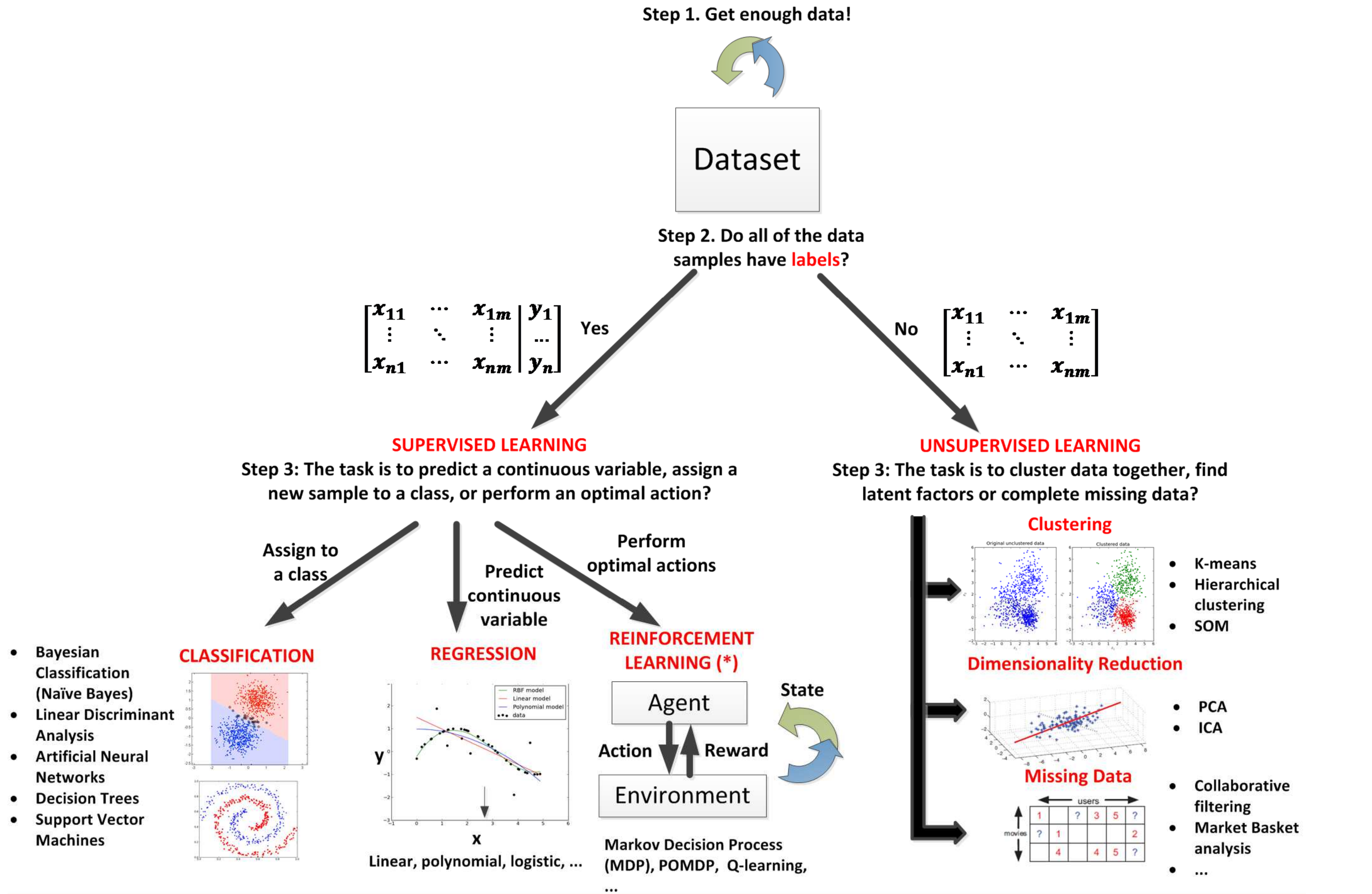
Lecture 14

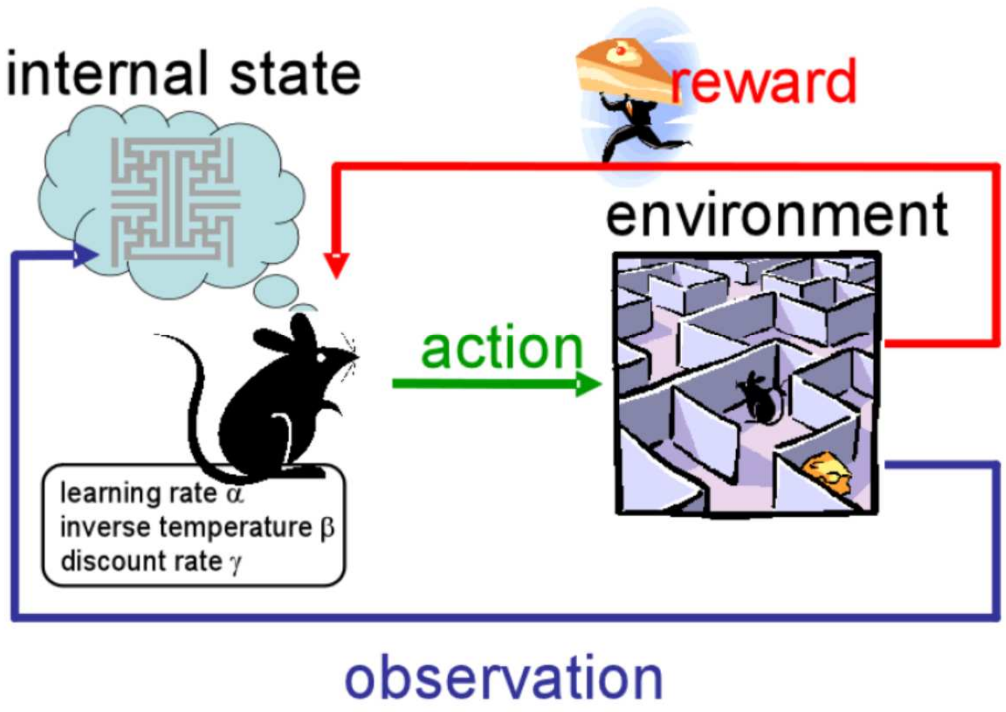
Reinforcement Learning

Instructor: Ilias Tagkopoulos
iliast@ucdavis.edu



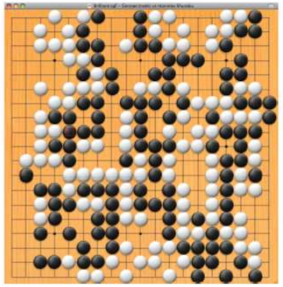
Step 1. Get enough data!
Dataset
Step 2. Do all of the data samples have labels?
Yes
$$\left[\begin{array}{ccc|c} x_{11} & \cdots & x_{1m} & y_1 \\ \vdots & \ddots & \vdots & \cdots \\ x_{n1} & \cdots & x_{nm} & y_n \end{array}\right]$$
No
$$\begin{bmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nm} \end{bmatrix}$$
SUPERVISED LEARNING
Step 3: The task is to predict a continuous variable, assign a new sample to a class, or perform an optimal action?
Assign to a class
Predict continuous variable
Perform optimal actions
CLASSIFICATION
- Bayesian Classification (Naïve Bayes)
- Linear Discriminant Analysis
- Artificial Neural Networks
- Decision Trees
- Support Vector Machines
REGRESSION
RBF model
Linear model
Polynomial model
data
y
x
Linear, polynomial, logistic, ...
