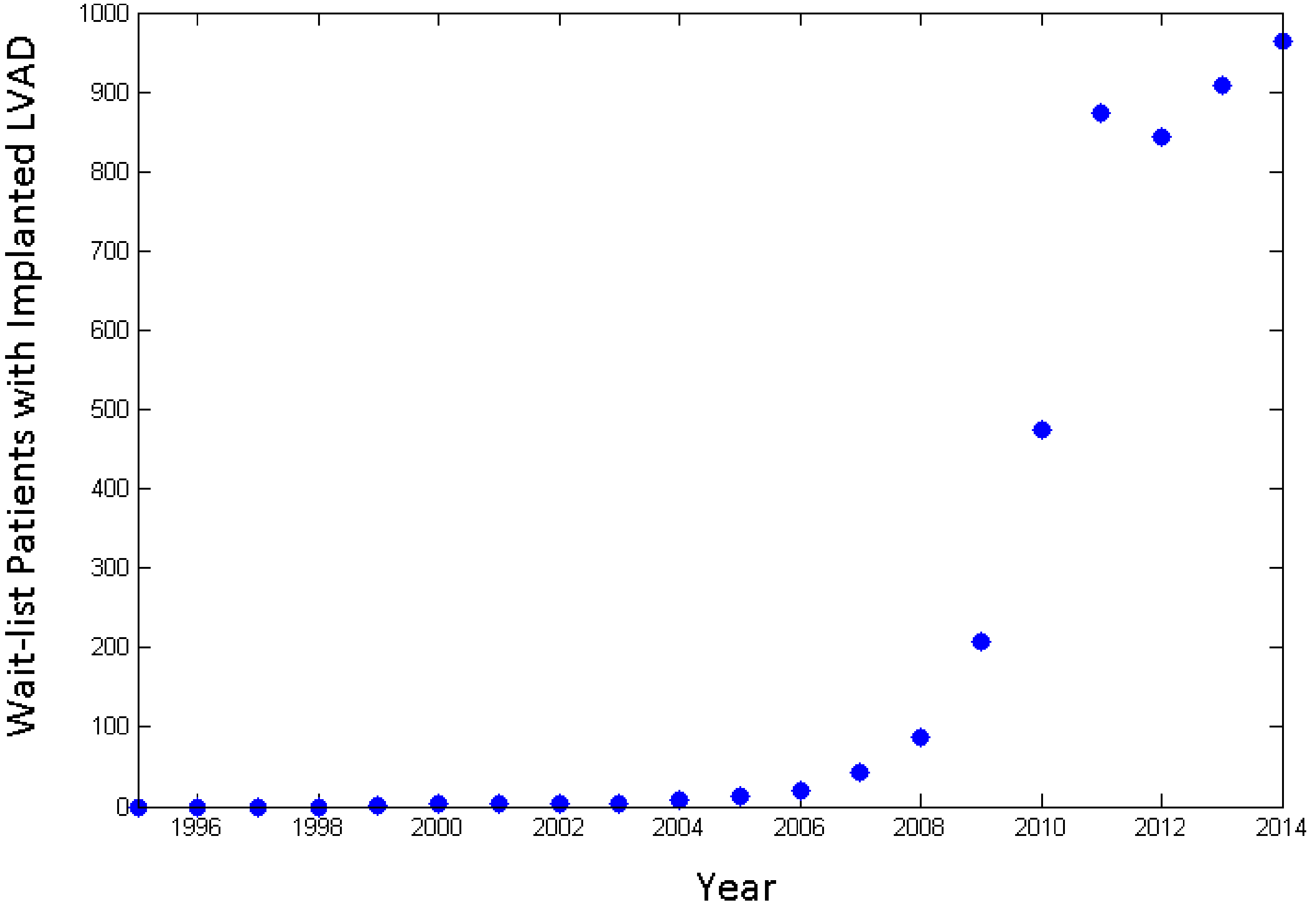
United Network for Organ Transplantation (UNOS)

- ALL patients registered for heart transplantation in US in **1985-2015**
- **60,000+** patients received heart transplant
- **35,000+** patients wait-listed but did not receive heart transplant
  - Date of waitlisting + survival
  - **33** features of patients

Intervention:

LVAD

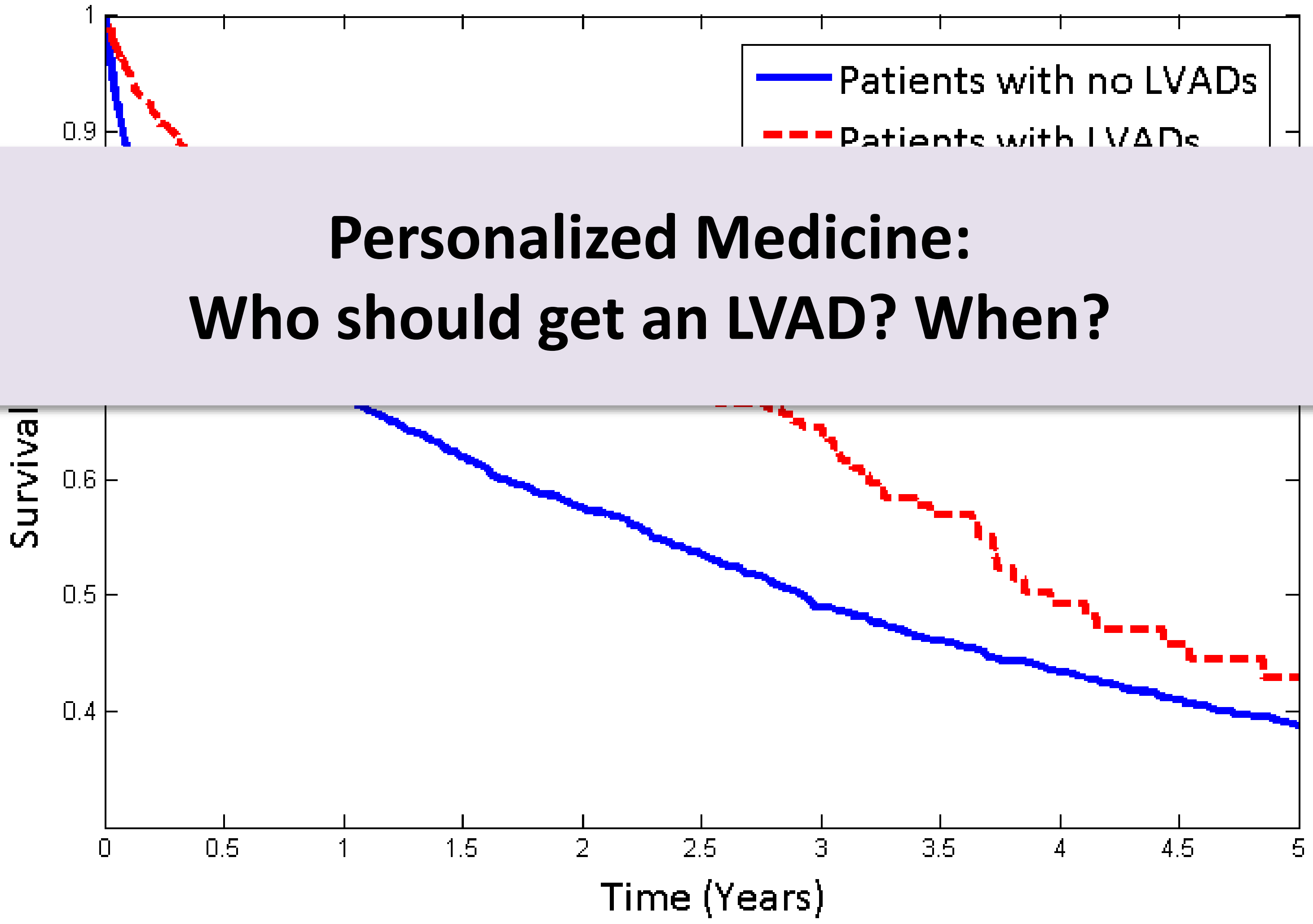
Number of LVADs increases in past decade



2017:

LVAD implantation cost **$175,000** for the procedure but carried a 6-year total price tag of **$726,000**

Population-level Survival Benefit of LVADs: Kaplan-Meier Estimates



Personalized Medicine:
Who should get an LVAD? When?

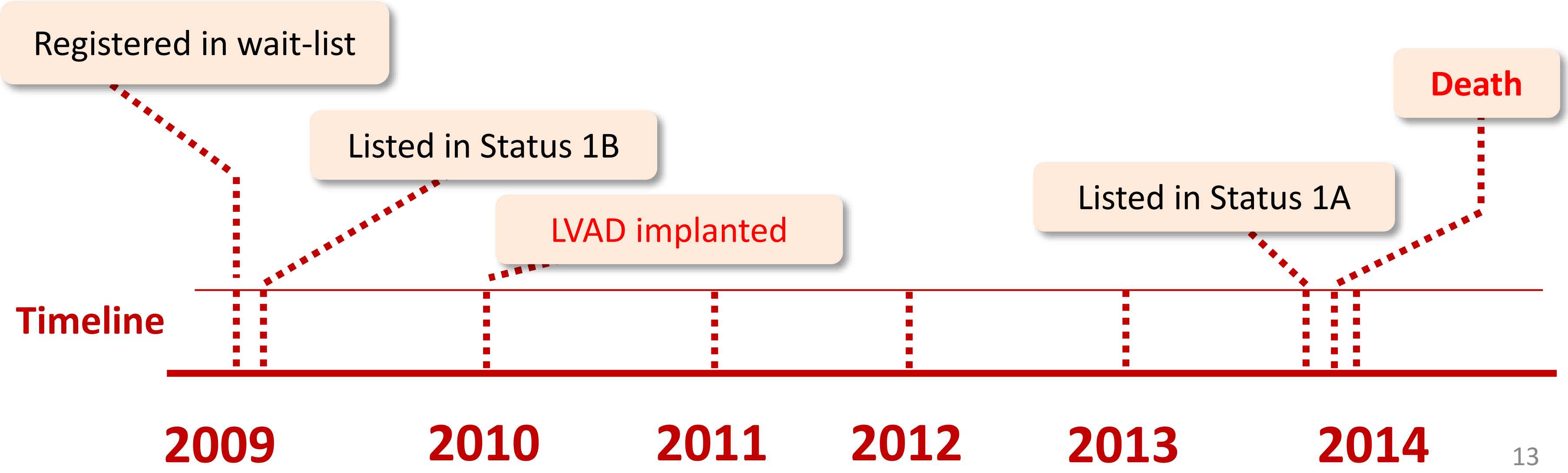
Life and Death for One Patient

A young diabetic patient in the wait-list had an LVAD implanted.

Her expected LVAD survival benefit was overestimated and she died before getting a transplant!

Covariates	
Age	34
Gender	Female
Comorbidities	Diabetes

What would have happened had we got a personalized estimate?

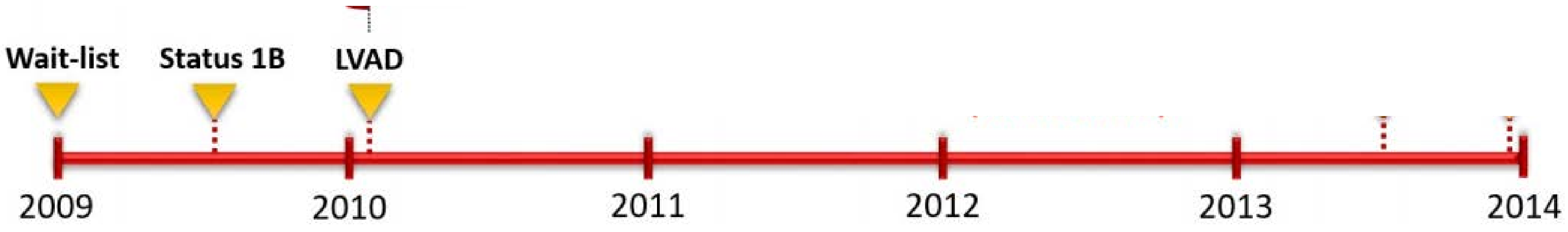


Life and Death for One Patient

This patient was assigned a low priority because survival was estimated based on the average ("population") estimate of LVAD benefits!

Personalized Estimate: For this specific patient, the posterior average survival benefit -> early 2013!

