Data Science in Biomedicine: Risk Estimators and Pattern Mining

Panos M. Pardalos[1,2]

[1]Center for Applied Optimization
Department of Industrial and Systems Engineering
University of Florida, USA
http://www.ise.ufl.edu/cao/

[2]Laboratory of Algorithms and Technologies for Networks Analysis (LATNA)
National Research University Higher School of Economics, Russia
https://nnov.hse.ru/en/latna/

BIOMAT 2017 (Moscow, Russia)
October 29 - November 3, 2017

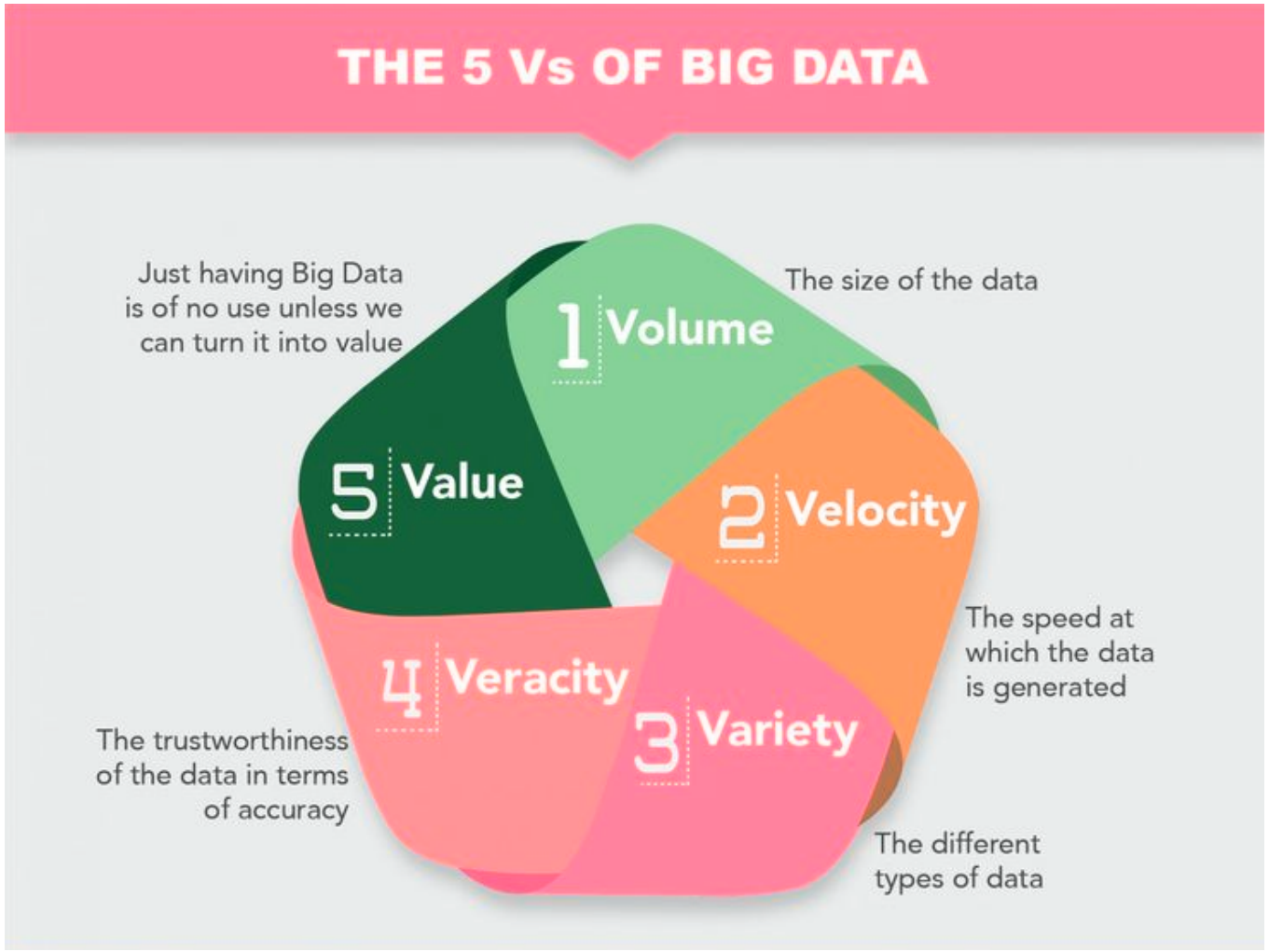
This talk is dedicated to the memory of my friend and colleague Chris Floudas

Outline

What is Big Data?

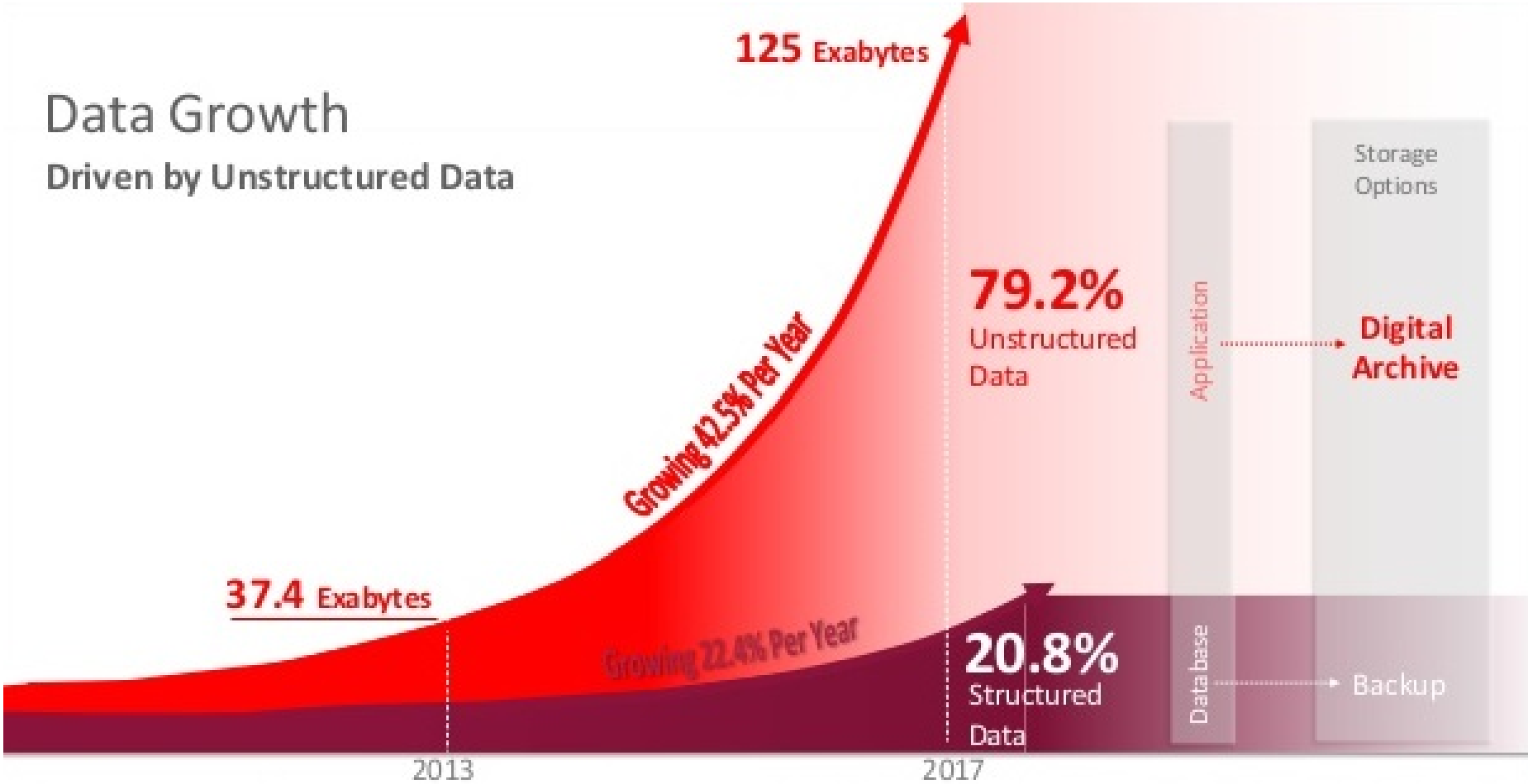
- The proliferation of massive datasets brings with it a series of special computational challenges.
- This data avalanche arises in a wide range of scientific and commercial applications.
- With advances in computer and information technologies, many of these challenges are beginning to be addressed

The 5 V's that define Big Data

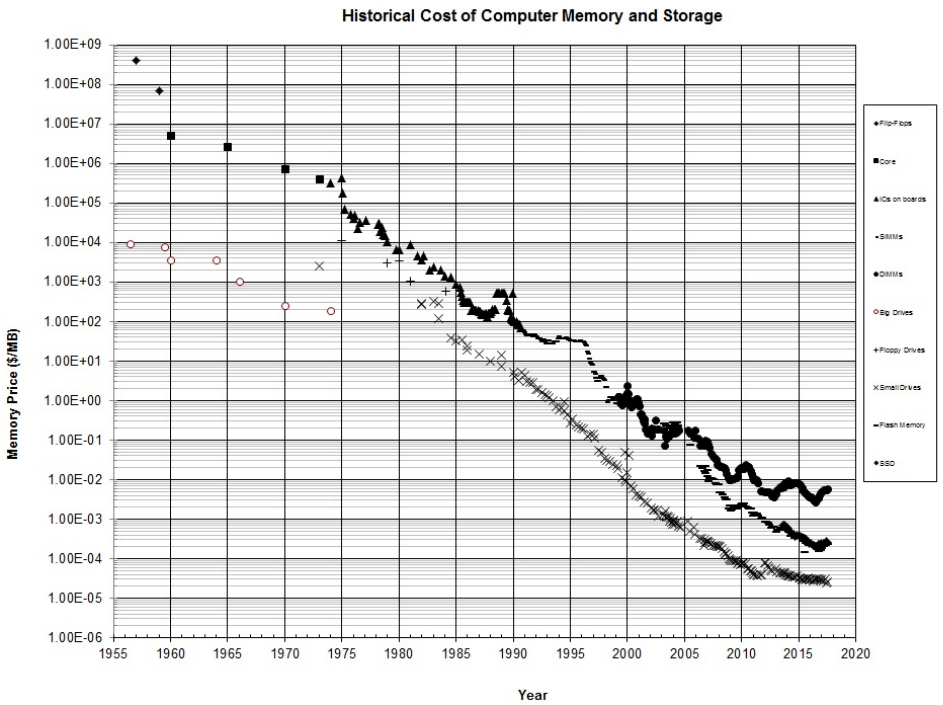


http://bigdata.black/featured/what-is-big-data

Data Growth


Data Growth
Driven by Unstructured Data
125 Exabytes
37.4 Exabytes
Growing 42.5% Per Year
Growing 22.4% Per Year
79.2%
Unstructured
Data
20.8%
Structured
Data
Application
Database
Storage
Options
Digital
Archive
Backup
2013
2017

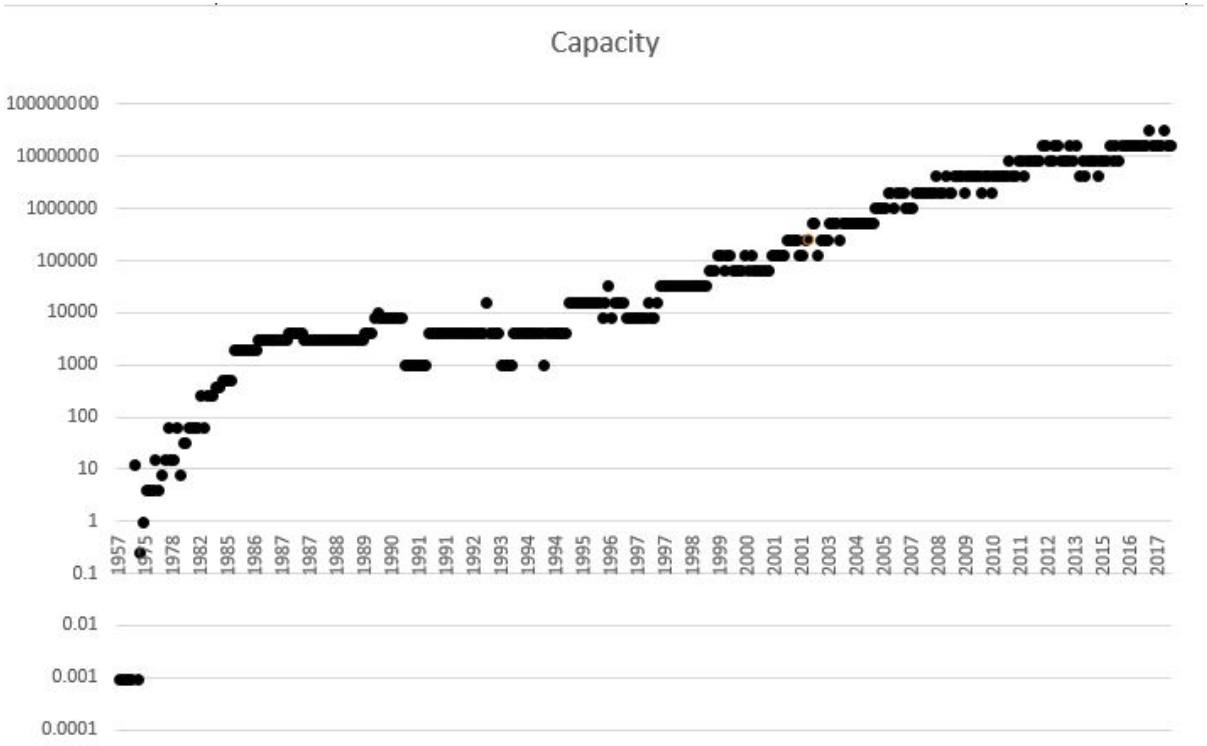
Historical Cost of Computer Memory and Storage