REINFORCEMENT LEARNING (*)
Agent
Action
Reward
State
Environment
Markov Decision Process (MDP), POMDP, Q-learning, ...
UNSUPERVISED LEARNING
Step 3: The task is to cluster data together, find latent factors or complete missing data?
Clustering
Original unclustered data
Clustered data
- K-means
- Hierarchical clustering
- SOM
Dimensionality Reduction
- PCA
- ICA
Missing Data
users
movies
- Collaborative filtering
- Market Basket analysis
- ...

What is Reinforcement Learning



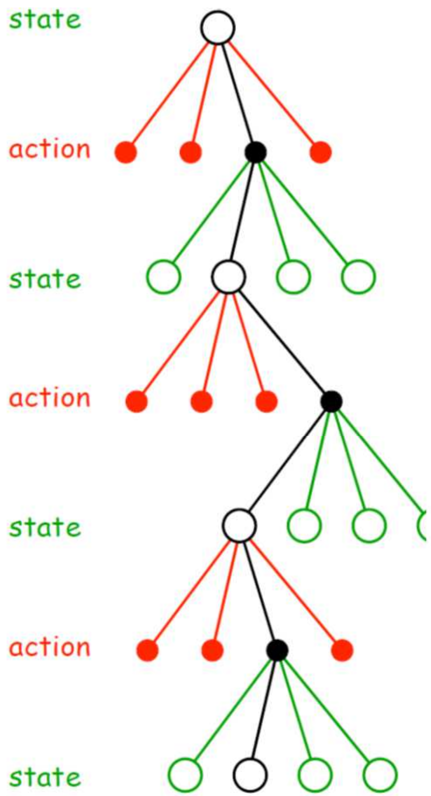
- **Define states (s), actions (a), observations (o), rewards (r)**
- **States can be observed (e.g. MDP) or Latent/Hidden (e.g. POMDP)**
- **The task is to maximize the cumulative reward**

Examples of reinforcement learning (Rl)

- Games (e.g., G. Tesauro, D. Silver)
- Operations research and scheduling (e.g., W. Powell, P. Tadepalli)
- Recently: robotics (e.g., P. Abbeel, J. Peters, P. Stone, M. Riedmiller)

What is the objective in Reinforcement Learning?



