Predictive Diagnosis

Clustering to Better Predict Heart Attacks

15.071x – The Analytics Edge

Heart Attacks

- Heart attack is a common complication of coronary heart disease resulting from the interruption of blood supply to part of the heart

- 2012 report from the American Heart Association estimates **about 715,000** Americans have a heart attack every year
  - **Every 20 seconds**, a person has a heart attack in the US
  - Nearly **half** occur without prior warning signs
  - **250,000** Americans die of Sudden Cardiac Death yearly

Heart Attacks

- Well-known symptoms
  - Chest pain, shortness of breath, upper body pain, nausea

- Nature of heart attacks makes it hard to predict, prevent and even diagnose
  - **25%** of heart attacks are silent
  - **47%** of sudden cardiac deaths occur outside hospitals, suggesting many do not act on early warning signs
  - **27%** of respondents to a 2005 survey recognized the symptoms and called 911 for help

Analytics Helps Monitoring

- Understanding the clinical characteristics of patients in whom heart attack was missed is key
- Need for an increased understanding of the patterns in a patient's diagnostic history that link to a heart attack
- Predicting whether a patient is at risk of a heart attack helps monitoring and calls for action
- Analytics helps **understand patterns** of heart attacks and provides **good predictions**

Claims Data

- Claims data offers an expansive view of a patient's health history
  - Demographics, medical history and medications
  - Offers insights regarding a patient's risk
  - May reveal **indicative signals and patterns**
- We will use health insurance claims filed for about 7,000 members from January 2000 – November 2007

Claims Data

- Concentrated on members with the following attributes
  - At least 5 claims with coronary artery disease diagnosis
  - At least 5 claims with hypertension diagnostic codes
  - At least 100 total medical claims
  - At least 5 pharmacy claims
  - Data from at least 5 years

- Yields patients with a high risk of heart attack and a reasonably rich history and continuous coverage

Data Aggregation

- The resulting dataset includes about **20 million health insurance entries** including individual medical and pharmaceutical records

- Diagnosis, procedure and drug codes in the dataset comprise tens of thousands of attributes

- Codes were aggregated into groups
  - 218 diagnosis groups, 180 procedure groups, 538 drug groups
  - 46 diagnosis groups were considered by clinicians as possible risk factors for heart attacks

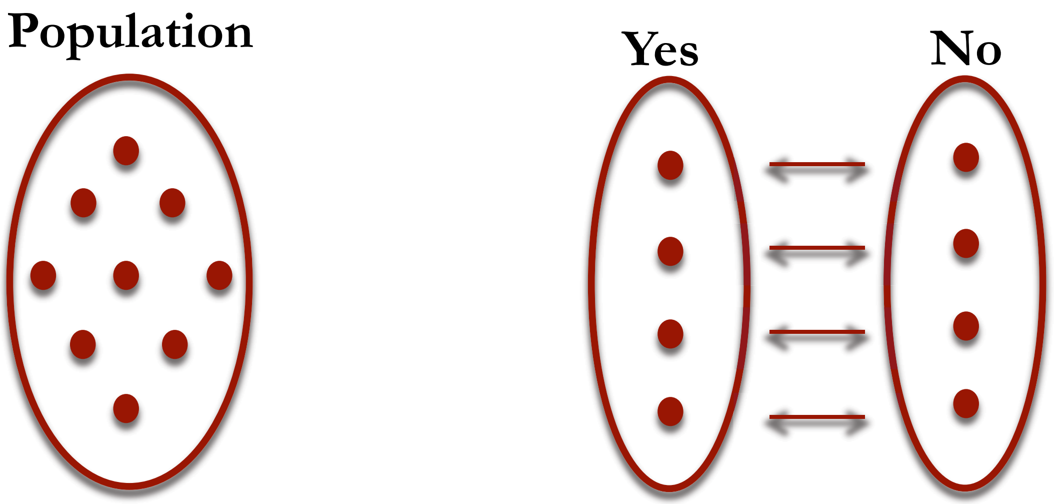
Diagnostic History

- We then compress medical records to obtain a chronological representation of a patient's diagnostic profile
    - Cost and number of medical claims and hospital visits by diagnosis

- Observations split into 21 periods, each 90 days in length
    - Examined 9 months of diagnostic history leading up to heart attack/no heart attack event
    - Align data to make observations date-independent while preserving the order of events
        - 3 months ~ 0-3 months before heart attack
        - 6 months ~ 3-6 months before heart attack
        - 9 months ~ 6-9 months before heart attack

Target Variable

- Target prediction is the first occurrence of a heart attack
  - Diagnosis on medical claim
  - Visit to emergency room followed by hospitalization
  - Binary Yes/No



Dataset Compilation

Variables	Description
1	Patient identification number
2	Gender
3-49	Diagnosis group counts 9 months before heart attack
50	Total cost 9 months before heart attack
51-97	Diagnosis group counts 6 months before heart attack
98	Total cost 6 months before heart attack
99-145	Diagnosis group counts 3 months before heart attack
146	Total cost 3 months before heart attack
147	Yes/No heart attack

