# Physiological Acuity Modelling with (Ugly) Temporal Clinical Data

Marzyeh Ghassemi
CSAIL PhD Candidate

Massachusetts Institute of Technology

# Agenda

- Techniques
  - Topic Models (LDA)
  - Gaussian Processes (GP)
- Applications
  - **KDD 2014 -** Unfolding Physiological State: Mortality Modeling in Intensive Care Units
  - **AAAI 2015 -** A Multivariate Timeseries Modeling Approach to Severity of Illness Assessment and Forecasting in ICU with Sparse, Heterogeneous Clinical Data

# Agenda

- Techniques
  - Topic Models (LDA)
  - Gaussian Processes (GP)
- Applications
  - **KDD 2014 -** Unfolding Physiological State: Mortality Modeling in Intensive Care Units
  - **AAAI 2015 -** A Multivariate Timeseries Modeling Approach to Severity of Illness Assessment and Forecasting in ICU with Sparse, Heterogeneous Clinical Data

# Topic Model Tutorial

- Content is from:
  - Steyvers & Griffiths 2006 paper
  - Blei ICML 2012 Tutorial

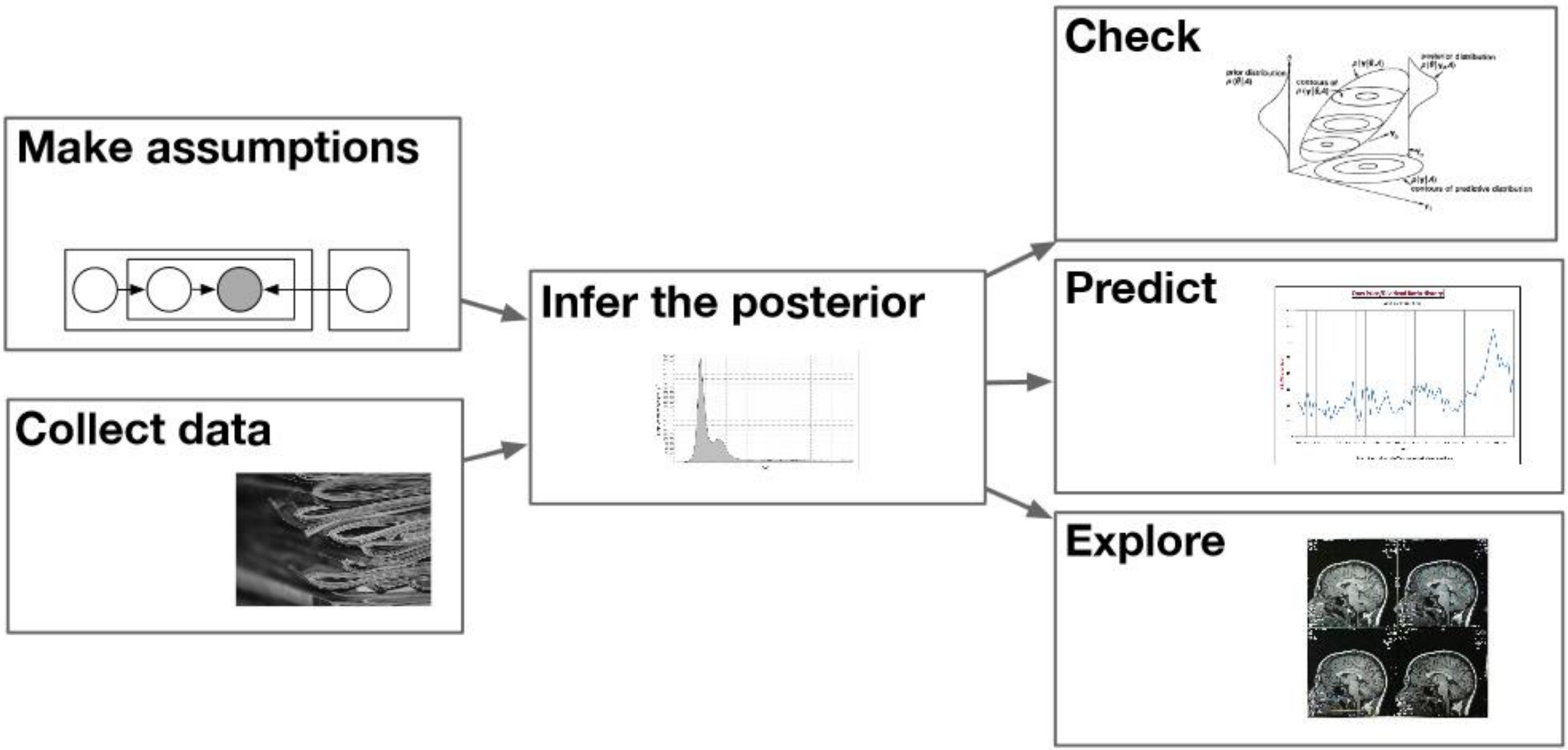
LDA

# Topic Models – Popularity is great

- All the right cliques:
  - Directed graphical models
  - Conjugate priors and nonconjugate priors
  - Time series modeling
  - Modeling with graphs
  - Hierarchical Bayesian methods
  - Approximate posterior inference (MCMC, variational methods)
  - Exploratory and descriptive data analysis
  - Model selection and Bayesian nonparametric methods
  - Mixed membership models
  - Prediction from sparse and noisy inputs

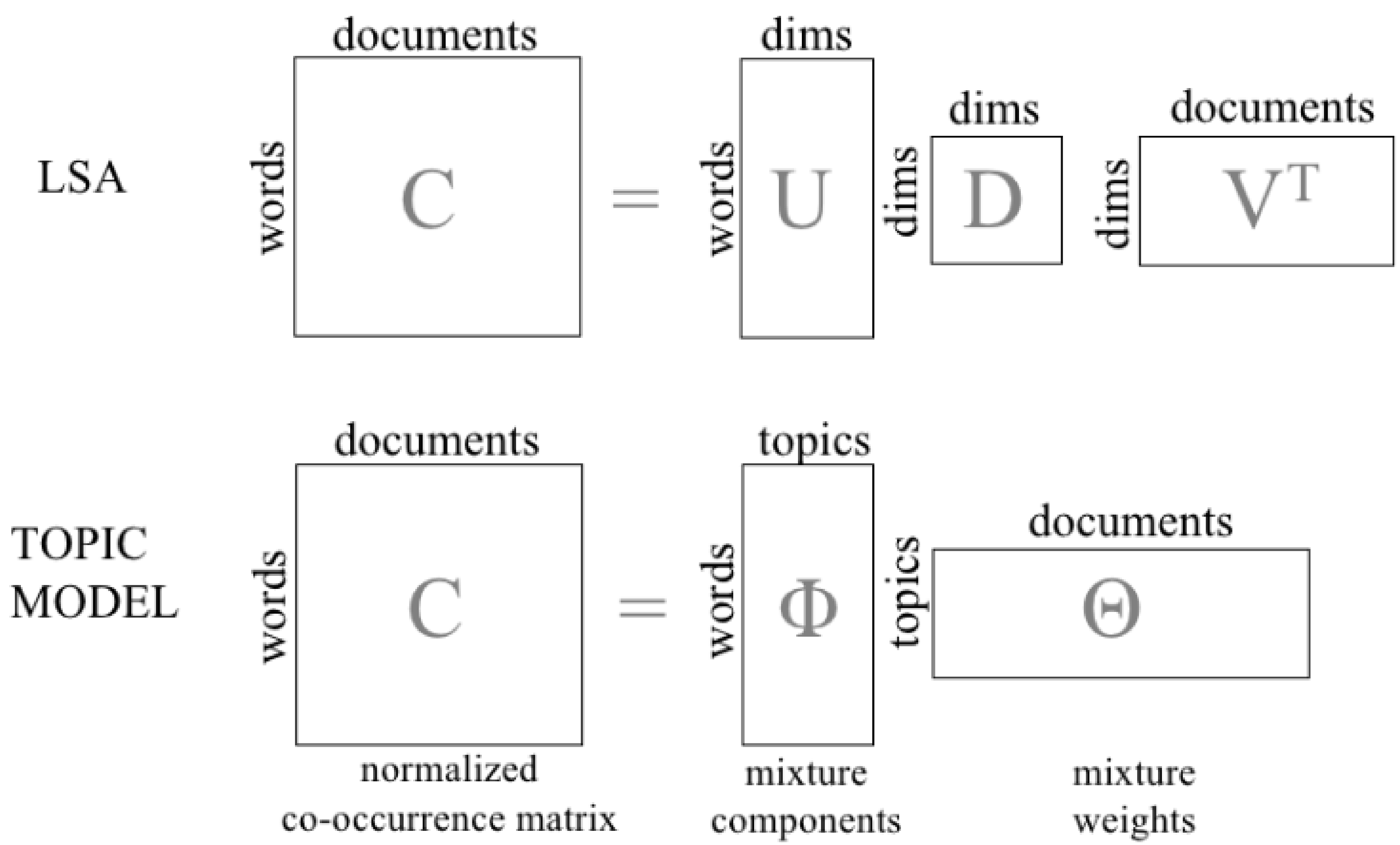
LDA

# Data/Discovery Process



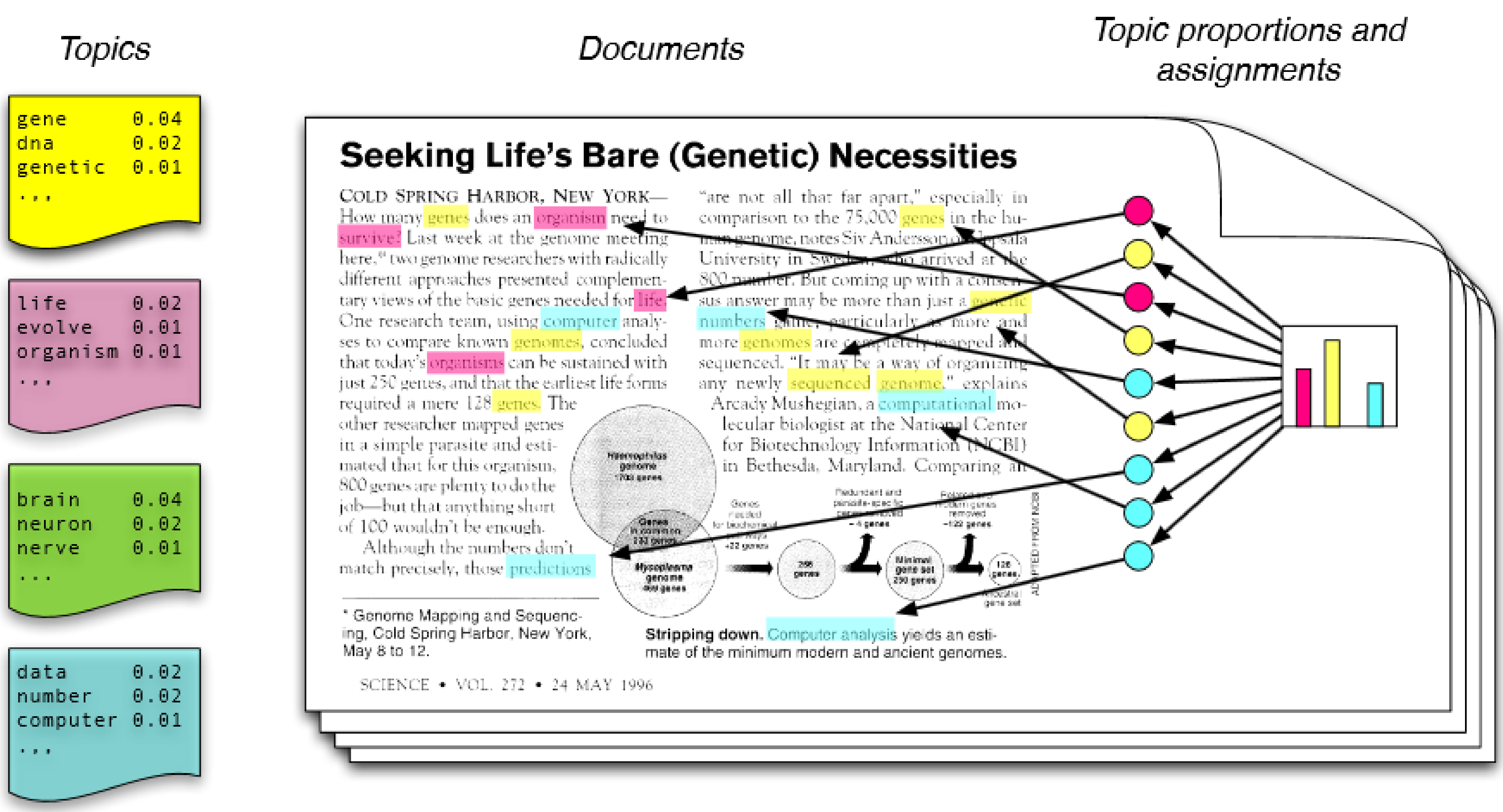
# How to Get the “Latent”?

- Graphical Models ~ Matrix Decomp ~ Tensor Decomp



# Intuition: Documents are made of Topics

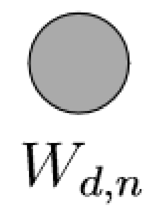
- Every document is a mixture of topics
- Every topic is a distribution over words
- Every word is a draw from a topic



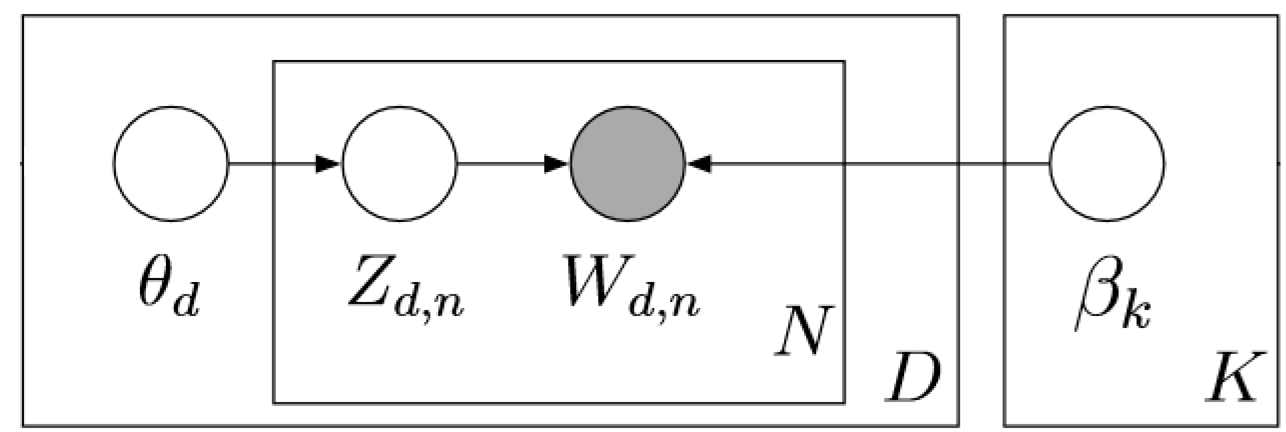
LDA

# Circles & Boxes

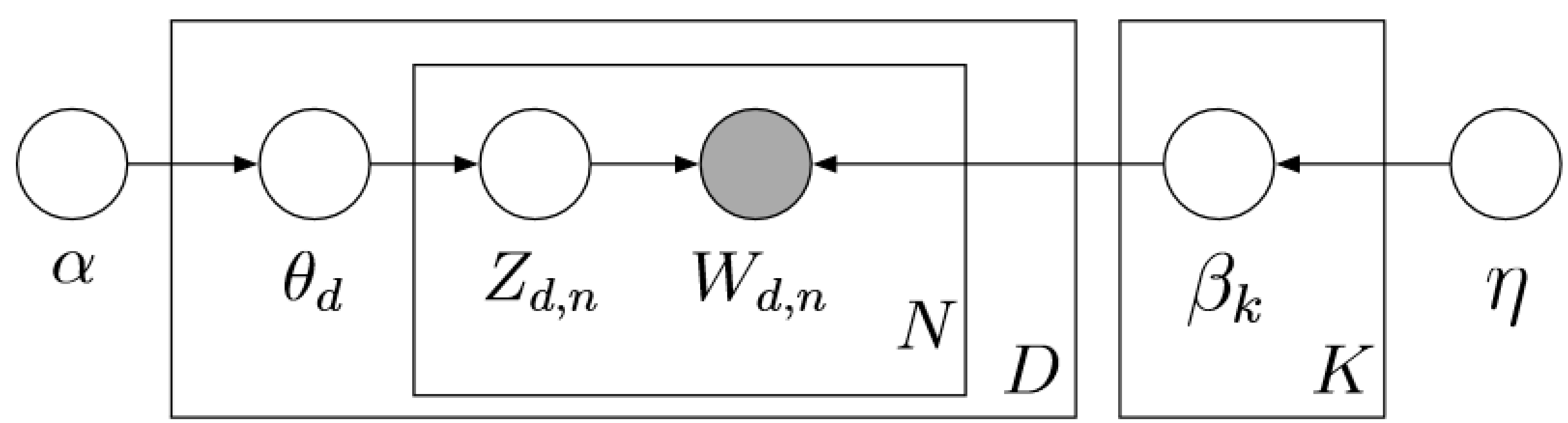
- Observe: N words over D documents




- Infer:
  - Per-word topic assignment $Z_{d,n}$
  - Per-doc topic proportion $\theta_d$
  - Corpus topic distribution $\beta_k$

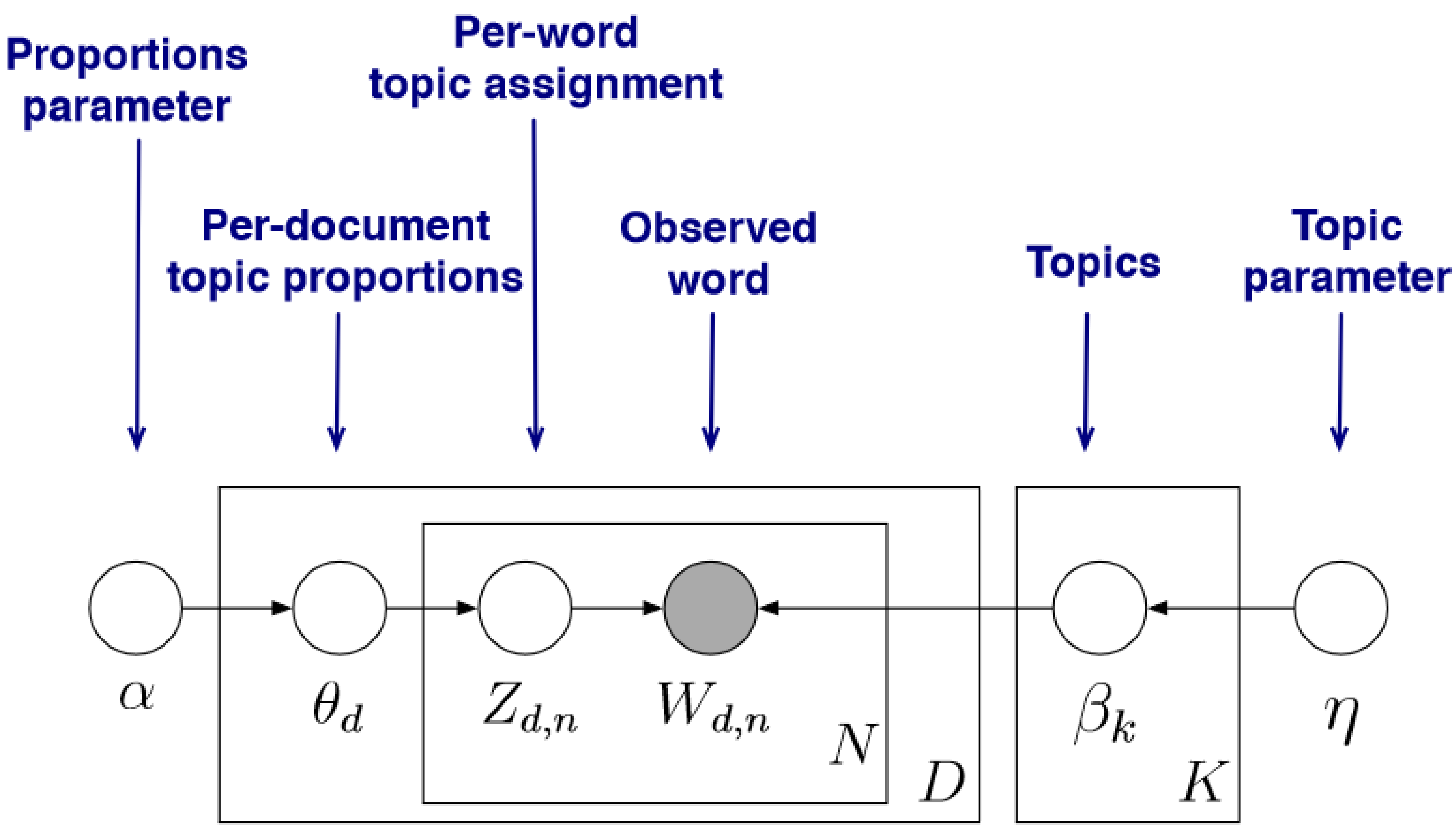



- Dirichlet Priors Give:
  - Sparsity $\alpha$
  - Exclusivity $\eta$



# LDA – Latent Dirichlet Allocation

- We observe words, we infer everything else, with our assumed structure




$$\prod_{i=1}^{K} p(\beta_i \mid \eta) \prod_{d=1}^{D} p(\theta_d \mid \alpha) \left( \prod_{n=1}^{N} p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid \beta_{1:K}, z_{d,n}) \right)$$

# “Dirich-let” It On Too Thick?

- What are $\alpha$ & $\eta$?
- Each hyperparameter is a prior “observation count”:
  - $\alpha$ is the number of times a topic is sampled in a document before having observed anything from the document.
  - $\eta$ is the number of times words are sampled from a topic before any words are observed from the corpus.

