A Visual Analysis Approach to Cohort Study of Electronic Patient Records

Chun-Fu Wang[1], **Jianping Li**[1], Kwan-Liu Ma[1],
Chih-Wei Huang[2], Yu-Chuan Li[2]

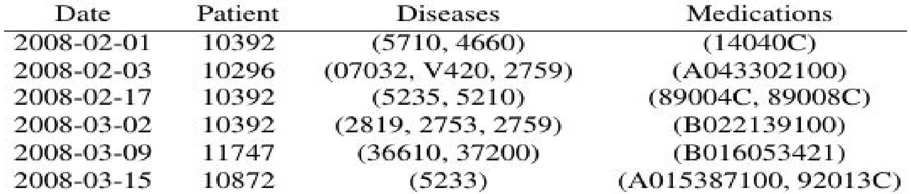
1 University of California at Davis
2 Taipei Medical University

ViDi
Visualization & Interaction
Design Institute

Electronic Medical Record (EMR)

- Rich information, great value
- 500 petabytes in 2012, 25000 petabytes expected in 2020
- Large and complex - challenges and opportunities

Challenges to Analyze EMRs

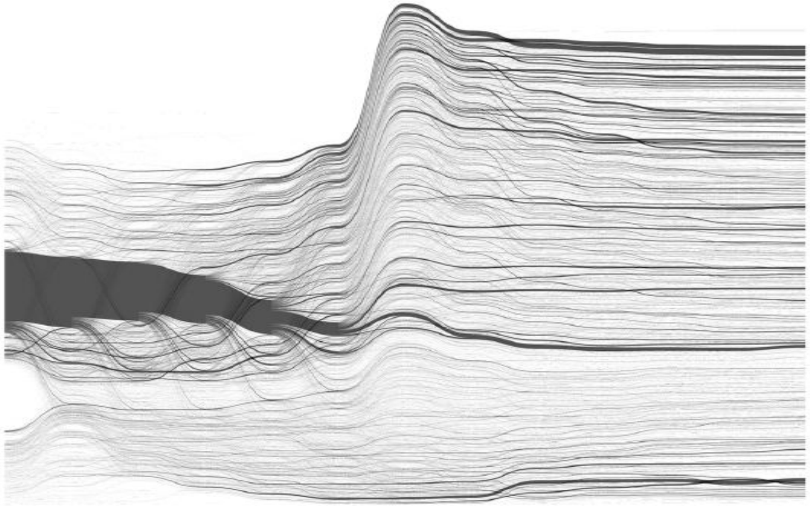


Date	Patient	Diseases	Medications
2008-02-01	10392	(5710, 4660)	(14040C)
2008-02-03	10296	(07032, V420, 2759)	(A043302100)
2008-02-17	10392	(5235, 5210)	(89004C, 89008C)
2008-03-02	10392	(2819, 2753, 2759)	(B022139100)
2008-03-09	11747	(36610, 37200)	(B016053421)
2008-03-15	10872	(5233)	(A015387100, 92013C)

...