Individualized Estimate



Estimating Causal Effects from Observational Data

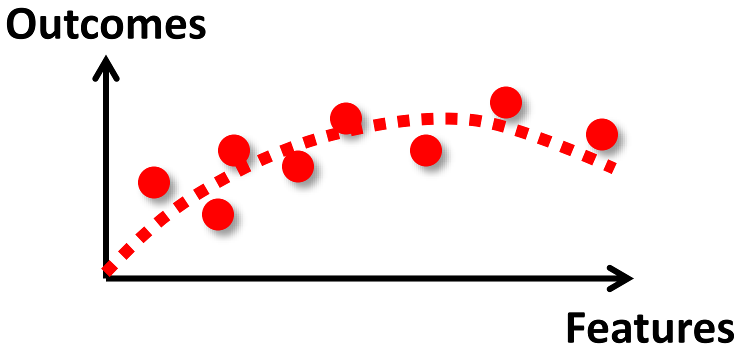
- Most works on causal inference focused on answering the following question: does ***X*** cause ***Y*** (***X***→***Y***)? **[Judea Pearl]**
- A coarse binary hypothesis!
- Does not quantify "context-specific" magnitude of causal effect
- **Much less work has focused on estimating the magnitude of the effect of *X*** on ***Y for an* individual subject *given his/her features!***

- **Individual-level inference of causal effects is a key problem in the area of precision medicine**
- **Recent advances in machine learning can estimate granular causal effects from observational data**

Not a conventional supervised learning problem!

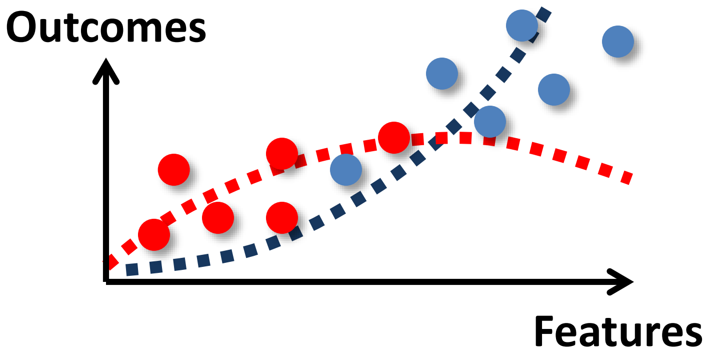
- **Observational data:** we only observe factual outcomes of treatment assignment, but we need counterfactual outcomes to estimate causal effects.

Supervised Learning



The goal is to estimate the underlying true function ▪▪▪▪▪ given the training examples ●

Causal Inference

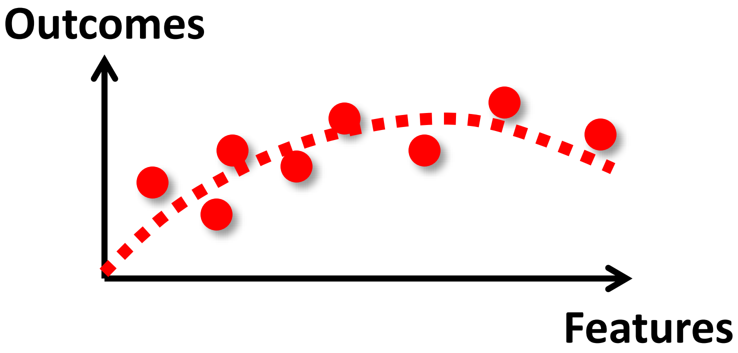


The goal is to estimate the difference between the true responses ▪▪▪▪ and ▪▪▪▪ given the factual outcomes of treated ● and untreated ● subjects

Not a conventional supervised learning problem!

- **Observational data:** we only observe factual outcomes of treatment assignment, but we need counterfactual outcomes to estimate causal effects.
- **Selection bias!**

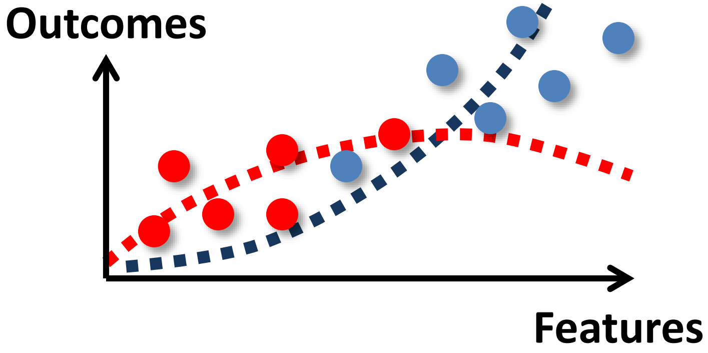
Supervised Learning



Straightforward regression problem

Performance can be assessed by out-of-sample testing

Causal Inference



Unbalanced dataset with unobserved outcomes

Inference problem: need a measure of uncertainty

Observational Data, not Randomized Trials

Observational EHR data:

$$\mathcal{D} = \left(X_i, W_i, W_i \cdot Y_i^{(1)} + (1 - W_i) \cdot Y_i^{(o)} \right)_{i=1}^{n}$$

Feature

Treatment assignment

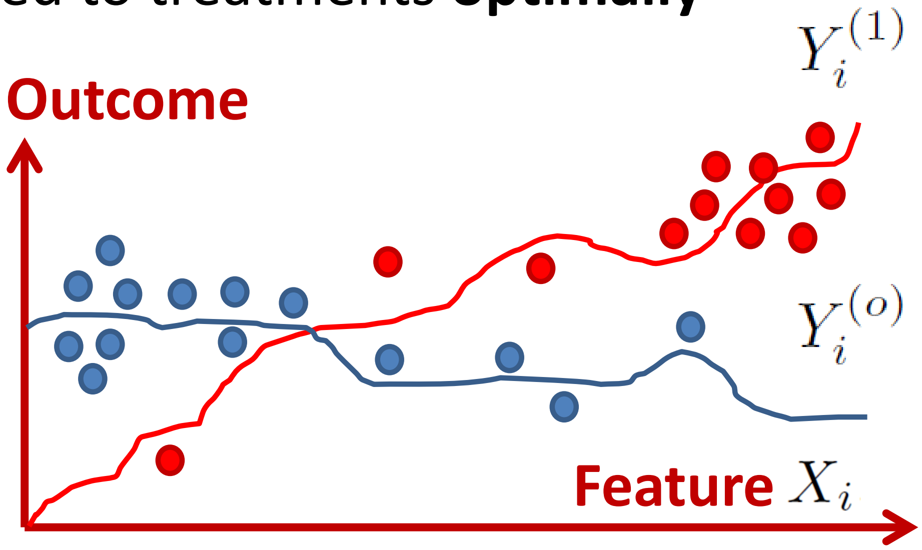
Treatment outcome

Current clinical practice:

- Patients not assigned to treatments **randomly**
- Patients (probably) not assigned to treatments **optimally**

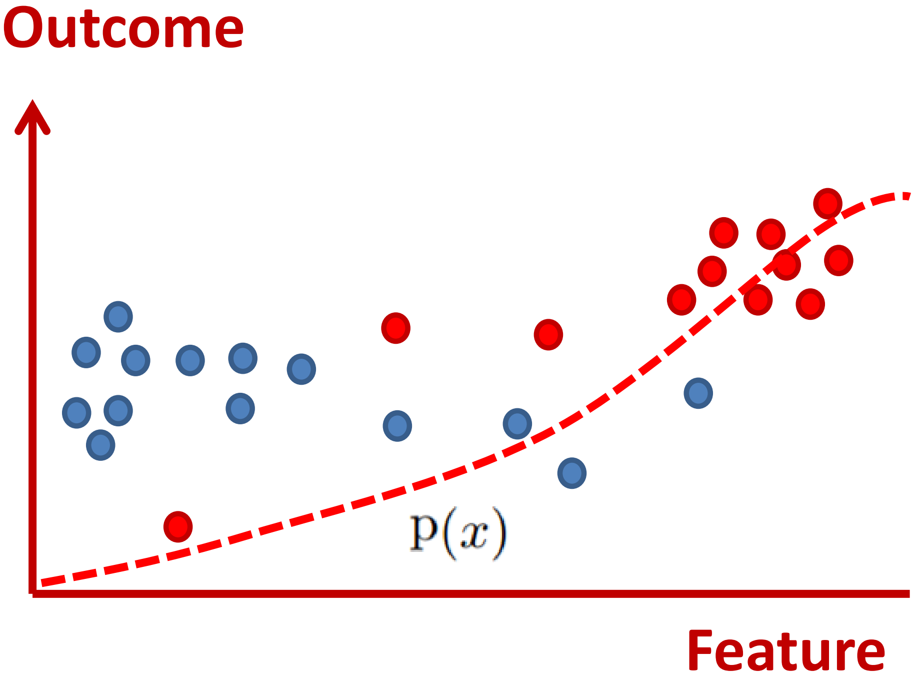
Treatment effect

$$T(x) = \mathbb{E}\left[Y_i^{(1)} - Y_i^{(o)} \middle| X_i = x \right]$$



Estimating *Average* Treatment Effects

Most medical studies estimate average treatment effects **->** **Solved problem!**



Estimate propensity score (e.g. using logistic regression)

$$\mathrm{p}(x) = \mathbb{E}\left[W_i \mid X_i = x\right]$$

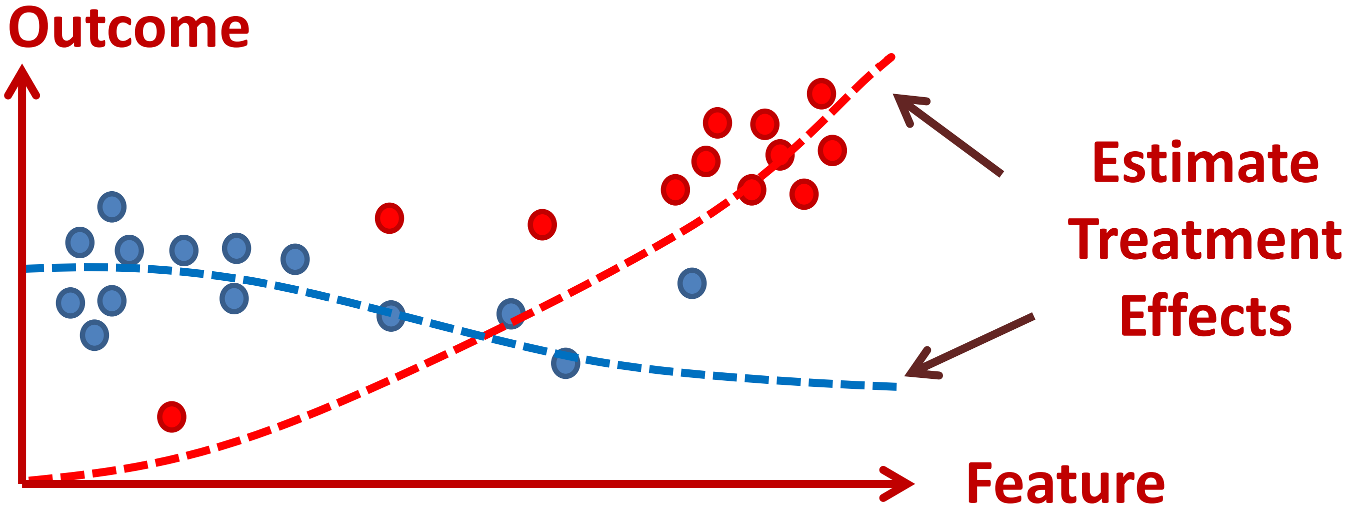
Unbiased estimator for the average treatment effect

$$\mathbb{E}\left[Y_i\left(\frac{W_i}{\mathrm{p}(x)} - \frac{1-W_i}{1-\mathrm{p}(x)}\right) \middle| \, X_i = x\right] = \mathrm{T}(x)$$