Cost Bucket Partitioning

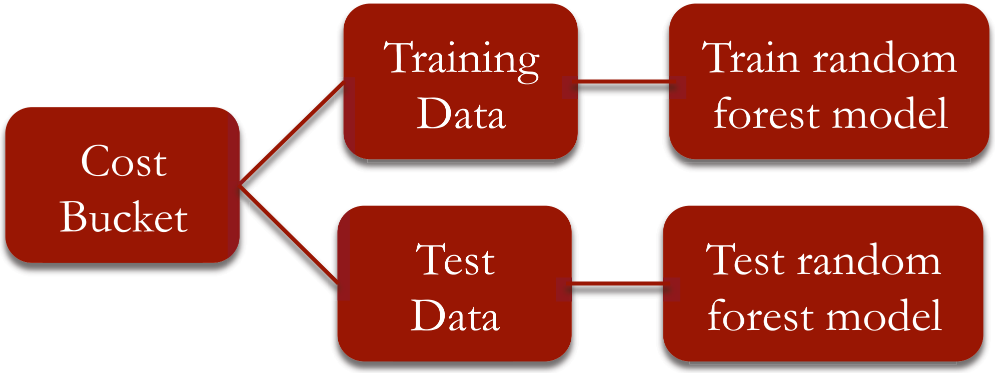
- Cost is a good summary of a person's overall health
- Divide population into similar smaller groups
  - Low risk, average risk, high risk

Bucket	Cost Range	% Data	Members	% with Heart Attack
1	< $2K	67.56	4,416	36.14
2	$2K - $10K	21.56	1,409	43.22
3	> $10K	10.88	711	38.12

- Build models for each group

Predicting Heart Attacks (Random Forest)

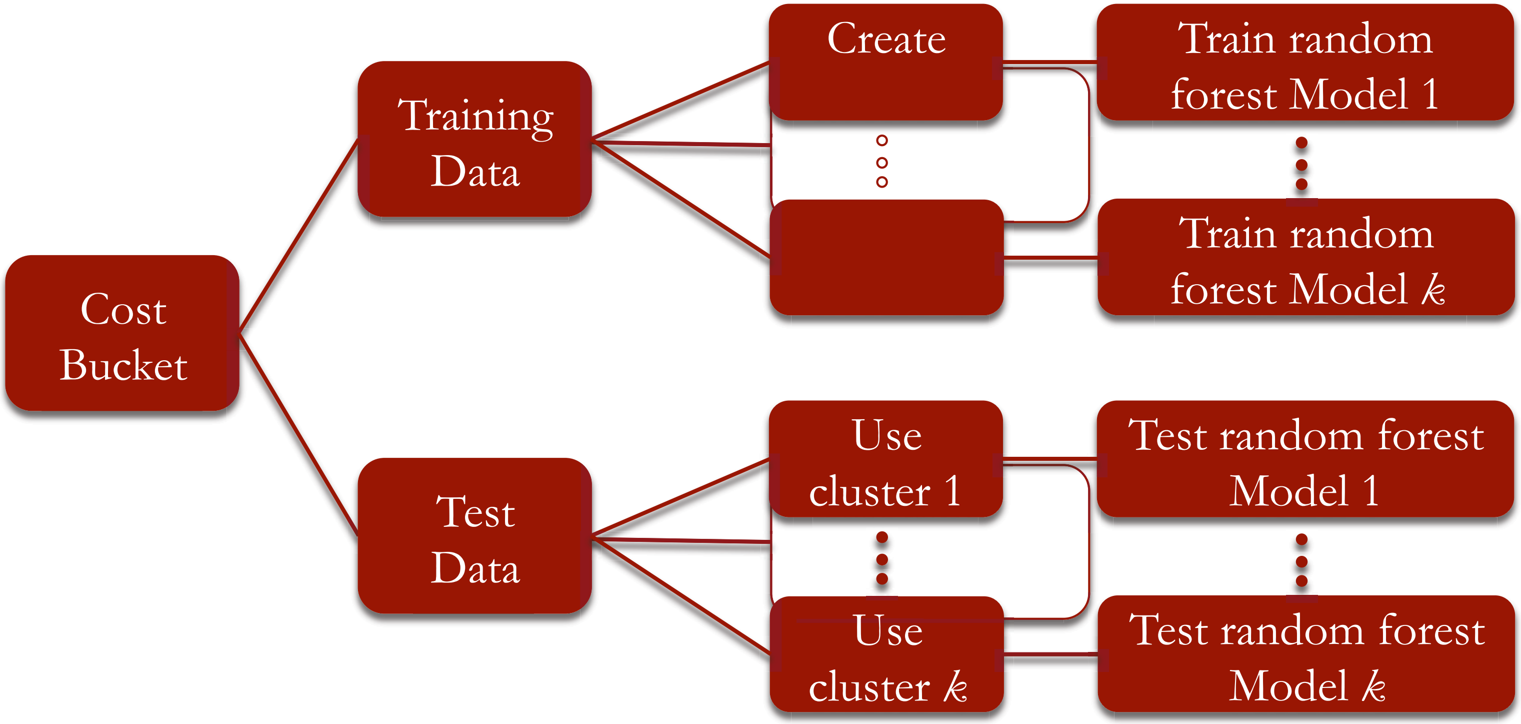
- Predicting whether a patient has a heart attack for each of the cost buckets using the random forest algorithm



Bucket	Random Forest
1	49.63%
2	55.99%
3	58.31%

Incorporating Clustering

- Patients in each bucket may have different characteristics


Cost Bucket
Training Data
Test Data
Create
Train random forest Model 1
Train random forest Model *k*
Use cluster 1
Use cluster *k*
Test random forest Model 1
Test random forest Model *k*

