Big Data Analytics for Healthcare

Jimeng Sun
Healthcare Analytics Department
IBM TJ Watson Research Center

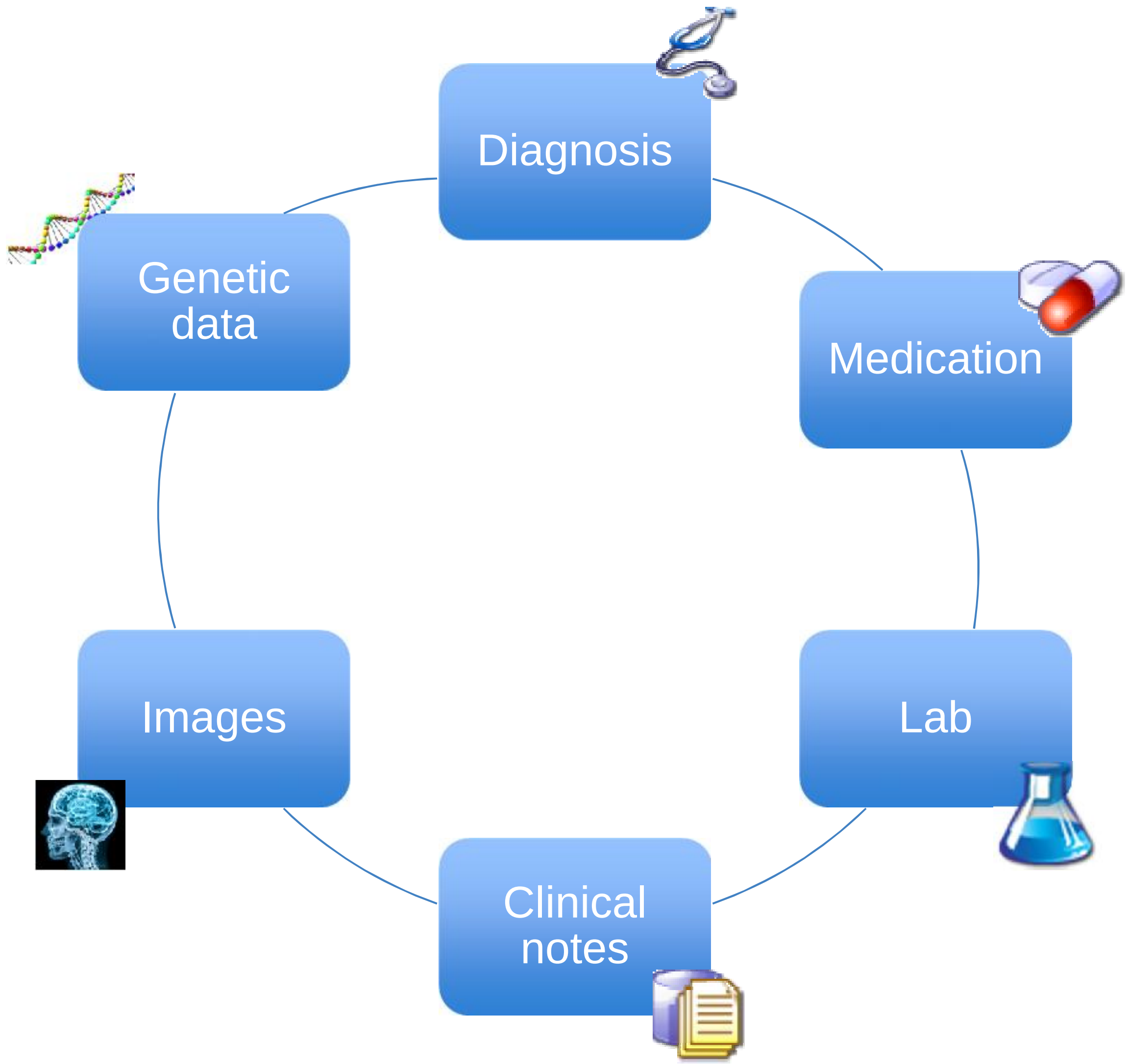
Chandan K. Reddy
Department of Computer Science
Wayne State University

Healthcare Analytics using Electronic Health Records (EHR)

- Old way: **Data are expensive and small**
  - Input data are from clinical trials, which is small and costly
  - Modeling effort is small since the data is limited
    - A single model can still take months

- EHR era: **Data are cheap and large**
  - Broader patient population
  - Noisy data
  - Heterogeneous data
  - Diverse scale
  - Complex use cases

Heterogeneous Medical Data

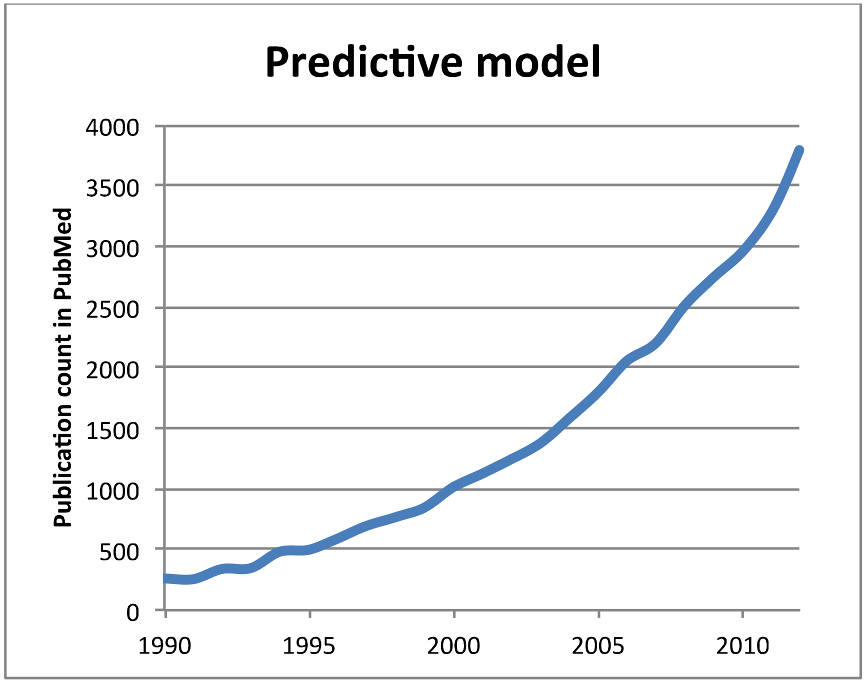


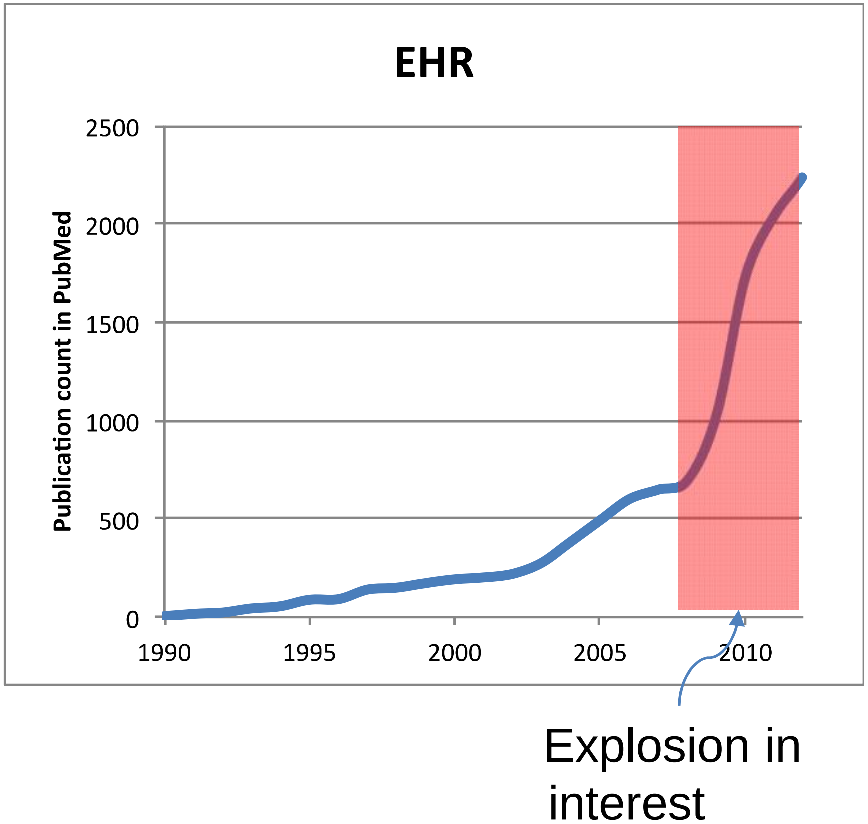
Challenges in Healthcare Analytics

- Collaboration across domains
- Analytic platform
- Intuitive results
- Scalable computation

PARALLEL MODEL BUILDING

Motivation – Predictive modeling using EHR is growing





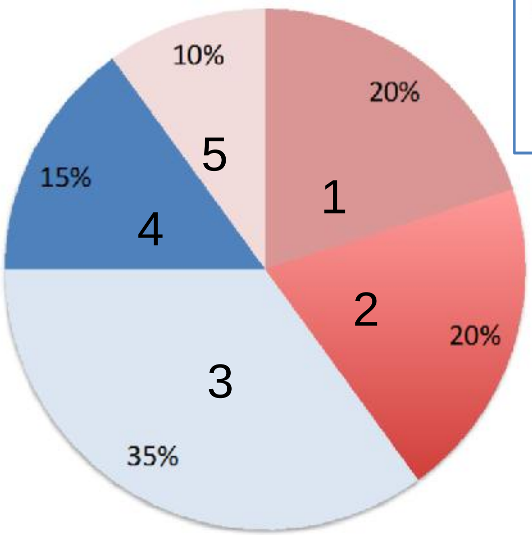
- Need for scalable predictive modeling platforms/systems due to increased computational requirements from:
  – Processing EHR data (due to volume, variability, and heterogeneity)
  – Building accurate models
  – Building clinically meaningful models
  – Validating models for accuracy and generalizability

What does it take to develop a predictive model using EHR?

Marina: IBM
Analytics Consultant

Within 3 months, we need to

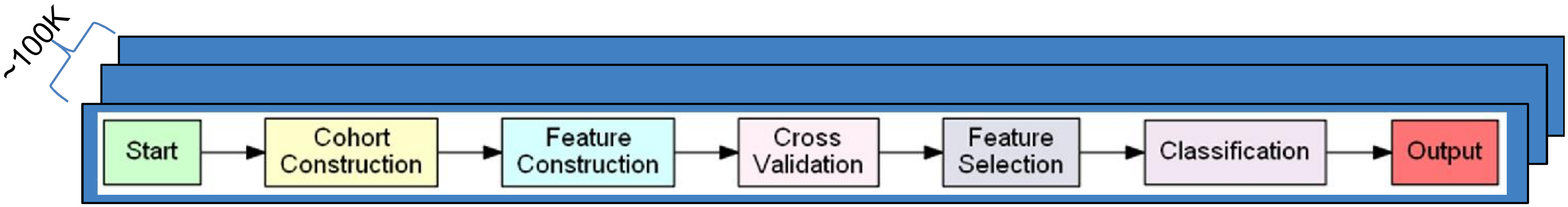
1. understand business case
2. obtain the data
3. prepare the data
4. develop predictive models
5. deliver the final model



How to help her to develop a good predictive model quickly?

David Gotz, Harry Starvropoulos, Jimeng Sun, Fei Wang.
ICDA: A Platform for Intelligent Care Delivery Analytics, AMIA 2012

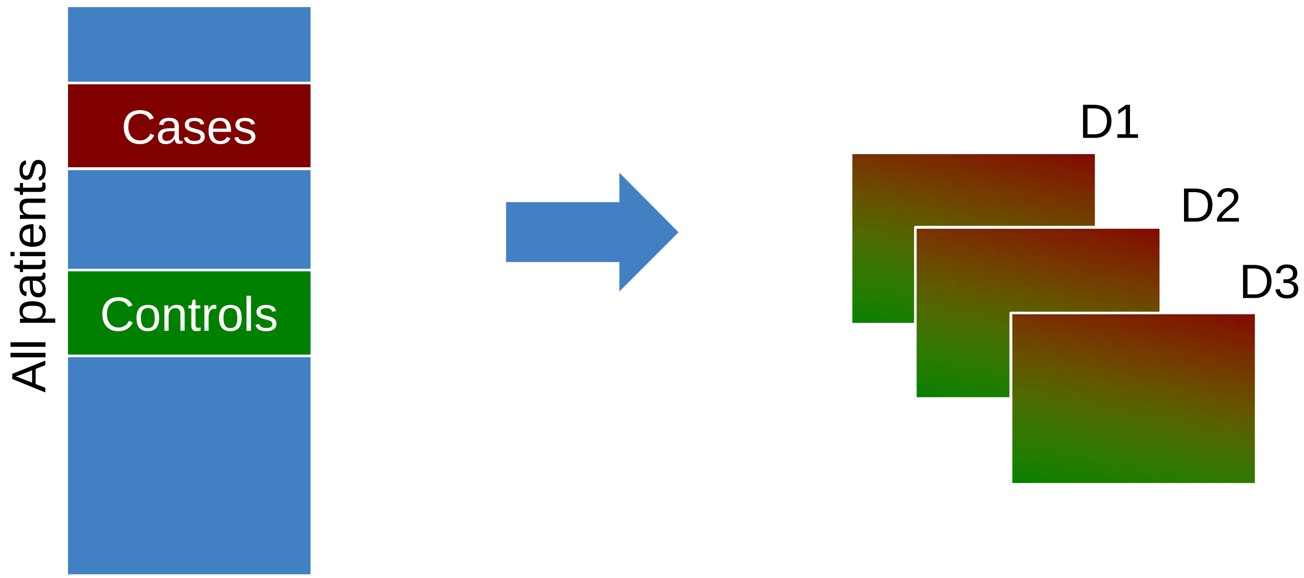
A Generalized Predictive Modeling Pipeline



Model specification

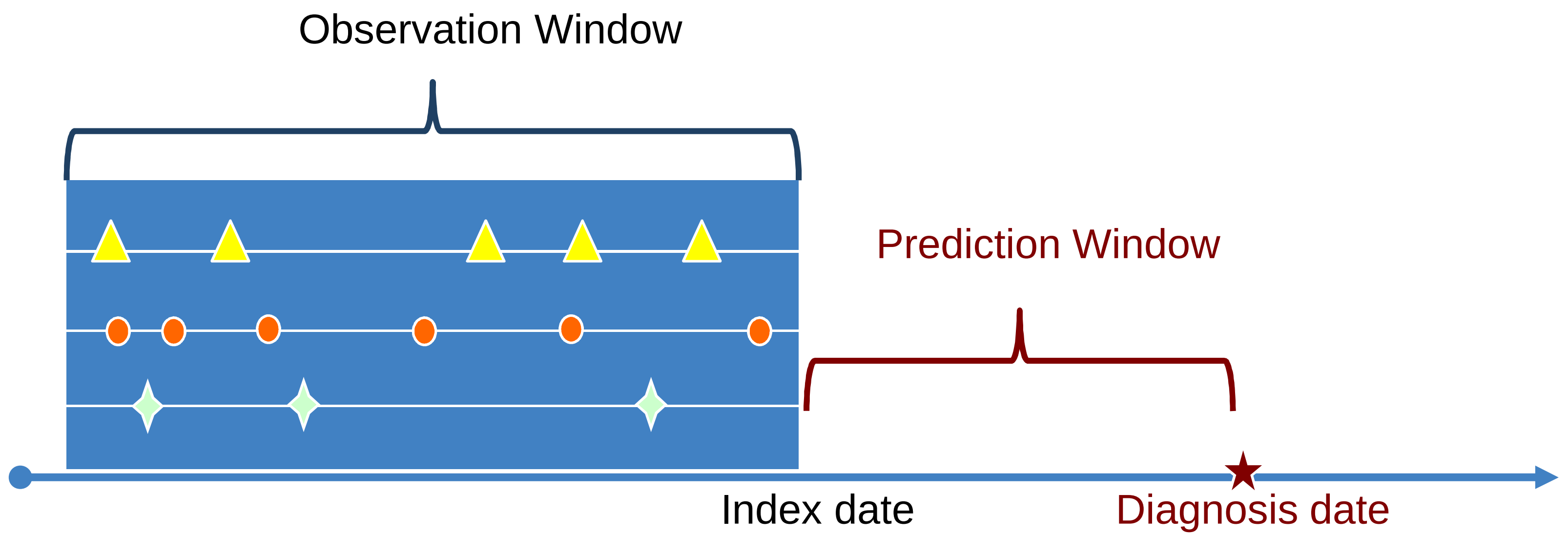
- **Cohort Construction**: Find an appropriate set of patients with the specified target condition and a corresponding set of control patients without the condition.
- **Feature Construction**: Compute a feature vector representation for each patient based on the patient's EHR data.
- **Cross Validation**: Partition the data into complementary subsets for use in model training and validation testing.
- **Feature Selection**: Rank the input features and select a subset of relevant features for use in the model.
- **Classification**: The training and evaluation of a model for a specific classifier.
- **Output**: Clean up intermediate files and to put results into their final locations.