- Make optimal decisions $\boldsymbol{a}^*$ by maximizing an expected utility

$$\boldsymbol{a}^* \in \arg\max_{\boldsymbol{a}} \mathbb{E}[r(\boldsymbol{a})] = \arg\max_{\boldsymbol{a}} \sum_{j=1}^{m} r(\boldsymbol{s}_j, \boldsymbol{a}) p(\boldsymbol{s}_j)$$

$\boldsymbol{a}$: decision

$\boldsymbol{s}$: information about environment/state

- Bayesian sequential decision theory (statistics)
- Optimal control theory (engineering)
- Reinforcement learning (computer science, psychology)

One example of optimal policy

Example: Winning the Lottery

Actions	Outcomes
a_1: play	s_1: Win the lottery
a_2: don't play	s_2: Don't win the lottery

Optimal action

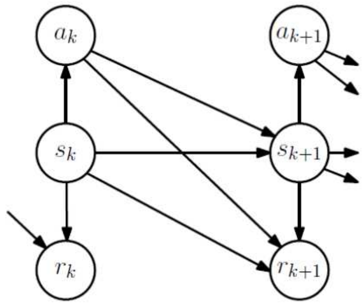
$$a^* = \arg\max_{a_i} \sum_{j=1}^{2} r_{ij} p(s_j | a_i)$$

$$\begin{aligned} p(s_1|a_1) &= 10^{-7} & r_{11} &= 500,000 \text{ USD} \\ p(s_2|a_1) &= 1 - 10^{-7} & r_{12} &= -1 \text{ USD} \\ p(s_1|a_2) &= 0 & r_{21} &= 0 \text{ USD} \\ p(s_2|a_2) &= 1 & r_{22} &= 0 \text{ USD} \end{aligned}$$

A RI Method: Markov Decision Process (MDP)

Markov Decision Process: Definition

- $\mathcal{S}$: State space (finite)
- $\mathcal{A}$: Action space (finite)
- $\mathcal{P}$: Transition probability $p(\boldsymbol{s}_{k+1}|\boldsymbol{s}_k, \boldsymbol{a}_k)$
- r: Reward function
- $\gamma \in [0, 1)$: Discount factor
- π: Policy
  - Deterministic: $\boldsymbol{a} = \pi(\boldsymbol{s})$
  - Stochastic: $\boldsymbol{a} \sim p_\pi(\boldsymbol{a}|\boldsymbol{s})$ alternative notation: $p_\pi(\boldsymbol{a}|\boldsymbol{s}) = \pi(\boldsymbol{a}|\boldsymbol{s})$



Objective

Find a policy π^* that maximizes the expected long-term reward

$$V^\pi(\boldsymbol{s}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k r_{k+1} \middle| \boldsymbol{s}_0 = \boldsymbol{s}, \pi\right], \qquad r_{k+1} = r_{k+1}(\boldsymbol{s}_k, \boldsymbol{a}_k, \boldsymbol{s}_{k+1})$$

How to solve a MDP

Value Functions

- State Value Function: How good is it to be in a particular state s? Well, this depends on the current policy:

$$V^{\pi}(s) = \mathbb{E}[R|s_0 = s] = \mathbb{E}[\sum_{k=0}^{\infty} \gamma^k r_{k+1}|s_0 = s, \pi]$$
$$= \mathbb{E}[r_1 + \gamma V^{\pi}(s_1)|s_0 = s, \pi] \qquad \text{Self-consistency}$$

- **What you want is to find and the policy that maximize the cumulative reward, calculated by the value of each state in the trajectory**

$$\pi^*(s) = \arg\max_a \mathbb{E}[r(s, a, s') + \gamma V^*(s')]$$

This leads to the Bellman equations

One for state value

$$
\begin{aligned}
V^*(s) &= \max_a Q^*(s, a) \\
&= \max_a \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k r_{k+1} | s_0 = s, a_0 = a, \pi^*\right] \\
&= \max_a \mathbb{E}\left[r_1 + \gamma \sum_{k=0}^{\infty} \gamma^k r_{k+2} | s_0 = s, a_0 = a, \pi^*\right] \\
&= \max_a \mathbb{E}\left[r_1 + \gamma V^*(s_1) | s_0 = s, a_0 = a, \pi^*\right]
\end{aligned}
$$

Another for state-action value

$$
\begin{aligned}
Q^*(s, a) &= \mathbb{E}[r_1 + \gamma \max_{a'} Q^*(s_1, a') | s_0 = s, a_0 = a] \\
&= \mathbb{E}[r_1 + \gamma V^*(s_1) | s_0 = s, a_0 = a]
\end{aligned}
$$