Estimating *Individualized* Treatment Effects

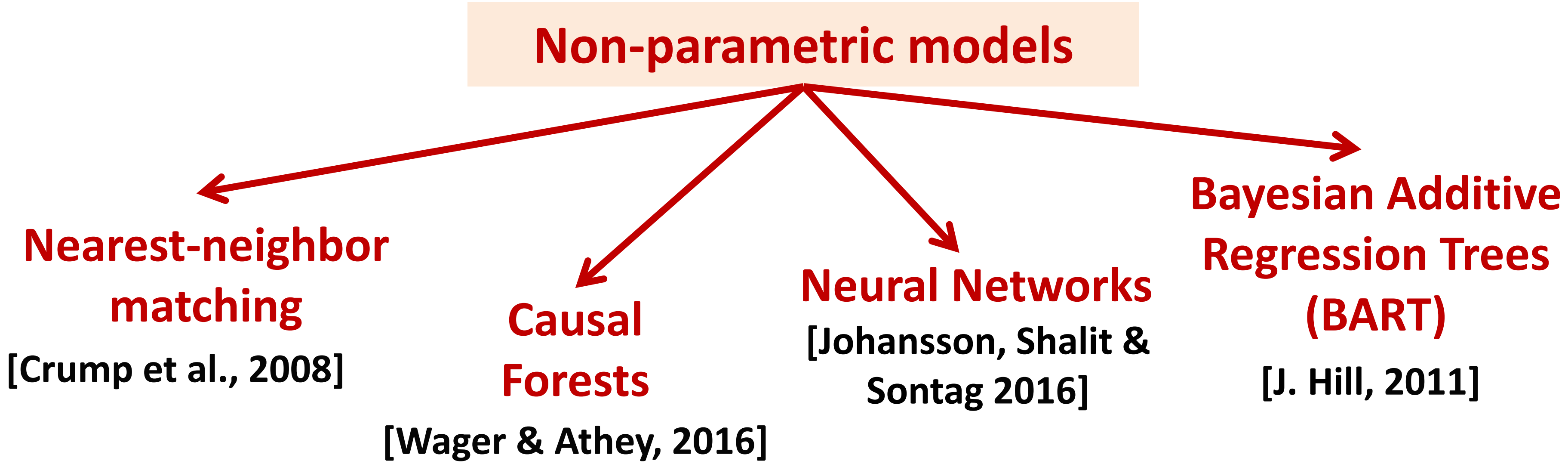
Response surface modeling/covariate adjustment:

- **for each outcome: data -> estimate a model for that outcome**
- **difference of outcomes = treatment effect**
- **difference of models = estimate of treatment effect**



Individualized Treatment Effects – State-of-the-art

Complexity of non-parametric models grows with the amount of available data (heterogeneous populations)



Non-parametric models

Nearest-neighbor matching

[Crump et al., 2008]

Causal Forests

[Wager & Athey, 2016]

Neural Networks

[Johansson, Shalit & Sontag 2016]

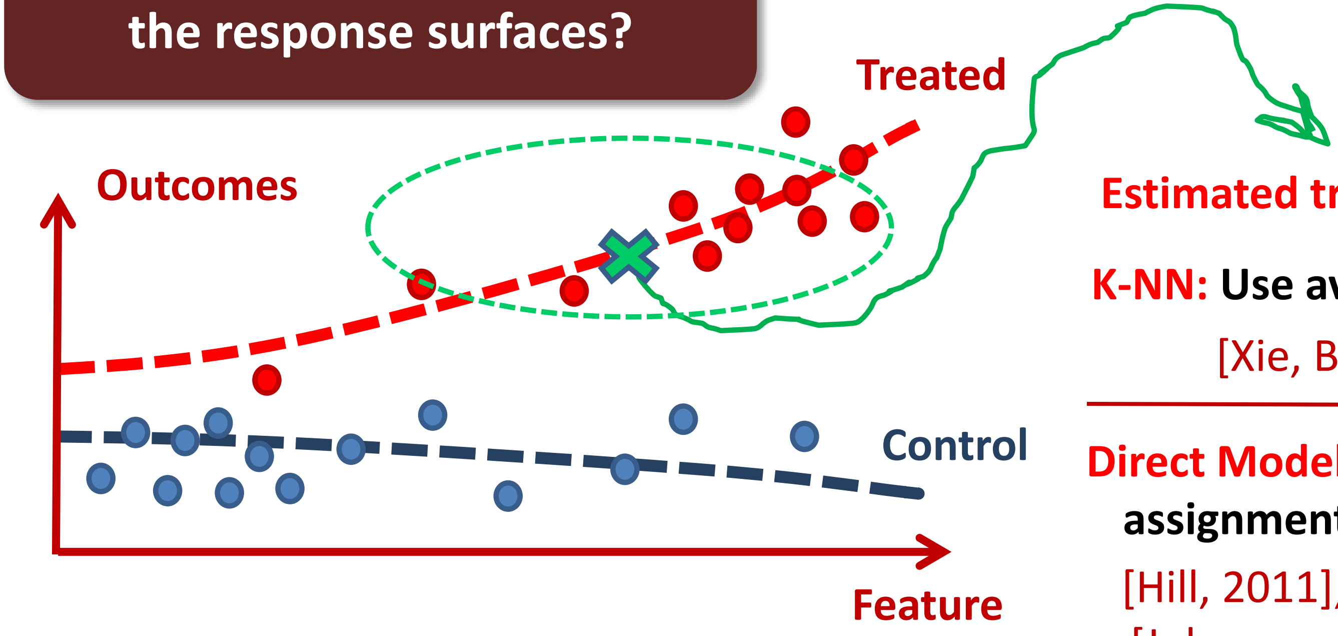
Bayesian Additive Regression Trees (BART)

[J. Hill, 2011]

Our method improves on these methods by using a multi-task learning approach!

Individualized Treatment Effects – State-of-the-art (II)

How did previous works model the response surfaces?



Estimated treatment outcome =

K-NN: Use average of *k* neighbors

[Xie, Brand, Jan, 2012]

Direct Modeling: Model treatment assignment as an input feature

[Hill, 2011], [Wager & Athey, 2016]
[Johansson, Shalit & Sontag 2016]

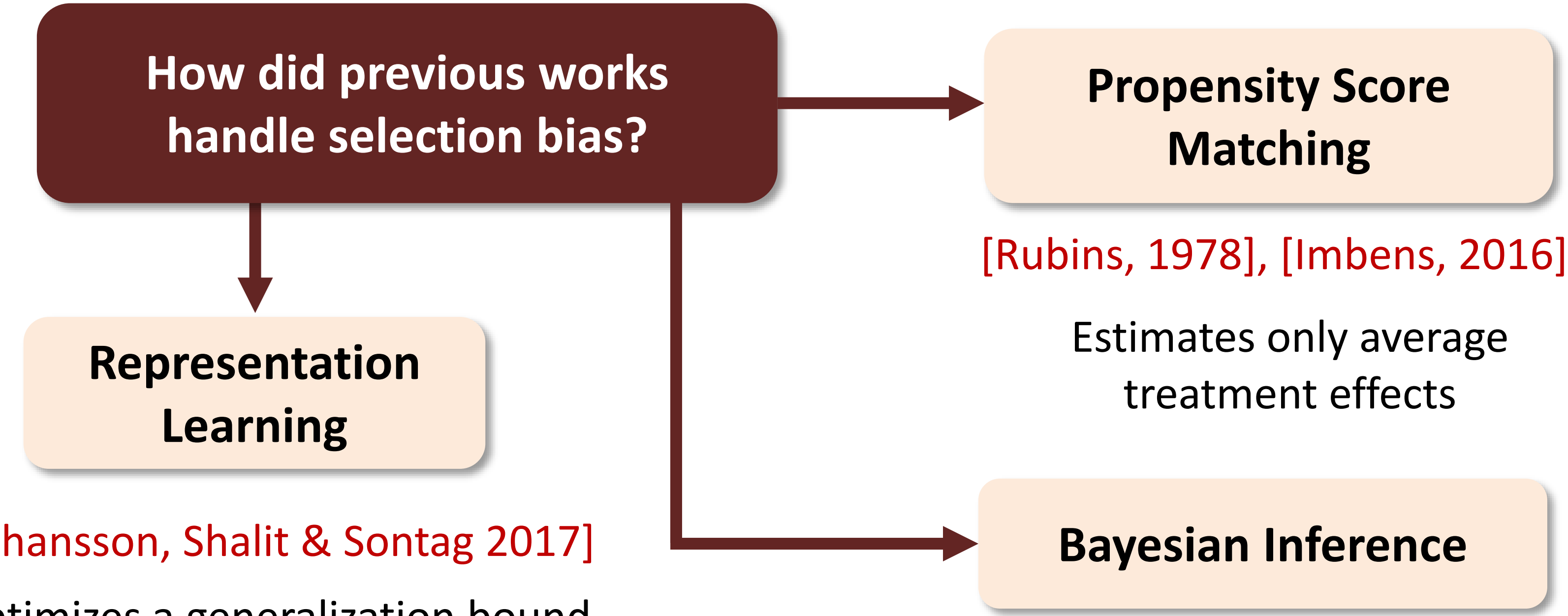
Virtual Twins: Fit separate regression models for treated and control populations

[Lu et al., 2017]

Previous methods ignore similarity of learning tasks

Multi-task Learning provides statistical efficiency

Individualized Treatment Effects – State-of-the-art (III)



How did previous works handle selection bias?

Propensity Score Matching

[Rubins, 1978], [Imbens, 2016]

Estimates only average treatment effects

Representation Learning

[Johansson, Shalit & Sontag 2017]

Optimizes a generalization bound on the true loss function that depends only on factual outcomes

Bayesian Inference

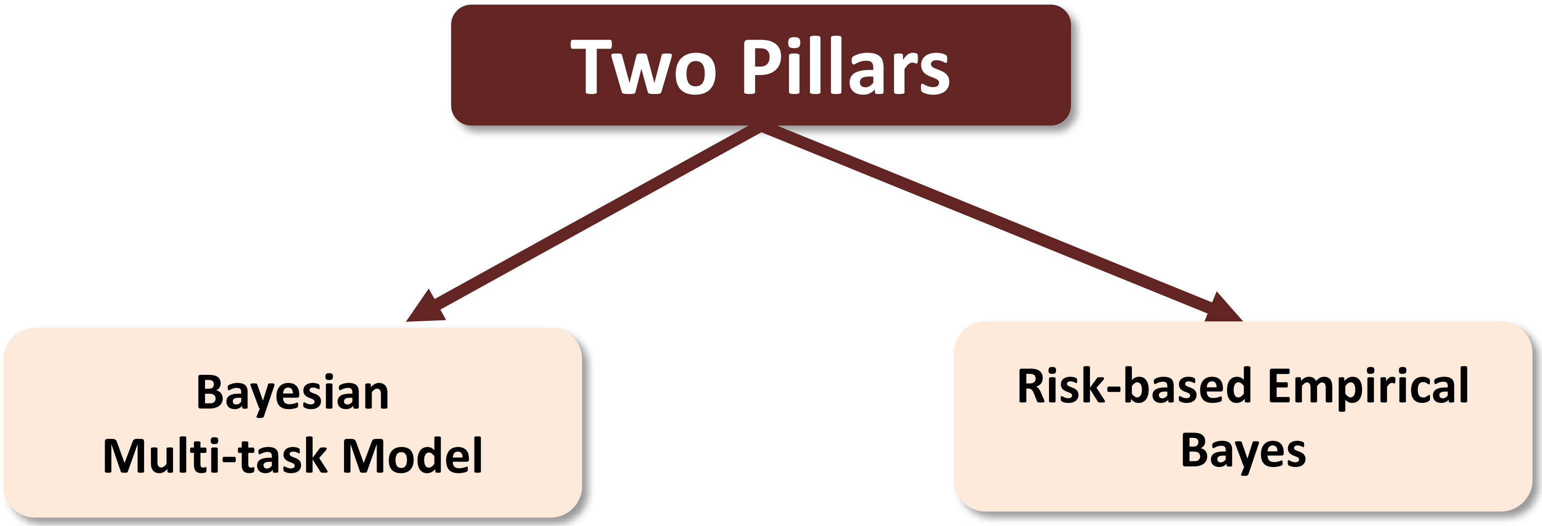
[Hill, 2011]

Tunes prior based only on factual outcomes

Our approach: Risk-based Empirical Bayes

We tune a *multi-task prior* to minimize the expected loss in *both* factual and counterfactual outcomes

How do we learn more effectively?