Cohort Construction

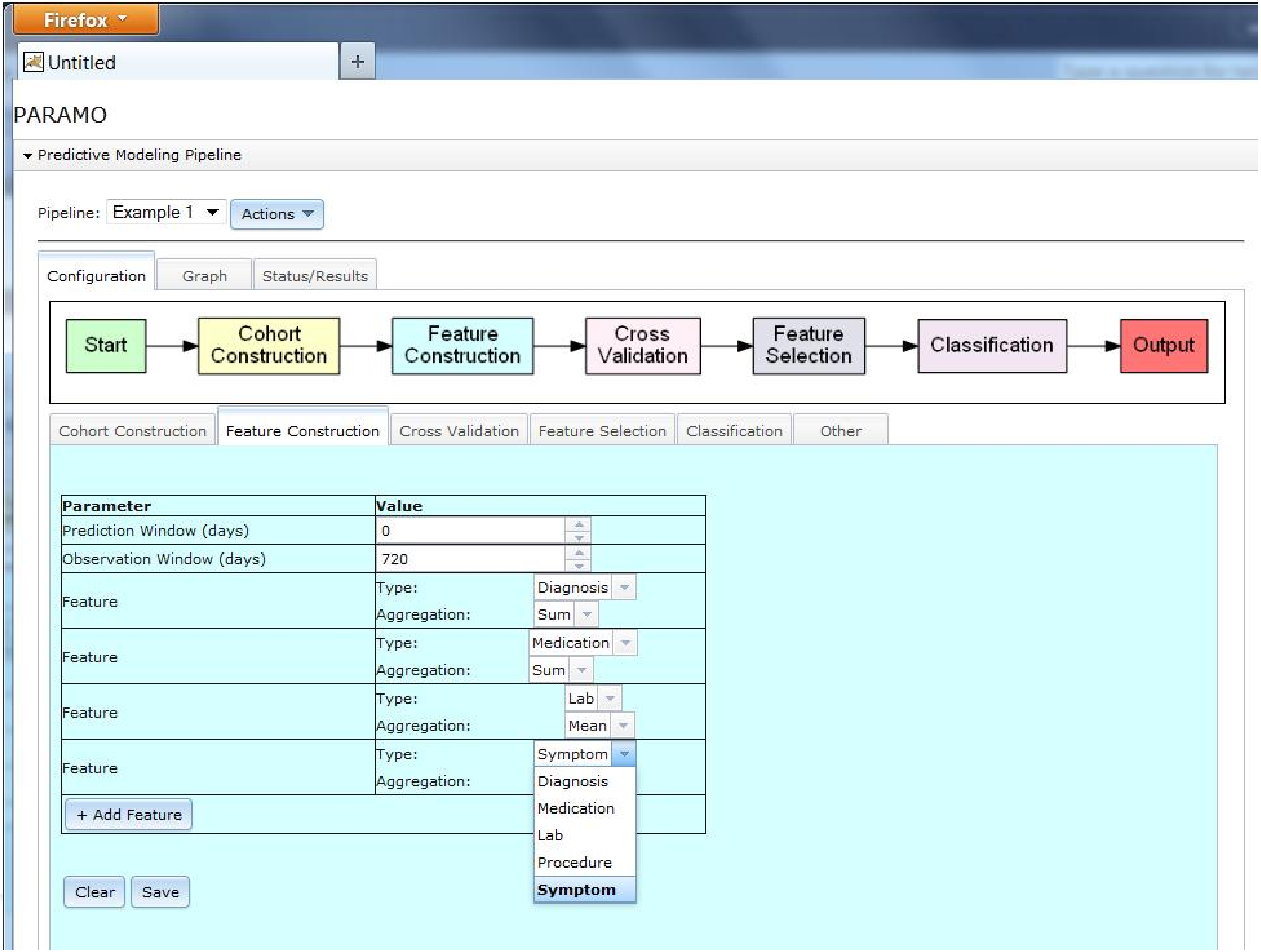


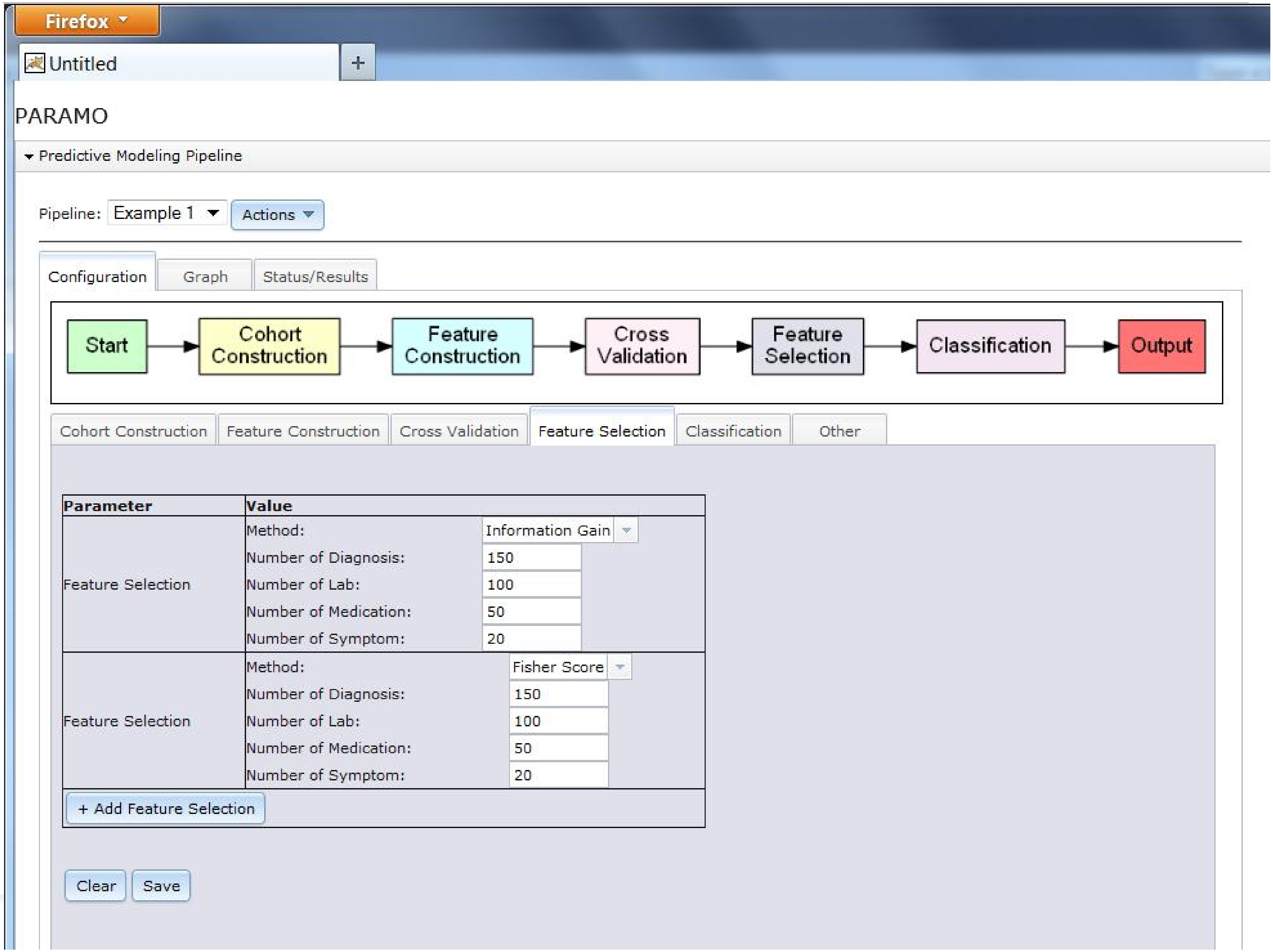
	Disease Target	samples
D1	Hypertension control	5000
D2	Heart failure onset	33K
D3	Hypertension diagnosis	300K

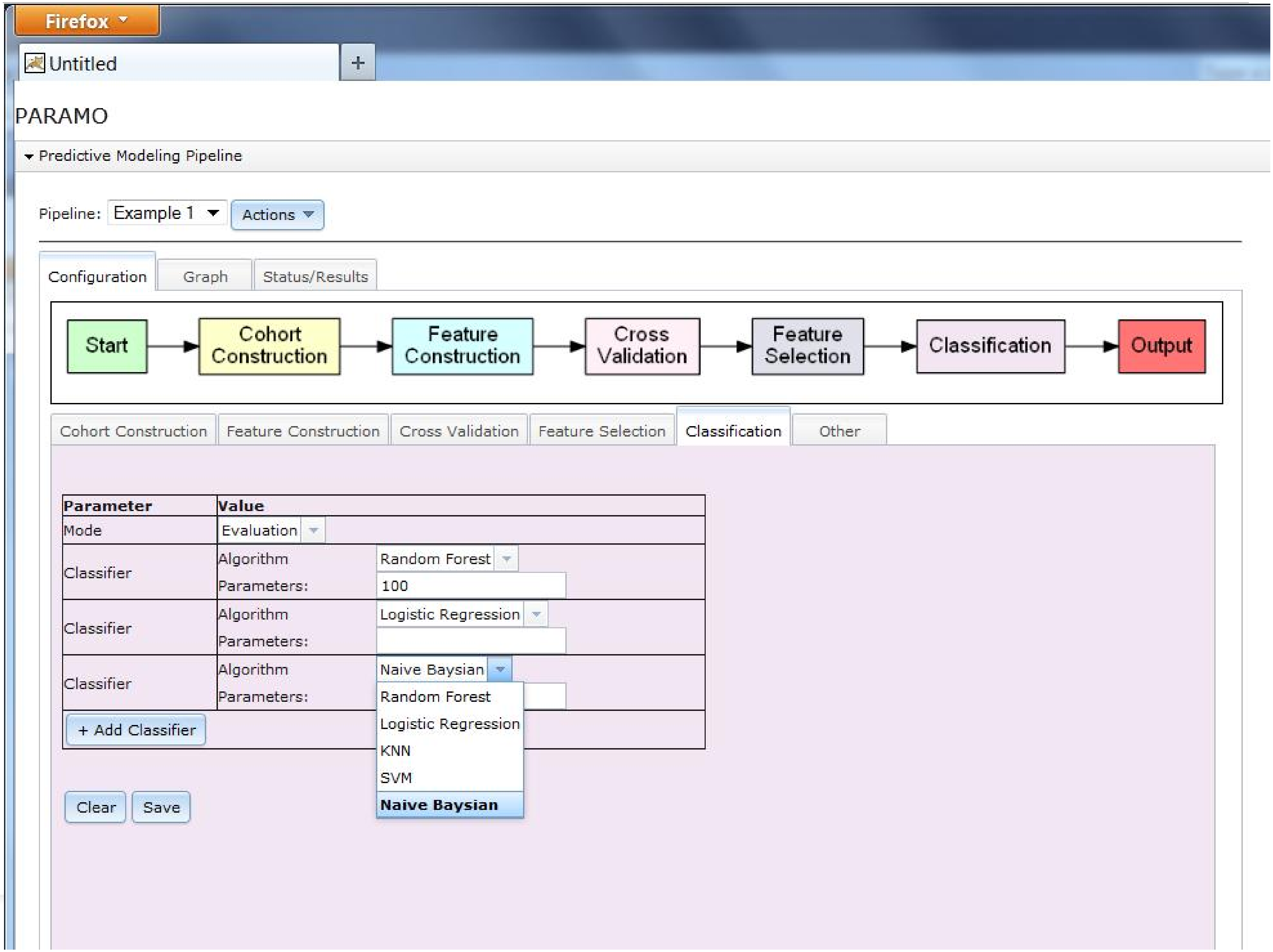
Feature Construction

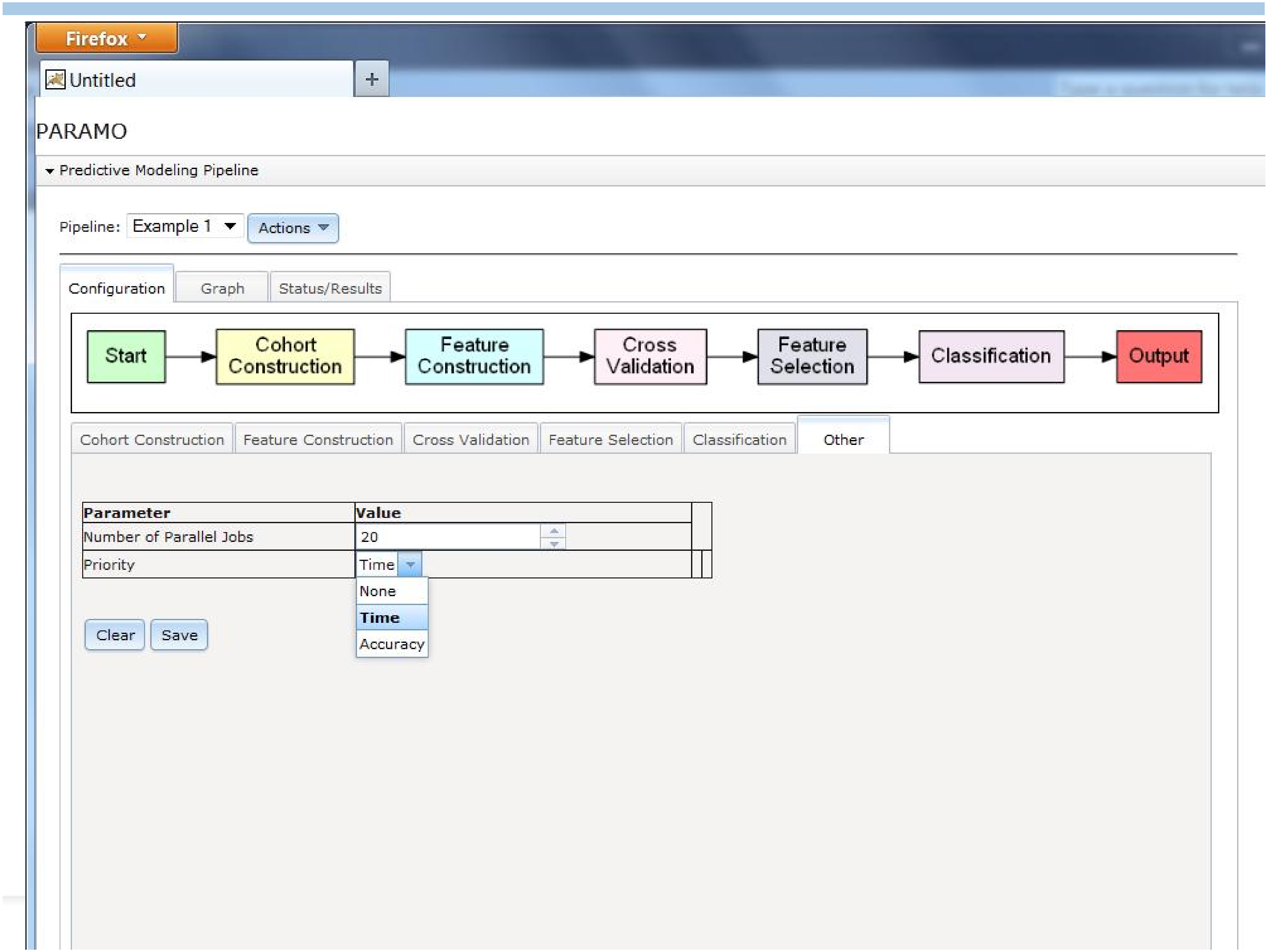


- We define
  - Diagnosis date and index date
  - Prediction and observation windows
- Features are constructed from the observation window and predict HF onset after the prediction window