This leads to the Bellman equations

One for state value

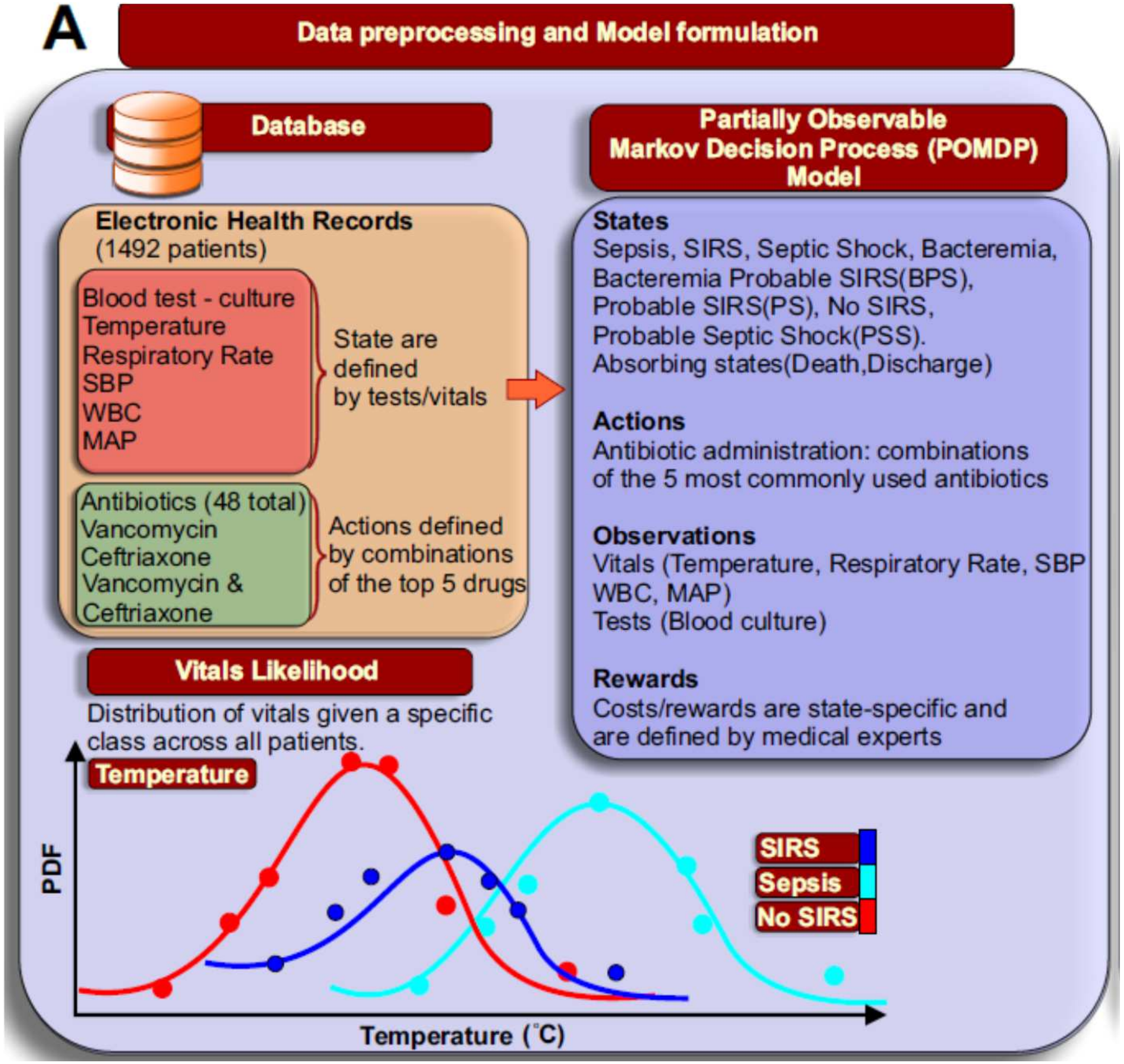
$$
\begin{aligned}
V^*(s) &= \max_a Q^*(s,a) \\
&= \max_a \mathbb{E}\Big[\sum_{k=0}^{\infty} \gamma^k r_{k+1} | s_0 = s, a_0 = a, \pi^*\Big] \\
&= \max_a \mathbb{E}\Big[r_1 + \gamma \sum_{k=0}^{\infty} \gamma^k r_{k+2} | s_0 = s, a_0 = a, \pi^*\Big] \\
&= \max_a \mathbb{E}\big[r_1 + \gamma V^*(s_1) | s_0 = s, a_0 = a, \pi^*\big]
\end{aligned}
$$

