Clinical Data:

Data Types and Challenges

What/Where Are the Data ?

EPIC:

- Electronic Health Care Record used Hopkins-wide
- Specialized to go fast & deep into a patient’s data

EPIC Clarity:

- EPIC data dumped to EPIC Clarity every 24 hours
- Center for Clinical Data Analysis (CDSA) does extracts

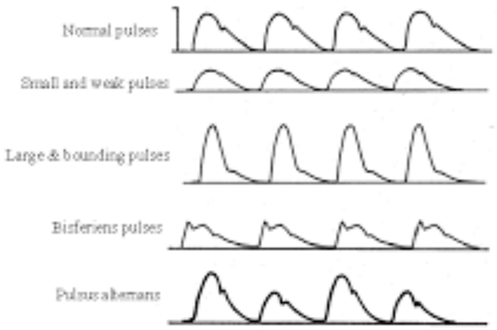
PhysioCloud:

- Data repository for physiological monitoring data
- Integrated into the Precision Medicine Analytic Platform

High-Frequency (HF) PTS Data

ECG

Pulse Oximetry



Respiratory Waveform

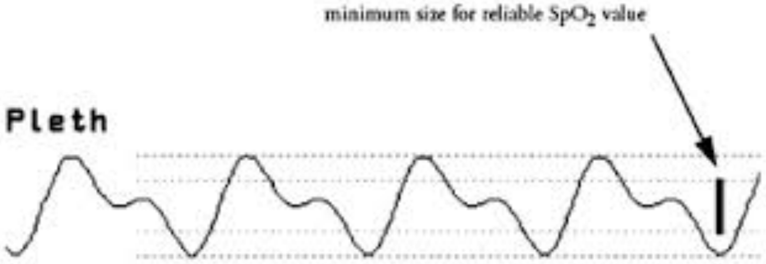
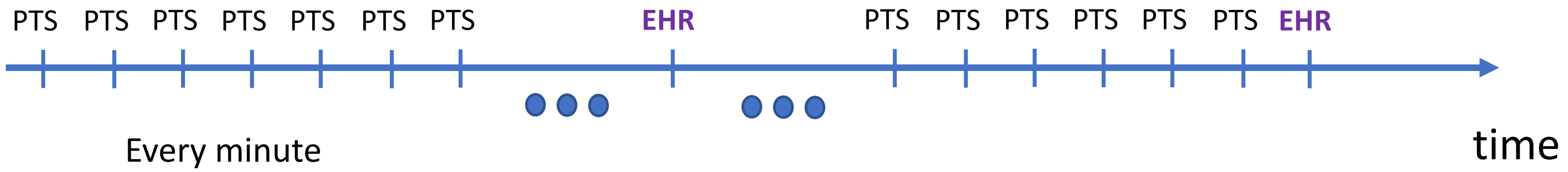


Figure 1 - The Pleth wave. The vertical bar indicates the minimum size for reliable peripheral oxygen saturation (SpO_2) value.