Clustering Cost Buckets

- Two clustering algorithms were used for the analysis as an alternative to hierarchal clustering
  - Spectral Clustering
  - ***k*-means clustering**

Clustering Cost Buckets

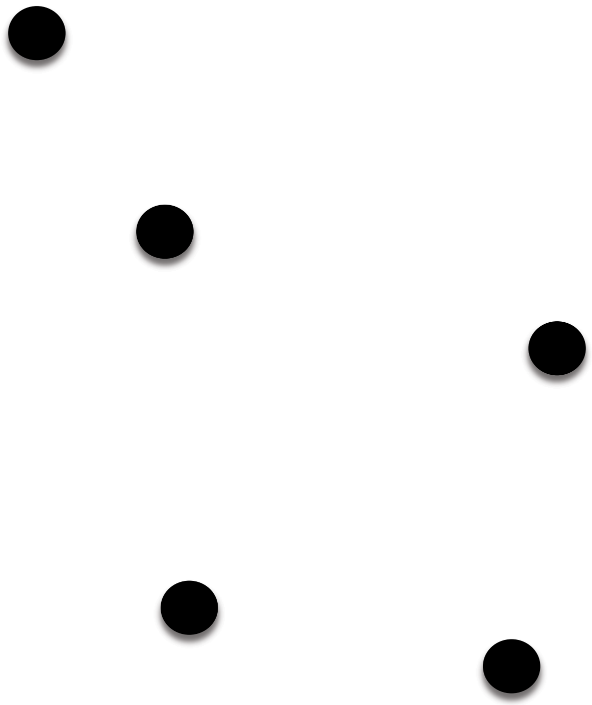
- Two clustering algorithms were used for the analysis as an alternative to hierarchal clustering
  - Spectral Clustering
  - ***k*-means clustering**

***k*-Means Clustering Algorithm**
1. Specify desired number of clusters *k*
2. Randomly assign each data point to a cluster
3. Compute cluster centroids
4. Re-assign each point to the closest cluster centroid
5. Re-compute cluster centroids
6. Repeat 4 and 5 until no improvement is made

k-Means Clustering

_k_-Means Clustering Algorithm

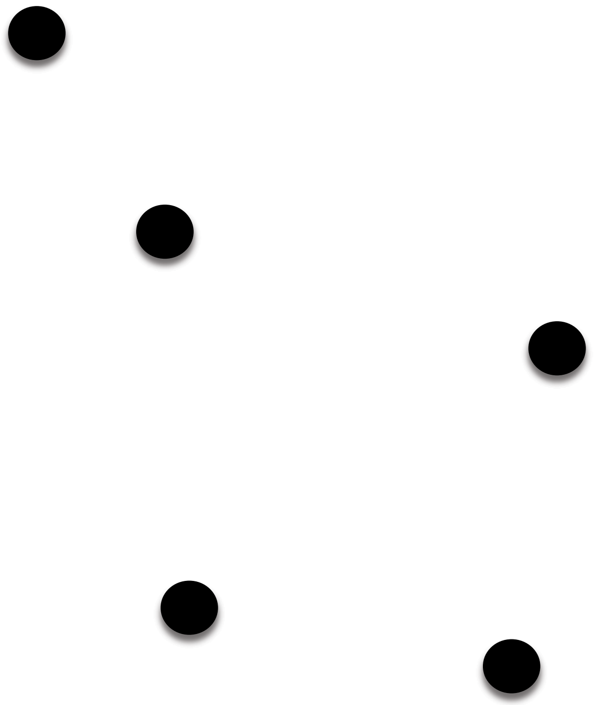
1. Specify desired number of clusters *k*



k-Means Clustering

_k_-Means Clustering Algorithm

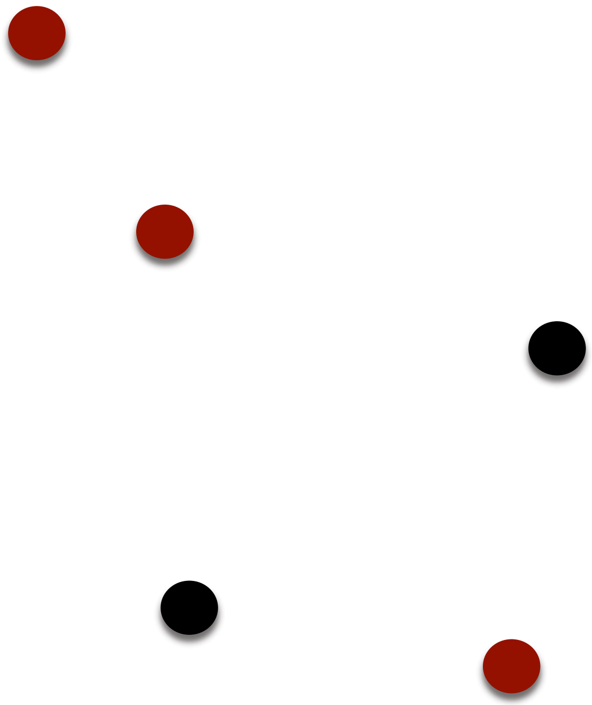
1. Specify desired number of clusters *k*
2. Randomly assign each data point to a cluster