Approximately **10 times** cheaper every **5 years**

http://www.jcmit.net/diskprice.htm

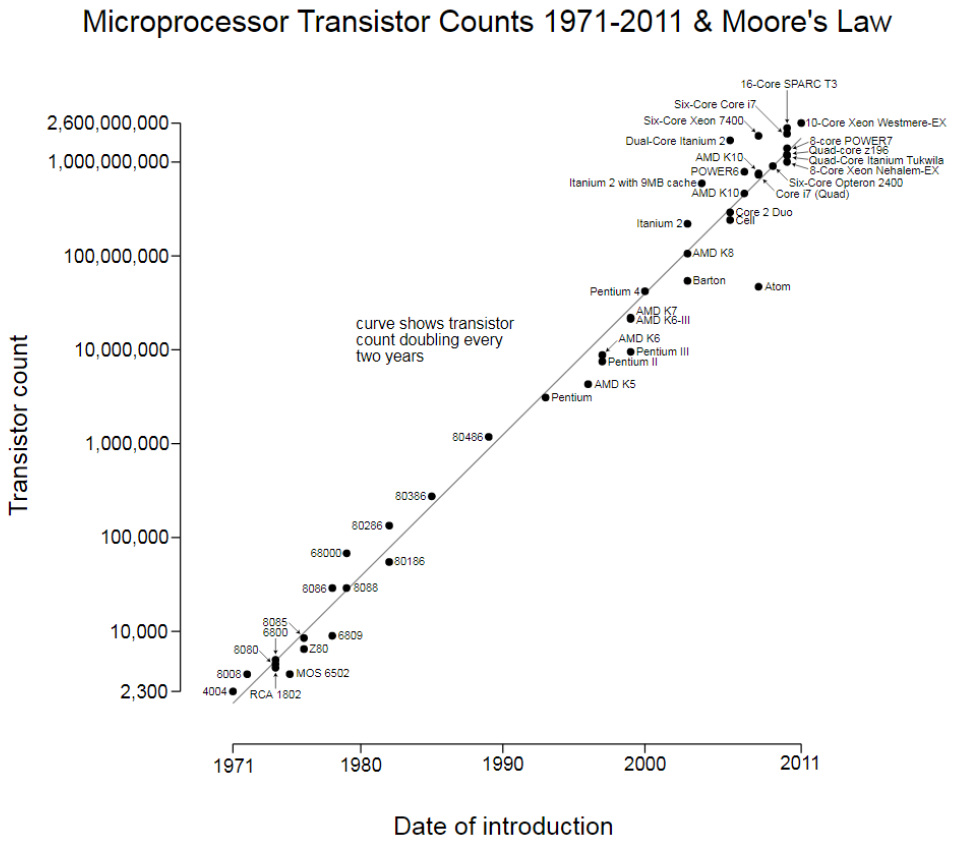
Memory Capacity



Approximately **10 times** more every **5 years**

http://www.jcmit.net

Processing Power



Number of transistors in CPUs **doubles** every two years

https://en.wikipedia.org/wiki/Transistor_count

Some Examples of Massive Data Sources

- WWW, Internet
- Weather
- Telecommunications
- Media
- Demographics Data
- Financial Data
- Social Networks
- Biological Networks
- Health Care

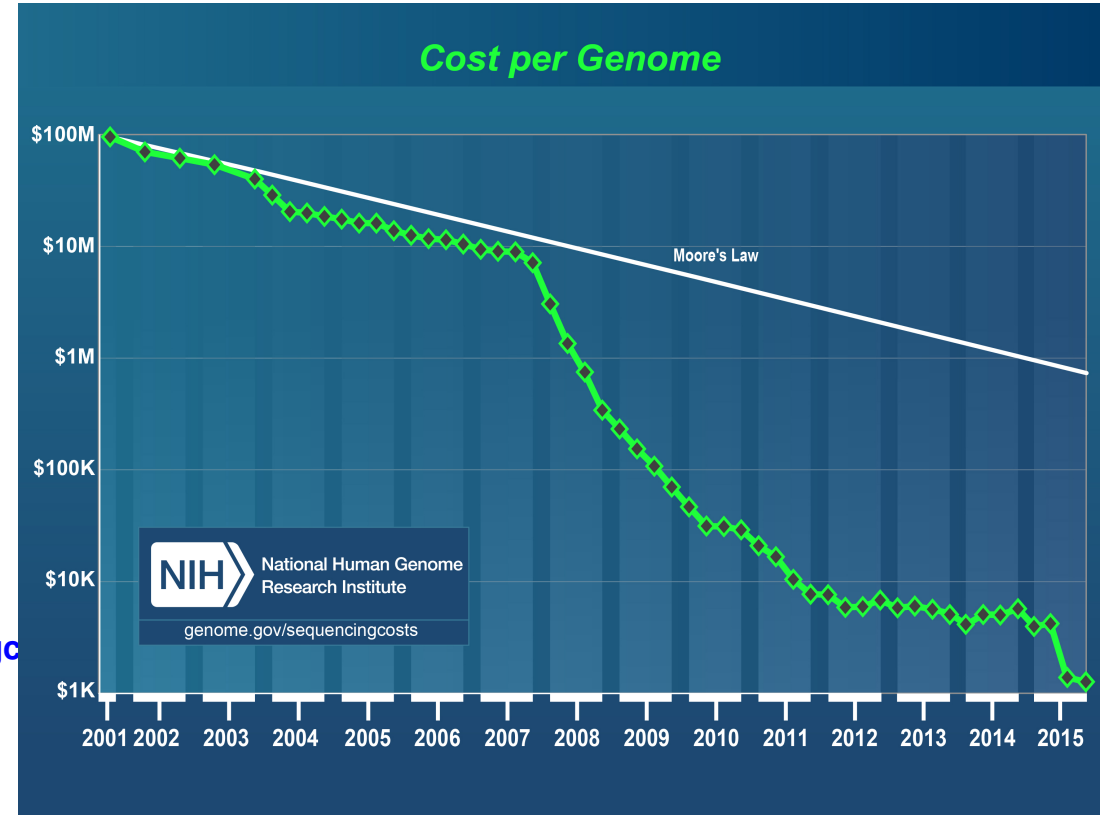
AT&T Call Graph

- **Call graph**:
  - vetrices are phone numbers
  - two are connected if a phone call was made from one to another
- **One-day** call graph had **53,767,087 vertices** and **170 million edges** [1].

[1] Abello, J., Pardalos, P.M, and Resende,M.G.C. On maximum clique problems in very large graphs. DIMACS series 50, 119-130 (1999).

Human Genome Project

- The first human genome was sequenced in 2003.
- The number of completely sequenced genomes is close to **1 million**. There will be **1 billion** of them in 10-15 years [1].
- Sequencing costs have dropped **100,000-fold** since 2001 from **100 million** dollars to **one thousand** per genome: **www.genome.gov/sequencingc**



[1] Eisenstein, M. Big data: The power of petabytes. Nature 527, S2S4 (2015).

Human Brain

- Enormous size and complexity: **85 billion** neurons and **100 trillion** synapses [1].
- Small volumes of brain tissue can be fully reconstructed [2].
- **175 exabytes** are projected to be required to store human brain connectome [3].

[1] Azevedo, F.A. *et al*. Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. J Comp Neurol.513(5), 532-541 (2009).

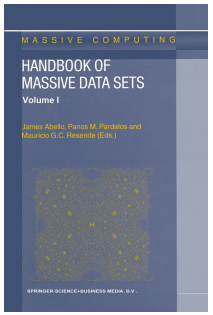
[2] Helmstaedter, M. Cellular-resolution connectomics: challenges of dense neural circuit reconstruction. Nature Methods 10, 501-507 (2013).

[3] Mikula, S. Progress Towards Mammalian Whole-Brain Cellular Connectomics. Frontiers in Neuroanatomy 10, 62 (2016).

Electronic Medical Records (EMRs)

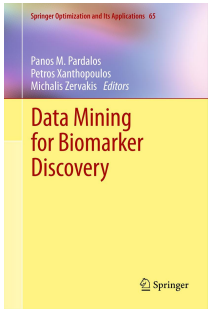
- EMRs are collections of **partially-structured** records of medical data.
- Data come in different formats:
  - demographic data
  - time series
  - imaging
  - plain text
  - etc.
- Examples:
  - MIMIC-III: deidentified health data associated with $\sim 40,000$ critical care patients https://mimic.physionet.org/
  - University of Florida Health Integrated Data Repository

Selected Books



- Abello, James, Panos M. Pardalos, and Mauricio GC Resende, eds.
  Handbook of massive data sets. Vol. 4. Springer, 2013.

- Pardalos, Panos M., Vladimir L. Boginski, and Vazacopoulos Alkis, eds.
  Data mining in biomedicine.Vol. 7. Springer Science & Business Media, 2008.

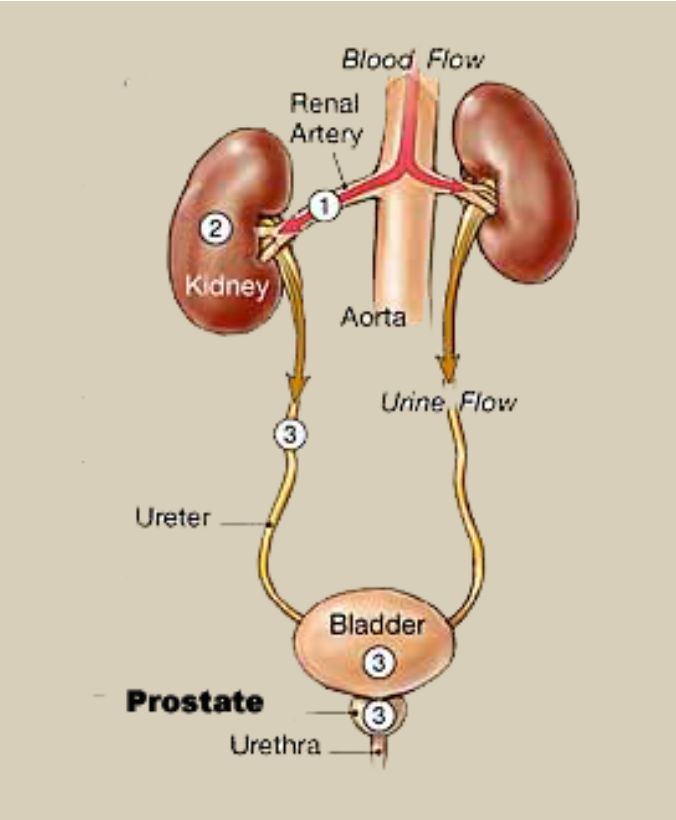


- Pardalos, Panos M., Petros Xanthopoulos, and Michalis Zervakis, eds.
  Data Mining for Biomarker Discovery. Vol. 65. Springer Science & Business Media, 2012.

AKI

- **Acute Kidney Injury** (AKI) is one of common complications in post-operative patients.
- The in-hospital mortality rate for patients with AKI may be as high as **60%**.
- An elevated **serum creatinine (sCr)** level in blood is commonly recognized as an indicator of AKI.
- The degree to which patterns of sCr change are associated with in-hospital mortality is unknown

Causes of AKI



1. **Prerenal**. Sudden and severe drop in blood pressure (shock) or interruption of blood flow to the kidneys from severe injury or illness.
2. **Intrarenal**. Direct damage to the kidneys by inflammation, toxins, drugs, infection, or reduced blood supply.
3. **Postrenal**. sudden obstruction of urine flow due to enlarged prostate, kidney stones, bladder tumor, or injury.

Sepsis

- **Sepsis** is a severe and dysregulated inflammatory response to infection characterized by end-organ dysfunction distant from the primary site of infection.
- Development of (AKI) during sepsis:
  - increases patient morbidity,
  - predicts higher mortality,
  - has a significant effect on multiple organ functions,
  - is associated with an increased length of stay in the intensive care unit,
  - and hence consumes considerable healthcare resources

[1] Zarjou, A., & Agarwal, A. (2011). Sepsis and acute kidney injury. Journal of the American Society of Nephrology, 22(6), 999-1006.

Study Objective

Objectives:

- To develop a comprehensive model for assessing the mortality risk in post-operative patients
- To establish a quantitative relationship between sCr pattern and mortality risk

Results:

- A probabilistic model was designed to assess mortality risk in post-operative patients
- The model provided high discriminative capacity and accuracy
- A quantitative association between sCr time series and mortality risk was established
- Simple and informative sCr risk factors were derived

Data Set

- We performed a retrospective study involving patients admitted to Shands Hospital (Gainesville, FL, USA) between January 1, 2000 and November 30, 2010.
- For each patient who underwent a surgery, detailed clinical and outcome data were collected.
- The analysis included **46,299** patients.

Other Covariates

- emergent surgery status
- type of surgical procedure
- age
- race
- gender
- intensive care unit (ICU) admission
- Charlson-Deyo comorbidity index (CCI)
- renal replacement therapy (RRT)
- time of surgery (days elapsed between admission and the surgical operation)
- operating surgeons' unique identifier

Two Cohorts of Patients

1. A subcohort of **7,766** patients with either
   - history of **chronic kidney disease (CKD)** prior to admission or
   - with the baseline **estimated glomerular filtration rate (eGFR)** of less than **60 ml/min/1.73 m2**.

   Since these patients may have **different sCr kinetics**, we performed a separate analysis for this cohort.
2. The remaining cohort consists of **38,533** patients.

Two Cohorts Comparison

Table 1

Baseline patients' characteristics stratified by ninety-day mortality.

	Baseline eGFR<60 ml/min/1.73 m^2 (n=7,766)		Baseline eGFR≥60 ml/min/1.73 m^2 (n=38,533)	
	Non-Survivors (n=848)	**Survivors (n=6,918)**	**Non-Survivors (n= 1,597)**	**Survivors (n= 36,936)**
Patients' Characteristics				
Age in years, median (25th, 75th)	73 (63, 80)	67 (55, 75) [a]	64 (52, 74)	54 (41, 65) [a]
Female gender, n (%)	360 (42)	3135 (45)	700 (44)	18,509 (50) [a]

Feature Extraction From sCr Time Series

- **Model0**: preoperative data only (**no sCr time series**). Used as a baseline.
- **Model1**: preoperative data and
  - the minimum of the sCr values available within six months of admission including the admission day value (**BaseCr**),
  - the absolute values for the maximum sCr (**MaxCr**), and
  - the last measured sCr during index hospitalization (**LastCr**).
- **Model2**: preoperative data and
  - **BaseCr**,
  - **MaxCr - BaseCr**,
  - **LastCr - BaseCr**.
- **Model3**: preoperative data and
  - **BaseCr**,
  - **MaxCr/BaseCr**,
  - **LastCr/BaseCr**.

Mortality Risk

Probabilistic score

$$\log \frac{P(C=1|X=x)}{P(C=0|X=x)} = \alpha + \sum_{i=1}^{m} f_i(x_i), \tag{1}$$

$C \in \{0,1\}$ is an outcome,
$X = (X_1, \ldots, X_m)$ - the risk factors,
$x = (x_1, \ldots, x_m)$ - the values of these factors,
f_i - is a nonlinear risk function associated with risk factor i, and
α is a free term.