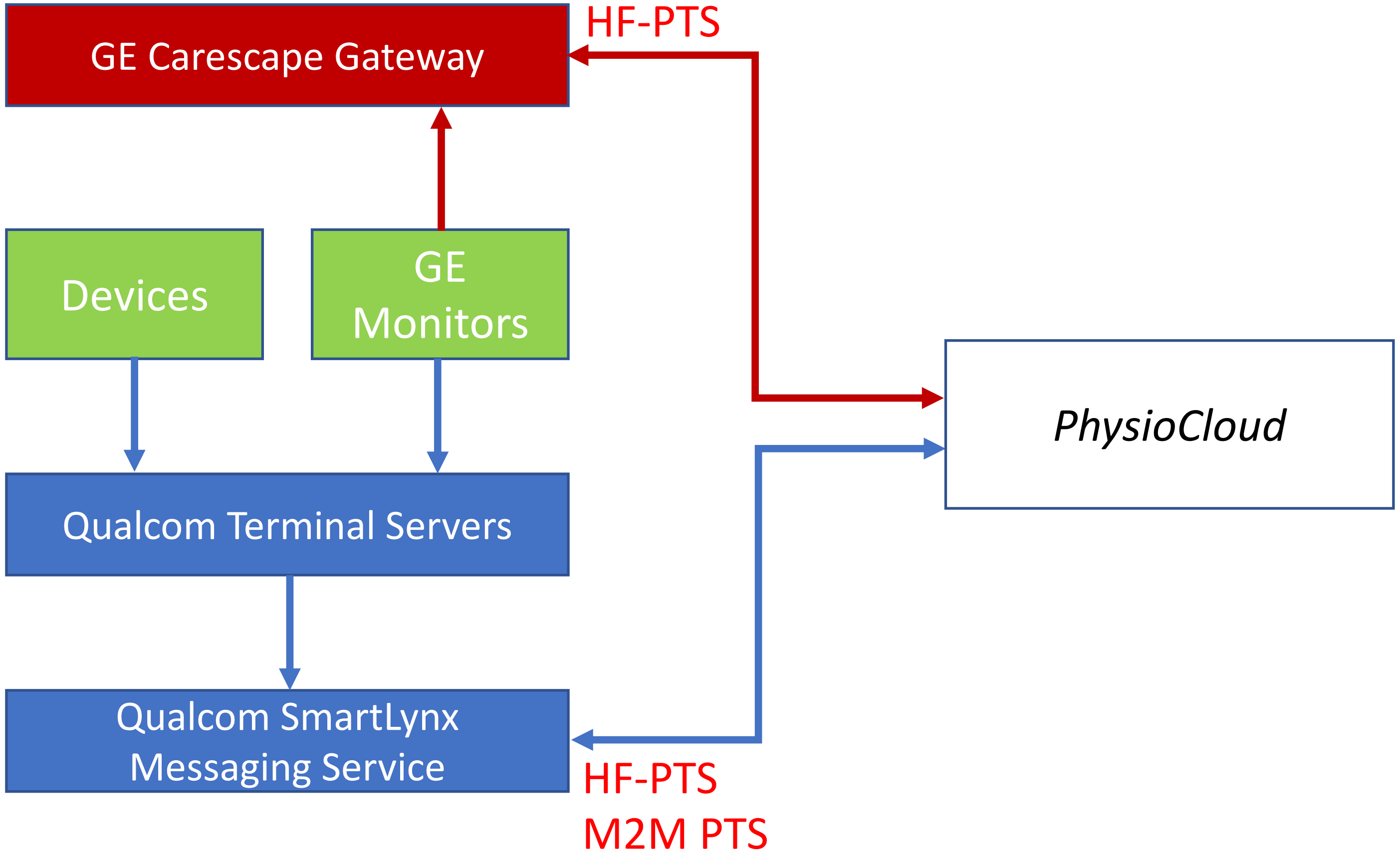
Minute to Minute (M2M) PTS Data

Average Heart Rate Over One Minute Intervals

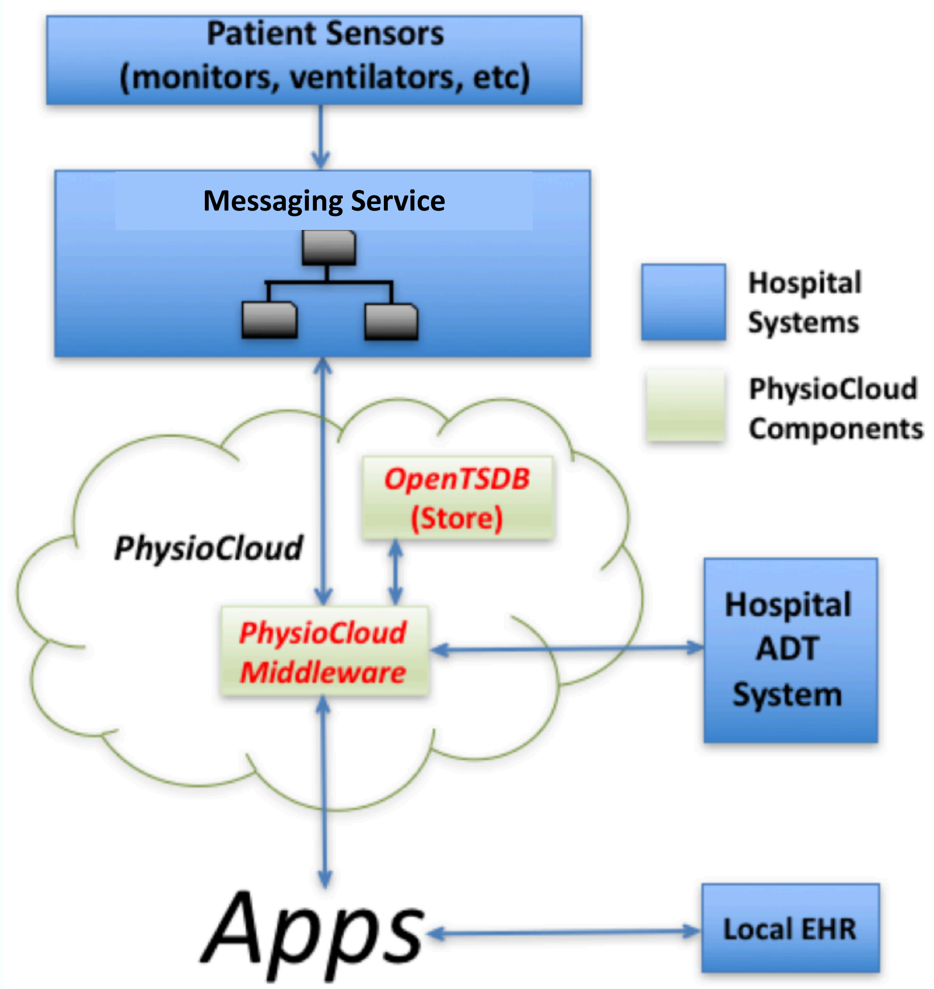


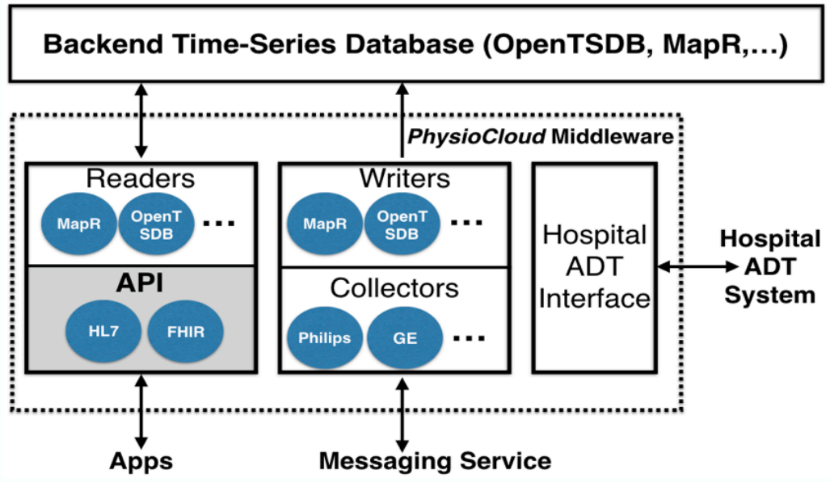
- Sampled waveforms (250 Hz), HF-PTS data
- M2M PTS data, “vital signs”
- Low-rate EHR data, irregular timing (hours)