Another for state-action value

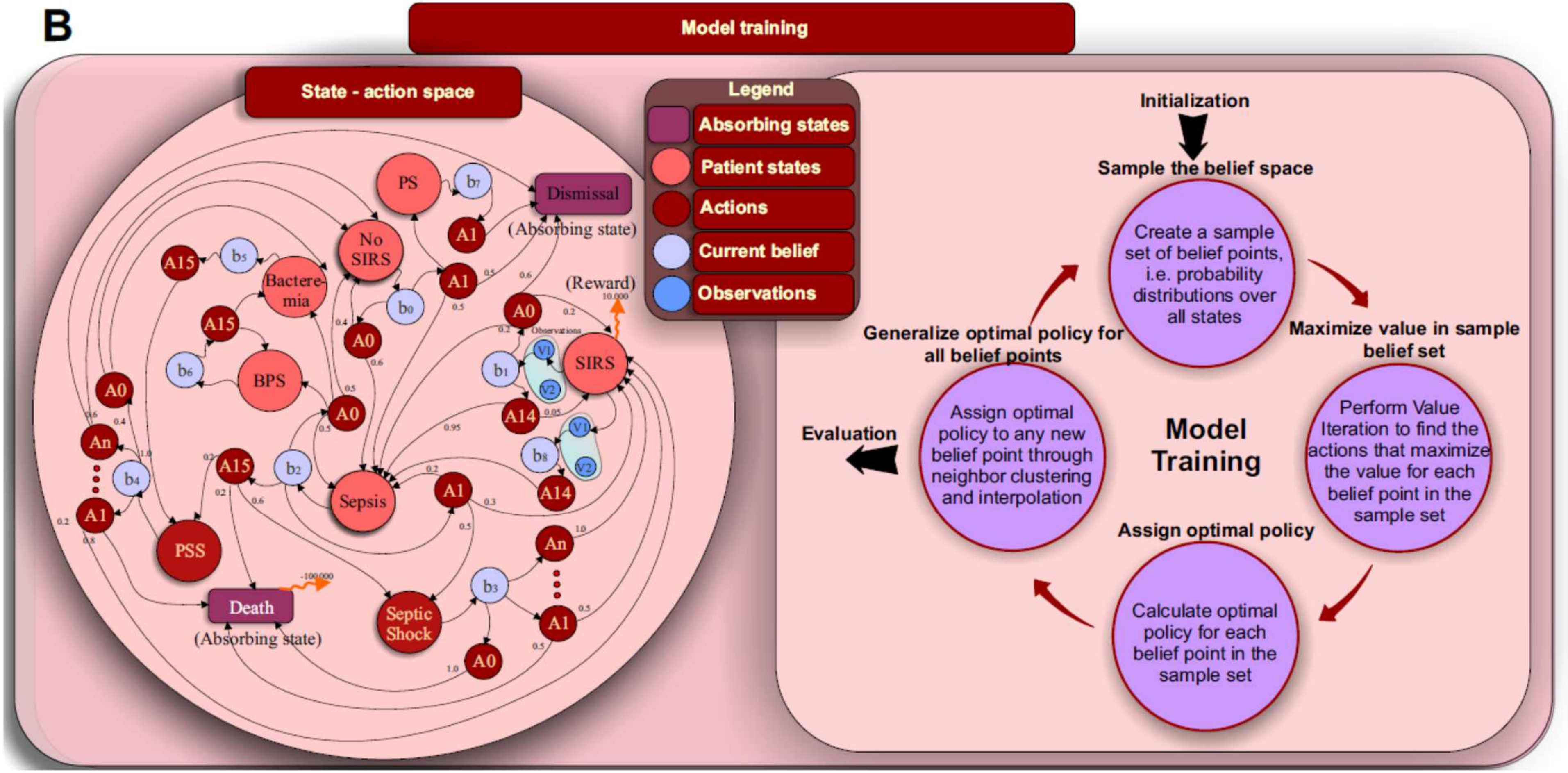
$$
\begin{aligned}
Q^*(s,a) &= \mathbb{E}\big[r_1 + \gamma \max_{a'} Q^*(s_1, a') | s_0 = s, a_0 = a\big] \\
&= \mathbb{E}\big[r_1 + \gamma V^*(s_1) | s_0 = s, a_0 = a\big]