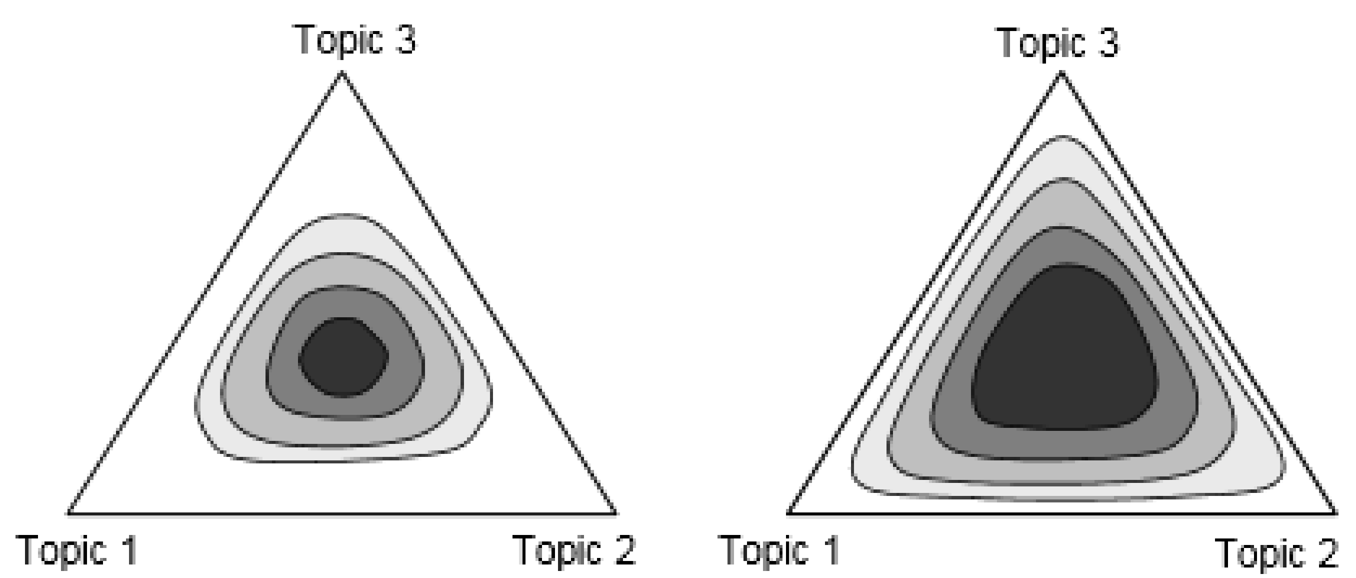
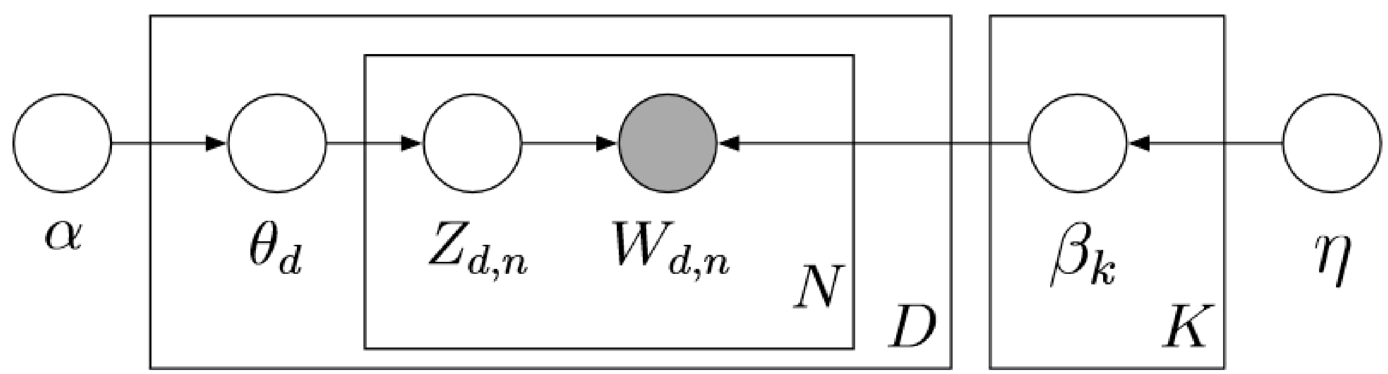



**Figure 3.** Illustrating the symmetric Dirichlet distribution for three topics on a two-dimensional simplex. Darker colors indicate higher probability. Left: $\alpha = 4$. Right: $\alpha = 2$.

# Why Do We Need Inference?

- Want the posterior distribution p(z|w) - assignment of word to topics
- We could estimate $\theta_d$ , $\beta_k$ using EM, or marginalize out with approx. inf.

$$\prod_{i=1}^{K} p(\beta_i \mid \eta) \prod_{d=1}^{D} p(\theta_d \mid \alpha) \left( \prod_{n=1}^{N} p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid \beta_{1:K}, z_{d,n}) \right)$$




- Many Approximate Methods
  - Sampling –*randomly* resample a *specific* tagging for each word, given specific taggings of all other words, and a specific value for θ.
  - Variational Inference - *deterministically* update the *distribution* over taggings for each word, given *distributions* over the taggings for other words and a *distribution* over θ.

# Agenda

- Techniques
  - Topic Models (LDA)
  - Gaussian Processes (GP)
- Applications
  - **KDD 2014 -** Unfolding Physiological State: Mortality Modeling in Intensive Care Units
  - **AAAI 2015 -** A Multivariate Timeseries Modeling Approach to Severity of Illness Assessment and Forecasting in ICU with Sparse, Heterogeneous Clinical Data

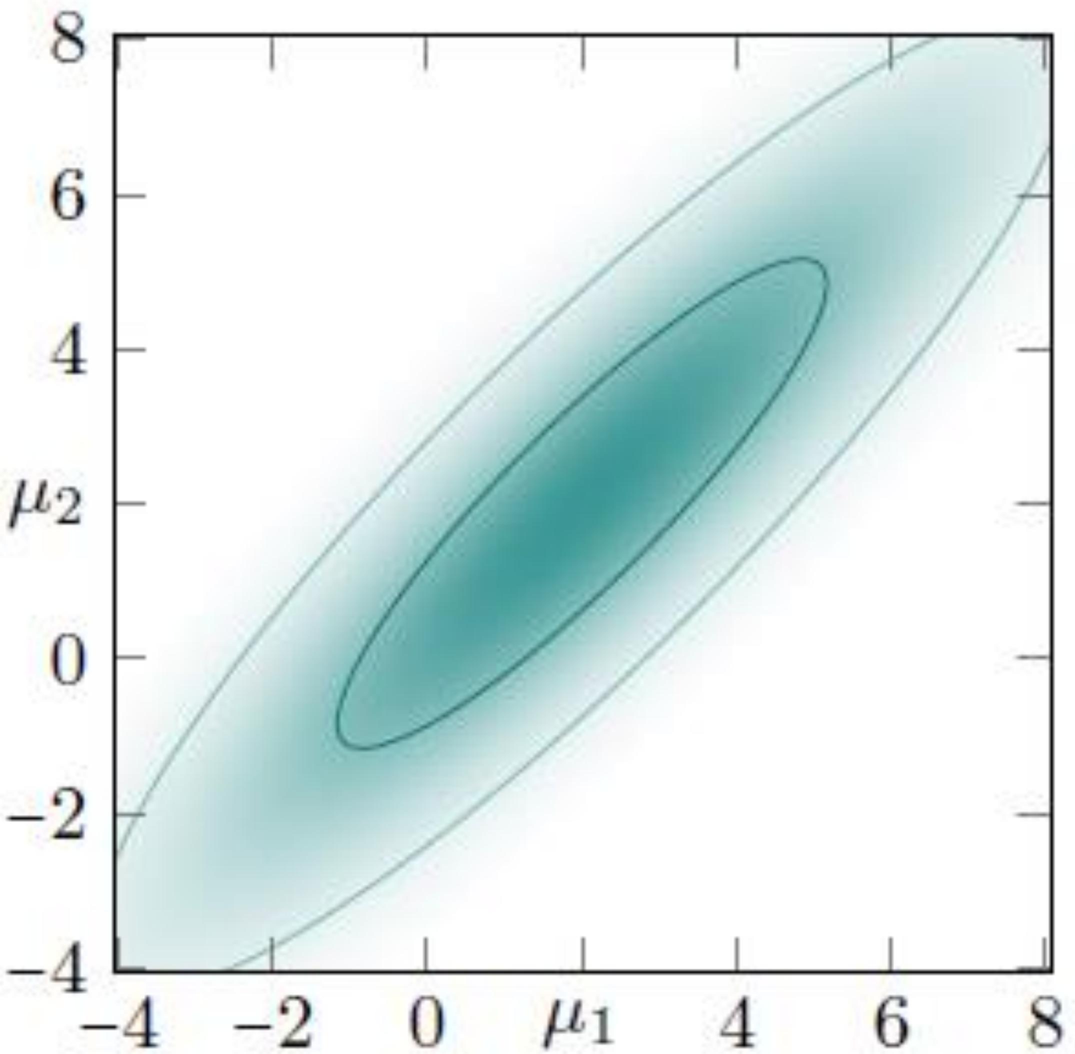
# GP Tutorial

- Content is from:
  - Phillip Henning MLSS 2013 Tutorial
  - Murphy’s Machine Learning Book (and code!)

# GPs?

- GPs define a prior over functions, which can be converted into a posterior over functions once we've seen some data.
- Assumes p(f(x1), … f(xn)) is jointly Gaussian, with some mean and covariance given by
- Computation is $O(N^3)$.
- GPs can be thought of as a Bayesian alternative to sparser/faster kernel methods (SVM), with probabilistic outputs.

# Multivariate Gaussian




$$\mathcal{N}(x;\mu,\Sigma) = \frac{1}{(2\pi)^{N/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu)^\intercal \Sigma^{-1}(x-\mu)\right]$$

- $x, \mu \in \mathbb{R}^N$, $\Sigma \in \mathbb{R}^{N \times N}$
- $\Sigma$ is positive semidefinite

# Why do we like them?

- Closure under multiplication
- Closure under linear maps
- Closure under marginalization

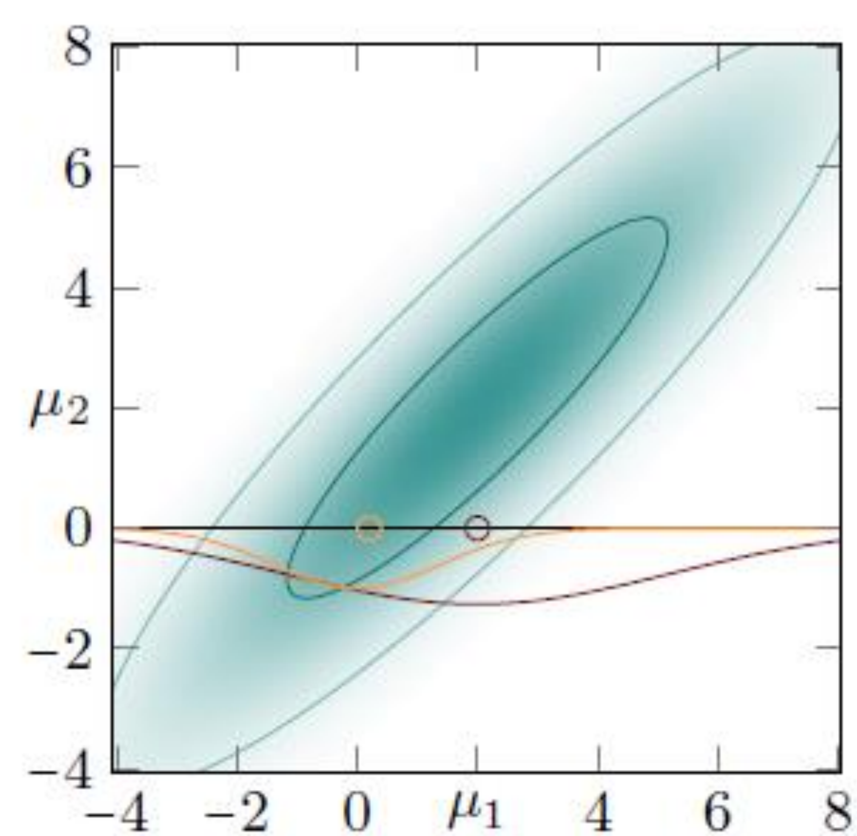
$$\begin{cases} \mathcal{N}(x; a, A)\mathcal{N}(x; b, B) = \mathcal{N}(x; c, C)\mathcal{N}(a; b, A+B) \\ C := (A^{-1} + B^{-1})^{-1} \qquad c := C(A^{-1}a + B^{-1}b) \end{cases}$$

$$\begin{cases} p(z) = \mathcal{N}(z; \mu, \Sigma) \\ p(Az) = \mathcal{N}(Az, A\mu, A\Sigma A^\top) \end{cases}$$

- Closure under conditioning

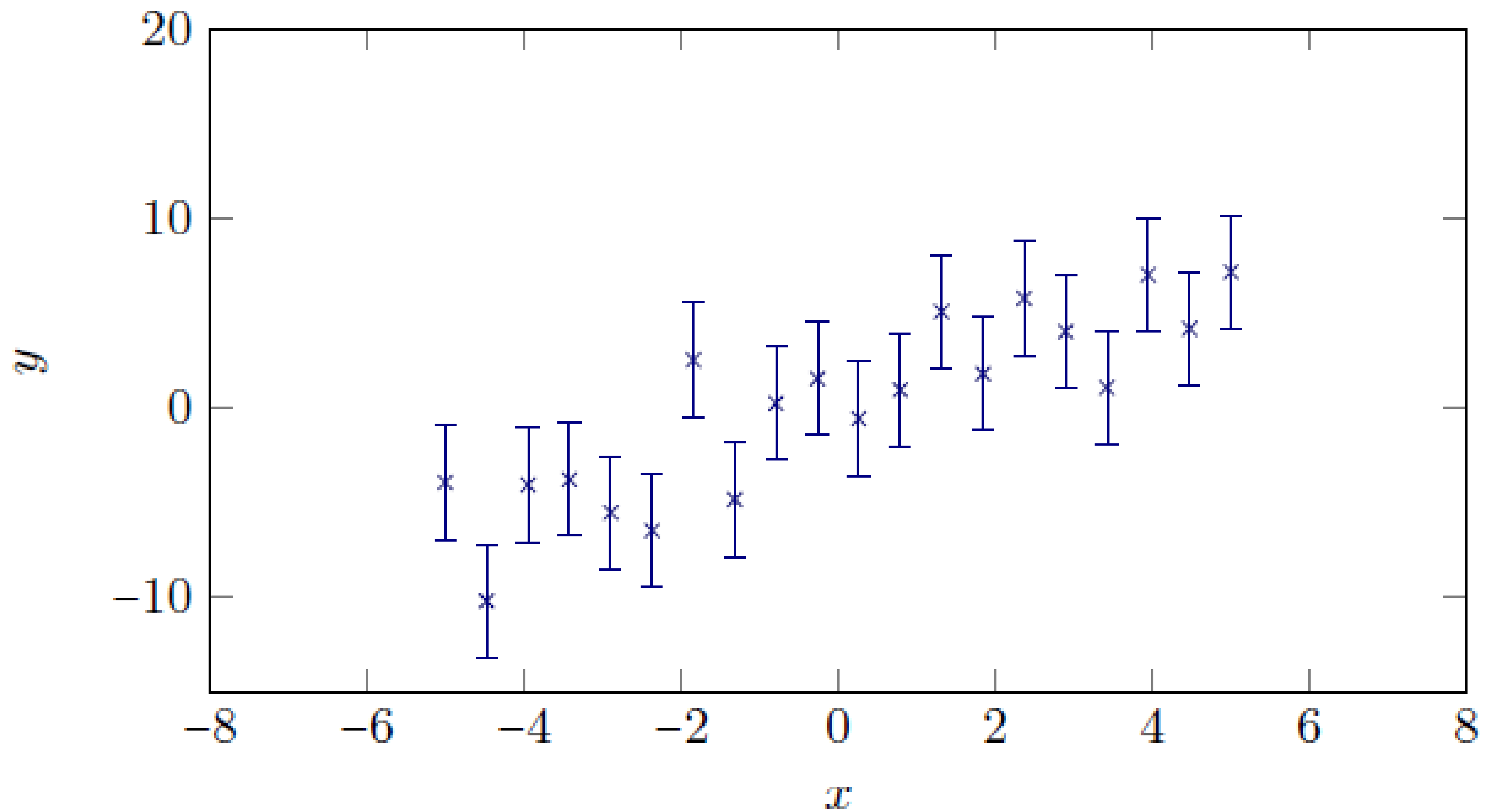
$$\int \mathcal{N}\left[\begin{pmatrix} x \\ y \end{pmatrix}; \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix}\right] \mathrm{d}y = \mathcal{N}(x; \mu_x, \Sigma_{xx})$$

$$p(x \mid y) = \frac{p(x, y)}{p(y)} = \mathcal{N}\left(x; \mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(y - \mu_y), \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}\right)$$



# What can we do?

given $y \in \mathbb{R}^N$, $p(y \mid f)$, what's $f$?

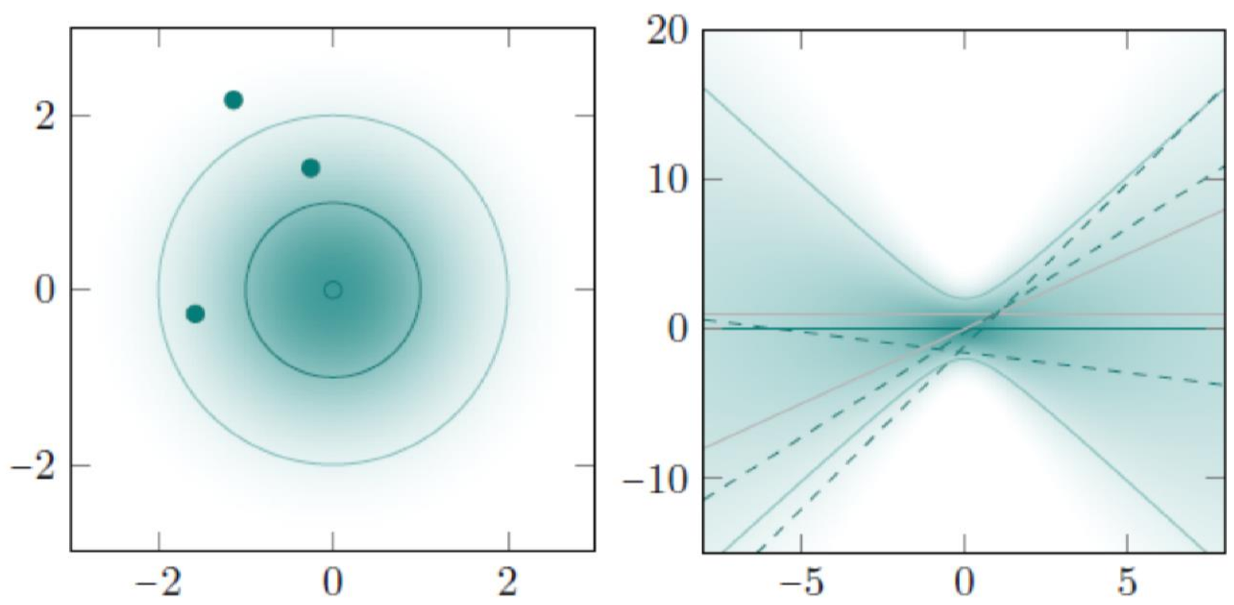


CSAIL

# Linear Regression

$$f(x) = w_1 + w_2 x = \phi_x^\intercal w \qquad p(w) = \mathcal{N}(w; \mu, \Sigma)$$

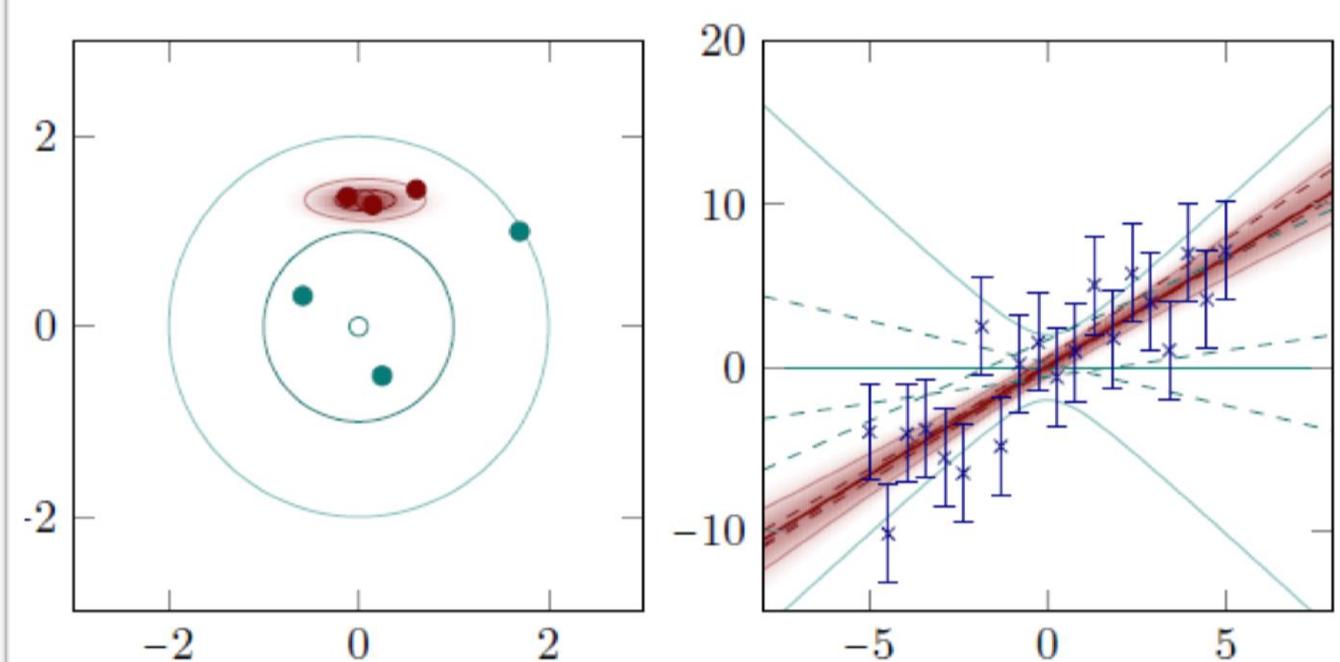
$$\phi_x = \begin{pmatrix} 1 \\ x \end{pmatrix} \qquad p(f) = \mathcal{N}(f; \phi_x^\intercal \mu, \phi_x^\intercal \Sigma \phi_x)$$




Prior over linear functions

$$p(y \mid w, \phi_X) = \mathcal{N}(y; \phi_X^\intercal w, \sigma^2 I)$$

$$p(f_x \mid y, \phi_X) = \mathcal{N}(f_x; \phi_x^\intercal \mu + \phi_x^\intercal \Sigma \phi_X (\phi_X^\intercal \Sigma \phi_X + \sigma^2 I)^{-1} (y - \phi_X^\intercal \mu),$$
$$\phi_x^\intercal \Sigma \phi_x - \phi_x^\intercal \Sigma \phi_X (\phi_X^\intercal \Sigma \phi_X + \sigma^2 I)^{-1} \phi_X^\intercal \Sigma \phi_x$$