Two Pillars

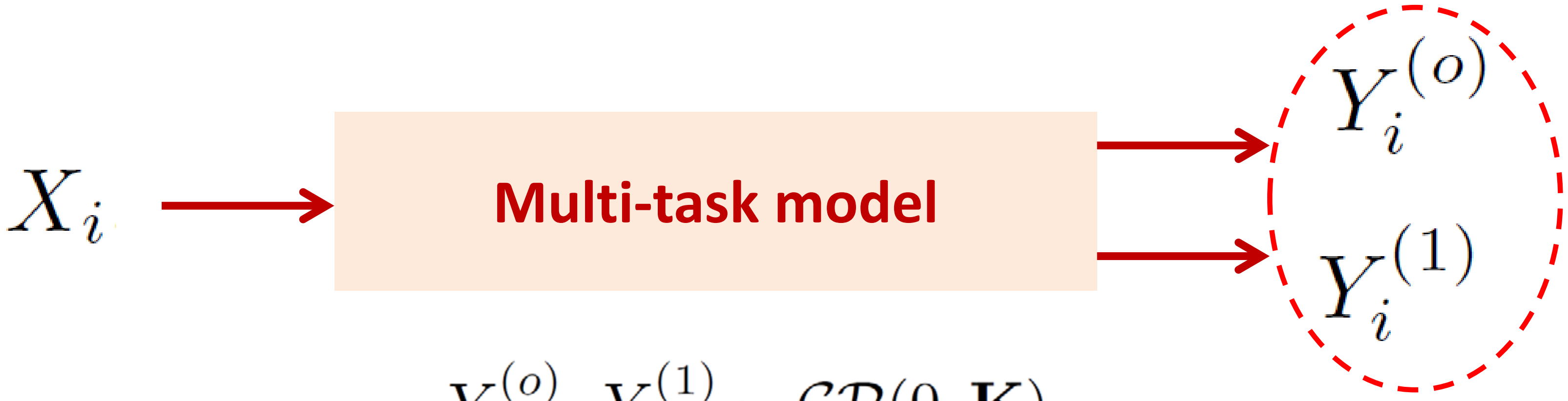
Bayesian Multi-task Model

Risk-based Empirical Bayes

- **Flexibility:** nonparametric interactions between covariates and treatment assignment
- **Data efficiency:** treated and control models have shared parameters

Selection bias handled by tuning prior so as to minimize posterior variance of counterfactuals

Multi-task Learning for Causal Inference (I)



X_i

Multi-task model

$Y_i^{(o)}$

$Y_i^{(1)}$

$$Y_i^{(o)}, Y_i^{(1)} \sim \mathcal{GP}(0, \mathbf{K})$$

Use a multi-task Gaussian process prior on the potential outcomes!

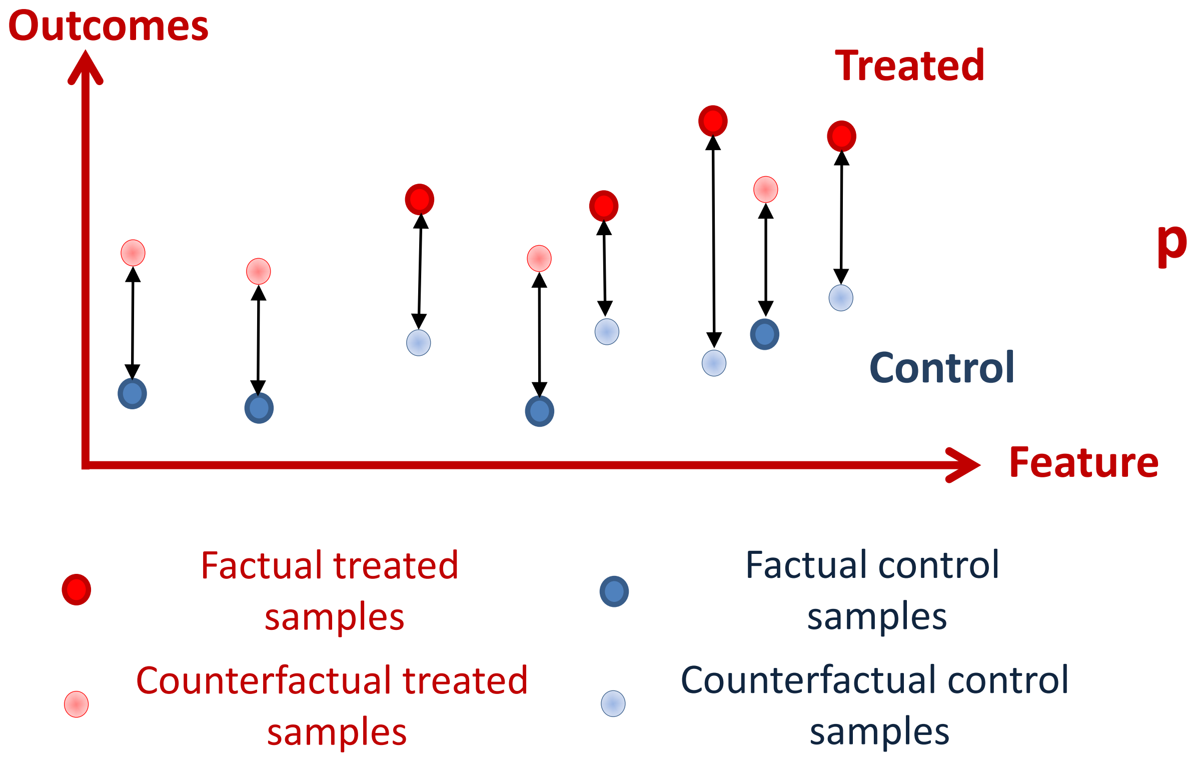
Multi-task Learning for Causal Inference (II)

Construct a “proxy” for the *error* in estimated treatment effect

Bayesian Risk →

$$R(\theta,\hat{\mathbf{f}};\mathcal{D}) = \mathbb{E}_\theta\left[\hat{\mathcal{L}}(\hat{\mathbf{f}};\mathbf{K}_\theta,\mathbf{Y}^{(\mathbf{W})},\mathbf{Y}^{(1-\mathbf{W})})\,\middle|\,\mathcal{D}\right]$$

Factuals **Counterfactuals**



Bayesian framework provides estimates of ITE through the posterior counterfactual distribution

Risk-based Empirical Bayes (I)

$$R(\theta, \hat{\mathbf{f}}; \mathcal{D}) = \mathbb{E}_\theta \left[\hat{\mathcal{L}}(\hat{\mathbf{f}}; \mathbf{K}_\theta, \mathbf{Y}^{(\mathbf{W})}, \mathbf{Y}^{(\mathbf{1}-\mathbf{W})}) \,\middle|\, \mathcal{D} \right]$$

Kernel Hyper-parameters

Theorem

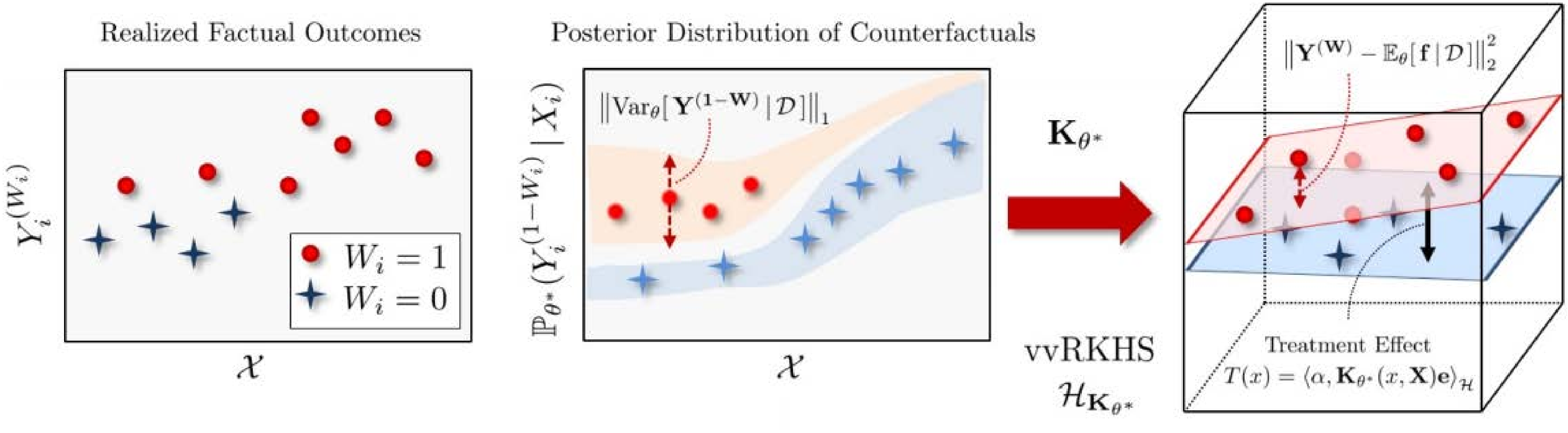
Optimal Kernel for the Prior in terms of Bayesian risk?

Optimal prior

$$\theta^* = \arg\min_{\theta \in \Theta} \left[\underbrace{\left\| \mathbf{Y}^{(\mathbf{W})} - \mathbb{E}_\theta[\, \mathbf{f} \,|\, \mathcal{D} \,] \right\|_2^2}_{\text{Empirical factual error}} + \underbrace{\left\| \mathrm{Var}_\theta[\, \mathbf{Y}^{(\mathbf{1}-\mathbf{W})} \,|\, \mathcal{D} \,] \right\|_1}_{\text{Posterior counterfactual variance}} \right]$$

Risk-based Empirical Bayes (II)

- Risk-based empirical Bayes is equivalent to learning a balanced linear representation (hyper-plane) in a vector-valued Reproducing Kernel Hilbert Space (vvRKHS)



Realized Factual Outcomes
$Y_i^{(W_i)}$
$W_i = 1$
$W_i = 0$
$\mathcal{X}$
Posterior Distribution of Counterfactuals
$\mathbb{P}_{\theta^*}(Y_i^{(1-W_i)} \mid X_i)$
$\|\mathrm{Var}_\theta[\mathbf{Y}^{(1-\mathbf{W})} \mid \mathcal{D}]\|_1$
$\mathcal{X}$
$\mathbf{K}_{\theta^*}$
vvRKHS
$\mathcal{H}_{\mathbf{K}_{\theta^*}}$
$\|\mathbf{Y}^{(\mathbf{W})} - \mathbb{E}_\theta[\mathbf{f} \mid \mathcal{D}]\|_2^2$
Treatment Effect
$T(x) = \langle \alpha, \mathbf{K}_{\theta^*}(x, \mathbf{X})\mathbf{e} \rangle_{\mathcal{H}}$

The Model (I)

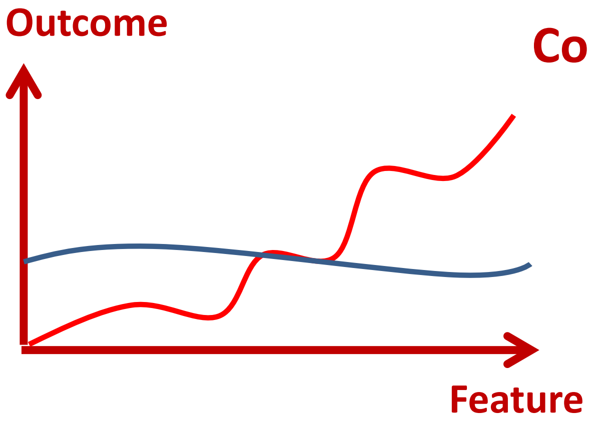
The response surface for the “no treatment” outcome and for the “treatment” outcome are **different!**

=> Construct a kernel function with different length-scales for each surface using a **linear coregionalization model**!

$$\mathbf{K}(x,x') = \mathbf{B}_o k_o(x,x') + \mathbf{B}_1 k_1(x,x')$$

Covariance function for the first potential outcome

Covariance function for the second potential outcome



The Model (II)

$$\mathbf{K}(x,x') = \mathbf{B}_o\, k_o(x,x') + \mathbf{B}_1\, k_1(x,x')$$

Outcome-specific Squared exponential kernel

$$k_W(x,x') = \exp\left(-\frac{1}{2}(x-x')^T\, \mathbf{R}_W\, (x-x')\right),$$

Relevance parameters

$$W \in \{0,1\},\, \mathbf{R}_W = \mathrm{diag}(\ell_{1,W}^{-2}, \ell_{2,W}^{-2}, \ldots, \ell_{d,W}^{-2}).$$

Length-scale of a feature determines its ***relevance*** to treatment outcomes

$$\mathbf{B}_o = \begin{bmatrix} b_{11}^o & b_{12}^o \\ b_{21}^o & 0 \end{bmatrix},\; \mathbf{B}_1 = \begin{bmatrix} 0 & b_{12}^1 \\ b_{21}^1 & b_{22}^1 \end{bmatrix}$$