\end{aligned}
$$

Solving them (Dynamic Programming, Monte Carlo) will find the optimal policy

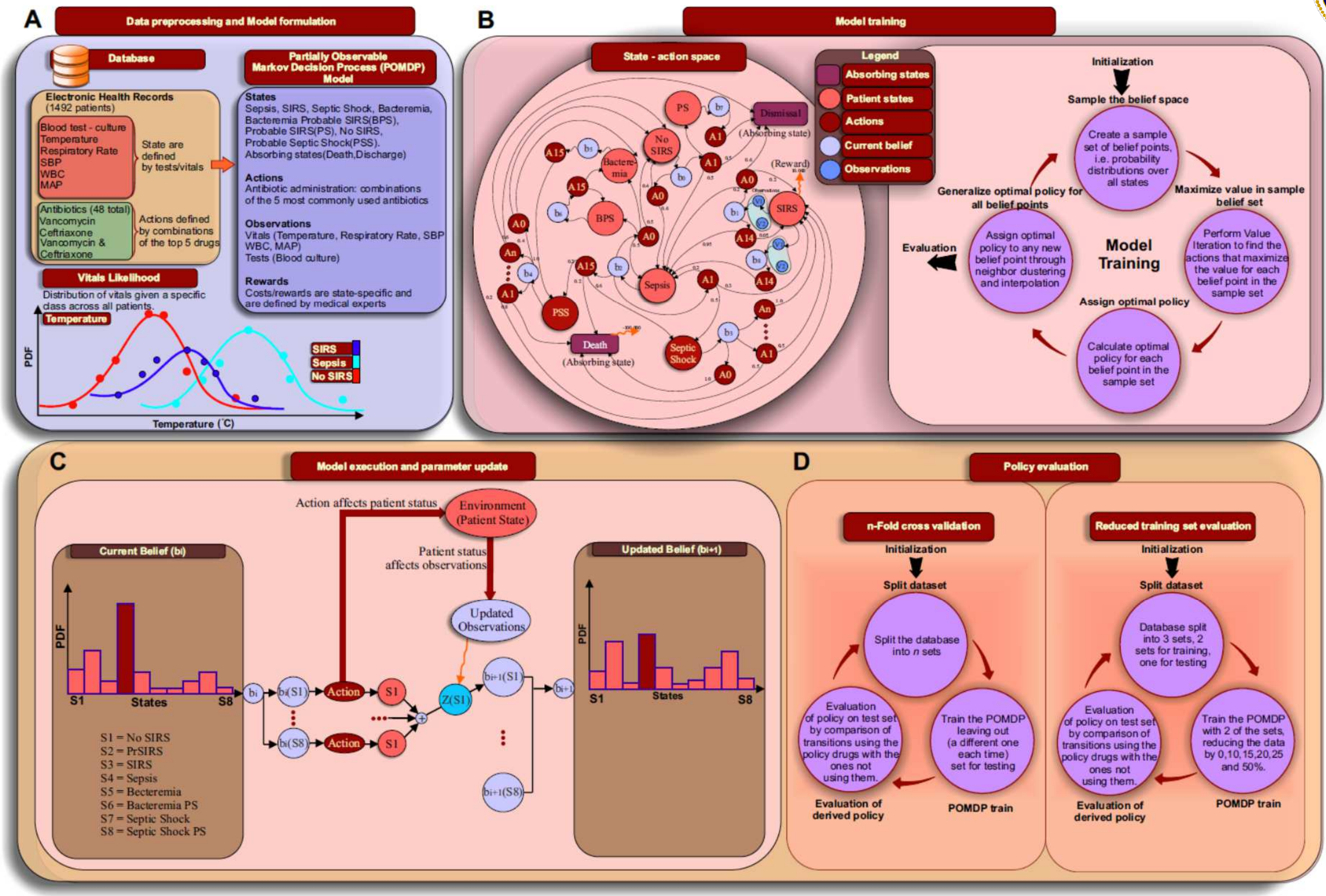
Example: Modeling Sepsis with Partially Observable MDP