Posterior over linear functions

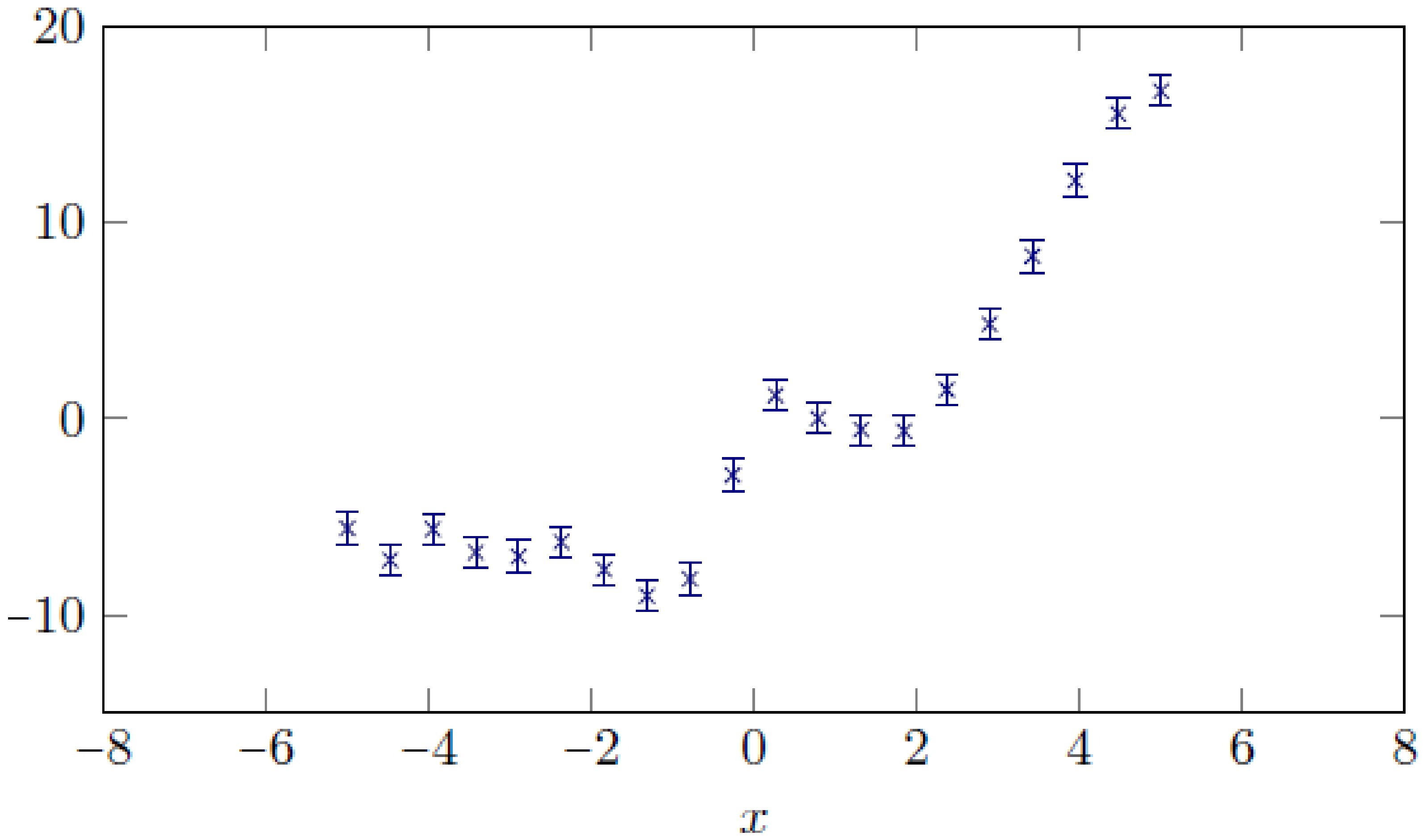
GPs

# Is this hard??

```matlab
% prior on $w$
F     = 2;                                              % number of features
phi   = @(a)(bsxfun(@power,a,0:F-1));                   % $\phi(a) = [1; a]$
mu    = zeros(F,1);
Sigma = eye(F);                                         % $p(w) = \mathcal{N}(\mu, \Sigma)$

% prior on $f(x)$
n     = 100; x = linspace(-6,6,n)';                     % 'test' points
phix  = phi(x);                                         % features of $x$
m     = phix * mu;
kxx   = phix * Sigma * phix';                           % $p(f_x) = \mathcal{N}(m, k_{xx})$
s     = bsxfun(@plus,m,chol(kxx + 1.0e-8 * eye(n))' * randn(n,3));  % samples from prior
stdpi = sqrt(diag(kxx));                                % marginal stddev, for plotting

load('data.mat'); N = length(Y);                        % gives Y,X,sigma

% prior on $Y = f_X + \epsilon$
phiX  = phi(X);                                         % features of data
M     = phiX * mu;
kXX   = phiX * Sigma * phiX';                           % $p(f_X) = \mathcal{N}(M, k_{XX})$

G     = kXX + sigma^2 * eye(N);                         % $p(Y) = \mathcal{N}(M, k_{XX} + \sigma^2 I)$
R     = chol(G);                                        % most expensive step: $\mathcal{O}(N^3)$

kxX   = phix * Sigma * phiX';                           % $\mathrm{cov}(f_x, f_X) = k_{xX}$
A     = kxX / R;                                        % pre-compute for re-use

mpost = m + A * (R' \ (Y-M));       % $p(f_x \mid Y) = \mathcal{N}(m + k_{xX}(k_{XX} + \sigma^2 I)^{-1}(Y - M),$
vpost = kxx - A * A';               % $k_{xx} - k_{xX}(k_{XX} + \sigma^2 I)^{-1} k_{Xx})$
spost = bsxfun(@plus,mpost,chol(vpost + 1.0e-8 * eye(n))' * randn(n,3));     % samples

stdpo = sqrt(diag(vpost));                              % marginal stddev, for plotting
```

# More realistic data

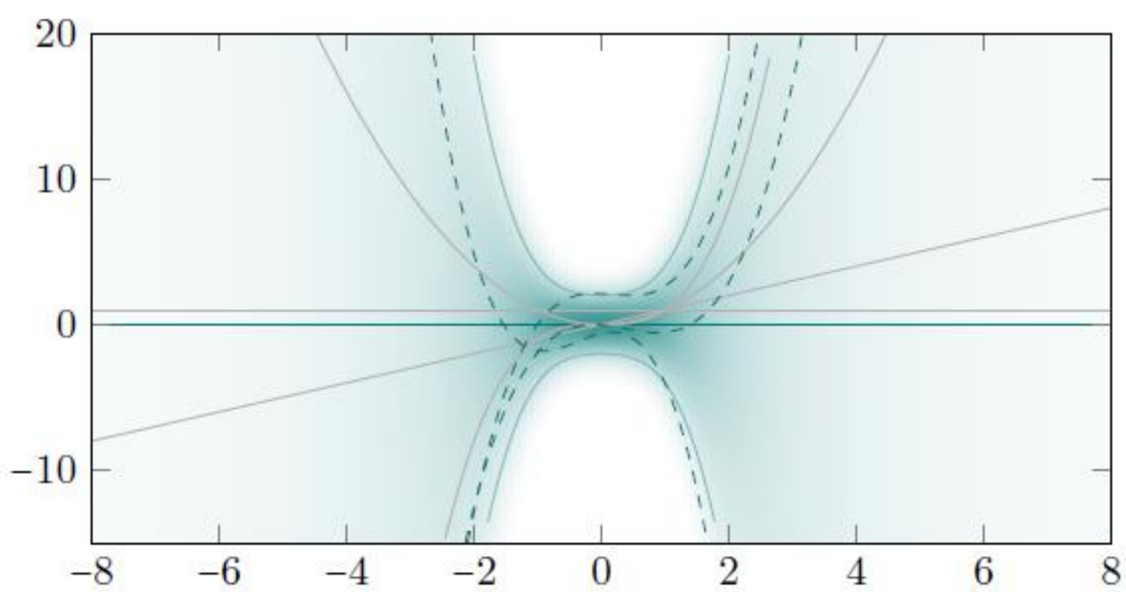
$$f(x) = \phi_x^\top w \qquad ?$$



GPs

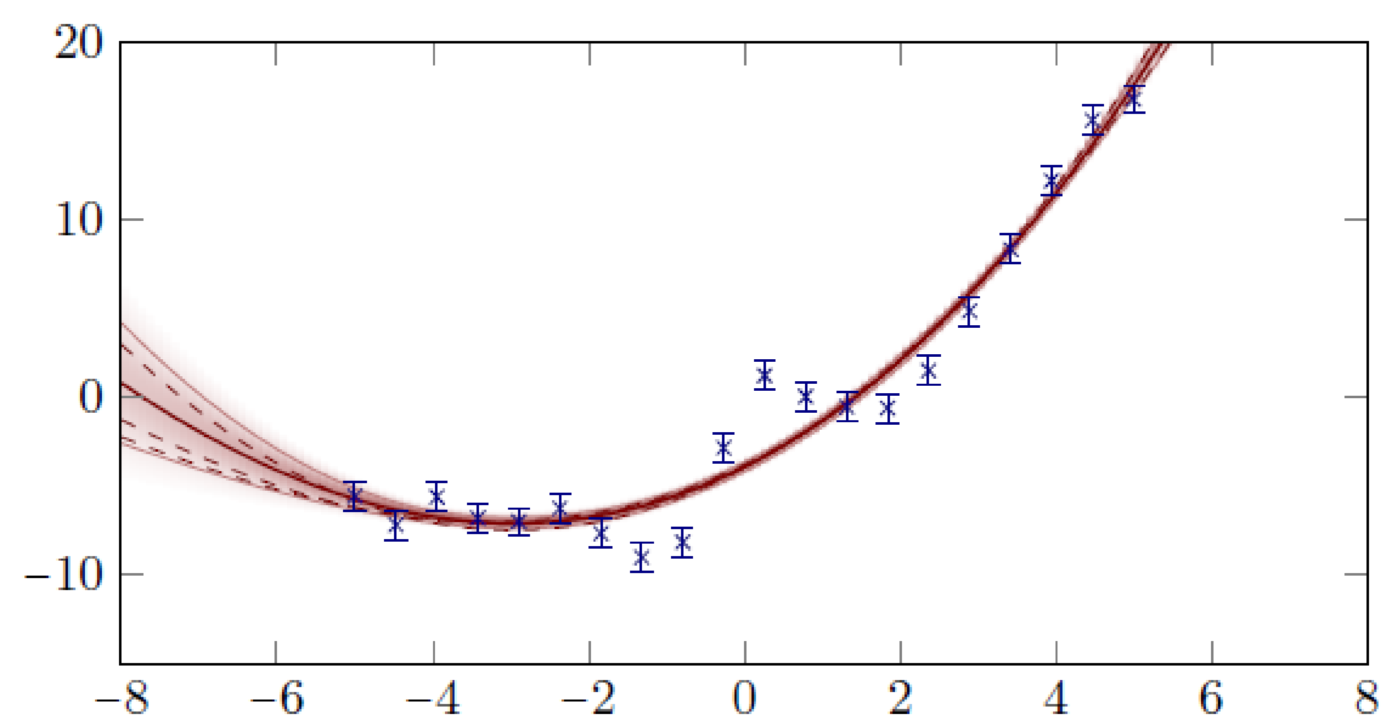
# Cubic Regression

$$f(x) = \phi(x)^\top w \qquad \phi(x) = \begin{pmatrix}1 & x & x.^2 & x.^3\end{pmatrix}^\top$$




Prior

$$f(x) = \phi(x)^\top w \qquad \phi(x) = \begin{pmatrix}1 & x & x.^2 & x.^3\end{pmatrix}^\top$$




Posterior

# Not any harder.

F = 4

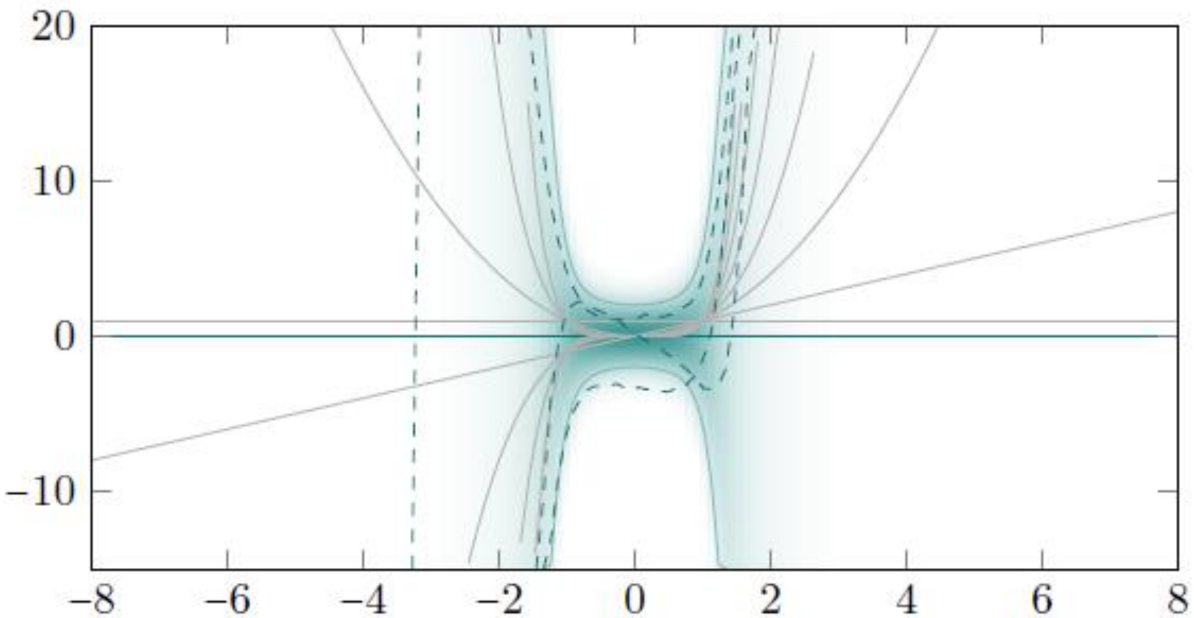
```
% prior on w
F     = 2;                                                % number of features
phi   = @(a)(bsxfun(@power,a,0:F-1));                     % φ(a) = [1;a]
mu    = zeros(F,1);
Sigma = eye(F);                                           % p(w) = N(μ,Σ)

% prior on f(x)
n     = 100; x = linspace(-6,6,n)';                       % 'test' points
phix  = phi(x);                                           % features of x
m     = phix * mu;
kxx   = phix * Sigma * phix';                             % p(f_x) = N(m,k_xx)
s     = bsxfun(@plus,m,chol(kxx + 1.0e-8 * eye(n))' * randn(n,3));  % samples from prior
stdpi = sqrt(diag(kxx));                                  % marginal stddev, for plotting

load('data.mat'); N = length(Y);                          % gives Y,X,sigma

% prior on Y = f_X + ε
phiX  = phi(X);                                           % features of data
M     = phiX * mu;
kXX   = phiX * Sigma * phiX';                             % p(f_X) = N(M,k_XX)

G     = kXX + sigma^2 * eye(N);                           % p(Y) = N(M,k_XX + σ^2 I)
R     = chol(G);                                          % most expensive step: O(N^3)

kxX   = phix * Sigma * phiX';                             % cov(f_x,f_X) = k_xX
A     = kxX / R;                                          % pre-compute for re-use

mpost = m + A * (R' \ (Y-M));          % p(f_x|Y) = N(m + k_xX(k_XX + σ^2 I)^{-1}(Y - M),
vpost = kxx - A * A';                  % k_xx - k_xX(k_XX + σ^2 I)^{-1} k_Xx)
spost = bsxfun(@plus,mpost,chol(vpost + 1.0e-8 * eye(n))' * randn(n,3));    % samples

stdpo = sqrt(diag(vpost));                                % marginal stddev, for plotting
```

GPs

```
phi = @(a)(bsxfun(@power,a,[0:7]));
```

# Septic Regression

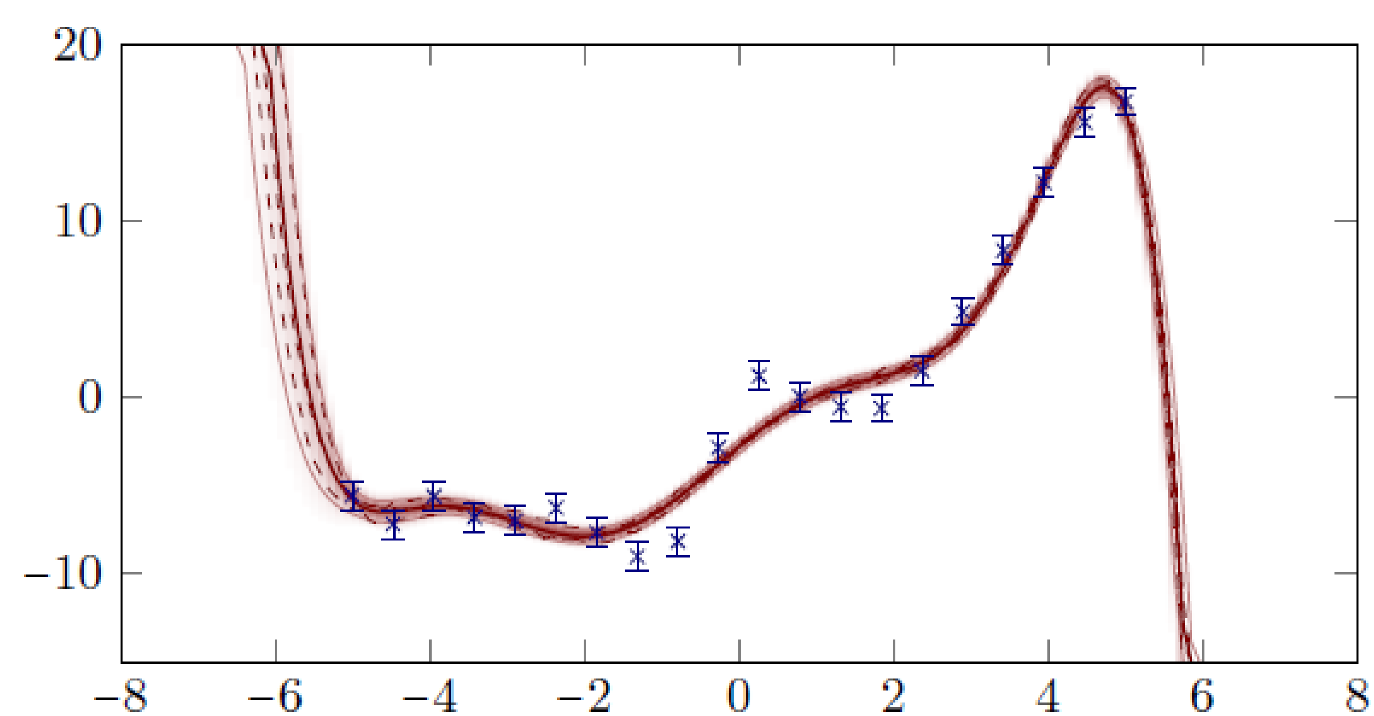
$$f(x) = \phi(x)^\top w \qquad \phi(x) = \begin{pmatrix}1 & x & x.^2 & \cdots & x.^7\end{pmatrix}^\top$$




Prior

Posterior

$$f(x) = \phi(x)^\top w \qquad \phi(x) = \begin{pmatrix}1 & x & x.^2 & \cdots & x.^7\end{pmatrix}^\top$$

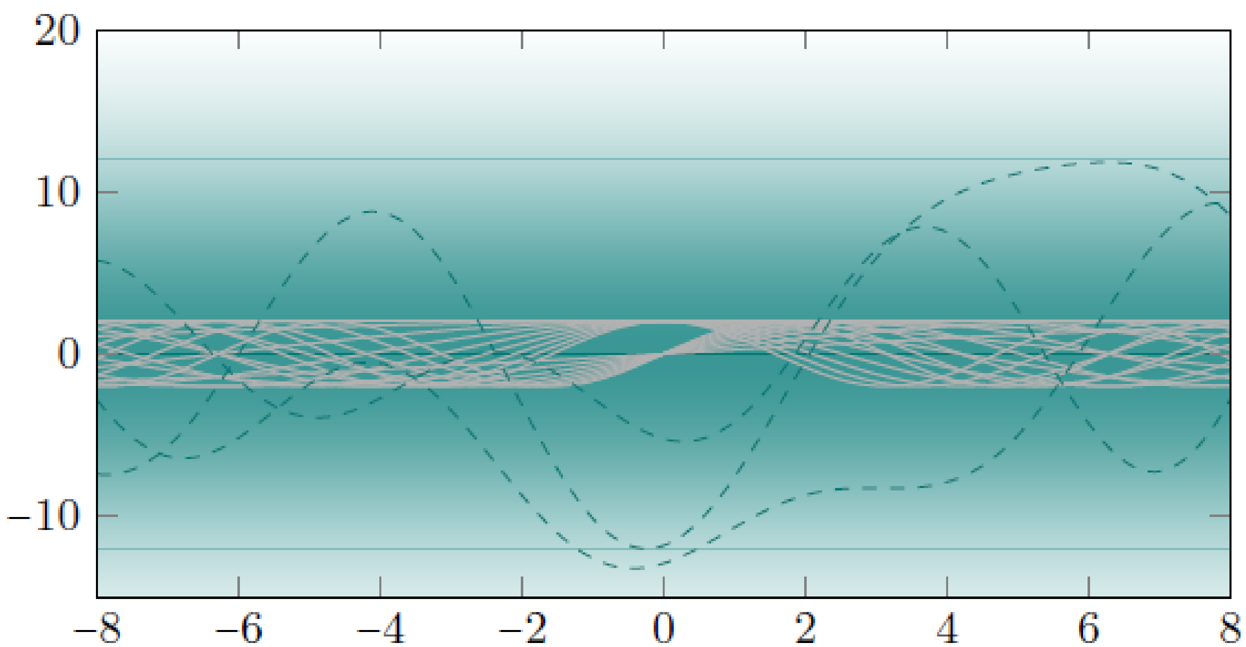


GPs

```
phi = @(a)(2 * [cos(bsxfun(@times,a/8,[0:8])), sin(bsxfun(@times,a/8,[1:8]))]);
```

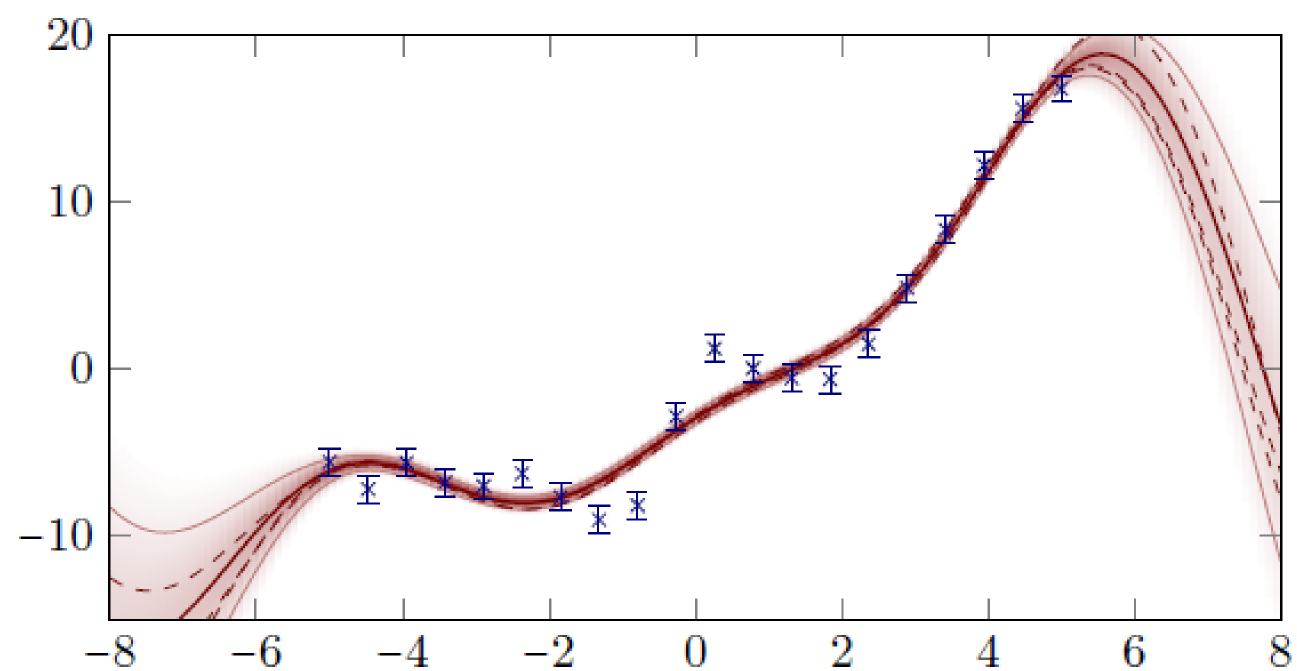
# Fourier Regression

$$\phi(x) = \begin{pmatrix}\cos(x) & \cos(2x) & \cos(3x) & \dots & \sin(x) & \sin(2x) & \dots\end{pmatrix}^{\top}$$