Cross-outcome correlations

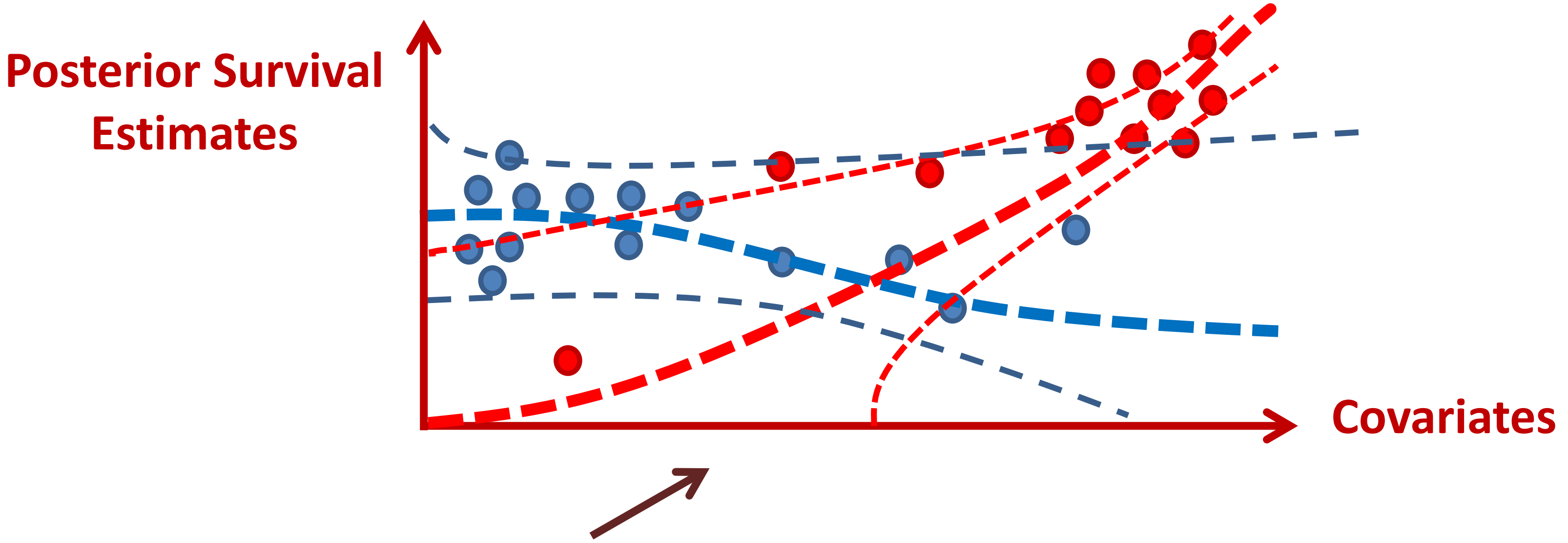
Bayesian Non-parametric Estimation of Individualized Treatment Effects

Specify prior over model parameters

Compute posterior distribution of parameters

Average over many models!

Allows computing posterior credible intervals for the survival estimates of every individual!



Posterior distribution of treatment effects → Final estimate is an average over posterior of parameters

Results: Infant Health Development Program

- **Subjects:** premature infants with low birth weight (747 subjects, 25 covariates)
- **Treatment:** educational and family support services and pediatric follow-up offered during the first 3 years of life.
- **Outcomes:** IQ test applied when infants reached 3 years.
- **All outcomes (response surfaces) are simulated**

Method	Out-of-sample Estimated Error
Bayesian Multi-task GPs	**1.0 ± 0.08**
Balancing Counterfactual Regression (Sontag)	2.2 ± 0.13
BART (Hill)	2.2 ± 0.17
Causal Forests (Athey)	2.4 ± 0.23
Nearest Neighbor Matching (Xie)	4.2 ± 0.22

Powerful methodology – many applications

Individualized treatment effects

- treatments, medications, procedures

- Which?
- When?

Will revolutionize the design of clinical trials

A. M. Alaa and M. van der Schaar, "Bayesian Inference of Individualized Treatment Effects using Multi-task Gaussian Processes," https://arxiv.org/pdf/1704.02801.pdf

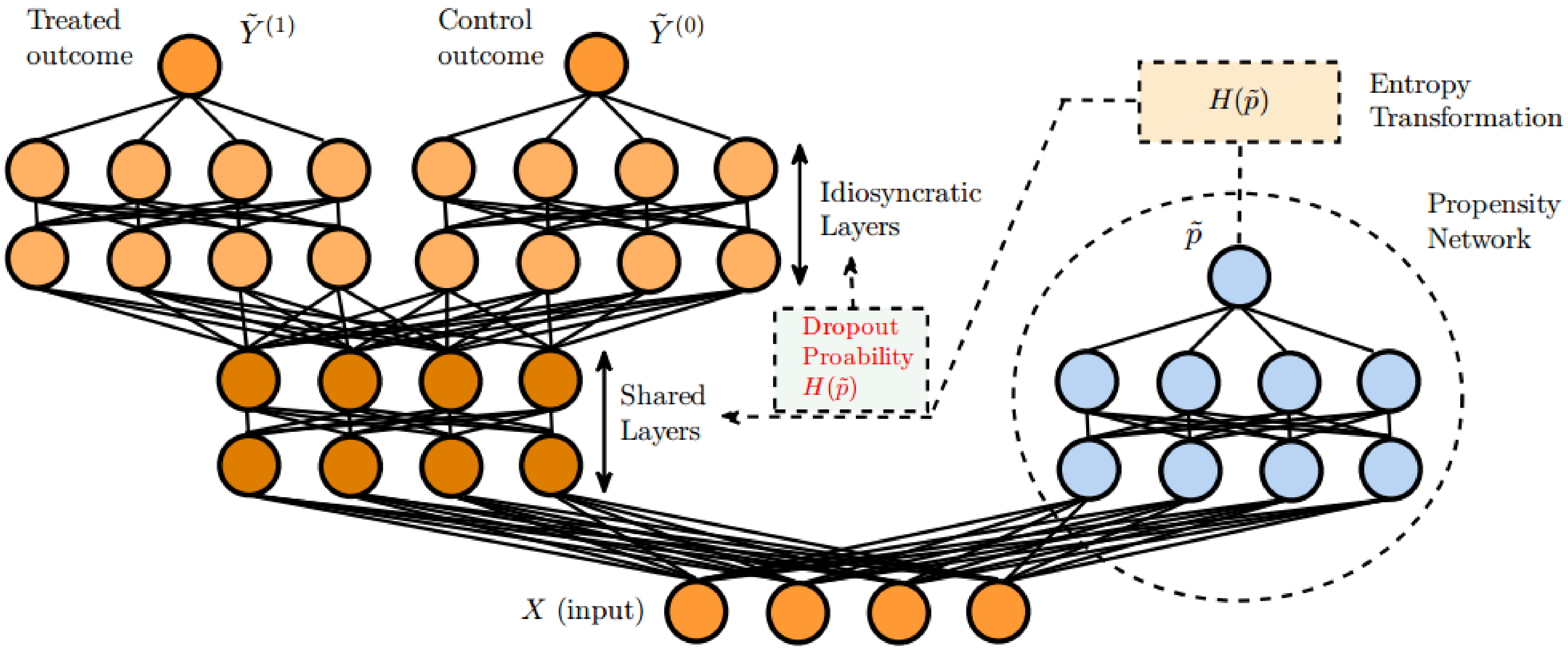
Deep Counterfactual Networks

We can use a deep learning implementation for our model as well!

Multi-task GP -> Multi-task Networks with Dropout

Risk-based Empirical Bayes -> Propensity-Dropout

The multi-task network has layers shared between treated and control patients, and dropout probability depends on propensity scores



Personalized Risk Scoring for Critical Care

ICML 2016, NIPS 2016
IEEE Trans. on Biomedical Engineering, 2016

Timely Prognosis and Intervention

In the US, every year

- 200,000 hospitalized patients experience cardio-pulmonary arrests
- 75% of those patients die
- 50% of those patients could have been saved
- 75,000 unnecessary deaths *in hospital*

Current risk assessment methods do not work well!

What is needed?

- **Timely intervention: earlier admission to Intensive Care Units**

What is the problem?

- **ICU space is scarce**
- **Hard to identify *which* patients must go to ICU *now***

Time is life - minutes matter

- **Our work (Forecast ICU) saves *hours*, hence *lives!***

What data is available to us?

Vital signs

- Diastolic blood pressure
- Systolic blood pressure
- Best motor response
- Best verbal response
- Eye opening
- Glasgow coma scale score
- Heart rate
- Respiratory rate
- Oxygen saturation
- Temperature
- Oxygen device assistance

1 measurement / 4 hours

Lab tests

- Chloride
- Creatinine
- Glucose
- Hemoglobin
- Platelet count
- Potassium
- Sodium
- Total CO2
- Urea nitrogen
- White blood cell count

1 measurement / 24 hours

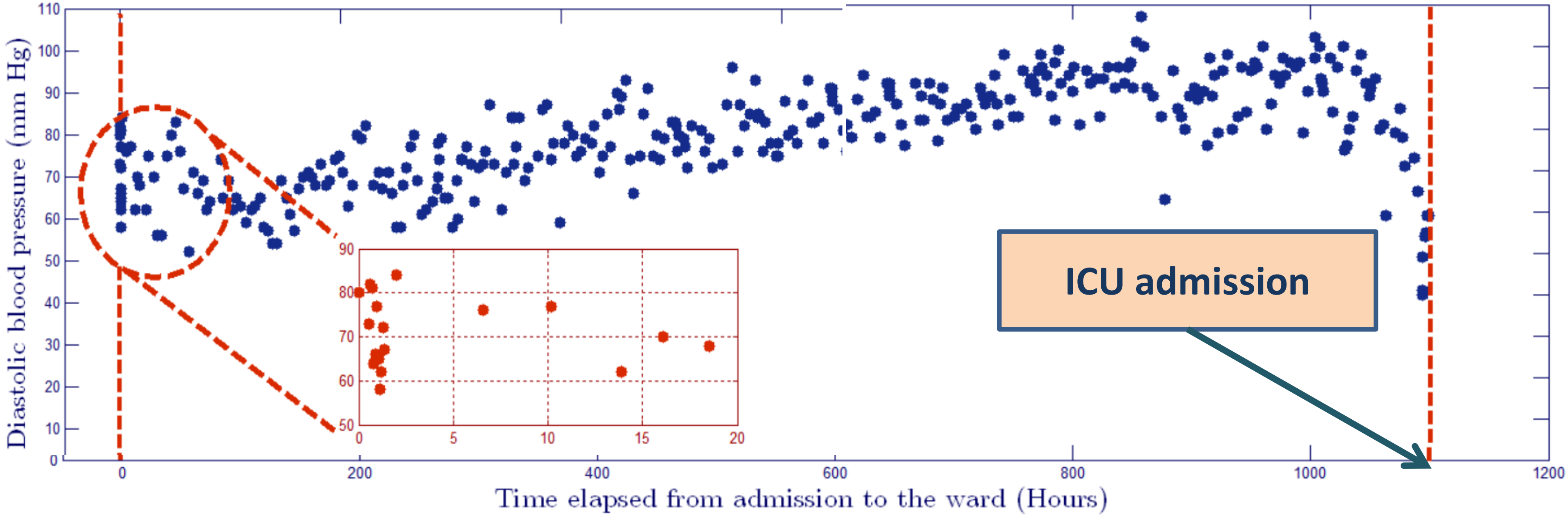
Admission information

- Transfer
- Age
- Floor ID
- Gender
- Ethnicity
- Race
- Stem cell transplant
- ICD-9 codes

Constant

Physiological time-series data

- **Example:** Diastolic blood pressure for a patient hospitalized in a regular ward for more than 1000 hours and then admitted to ICU