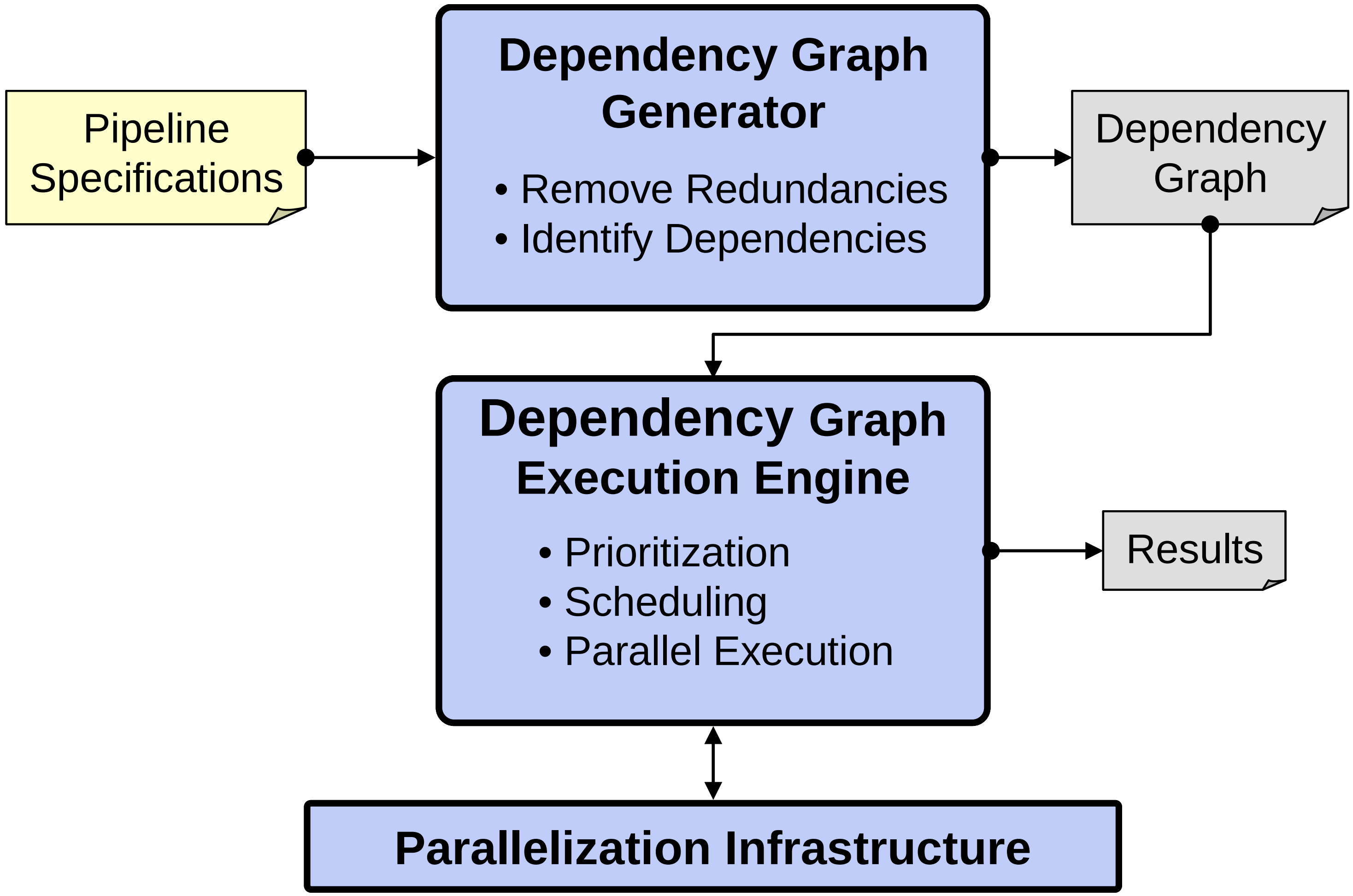

Firefox
Untitled
PARAMO
Predictive Modeling Pipeline
Pipeline: Example 1
Actions
Configuration
Graph
Status/Results
Start
Cohort Construction
Feature Construction
Cross Validation
Feature Selection
Classification
Output
Cohort Construction
Feature Construction
Cross Validation
Feature Selection
Classification
Other
Parameter
Value
Prediction Window (days)
0
Observation Window (days)
720
Feature
Type: Diagnosis
Aggregation: Sum
Feature
Type: Medication
Aggregation: Sum
Feature
Type: Lab
Aggregation: Mean
Feature
Type: Symptom
Aggregation:
Diagnosis
Medication
Lab
Procedure
Symptom
+ Add Feature
Clear
Save


Firefox
Untitled
PARAMO
Predictive Modeling Pipeline
Pipeline: Example 1
Actions
Configuration
Graph
Status/Results
Start
Cohort Construction
Feature Construction
Cross Validation
Feature Selection
Classification
Output
Cohort Construction
Feature Construction
Cross Validation
Feature Selection
Classification
Other
Parameter
Value
Feature Selection
Method: Information Gain
Number of Diagnosis: 150
Number of Lab: 100
Number of Medication: 50
Number of Symptom: 20
Feature Selection
Method: Fisher Score
Number of Diagnosis: 150
Number of Lab: 100
Number of Medication: 50
Number of Symptom: 20
+ Add Feature Selection
Clear
Save

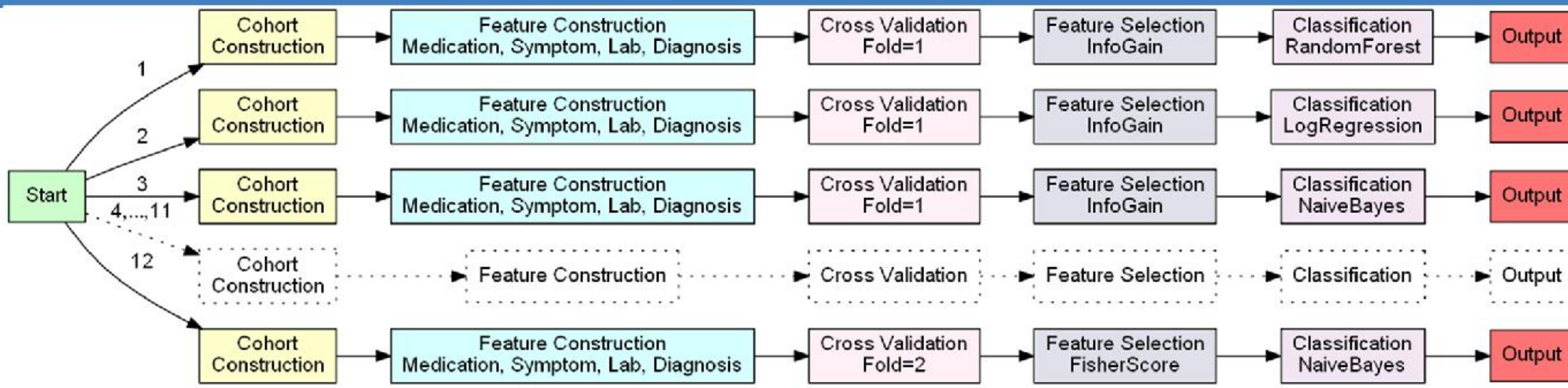

Firefox
Untitled
PARAMO
Predictive Modeling Pipeline
Pipeline: Example 1
Actions
Configuration
Graph
Status/Results
Start
Cohort Construction
Feature Construction
Cross Validation
Feature Selection
Classification
Output
Cohort Construction
Feature Construction
Cross Validation
Feature Selection
Classification
Other
Parameter
Value
Mode
Evaluation
Classifier
Algorithm
Random Forest
Parameters:
100
Classifier
Algorithm
Logistic Regression
Parameters:
Classifier
Algorithm
Naive Baysian
Parameters:
Random Forest
Logistic Regression
KNN
SVM
Naive Baysian
+ Add Classifier
Clear
Save


Firefox
Untitled
PARAMO
Predictive Modeling Pipeline
Pipeline: Example 1
Actions
Configuration
Graph
Status/Results
Start
Cohort Construction
Feature Construction
Cross Validation
Feature Selection
Classification
Output
Cohort Construction
Feature Construction
Cross Validation
Feature Selection
Classification
Other
Parameter
Value
Number of Parallel Jobs
20
Priority
Time
None
Time
Accuracy
Clear
Save

PARAMO: A Parallel Predictive Modeling Platform



An Example Set of Pipeline Specifications

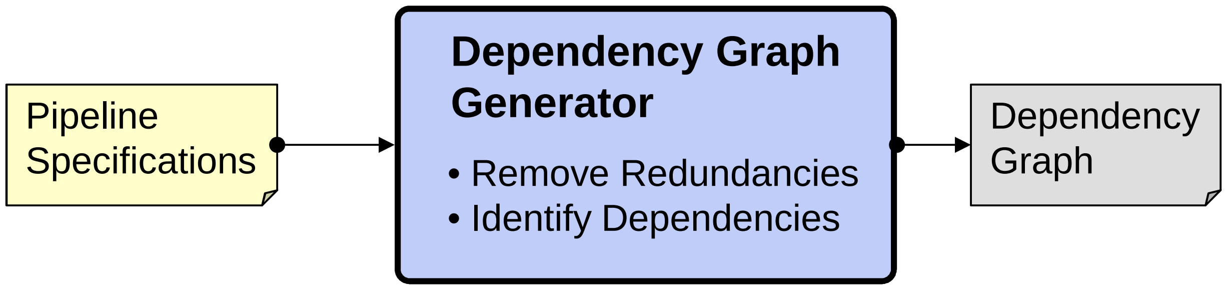


- Cohort construction: One patient data set
- Feature construction:

Feature Type	**Aggregation**
Diagnoses	Count
Medications	Count
Symptoms	Count
Labs	Mean

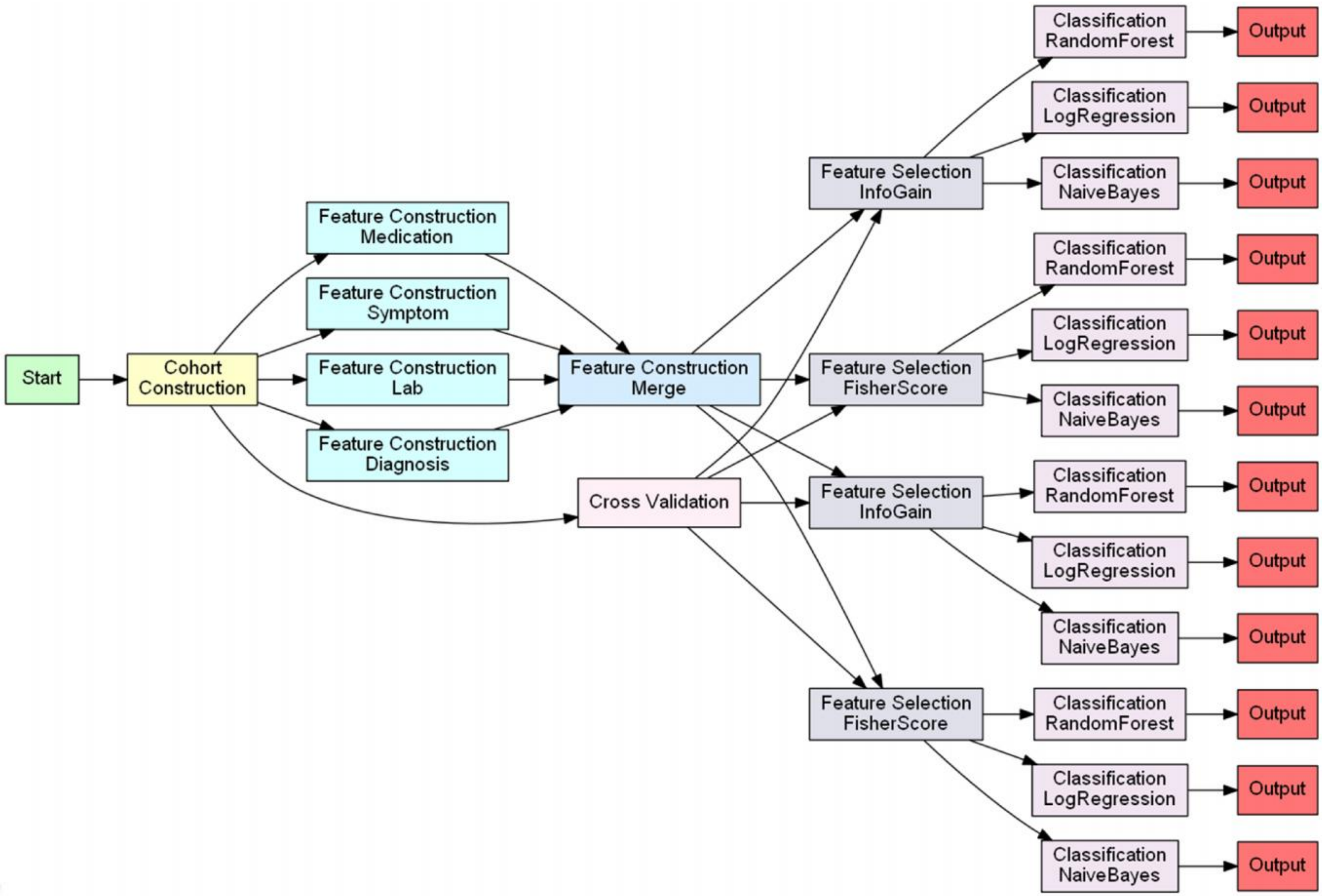
- Cross-validation: 2-fold cross-validation
- Feature selection: Information Gain, Fisher Score
- Classification: Naïve Bayes, Logistic Regression, Random Forest

Dependency Graph Generator

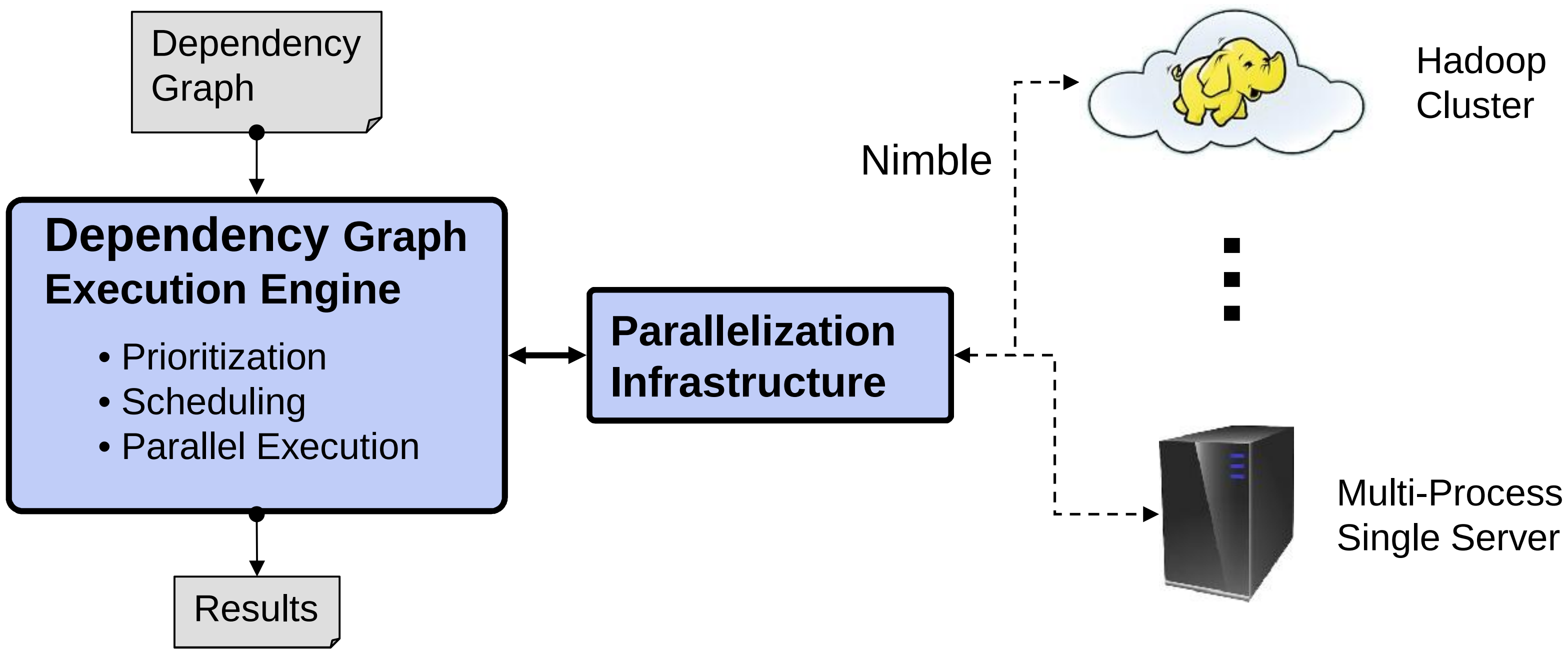


- Input: Pipeline specifications
- Output: Dependency graph
- Function:
  – Remove Redundancies
  – Identify Dependencies
  – Encode parallel jobs