Example: State-action-observation-reward space

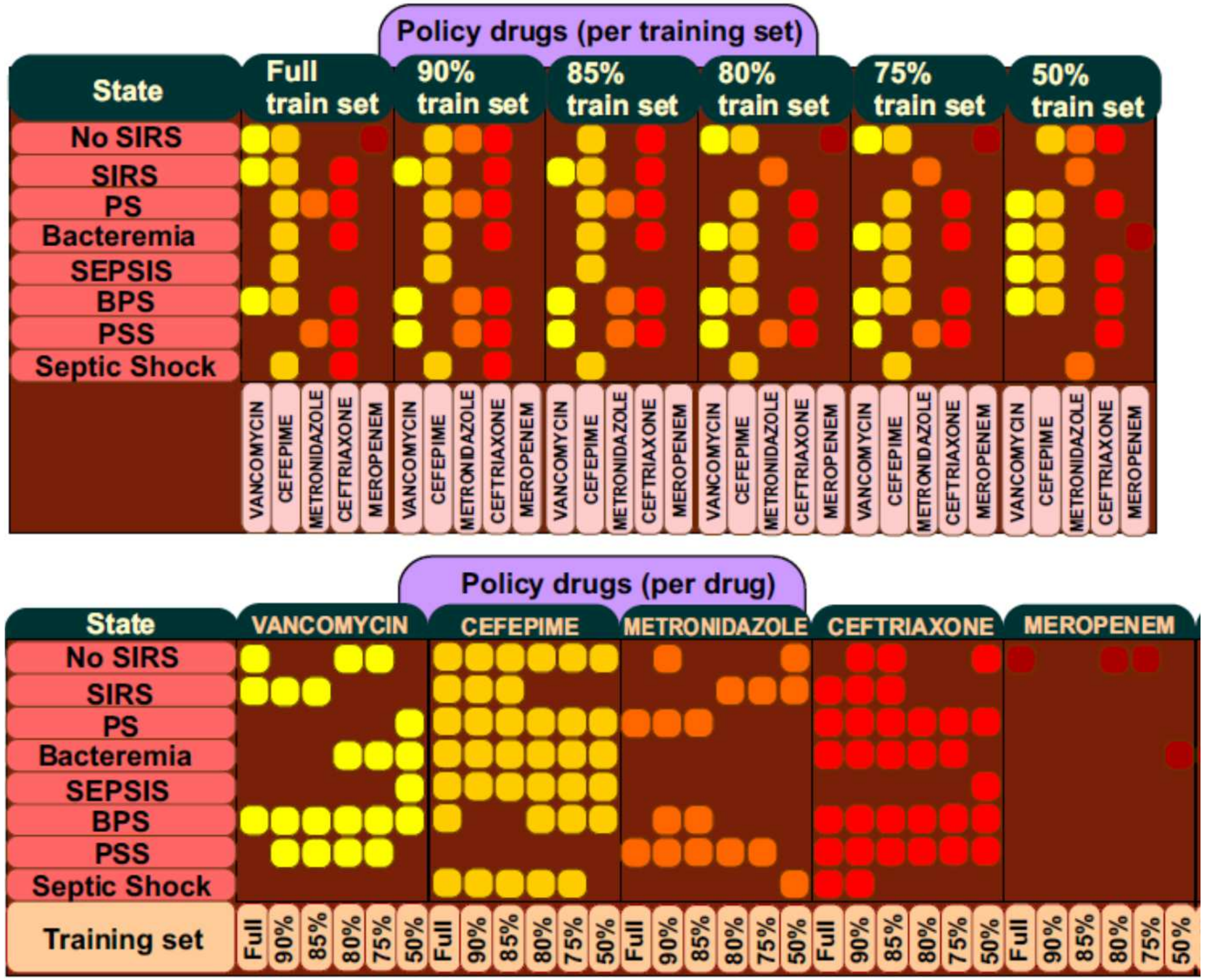


Training and Testing

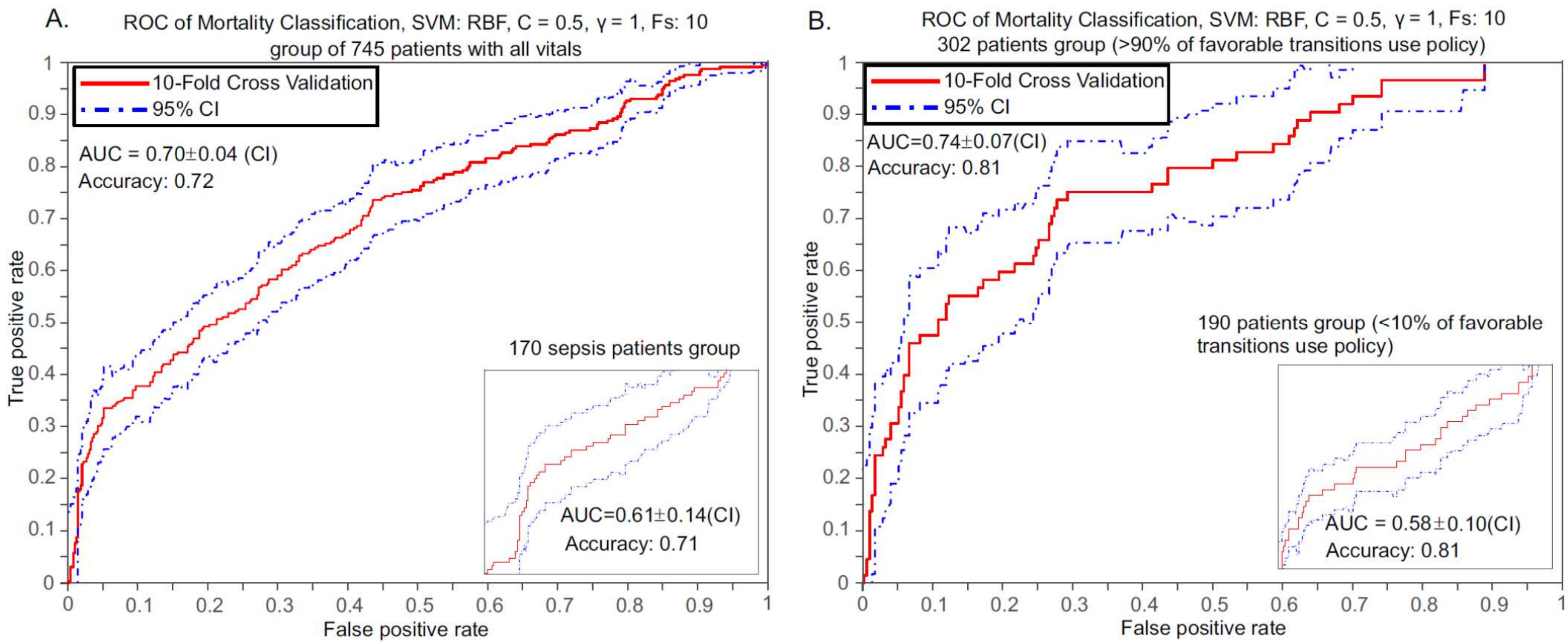


Gultepe E., Green J., Nguyen H., Adams J., Albertson T., **Tagkopoulos I.** (2014) From vital signs to clinical outcomes for patients with sepsis: A machine learning basis for a clinical decision support system. *Journal of American Medical Informatics Association (JAMIA)*, 21(2):315-25, doi:10.1136/amiajnl-2013-1815