Complexities in EMR

- multidimensional
- high variance



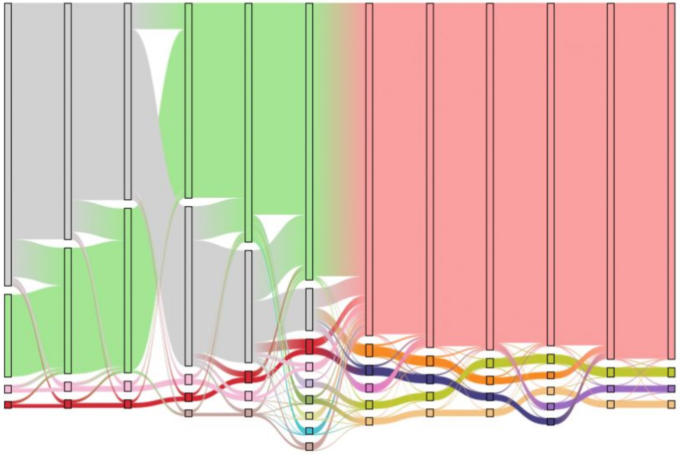
14,567 patients histories in 24 years

Opportunities for EMR Visual Analysis

EMRs

+

Domain Expertises



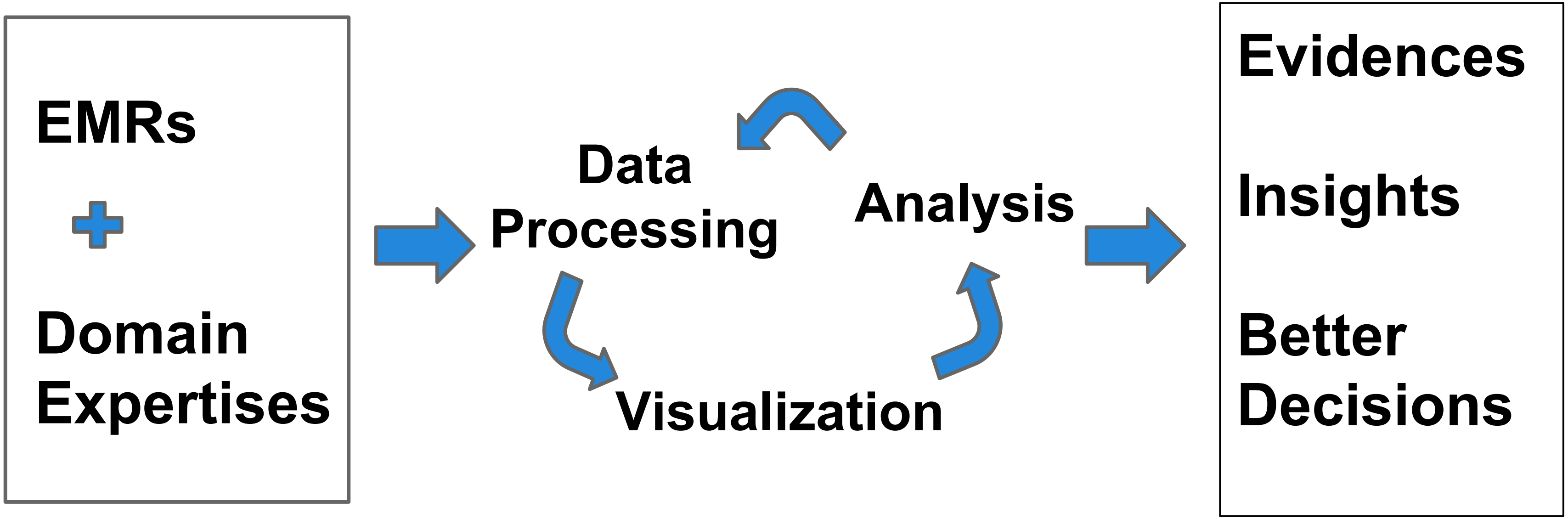
Visual Mining and Analysis

Evidences

Insights

Best Decisions

Iterative Visual Mining



EMRs

+

Domain Expertises

Data Processing

Analysis

Visualization

Evidences

Insights

Better Decisions

Our Approach

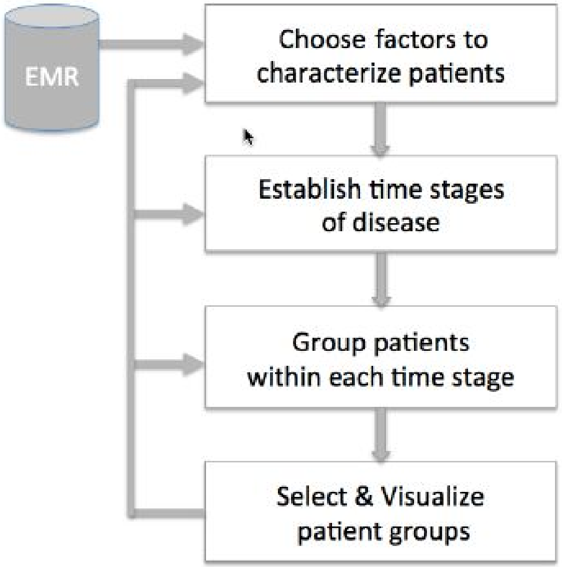
- An iterative workflow for analyzing large and complex EMR data.
- An interactive visualization system to support exploration of EMRs

Related Work

- **LifeFlow** - novel visualization tool to simplify and aggregate temporal event sequences into a tree-based summary
- **V-model** - compressed causal relationship along the linear time-scale to an ordinal representation
- **LifeLines2** - visual summary of prevalence and comparison of multiple groups

Workflow

- Choose factors based user knowledge
- Filter patient records using the factors
- Define time stages by partition and align
- Aggregate patients into groups(cohorts)