Prior

$$\phi(x) = \begin{pmatrix}\cos(x) & \cos(2x) & \cos(3x) & \dots & \sin(x) & \sin(2x) & \dots\end{pmatrix}^{\top}$$



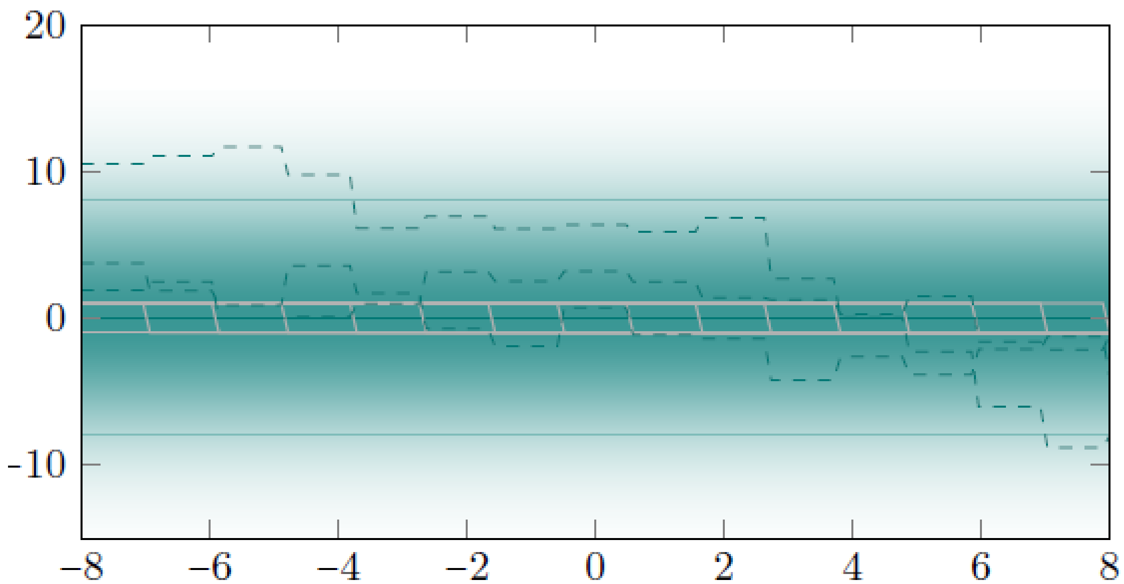

Posterior

GPs

```
phi = @(a)(-1 + 2 * bsxfun(@lt,a,linspace(-8,8,16)));
```

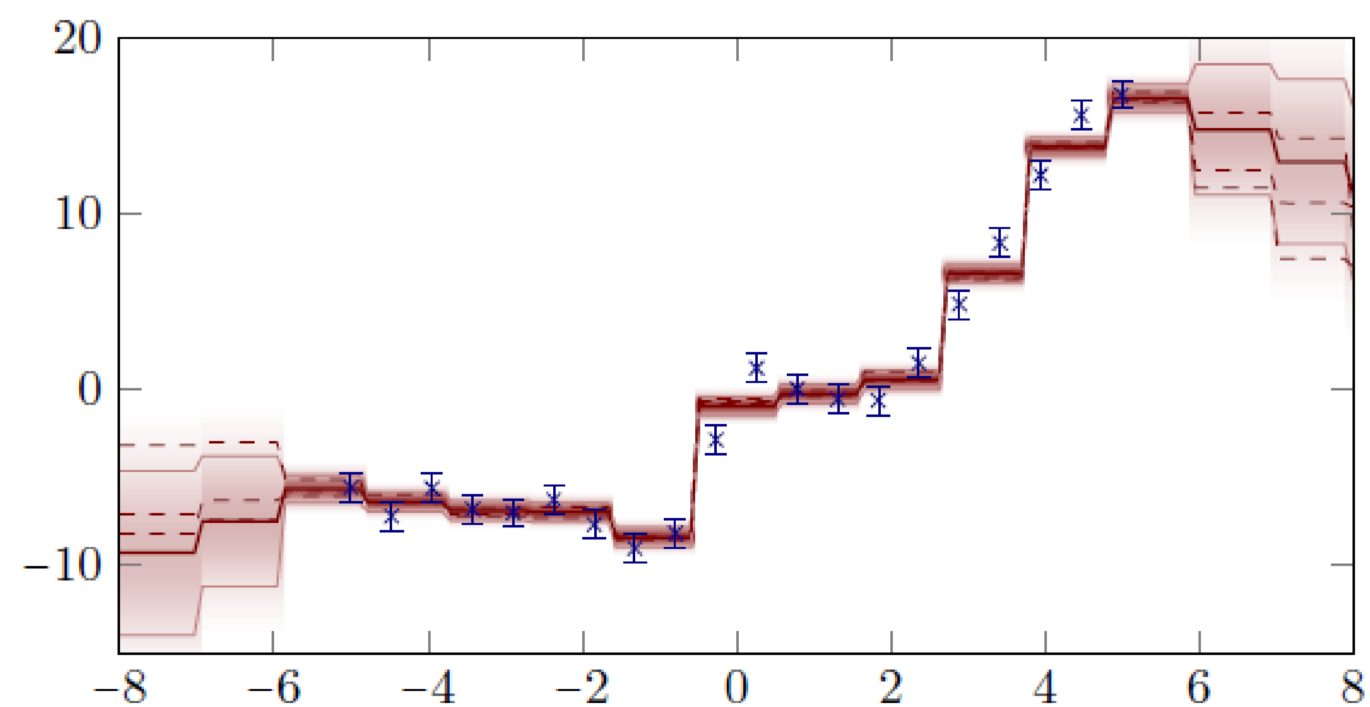
# Step Regression

$$\phi(x) = -1 + 2\left(\theta(x-8) \quad \theta(8-x) \quad \theta(x-7) \quad \theta(7-x) \quad \ldots\right)^{\top}$$




Prior

$$\phi(x) = -1 + 2\left(\theta(x-8) \quad \theta(8-x) \quad \theta(x-7) \quad \theta(7-x) \quad \ldots\right)^{\top}$$




Posterior

# How many features should we use?

$$p(f_x \mid y, \phi_X) = \mathcal{N}(f_x; \phi_x^\intercal \mu + \phi_x^\intercal \Sigma \phi_X (\phi_X^\intercal \Sigma \phi_X + \sigma^2 I)^{-1} (y - \phi_X^\intercal \mu),$$
$$\phi_x^\intercal \Sigma \phi_x - \phi_x^\intercal \Sigma \phi_X (\phi_X^\intercal \Sigma \phi_X + \sigma^2 I)^{-1} \phi_X^\intercal \Sigma \phi_x)$$

all objects involving $\phi$ are of the form

- $\phi^\intercal \mu$ — the mean function
- $\phi^\intercal \Sigma \phi$ — the kernel

once these are known, cost is independent of the number of features

remember the code:

```
M      = phiX * mu;
m      = phix * mu;
kXX    = phiX * Sigma * phiX';                  % p(f_X) = N(M, k_XX)
kxx    = phix * Sigma * phix';                  % p(f_x) = N(m, k_xx)
kxX    = phix * Sigma * phiX';                  % cov(f_x, f_X) = k_xX
```

# Pre-compute the kernel

```
% prior
F     = 2;                                                         % number of features
phi   = @(a)(bsxfun(@power,a,0:F));                                % \phi(a) = [1; a]
k     = @(a,b)(phi(a)' * phi(b));                                  % kernel
mu    = @(a)(zeros(size(a,1)));                                    % mean function

% belief on f(x)
n     = 100; x = linspace(-6,6,n)';                                % 'test' points
m     = mu(x);
kxx   = k(x,x);                                                    % p(f_x) = N(m, k_xx)
s     = bsxfun(@plus,m,chol(kxx + 1.0e-8 * eye(n))' * randn(n,3)); % samples from prior
stdpi = sqrt(diag(kxx));                                           % marginal stddev, for plotting

load('data.mat'); N = length(Y);                                   % gives Y,X,sigma

% prior on Y = f_X + \epsilon
M     = mu(X);
kXX   = k(X,X);                                                    % p(f_X) = N(M, k_XX)

G     = kXX + sigma^2 * eye(N);                                    % p(Y) = N(M, k_XX + \sigma^2 I)
R     = chol(G);                                                   % most expensive step: O(N^3)

kxX   = k(x,X);                                                    % cov(f_x, f_X) = k_xX
A     = kxX / R;                                                   % pre-compute for re-use

mpost = m + A * (R' \ (Y-M));                  % p(f_x | Y) = N(m + k_xX (k_XX + \sigma^2 I)^{-1} (Y - M),
vpost = kxx - A * A';                          % k_xx - k_xX (k_XX + \sigma^2 I)^{-1} k_Xx)
spost = bsxfun(@plus,mpost,chol(vpost + 1.0e-8 * eye(n))' * randn(n,3));     % samples

stdpo = sqrt(diag(vpost));                                         % marginal stddev, for plotting
```

# Kernalization

## Definition

*A function $k : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ is a Mercer kernel if, for any finite collection $X = [x_1, \dots, x_N]$, the matrix $k_{XX} \in \mathbb{R}^{N \times N}$ with elements $k_{XX,(i,j)} = k(x_i, x_j)$ is positive semidefinite.*

## Lemma

*Any kernel that can be written as*

$$k(x, x') = \sum\!\!\!\!\!\!\int \phi_\ell(x)\phi_\ell(x')\, d\ell$$

*is a Mercer kernel.* *(assuming integral over positive set)*

**Proof:** $\forall X \in \mathbb{X}^N, v \in \mathbb{R}^N$

$$v^\top k_{XX} v = \sum\!\!\!\!\!\!\int \sum_i^N v_i \phi_\ell(x_i) \sum_j^N v_j \phi_\ell(x_j)\, d\ell = \sum\!\!\!\!\!\!\int \left[\sum_i v_i \phi_\ell(x_i)\right]^2 d\ell \geq 0 \quad \square$$

# Gaussian Process Priors

## Definition

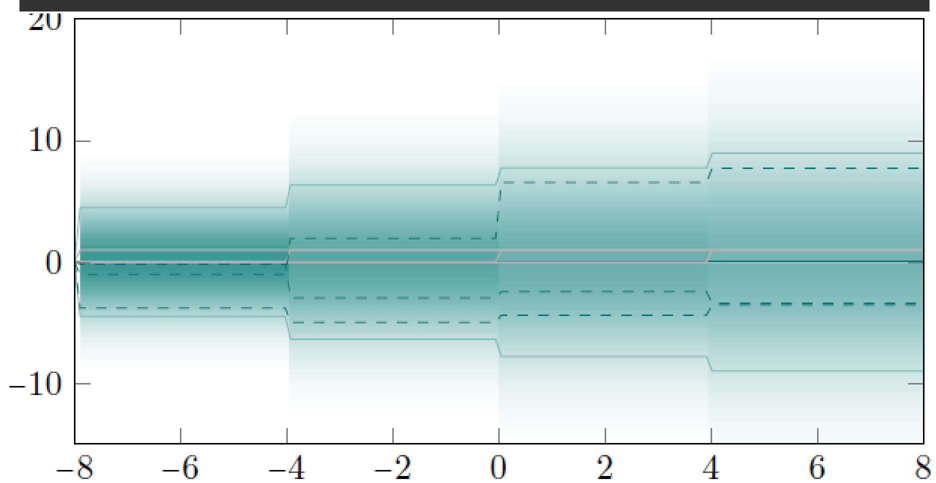
*A function* $k : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ *is a Mercer kernel if, for any finite collection* $X = [x_1, \dots, x_N]$*, the matrix* $k_{XX} \in \mathbb{R}^{N \times N}$ *with elements* $k_{XX,(i,j)} = k(x_i, x_j)$ *is positive semidefinite.*

## Definition

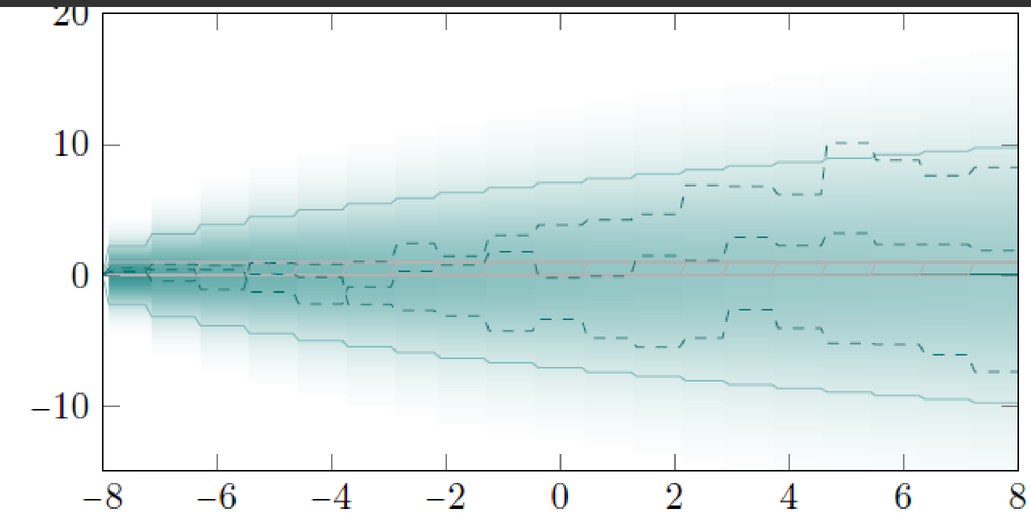
*Let* $\mu : \mathbb{X} \to \mathbb{R}$ *be any function,* $k : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ *be a Mercer kernel.*
*A Gaussian process* $p(f) = \mathcal{GP}(f; \mu, k)$ *is a probability distribution over the function* $f : \mathbb{X} \to \mathbb{R}$*, such that every finite restriction to function values* $f_X := [f_{x_1}, \dots, f_{x_N}]$ *is a Gaussian distribution* $p(f_X) = \mathcal{N}(f_X; \mu_X, k_{XX})$*.*

# E.g. Kernelization of Step Functions

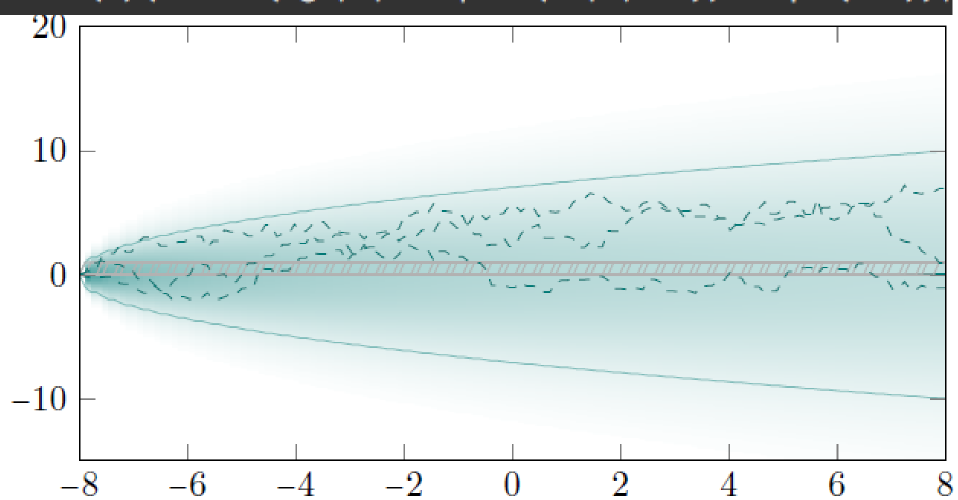
```
phi = @(a)(bsxfun(@gt,a,linspace(-8,8,5))./sqrt(5));
```



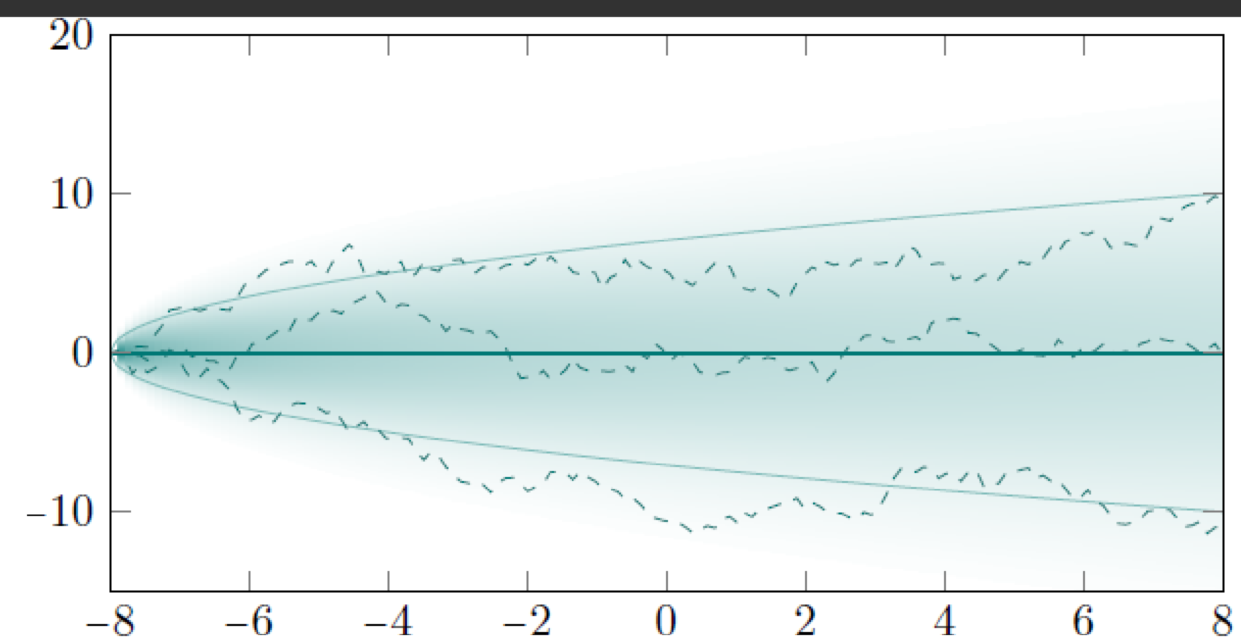

```
phi = @(a)(bsxfun(@gt,a,linspace(-8,8,20))./sqrt(20));
```




```
phi = @(a)(bsxfun(@gt,a,linspace(-8,8,100))./sqrt(100));
```




```
k = @(a,b)(theta.^2 * bsxfun(@min,a+8,b'+8)/16);
```