Dependency Graph

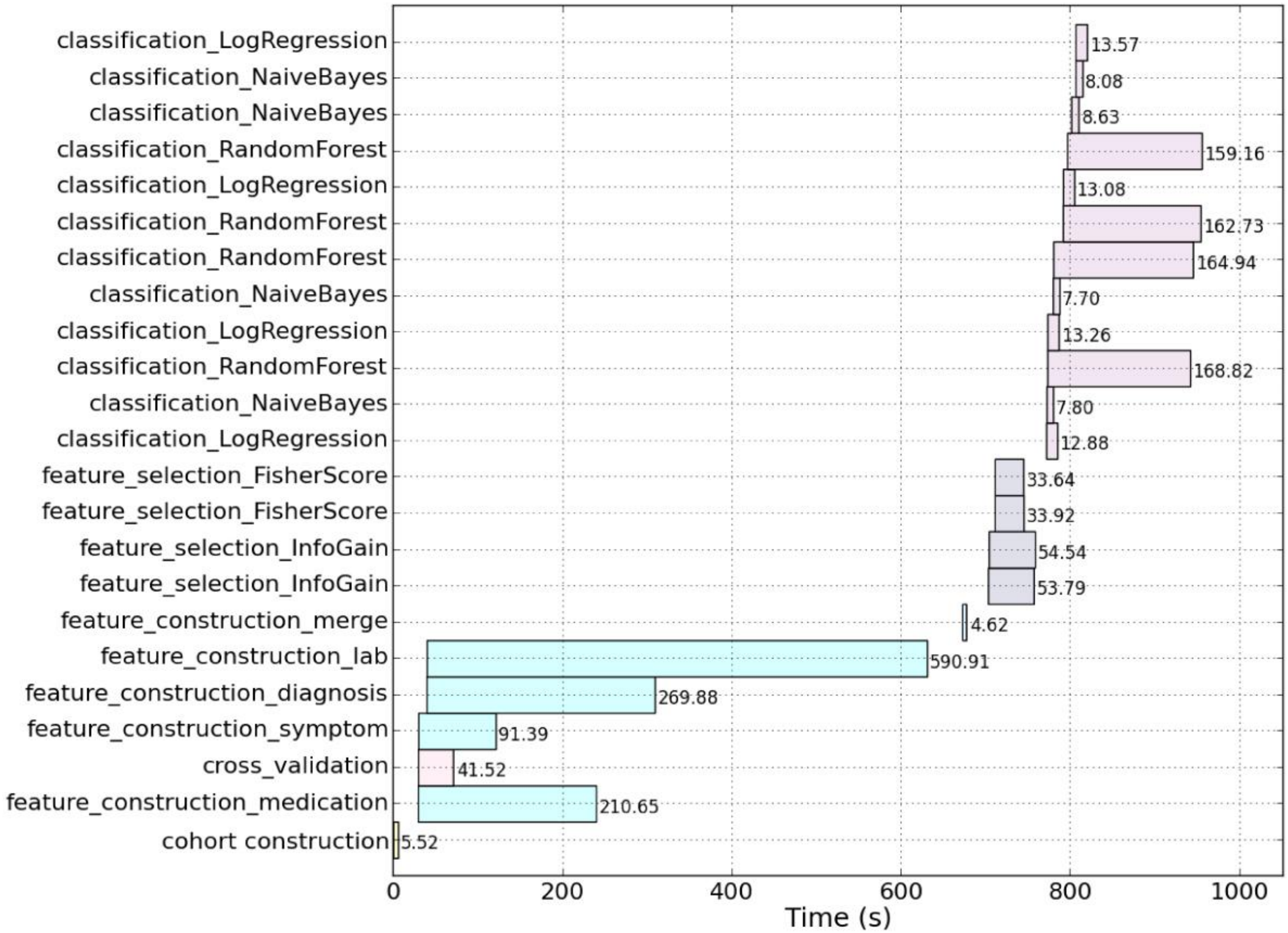


Dependency Graph Execution Engine



- Input: Dependency graph
- Output: Results (models, scores, etc.)
- Function:
  - Schedules tasks in a topological ordering of the graph
  - Prioritizes pending tasks using information from already completed tasks
  - Executes tasks in parallel via the parallelization infrastructure

Dependency Graph Runtime Analysis (Hadoop, 20 Concurrent Tasks)



Experimental Data Sets

Data Set	Years of Data	Number of Patients	Number of Features	Number of Records	Number of Cases	Number of Controls	Target Condition
Small	3	4,758	25932	3,312,558	615	949	Hypertension Control
Medium	10	32,675	46117	24,719,809	4644	28031	Heart Failure Onset
Large	4	319,650	49269	33,531,311	16385	164743	Hypertension Onset

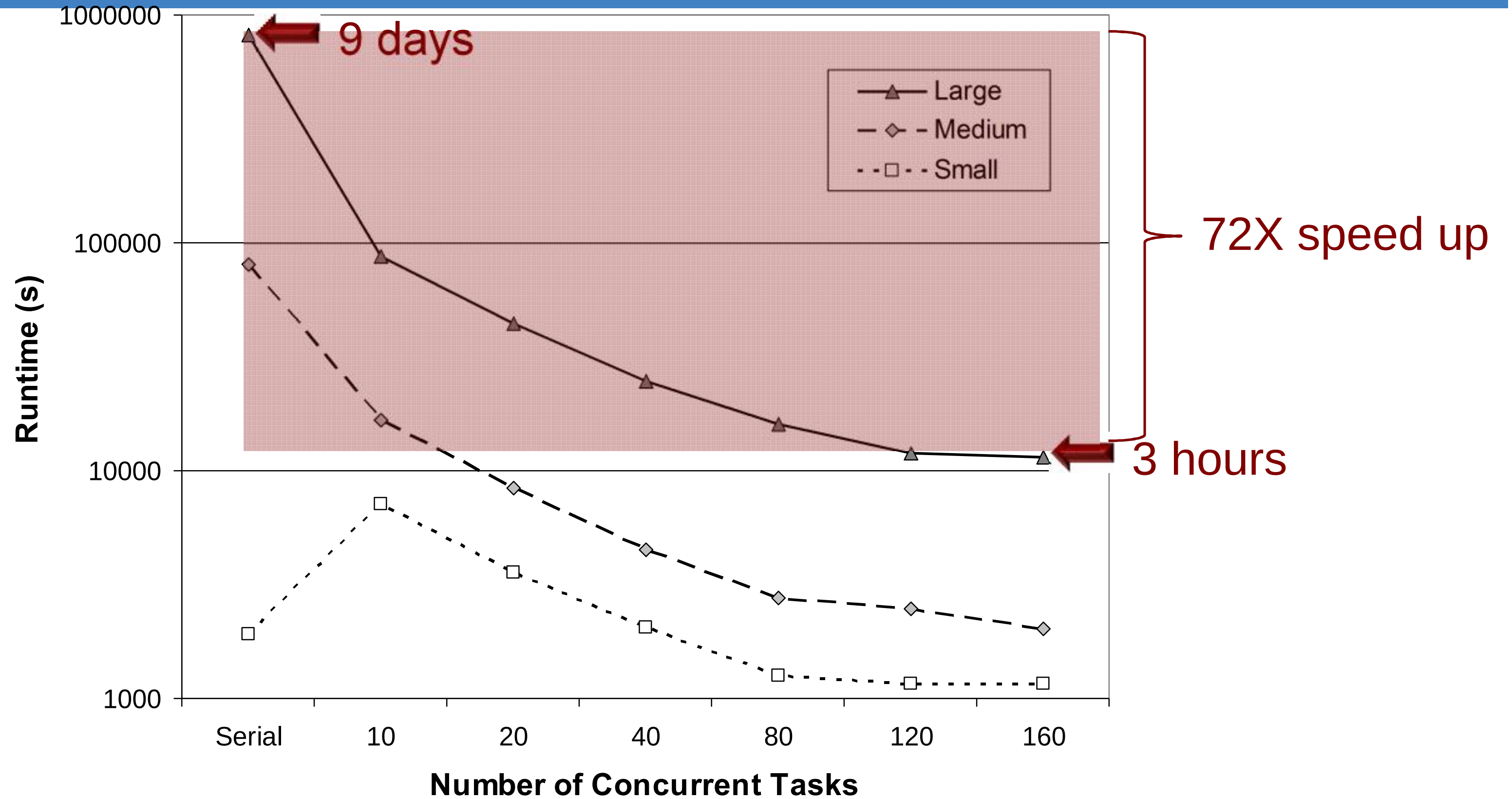
Experimental Pipeline Specifications

- Cohort construction: One patient data set: Small, Medium, or Large
- Feature construction:

Feature Type	Aggregation
Diagnoses	Count
Medications	Count
Procedures	Count
Symptoms	Count
Labs	Mean

- Cross-validation: 10 x 10-fold cross-validation
- Feature selection: Information Gain, Fisher Score
- Classification: k-NN, Naïve Bayes, Logistic Regression, Random Forest

Running Time vs. Parallelism level



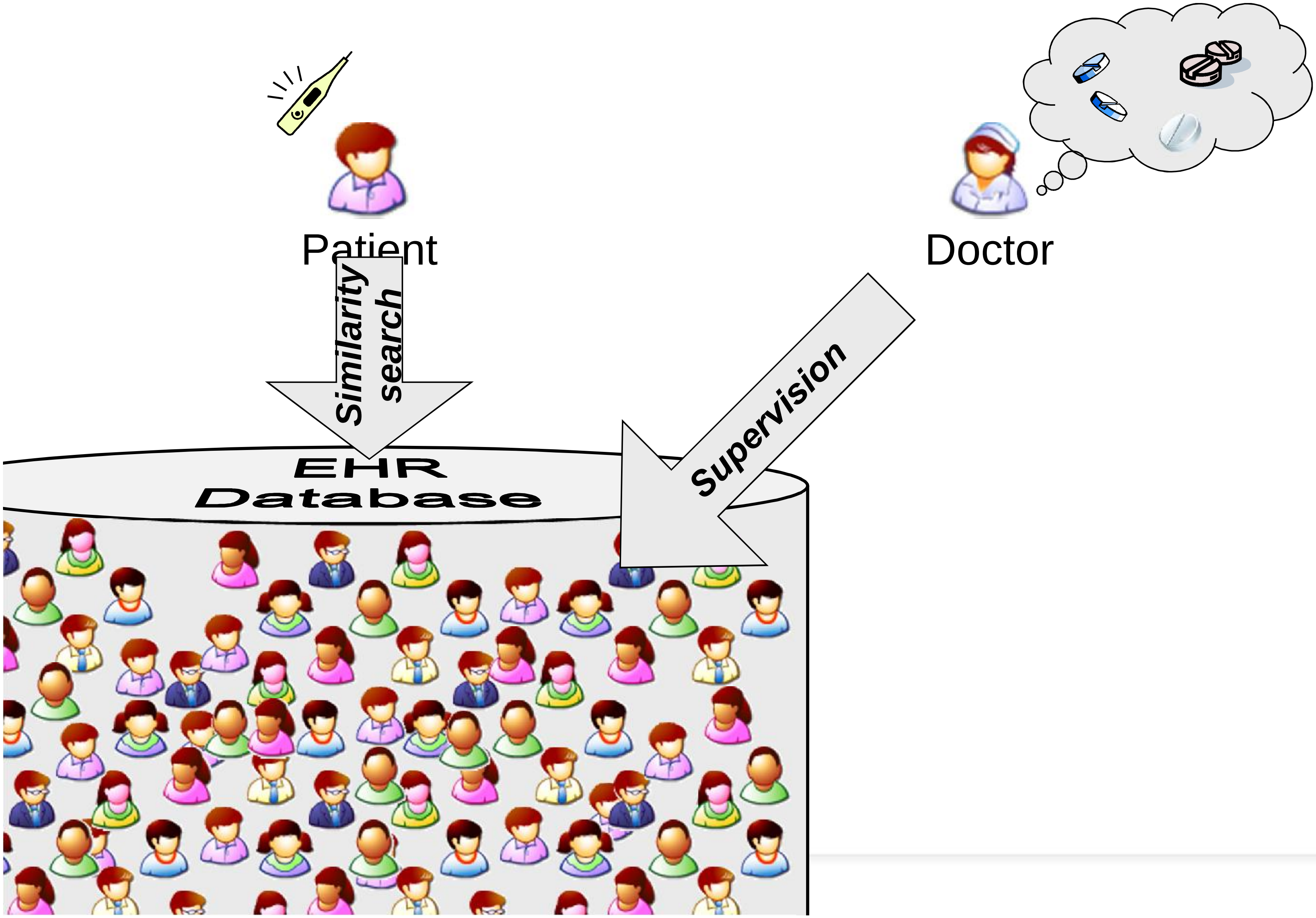
- Small: 5,000 patients, Medium: 33K, Large: 319K
- 10 times 10-fold cross validation
- Dependency graph: 1808 nodes and 3610 edges

Summary

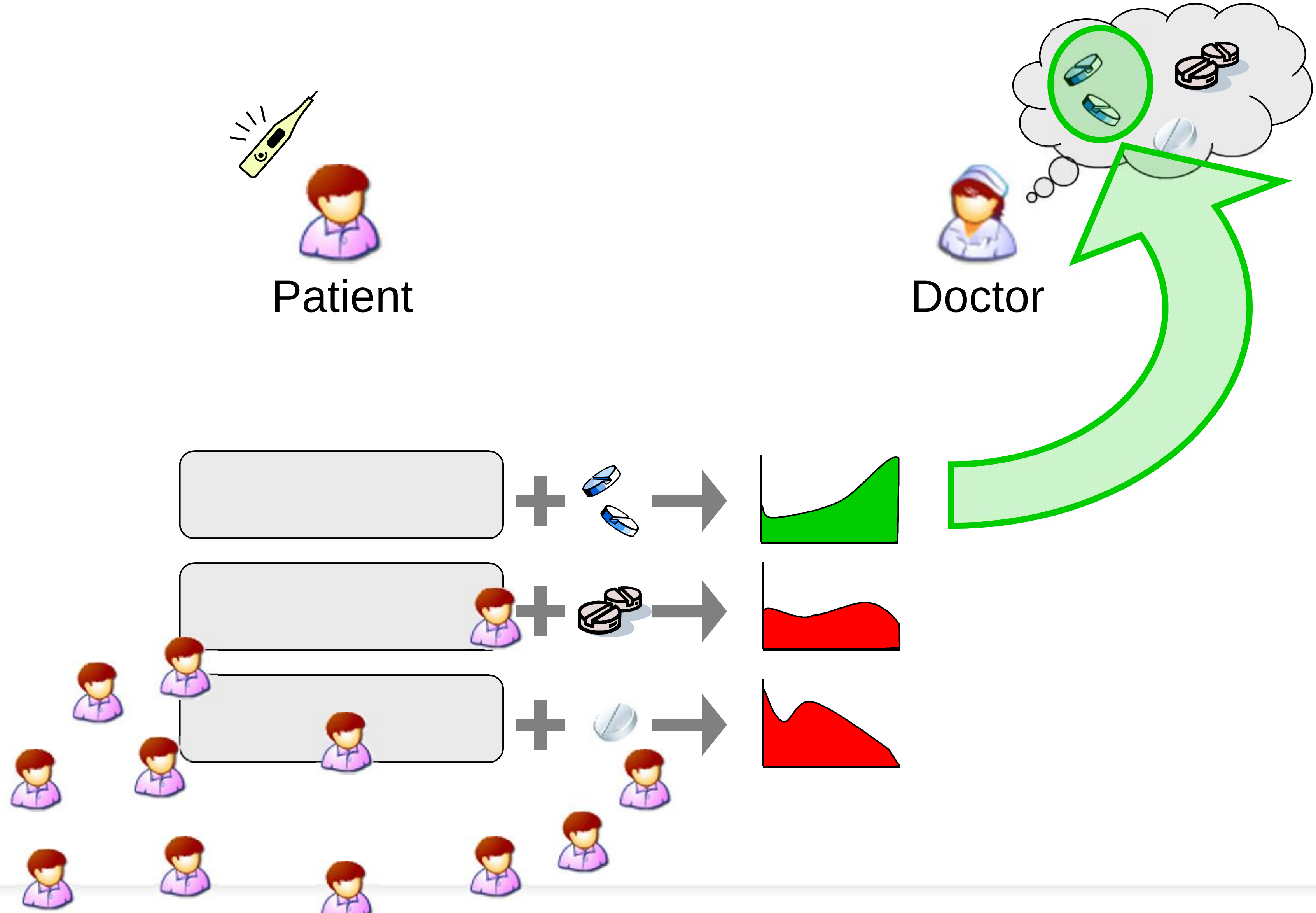
- Predictive models in healthcare research is becoming more prevalent
- Electronic health records (EHR) adoption continues to accelerate
- Need for scalable predictive modeling platforms/systems
- PARAMO is a parallel predictive modeling platform for EHR data
- PARAMO can facilitate large-scale modeling endeavors and speed-up the research workflow
- Tests on real EHR data show significant performance gains

