

Modeling heterogeneous data at scale: applications in sepsis prediction

CASIS 2016

Todd Wasson

May 18, 2016



LLNL-PRES-691883

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344. Lawrence Livermore National Security, LLC

 Lawrence Livermore
National Laboratory

Outline

- **Sepsis**
- Kaiser Permanente electronic medical record dataset
- Density estimation in heterogeneous feature spaces
- SparkPlug
- Bayesian extensions
- Preliminary EMR analysis

Sepsis is lethal, expensive, and ripe for data science

- Sepsis
 1. Affects 750,000 Americans annually
 2. Causes ~50% of all in-hospital deaths
 3. Costs \$17B each year
- *Mortality from septic shock **increases by 7.6%** with **every hour** that treatment is delayed after onset of hypotension*
 - Even incremental improvements in early warning would reap great rewards
- Rich electronic medical record (EMR) data has been gathered but analysis is still in its infancy
 - This is particularly true outside of the ICU
- ***Our goal is to predict sepsis onset hours in advance, facilitating intervention and saving lives***

Outline

- Sepsis
- **Kaiser Permanente electronic medical record dataset**
- Density estimation in heterogeneous feature spaces
- SparkPlug
- Bayesian extensions
- Preliminary EMR analysis

Analysis driven by data: Kaiser Permanente data set

- 244k hospitalizations with suspected infections
 - 5 years, 21 hospitals
- ~50 variables**
- Data characteristics
 - Static
 - Time series
 - Missing
 - Heterogeneous

DIVISION OF RESEARCH

Category	Examples
Patient	<i>Demographics, prior healthcare utilization and code status directives, nursing home residence, prior medication usage</i>
Disease	<i>Admission and discharge diagnoses</i>
Physiologic	<i>Vital signs, neurologic status, general and organ-specific laboratory test results, composite severity scores</i>
Infectious	<i>Microbiological culture results, chest imaging report results, sites of infection, antibiotic type and route</i>
Structural	<i>Size/acuity of hospital, teaching status, daily hospital capacity, hospital shift, time/season of year, risk-adjusted hospital outcomes</i>
Treatment	<i>Level of care/ICU transfers, fluid administration volumes, central line placement, mechanical ventilation, dialysis</i>

Time series
Categorical

Image or text
Numerical

Desired tasks in sepsis prediction

- Predict a variety of quantities of interest
 - $P(\text{Death} \mid X)$
 - $P(\text{Latent sepsis state} \mid X)$
 - Identify structure within putative septic patient population
- Predict irrespective of which subset of features is available for a given patient
 - This includes temporal data... or the absence thereof
- Account for growth in both **scale** and **complexity**

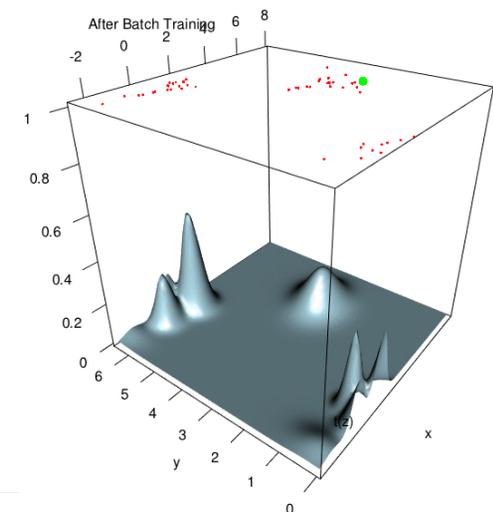
Outline

- Sepsis
- Kaiser Permanente electronic medical record dataset
- **Density estimation in heterogeneous feature spaces**
- SparkPlug
- Bayesian extensions
- Preliminary EMR analysis