PTS Data Flow

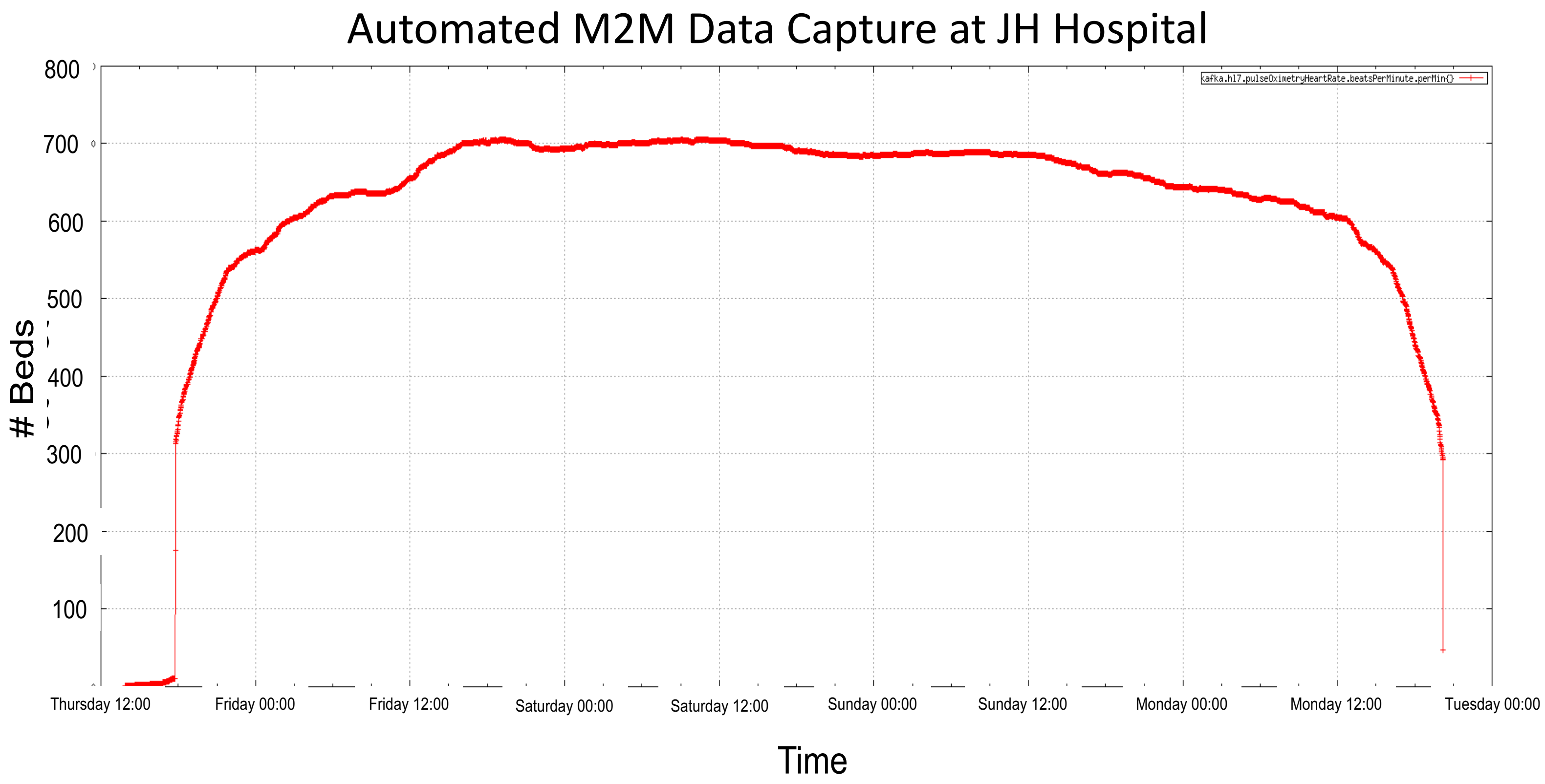


PTS Data Flow

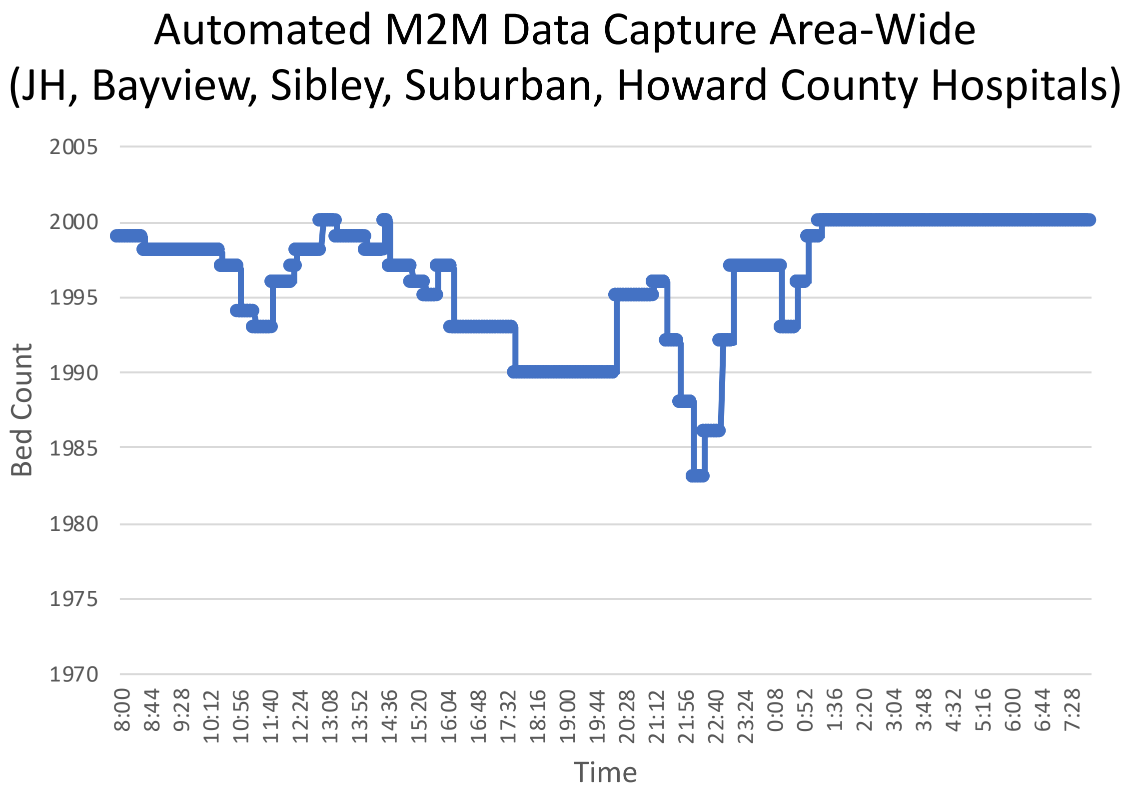




PhysioCloud in Operation

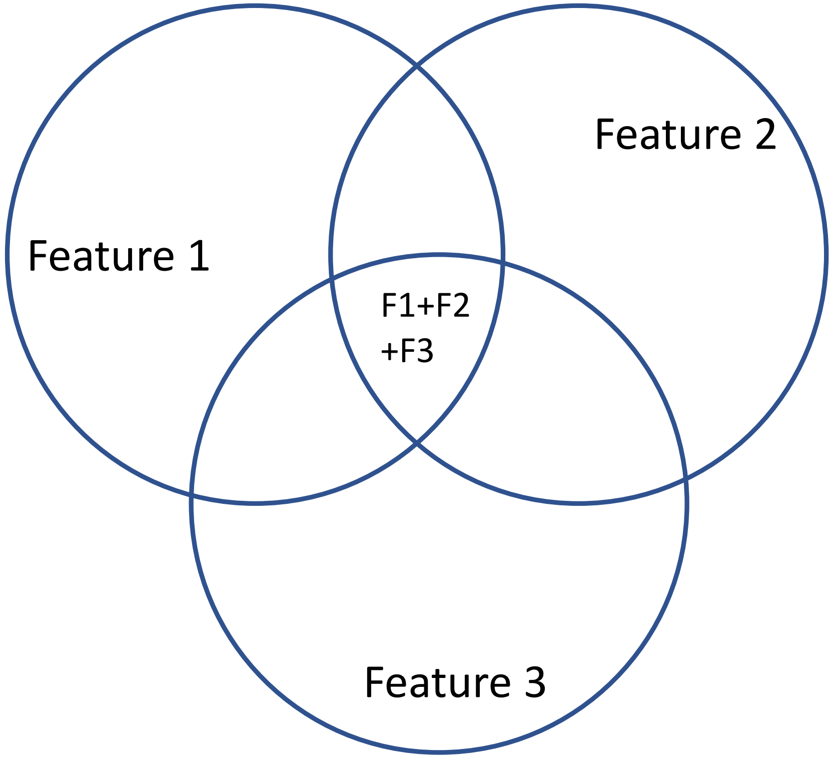


PhysioCloud in Operation

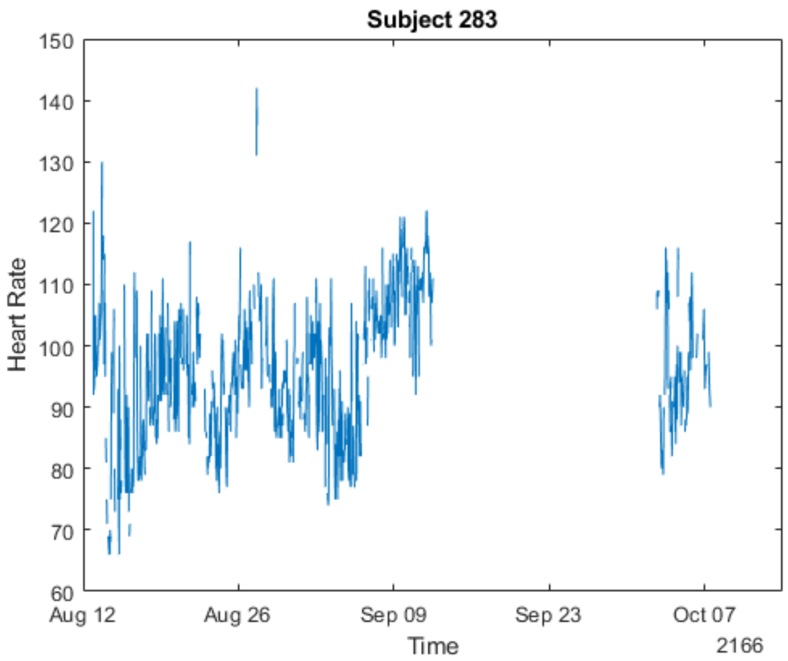


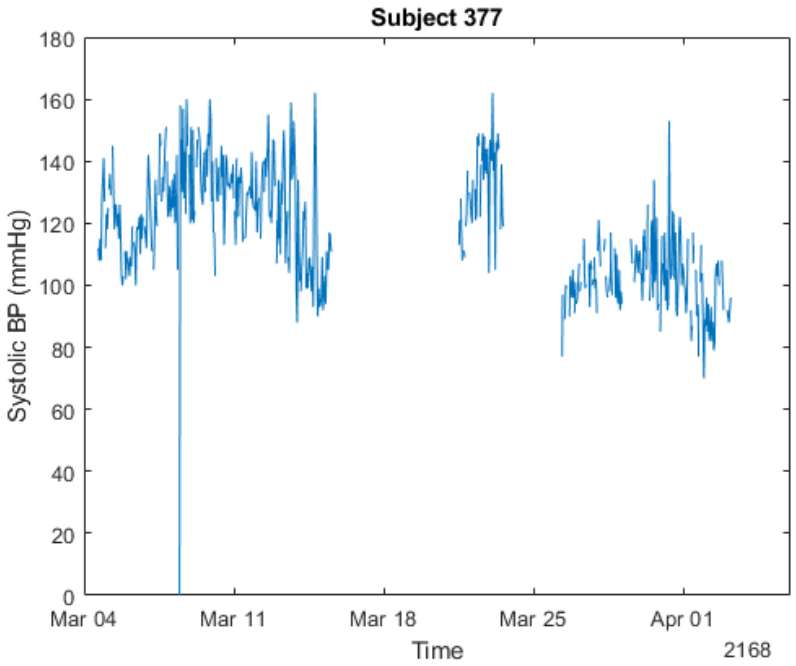
PhysioCloud is being integrated with the Precision Medicine Analytic Platform

The Missing Data Problem



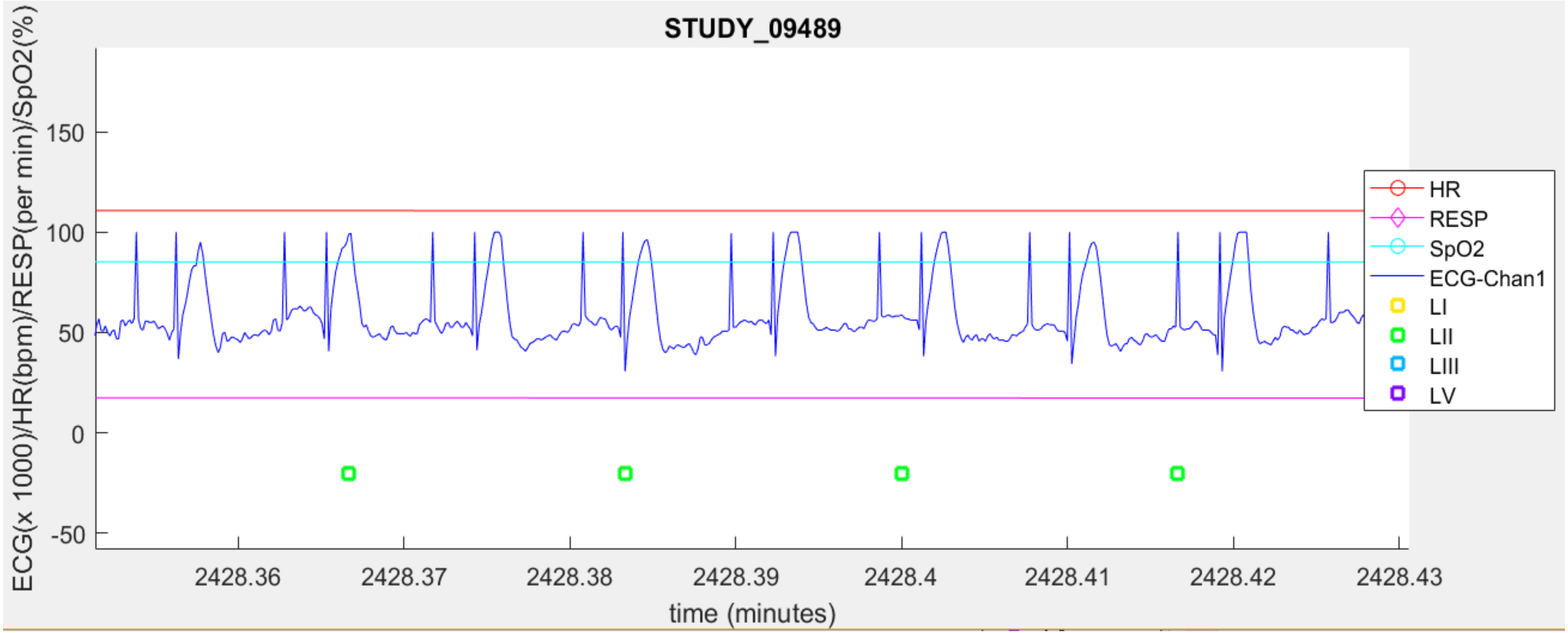
The Missing Data Problem –
Another variant