PATIENT SIMILARITY PLATFORM

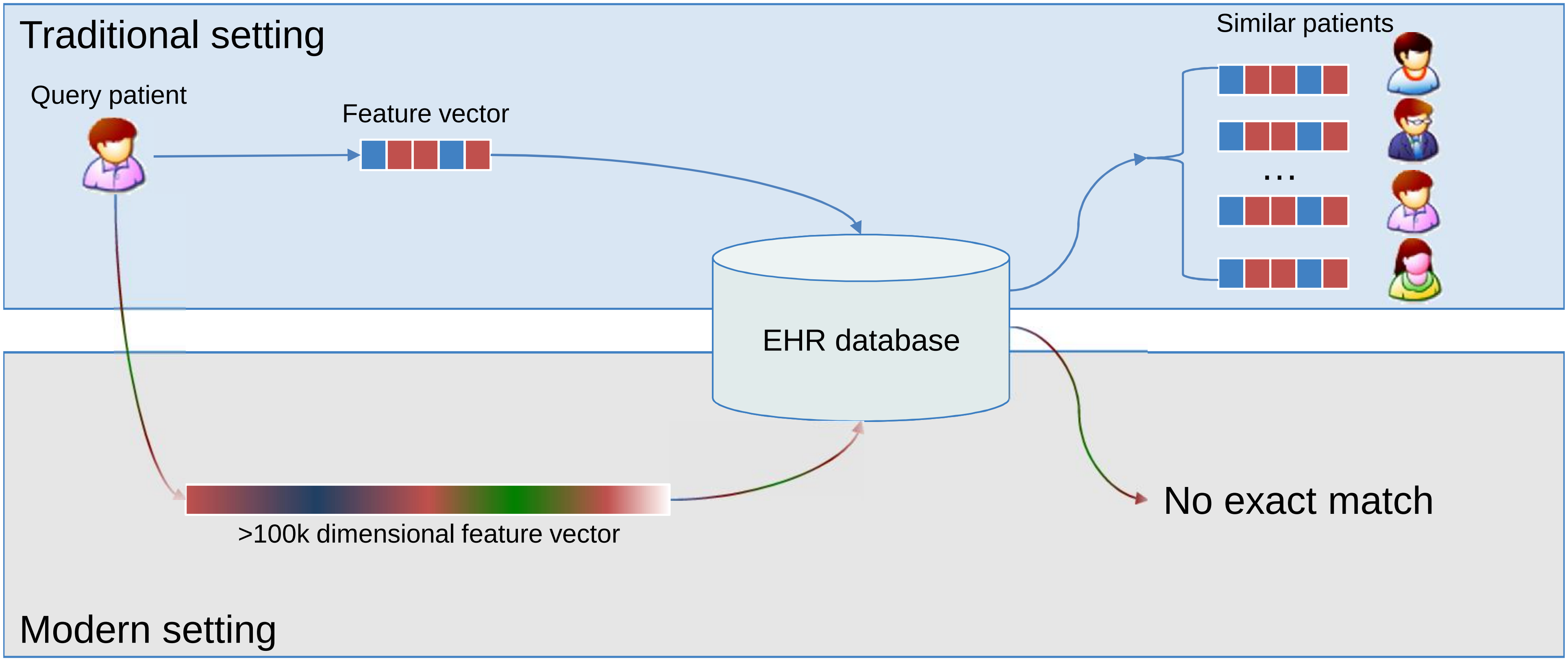
Patient Similarity Problem



Patient Similarity Problem

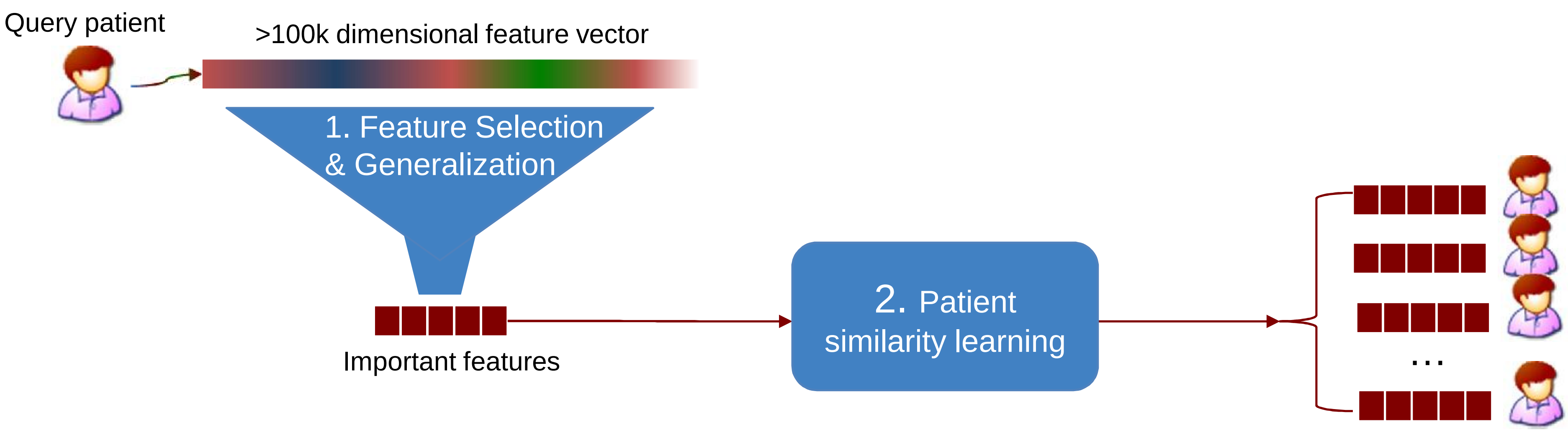


Challenges of Patient Similarity



- As the size of a feature vector increases, it is hard to find exact match on all features
- How to find relevant patients to a query for a specific clinical context?

Our Approach



- For a clinical context,
  1. What are important features?
  2. What is the right similarity measure?

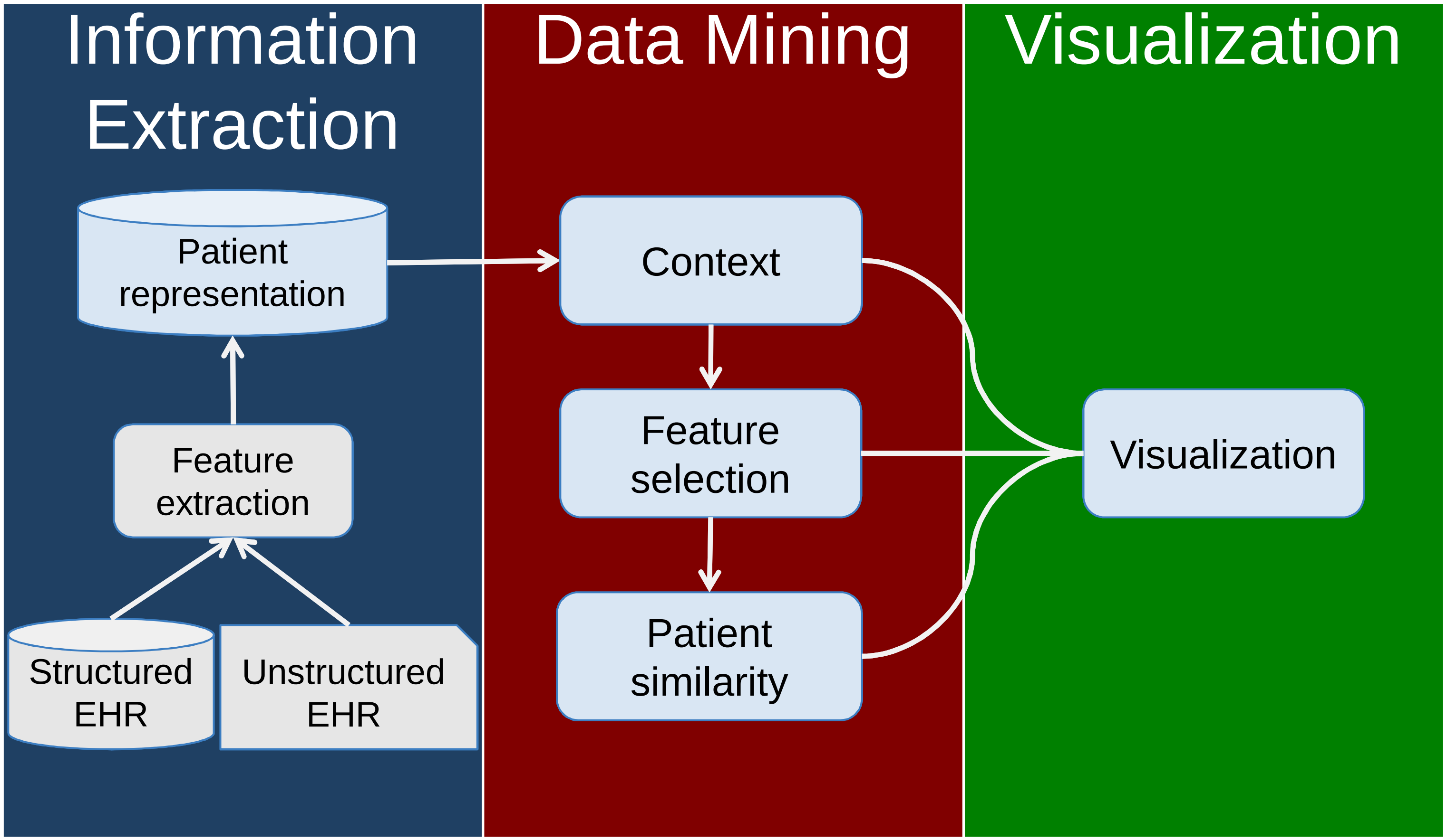
Healthcare Analytic Platform

Large-scale Analytics Platform

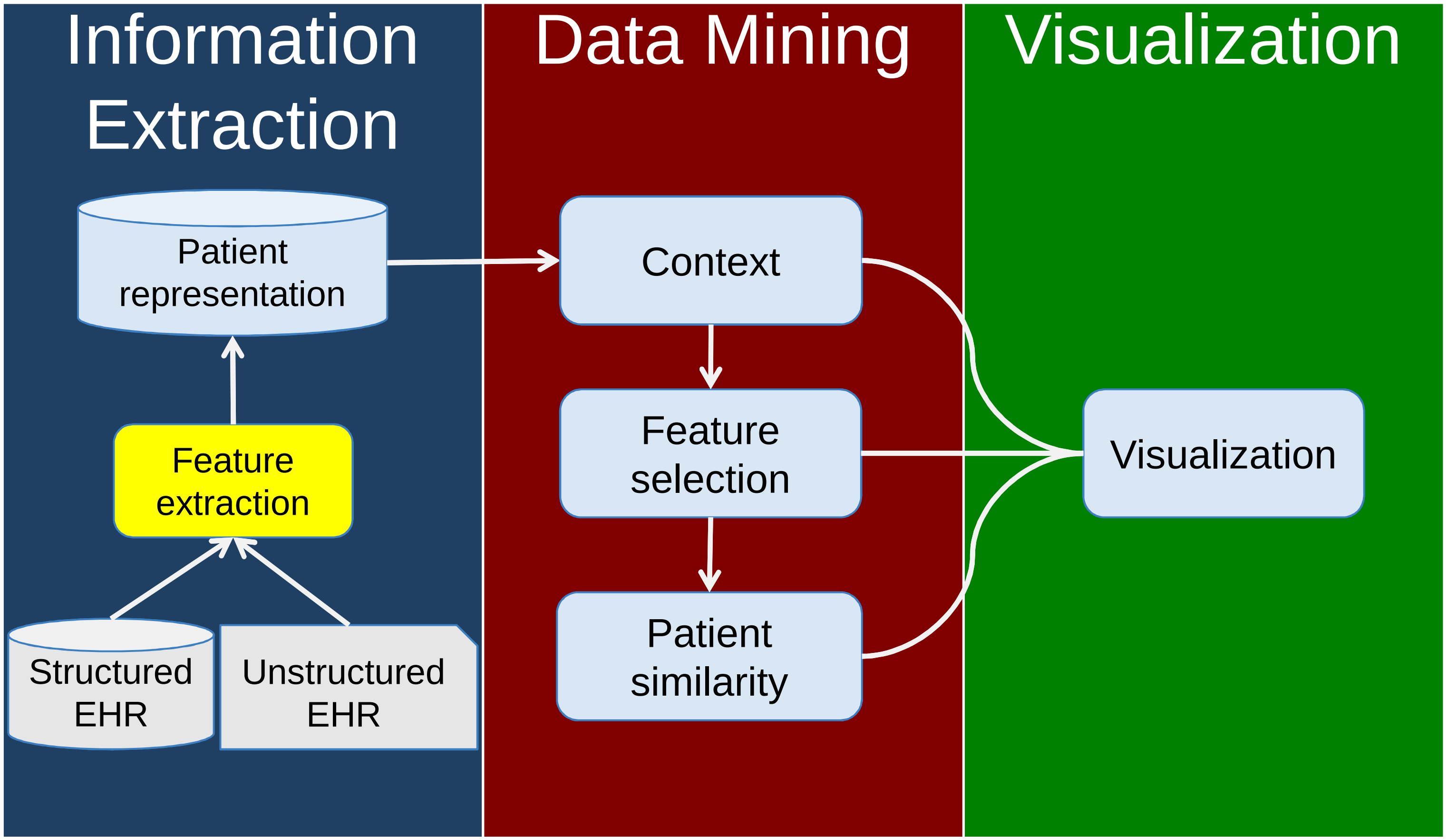
Healthcare Analytic Platform

Information Extraction	Data Mining	Visualization

Healthcare Analytic Platform



Feature Extraction from Unstructured EHR data



Motivations for Early Detection of Heart Failure

- Heart failure (HF) is a complex disease
- Huge Societal Burden

5 millions

0.5 millions

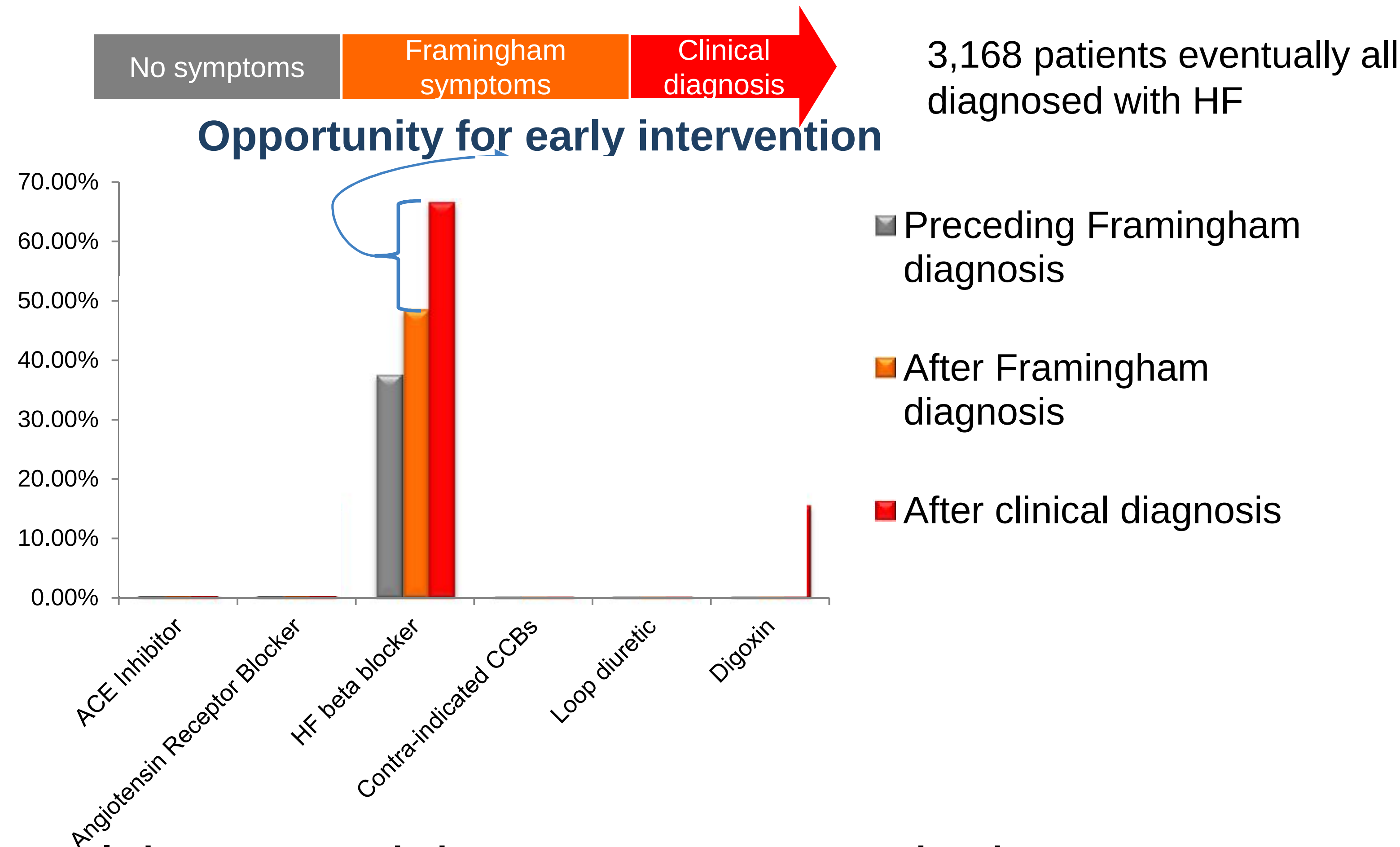
20%

48%

- Diagnoses are usually made late, despite there are symptoms documented in clinical notes prior
- Our method exacting HF symptoms achieves precision 0.925, recall 0.896, and *F*-score 0.910