EMR

Choose factors to characterize patients

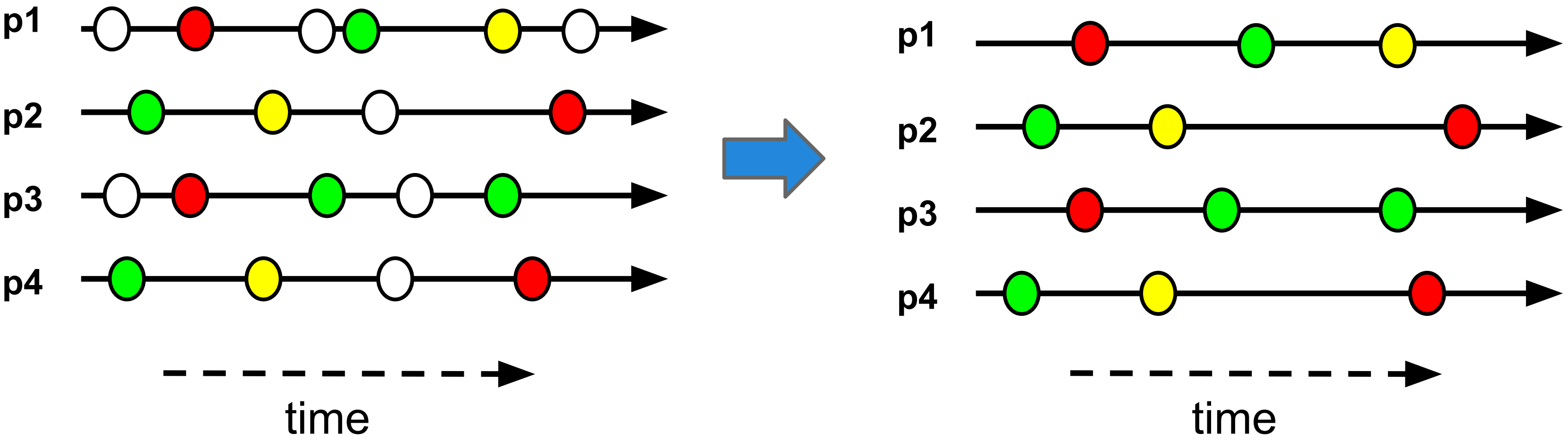
Establish time stages of disease

Group patients within each time stage

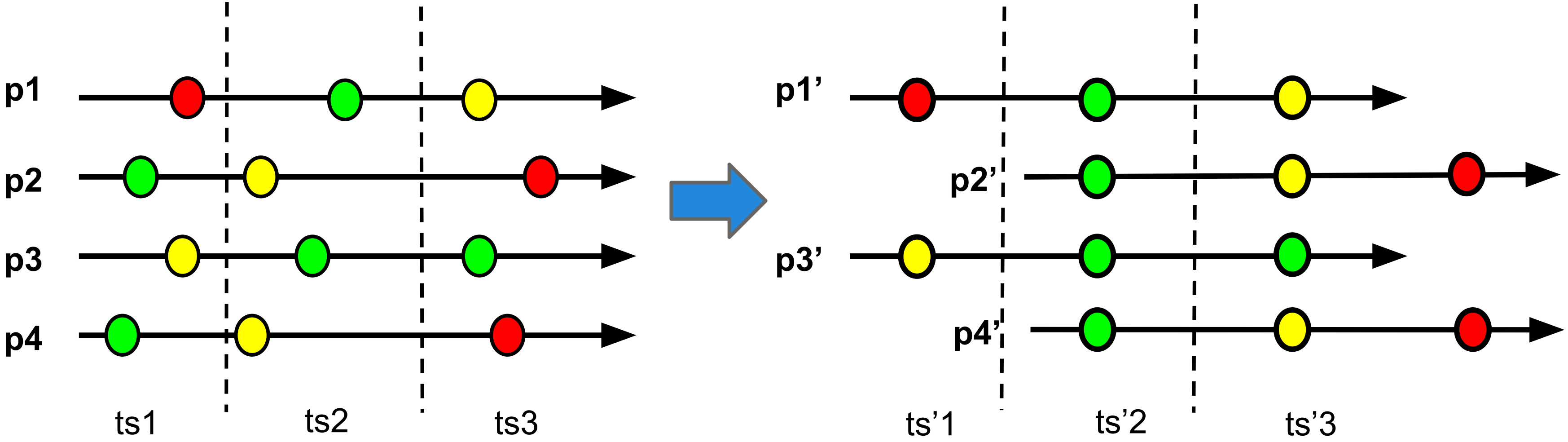
Select & Visualize patient groups

Factors and Filtering