k-Means Clustering

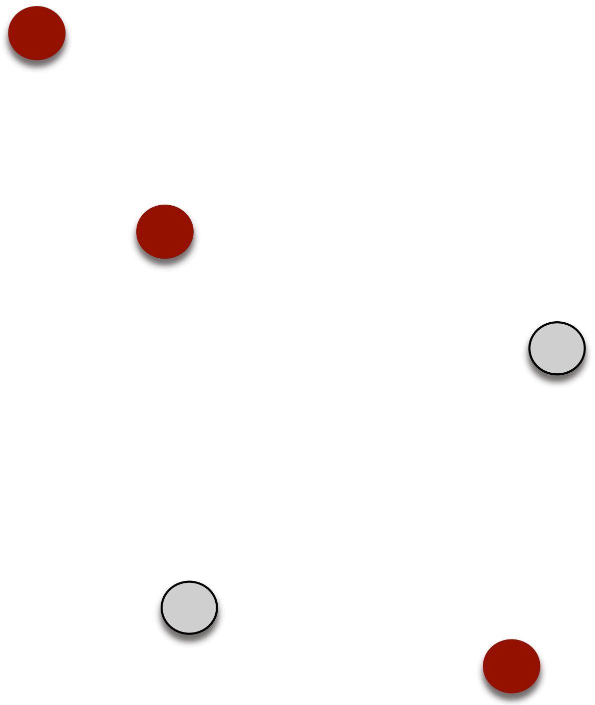
_k_-Means Clustering Algorithm

1. Specify desired number of clusters *k*
2. Randomly assign each data point to a cluster



k-Means Clustering

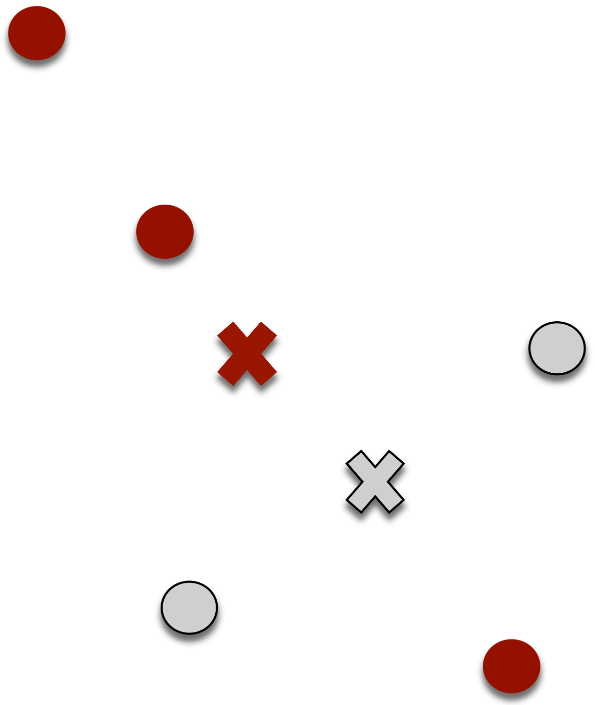
***k*-Means Clustering Algorithm**
1. Specify desired number of clusters *k*
2. Randomly assign each data point to a cluster