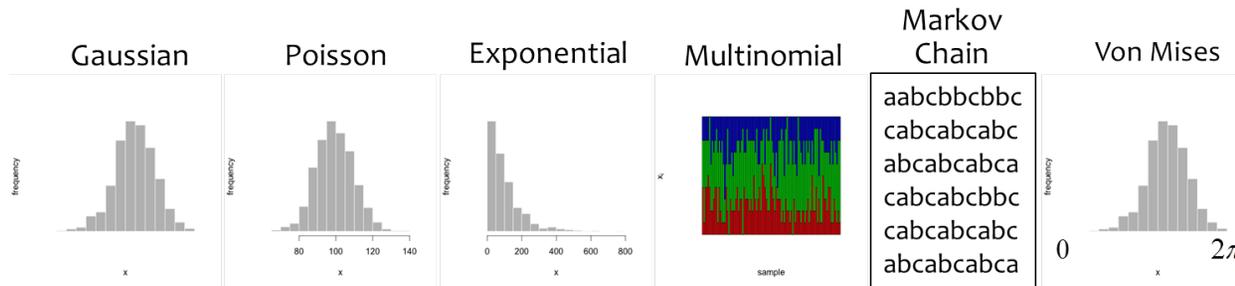
Density estimation R&D enables clustering, classification, prediction, and anomaly detection of large heterogeneous and complex data

Heterogeneous feature vectors are efficiently modeled with composite mixture models

$$P(\mathbf{X}) = \sum_{k=1}^K \underbrace{\pi_k}_{k^{\text{th}} \text{ mixture weight}} \prod_{i=1}^{|\mathbf{X}|} \underbrace{P_k(x_i | \theta_k)}_{k^{\text{th}} \text{ mixture component}}$$



Streaming Density Estimation Cartoon



Composite Mixture Model Base Distributions (p_k)