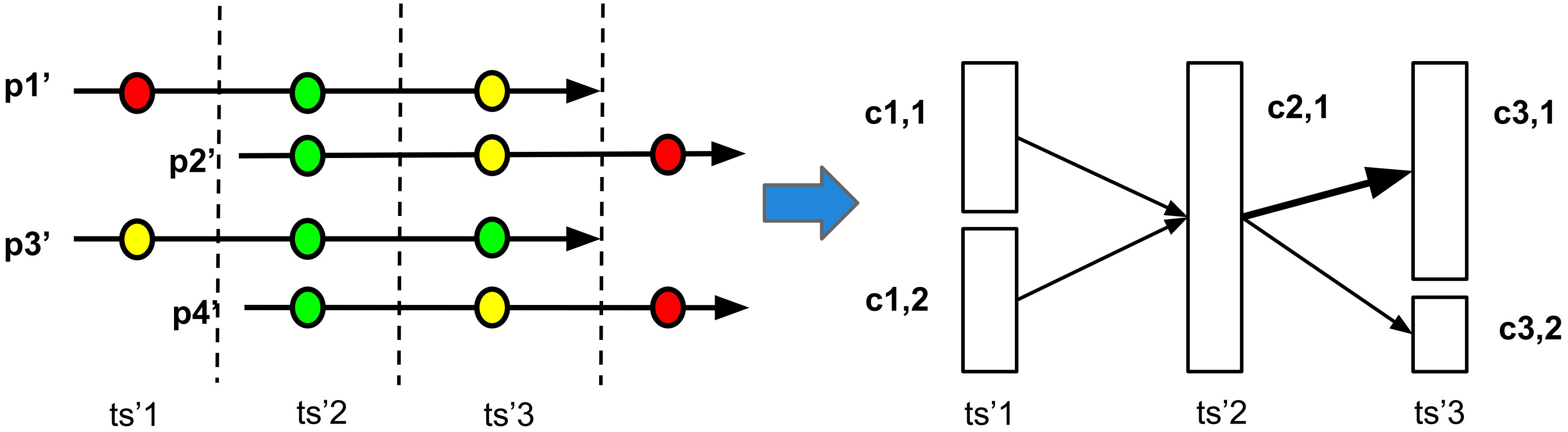
Factors (diseases): 🔴 🟢 🟡



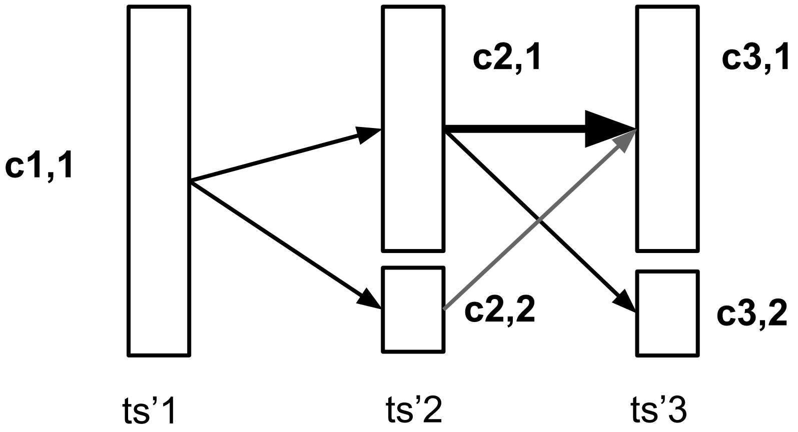
Partitioning and Aligning

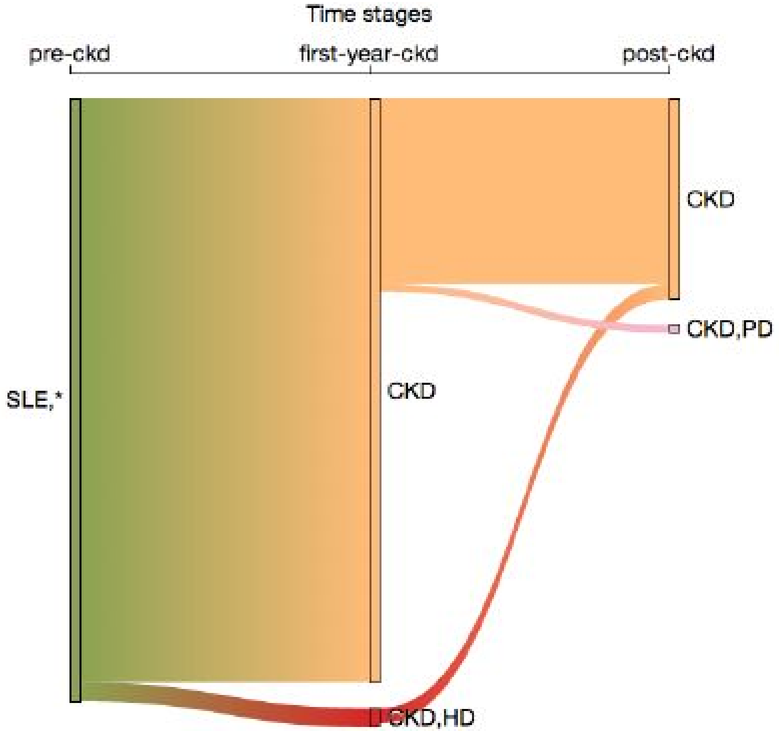


Aggregating to Cohorts

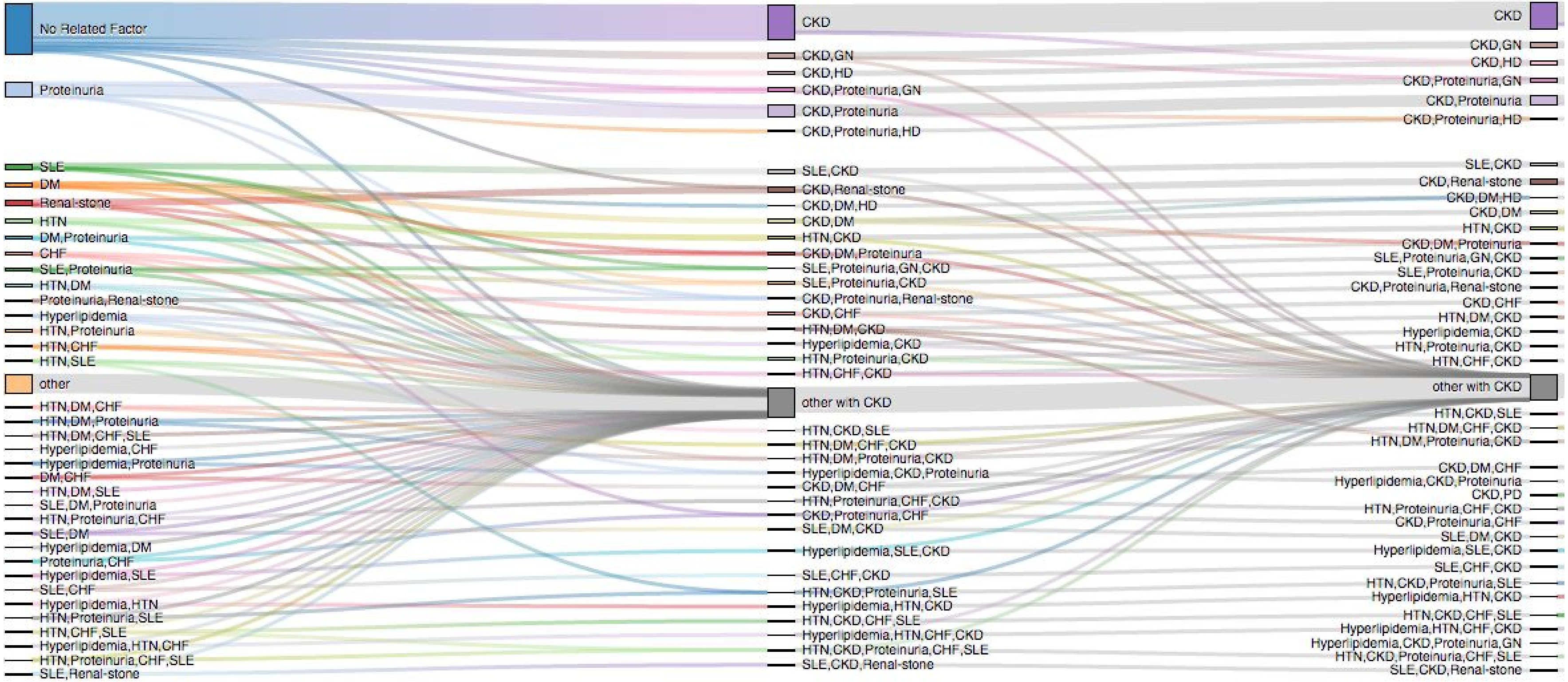


Visual Representation





But with Big Data ...