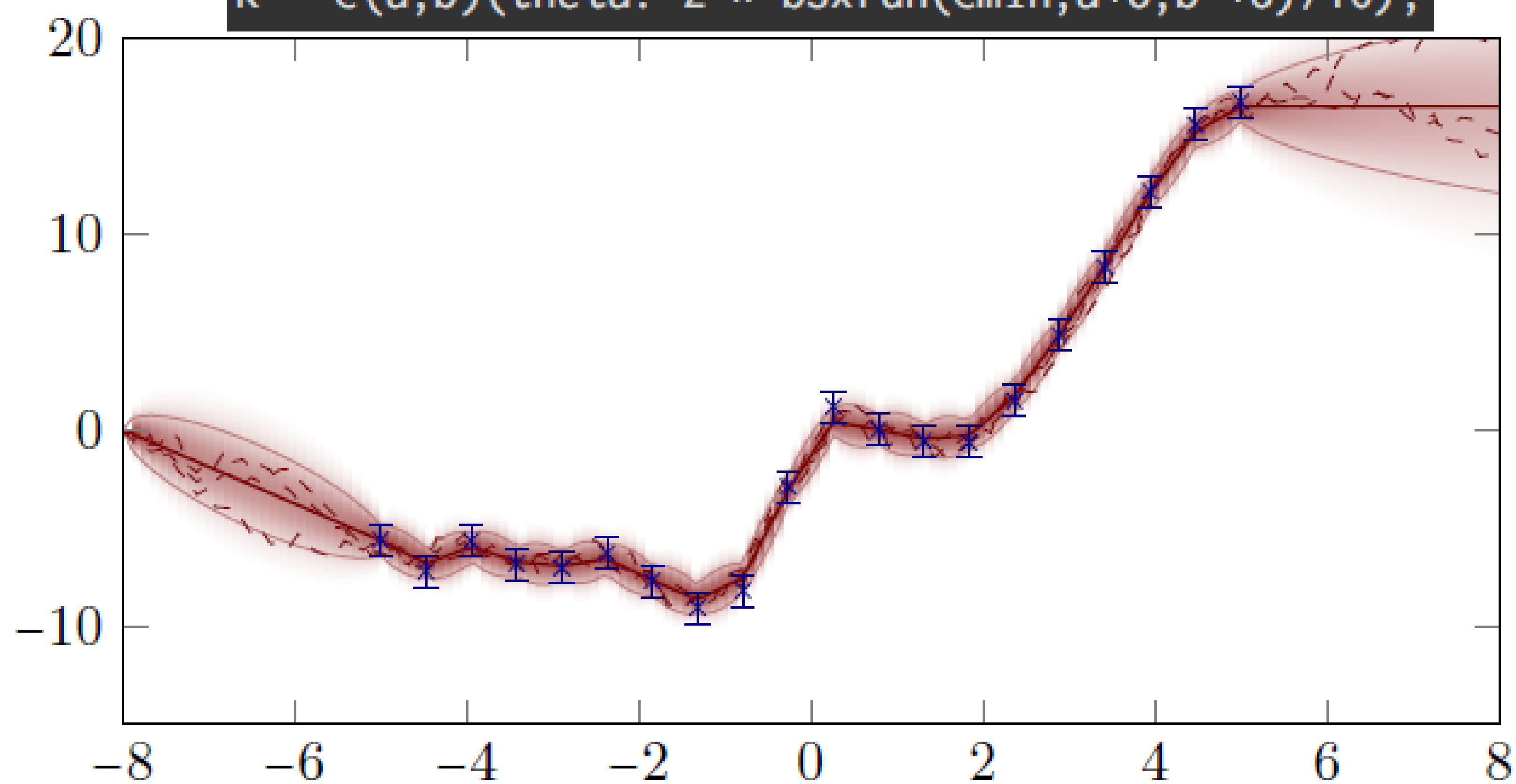


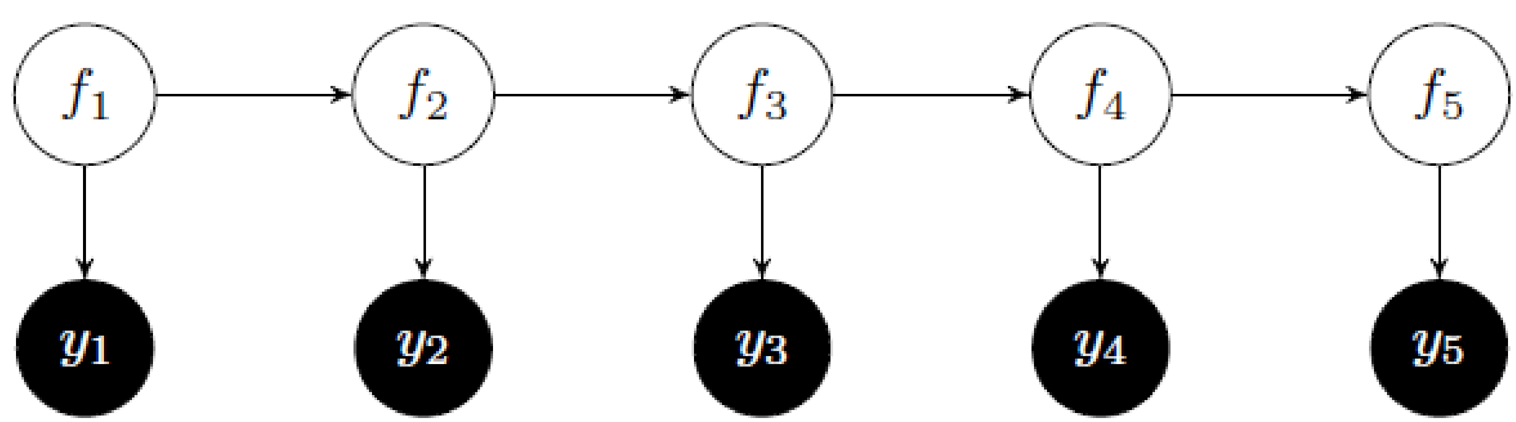

$$\operatorname{cov}(f_{x_i}, f_{x_j}) = \int_{c_{\min}}^{\infty} \theta(x_i - c)\theta(x_j - c)\,\mathrm{d}c = \min(x_i, x_j) - c_{\min}$$

aka. the Wiener process

# Applying It

```
k = @(a,b)(theta.^2 * bsxfun(@min,a+8,b'+8)/16);
```



# Summary

Gaussians are closed under

- linear projection / marginalization / sum rule
- linear restriction / conditioning / product rule

they provide the linear algebra of inference

combine with nonlinear features $\phi$, get nonlinear regression

in fact, number of features can be infinite

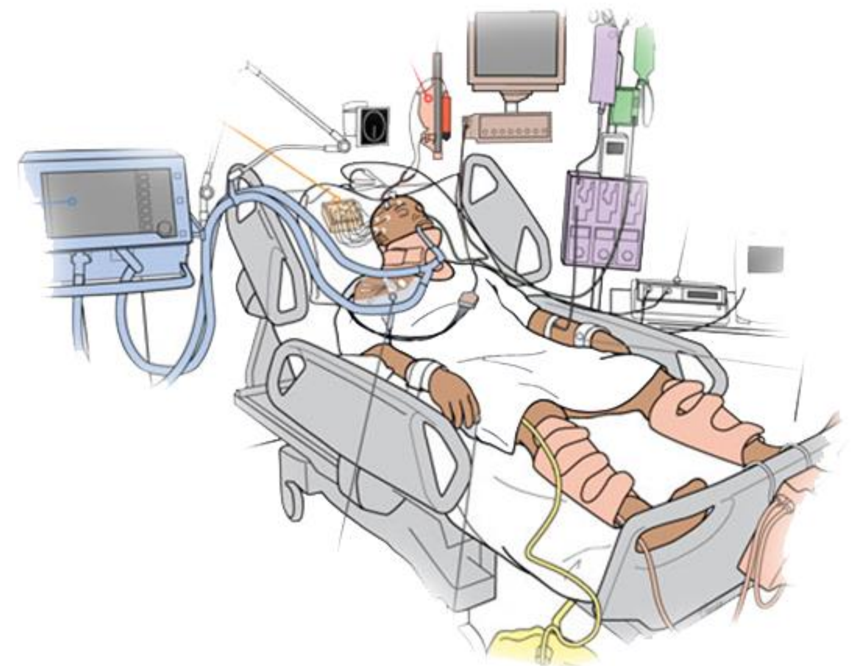
(nonparametric) Gaussian process regression

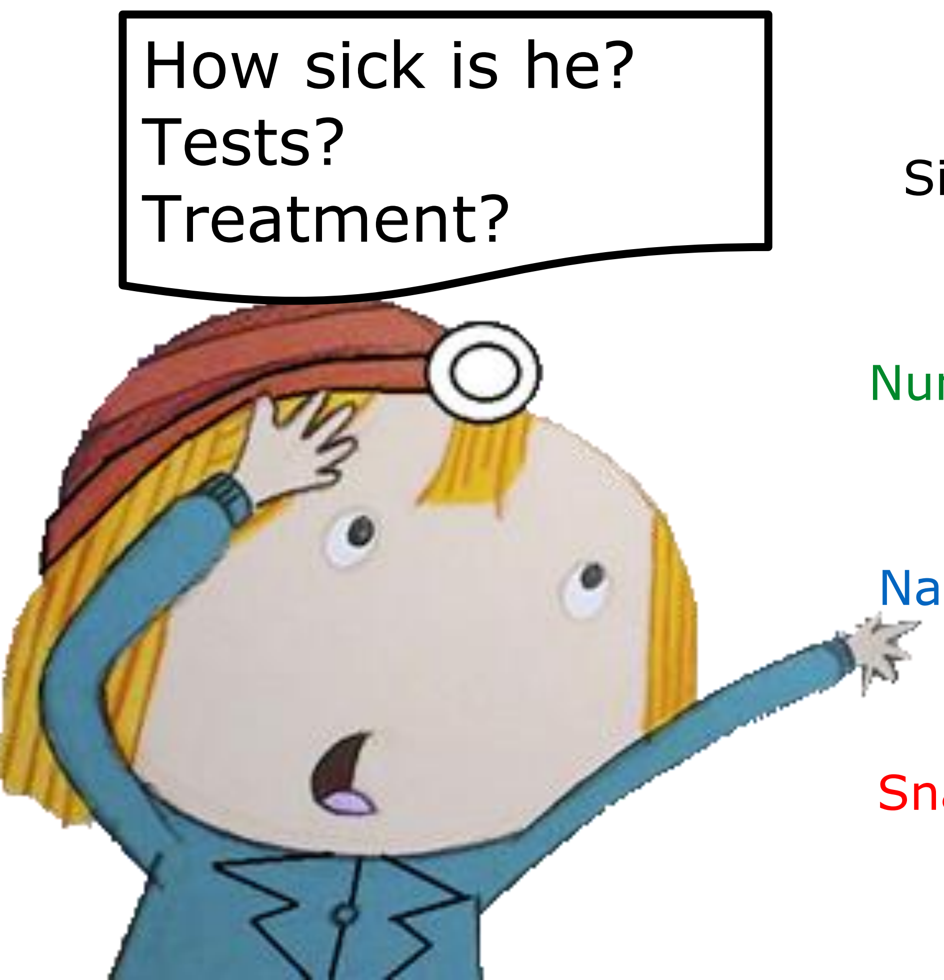
# Agenda

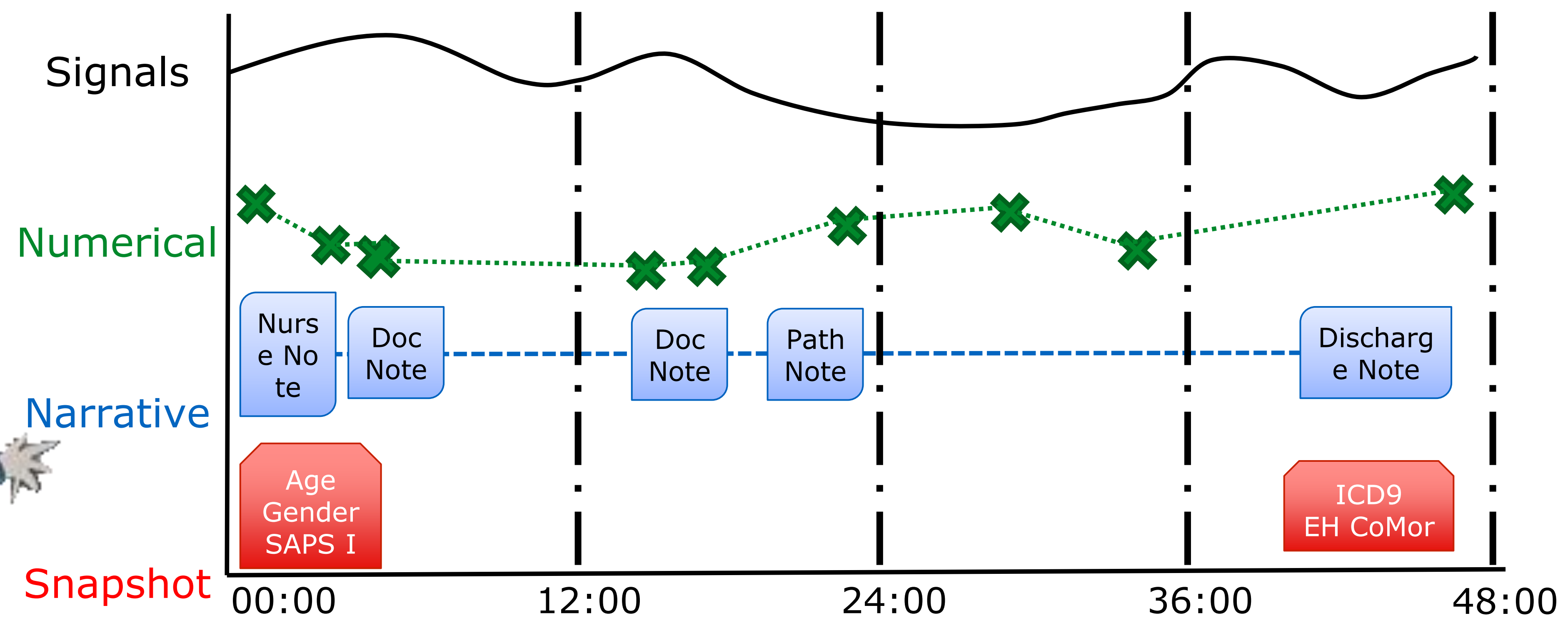
- Techniques
  - Topic Models (LDA)
  - Gaussian Processes (GP)
- Applications
  - **KDD 2014 -** Unfolding Physiological State: Mortality Modeling in Intensive Care Units
  - **AAAI 2015 -** A Multivariate Timeseries Modeling Approach to Severity of Illness Assessment and Forecasting in ICU with Sparse, Heterogeneous Clinical Data

# We've Got A Really Big Problem

- ICUs are busy, and carestaff are often inundated with information.
- Which patient needs attention?





# What Do We Already Know?

- In 2009, 118 validated mortality prediction tools published.**
  - Modest accuracy
  - Large variability
  - Models based on numeric, waveform, or snapshot data
  - Snapshot data (e.g. ICD9) is not “realtime” or actionable

- A good predictive rule must be*:
  - Accurate in a wide variety of clinical settings
  - Easy to incorporate into routine clinical practice
  - Improves prognostic accuracy

* Grady, Deborah, and Seth A. Berkowitz. "Why is a good clinical prediction rule so hard to find?." *Archives of internal medicine* 171.19 (2011): 1701-1702.

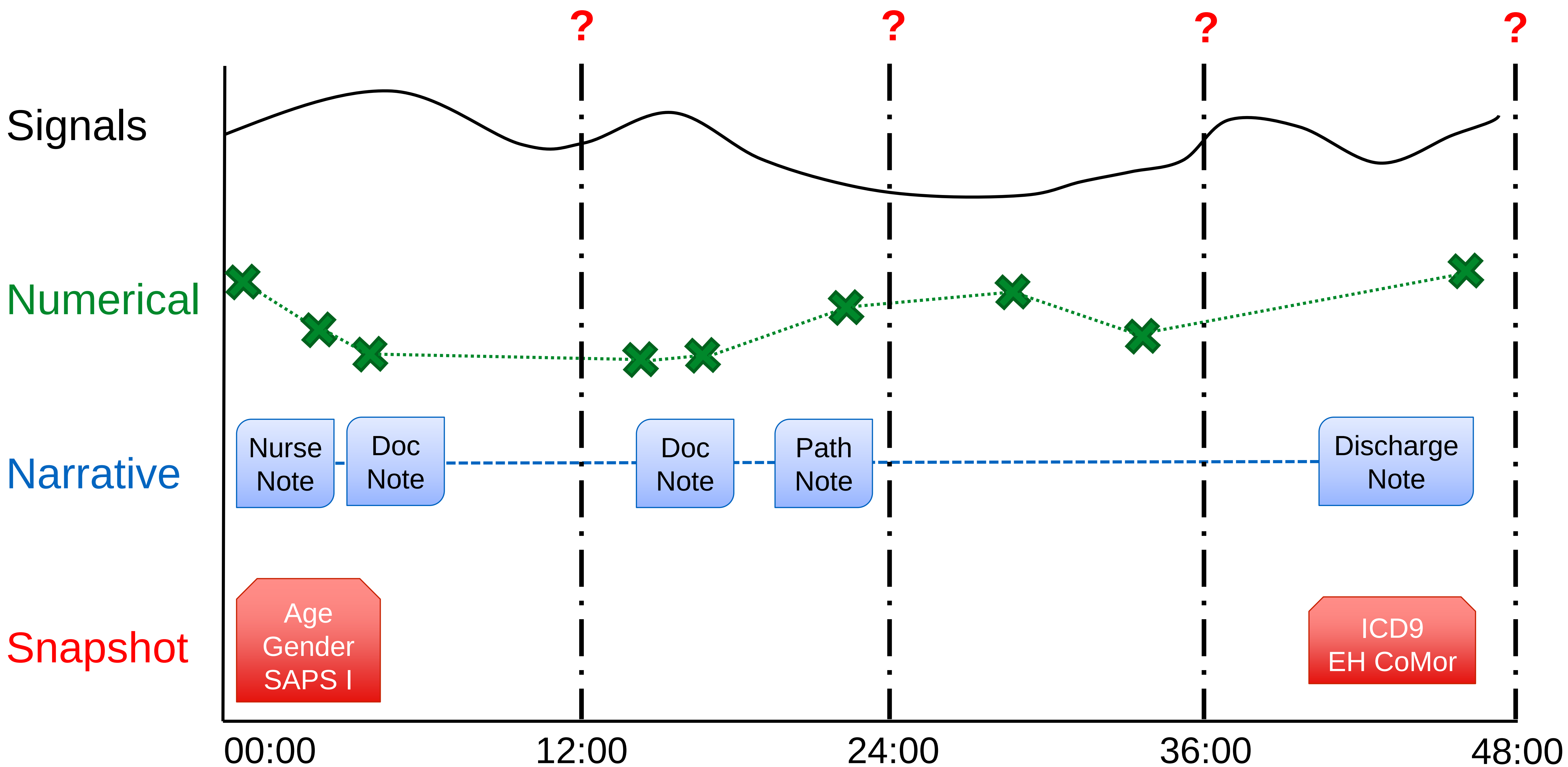
** Siontis, George CM, Ioanna Tzoulaki, and John PA Ioannidis. "Predicting death: an empirical evaluation of predictive tools for mortality." *Archives of internal medicine* 171.19 (2011): 1721-1726.

# Unfolding Physiological State: Mortality Modeling in Intensive Care Units

- KDD 2014

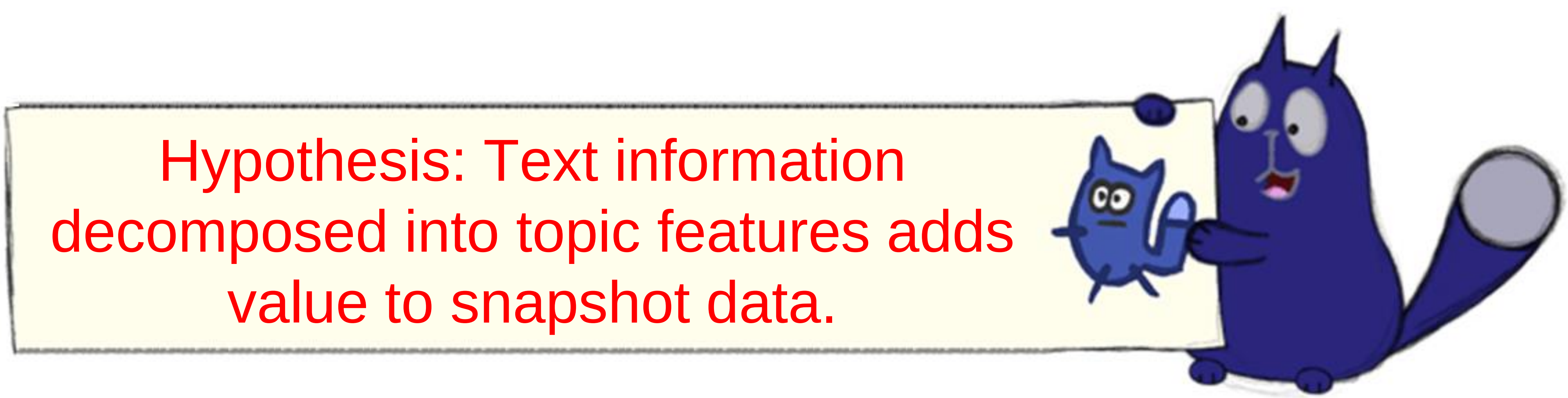
- Marzyeh Ghassemi, Tristan Naumann, Finale Doshi-Velez, Nicole Brimmer, Rohit Joshi, Anna Rumshisky, Peter Szolovits

# Lots of Data Sources

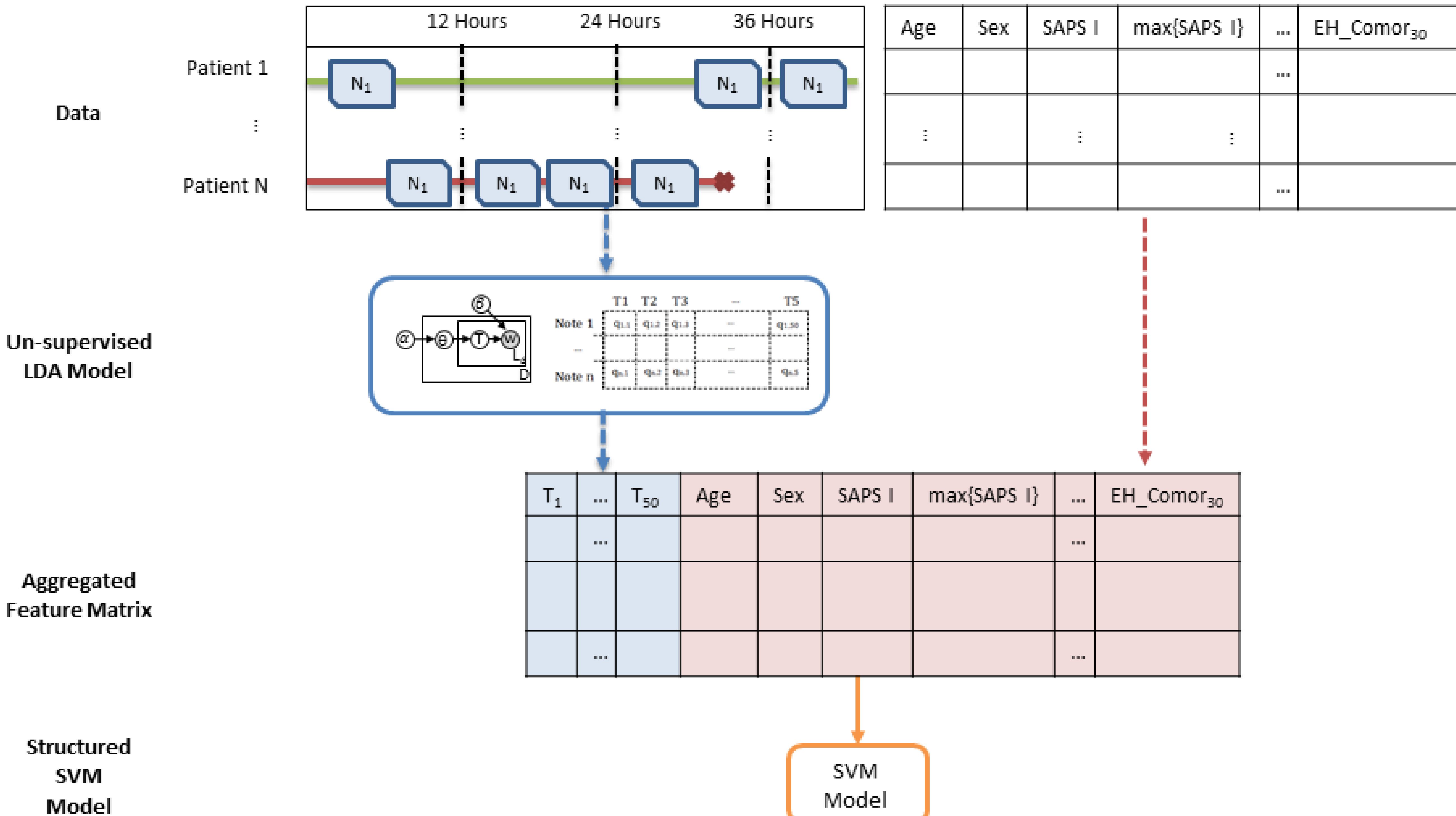


# Every Cat Needs a Plan

- Create forward-facing models every 12 hours that only use data what would have actually been available, or “realtime” data.
- Incorporate clinical text with snapshot data.
- Measure performance on mortality prediction in-hospital, at 30-days and 1-year post-discharge.