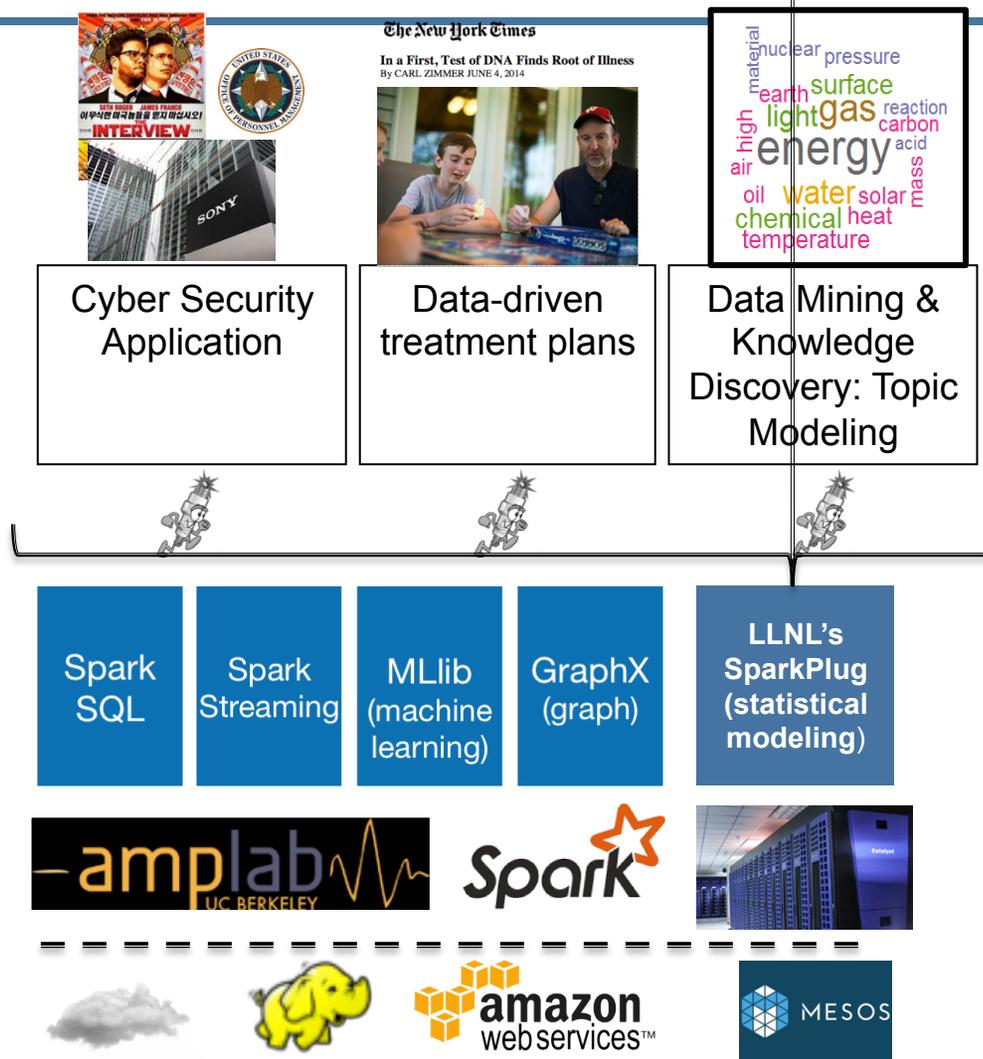
Outline

- Sepsis
- Kaiser Permanente electronic medical record dataset
- Density estimation in heterogeneous feature spaces
- **SparkPlug**
- Bayesian extensions
- Preliminary EMR analysis



SparkPlug is our density estimation toolbox for Big-Data Machine Learning at scale

- Recurring data challenges in mission applications
 - Huge data sets
 - Sparse labels
 - Heterogeneous
 - Complex structure
- **SparkPlug** addresses these challenges
 - Allows complex models able to utilize application specific understanding
 - Scalable design supports large data sizes
 - Modeling does not require advanced software development background

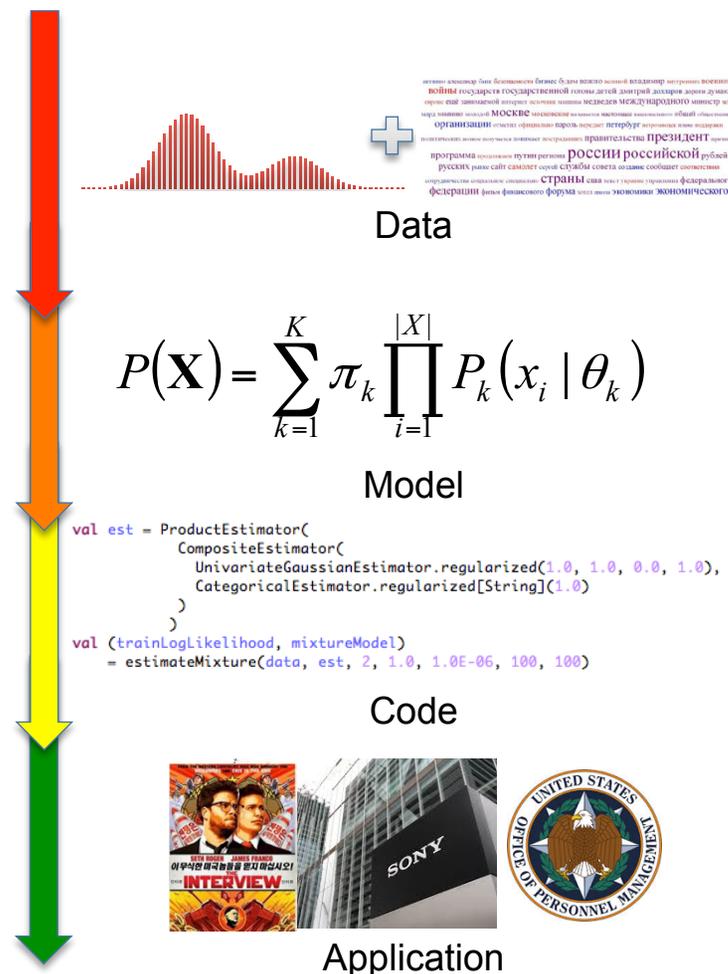




SparkPlug is a flexible and easy to use tool for statistical machine learning at scale

■ SparkPlug's Lego Bricks:

- **Distributions** – Evaluate probability densities given data
- **Estimators** – Fit model parameters given data
- **Combinators** - Combining distributions/estimators to build up new complex models
- **Graphical model templates** – Estimation for rich models with unobservable information
- **Samplers** - Draw samples from models for use in applications



Outline

- Sepsis
- Kaiser Permanente electronic medical record dataset
- Density estimation in heterogeneous feature spaces
- SparkPlug
- **Bayesian extensions**
- Preliminary EMR analysis

SparkPlug adapted to EMR modeling and prediction

- In our EMR task, we have some specific requirements
 - Posterior parameter prediction with confidence given partially-observed feature vectors
 - Simultaneous inference on other parameters, including non-randomly-missing data structure models

- Hence, we are extending SparkPlug in several key ways
 - Gaussian Process modeling of irregularly-sampled time series
 - MCMC fitting of models
 - Distributed MCMC approaches

Parallel MCMC

- The basic idea:
 - Split your data into “shards” or “slices” and distribute to each node
 - Run MCMC chain on each node (fit sub-posterior)
 - Combine sub-posterior samples at driver node to approximate full posterior

- Where current methods differ is in the last step
 - We’ve implemented four approaches (Neiswanger et al., 2014, van Derwerken et al., 2013)

SparkPlug functionality to date

- Distributions / samplers
 - Beta
 - Binomial
 - Categorical
 - CensoredExponential
 - CensoredGeometric
 - Composite
 - Conditional
 - Either
 - Exponential
 - Gamma
 - Geometric
 - HMM
 - Hierarchical mixture
 - Inverse gamma
 - Inverse Wishart
 - Markov chain
 - Multinomial
 - Multinomial logistic regression
 - Multivariate Gaussian
 - Negative binomial
 - Normal inverse gamma
 - Pareto
- Poisson
- Product
- Two-level mixture
- Uniform
- Univariate Gaussian
- von Mises
- Zero-altered negative binomial
- Zero-altered Poisson
- EM estimators
 - Binomial
 - Categorical
 - CensoredExponential
 - CensoredGeometric
 - Composite
 - Conditional
 - Either
 - Exponential
 - Gamma
 - Geometric
 - HMM
 - Hierarchical mixture
 - Linear regression
- Markov chain
- Multinomial
- Multinomial logistic regression
- Multivariate Gaussian
- Negative binomial
- Pareto
- Poisson
- Product
- Two-level mixture
- Uniform
- Univariate Gaussian
- von Mises
- Zero-altered negative binomial
- Zero-altered Poisson
- MixtureModel
- HierarchicalMixtureModel
- Clustering
 - K-means
- MCMC estimators
 - Binomial
 - Categorical
 - Composite
 - Dirichlet
 - Exponential
 - Geometric
 - Multinomial
 - Multinomial logistic regression
 - Multivariate Gaussian
 - Pareto
 - Poisson
 - Uniform
 - Univariate Gaussian
- Distributed MCMC
 - Neiswanger nonparametric
 - Neiswanger semiparametric
 - Neiswanger parametric
 - Dunson median posterior

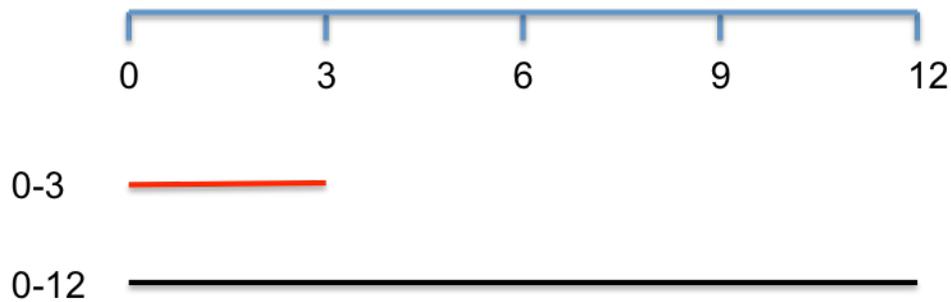
Outline

- Sepsis
- Kaiser Permanente electronic medical record dataset
- Density estimation in heterogeneous feature spaces
- SparkPlug
- Bayesian extensions
- **Preliminary EMR analysis**