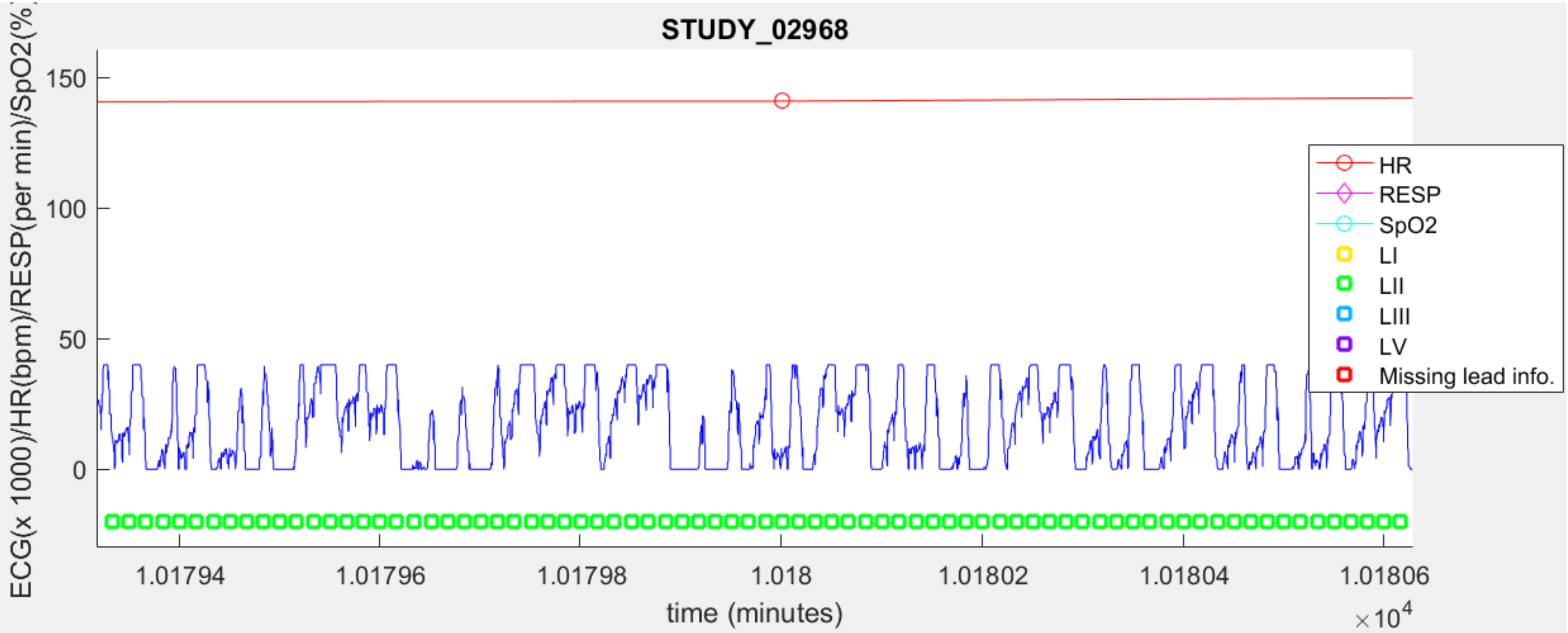
Data Artifacts:

Corrupting Signals



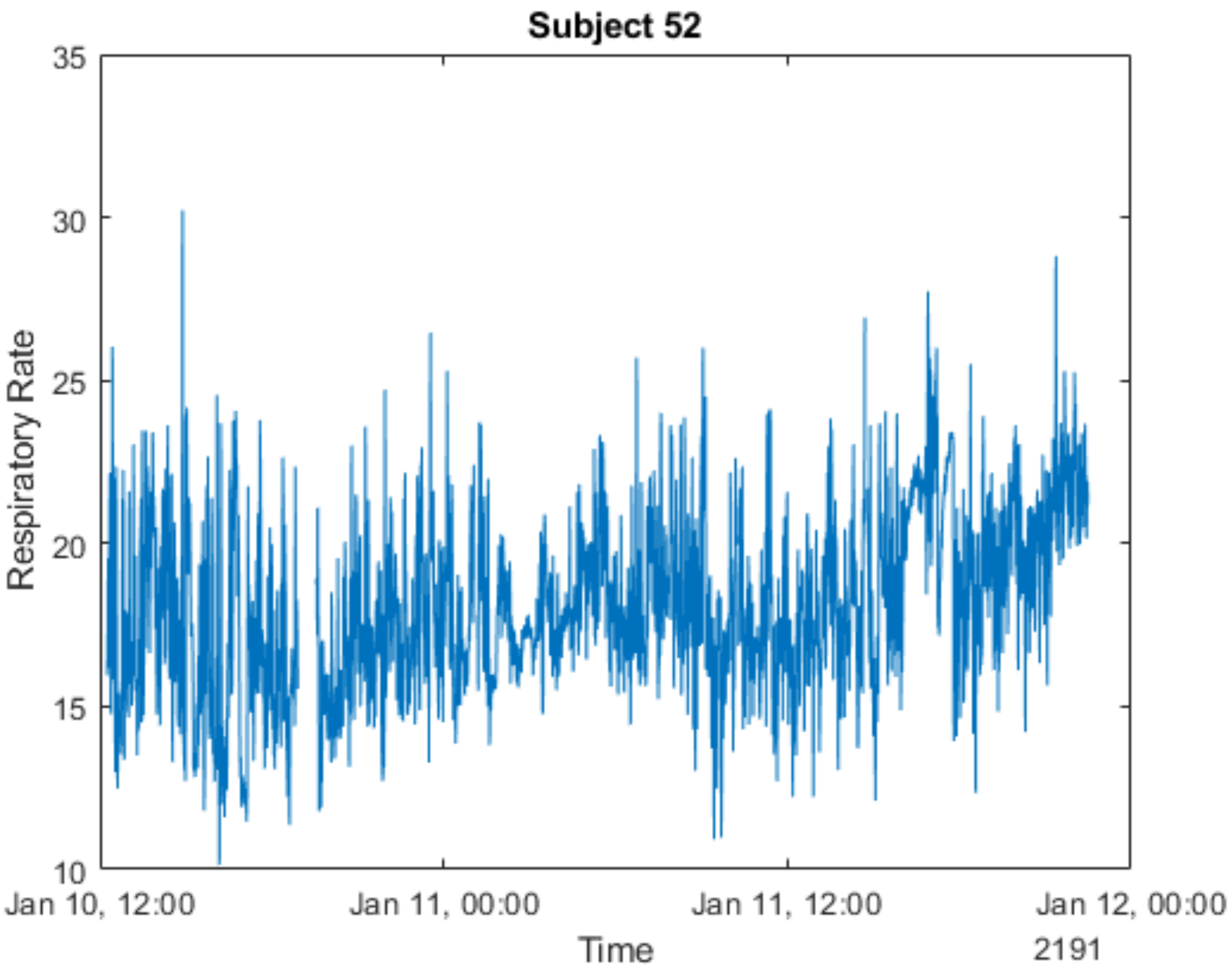
Data Artifacts:

Signal Saturation



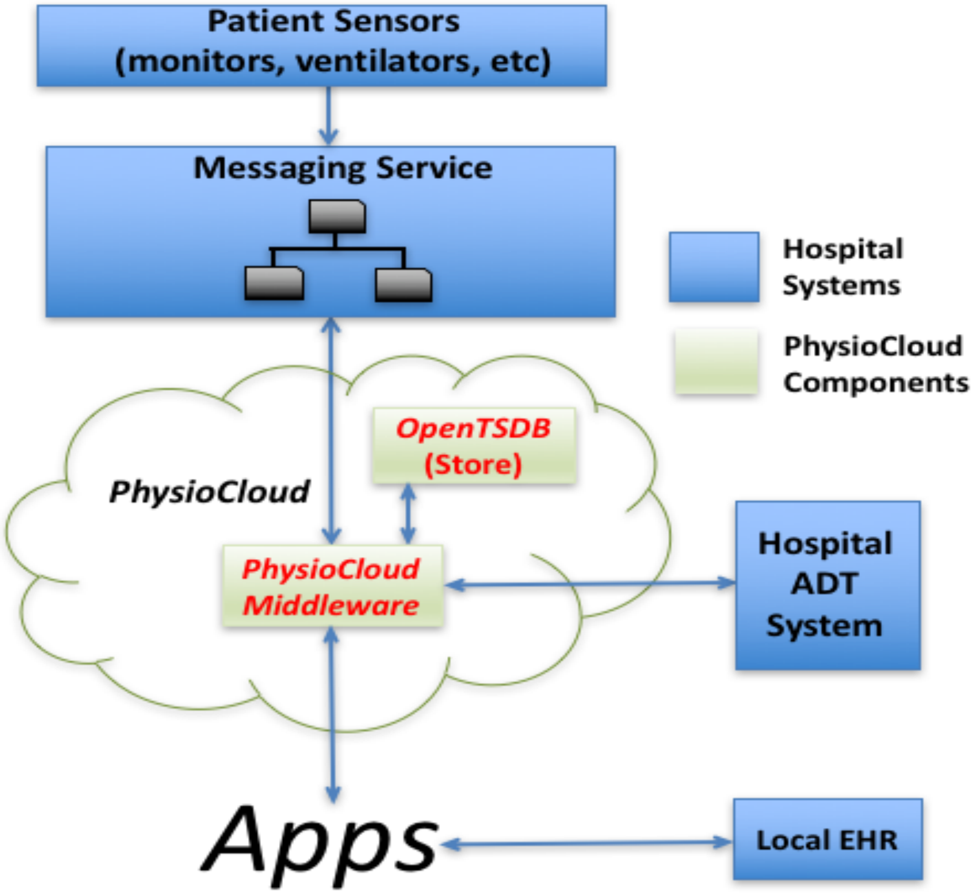
Data Artifacts:

Noise



Here's A Real Bizarre Problem:
Who is the Patient?