Roy J. Byrd, Steven R. Steinhubl, Jimeng Sun, Shahram Ebadollahi. Automatic identification of heart failure diagnostic criteria, using text analysis of clinical notes from electronic health records. International Journal of Medical Informatics 2013

Potential Impact on Evidence-based Therapies

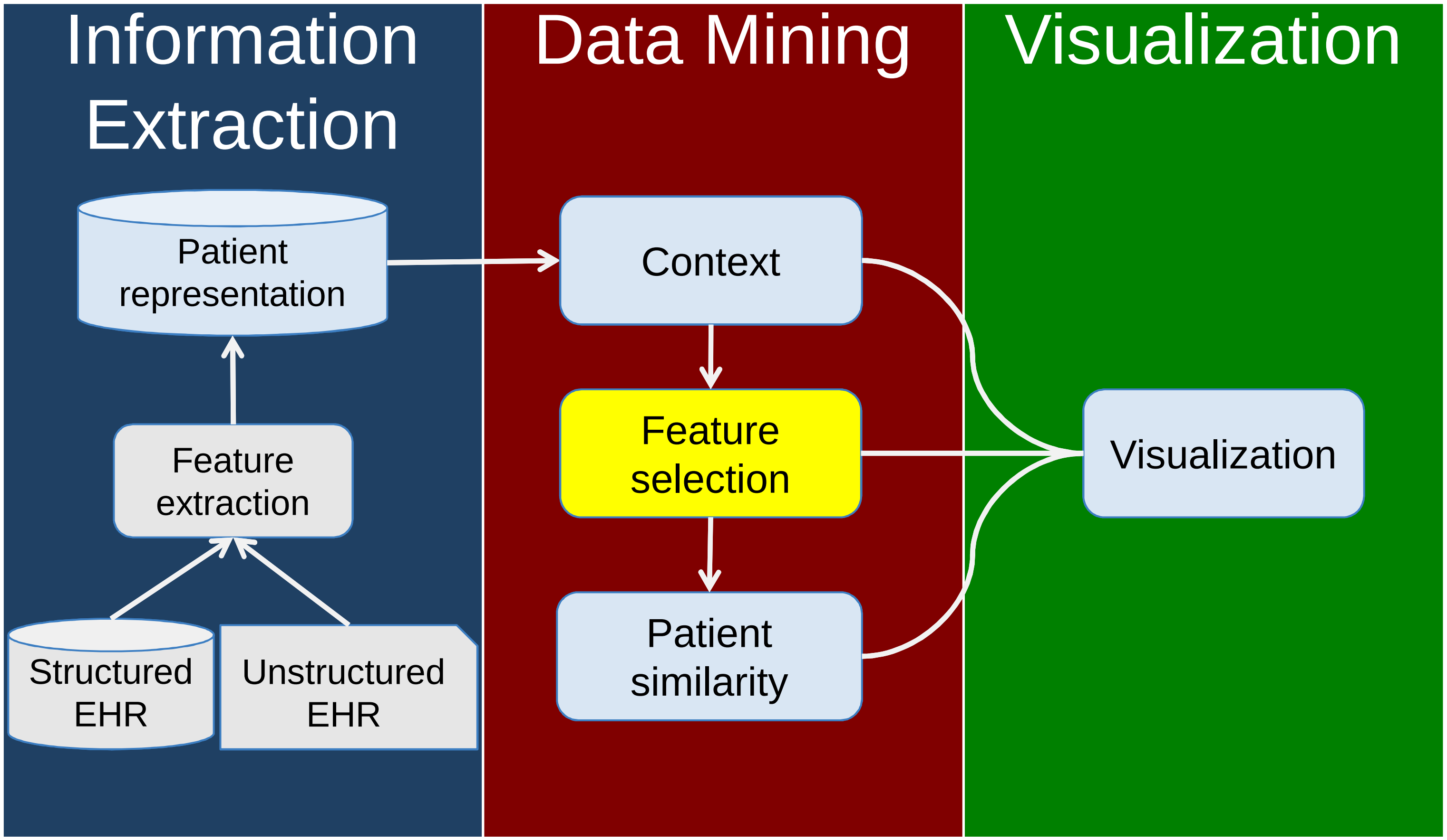


3,168 patients eventually all diagnosed with HF

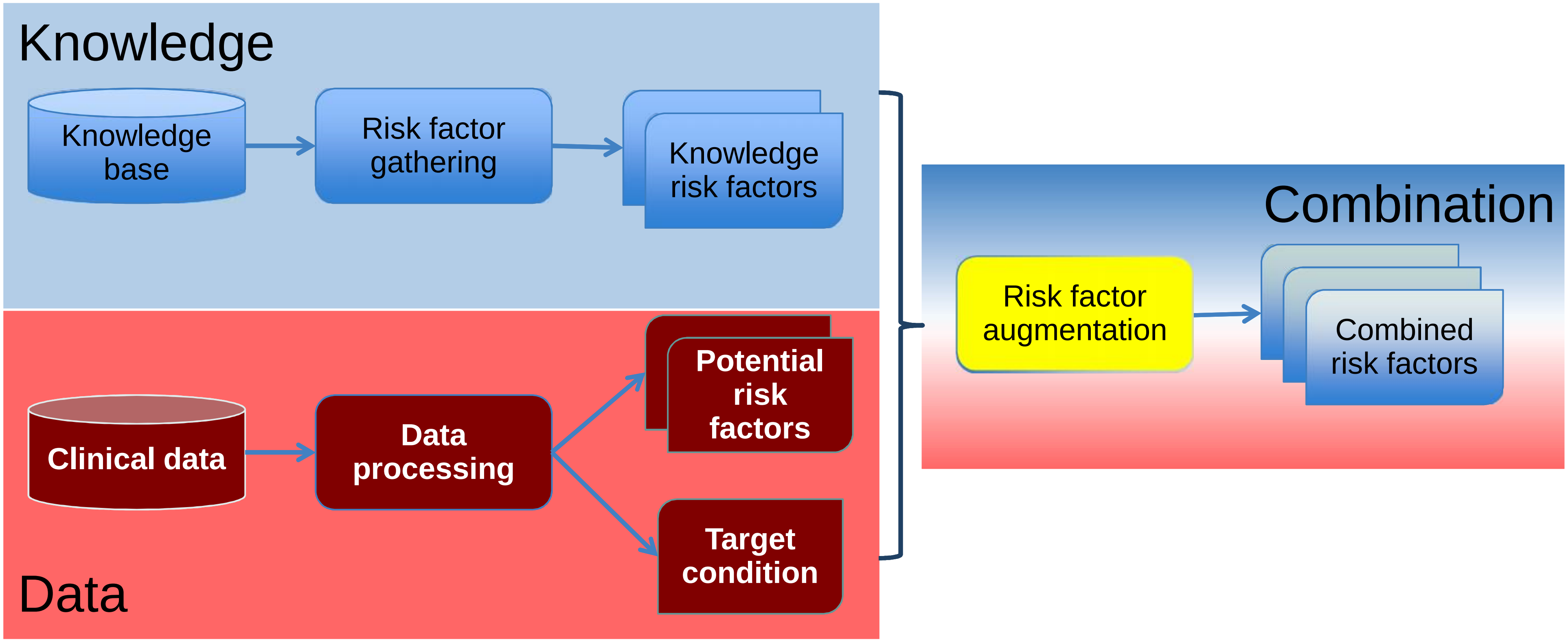
- Applying text mining to extract Framingham symptoms can help trigger early intervention

Vhavakrishnan R, Steinhubl SR, Sun J, et al. Potential impact of predictive models for early detection of heart failure on the initiation of evidence-based therapies. J Am Coll Cardiol. 2012;59(13s1):E949-E949.

Knowledge plus Data Feature Selection



Combining Knowledge- and Data-driven Risk Factors



Jimeng Sun, Jianying Hu, Dijun Luo, Marianthi Markatou, Fei Wang, Shahram Ebadollahi, Steven E. Steinhubl, Zahra Daar, Walter F. Stewart. Combining Knowledge and Data Driven Insights for Identifying Risk Factors using Electronic Health Records. AMIA2012

Dijun Luo, Fei Wang, Jimeng Sun, Marianthi Markatou, Jianying Hu,Shahram Ebadollahi, SOR: Scalable Orthogonal Regression for Low-Redundancy Feature Selection and its Healthcare Applications. SDM'12

Risk Factor Augmentation

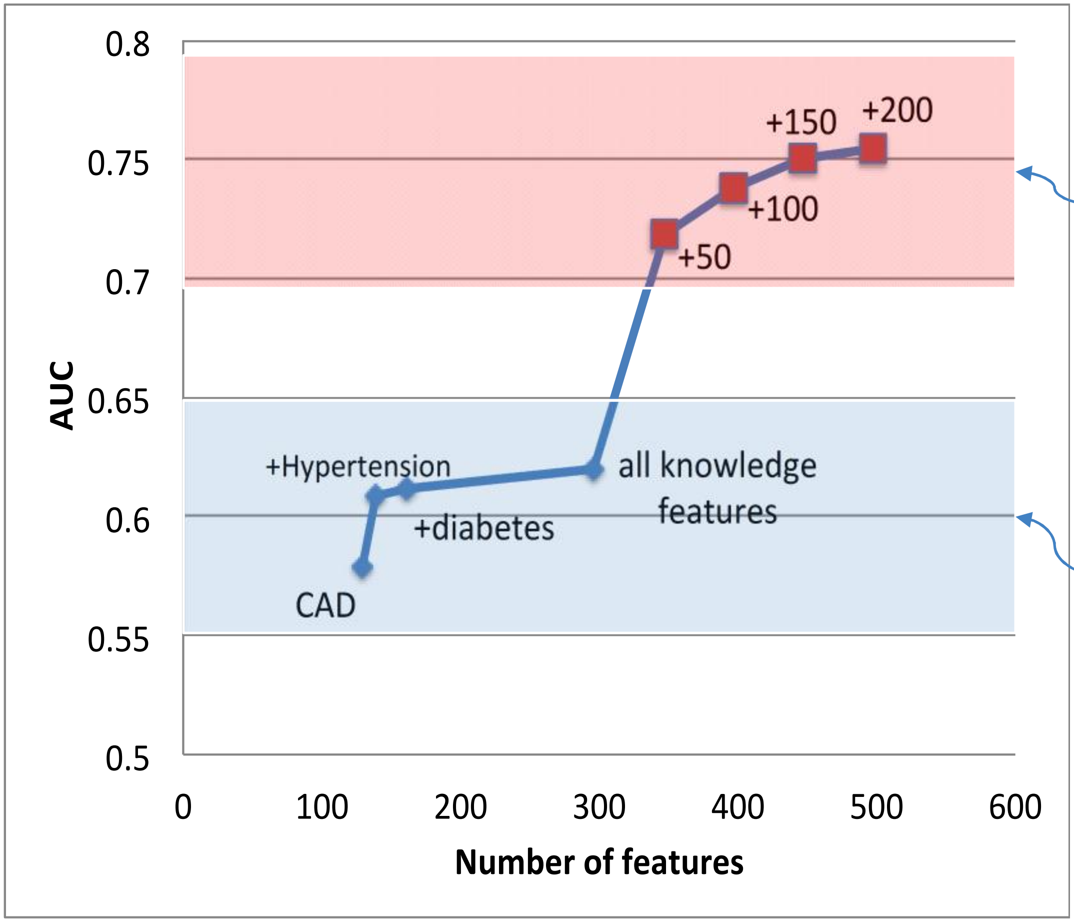
- Sparse learning objective formulation:

$$f(\boldsymbol{\alpha}) = \underbrace{\frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\alpha}\|^2}_{\text{Model error}}$$

Jimeng Sun, Jianying Hu, Dijun Luo, Marianthi Markatou, Fei Wang, Shahram Ebadollahi, Steven E. Steinhubl, Zahra Daar, Walter F. Stewart. Combining Knowledge and Data Driven Insights for Identifying Risk Factors using Electronic Health Records. AMIA2012

Dijun Luo, Fei Wang, Jimeng Sun, Marianthi Markatou, Jianying Hu,Shahram Ebadollahi, SOR: Scalable Orthogonal Regression for Low-Redundancy Feature Selection and its Healthcare Applications. SDM'12

Prediction Results using Selected Features



Data driven features improve AUC significantly

Knowledge driven features improve AUC slightly

Top-10 Selected Data-driven Features

Feature	Relevancy to HF
Dyslipidemia	✓
Thiazides-like Diuretics	✓
Antihypertensive Combinations	✓
Aminopenicillins	✓
Bone density regulators	✗
Naturietic Peptide	✓
Rales	✓
Diuretic Combinations	✓
S3Gallop	✓
NSAIDS	✓

- 9 out of 10 are considered relevant to HF

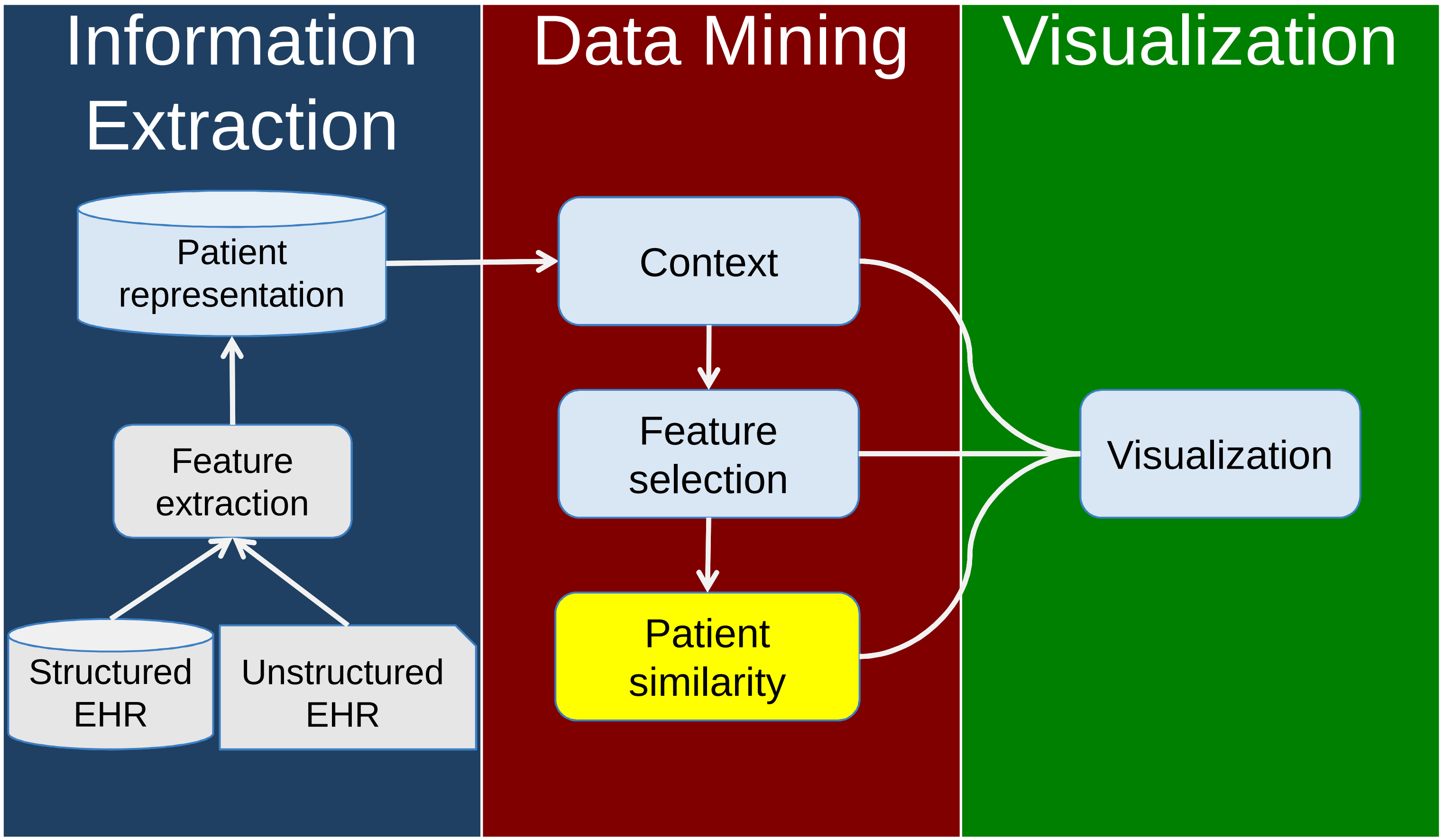
Top-10 Selected Data-driven Features

Category
Diagnosis
Medication
Lab
Symptom

Feature	Relevancy to HF
Dyslipidemia	✔
Thiazides-like Diuretics	✔
Antihypertensive Combinations	✔
Aminopenicillins	✔
Bone density regulators	✖
Naturietic Peptide	✔
Rales	✔
Diuretic Combinations	✔
S3Gallop	✔
NSAIDS	✔