k-Means Clustering

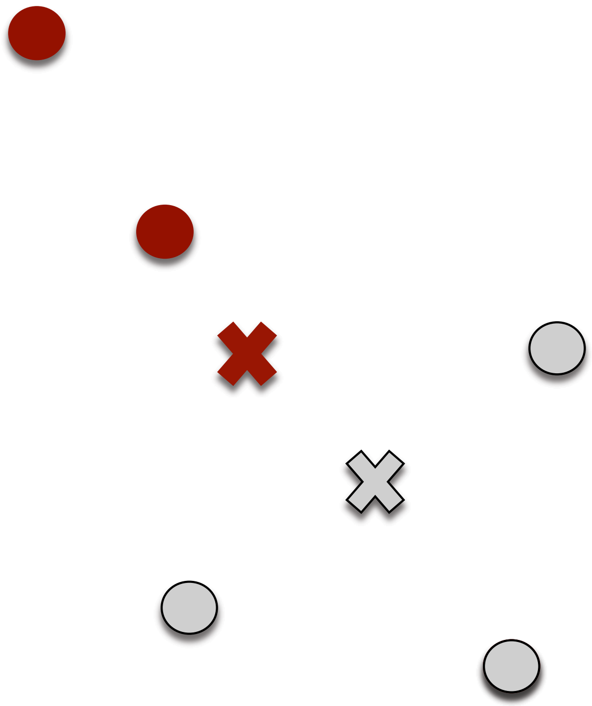
_k_-Means Clustering Algorithm

1. Specify desired number of clusters *k*
2. Randomly assign each data point to a cluster
3. Compute cluster centroids



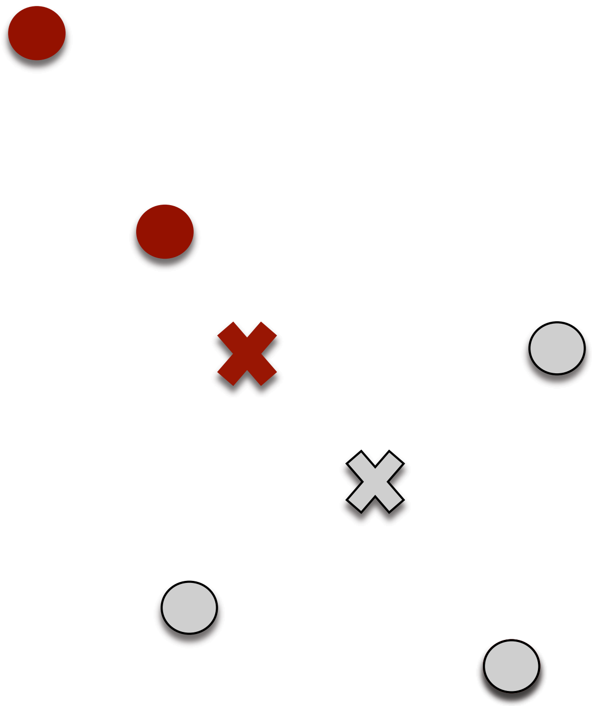
k-Means Clustering

k-Means Clustering Algorithm
1. Specify desired number of clusters *k*
2. Randomly assign each data point to a cluster
3. Compute cluster centroids
4. Re-assign each point to the closest cluster centroid



k-Means Clustering

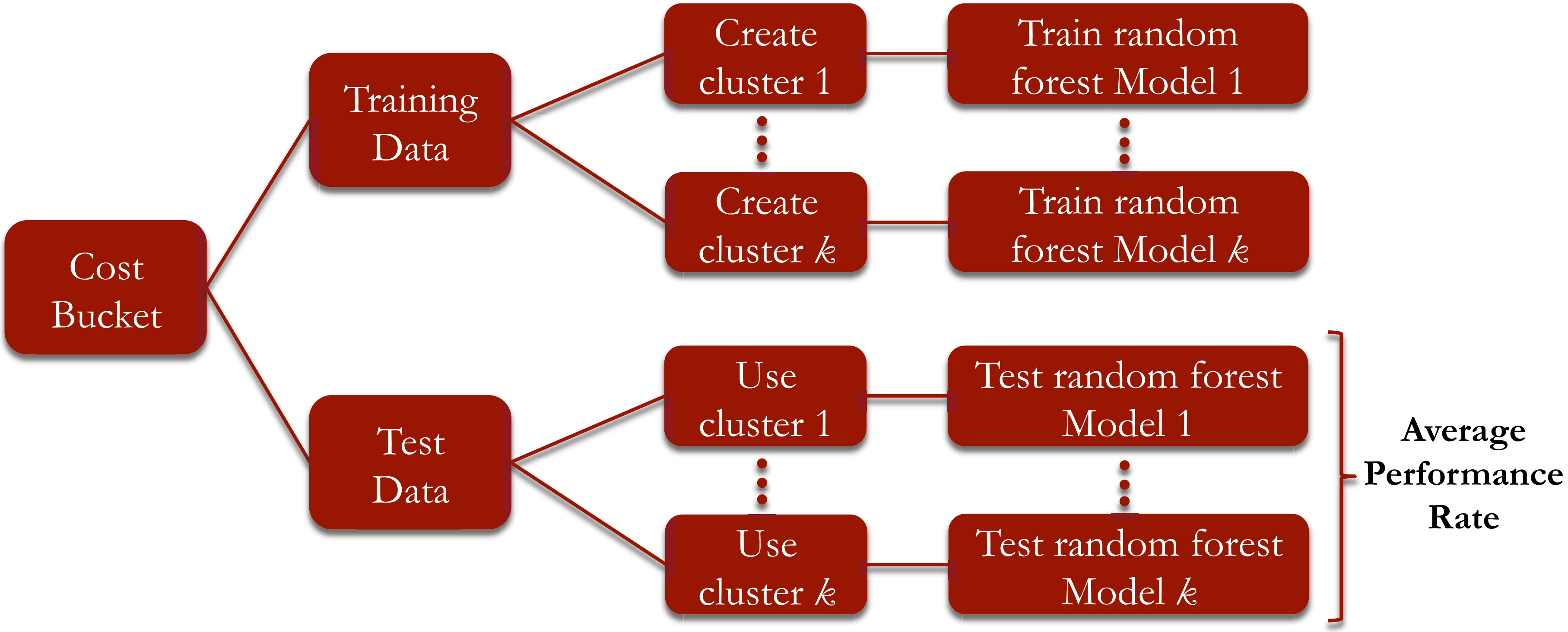
***k*-Means Clustering Algorithm**
1. Specify desired number of clusters *k*
2. Randomly assign each data point to a cluster
3. Compute cluster centroids
4. Re-assign each point to the closest cluster centroid
5. Re-compute cluster centroids
6. Repeat 4 and 5 until no improvement is made



Practical Considerations

- The number of clusters *k* can be selected from previous knowledge or experimenting
- Can strategically select initial partition of points into clusters if you have some knowledge of the data
- Can run algorithm several times with different random starting points
- In recitation, we will learn how to run the *k*-means clustering algorithm in R

Random Forest with Clustering



Predicting Heart Attacks

- Perform clustering on each bucket using k=10 clusters
- Average prediction rate for each cost bucket

Cost Bucket	Random Forest without Clustering	Random Forest with Clustering
1	49.63%	64.75%
2	55.99%	72.93%
3	58.31%	78.25%