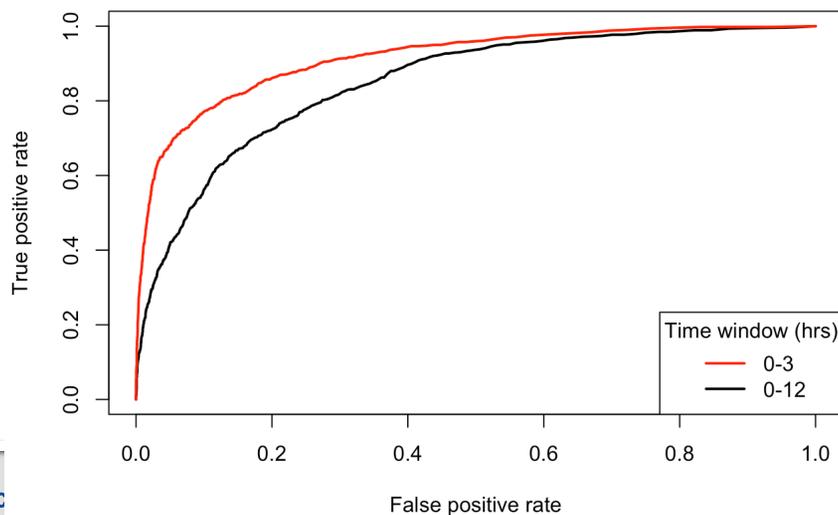
Predicting outcome based on initial X hours of hospitalization: is there an ideal window of time?

- Input variables:
 - 11 admission variables: ADMIT_TYPE, AGE, HOSP_TYPE, MEMBER, LAPS, COPS, ADMIT_TIME, SNF, TRANSPORT, SEX, HET
 - 56 vitals variables:
 - 7 vitals: BPDIA, BPSYS, HRTRT, O2SAT, PP, RPSRT, TEMP
 - 8 summary stats: N_HIGH_OUTLIERS, N_LOW_OUTLIERS, IQR, SD, MAX, MIN, MEAN, MEDIAN

Predicting outcome based on initial X hours of hospitalization: is there an ideal window of time?