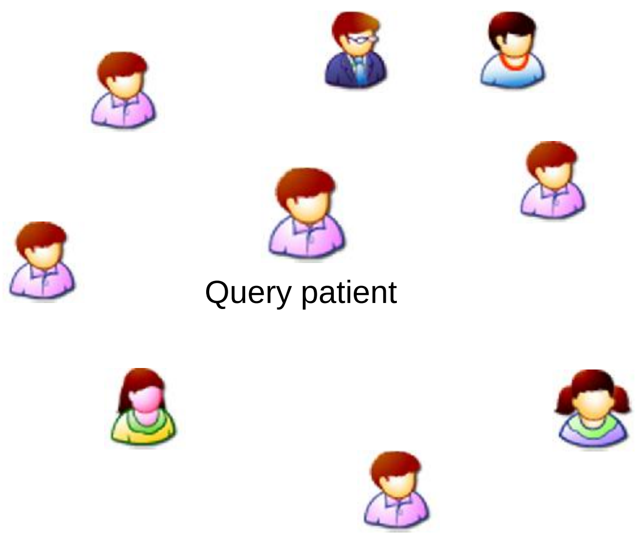
- 9 out of 10 are considered relevant to HF
- The data driven features are complementary to the existing knowledge-driven features

Patient Similarity



Patient Similarity through Locally Supervised Metric Learning

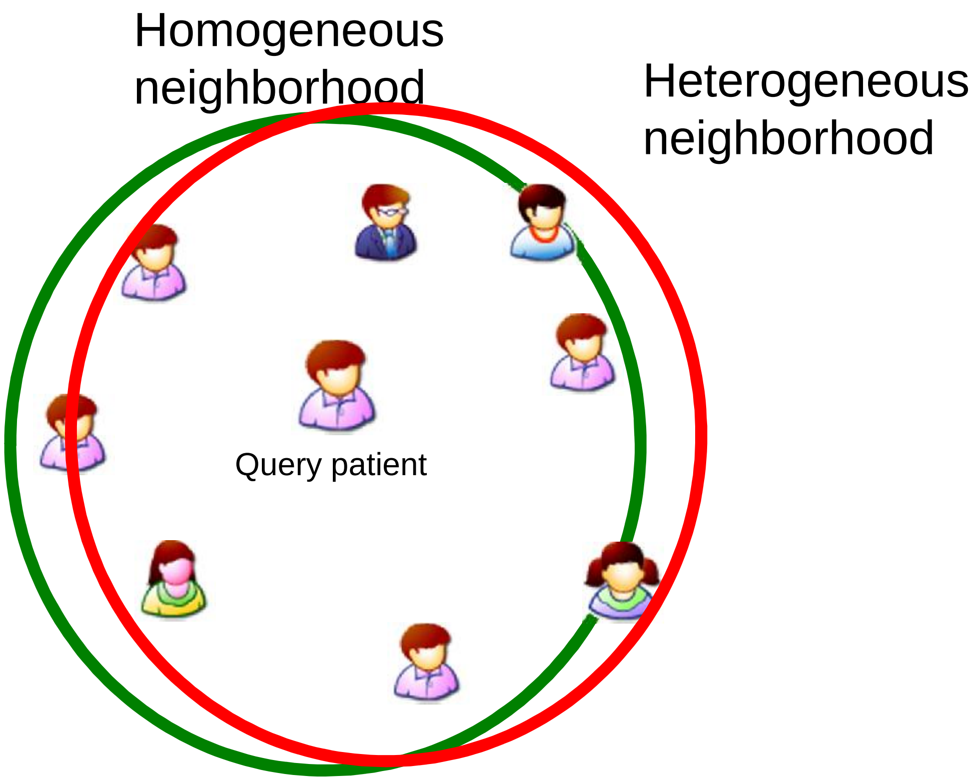
Under a specific clinical context



Jimeng Sun, Fei Wang, Jianying Hu, Shahram Edabollahi: Supervised patient similarity measure of heterogeneous patient records. SIGKDD Explorations 14(1): 16-24 (2012)

Patient Similarity through Locally Supervised Metric Learning

Under a specific clinical context

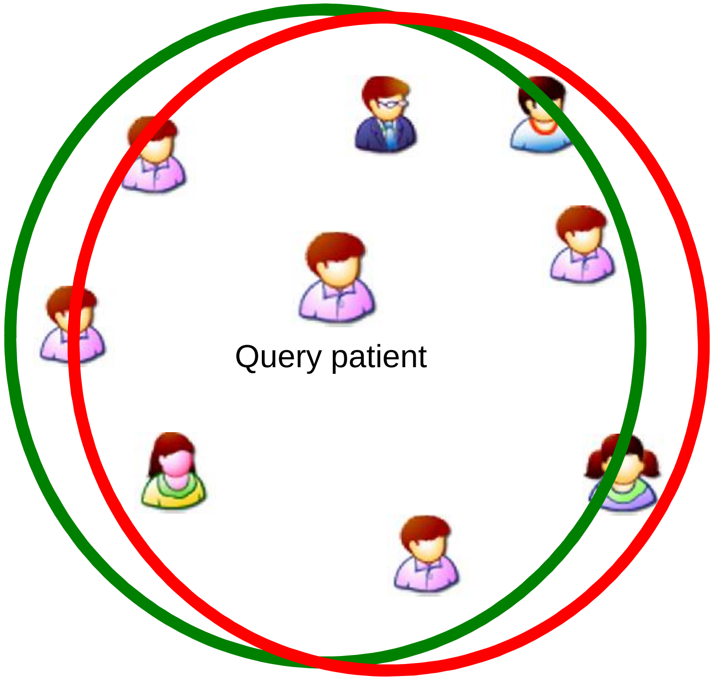


- Homogeneous neighbors: true positives
- Heterogeneous neighbors: false positives

Jimeng Sun, Fei Wang, Jianying Hu, Shahram Edabollahi: Supervised patient similarity measure of heterogeneous patient records. SIGKDD Explorations 14(1): 16-24 (2012)

Patient Similarity through Locally Supervised Metric Learning

Under a specific clinical context



- Shrink homogeneous neighborhood
- Grow heterogeneous neighborhood

Jimeng Sun, Fei Wang, Jianying Hu, Shahram Edabollahi: Supervised patient similarity measure of heterogeneous patient records. SIGKDD Explorations 14(1): 16-24 (2012)

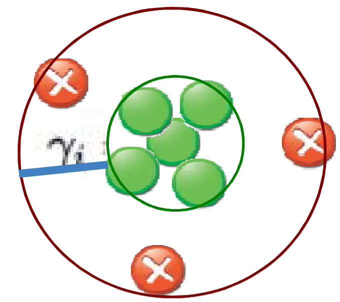
Locally Supervised Metric Learning (LSML)

Goal: Learn a generalized Mahalanobis distance for a specific clinical context (target label)

$$d_{\Sigma}(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^{\top} \Sigma (\mathbf{x}_i - \mathbf{x}_j)} \qquad \Sigma = \mathbf{W}\mathbf{W}^{\top}$$

Margin for x_i

$$\gamma_i = \underbrace{\sum_{k:\mathbf{x}_k \in \mathcal{N}_i^e} \left\|\mathbf{W}^T\mathbf{x}_i - \mathbf{W}^T\mathbf{x}_k\right\|^2}_{\text{Total distance to heterogeneous neighbors}} - \underbrace{\sum_{j:\mathbf{x}_j \in \mathcal{N}_i^o} \left\|\mathbf{W}^T\mathbf{x}_i - \mathbf{W}^T\mathbf{x}_j\right\|^2}_{\text{Total distance to homogeneous neighbors}}$$



$\mathcal{N}_i^o$ Homogeneous neighborhood for x_i

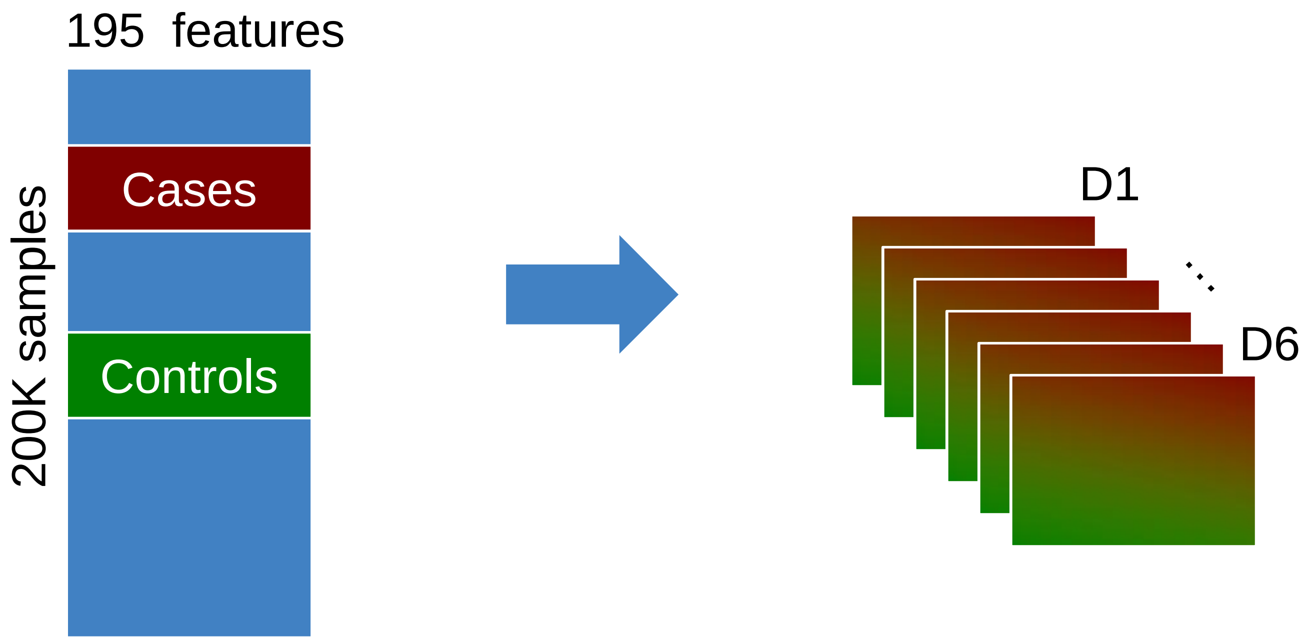
$\mathcal{N}_i^e$ Heterogeneous neighborhood for x_i

Maximize the total margin

$$\gamma = \sum_i \sum_{k:\mathbf{x}_k \in \mathcal{N}_i^e} \mathbf{W}^T(\mathbf{x}_i - \mathbf{x}_k)(\mathbf{x}_i - \mathbf{x}_k)^T\mathbf{W} - \sum_i \sum_{j:\mathbf{x}_j \in \mathcal{N}_i^o} \mathbf{W}^T(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T\mathbf{W}$$

Jimeng Sun, Fei Wang, Jianying Hu, Shahram Edabollahi: Supervised patient similarity measure of heterogeneous patient records. SIGKDD Explorations 14(1): 16-24 (2012)

Patient Similarity Experiment Design



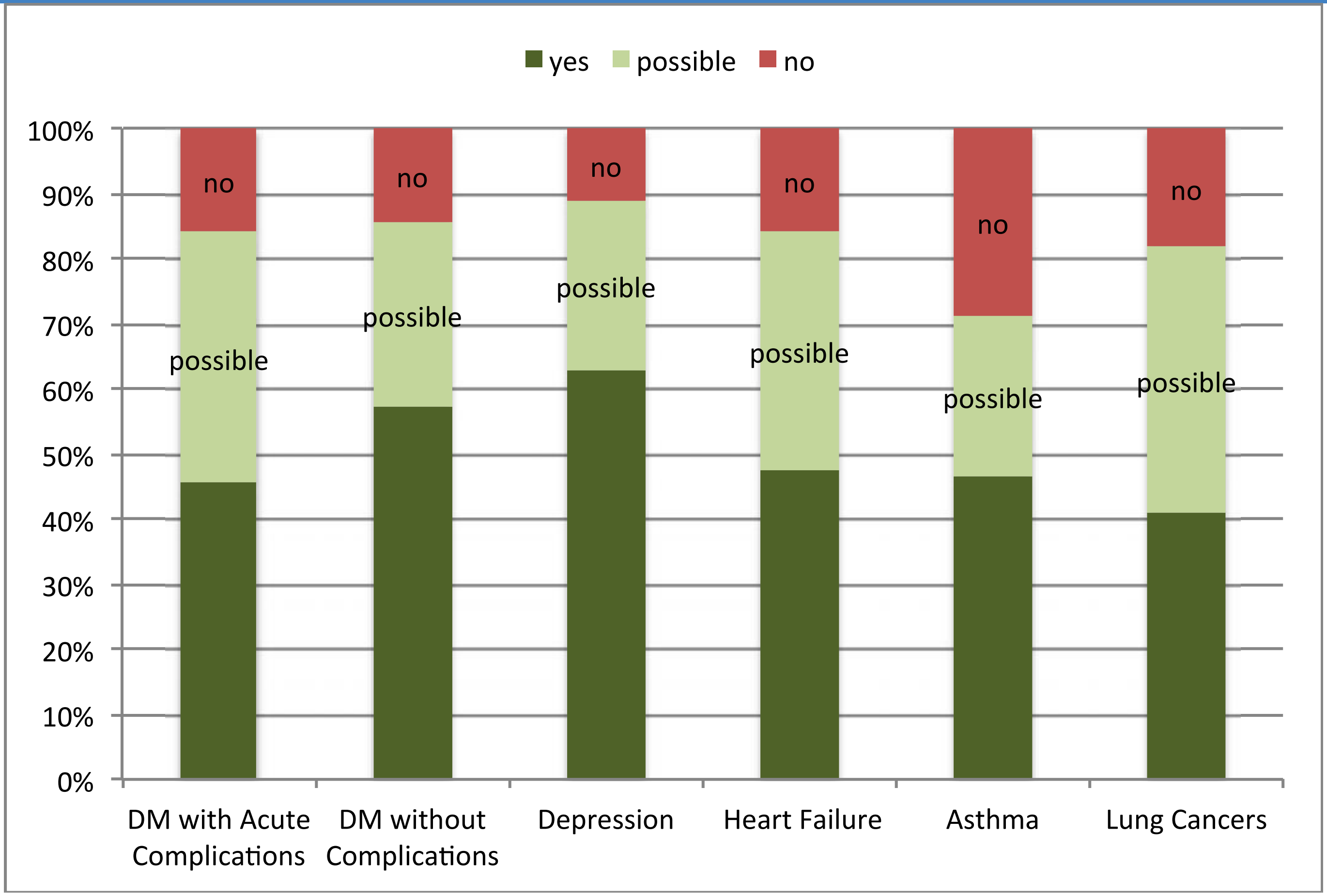
	Disease Target	samples
D1	DM with Acute Complications	4,392
D2	DM without Complications	10,734
D3	Depression	6,794
D4	Heart Failure	5,262
D5	Asthma	6,606
D6	Lung Cancers	1,172

Prediction Results on Patient Similarity

- Baselines:
  - EUC: Euclidean distance
  - PCA: Principal component analysis
  - LDA: Linear discriminant analysis
- Observations:
  - LDA does not perform well, because of the resulting dimensionality is too low
  - LSML algorithm performs the best among all

	DM with Acute Complications	DM without Complications	Depression	Congestive Heart Failure	Asthma	Lung Cancers
Euclidean	0.539	0.638	0.609	0.688	0.602	0.645
LDA	0.541	0.604	0.589	0.564	0.584	0.595
PCA	0.57	0.639	0.625	0.697	0.617	0.664
LSML	**0.576**	**0.669**	**0.632**	**0.723**	**0.625**	**0.677**