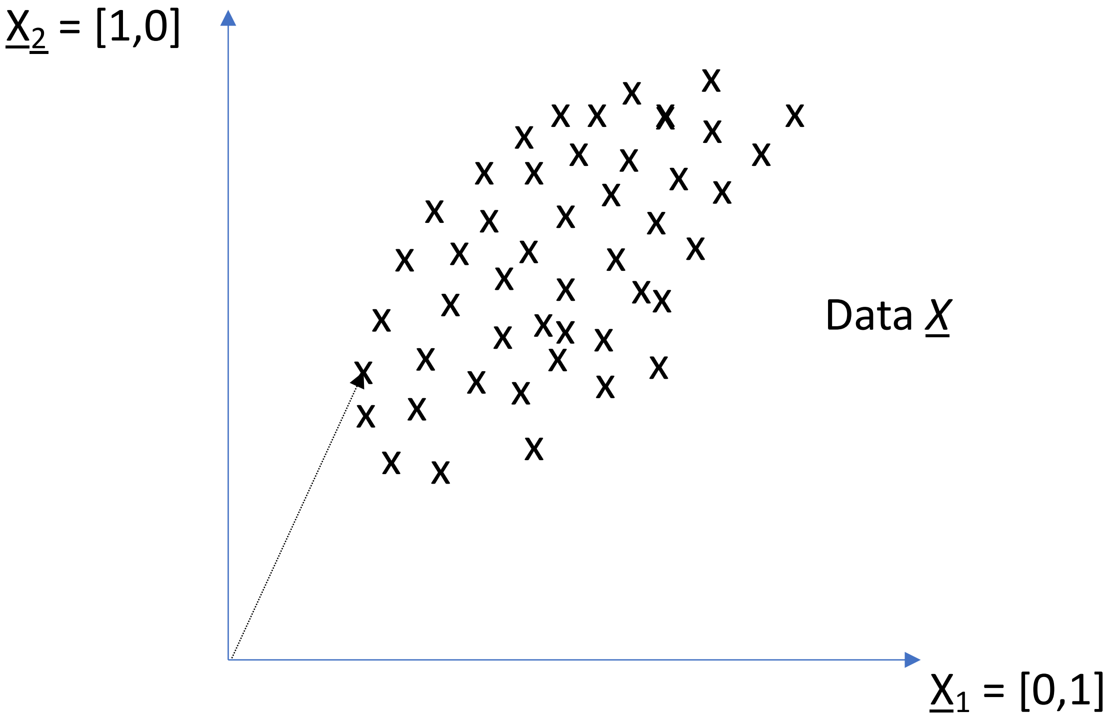
- In the past we've had collaborations in which the mapping from dataset to patient isn't ***truly*** known !

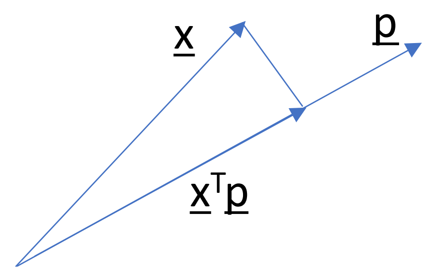


Data Exploration in the Wild

Principle Components Analysis

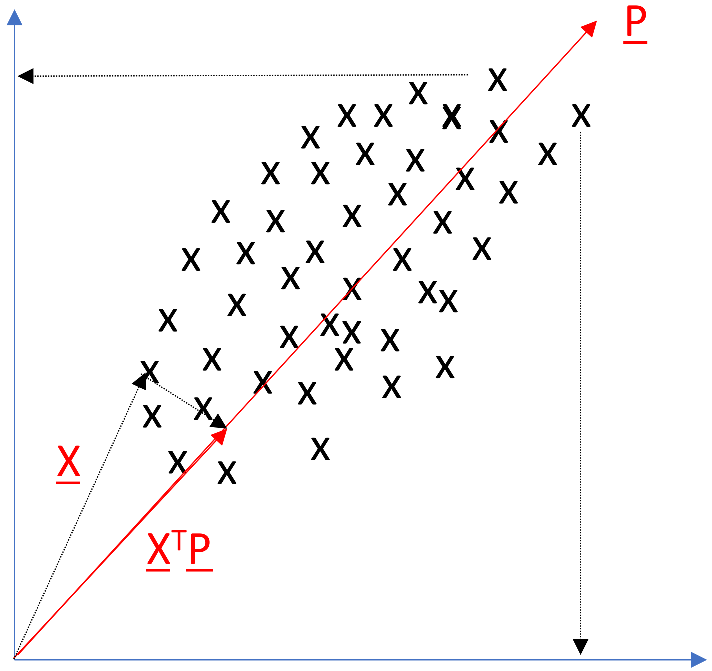
Principle Component Analysis: Motivation





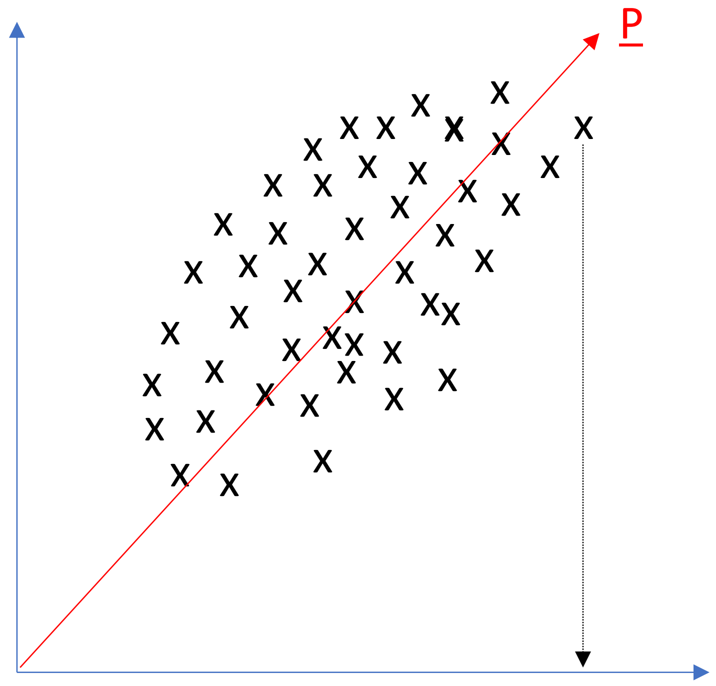
- Bases are the directions $\underline{x}_1$ and $\underline{x}_2$. Two features are plotted in this 2-dimensional vector-space
- We want to find a vector $\underline{P}$ such that the collection of random variables $\{xp_1, xp_2, .. ,xp_n\}$ that are the magnitudes of the projections of the data $\underline{X}$ onto $\underline{P}$ have maximum variance
- What is your guess for this “best” direction (basis vector) $\underline{P}$?