Nonlinear Risk Factor Estimation

From the probabilistic score equation we can derive

$$\alpha + \sum_{i=1}^{m} f_i(x_i) = \log \frac{P(X = x|C = 1)}{P(X = x|C = 0)} + \log \frac{P(C = 1)}{P(C = 0)}. \quad (2)$$

Under assumption of risk factors independence:

$$\alpha + \sum_{i=1}^{m} f_i(x_i) = \sum_{i=1}^{m} \log \frac{P(X_i = x_i|C = 1)}{P(X_i = x_i|C = 0)} + \log \frac{P(C = 1)}{P(C = 0)}. \quad (3)$$

Nonlinear Risk Factor Estimation

$$P(X_i|C = c) = \frac{\#\{j : C^j = c, x_i^j = x\}}{\#\{j : C^j = c\}} \tag{4}$$

For continuous factors, $P(X_i = x_i|C)$ implies $P(X_i \in [x_i - \epsilon_i, x_i + \epsilon_i])$, where ϵ_i's are chosen appropriately.

- Nonlinear risk functions f_i were estimated with cubic splines via a local scoring algorithm.
- This back fitting algorithm uses an iterative method which approximates the risk function, where the change in residuals is less than a prespecified threshold.
- The degrees of freedom for each spline were estimated by maximizing a restricted likelihood function.
- From the initial data set the predetermined set of variables was used for backward-stepwise selection based on AIC criteria.

Degrees of freedom for f_i's after backward-stepwise selection

	Baseline eGFR≥60 ml/min/1.73 m^2				Baseline eGFR<60 ml/min/1.73 m^2			
Variables	Model 0[a]	Model 1[b]	Model 2[c]	Model 3[d]	Model 0[a]	Model 1[b]	Model 2[c]	Model 3[d]
Categorical or Nominal								
Type of surgical procedure	8.4	8.6	8.6	8.5	8.2	1.6	1.6	1.2
Emergent surgery	1	1	1	1	1	1	1	1
Gender	1	1	1	1	1	1	1	1
Race	1	1	1	1	1	1	1	1
Operating surgeon	1.8	1.4	1.5	1.7	1	1.5	1.9	1.6
Intensive care unit admission	1	1	1	1	1	1	1	1
Charlson comorbidity index	5.2	4.9	5.1	5.2	4.5	5.1	1	1
Time of surgery	X	X	X	X	X	X	X	X
Renal Replacement therapy	NA	NA	NA	NA	NA	NA	NA	NA
Continuous								
Age	3.8	3.9	3.9	3.9	3.6	4.1	4.0	3.75
BaseCr		1.7	2.4	X		3.4	1	X
MaxCr		6				4.3		
LastCr		7.1				3.8		
MaxCr - BaseCr			5				4.2	
LastCr - BaseCr			7				4.1	
MaxCr/BaseCr				5.4				2.5
LastCr/BaseCr				7				5.9

The cohort excludes patients requiring renal replacement therapy. All values represent degree of freedom except for variables with the symbol "X" that were discarded during the feature selection procedure.

Final Models

Risk factors with estimated degrees of freedom close to 1 were not smoothed in a final model, instead the original values of risk factors were used. Therefore, the final model has the following form:

$$\log \frac{P(C=1|X=x)}{P(C=0|X=x)} = \alpha + \sum_{i \in I} w_i x_i + \sum_{i \notin I} f_i(x_i), \tag{5}$$

where I is a set of risk factors with estimated degrees of freedom close to 1.

Classification Accuracy Evaluation

- 70/30 cross-validation analysis has been performed
- Random split into 70% training subset and 30% testing subset
- The average over 100 runs results are reported

Resulting Non-linear Risk Functions

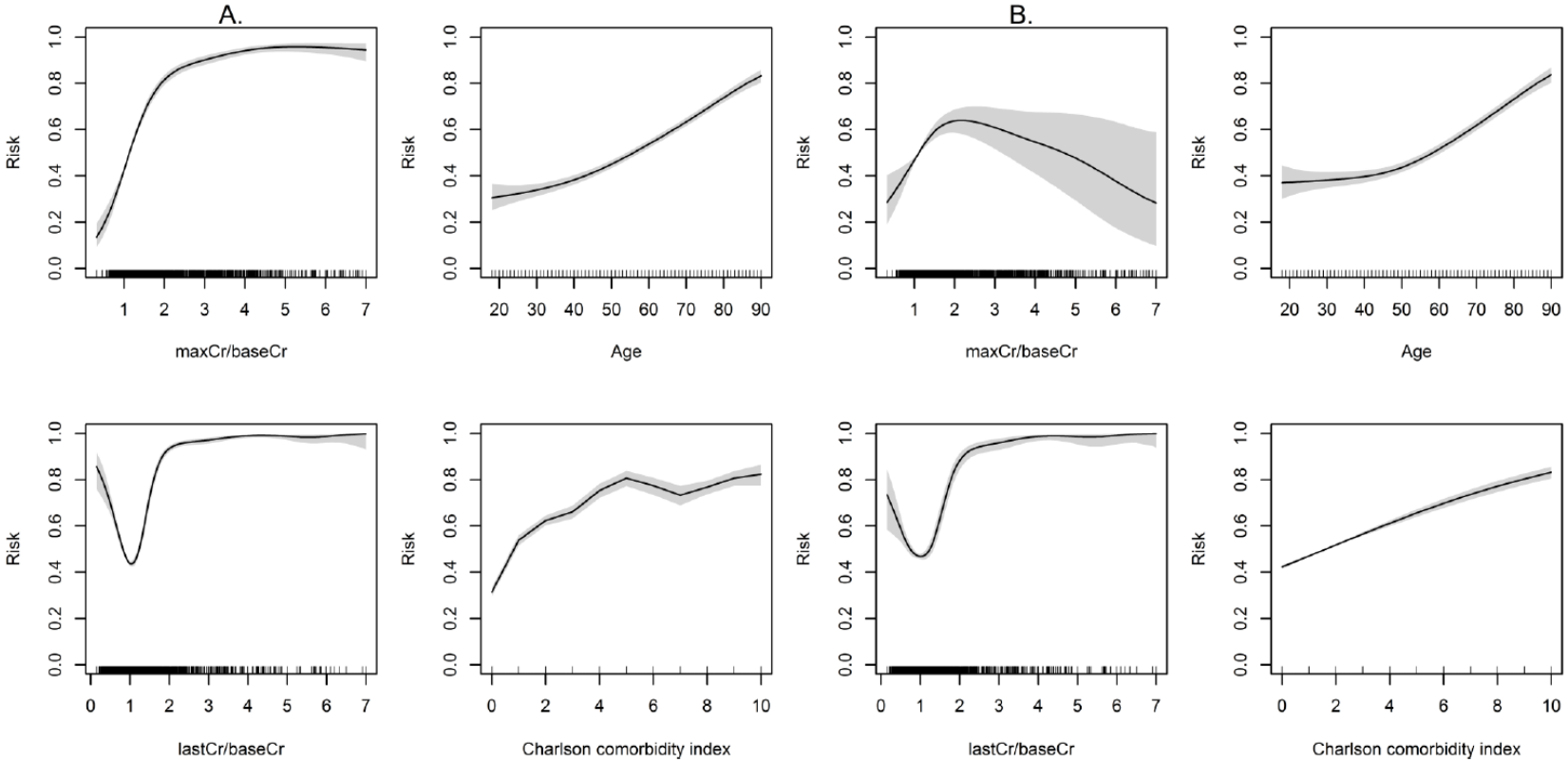


Figure: Nonlinear functions for the association between maxCr/baseCr, llastCr/baseCr, age, and CCI and ninety-day mortality for the patients with chronic kidney disease or baseline estimated glomerular filtration rate < 60 ml/min/1.73 m2 after excluding patients on renal replacement therapy. Panels (A) and (B) show unadjusted and adjusted nonlinear functions, respectively. The Y axis represents risk probability for ninety-days mortality ranging from 0 to 1. The shaded areas represent 95% prediction intervals for the function values.

Resulting Non-linear Risk Functions

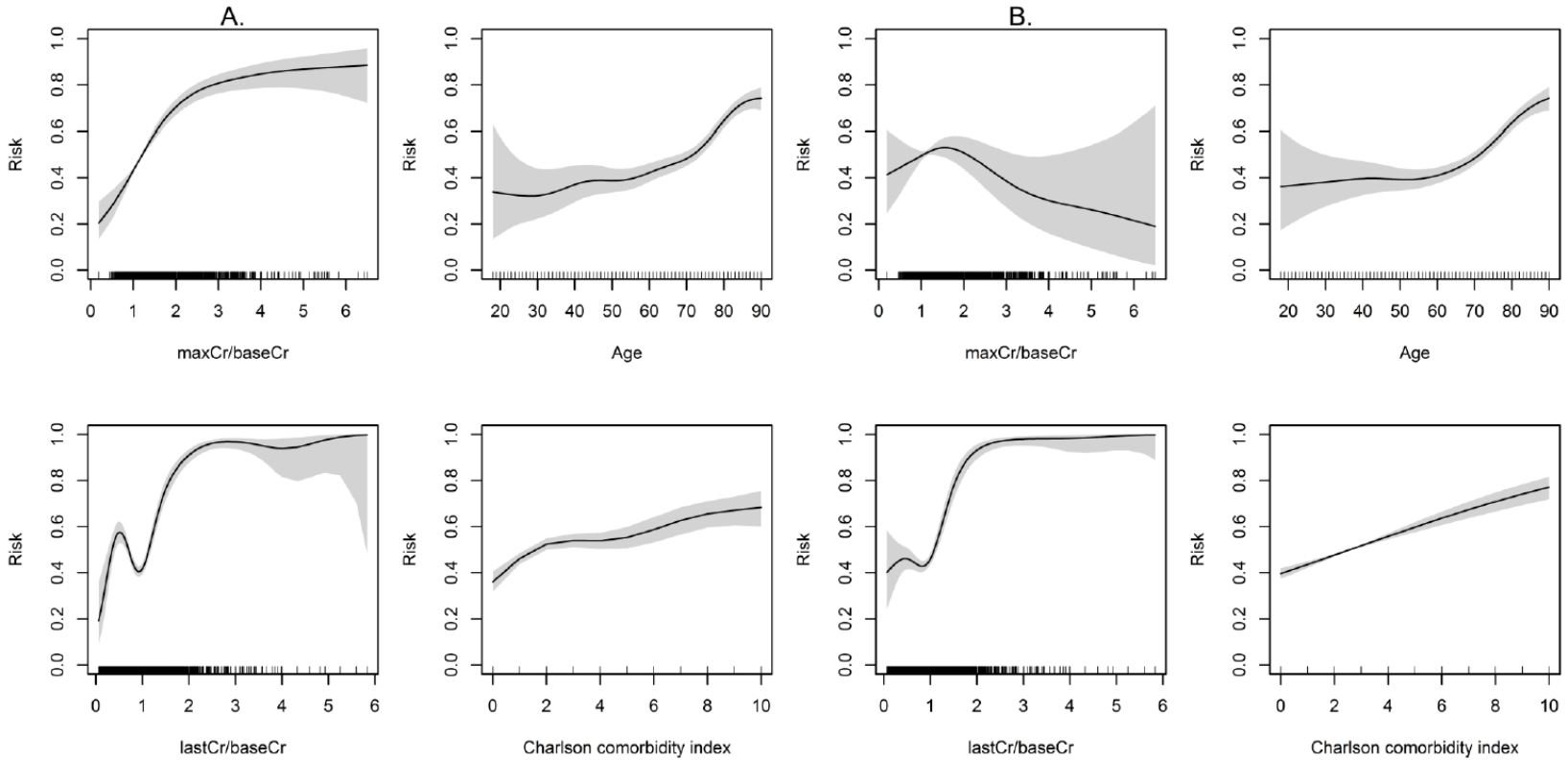


Figure: Nonlinear functions for the association between maxCr/baseCr, lastCr/baseCr, age, and CCI and ninety-day mortality for the patients with baseline EGFT $\geq$ 60 ml/min/1.73 m2 after excluding patients on renal replacement therapy. Panels (A) and (B) show unadjusted and adjusted nonlinear functions, respectively. The Y axis represents risk probability for ninety-days mortality ranging from 0 to 1. The shaded areas represent 95% prediction intervals for the function values.

Association Between pattern in sCr change and 90-day mortality

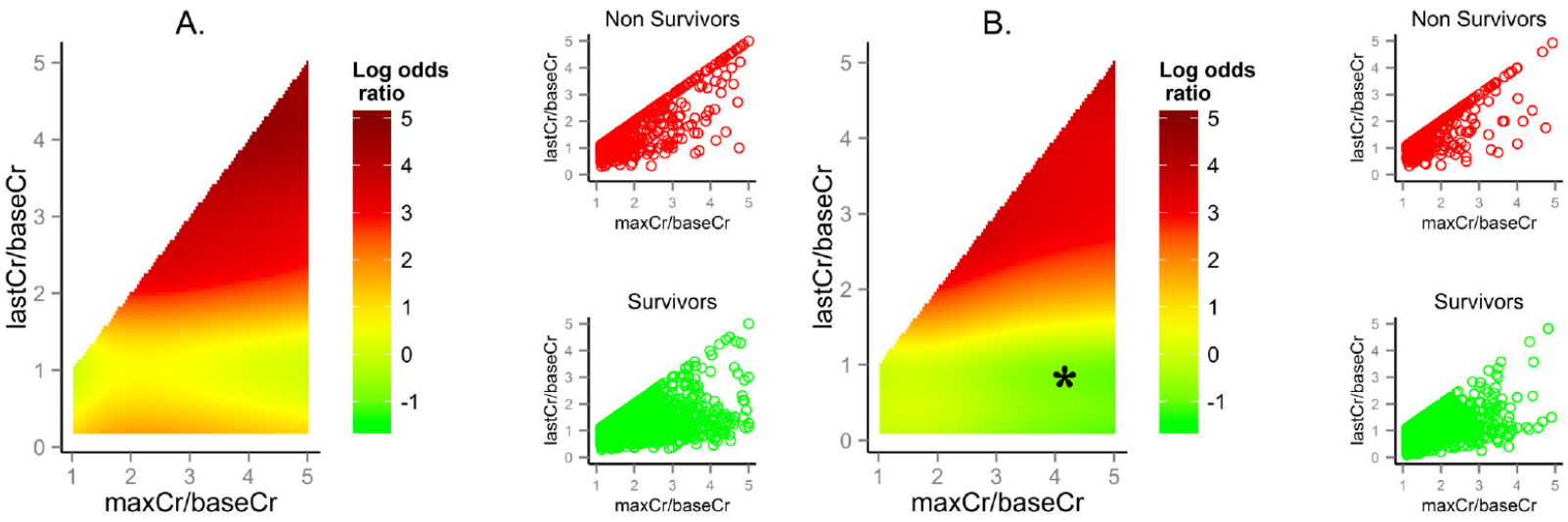


Figure: Association between pattern in sCr change and 90-day mortality adjusted for preoperative clinical factors among (A) patients with baseline estimated glomerular filtration rate $\geq$ 60 ml/min/1.73 m2 and (B) patients with chronic kidney disease or baseline estimated glomerular filtration rate $<$ 60 ml/min/1.73 m2. Left panel: The log odds ratios for 90-day mortality based on adjusted nonlinear functions for maxCr/baseCr and lastCr/baseCr with respect to a pattern with no change in sCr. Right panel: Distribution of maxCr/baseCr and lastCr/baseCr values for non-survivors and survivors. *There were no patients in the dataset with the combination of the maxCr/baseCr and lastCr/baseCr in the marked range.