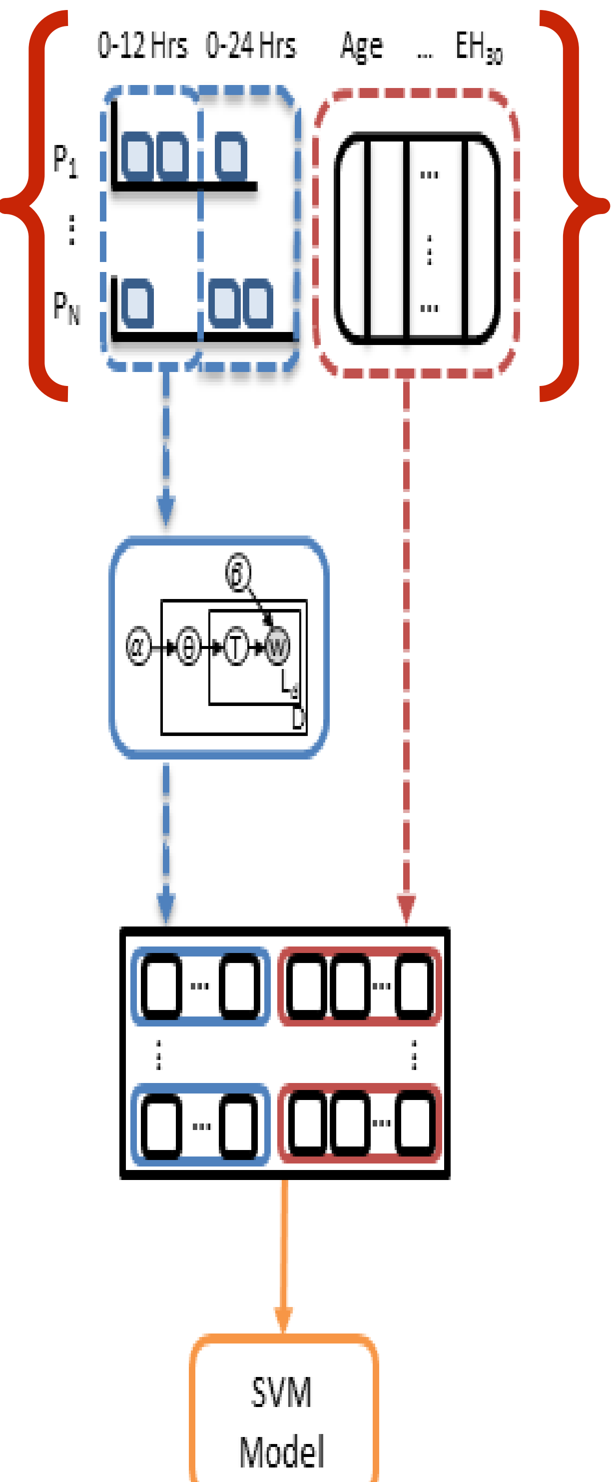


# Model Setup: Overview

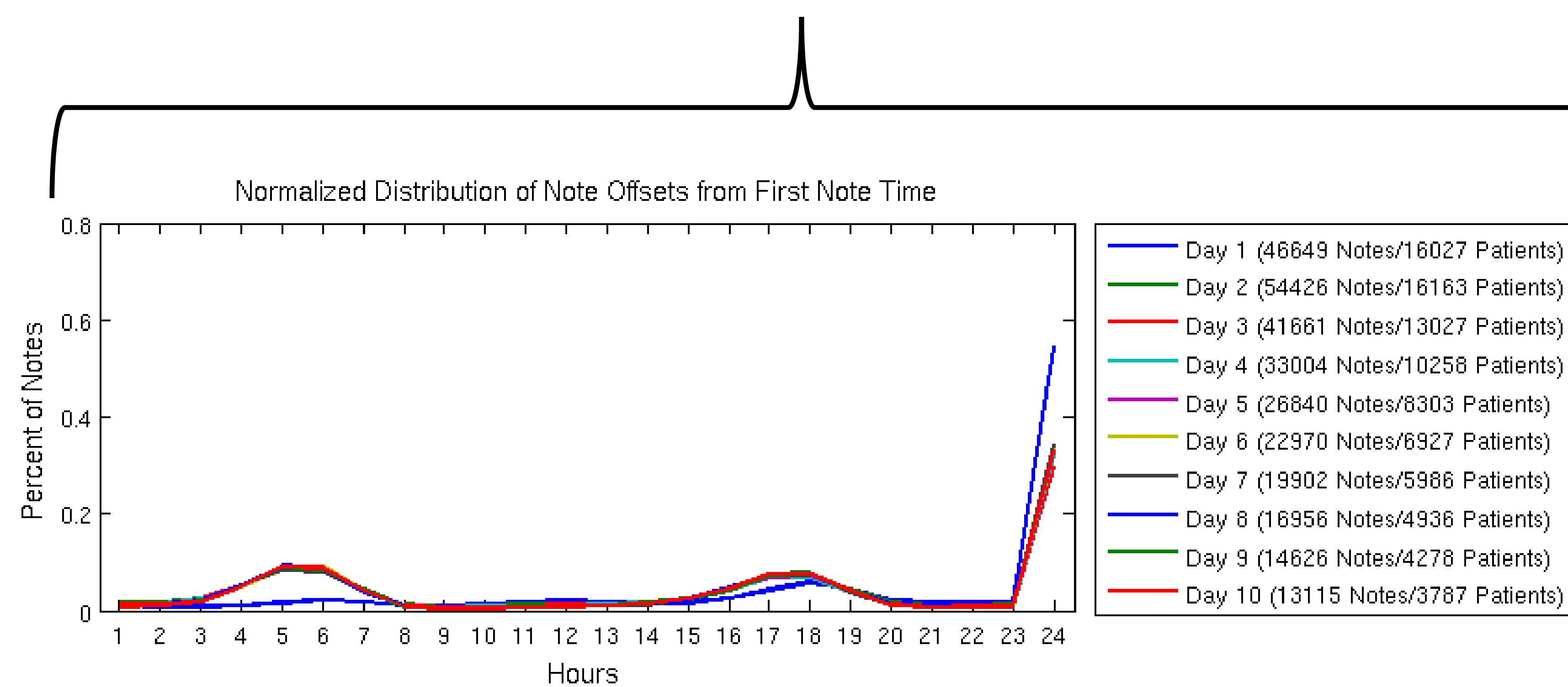


KDD 2014

# Model Setup: Data

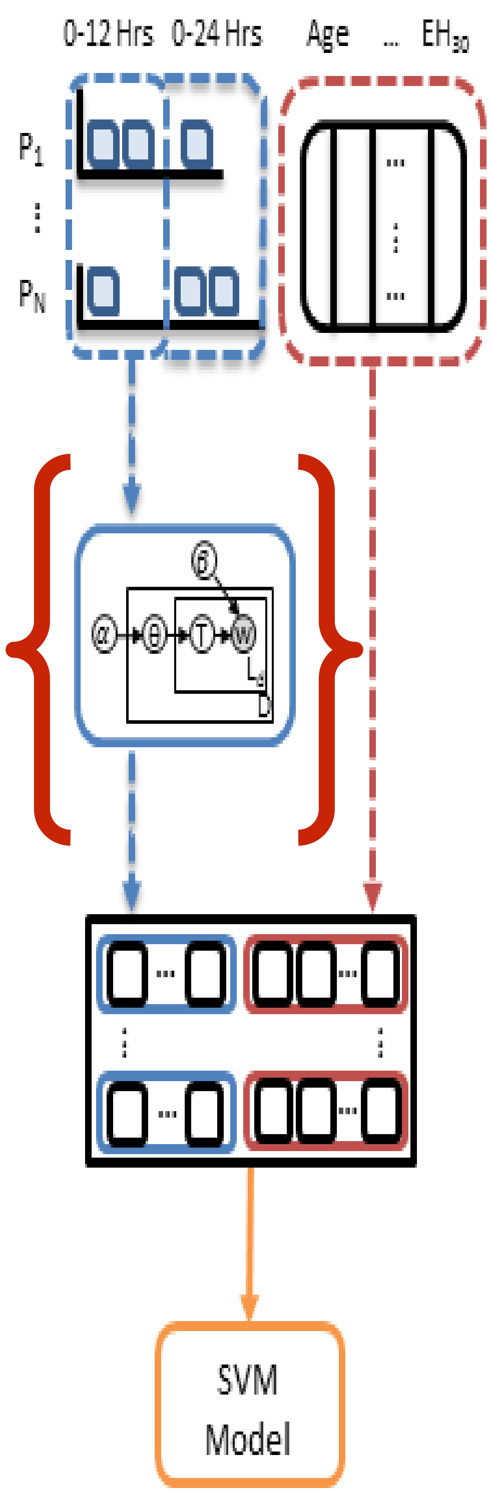


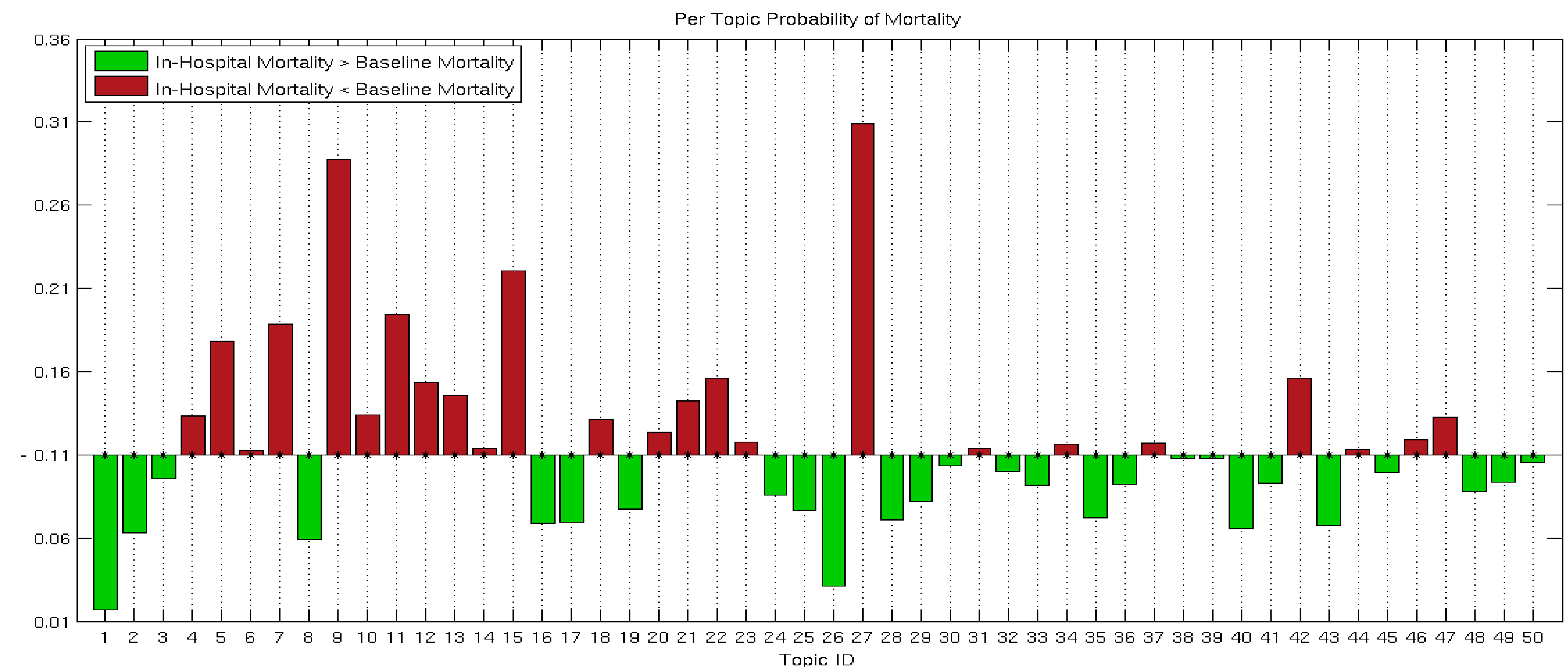

- Use 19,308 adult patient records
- Gather per-patient snapshot information
- Collect 473,764 notes
  - Use only first admissions
  - Ignore discharge summaries



KDD 2014

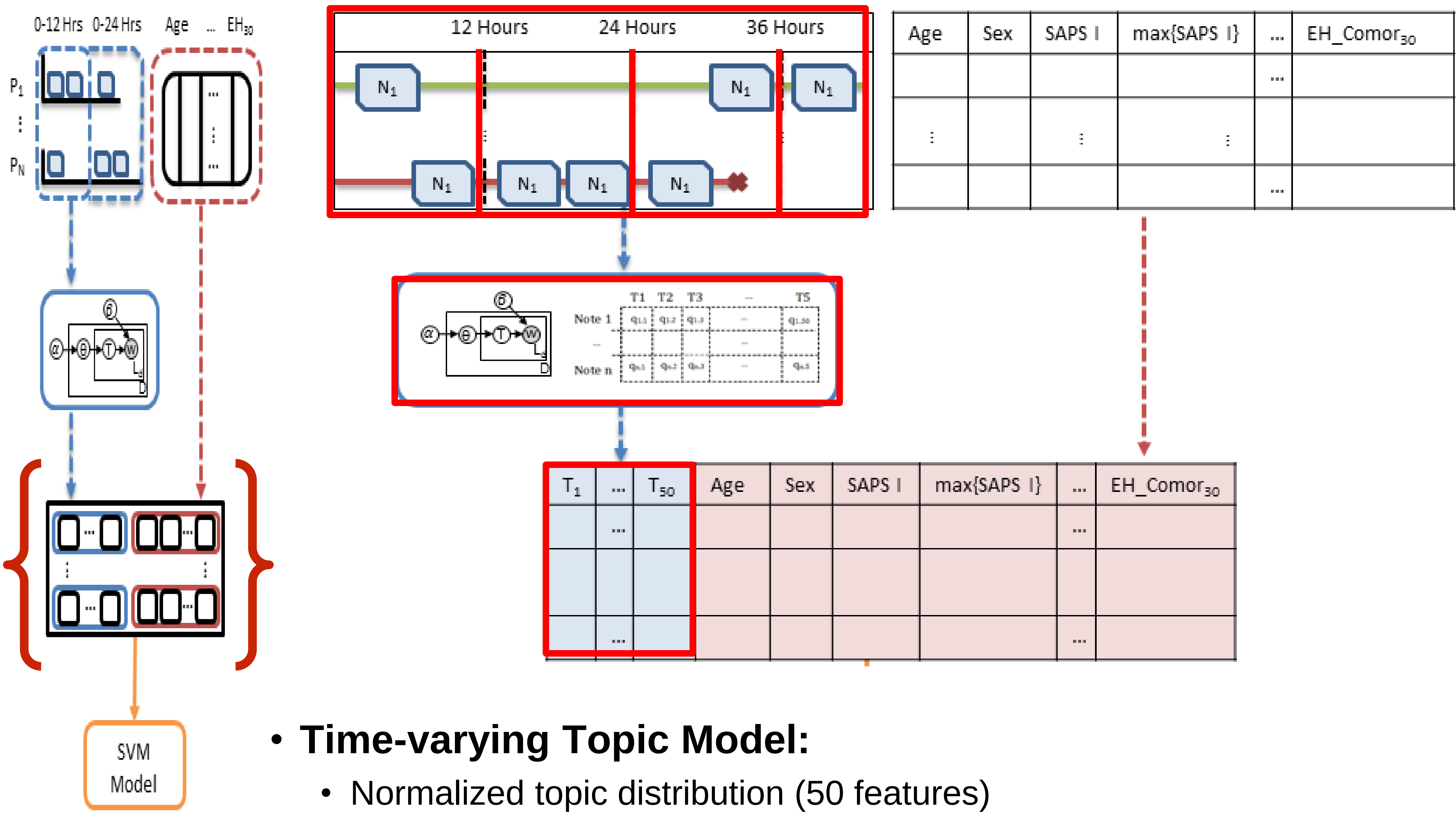
# Model Setup: Latent Topic Features



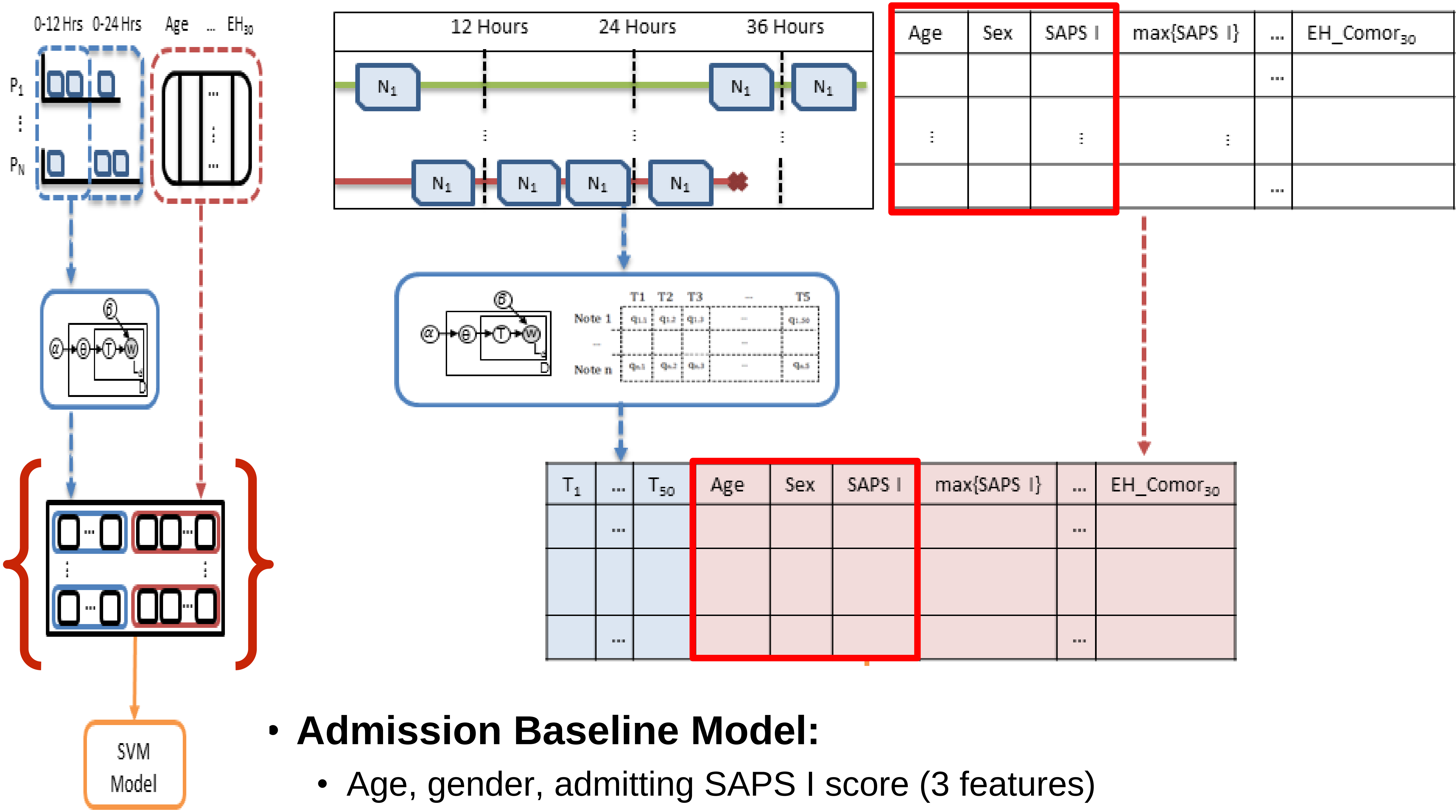

| | Topic # | Top Ten Words | Possible Topic |
|---|---|---|---|
| **In-Hospital Mortality** | 27 | name family neuro care noted status plan stitle dr remains | Discussion of end-of-life care |
| | 15 | intubated vent ett secretions propofol abg respiratory resp care sedated | Respiratory failure |
| | 7 | thick secretions vent trach resp tf tube coarse cont suctioned | Respiratory infection |
| | 5 | liver renal hepatic ascites dialysis failure flow transplant portal ultrasound | Renal failure |
| **Hospital Survival** | 1 | cabg pain ct artery coronary valve post wires chest sp neo | Cardiovascular Surgery |
| | 40 | left fracture ap views reason clip hip distal lat report joint | Fracture |
| | 16 | gtt insulin bs lasix endo monitor mg am plan iv | Chronic diabetes |
| **1 Year Mortality** | 3 | picc line name procedure catheter vein tip placement clip access | PICC line insertion |
| | 4 | biliary mass duct metastatic bile cancer left ca tumor clip | Cancer treatment |
| | 45 | catheter name procedure contrast wire french placed needle advanced clip | Coronary catheterization |

# Model Setup: Time-varying Topics




- **Time-varying Topic Model:**
  - Normalized topic distribution (50 features)

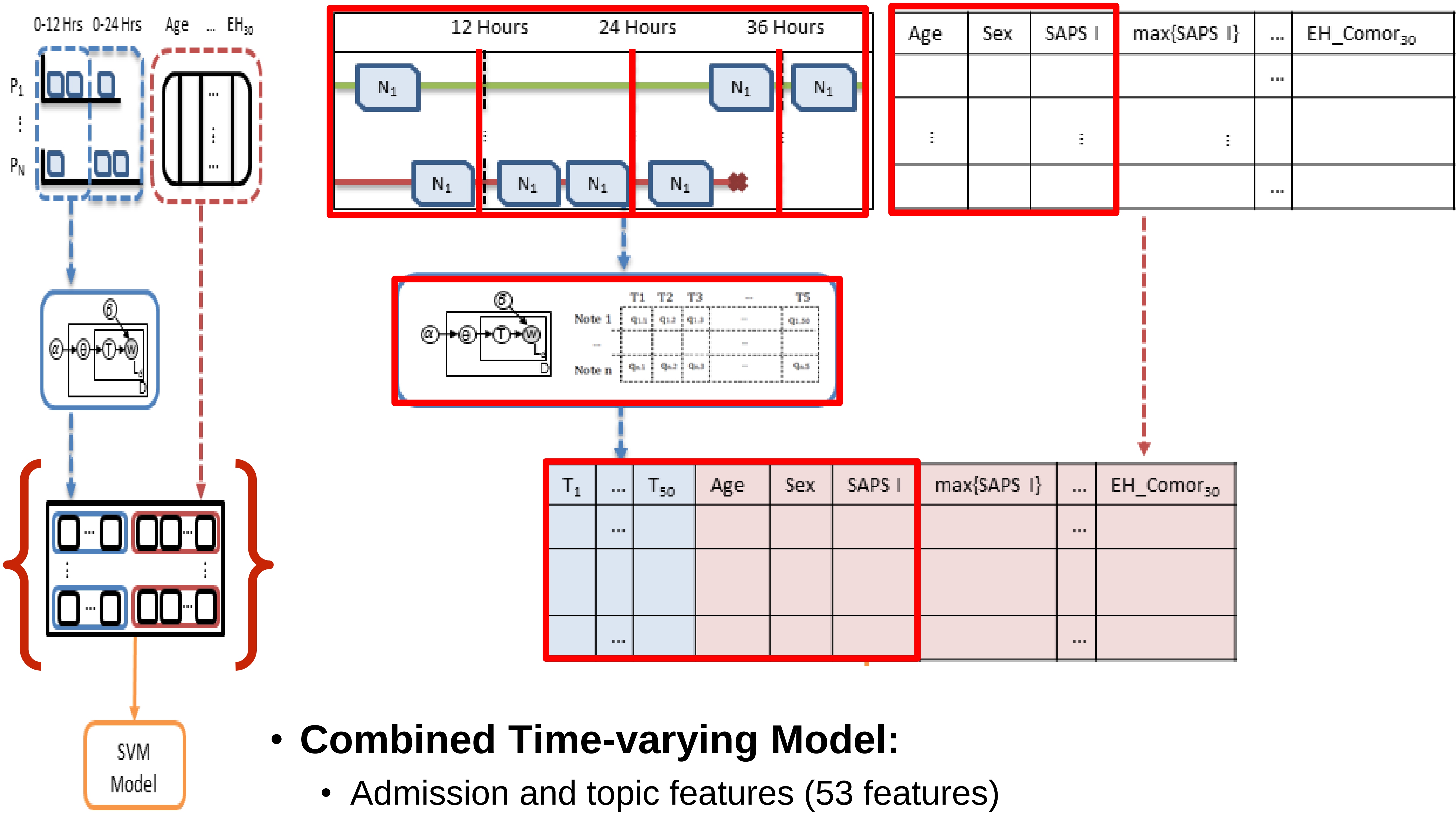
# Model Setup: Admission Baseline




- **Admission Baseline Model:**
  - Age, gender, admitting SAPS I score (3 features)