Principle Component Analysis: Motivation



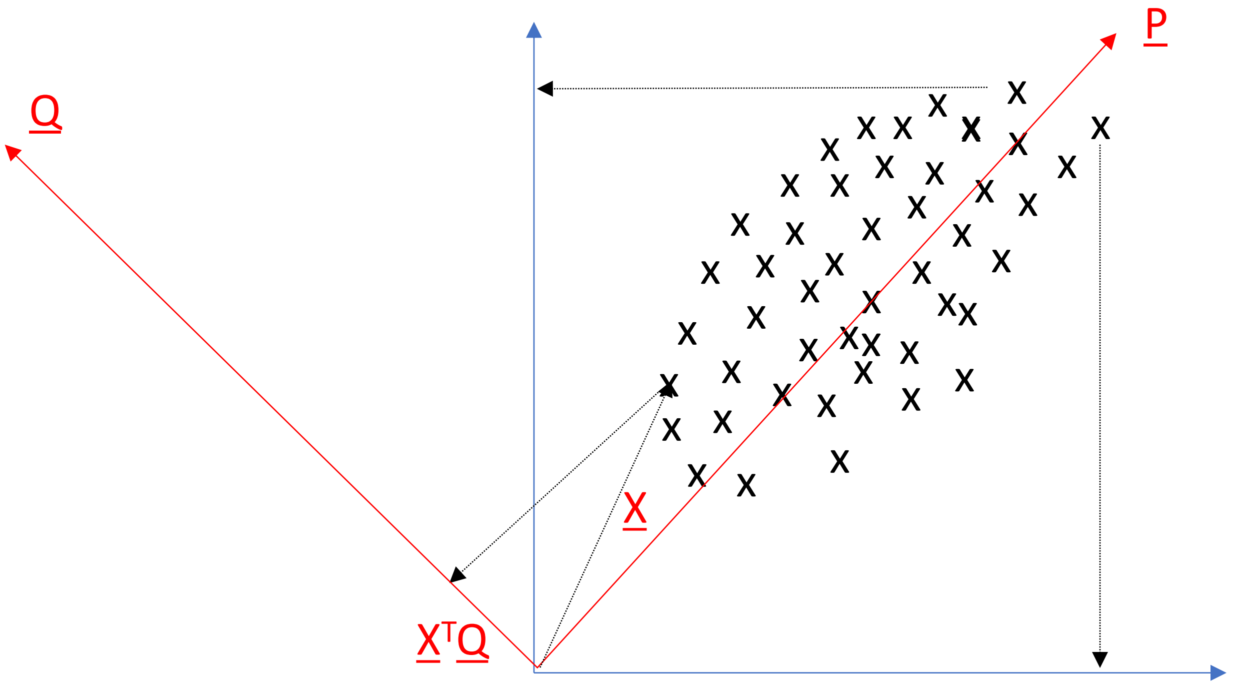
- Hopefully you chose something like the P I have drawn !
- To simplify data analysis, we might choose to project our data onto P so that we can deal with analyzing our data as scalars rather than as pairs of numbers
- *Feature reduction*

Principle Component Analysis: Motivation



- Next, we want to find a vector $\underline{Q}$ such that the collection of random variables $\{xq_1, xq_2, .. ,xq_n\}$ that are the magnitudes of the projections of the data $\underline{X}$ onto $\underline{Q}$ have the second largest variation possible
- What would be your choice of $\underline{Q}$?

Principle Component Analysis: Motivation



- I think it looks like the one I've drawn here
- P and Q are a new orthogonal basis set in which to represent the data
- If we choose to reduce the dimension of our data $\underline{X}$, P may be a good choice of a single basis vector with which to do that because it captures the largest variation in $\underline{X}$

“The Good, the Bad, and the Ugly” of PCA

1966 “Spaghetti” Western (Sergio Leone)
Clint Eastwood

- The *Good* - May be able to represent n-dimensional data in a space of much lower dimension, hopefully 2 or 3 because these projections are then easy to visualize. Also classes may separate naturally

- The *Bad and Ugly* – Classes may not separate, classes may overlap considerably, no guarantee

- Even if classes are not separable in this basis, don’t give up, classes may be separable using other approaches that we will learn about later

Let's Get Formal

- Let $\underline{X}$ be an n-dimensional observation vector of n features
- $\underline{X}$ is a random vector, that is, it's elements x_i are drawn from some underlying probability distribution
- The elements x_i of $\underline{X}$ have 0 mean, that is:

$$\underline{X} = \begin{bmatrix} x_1 \\ . \\ . \\ x_n \end{bmatrix}$$

$$E\{\underline{X}\} = \begin{bmatrix} E\{x_1\} \\ . \\ . \\ E\{x_n\} \end{bmatrix} = \begin{bmatrix} 0 \\ . \\ . \\ 0 \end{bmatrix}$$

E{ } is the expected value operator

$$E(X) = \sum_{x_i \in \Omega} x_i p(x_i)$$ X discrete

$$E(X) = \int_{\Omega} xp(x)\,dx$$ X continuous

The Data Covariance Matrix

$$Cov(\underline{X}) = E\{\underline{X}\underline{X}^T\} = C = \begin{bmatrix} E\{x_1x_1\} & E\{x_1x_2\} & . & . & E\{x_1x_n\} \\ E\{x_2x_1\} & . & . & . & E\{x_2x_n\} \\ . & . & . & . & . \\ . & . & . & . & . \\ E\{x_nx_1\} & . & . & . & E\{x_nx_n\} \end{bmatrix}$$

- By construction, what are important properties of C ?
- The *n x n* matrix C is by definition a real, symmetric, and positive-semidefinite matrix
- These facts will be relevant

Let's Get Formal

- Suppose we choose, as we did in the example, to project our data $\underline{X}$ onto some vector $\underline{P}$.
- We want to choose $\underline{P}$ to maximize the variance of the set of scalar RVs that are these projections
- What is this variance ?

$$Var(\underline{P}^T\underline{X}) = E\{(\underline{P}^T\underline{X})(\underline{P}^T\underline{X})^T\} = E\{(\underline{P}^T\underline{X})(\underline{X}^T\underline{P})\} = \underline{P}^T E\{\underline{X}\underline{X}^T\}\underline{P} = \underline{P}^T C\underline{P}$$

This variance is not well-defined because we could scale the vector $\underline{P}$ and thereby scale the variance arbitrarily

To deal with this, let's impose the constraint $\underline{P}^T\underline{P} = 1$, that is, $\underline{P}$ is a unit normal vector

Let’s Get Formal

We want to find the $\hat{\underline{P}}$ that satisfies the following

$$\hat{\underline{P}} = \underset{\underline{P}^T \underline{P}=1}{\arg\max}\left(\underline{P}^T C \underline{P} \right)$$

For a real, symmetric, positive-semidefinite matrix C, the maximizer is the unit eigenvector of C corresponding to the largest eigenvalue

Since C is a real symmetric matrix, it has a set of n orthonormal eigenvectors $\{\underline{e}_1, \underline{e}_2, \ldots, \underline{e}_n\}$
Use $\{\underline{e}_1, \underline{e}_2, \ldots, \underline{e}_n\}$ as the basis with which to represent our optimal choice of vector

Define the modal matrix M as $M = \begin{bmatrix} \underline{e}_1 & \underline{e}_2 & \cdots & \underline{e}_n \end{bmatrix}$

The vector $\underline{P}$ has a representation, call it $\underline{r}$, with respect to this basis:

$$\underline{P} = M\underline{r}$$

$$\underline{P}^T C \underline{P} = \left(M \underline{r} \right)^T CM\underline{r} = \underline{r}^T M^T CM\underline{r} = \underline{r}^T \Lambda \underline{r}$$

$$\underline{r}^T \Lambda \underline{r} = \sum_{i=1}^{n} r_i^2 \lambda_i \leq \lambda_1 \sum_{i=1}^{n} r_i^2 = \lambda_1$$

The above follows since the similarity transformation $M^T CM$ diagonalizes the matrix C.

Λ is this diagonalization. This is easy to show using the definition of M

The above gives the maximum value of $\underline{P}^T C\underline{P} = \underline{r}^T \Lambda \underline{r}$. This maximum value is achieved with $\underline{P} = \underline{e}_1$

$$\underline{e}_1^T C \underline{e}_1 = \underline{e}_1^T \lambda_1 \underline{e}_1 = \lambda_1 \underline{e}_1^T \underline{e} = \lambda_1$$

(see last 3 slides for solution using Lagrange Multipliers)

- Our goal was to find a vector $\underline{P}$ such that the projection of our data $\underline{X}$ onto $\underline{P}$ would have the maximum variance possible.

- The solution is to choose $\underline{P}$ to be the eigenvector of $C=E\{\underline{XX}^T\}$ corresponding to the largest eigenvalue of C