References

[1] Korenkevych, D., Ozrazgat-Baslanti, T., Thottakkara, P., Hobson, C. E., Pardalos, P., Momcilovic, P., & Bihorac, A. (2016). The pattern of longitudinal change in serum creatinine and 90-day mortality after major surgery. Annals of Surgery, 263(6), 1219-1227.

[2] Rosenthal, M., Korenkevych, D., Baslanti, T. O., Webel, A., Glerum, A., Momcilovic, P., ... & Bihorac, A. (2013). 703: Algorithms can accurately predict risk for major postoperative complications using preoperative data. Critical Care Medicine, 41(12), A173.

[3] Bihorac, A., Korenkevych, D., Baslanti, T. O., Momcilovic, P., Pardalos, P., Segal, M., & Moore, F. (2013). 714: Database Communication Enables Machine Learning Classifiers to Predict Postoperative AKI in ICU. Critical Care Medicine, 41(12), A176.

[4] Baslanti, T. O., Korenkevych, D., Momcilovic, P., Pardalos, P., Hobson, C., & Bihorac, A. (2012). Mathematical modeling of the association between the pattern of change in postoperative serum creatinine and hospital mortality. In CRITICAL CARE MEDICINE (Vol. 40, No. 12, pp. U131-U131). 530 WALNUT ST, PHILADELPHIA, PA 19106-3621 USA: LIPPINCOTT WILLIAMS & WILKINS.

Study Outline

Objective:

- To compare performance of risk prediction models for forecasting postoperative sepsis and acute kidney injury.

Design:

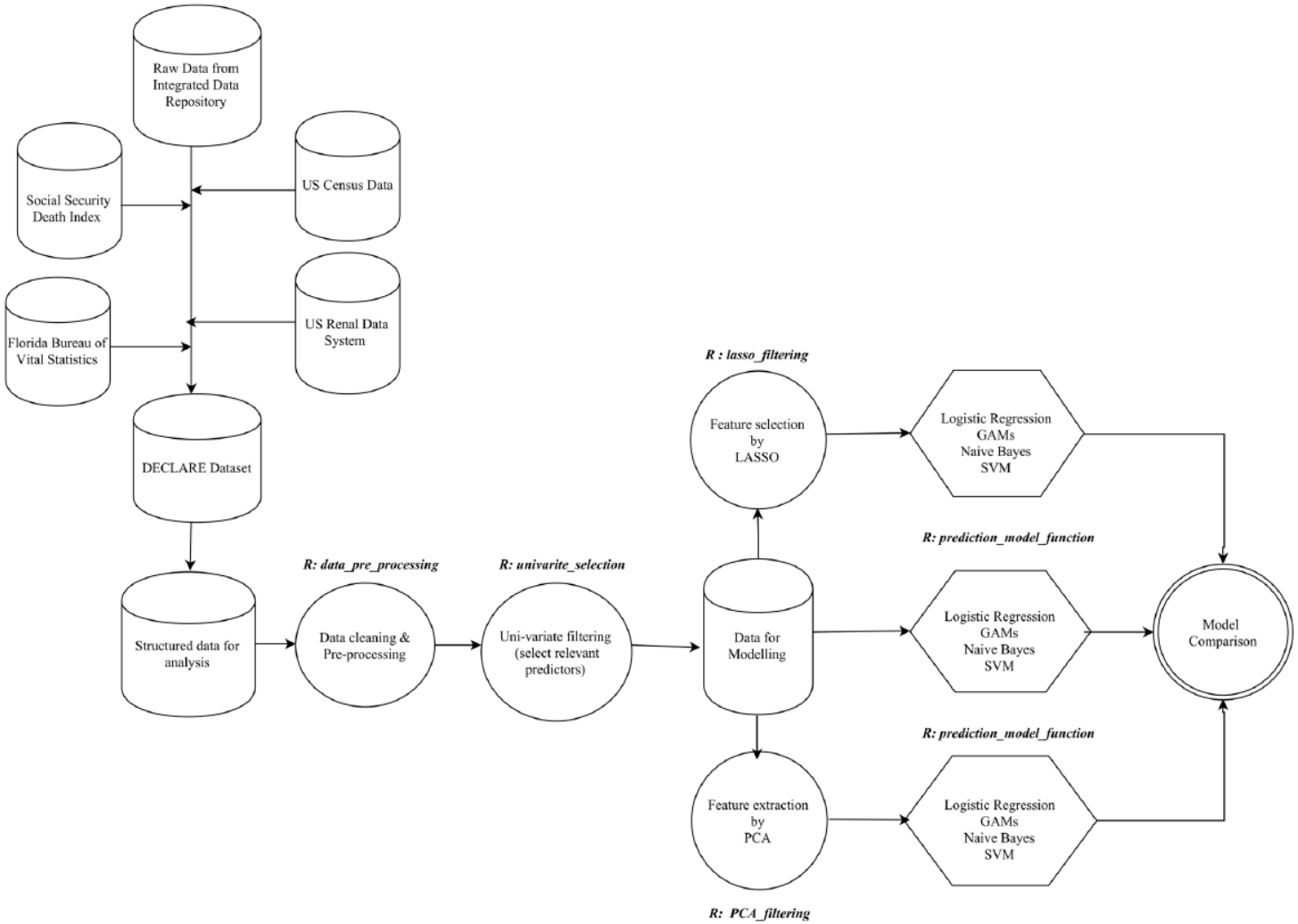
- Retrospective single center cohort study of adult surgical patients admitted between 2000 and 2010.
- 50,318 adult patients undergoing major surgery.

Study Outline

Algorithms:

- Logistic Regression (LR)
- Generalized Additive Models (GAMs)
- Naïve Bayes (NB) and
- Support Vector Machines (SVMs)

Development Flow from Raw Data to Model Building



Model Comparison

Model	Acute Kidney Injury			Severe Sepsis		
	Accuracy (95% CI)	AUC (95% CI)	PPV (95% CI)	Accuracy (95% CI)	AUC (95% CI)	PPV (95% CI)
Logistic Regression Model	0.752 (0.746,0.758)	0.824 (0.818,0.828)[b]	0.725 (0.714,0.737)	0.773 (0.762,0.781)	0.851 (0.840,0.8560)	0.811 (0.785,0.833)
GAMs	0.756 (0.751,0.761)	0.827 (0.821,0.832)[a]	0.719 (0.706,0.729)	0.775 (0.766,0.783)	0.852 (0.840,0.863)	0.806 (0.779,0.832)
Naïve Bayes Model	0.744 (0.738,0.749)	0.797 (0.791,0.803)[a,b]	0.545 (0.534,0.558)	0.805 (0.798,0.811)	0.83 (0.819,0.841)[a,b]	0.689 (0.659,0.716)
SVM	0.767 (0.757,0.774)	0.819 (0.811,0.828)[a,b]	0.662 (0.648,0.676)	0.71 (0.689,0.731)	0.762 (0.733,0.782)[a,b]	0.677 (0.619,0.722)
After feature selection with LASSO						
Logistic Regression Model	0.753 (0.747,0.757)	0.824 (0.818,0.830)[b]	0.726 (0.714,0.738)	0.772 (0.760,0.780)	0.85 (0.838,0.863)[b]	0.812 (0.781,0.838)
GAMs	0.757 (0.752,0.762)	0.828 (0.822,0.833)[a]	0.72 (0.706,0.732)	0.774 (0.766,0.780)	0.851 (0.842,0.862)	0.806 (0.783,0.831)
Naïve Bayes Model	0.744 (0.737,0.750)	0.797 (0.789,0.804)[a,b]	0.545 (0.533,0.556)	0.806 (0.800,0.813)	0.831 (0.817,0.841)[a,b]	0.69 (0.659,0.711)
SVM	0.767 (0.759,0.774)	0.82 (0.812,0.829)[a,b]	0.665 (0.646,0.685)	0.697 (0.684,0.713)	0.757 (0.736,0.779)[a,b]	0.689 (0.652,0.732)
After feature extraction with 5 principal components						
Logistic Regression Model	0.774 (0.769,0.781)	0.853 (0.849,0.859)[a,b]	0.758 (0.746,0.767)	0.818 (0.809,0.824)	0.904 (0.895,0.913)[a,b]	0.854 (0.841,0.880)
GAMs	0.773 (0.768,0.777)	0.858 (0.853,0.862)[a,b]	0.784 (0.771,0.793)	0.826 (0.819,0.833)	0.909 (0.902,0.917)[a,b]	0.86 (0.843,0.878)
Naïve Bayes Model	0.741 (0.735,0.747)	0.819 (0.814,0.826)[a,b]	0.666 (0.651,0.677)	0.805 (0.797,0.815)	0.882 (0.874,0.890)[a,b]	0.839 (0.822,0.866)
SVM	0.777 (0.767,0.782)	0.857 (0.850,0.862)[a,b]	0.735 (0.725,0.750)	0.85 (0.737,0.897)	0.877 (0.828,0.904)[a,b]	0.751 (0.667,0.850)

[a] $p<0.05$ for AUC comparison with respect to logistic regression model without any data reduction.
[b] $p<0.05$ for AUC comparison with respect to GAMs model without any data reduction.

GAMs: Predicted Risk Functions

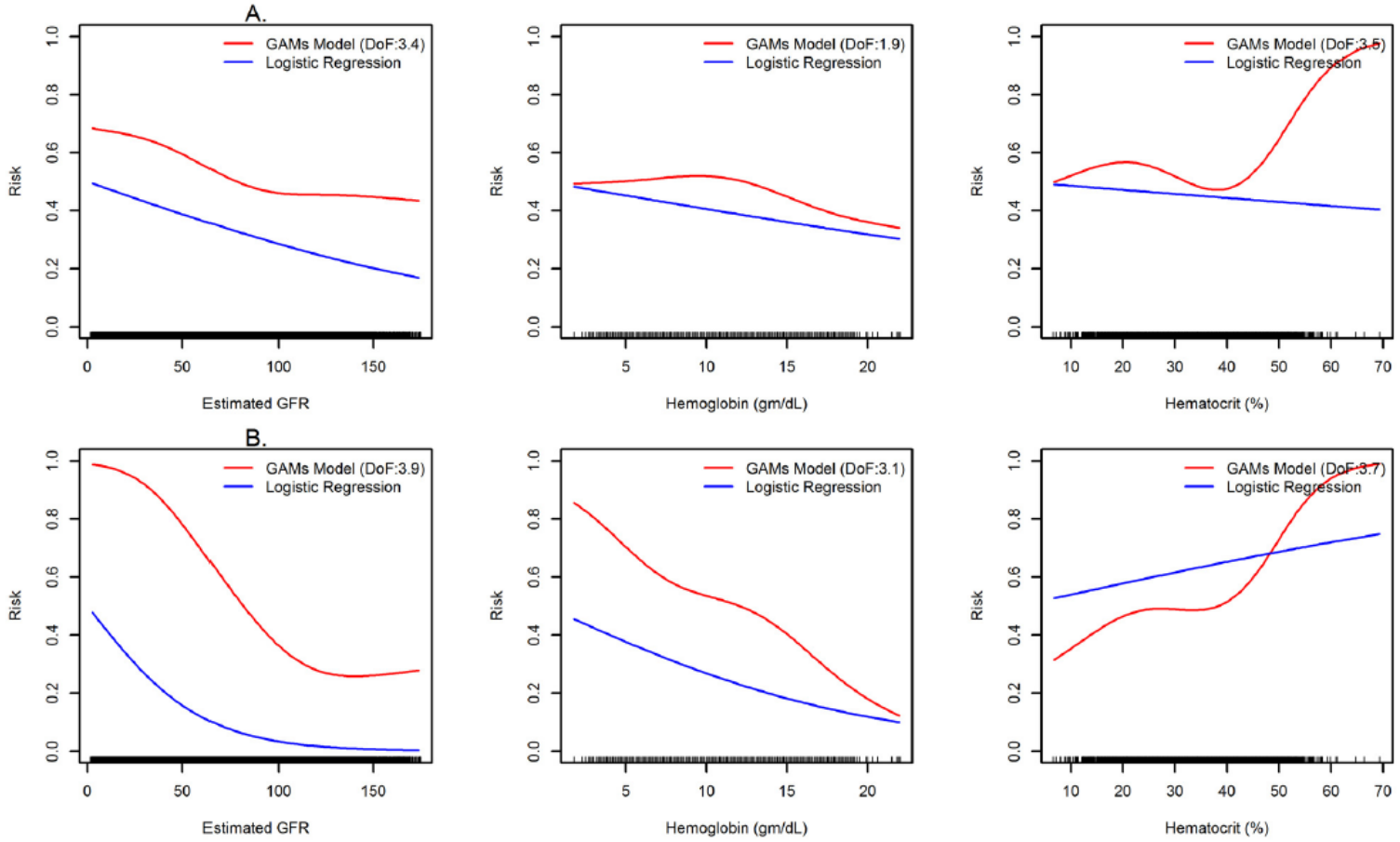


Figure: Predicted risk functions for the association between (A) acute kidney injury and (B) severe sepsis and continuous variables. Risk functions were generated from multivariate generalized additive models and logistic regression models. GAM, generalized additive model; DoF, degree of freedom; GFR, glomerular filtration rate.