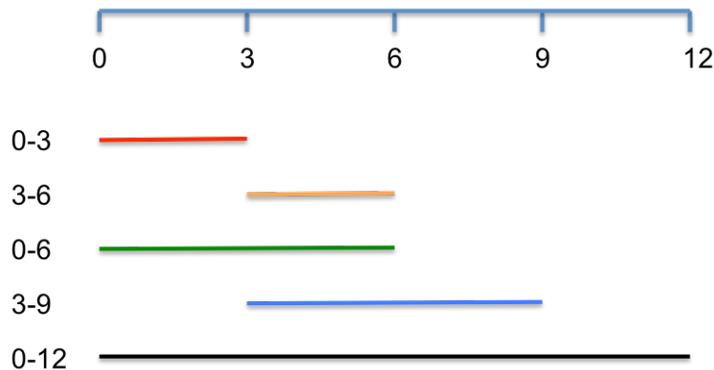


- Best predictors:
 - TEMP (MEAN, IQR, MIN, MAX)
 - LAPS2
 - FAC
 - RSPRT_MEAN

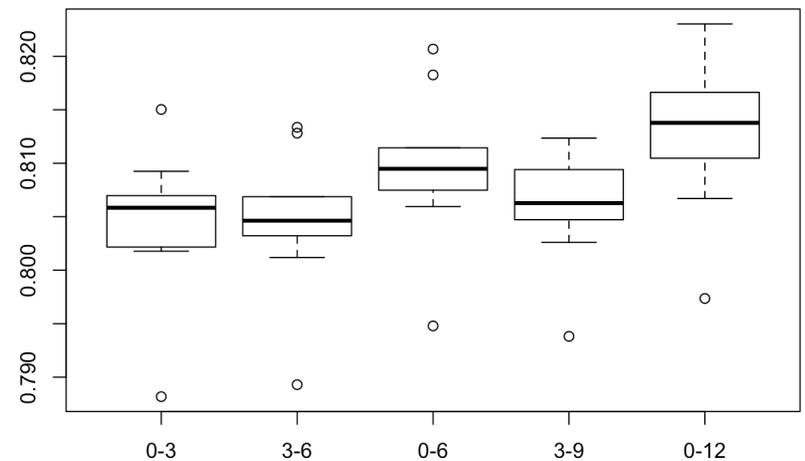
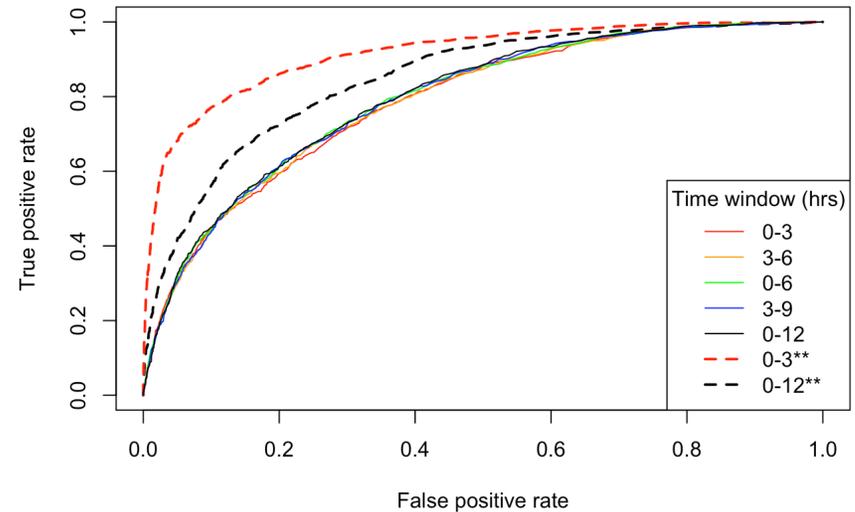


Hypothesis:
Because we're using only summary statistics, perhaps in longer time window signal is drowned out

Predicting outcome based on initial X hours of hospitalization: is there an ideal window of time?