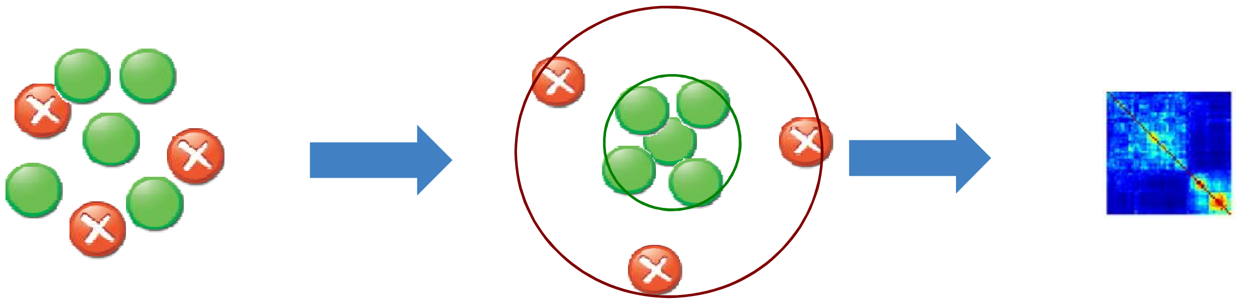
Clinical Relevancy of Patient Similarity Results



- Retrieve the top ranked comorbidities among similar patients
- **over 80% of those are considered as relevant** to target disease

Summary on Patient Similarity

- LSML learns a customized distance metric

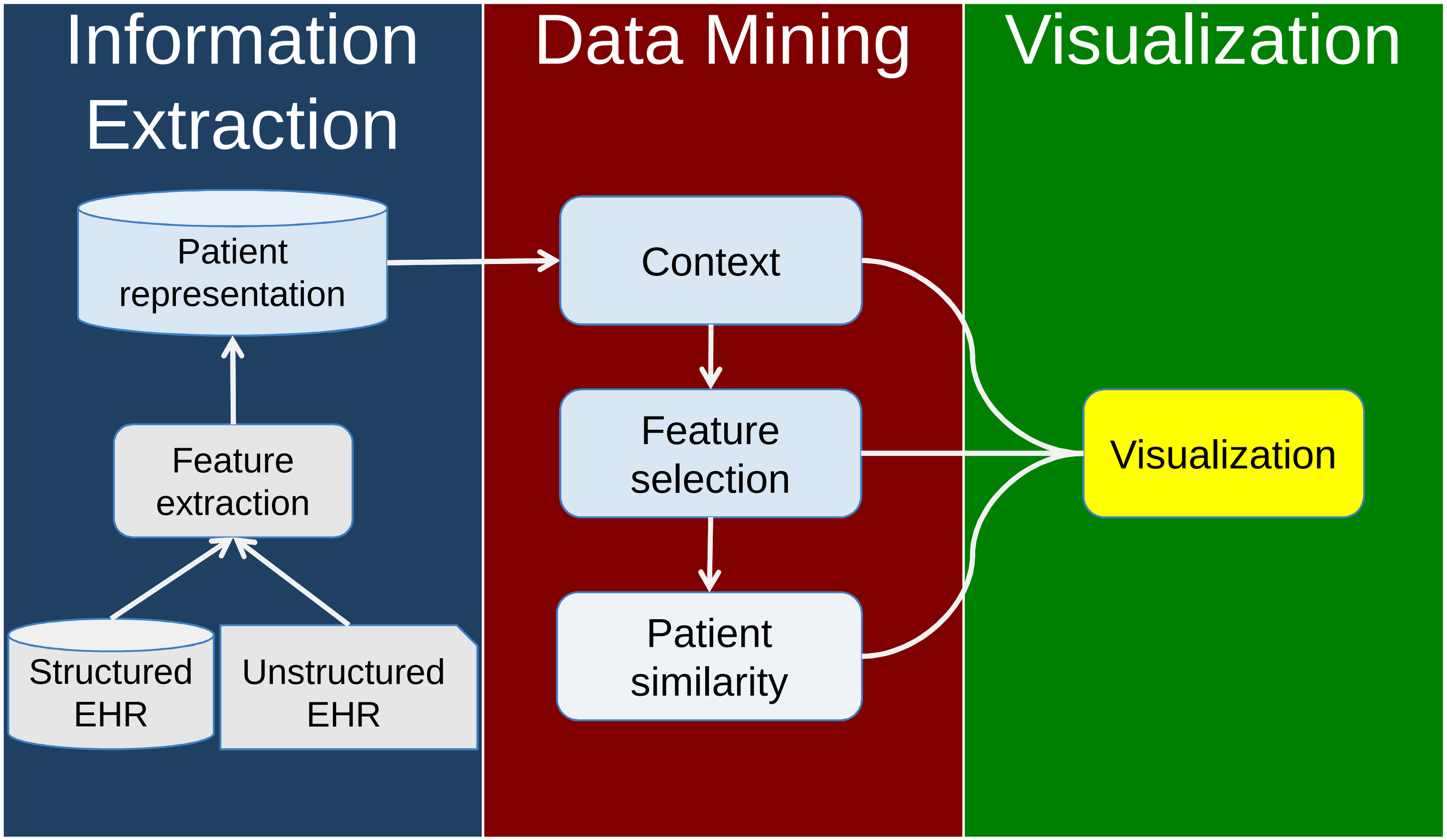


- Extension 1: Composite distance integration (Comdi) [1]
  – How to combine multiple patient similarity measures?

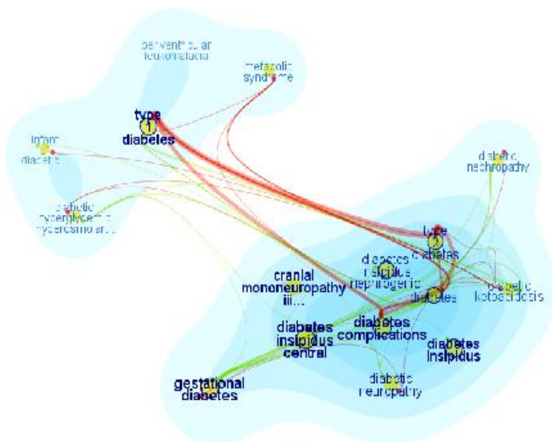
- Extension 2: Interactive metric update (iMet) [2]
  – How to update an existing distance measure?

1. Fei Wang, Jimeng Sun, Shahram Ebadollahi: Integrating Distance Metrics Learned from Multiple Experts and its Application in Inter-Patient Similarity Assessment. SDM 2011: 59-70 56
2. Fei Wang, Jimeng Sun, Jianying Hu, Shahram Ebadollahi: iMet: Interactive Metric Learning in Healthcare Applications. SDM 2011: 944-955

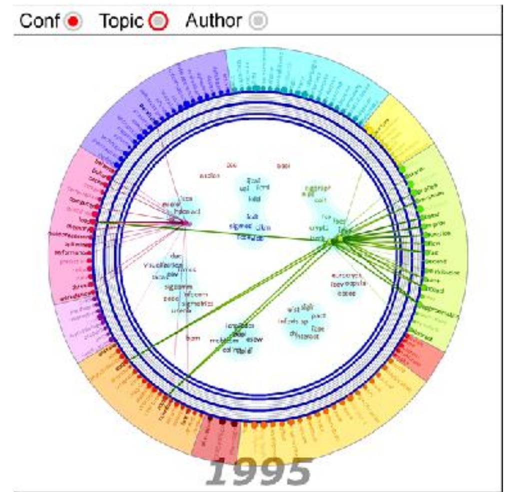
Visualization



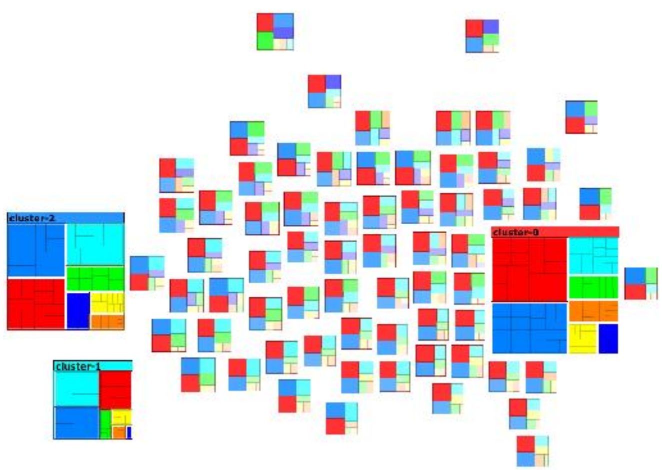
Visualization



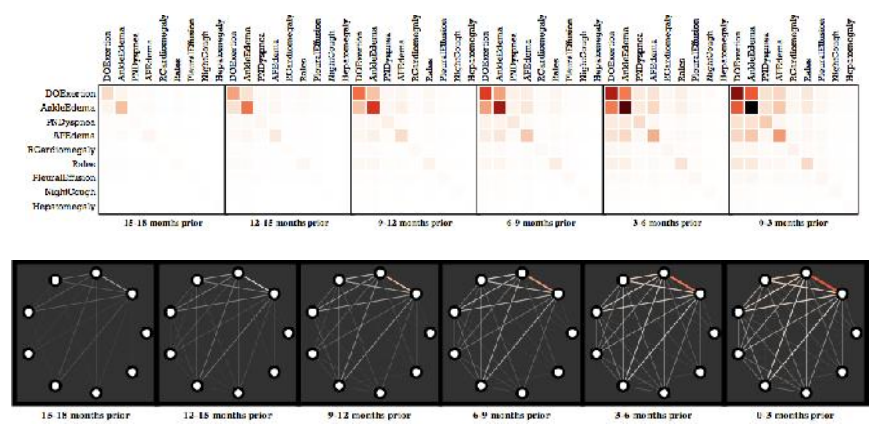
FacetAtlas (InfoVis'10)



SolarMap(ICDM'11)

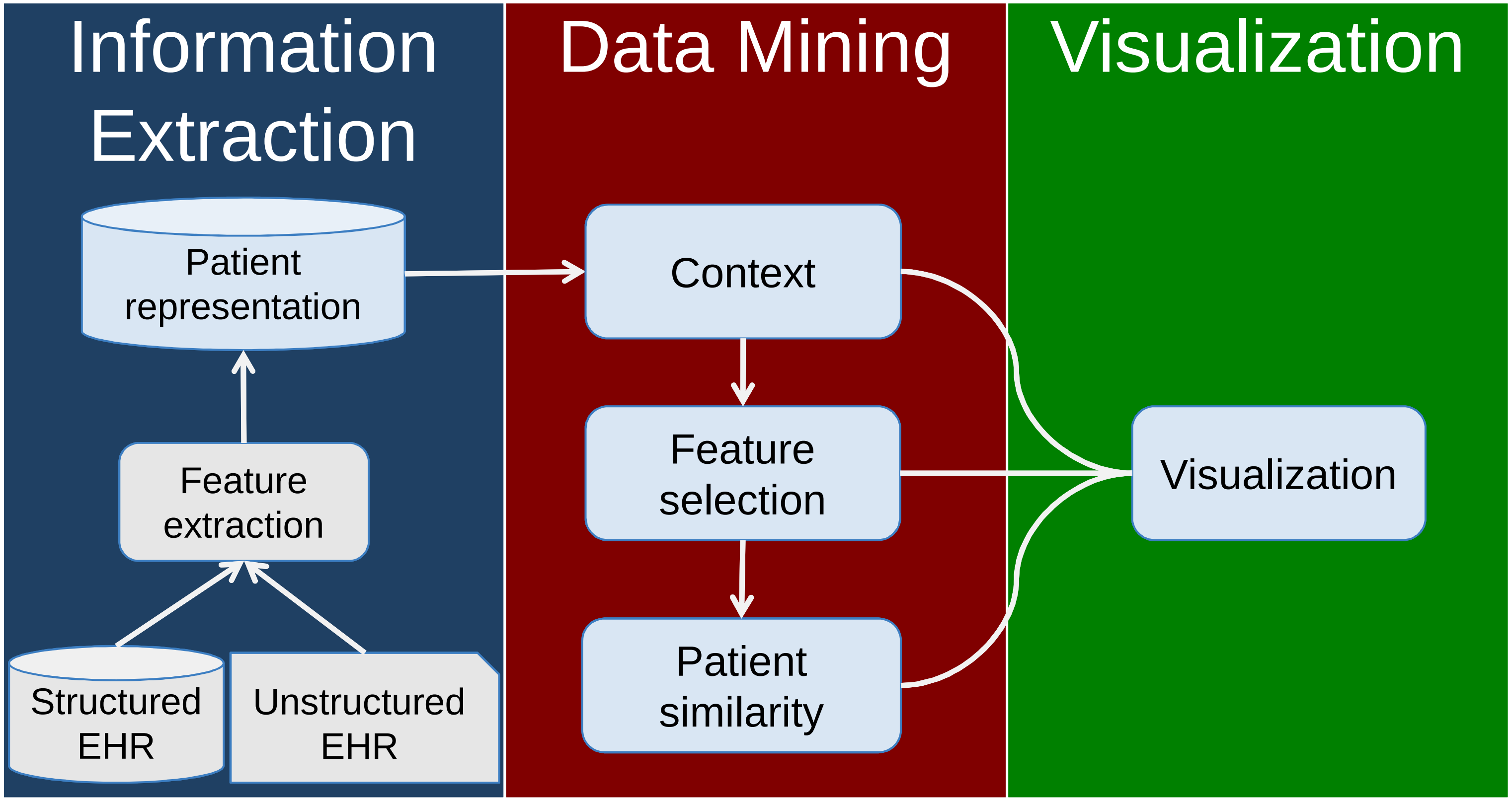


DICON (InfoVis'11)



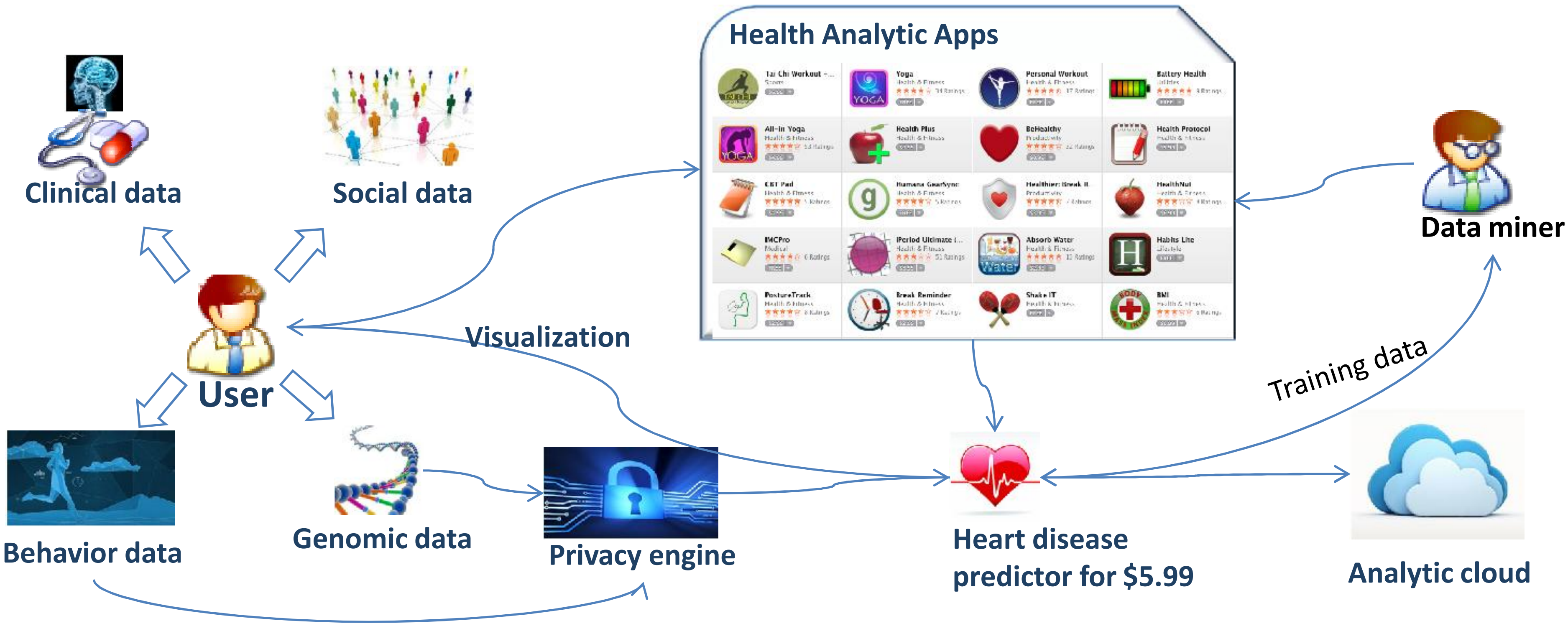
MatrixFlow (AMIA'12)

Conclusions



- Scalable healthcare analytic research platform
  - Enables efficient collaboration across domains
  - Extensible analytic platform
  - Intuitive analytic results
  - Scalable computation engine

Future of Healthcare Analytics



Research Challenges

- Data analytic techniques for each data modality
- Privacy preserving data sharing
- Visual analytic techniques