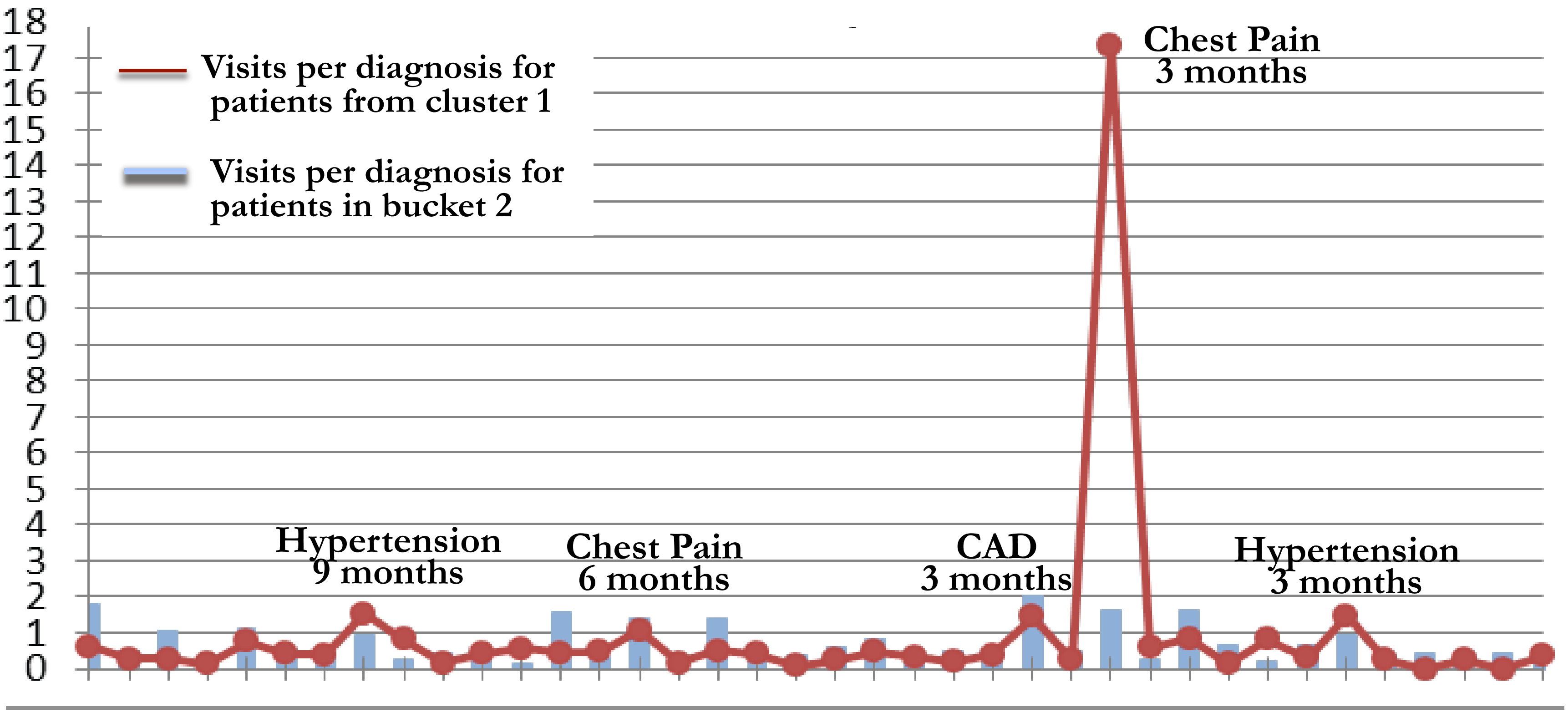
Understanding Cluster Patterns

- Clusters are interpretable and reveal unique patterns of diagnostic history among the population

Cost Bucket 2	
Cluster 1	Chest Pain (3 months)
Cluster 6	Coronary Artery Diseases (3 months)
Cluster 7	Chronic Obstructive Pulmonary Disease
Cost Bucket 3	
Cluster 4	Anemia (3, 6, 9 months)
Cluster 5	Hypertension and Cerebrovascular Disease
Cluster 10	Diabetes (3, 6, 9 months)