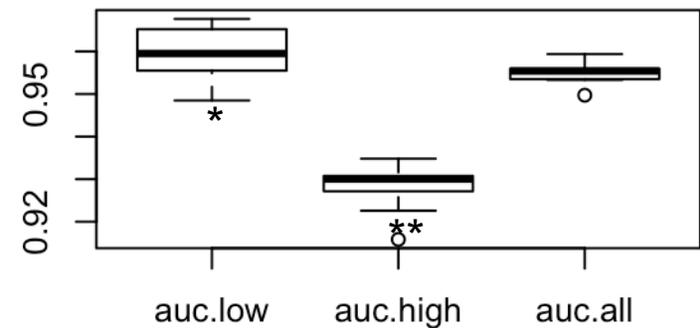
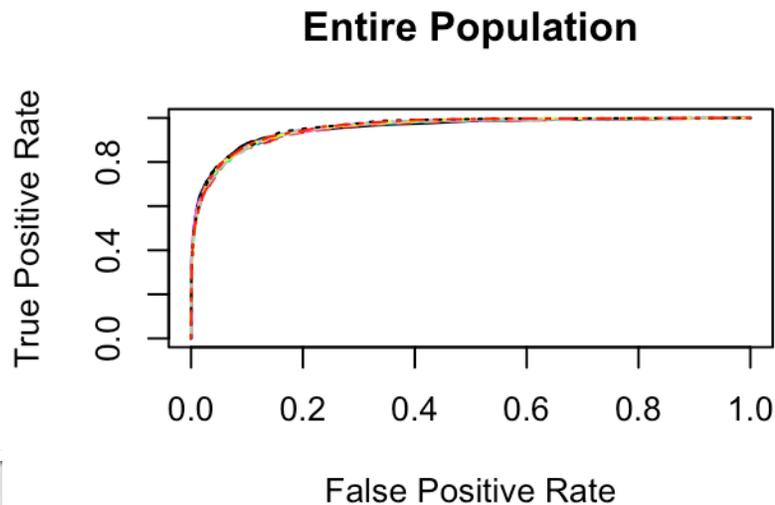
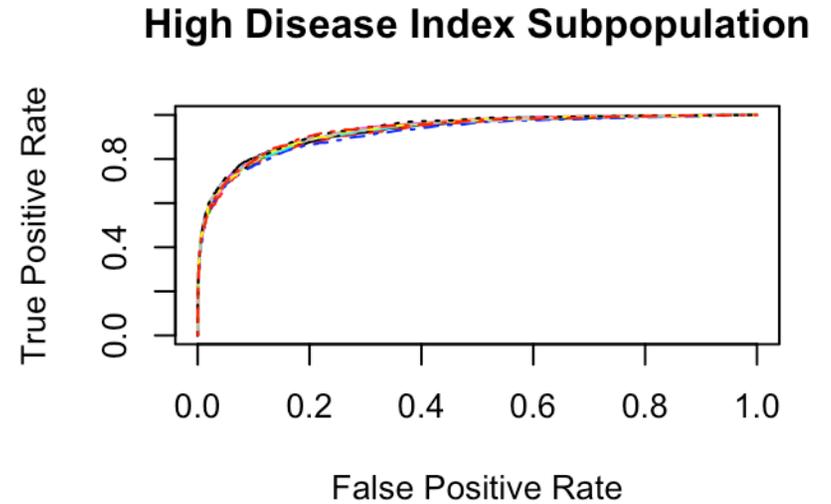
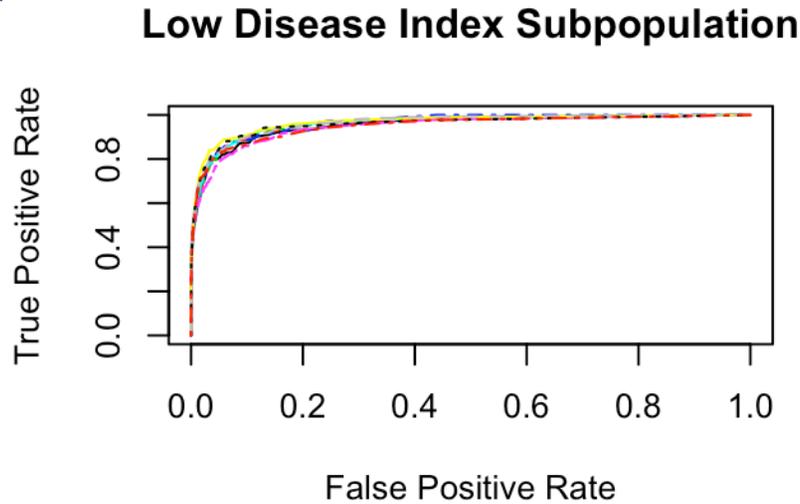
- Fewer predictors (27 out of 68) in order to have complete case analysis.



Can outcome prediction be improved by modeling subpopulations separately?

	Number of episodes	Number of patients	Mortality rate (%)
Entire population	243,233	164,844	5.12
Low disease index subpopulation	147,390	119,183	1.82
High disease index subpopulation	95,843	66,353	10.19

Can outcome prediction be improved by modeling subpopulations separately?



** 0.1, *** 0.05

Concluding remarks and take-home points

- We're developing scalable methods for a wide variety of problem domains and applications
 - In the process of open-sourcing SparkPlug
- In particular, sepsis an especially highly-rewarding target
- Initial work suggests that we can automatically infer substructure in the sepsis population and yield improved predictions in this manner

Acknowledgments and thanks

Vincent Liu, Kaiser Permanente Northern California Division of Research



Ana Paula Sales, LLNL



Michael Mayhew, LLNL



LLNL LDRD 15-ERD-053



**Lawrence Livermore
National Laboratory**