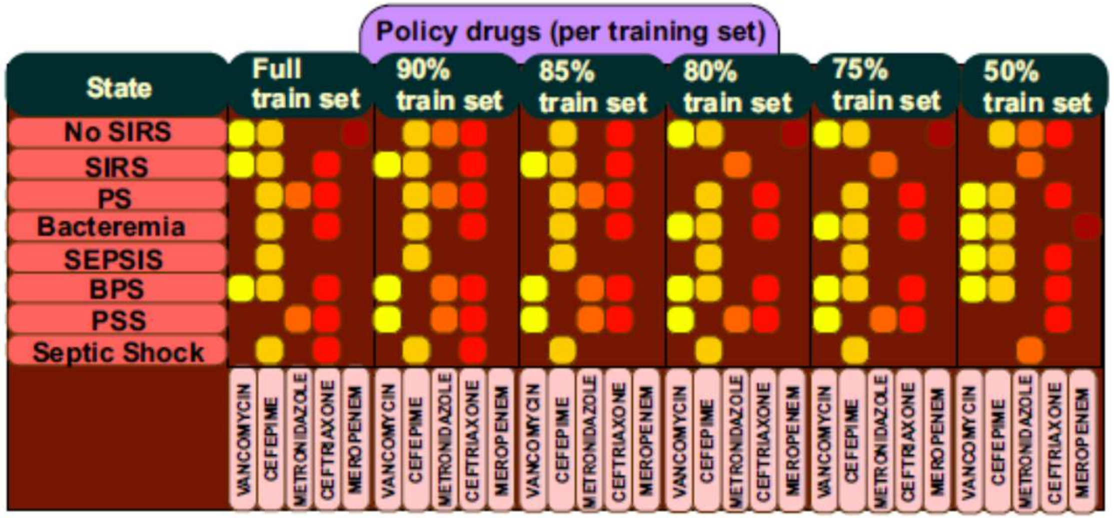
Optimal Policy



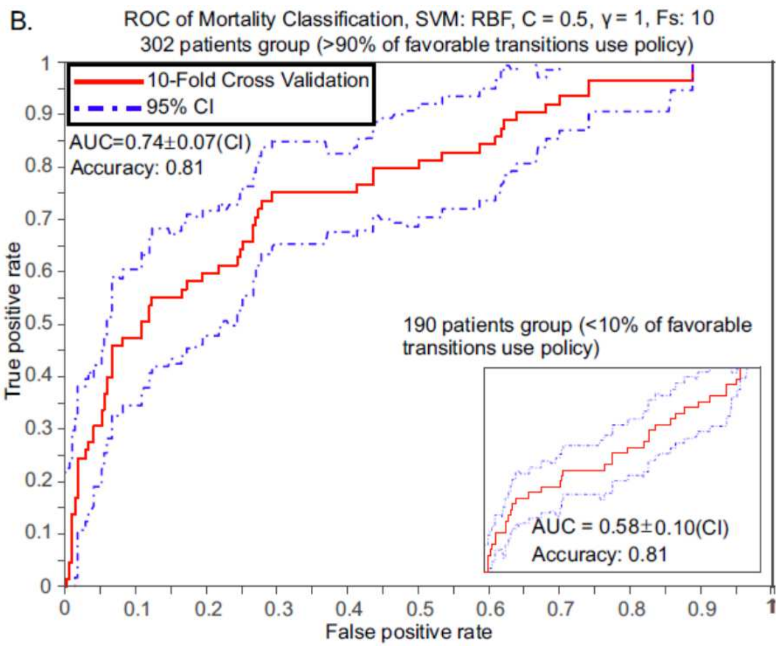
Predicting Mortality with SVMs

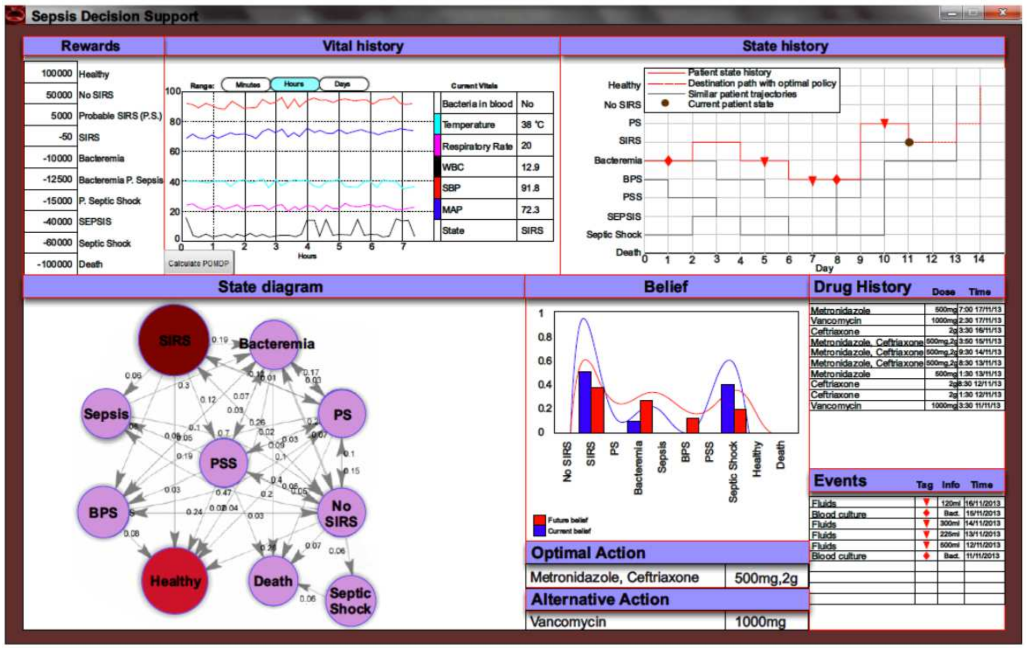


Medical Informatics: Mining the Electronic Health Records



Data-derived treatment led to better outcomes (47% vs. 36%)





Another example: agricultural informatics

1498 crop fields in CA San Joaquin valley, 1997 to 2008
Optimal pest management (*Lygus Hesperus*)

Only ~82% of the farmers use the data-predicted optimal management

	5-11Jun	12-18Jun	19-25Jun	26Jun-2Jul	3-9Jul	10-16Jul	17-23Jul
Low	0.97	0.98	0.95	0.97	0.99	0.99	1.00
Medium	0.08	0.89	0.89	0.96	0.98	0.97	0.97
High	0.18	0.67	0.74	0.83	0.88	0.90	0.87

Cost from sub-optimal pest management per acre (yield-related cost only)

cotton.

	5-11Jun	12-18Jun	19-25Jun	26Jun-2Jul	3-9Jul	10-16Jul	17-23Jul
Low	\$14.50	\$17.51	\$20.81	\$19.79	\$4.04	\$54.70	\$35.50
Medium	\$45.59	\$2.79	\$18.75	\$18.55	\$22.46	\$17.95	\$45.61
High	\$33.81	\$2.64	\$20.06	\$12.96	\$23.89	\$19.99	\$42.99

End of Lecture 14