Cohorts Clustering

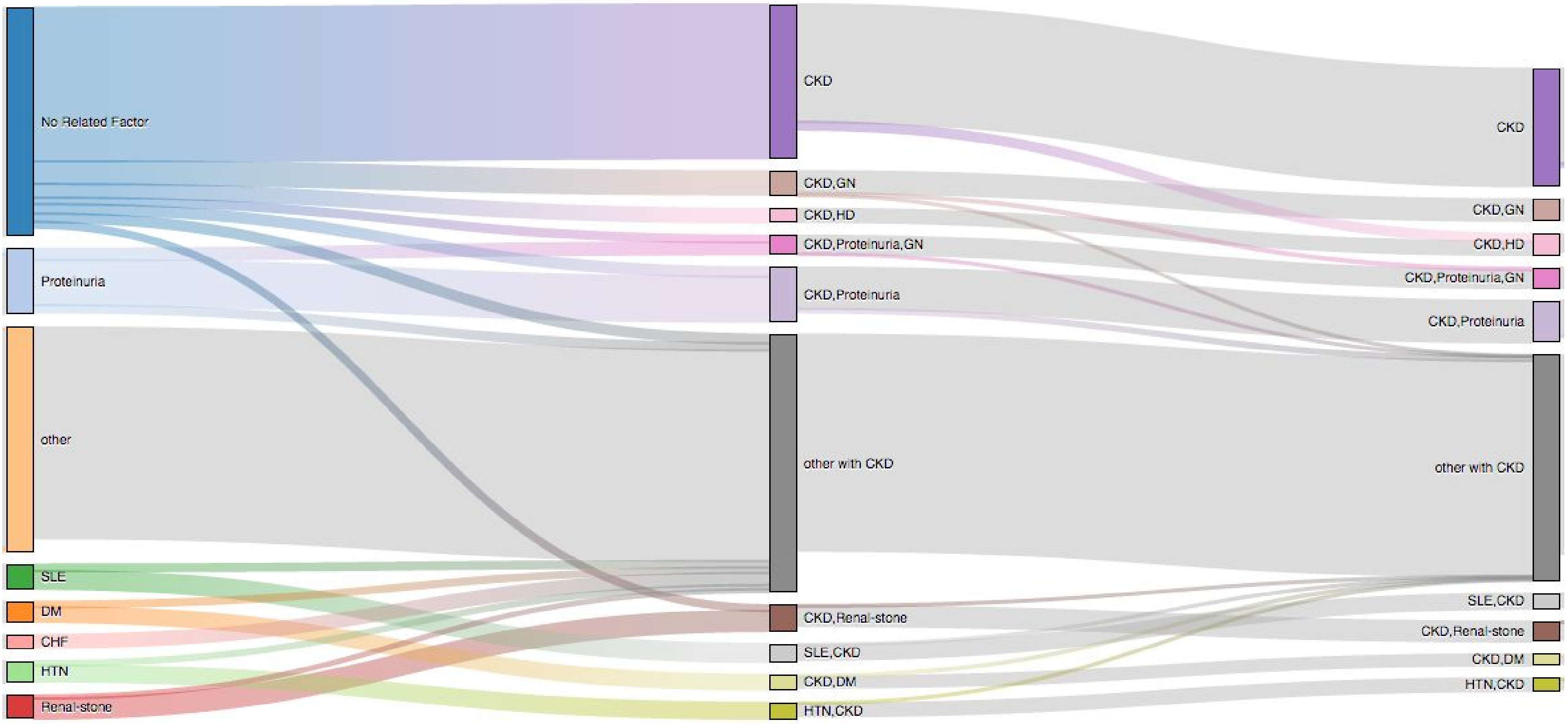
- Frequency-based clustering:
  - aggregate small cohorts into a cluster if the number of patient in the cohort is below the threshold

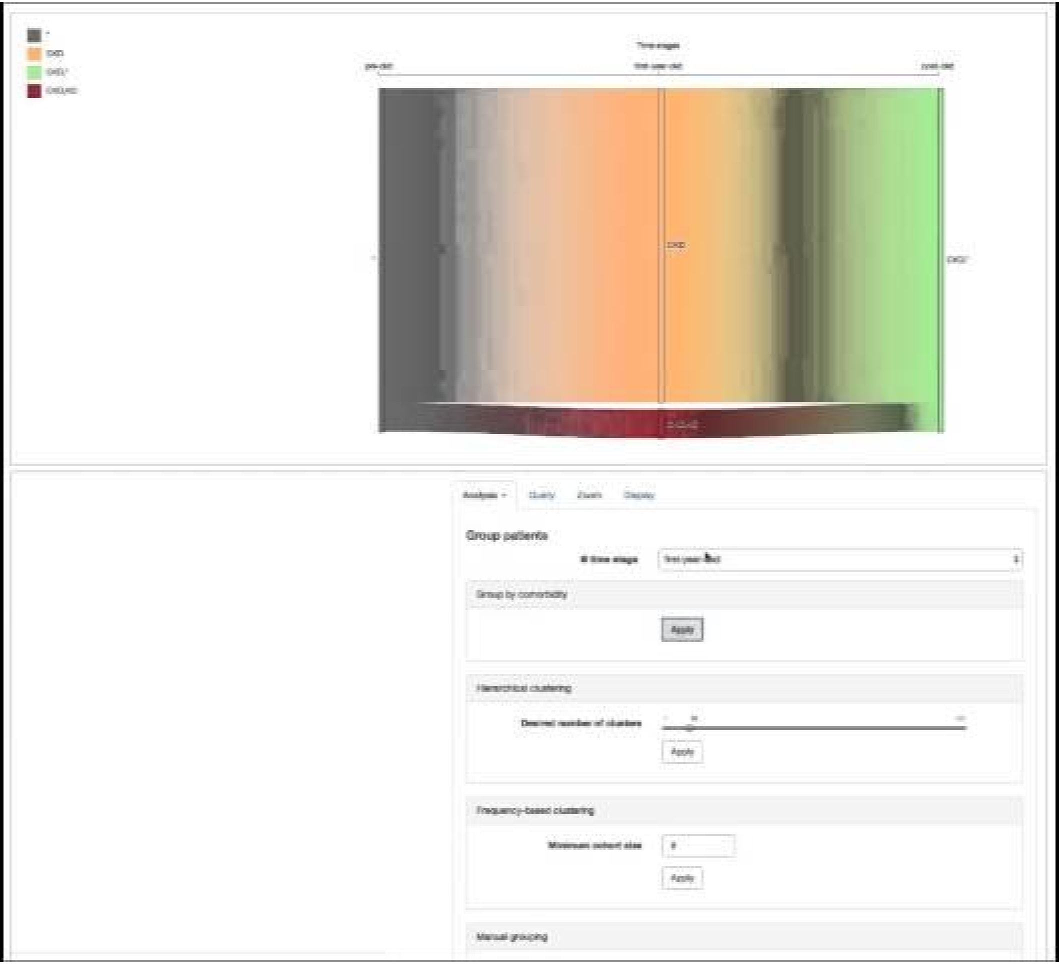
$$\text{cluster}\left(\mathbf{C}_l\right) = \begin{cases} \mathbf{c}_{l,h} & \text{if } \left|\mathbf{c}_{l,h}\right| \geq x \\ \text{others} & \text{if } \left|\mathbf{c}_{l,h}\right| < x \end{cases}$$