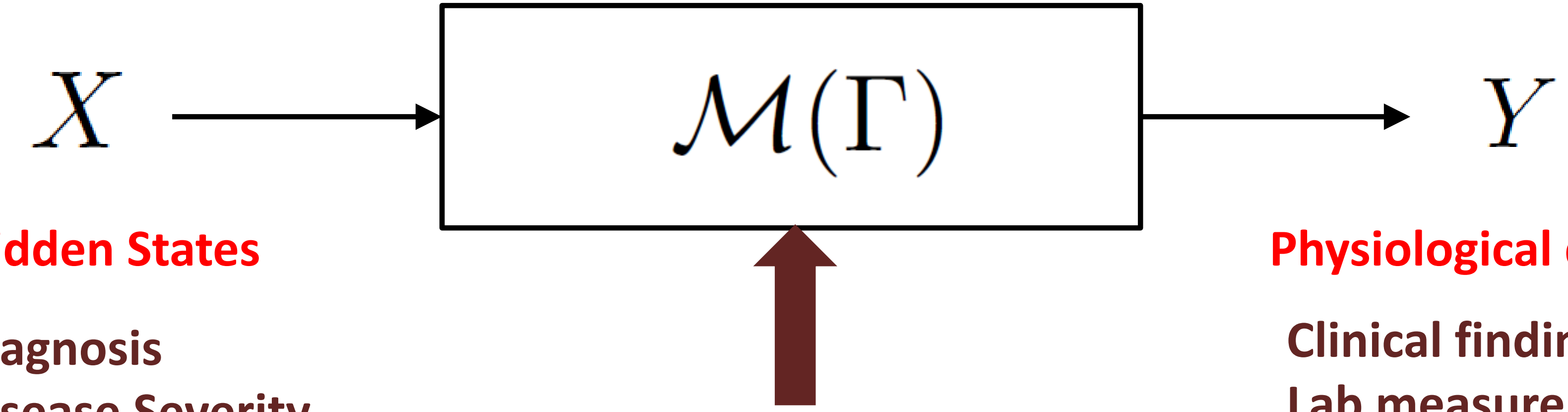
ICU admission

- Patient appeared stable, but was actually deteriorating – the *true* state was *hidden*

A general framework

- **Physiological modeling:** **general** model for mapping **hidden (clinical) states** to **observable (physiological) data**

Model



$X \longrightarrow \boxed{\mathcal{M}(\Gamma)} \longrightarrow Y$

Hidden States

Diagnosis
Disease Severity
Disease Progression Stage
Clinical deterioration

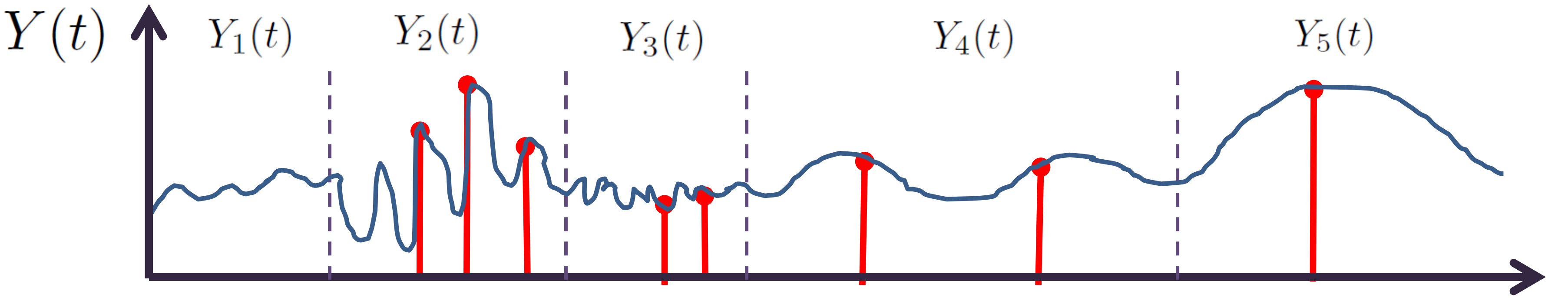
Observational EHR data

Physiological data

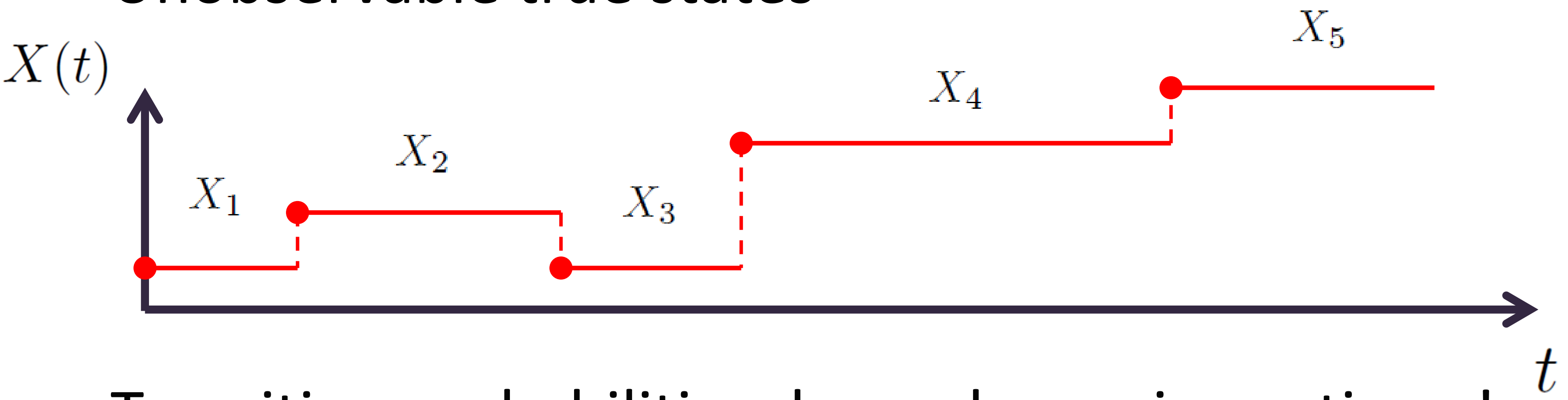
Clinical findings
Lab measurements
Vital signs
Observation times

Observable Process, Unobservable States

- Observable physiological process



- Unobservable true states



- Transition probabilities depend on sojourn times! (Semi-Markov)

Limitations of standard approaches

- Markov models?
  - *Not adequate*
  - True state not observed
- Hidden Markov Models?
  - *Not adequate*
  - Transition probabilities depend on sojourn time
  - Conditionally dependent observations
  - Irregularly but informatively sampled observations
- Informative censoring - absorbing states (observed)

Our New Model: Hidden Absorbing Semi-Markov Model (HASMM)

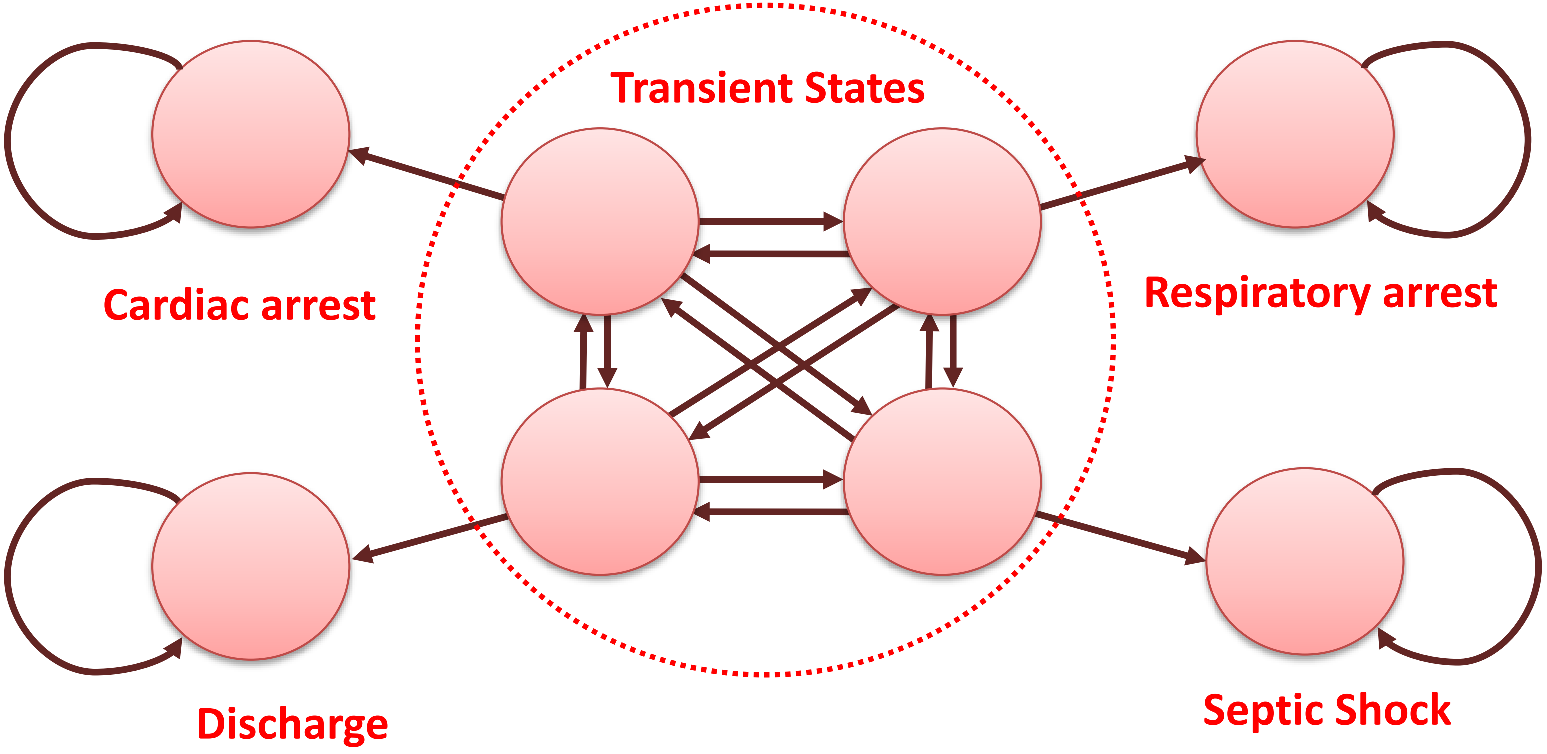
- A versatile model
- Generalizes previous models
- Captures (patient) heterogeneity
- Models the continuous-time data gathering process

Medical Applications

- **Prognosis**
- **Disease progression**
- **Disease trajectories**

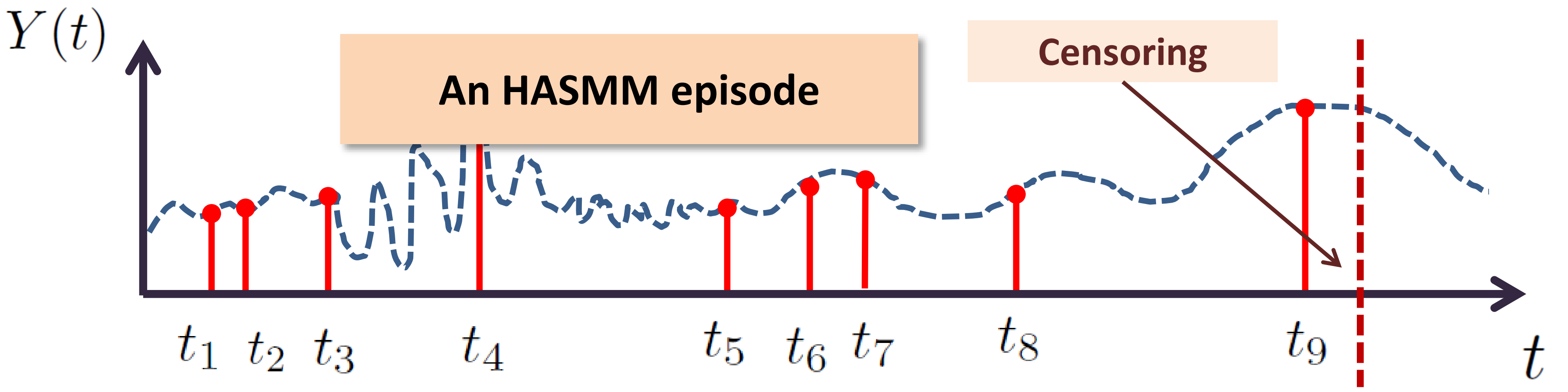
The Hidden Absorbing Semi-Markov Model

- Hidden (true) state space: $\mathcal{X} = \{1, 2, \ldots, N\}$
  - one or more absorbing states **(competing risks!)**



Learn risks/transition probabilities!