KDD 2014

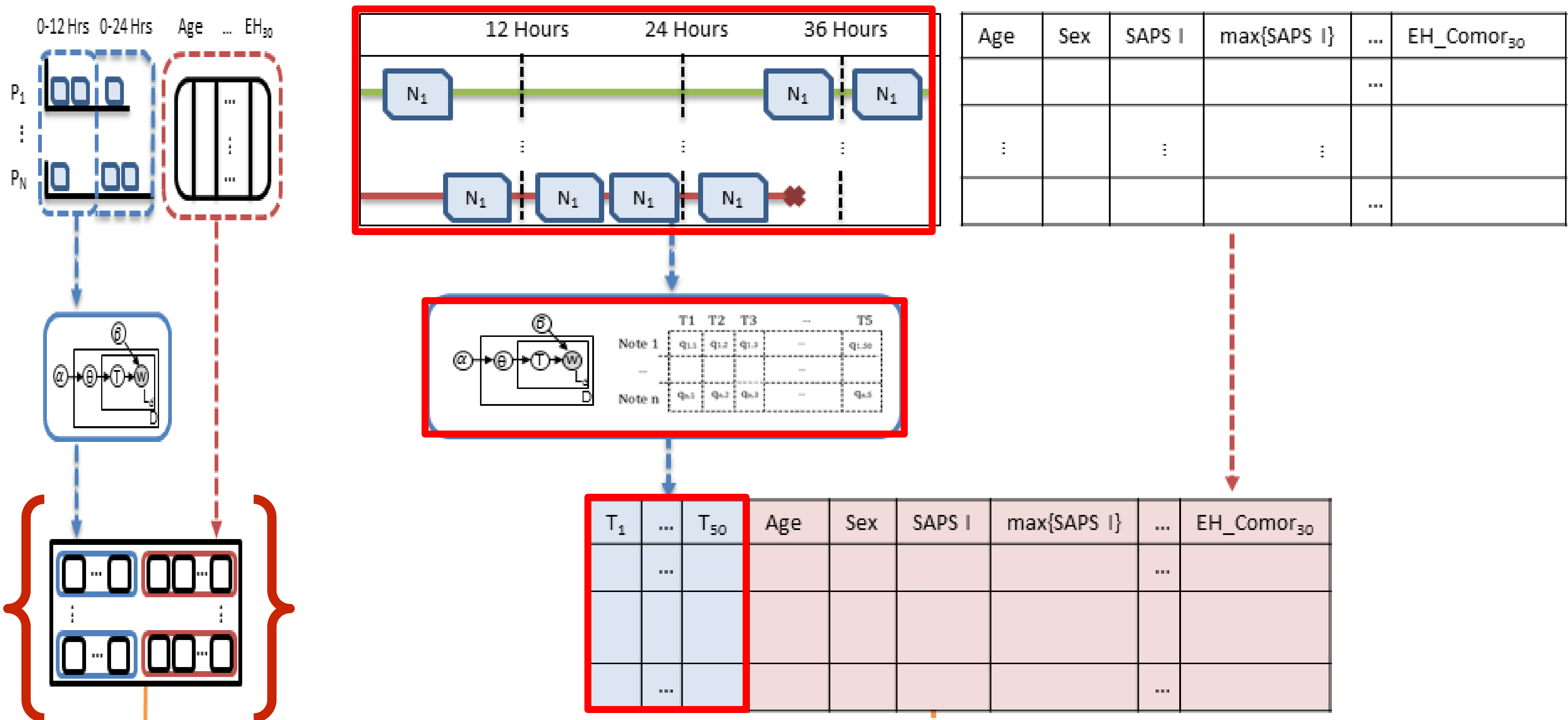
# Model Setup: Combined Time-varying

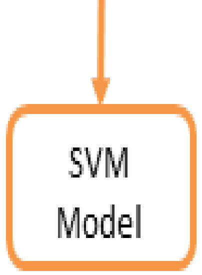



- **Combined Time-varying Model:**
  - Admission and topic features (53 features)

CSAIL

# Model Setup: Retrospective Topics

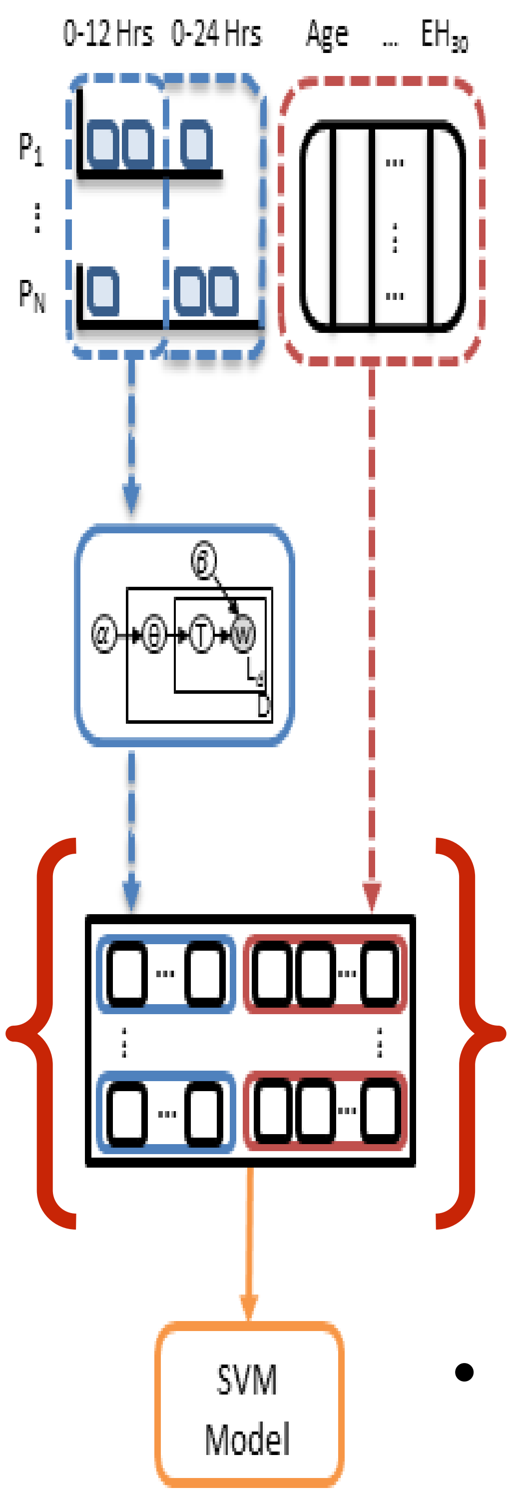


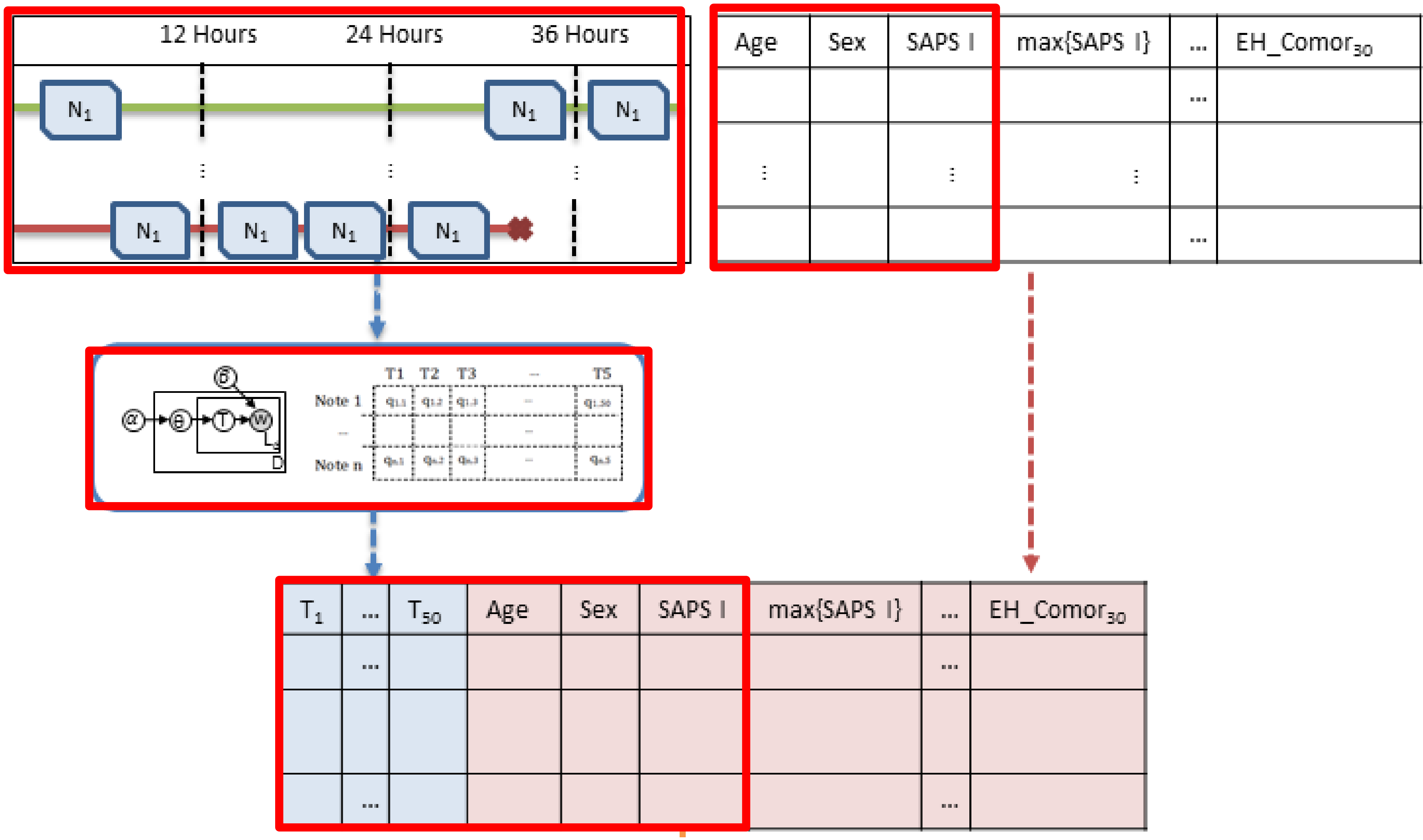




- **Retrospective Topic Model:**
  - Retrospective note features from entire patient stay (50 features).

KDD 2014

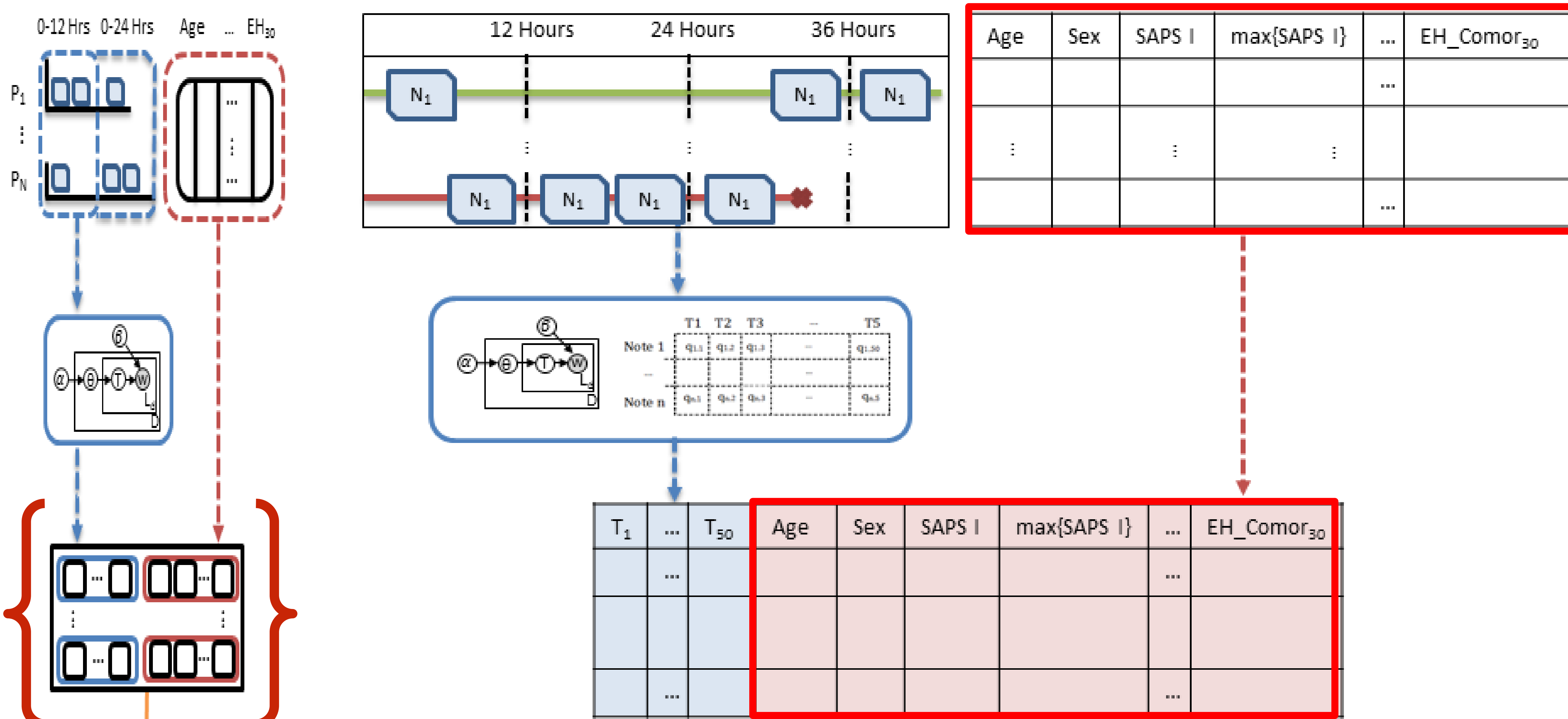
# Model Setup: Retrospective Topics + Admission

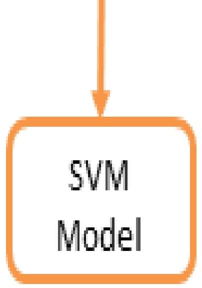



- **Retrospective Topic + Admission Model:**
  - Combined topic and admission feature (53 features).

# Model Setup: Retrospective Derived

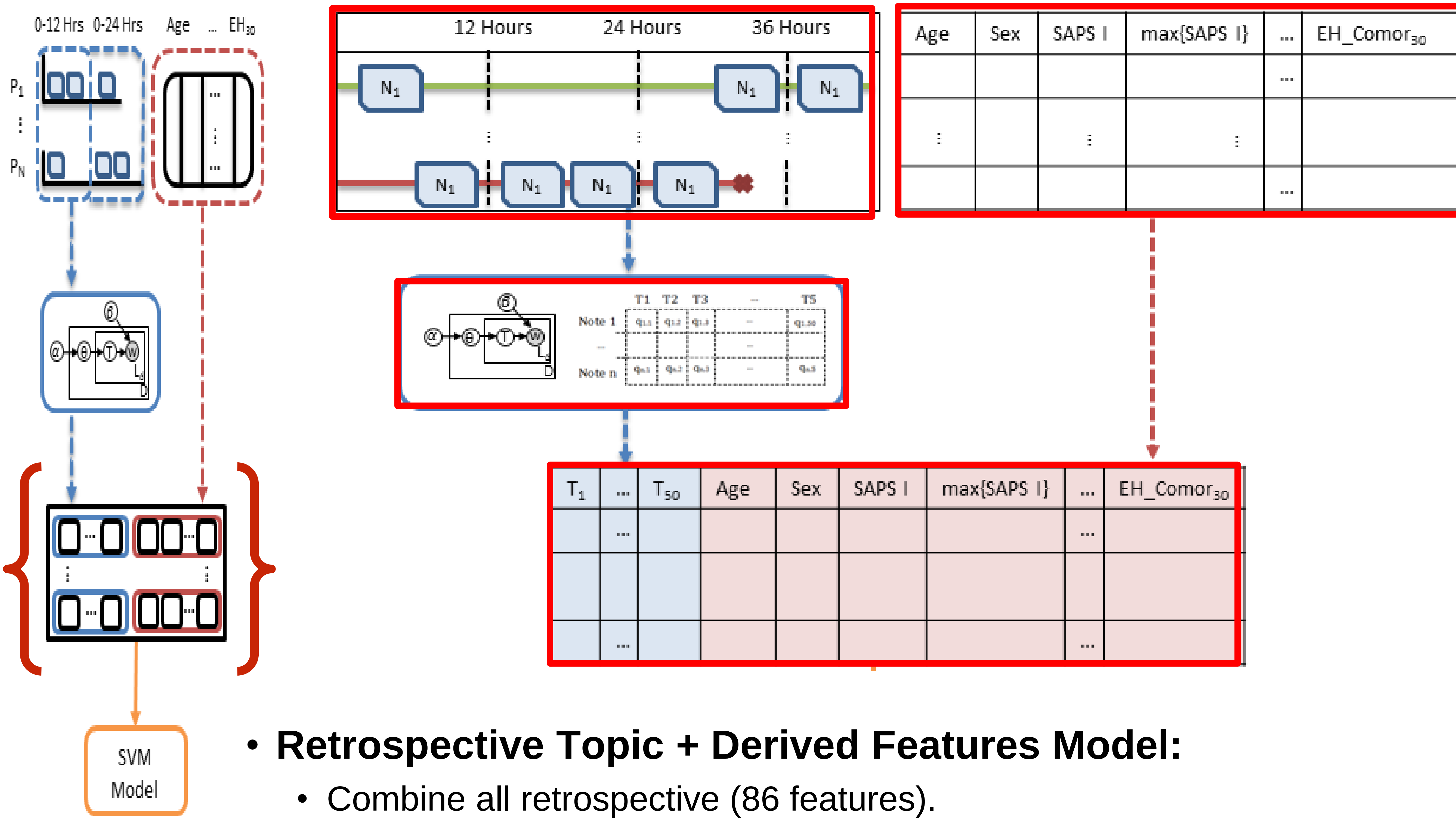



- **Retrospective Derived Features Model:**
  - Age, gender, admitting/min/max/final{SAPS I} and Elixhauser co-morbidity scores (36 features).

KDD 2014

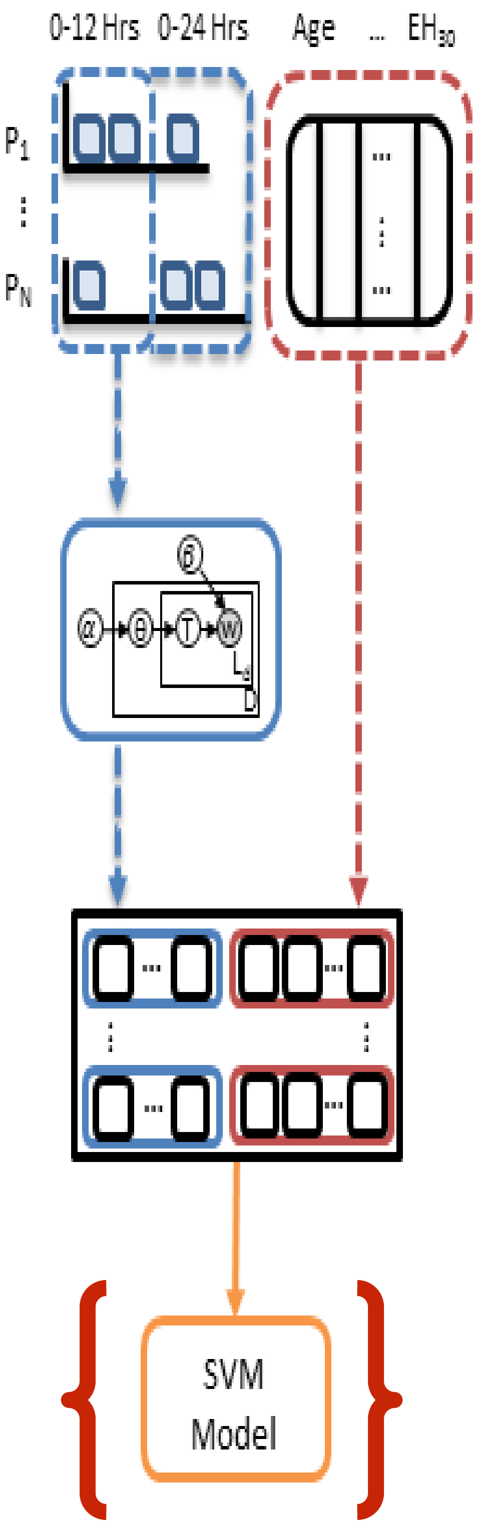
# Model Setup: Retrospective Topics + Derived

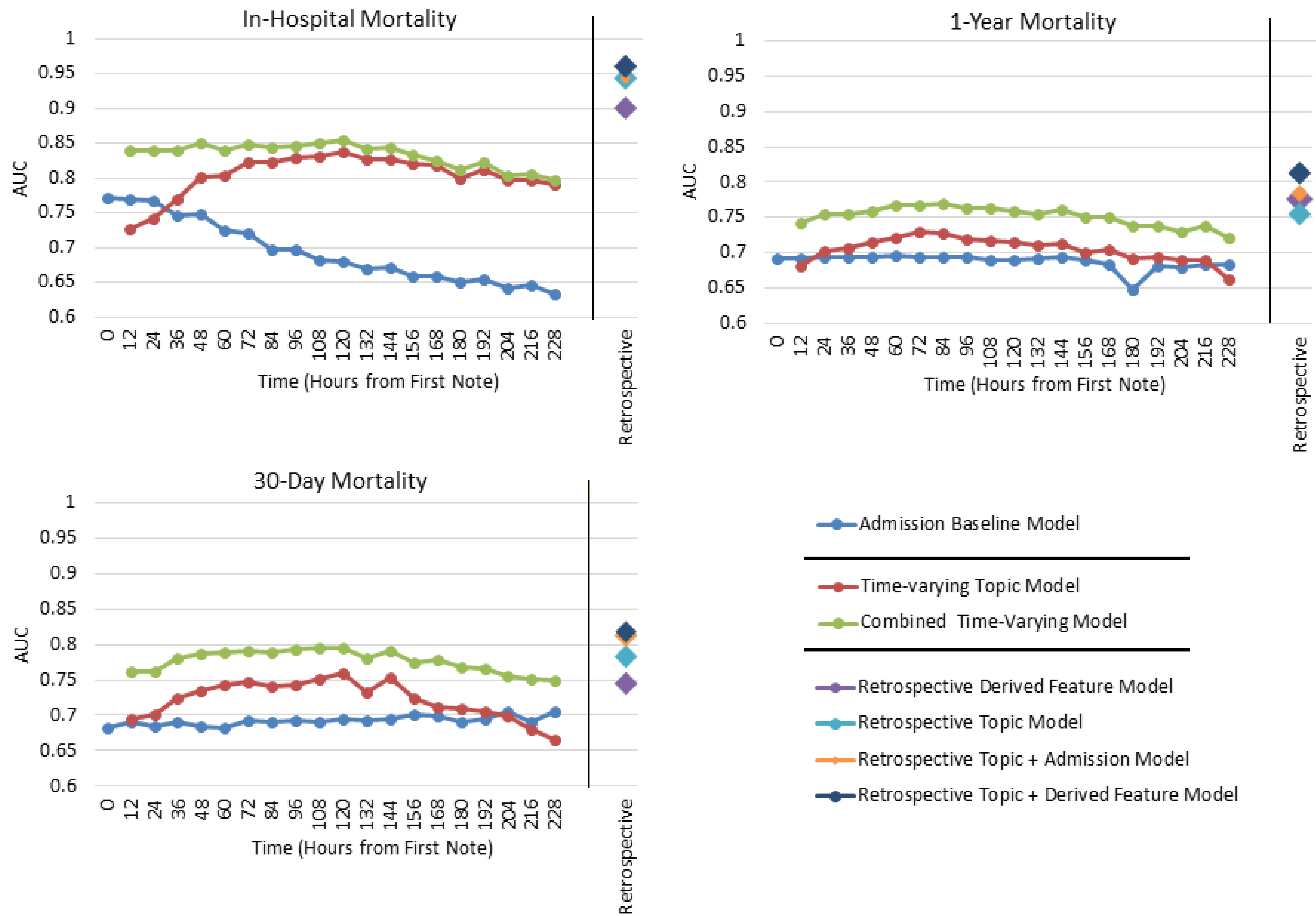



- **Retrospective Topic + Derived Features Model:**
  - Combine all retrospective (86 features).