Results

- AUC ROC for different models ranged between 0.797 and 0.858 for AKI and between 0.757 and 0.909 for severe sepsis.
- LR, GAM, and SVM had better performance compared to NB model
- GAMs additionally accounted for non-linearity of continuous clinical variables as depicted in their risk patterns plots
- Reducing the input feature space with LASSO had minimal effect on prediction performance, while feature extraction using principal component analysis improved performance of the models.

References

[1] Thottakkara, P., Ozrazgat-Baslanti, T., Hupf, B. B., Rashidi, P., Pardalos, P., Momcilovic, P., & Bihorac, A. (2016). Application of machine learning techniques to high-dimensional clinical data to forecast postoperative complications. PloS one, 11(5), e0155705.

Objectives

Objectives:

- Utilize information hidden in multivariate time series (heart rate, blood pressure, etc.) of physiological recordings collected during a surgical procedure.
- A faster algorithm for pattern mining form multivariate time series for on-line applications.

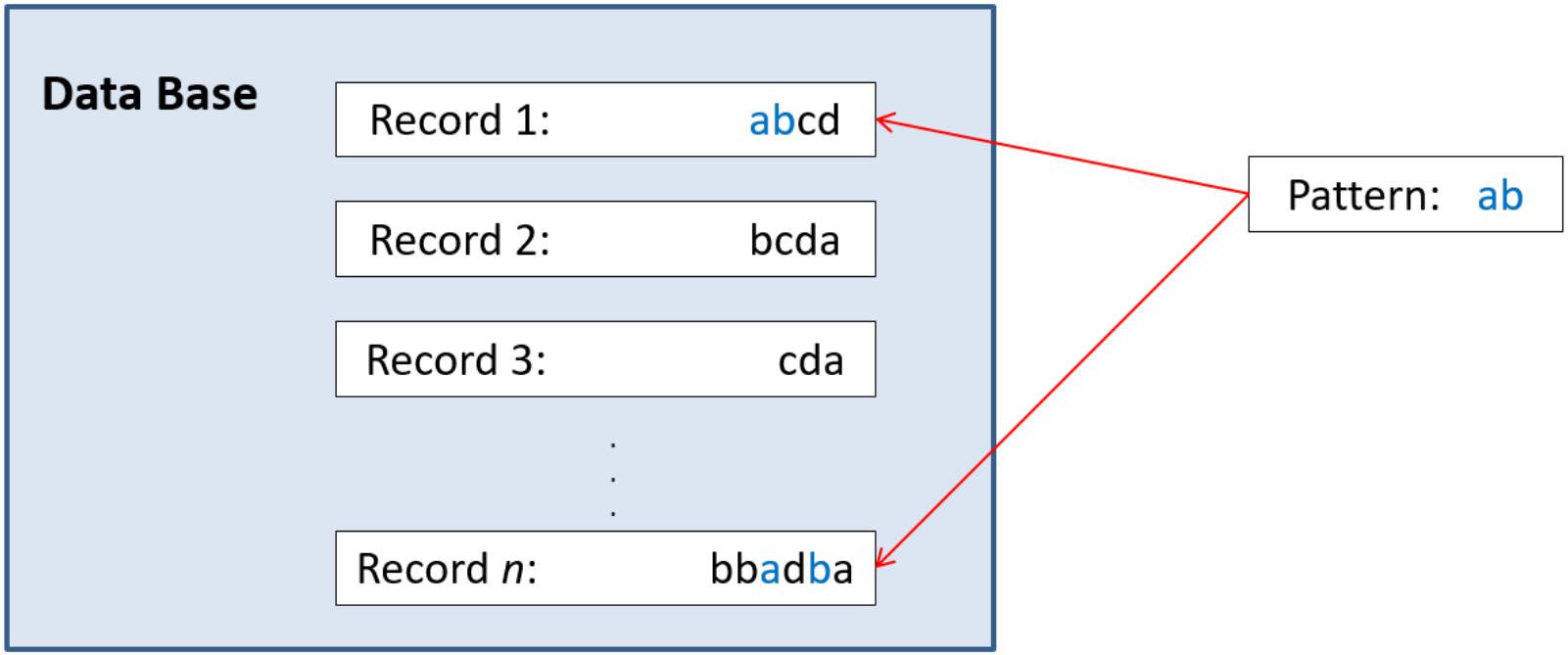
Augmented AKI Data Set

- For **5202** patients, intra-operative recordings were available
- We decided to concentrate on two time series:
  - Blood Pressure
  - Heart Rate

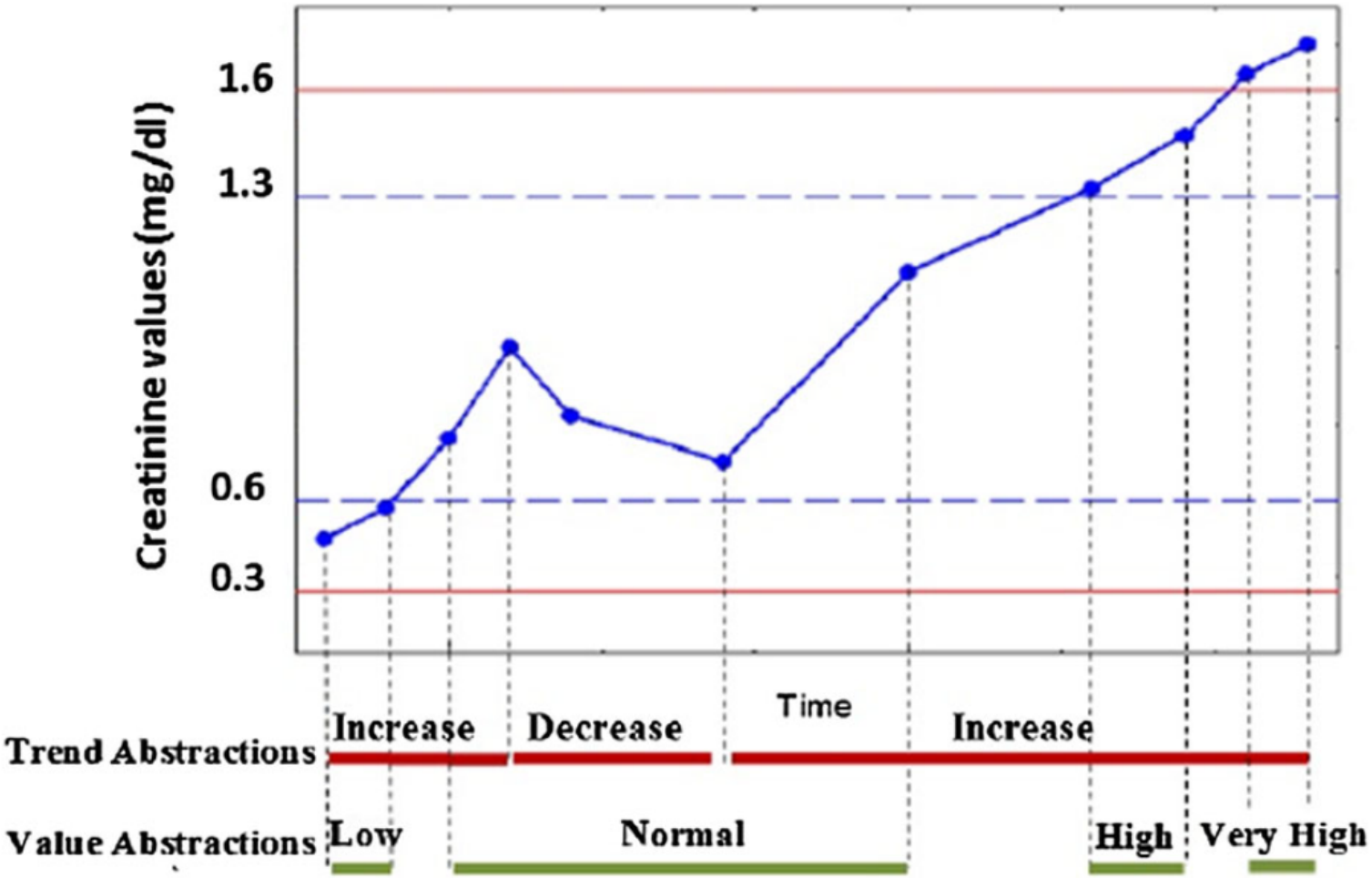
Intersection of two Problems

- Time Series Classification
  - Data is a collection of records, where each record consists of a number of time series and an outcome associated with it
  - Time series are sampled unevenly in time
  - Time series can be both numerical and categorical
- Pattern Mining as Class-Specific Feature Extraction

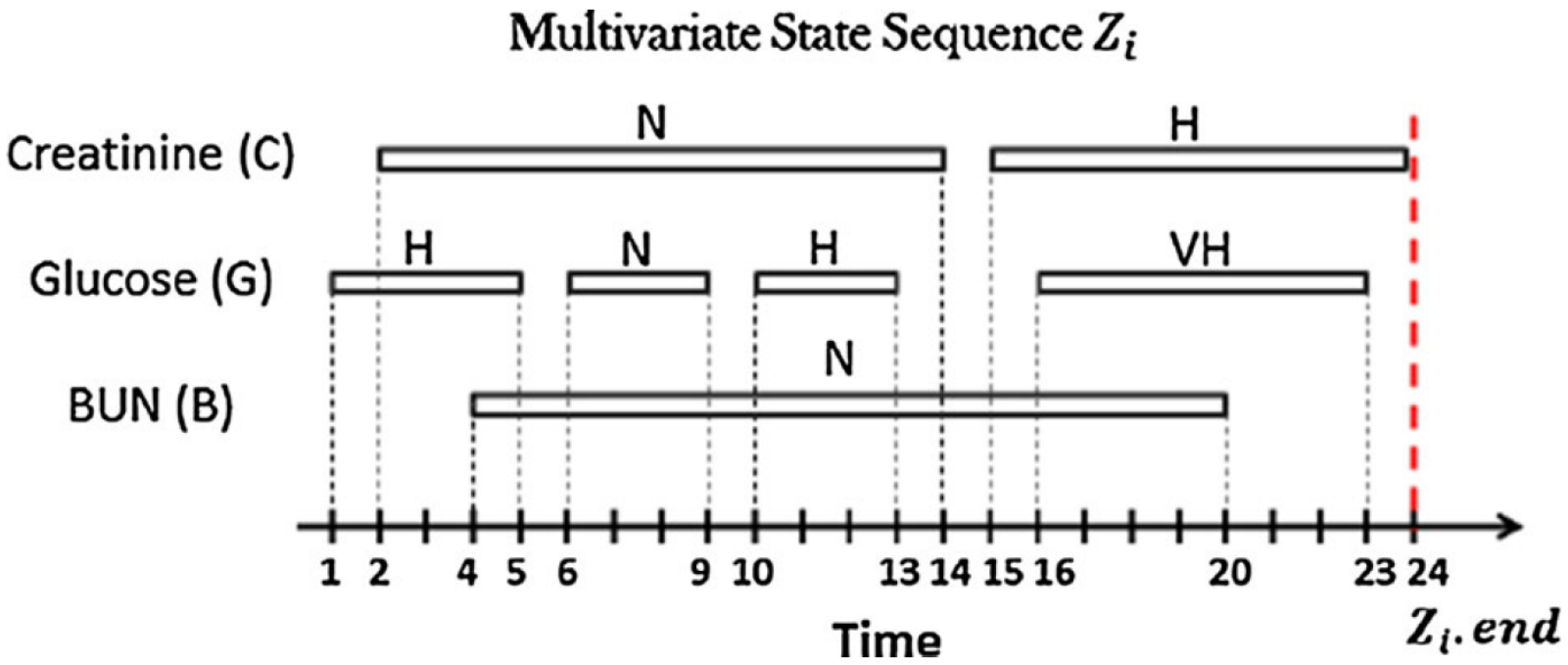
Mining Frequent Patterns



From Time Series to Temporal Abstractions



From Temporal Abstractions to Multivariate State Sequences



Multivariate State Sequences: Definitions

- $S = (F, V)$ is a **state** where is F is a variable label and $V \in \Sigma$ is an abstraction value.
- $E = (F, V, s, e)$ is a **state interval** where (F, V) is a state, and s and e are the start and end times of the state interval.
- $Z = \langle E_1, \ldots, E_l \rangle,\ E_i.s \leq E_{i+1}.s, 1 \leq i \leq l-1$, is a **Multivariate State Sequence (MSS)** consisting of l state intervals which are arranged in the non-decreasing order of their start times.