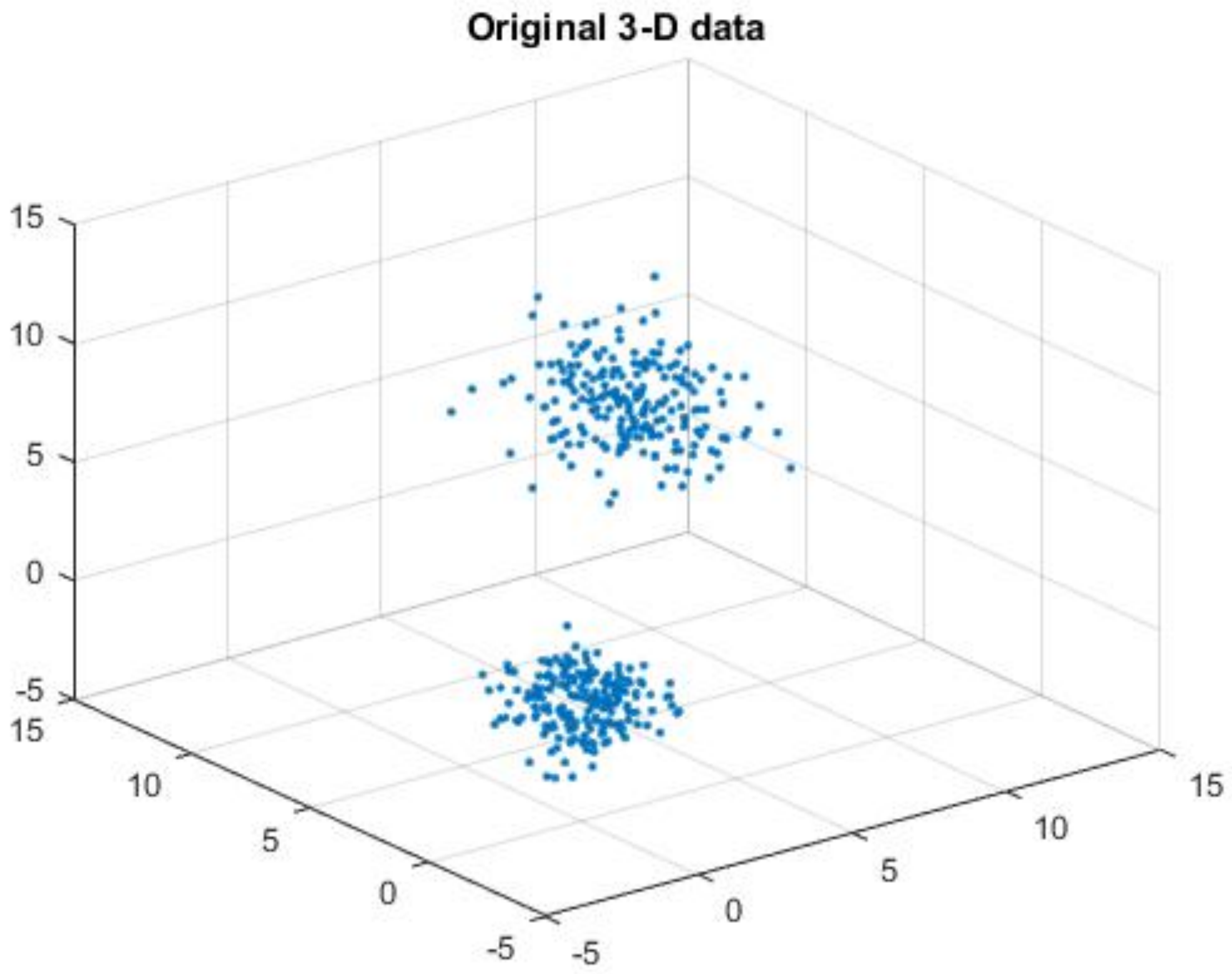
- Denote these as $\underline{P}_1$ (the primary eigenvector) and λ_1 (the primary eigenvalue) of C

- $\underline{P}_1$ is known as the first principle component of our data $\underline{X}$.

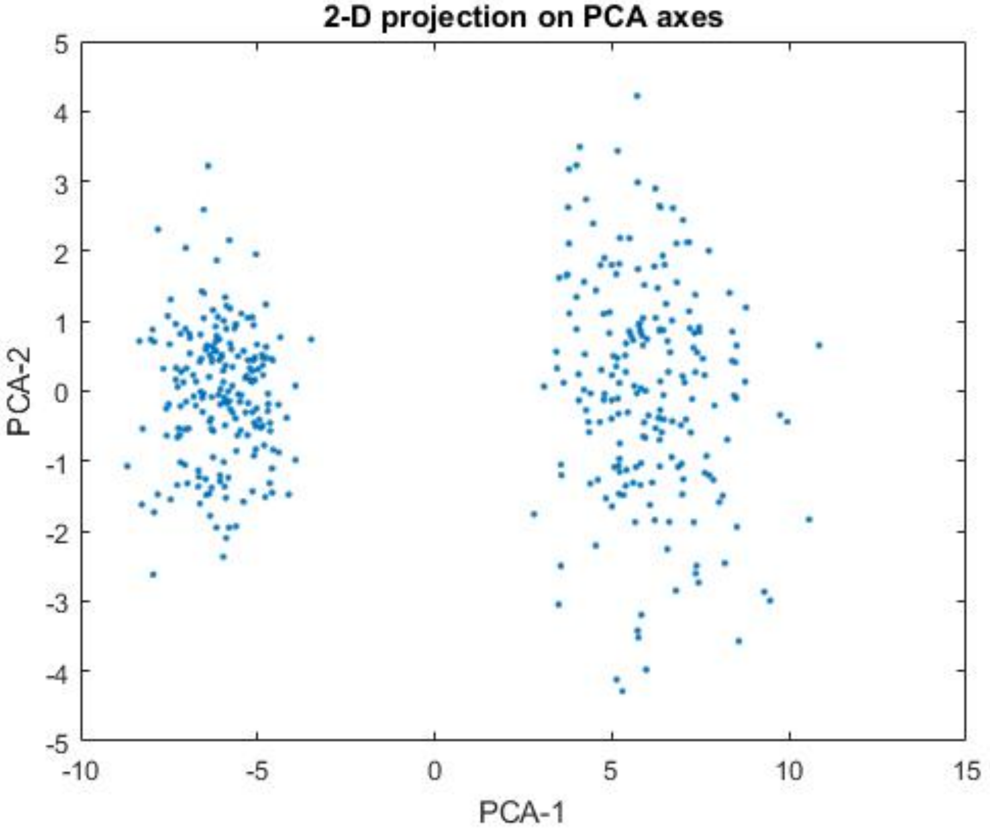
- If $\lambda_1 > \lambda_2 > \ldots > \lambda_n$ (and note strict inequality need not hold), then choose $\underline{e}_2, \ldots \underline{e}_n$ as successive basis vectors that account for the largest amount of variance in $\underline{X}$ of any basis choice

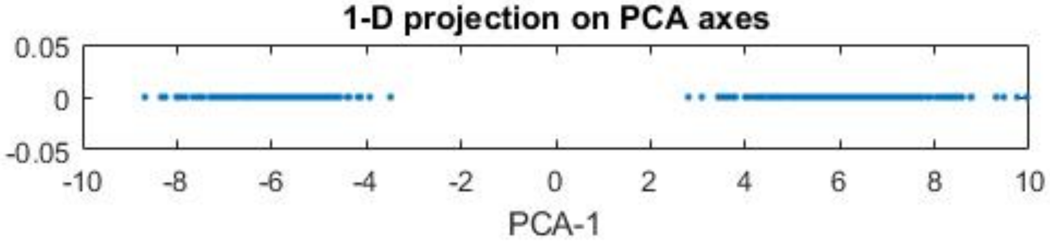
Note: None of this depended on assumptions regarding the distribution of the data $\underline{X}$.

The “Good” Example 1: Simulated 3-D Normal Data, Different Means



The “Good” Example 1: Simulated 3-D Normal Data





The "Bad and Ugly" Example 2: Predicting Risk of Cardiac Arrest

Variables:

- age-corrected HR, HRV, age-corrected RESP, SpO2, PVC rate per minute, age, gender from 13 CA and 13 NCA subjects.
- The PCA features are normalized by subtracting their feature(column) mean and dividing by the feature(column) standard deviation to avoid bias towards any particular feature.
- % of variance explained by the PCs:

  PC-1: 24.7761

  PC-2: 21.7598

  PC-3: 14.8896

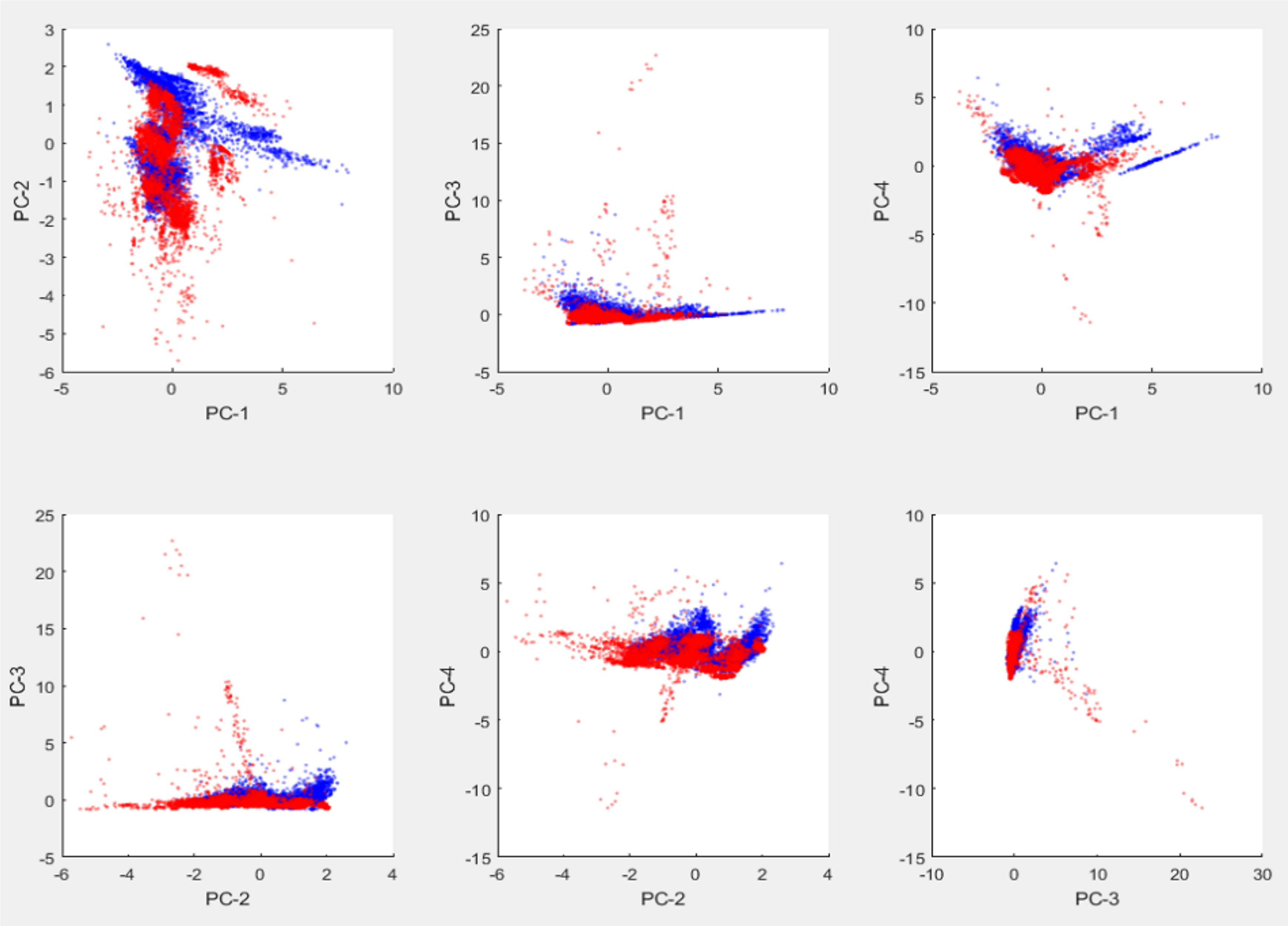
  PC-4: 13.0659

  PC-5: 11.1371

  PC-6: 8.2437

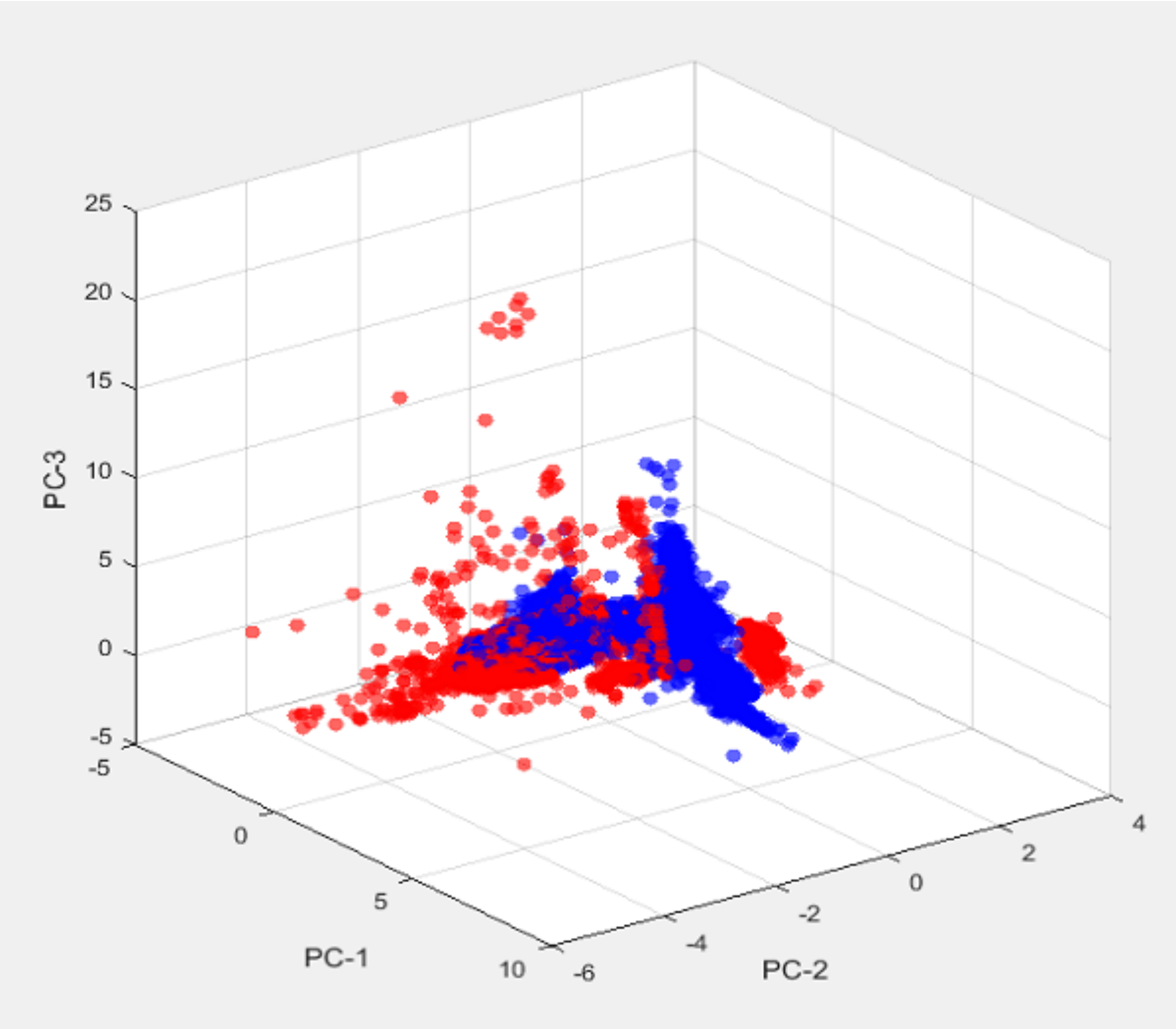
  PC-7: 6.1278

The “Bad and Ugly” Example 2: Predicting Risk of Cardiac Arrest



Red : CA subjects
Blue: NCA subjects

The “Bad and Ugly” Example 2: Predicting Risk of Cardiac Arrest



Red : CA subjects
Blue: NCA subjects

Here is a Solution for the 1st Principle Component Using Lagrange Multipliers

Let's use the method of Lagrange multipliers to find $\underline{P}$.
Define the Lagrangian Function u and the Lagrange multiplier α :

$$u(\underline{P},a)=\underline{P}^T C\underline{P}-a(\underline{P}^T\underline{P}-1)$$

We want to find the α and $\underline{P}$ that maximize the above subject to constraint $\underline{P}^T\underline{P}=1$.
To find solutions, set the derivatives wrt α and $\underline{P}$ to zero.

$$u(\underline{P},a)=\underline{P}^T C\underline{P}-a(\underline{P}^T\underline{P}-1)$$

$$\frac{\partial u(\underline{P},a)}{\partial a}=0\Rightarrow-(\underline{P}^T\underline{P}-1)=0\Rightarrow\underline{P}^T\underline{P}=1$$ (setting this derivative to 0 captures the constraint on $\underline{P}$)

$$\frac{\partial u(\underline{P},a)}{\partial \underline{P}}=??$$ (whatever this derivative is, we will set it to 0)

Let’s Get Formal

$$u(\underline{P},a) = \underline{P}^T C\underline{P} - a(\underline{P}^T \underline{P} - 1)$$

$$\frac{\partial u}{\partial \underline{P}} = \begin{bmatrix} \frac{\partial u}{\partial p_1} & \frac{\partial u}{\partial p_2} & \dots & \frac{\partial u}{\partial p_n} \end{bmatrix}$$

$$\frac{\partial \underline{P}^T \underline{P}}{\partial \underline{P}} = \begin{bmatrix} \frac{\partial(p_1^2 + p_2^2 + \ldots + p_n^2)}{\partial p_1} & \dots & \frac{\partial(p_1^2 + p_2^2 + \ldots + p_n^2)}{\partial p_n} \end{bmatrix}$$

$$= \begin{bmatrix} 2p_1 & \dots & 2p_n \end{bmatrix} = 2\underline{P}$$

$$\frac{\partial \underline{P}^T C\underline{P}}{\partial \underline{P}} = 2C\underline{P}$$

Let's Get Formal

$$\frac{\partial u(\underline{P},a)}{\partial \underline{P}} = 2C\underline{P} - a2\underline{P} = 0 \Rightarrow C\underline{P} = a\underline{P}$$

Therefore we have

$$\frac{\partial u(\underline{P},a)}{\partial a} = -(\underline{P}^T\underline{P} - 1) = 0 \Rightarrow \underline{P}^T\underline{P} = 1$$

$$\frac{\partial u(\underline{P},a)}{\partial \underline{P}} = 2C\underline{P} - a2\underline{P} = 0 \Rightarrow C\underline{P} = a\underline{P}$$

What do these relationships tell us?

$\underline{P}$ is a unit eigenvector of C corresponding to the maximum eigenvalue of C λ_1