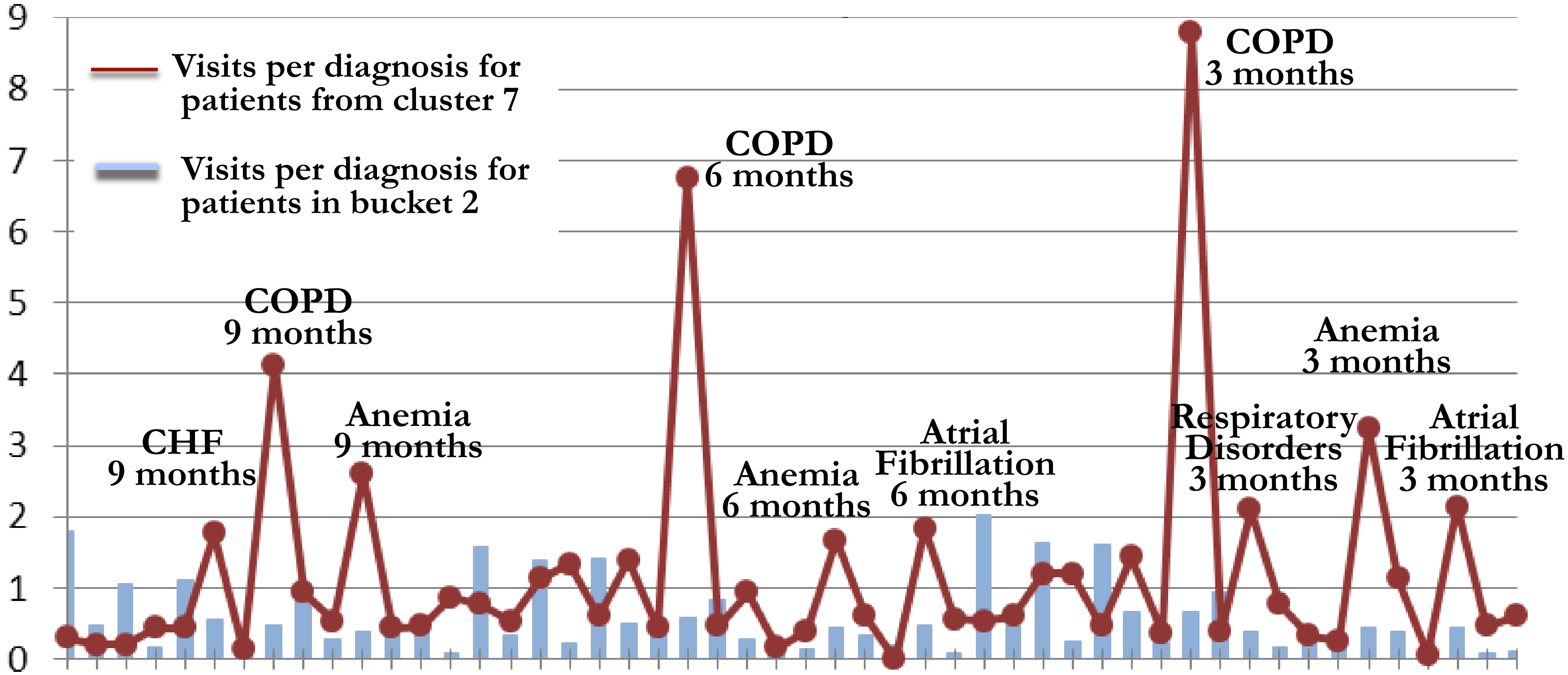
Occurrence of Chest Pain

- Cluster 1 in bucket 2 reflects a temporal pattern of chest pain diagnosed 3 months before a heart attack



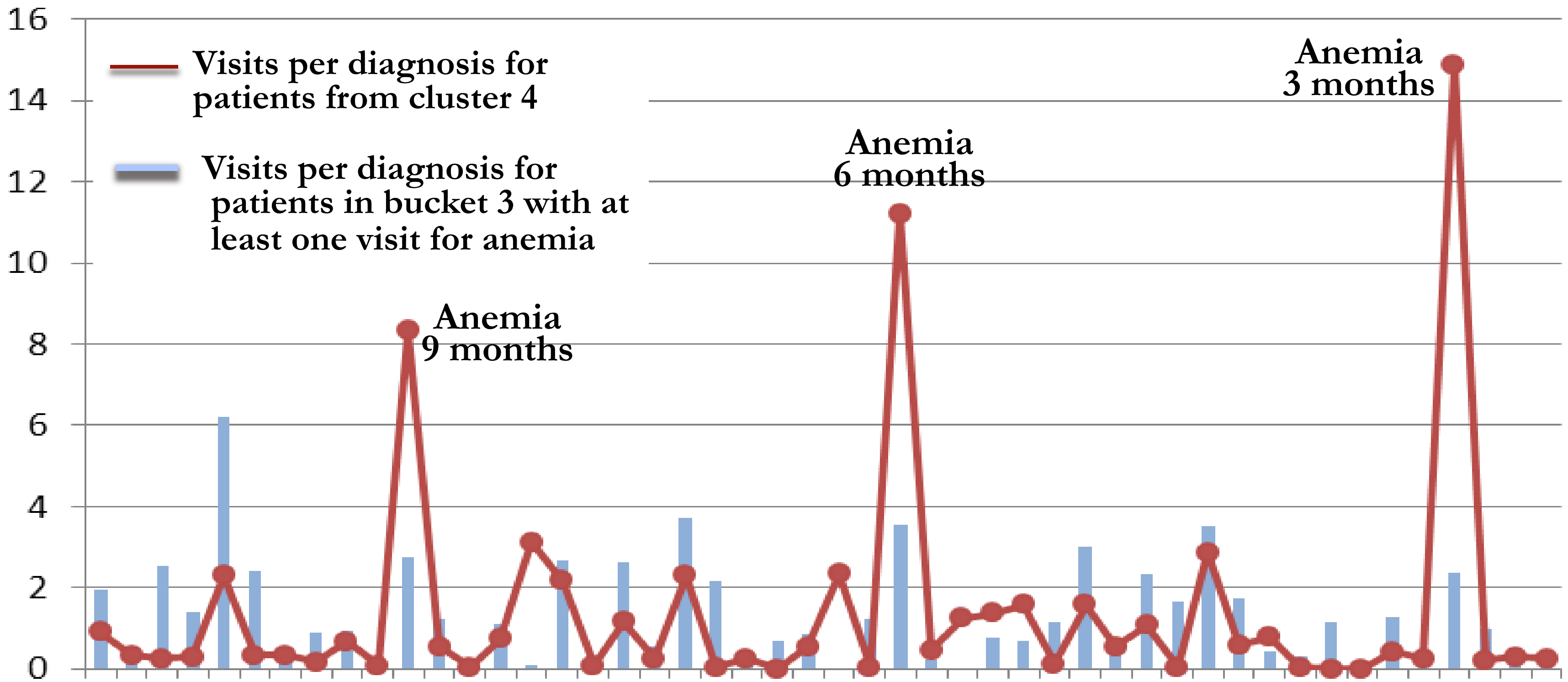
Chronic Obstructive Pulmonary Disease (COPD)