Temporal Patterns

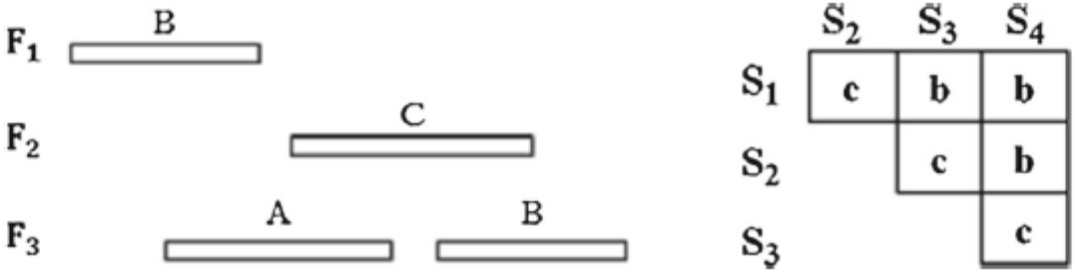
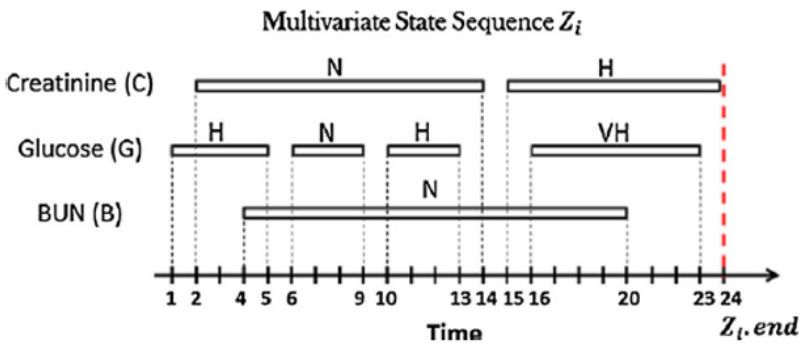


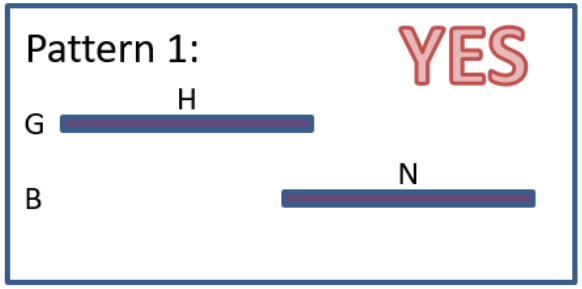
Fig. 4 A temporal pattern with states $\langle (F_1, B), (F_3, A), (F_2, C), (F_3, B) \rangle$ and temporal relations $R_{1,2} = \mathbf{c}$, $R_{1,3} = \mathbf{b}$, $R_{1,4} = \mathbf{b}$, $R_{2,3} = \mathbf{c}$, $R_{2,4} = \mathbf{b}$ and $R_{3,4} = \mathbf{c}$

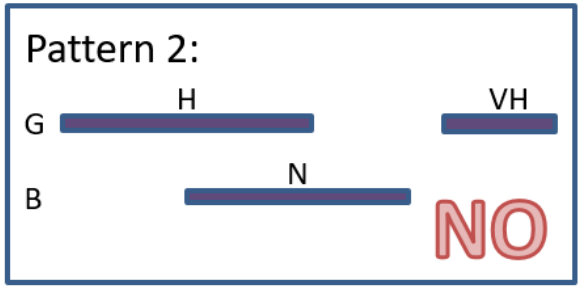
Temporal Pattern Definition

- $P = (\langle S_1, \ldots, S_k \rangle, R)$ is a **Temporal Pattern** (TP) of length k ($|P| = k$) with states $S_1, \ldots, S_k$, where R is a (upper-triangular) matrix describing pair-wise temporal relationships between the states: $R_{i,j} \in \{b, c\},\ 1 \leq i < j \leq k$.
- Given an MSS $Z = \langle E_1, E_2, \ldots, E_l \rangle$ and a temporal pattern $P = (\langle S_1, \ldots, S_k \rangle, R)$, we say that Z **contains** P, denoted as $P \in Z$, if there is an injective mapping $\pi : \{1, \ldots, k\} \rightarrow \{1, \ldots, l\}$ ($k \leq l$) that matches every state S_i in P to a state interval $E_{\pi(i)}$ in Z such that:
  1. $S_i.F = E_{\pi(i)}.F$ and $S_i.V = E_{\pi(i)}.V,\ 1 \leq i \leq k$,
  2. $\pi(i) < \pi(j),\ i < j$,
  3. $R(E_{\pi(i)}, E_{\pi(j)}) = R_{i,j},\ i < j$.

Temporal Patterns in MSSs

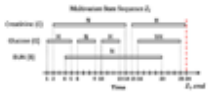


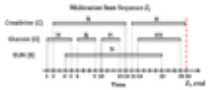


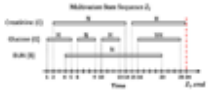


Frequent Temporal Patterns

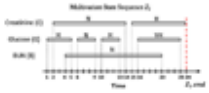
Class A:

record 1  YES

record 2  YES

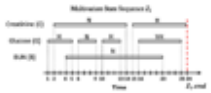
record 3  NO

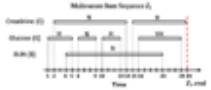
⋮

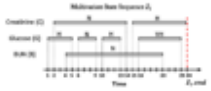
record k_A  YES

$$\frac{\#\text{"yes answers"}}{k_A} < \Theta$$

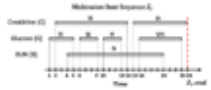
Class B:

record 1  NO

record 2  YES

record 3  NO

⋮

record k_B  YES

$$\frac{\#\text{"yes answers"}}{k_B} \geq \Theta$$

frequent pattern

Frequent Temporal Pattern Mining (FTPM)

Input: D, $1\text{-}FTPs$
Output: $FTPs$
$FTPs \leftarrow 1\text{-}FTPs$;
$k \leftarrow 1$;
while $|k\text{-}FTPs| > 0$ ***and no*** *other stopping criteria are met* **do**
  $(k+1)\text{-}FTPs \leftarrow \emptyset$;
  $(k+1)\text{-}candidates \leftarrow CreateCandidates(k\text{-}FTPs,\ 1\text{-}FTPs)$;
  forall the $TP \in (k+1)\text{-}candidates$ **do**
    if TP *is* FTP *in* D **then**
      $(k+1)\text{-}FTPs \leftarrow (k+1)\text{-}FTPs \cup \{TP\}$;
    end
  end
  $FTPs \leftarrow FTPs \cup (k+1)\text{-}FTPs$;
  $k \leftarrow k+1$;
end

Algorithm 1: FTPM algorithm : the general framework.

References

[1] Batal, I. et al. An efficient pattern mining approach for event detection in multivariate temporal data. Knowl Inf Syst 46, 115–150 (2016).

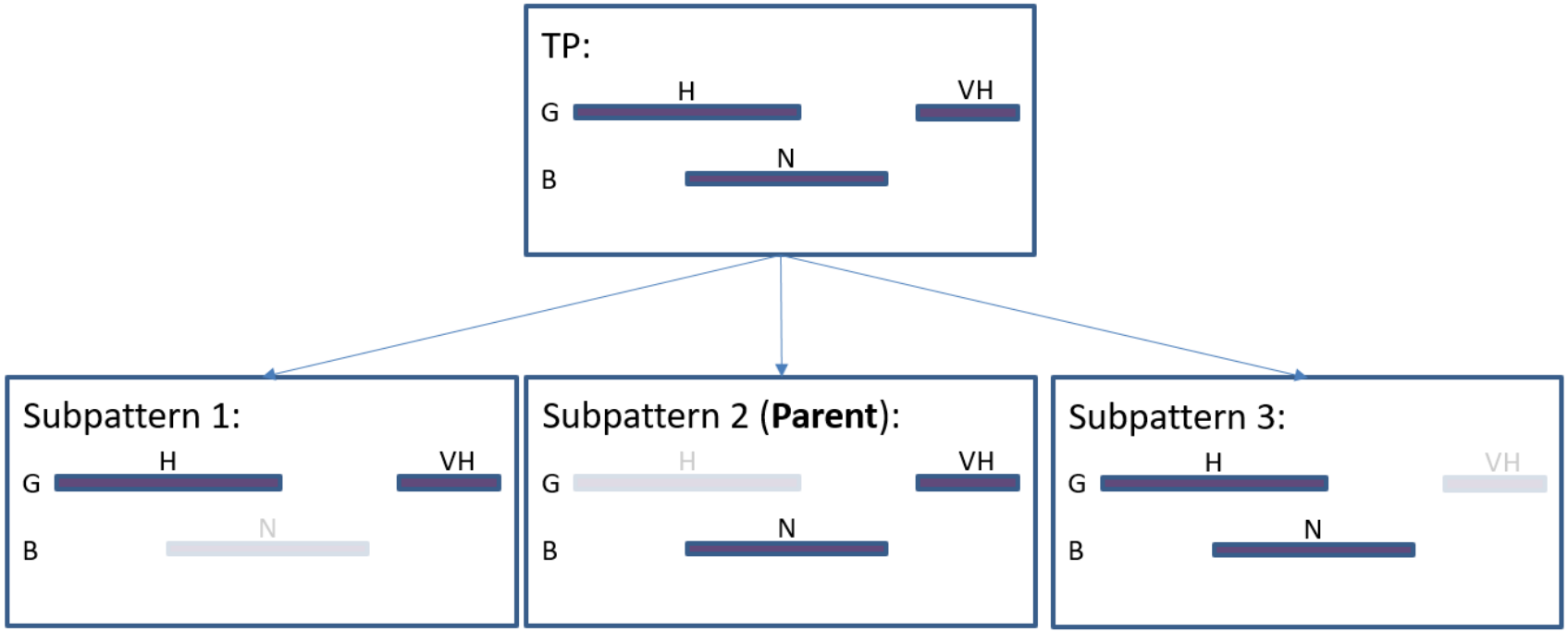
[2] Batal, I. et al. A Pattern Mining Approach for Classifying Multivariate Temporal Data. Proceedings (IEEE Int Conf Bioinformatics Biomed), 358–365 (2011).

[3] Batal, I. et al. Multivariate time series classification with temporal abstractions. Proceedings of the Twenty-Second International FLAIRS Conference, 344-349 (2009).

Apriory Property

Apriory Property:

$$P.p\_ids = \bigcap_{\widetilde{P} \in \mathrm{sub}(P)} \widetilde{P}.ids = \bigcap_{\widetilde{P} \in \mathrm{sub}_k(P)} \widetilde{P}.ids$$



Frequent Temporal Pattern Mining with Extended Lists (FTPMwEL)

Main Idea:

- Store locations where the first state of the pattern appears inside each record

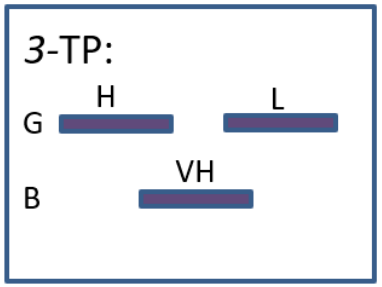
Definitions:

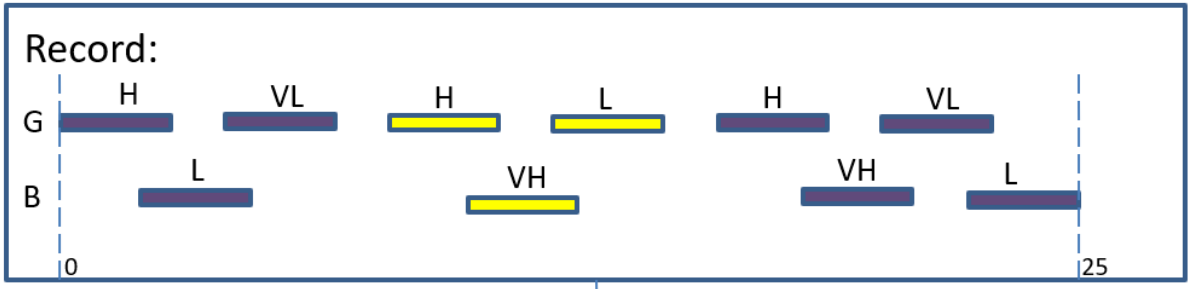
- Let $P.pos[i]$ and $P.p\_pos[i]$ denote all **positions** and **potential positions** at which P appears and may appear inside Z_i, respectfully.
- Thus, $P.ppos[i]$ can be computed as the following:

$$P.p\_pos[i] = \bigcap_{\widetilde{P} \in X} \widetilde{P}.pos[i],$$

where $X = \mathrm{sub}_k(P) \setminus \mathrm{parent}(P)$.