KDD 2014

# Mortality Prediction Results



# We Solved A Problem, Everything Is Awesome

- Text Data Is Valuable
  - A combination of latent topic features and snapshot features worked best
- Long-term Predictions Are Harder
  - Combinations of features were best able to per
    over first 24 hours.
- “Realtime” Models Are More Valuable
  - Retrospective models out-performed continuou
    actionable.

# A Multivariate Timeseries Modeling Approach to Severity of Illness Assessment and Forecasting in ICU with Sparse, Heterogeneous Clinical Data

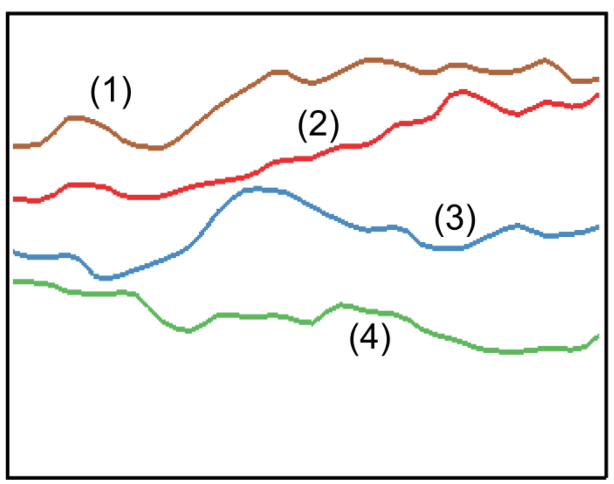
- AAAI 2015

- Marzyeh Ghassemi, Marco A. F. Pimentel, Tristan Naumann, Thomas Brennan, David A. Clifton, Peter Szolovits, Mengling Feng

AAAI 2015

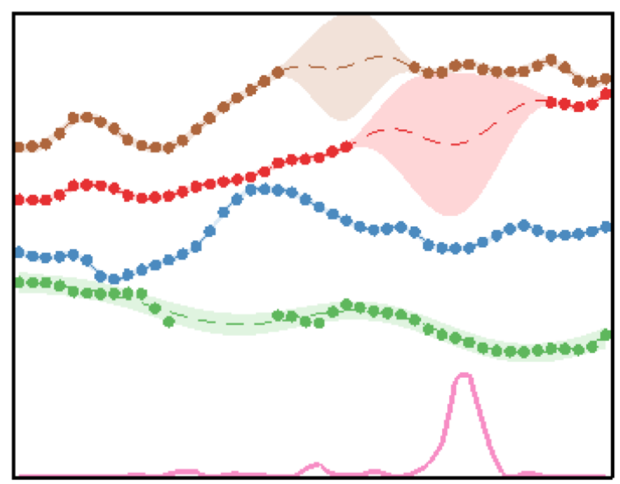
# Noisy, Sparse, Irregularly Sampled Data

- We use MTGPs to model the movements between and within multiple signals. This transforms a variety of irregularly-sampled clinical data into a new latent space using the MTGP hyperparameter



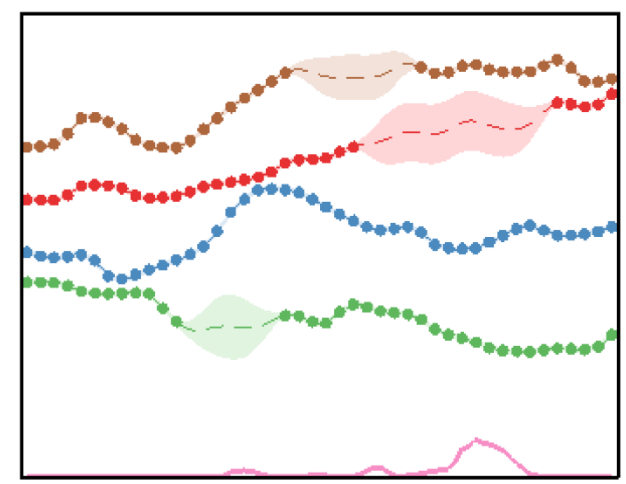

(a)

A sample function with 4 tasks. Tasks 1 and 2 were correlated; 4 was anti-correlated with 1 and 2; and 3 was uncorrelated.



(b)

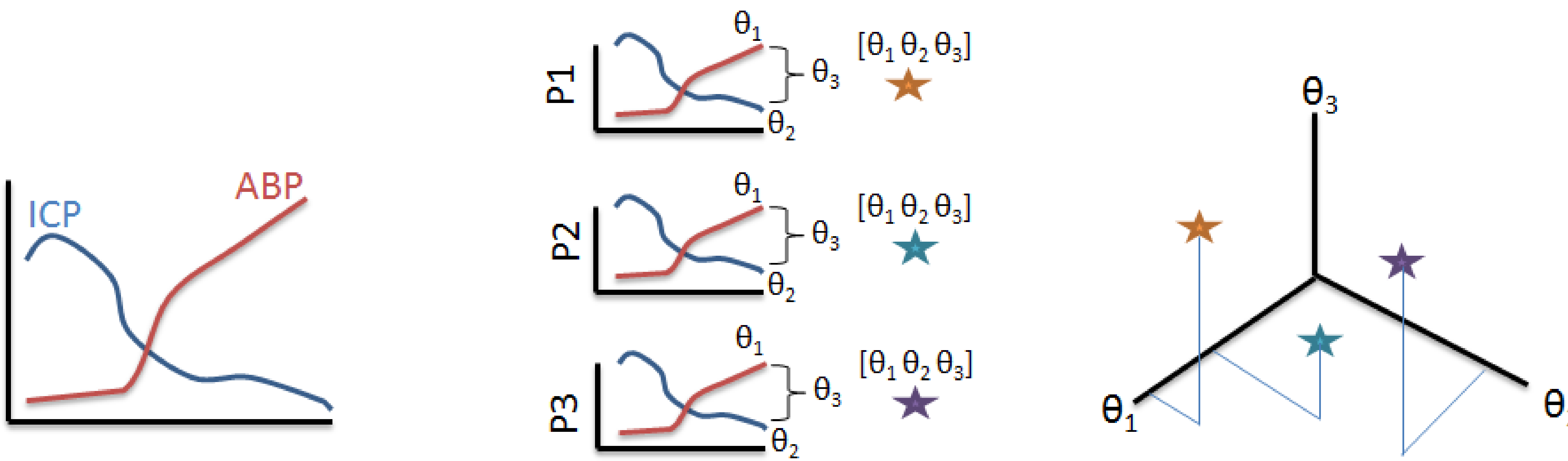
STGP predictions on all tasks. Mean absolute prediction error (over the 4 tasks) doesn't use signal interaction.



(c)

MTGP predictions on all tasks. Predictions improved by taking into account the correlation between the different tasks

# Projection to Latent Space




**Clinical Signals** → **MTGP Hyperparameters** → **Latent Space**

MTGP models capture the movement within and between signals.

We transform a variety of irregularly-sampled signals into a new latent space using MTGP hyperparameters.

Patients are compared based on their similarity in the new latent space.

θ provides a new latent search space to examine and evaluate the similarity of any two given multi-dimensional functions.

AAAI 2015

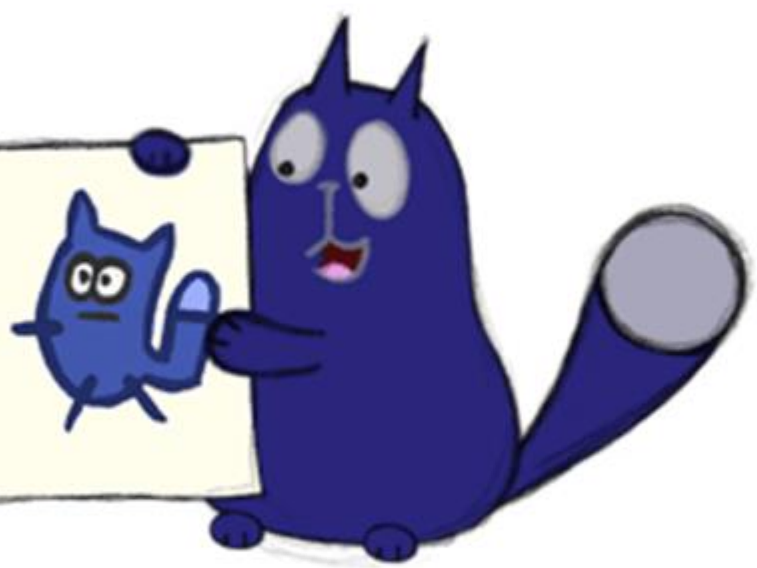
# But What Are These Hyperparameters?

- We use the squared exponential kernel, so each signal's θ govern the function input/output scale, and the inter-signal θ correspond to the correlation between the different outputs.

- These parameters are:
  1. a means of representing the functional behavior
  2. a set of observations learned directly from data; and
  3. generalizable to any type of longitudinal data, including categorical and numerical types

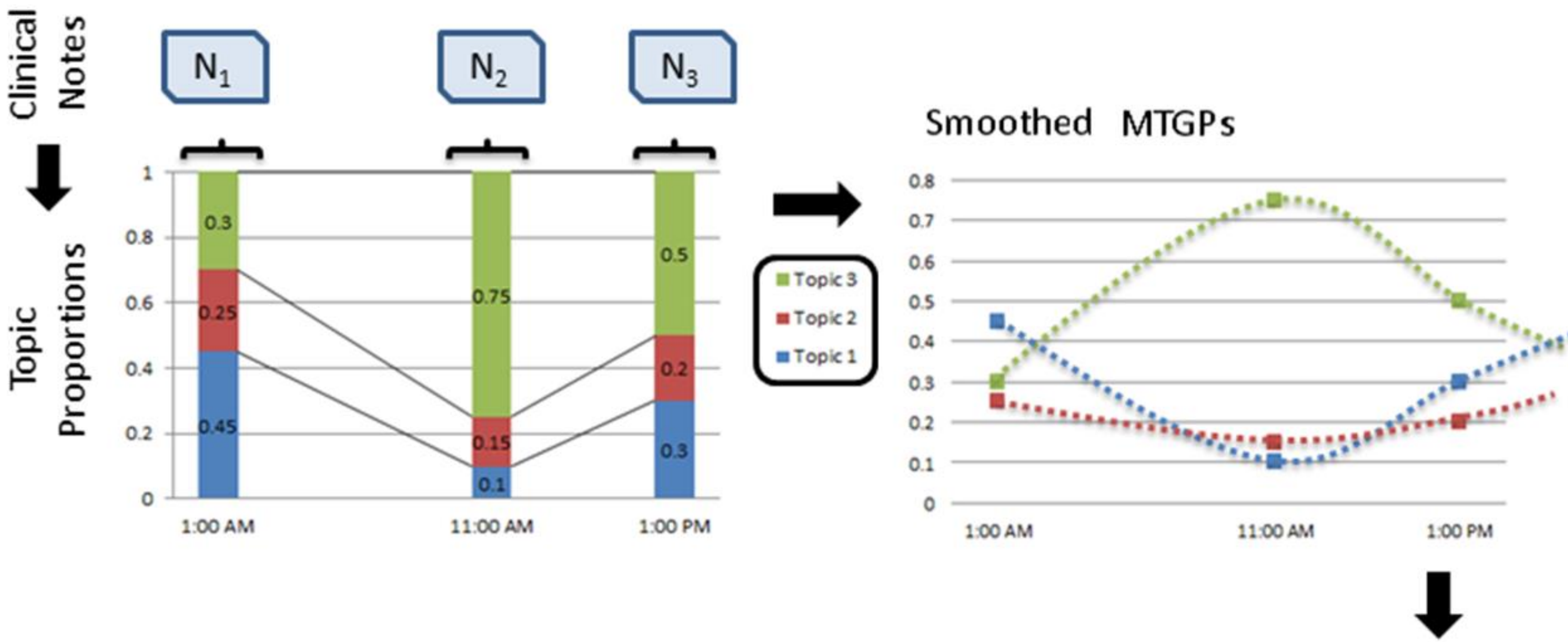
$$\mathbf{K}_{MT}(\mathbf{X}_n, \mathbf{l}, \boldsymbol{\theta}_c, \boldsymbol{\theta}_t) = \mathbf{K}_c(\mathbf{l}, \boldsymbol{\theta}_c) \otimes \mathbf{K}_t(\mathbf{X}_n, \boldsymbol{\theta}_t)$$

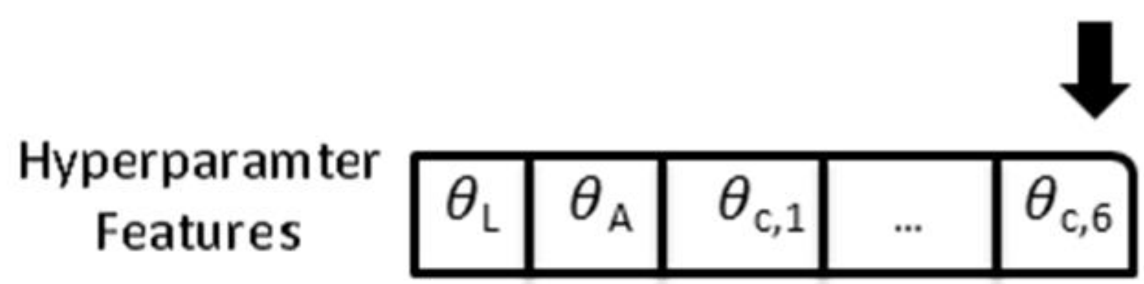
$$\mathbf{K}_t = \theta_A^2 \exp\left\{-\frac{\| x - x' \|^2}{2\theta_L^2}\right\} \qquad \mathbf{K}_c = \mathbf{L}\mathbf{L}^\top, \quad \mathbf{L} = \begin{bmatrix} \theta_{c,1} & 0 & \cdots & 0 \\ \theta_{c,2} & \theta_{c,3} & & 0 \\ \vdots & & \ddots & \vdots \\ \theta_{c,k-m+2} & \theta_{c,k-m+2} & \cdots & \theta_{c,k} \end{bmatrix}$$



CSAIL

# Incorporating Text Data




- Perform a “pre-projection step”
- Clinical notes are transformed into timeseries using LDA
- New set of topic proportion timeseries are fitted using the MTGPs
- Inferred hyperparameters θ are derived, projecting into the new latent space.

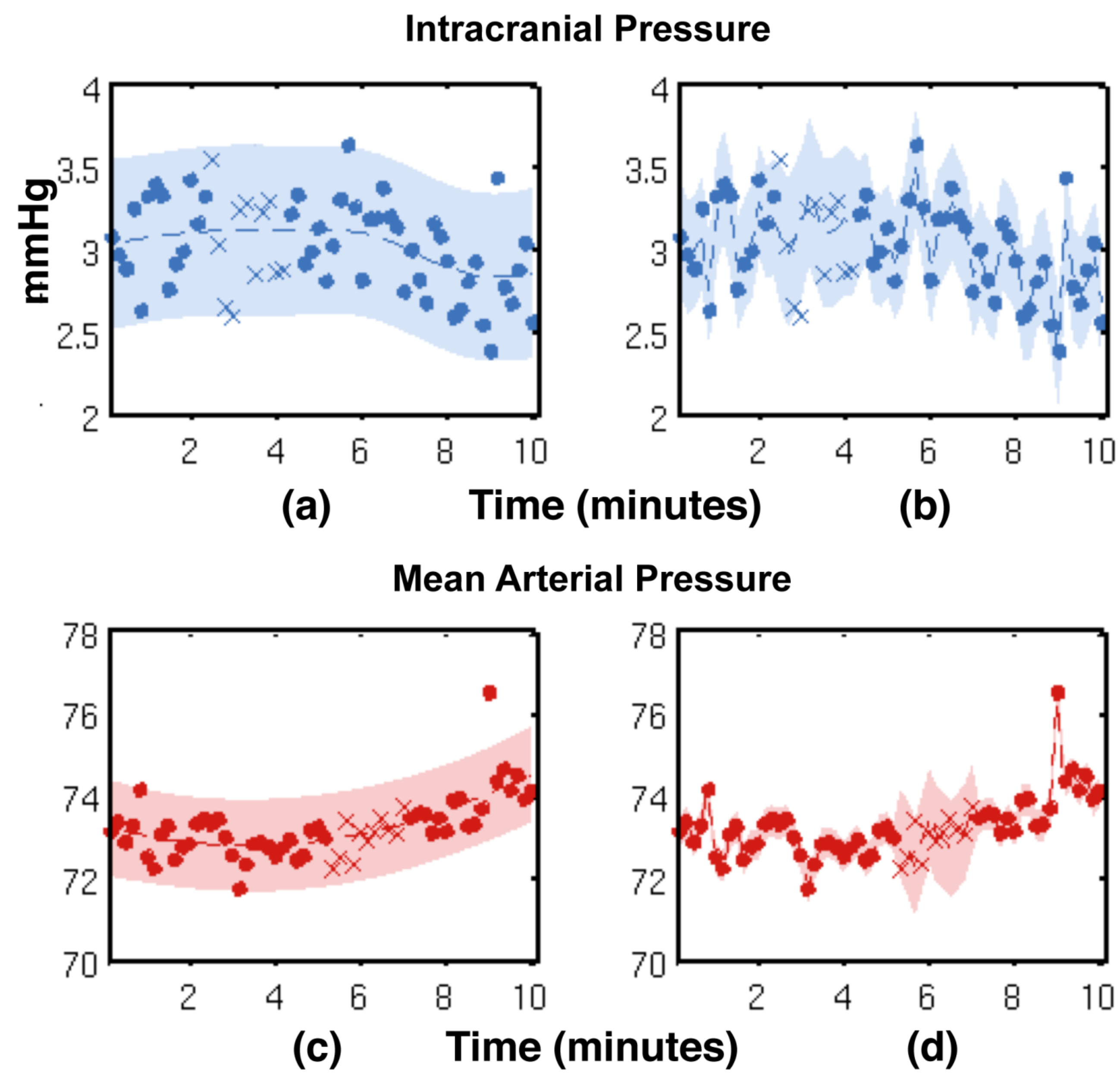
# Case Studies

- Estimating Signal in Traumatic Brain Injury Patients
- Mortality Prediction Using Clinical Progress Notes

# Estimating Signal in Traumatic Brain Injury

- ICP and ABP data were collected from 35 TBI patients who were monitored for more than 24-hours in a Neuro-ICU.
- Our goal was to forecast the MAP and ICP signals as well as estimate cerebrovascular pressure reactivity (PRx)

| Signal | Measure | STGP | MTGP |
|---|---|---|---|
| ICP | RMSE | 0.91 | 0.69 |
| | MSLL | 0.6 | 0.45 |
| ABP | RMSE | 2.77 | 1.98 |
| | MSLL | 0.65 | 0.55 |
| PRx-PRx* | RMSE | - | 0.09 |



AAAI 2015

# Mortality Prediction Using Clinical Notes

- 10,202 patients with 313,461 notes.
- Chose the 9 topics with a posterior likelihood above or below 5% of the population baseline likelihood across topics.
- Using MTGP hyperparameters as additional classification features also gave us improved results for mortality prediction (0.812 vs 0.788 AUC).

| | Top Five Words | Possible Topic |
|---|---|---|
| In-hospital Mortality | liver, renal, hepatic, ascites, dialysis | Renal Failure |
| | thick, secretions, vent, trach, resp | Respiratory infection |
| | remains, family, gtt, line, map | Systematic organ failure |
| | increased, temp, hr, pt, cc | Multiple physiological changes |
| | intubated, vent, ett, secretions, propofol | Respiratory failure |
| | name, family, neuro, care, noted | Discussion of end-of-life care |
| Survival | cabg, pain, ct, artery, coronary | Cardio-vascular surgery |
| | chest, pneumothorax, tube, reason, clip | |
| | pain, co, denies, oriented, neuro | Responsive patient |

| Features | Hospital Mortality | 1-Year Mortality |
|---|---|---|
| SAPS-I | 0.702 | 0.500 |
| Ave. Topics | 0.759 | 0.653 |
| SAPS-I + MTGP | 0.775 | 0.624 |
| Ave. Topics + MTGP | 0.788 | 0.673 |
| SAPS-I + Ave. Topics + MTGP | 0.812 | 0.686 |

# Acknowledgements

Thanks to:

Intel Science and Technology Center for Big Data

NIH NLM Biomedical Institute Research Training