- Hierarchical clustering
  - cluster the cohorts based on the common factors

$$\text{similarity} = \frac{\left|\mathbf{s}_1 \cap \mathbf{s}_2\right|}{\sqrt{\left|\mathbf{s}_1\right| \left|\mathbf{s}_2\right|}}$$

Frequency-based Clustering (threshold=300)





Group patients

At time stage: first-year-dial

Group by comorbidity

Hierarchical clustering

Desired number of clusters

Frequency-based clustering

Minimum cohort size: 8

Manual grouping

Case Study - Chronic Kidney Disease(CKD)

14,567 CKD patients extracted from Taiwan NHIDB with over 1 million patients

- Dataset:
  - 6 million records
  - from 1998 to 2011
- Codes:
  - ICD 9-CM
  - NHIDB procedure/drug

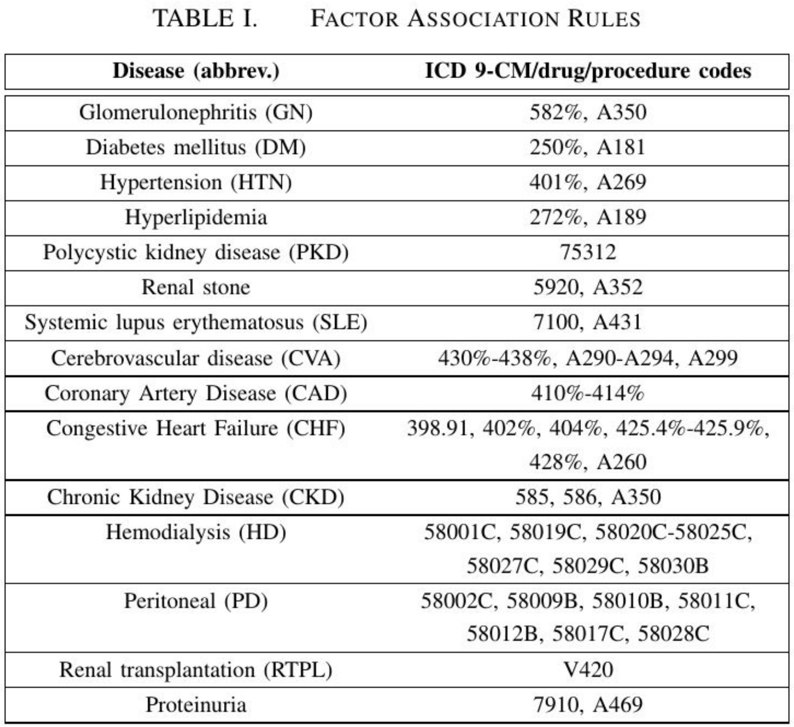
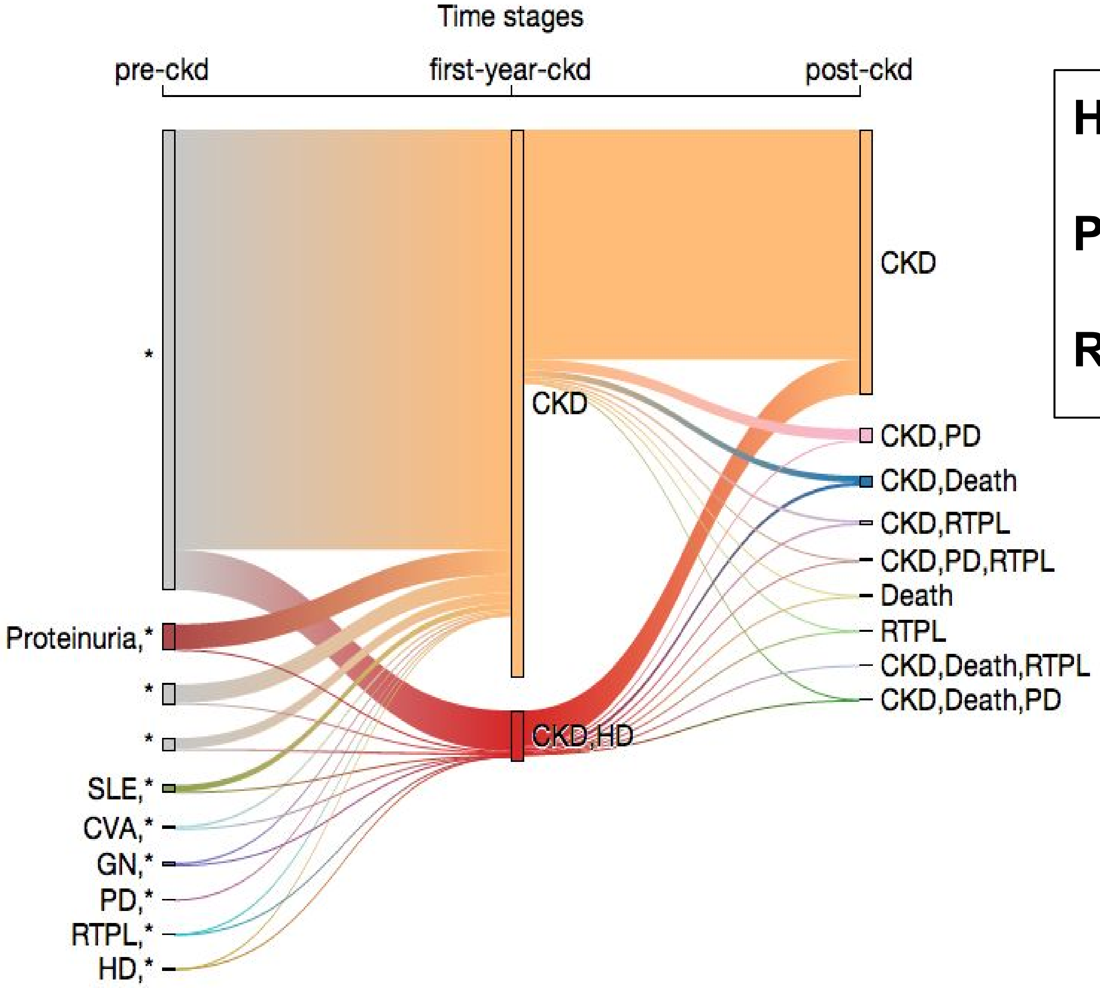