Extended Lists: Example

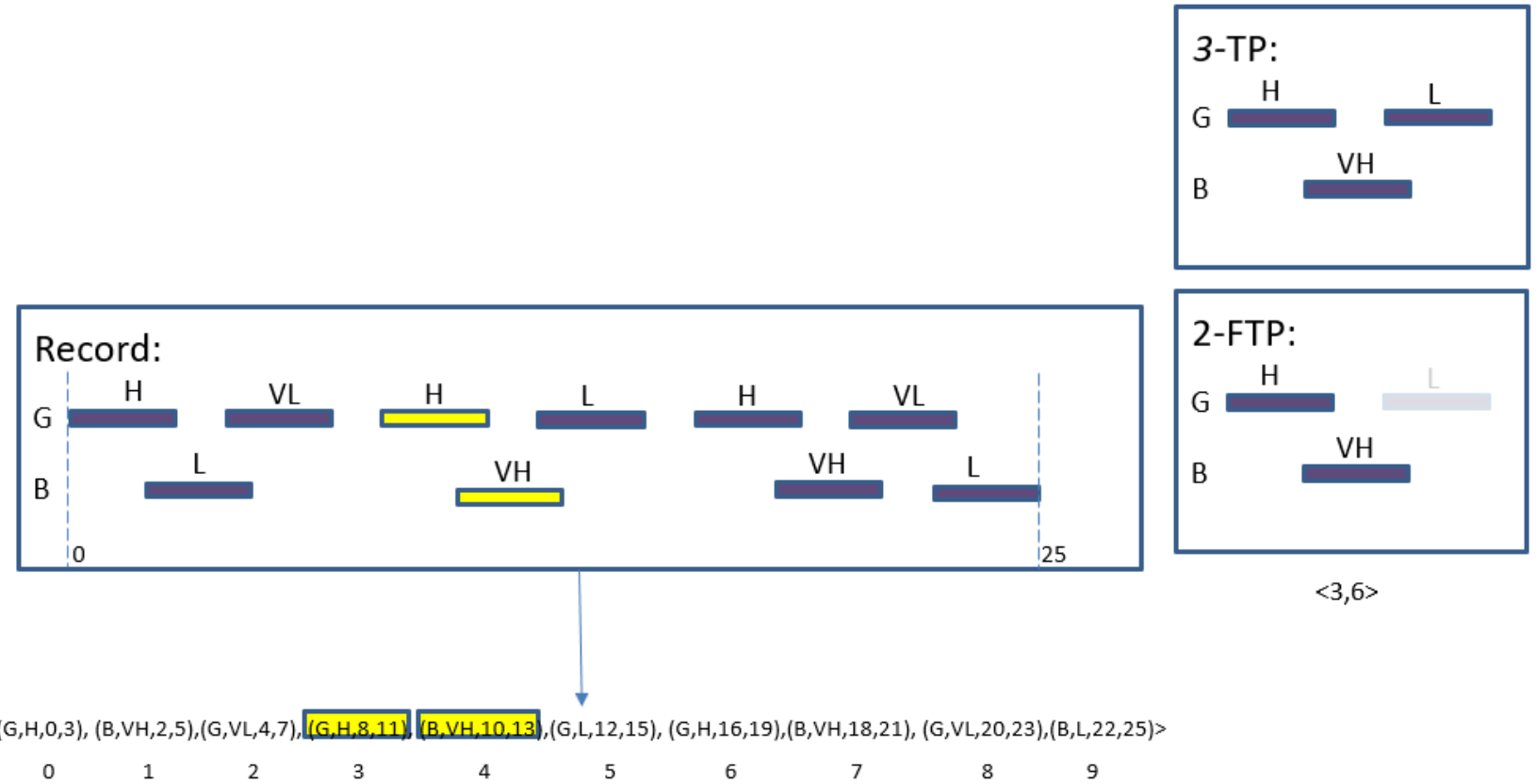




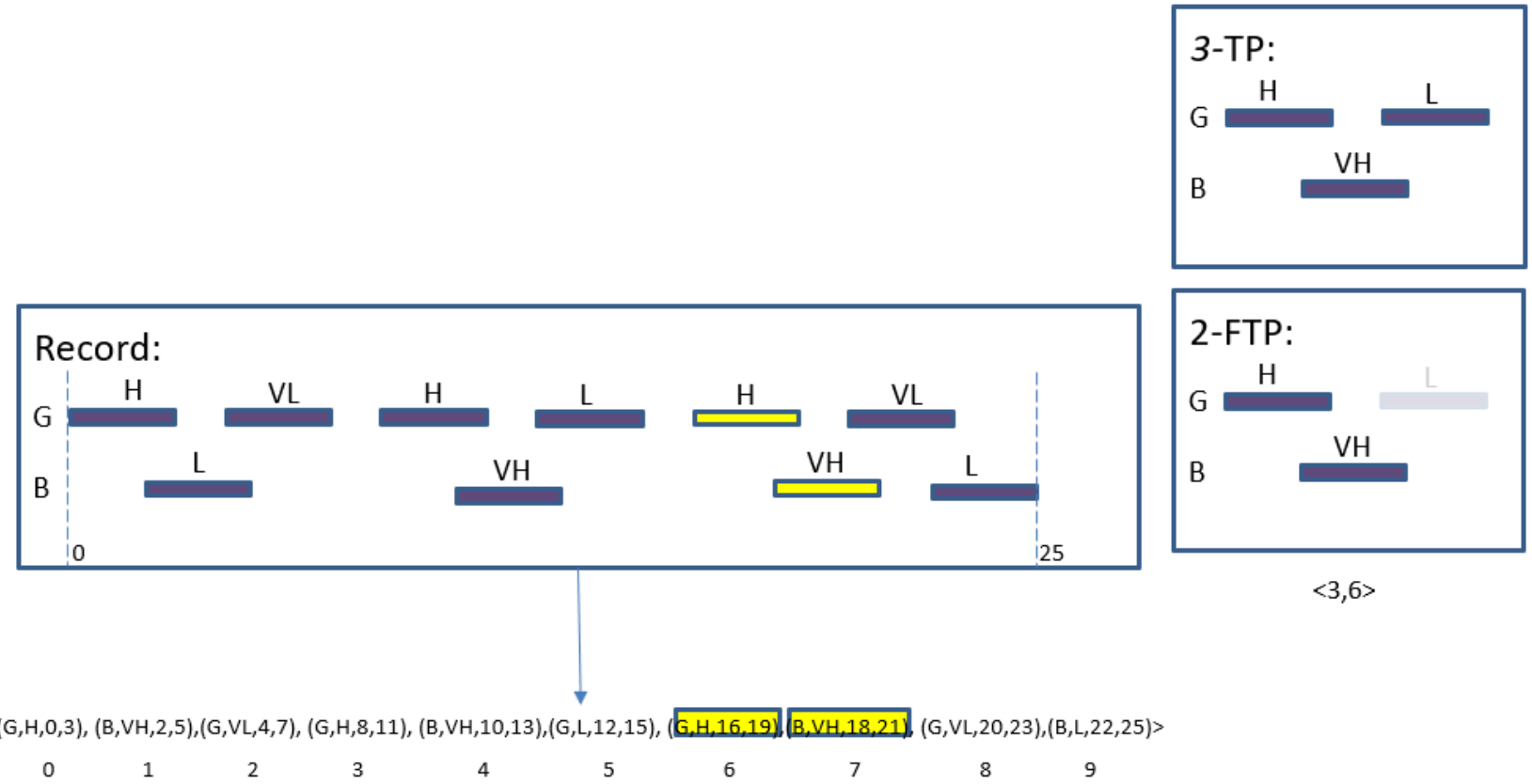
<(G,H,0,3), (B,VH,2,5),(G,VL,4,7), (G,H,8,11), (B,VH,10,13), (G,L,12,15), (G,H,16,19),(B,VH,18,21), (G,VL,20,23),(B,L,22,25)>