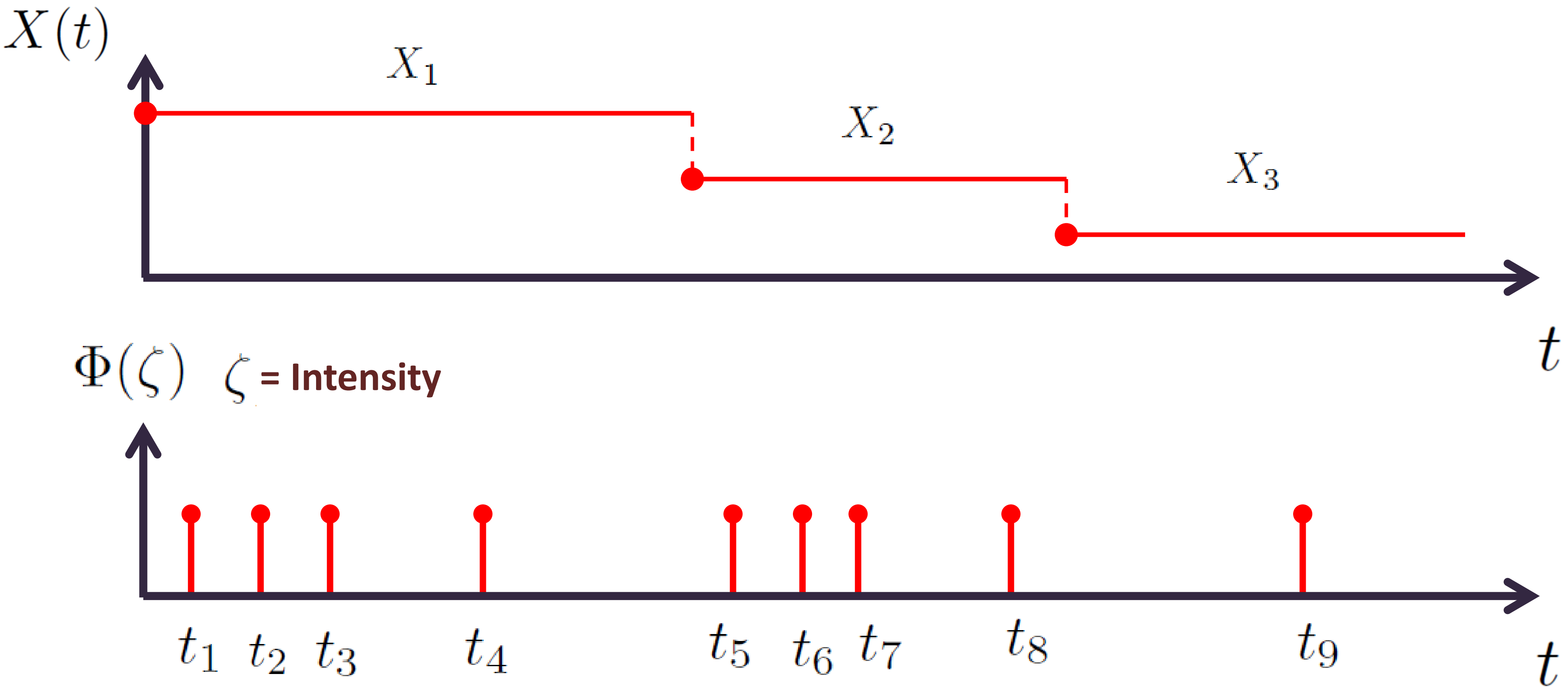
Informative observation times and censoring



Informative observation times

Observation times are modeled as a **Hawkes process**

- **Continuous-time jump process (like Poisson)**
- **Jump intensities depend on true physiological state (unlike Poisson)**



HASMM parameters

- **Sojourn time distribution**

Gamma distribution

$$v_i(s|\lambda_i = \{\lambda_{i,s}, \lambda_{i,r}\}) = \frac{1}{\Gamma(\lambda_{i,s})} \cdot \lambda_{i,r}^{\lambda_{i,s}} \cdot s^{\lambda_{i,s}} \cdot e^{-s\cdot\lambda_{i,r}}, s \geq 0$$

$V_i(.)$ - cumulative distribution function of state i’s sojourn time

- **Semi-Markov transition functions**

$$g_{ij}(s) = \frac{e^{\pi_{ij}(1+\beta_i\ s)}}{\sum_{k=1}^{N} e^{\pi_{ik}(1+\beta_i\ s)}}$$

Multinomial logistic

- **Sampling times of physiological streams: Hawkes point process**
- **Observed physiological data: multi-task Gaussian Process**

$$Y_n(t)|X_n = i \sim \mathcal{GP}(\Theta_i)$$

HASMM parameters

- **Sojourn time distribution**

Gamma distribution

$$v_i(s|\lambda_i = \{\lambda_{i,s}, \lambda_{i,r}\}) = \frac{1}{\Gamma(\lambda_{i,s})} \cdot \lambda_{i,r}^{\lambda_{i,s}} \cdot s^{\lambda_{i,s}} \cdot e^{-s \cdot \lambda_{i,r}}, s \geq 0$$

$V_i(.)$ - cumulative distribution function of state i's sojourn time

- **Semi-Markov transition functions**

$$g_{ij}(s) = \frac{e^{\pi_{ij}(1+\beta_i\ s)}}{\sum_{k=1}^{N} e^{\pi_{ik}(1+\beta_i\ s)}}$$

Multinomial logistic

- **Sampling times of physiological streams: Hawkes point process**
- **Observable process is a marked Hawkes process (with Gaussian Process as the mark process)**

$$Y_n(t)|X_n = i \sim \mathcal{GP}(\Theta_i)$$

Forecast ICU in practice

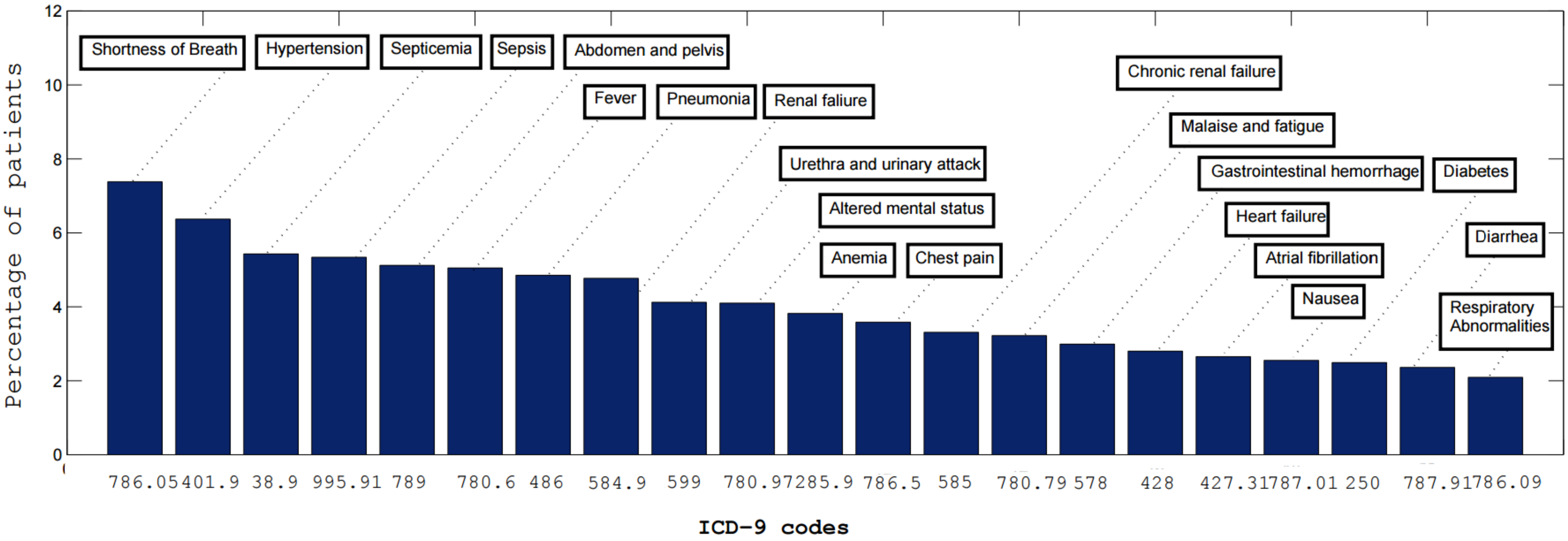
- Hospital: UCLA Ronald Reagan Medical Center

- **Cohort of 6,094 patients**

- Period: March 2013 ~ June 2015 (tested July 2015 – July 2016)
- Age: 18 ~ 100+ years
- Gender:
  - Male (3,018 patients, 49.5%)
  - Female (3,076 patients, 50.5%)
- Length of stay: 1.5 hours ~ 159 days

Wide Variety of Diagnoses

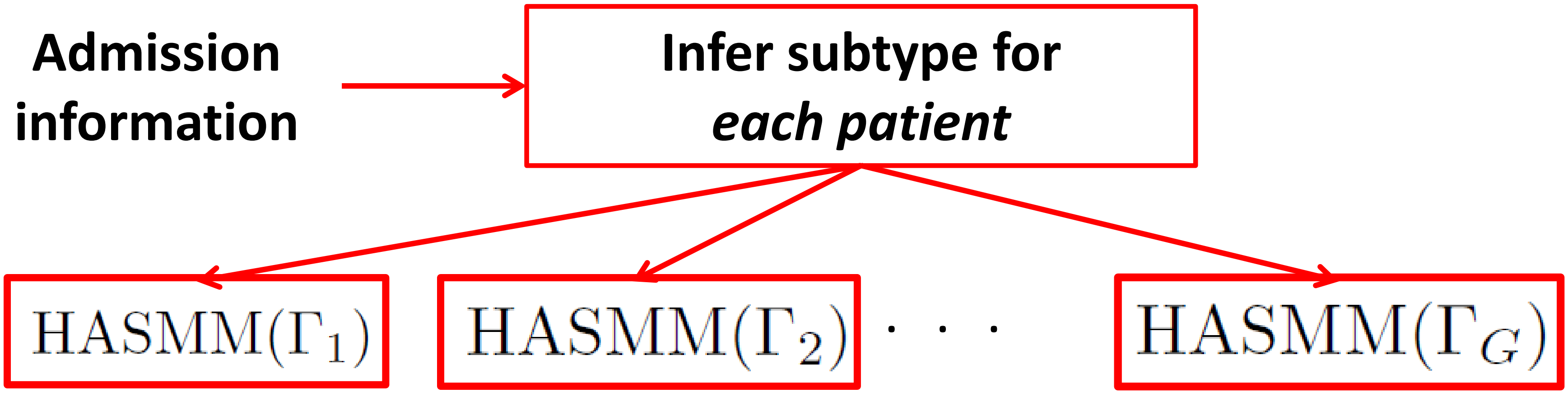
Percentage of patients in top 20 ICD 9 codes



Among 6,094 patients, 306 patients (5.0%) admitted to ICU unexpectedly; 5,788 patients (95.0%) discharged

Subtyping (Phenotyping)

- Discovering the different ways in which a disease manifests in different patients
- Key approach for **personalized medicine**



Admission information

Infer subtype for *each patient*

$\mathrm{HASMM}(\Gamma_1)$ $\mathrm{HASMM}(\Gamma_2)$ $\cdots$ $\mathrm{HASMM}(\Gamma_G)$