TABLE I. FACTOR ASSOCIATION RULES

Disease (abbrev.)	ICD 9-CM/drug/procedure codes
Glomerulonephritis (GN)	582%, A350
Diabetes mellitus (DM)	250%, A181
Hypertension (HTN)	401%, A269
Hyperlipidemia	272%, A189
Polycystic kidney disease (PKD)	75312
Renal stone	5920, A352
Systemic lupus erythematosus (SLE)	7100, A431
Cerebrovascular disease (CVA)	430%-438%, A290-A294, A299
Coronary Artery Disease (CAD)	410%-414%
Congestive Heart Failure (CHF)	398.91, 402%, 404%, 425.4%-425.9%, 428%, A260
Chronic Kidney Disease (CKD)	585, 586, A350
Hemodialysis (HD)	58001C, 58019C, 58020C-58025C, 58027C, 58029C, 58030B
Peritoneal (PD)	58002C, 58009B, 58010B, 58011C, 58012B, 58017C, 58028C
Renal transplantation (RTPL)	V420
Proteinuria	7910, A469

Case Study Objectives

- Investigate CKD related diseases co-occurrence (comorbidity)
- Explore the causal relationship between hemodialysis (HD) and other factors in early stage of CKD and identify the common driving factors of HD
- Explore global structures of cohorts and their changes over time stages