0 1 2 3 4 5 6 7 8 9

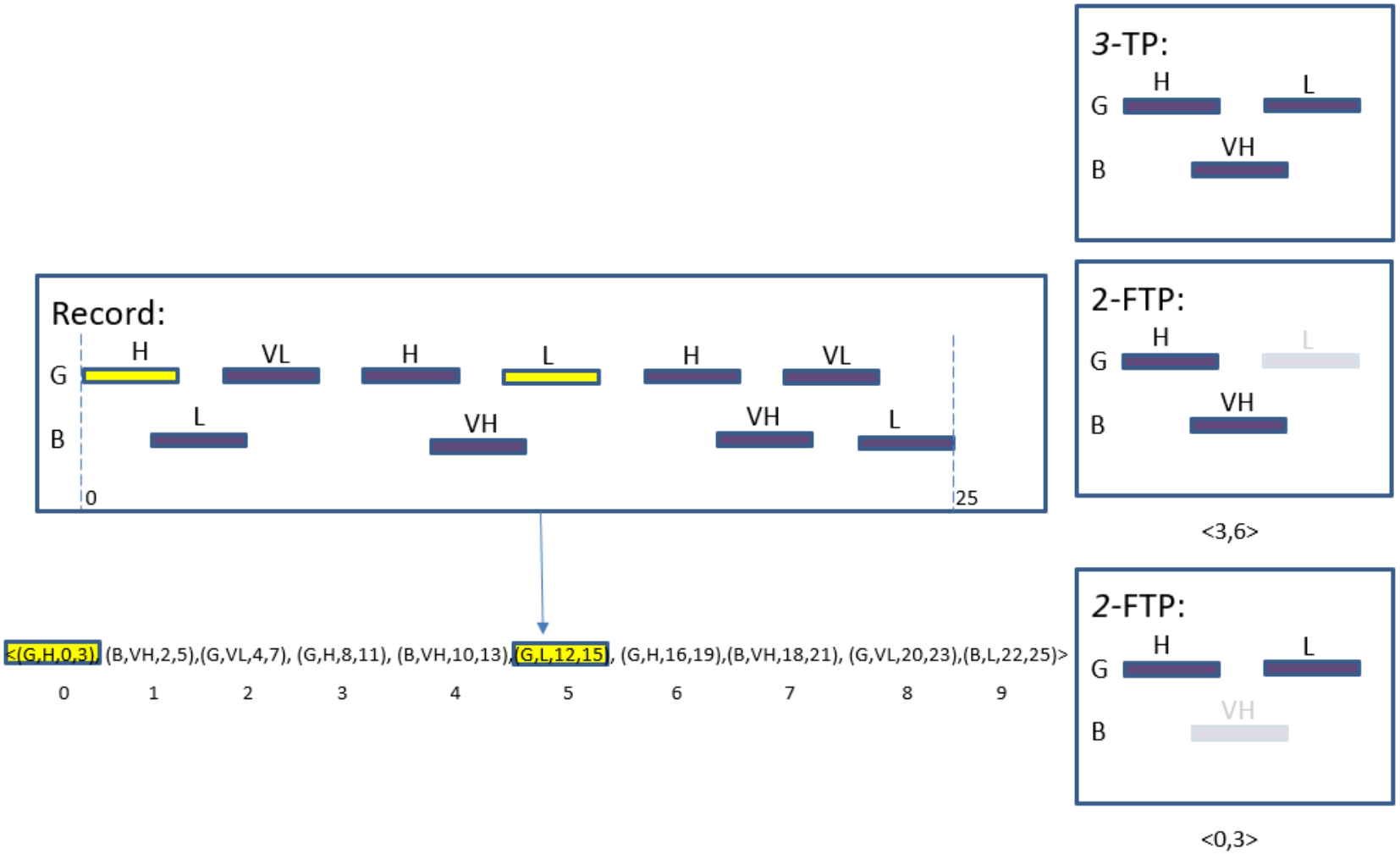
Extended Lists: Example



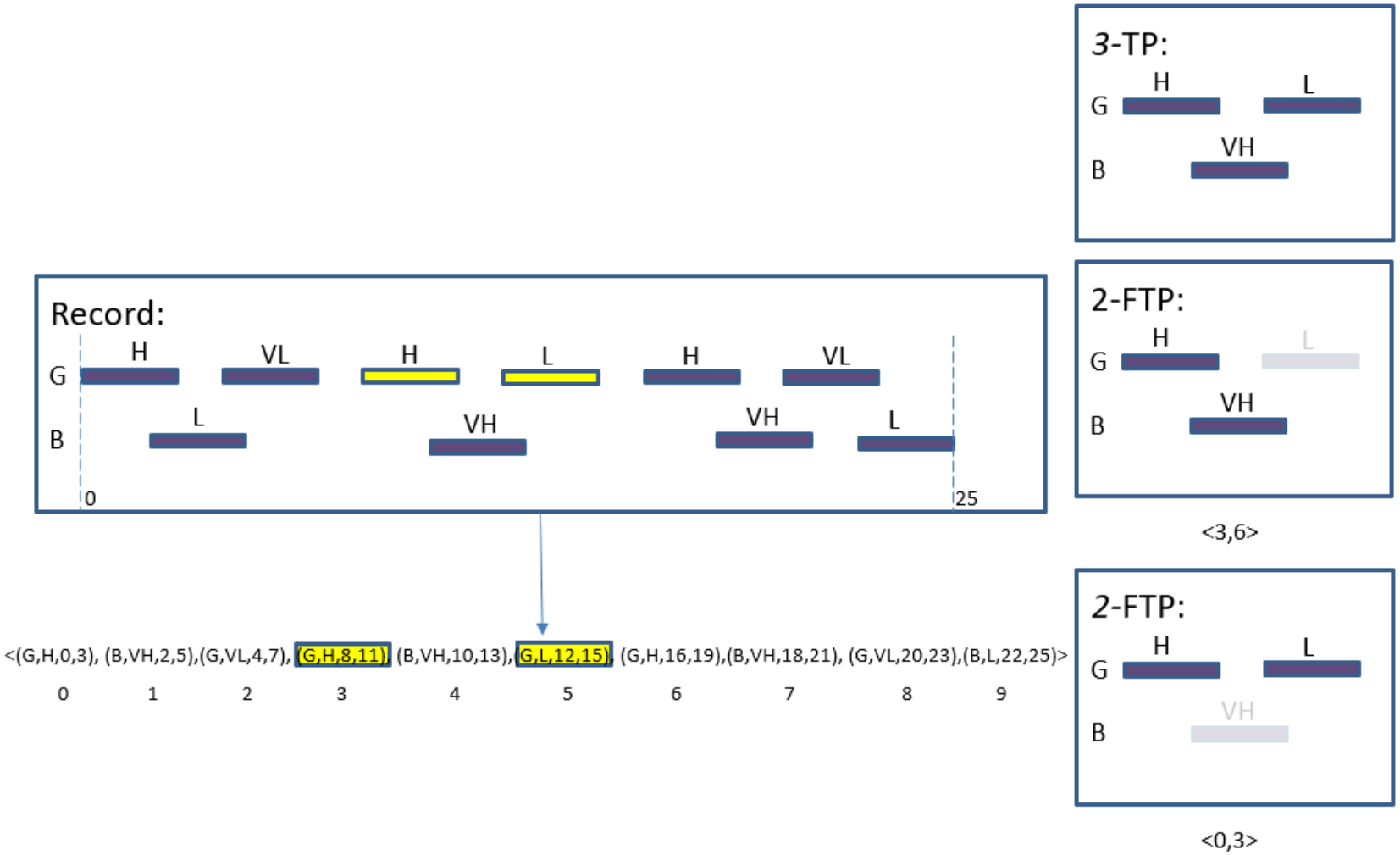
Extended Lists: Example



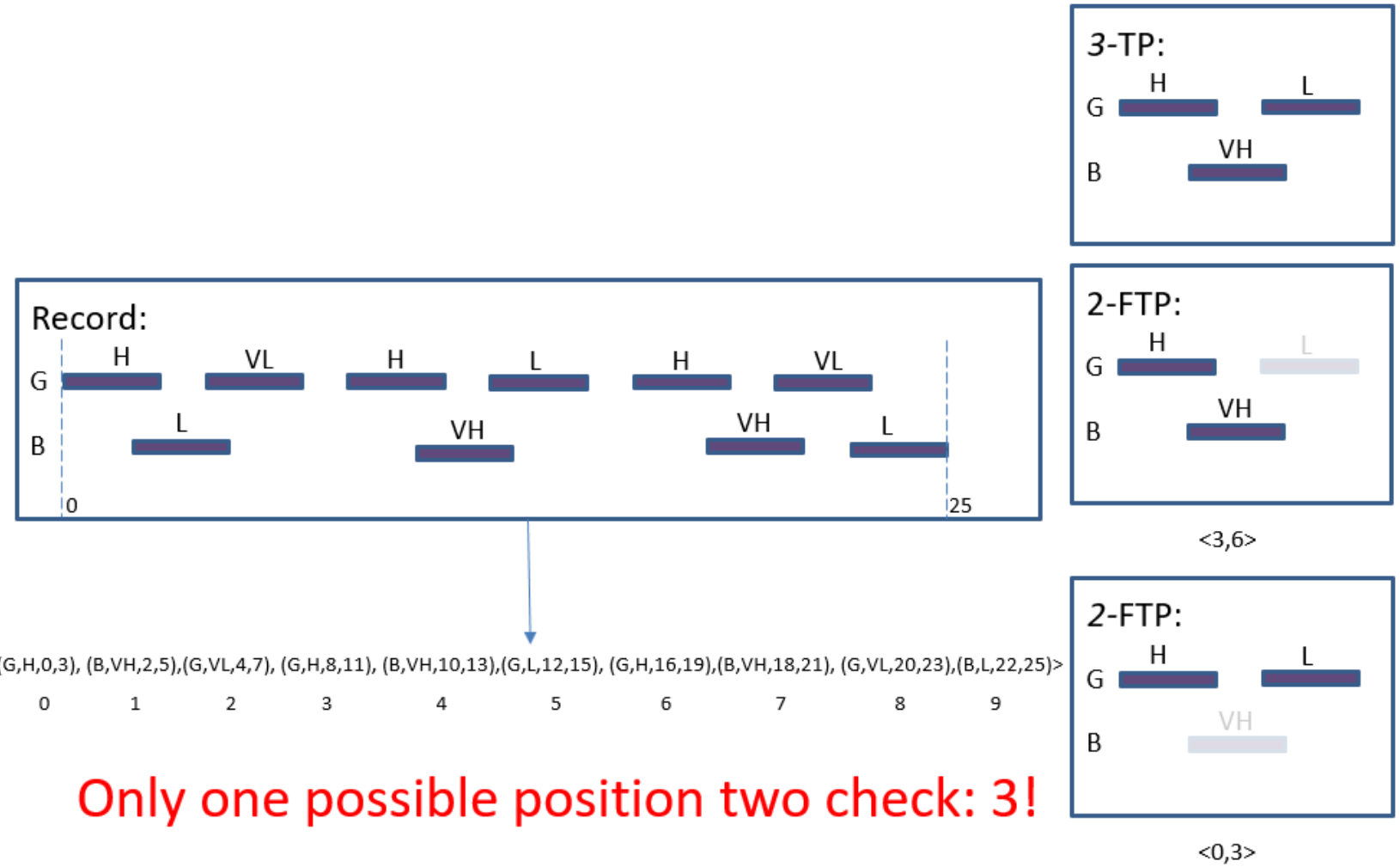
Extended Lists: Example



Extended Lists: Example

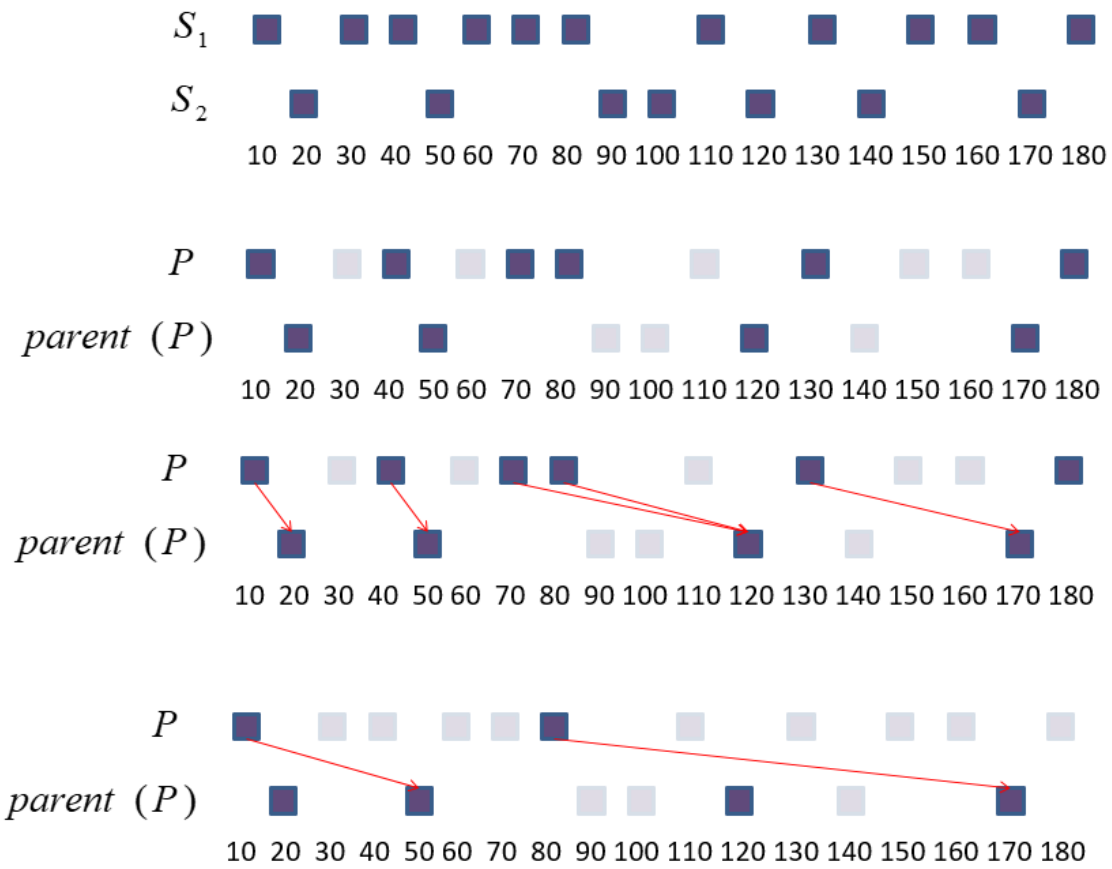


Extended Lists: Example

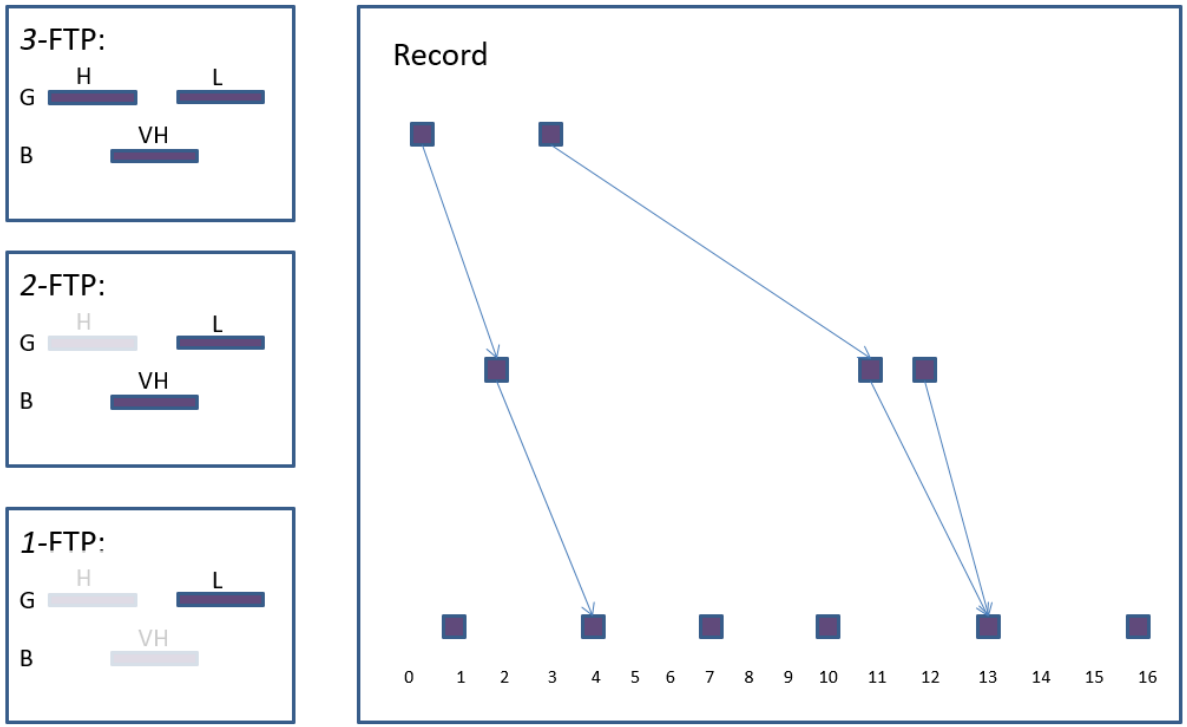


Extended Lists

$$P = (< S_1, S_2, \ldots, S_k >, R)$$



Speeding-up Search



FTPMwEL Algorithm

```
Input: D, 1-FTPs
Output: FTPs
FTPs ← 1-FTPs;
k ← 1 ;
while |k-FTPs| > 0 and no other stopping criteria are met do
    (k+1)-FTPs ← ∅ ;
    (k+1)-candidates ← CreateCoherentCandidates(k-FTPs, 1-FTPs) ;
    forall the P ∈ (k+1)-candidates do
        if ∃P̃ ∈ sub_k(P) such that P̃ isn't frequent then
            continue
        end
        P.p_ids ← ∩_{P̃∈sub_k(P)} P̃.ids;
        if not PotentiallyFrequent(P) then
            continue
        end
        forall the id ∈ P.p_ids do
            P.p_pos[id] = ∩_{P̃∈sub_k(P)\parent(P)} P̃.pos[id] ;
            if P.p_pos[id] = ∅ then
                P.p_ids ← P.p_ids \ id
            end
            else
                P.p_ex_list[id] ← CreateLinks(P.p_pos[id], P.parent.pos[id]) ;
                if P.p_ex_list[id] = ∅ then
                    P.p_ids ← P.p_ids \ id
                end
            end
        end
        if not PotentiallyFrequent(P) then
            continue
        end
        P.ids ← P.p_ids ;
        forall the id ∈ P.ids do
            P.ex_list[id] ← FindPositionsAndLinks(P.p_ex_list[id]) ;
            if P.ex_list[id] = ∅ then
                P.ids ← P.ids \ id
            end
        end
        if P is FTP in D then
            (k+1)-FTPs ← (k+1)-FTPs ∪ {P} ;
        end
    end
    FTPs ← FTPs ∪(k+1)-FTPs ;
    k ← k+1 ;
end
```

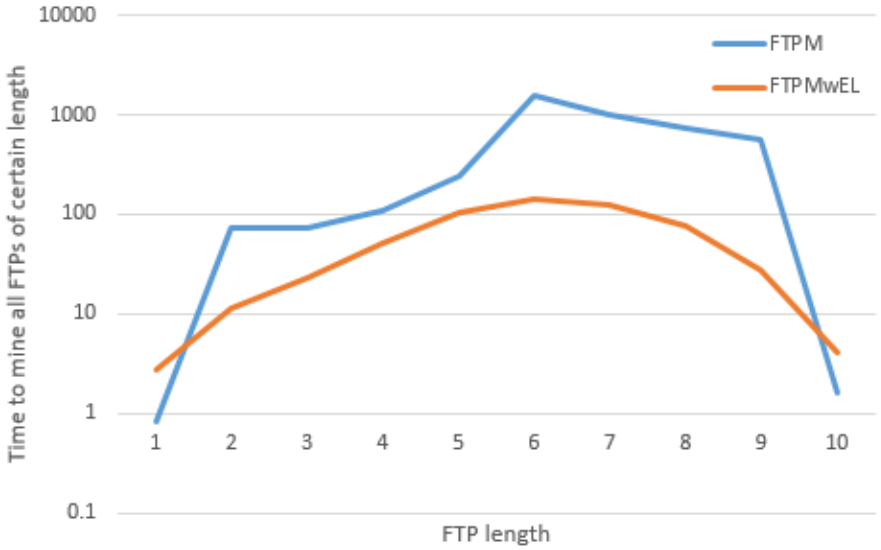
Algorithm 2: Frequent Temporal Pattern Mining with Extended Lists

Mining FTPs on AKI dataset

θ	max k	found k	FTPM wEL, sec	FTPM, sec	Speed-up Ratio	FTPM wEL, MB	FTPM, MB	Memory Ratio
0.5	5	5	5013.54	22768.01	4.54	14300.29	398.83	112.66
0.6	5	5	2097.22	11146.29	5.31	5105.14	126.93	40.22
0.7	Inf	7*(13)	8156.93	> 86400	> 10.59	179.8	NA	NA
0.8	Inf	10	569.16	4404.06	7.74	17.24	1.26	13.7
0.9	Inf	6	36.64	82.05	2.24	11.42	1.68	6.78

Table 1: Computational time and memory usage comparison of FTPMwEL and FTPM on the AKI dataset for a varying threshold level.

Mining FTPs on AKI dataset



Mining FTPs on datasets from UCR archive

	Dataset	FTPM wEL, sec	FTPM, sec	Speed-up Ratio	FTPM wEL, MB	FTPM, MB	Memory Ratio
1	BeetleFly	114.13	120.11	1.05	209.57	14.88	13.08
2	BirdChicken	38.38	53.37	1.39	105.48	5.89	16.91
3	Coffee	42.09	46.84	1.11	86.98	1.17	73.47
4	Computers	299.14	672.23	2.25	175.04	0.52	333.41
5	DistalPhalanx OutlineCorrect	13.16	24.34	1.85	39.91	0.52	75.24
6	Earthquakes	1661.45	666.89	-2.49	993.54	7.2	137.01
7	ECG200	58.06	115.58	1.99	128.51	3.46	36.09
8	ECGFiveDays	13.44	15.87	1.18	33.62	0.53	62.76
9	FordA	26578.2	21167.66	-1.26	23044.7	401.63	56.38
10	FordB	15227.69	12381.59	-1.23	13714.19	250.28	53.8
11	Gun_Point	1.25	2.28	1.82	0.07	0	NA
12	Ham	1001.84	1104.86	1.1	1318.66	44.4	28.7
13	HandOutlines	42.47	38.55	-1.1	116.25	4.2	26.71
14	Herring	92.62	164.44	1.78	286.96	11.68	23.56
15	ItalyPowerDemand	1.28	2.33	1.82	1.51	0	NA
16	Lighting2	737.92	2329.12	3.16	608.77	13.39	44.46
17	MiddlePhalanx OutlineCorrect	101.23	120.23	1.19	304.71	9.5	31.07
18	MoteStrain 2.48	3.61	1.46	2.28		0	NA
19	PhalangesOutlines Correct	95.14	159.27	1.67	405.23	14.16	27.62
20	ProximalPhalanx OutlineCorrect	42.34	63.23	1.49	174.46	4.72	35.97
21	ShapeletSim	1584.8	9980.11	6.3	878.67	18.49	46.52
22	SonyAIBORobot Surface	49.41	74.31	1.5	145.55	10.27	13.18
23	SonyAIBORobot SurfaceII	99.64	135.33	1.36	236.55	14.56	15.24
24	Strawberry	91.41	137.34	1.5	331.02	10.29	31.18
25	ToeSegmentation1	198.22	272.45	1.37	340.24	15.95	20.33
26	ToeSegmentation2	173.06	235.95	1.36	321.97	17.63	17.27
27	TwoLeadECG	3.25	4.19	1.29	3.88	0	NA
28	wafer	215.3	494.89	2.3	881.46	35.67	23.71
29	Wine	31.67	32.91	1.04	65.16	3.26	18.98
30	WormsTwoClass	854.77	2412.22	2.82	930.53	30.74	29.27
31	yoga	146.48	210.16	1.43	445.57	14.67	29.38

Other results

- The best pattern has accuracy of 76.8% alone
- FTPs are almost the same after sampling of records

Project participants:

- Anton Kocheturov
- Petar Momcilovic
- Azra Bihorac
- Dmytro Korenkevych
- Panos M Pardalos

Acknowledgments:

AK was supported by the grant by UF Informatics Institute. AB, PP and PM were supported by grant R01 GM-110240 by the National Institute of General Medical Sciences - National Institutes of Health.

The End

Thank you!