- Patients from Cluster 7 in cost bucket 2 who suffered a heart attack have regular doctor visits for COPD



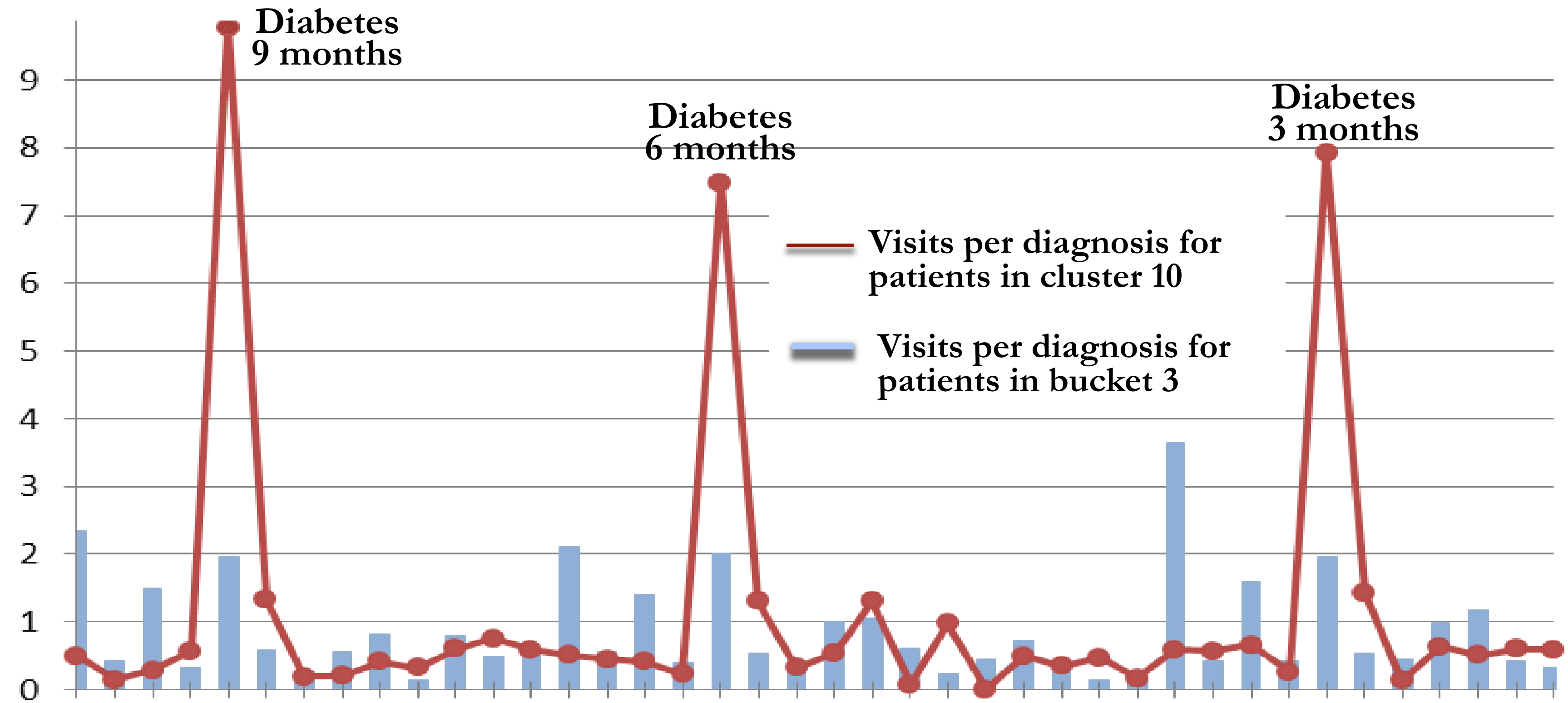
Gradually Increasing Occurrence of Anemia

- Cluster 4 in bucket 3 shows a temporal diagnosis pattern of anemia



Occurrence of Diabetes

- Cluster 10 in bucket 3 shows a temporal diagnosis of diabetes



Impact of Clustering

- Clustering members within each cost bucket yielded better predictions of heart attacks within clusters
- Grouping patients in clusters exhibits temporal diagnostic patterns within 9 months of a heart attack
- These patterns can be incorporated in the diagnostic rules for heart attacks
- Great research interest in using analytics for early heart failure detection through pattern recognition

Analytics for Early Detection

- IBM, Sutter Health and Geisinger Health System partnered in 2009 to research analytics tools in view of early detection

"Our earlier research showed that signs and symptoms of heart failure in patients are often documented **years before** a diagnosis"

" The **pattern of documentation** can offer clinically useful **signals for early detection** of this deadly disease"

Steve Steinhubl (2013), Cardiologist MD from Geisenger