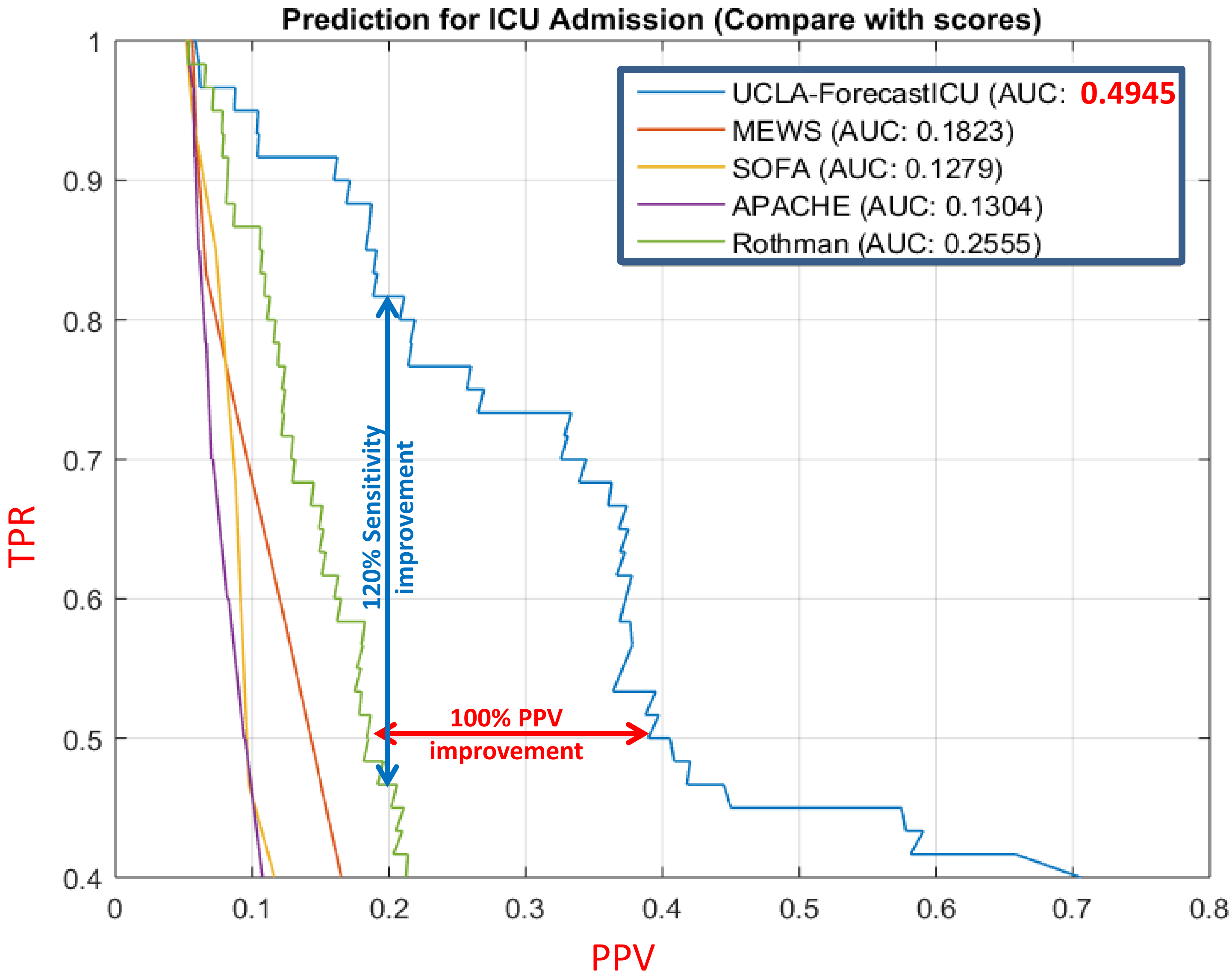
ICML-W 2016

Performance Metrics

- **TPR** (True Positive Rate, i.e. **Sensitivity**) = True Positive/True ICU Patients
- **TNR** (True Negative Rate, i.e. **Specificity**) = True Negative/True Discharge patients
- **PPV** (Positive Predictive Value, i.e. **Precision**) = True Positive/Predicted ICU Patients
- **NPV** (Negative Predictive Value) = True Negative/Predicted Discharge patients

	Predicted ICU patients	**Predicted Discharge patients**
True ICU patients	True Positive	False Negative
True Discharge patients	False Positive	True Negative

Results: TPR vs. PPV

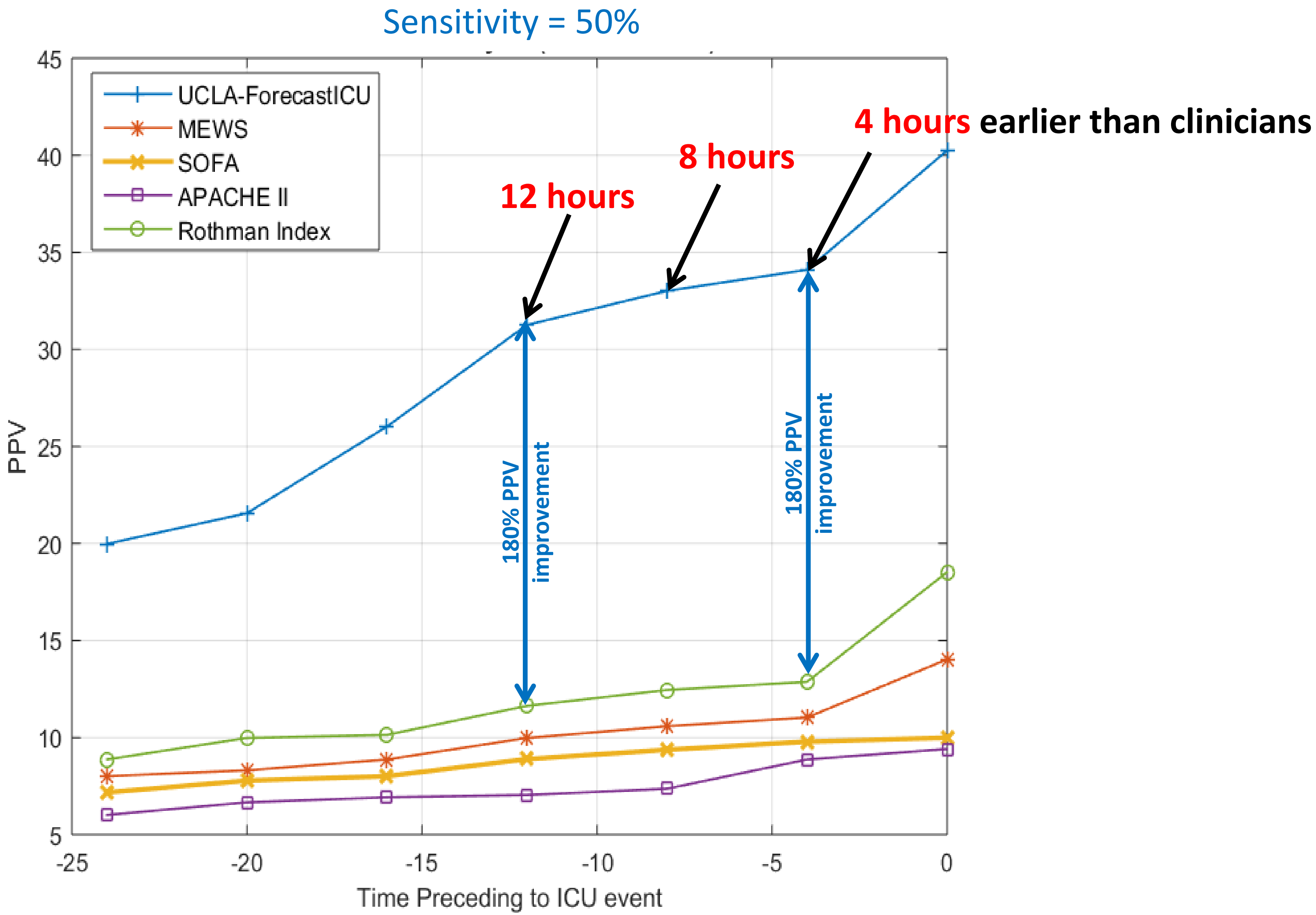


Results: Sensitivity vs PPV

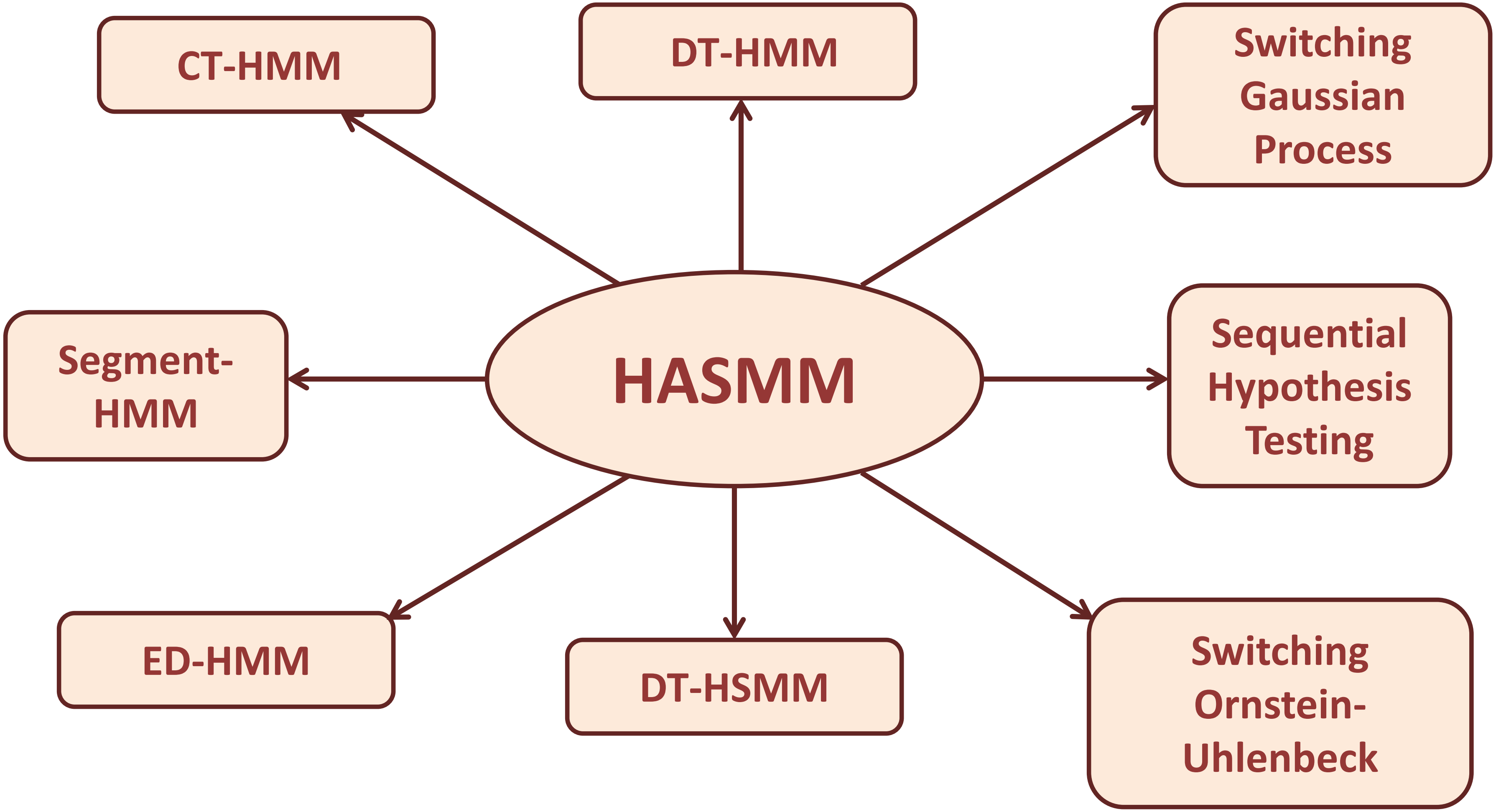
Algorithm	AUC (TPR vs PPV)
HASMM	**0.49**
(Sequential) Random Forest	0.36
(Sequential) Logistic Regression	0.27
(Sequential) LASSO	0.26
HMM (Gaussian emission)	0.32
Multitask Gaussian Processes	0.30
Recurrent Neural Networks	0.29
Rothman	0.25
MEWS	0.18
APACHE II	0.13
SOFA	0.13

http://medianetlab.ee.ucla.edu/MedAdvance

Results: Timeliness



New methodology for learning from time-series data


CT-HMM
DT-HMM
Switching Gaussian Process
Segment-HMM
HASMM
Sequential Hypothesis Testing
ED-HMM
DT-HSMM
Switching Ornstein-Uhlenbeck

Applications beyond medicine (e.g. finance)

Join the revolution!

“Augmented” MD

- through machine learning and artificial intelligence

- Which diseases/medical problems?

- General Practice
- Emergency care, Hospital care, ICU
- Cardiovascular diseases
- Chronic diseases
- Cystic Fibrosis
- Surgery
- Cancers

- Many lives saved
- Many resources saved
- Scientific breakthroughs: disease understanding