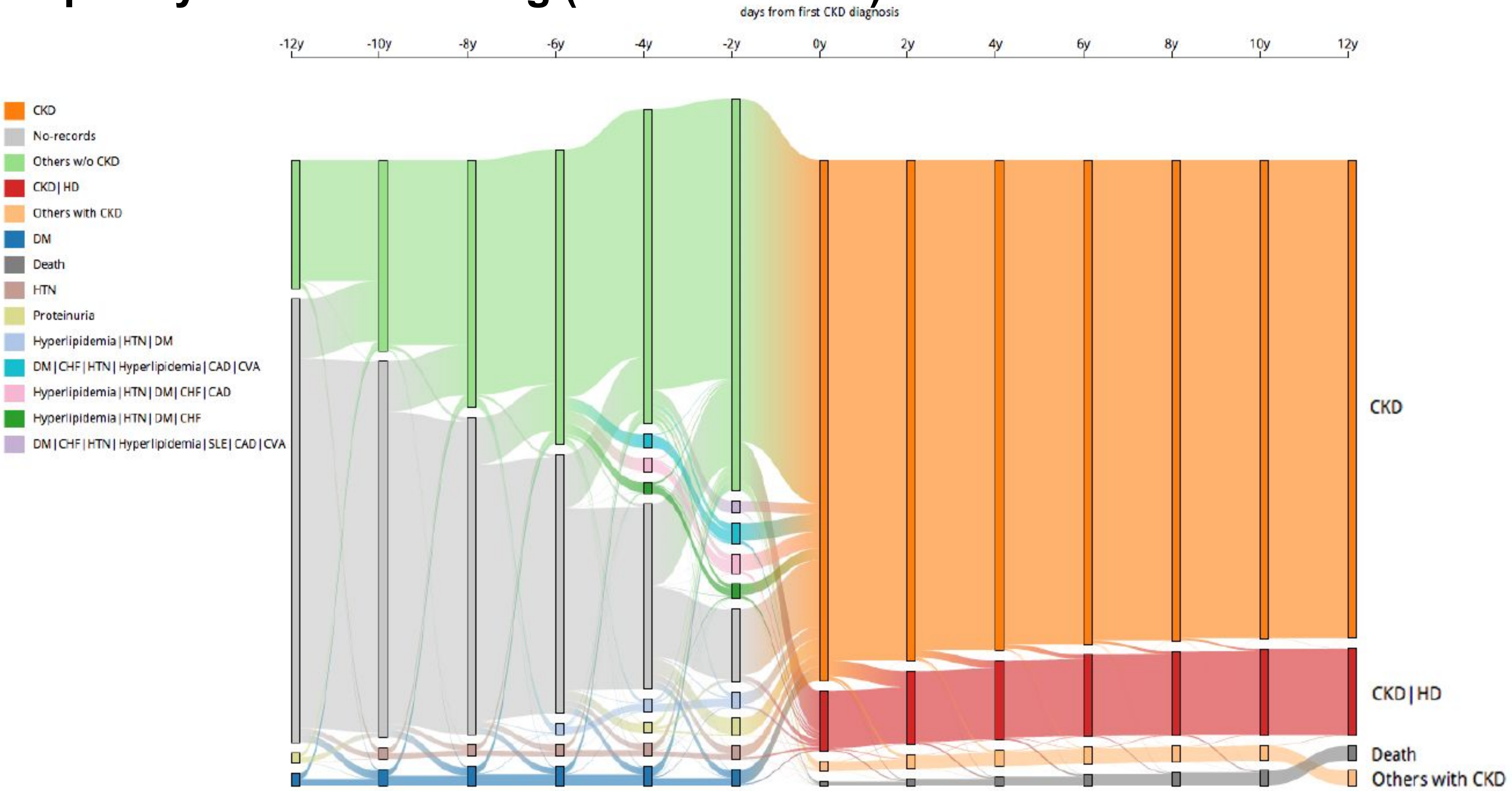


HD - Hemodialysis

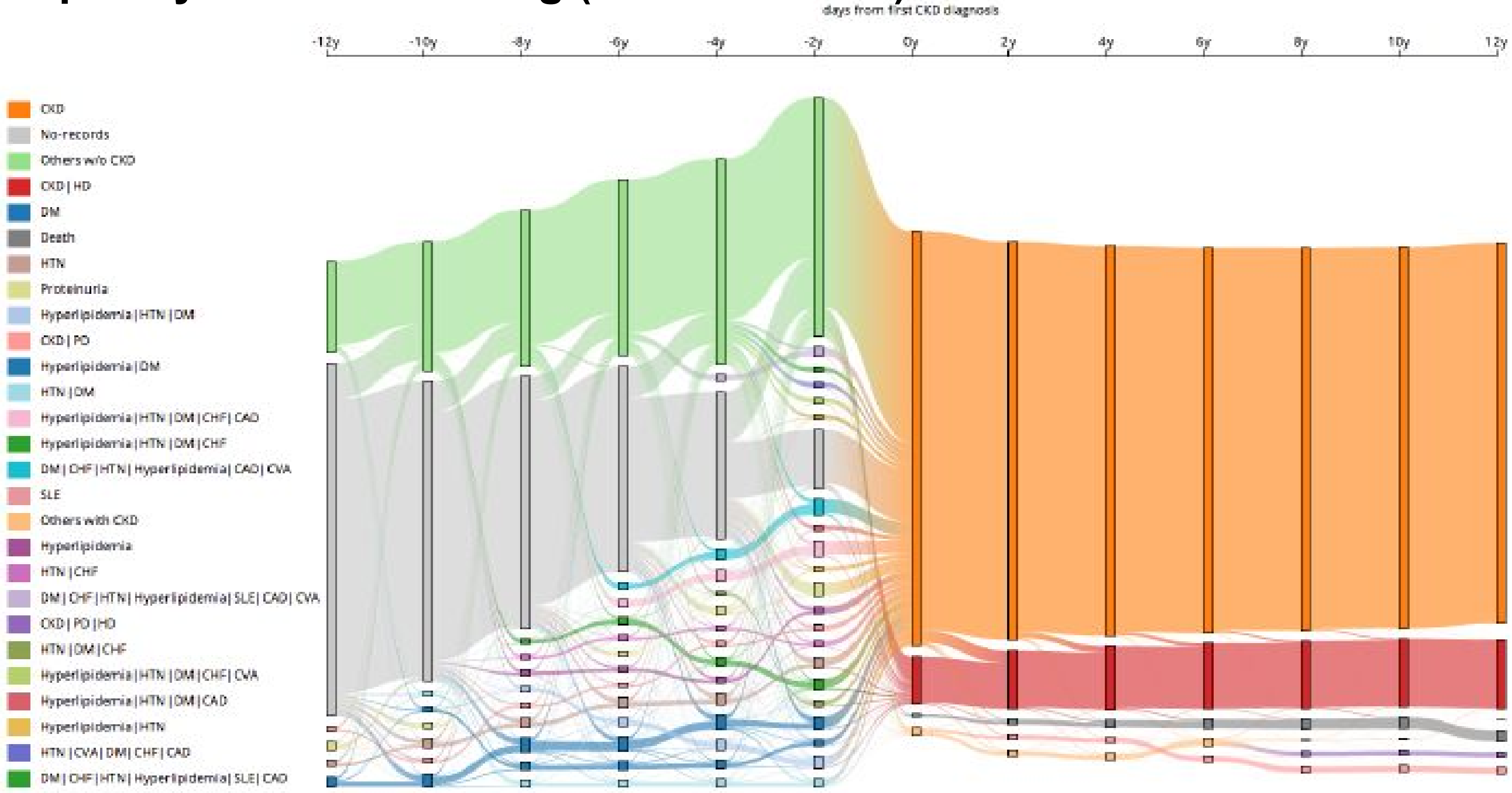
PD - Peritoneal

RTPL - Renal Transplantation

Frequency-based clustering (threshold=250)



Frequency-based clustering (threshold=150)



Conclusion

- EMRs
  - Large and complex
  - Rich and valuable information
- A new EMR visualization tool
  - An iterative process for EMR visual mining
  - An interactive system for visual analysis of EMR

Future Work

- Usability evaluation and improvement
- Comparative visualization
- High performance and scalable visual analytics system for large scale EMR data

Acknowledgement

This research is sponsored in part by the U.S. National Science Foundation and UC Davis RISE program.

Thank You