Evaluating an Integrated Time-Series Data Mining Environment

~A Case Study on a Chronic Hepatitis Data Mining~

Hidenao Abe (Department of Medical Informatics,
Shimane University, School of Medicine)
Miho Ohsaki (Faculty of Engineering, Doshisha University)
Hideto Yokoi (Department of Medical Informatics, Kagawa University Hospital)
Takahira Yamaguchi (Faculty of Science and Technology, Keio University)

Contents

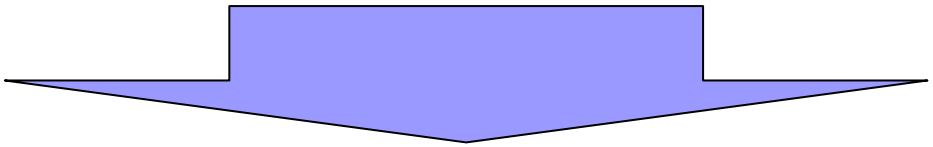
- Background
- The Integrated Time-Series Data Mining Environment
- Case Study on Chronic Hepatitis Data Mining
- Conclusion

Background

- KDD (Knowledge Discovery in Databases) has been widely known as a powerful process to extract useful knowledge.
- Collaboration of data miners, domain experts and system developers is important to success a data mining process.
- Knowledge depended on time stream is useful to predict some risk in future.

Issues and Our approach

- Many DM tools only supply DM methods
- There are no systematic support to carry out time-series data mining processes.



- Systematic support with preparing data mining methods from systematic analysis
- Human-system interaction

Map of our research

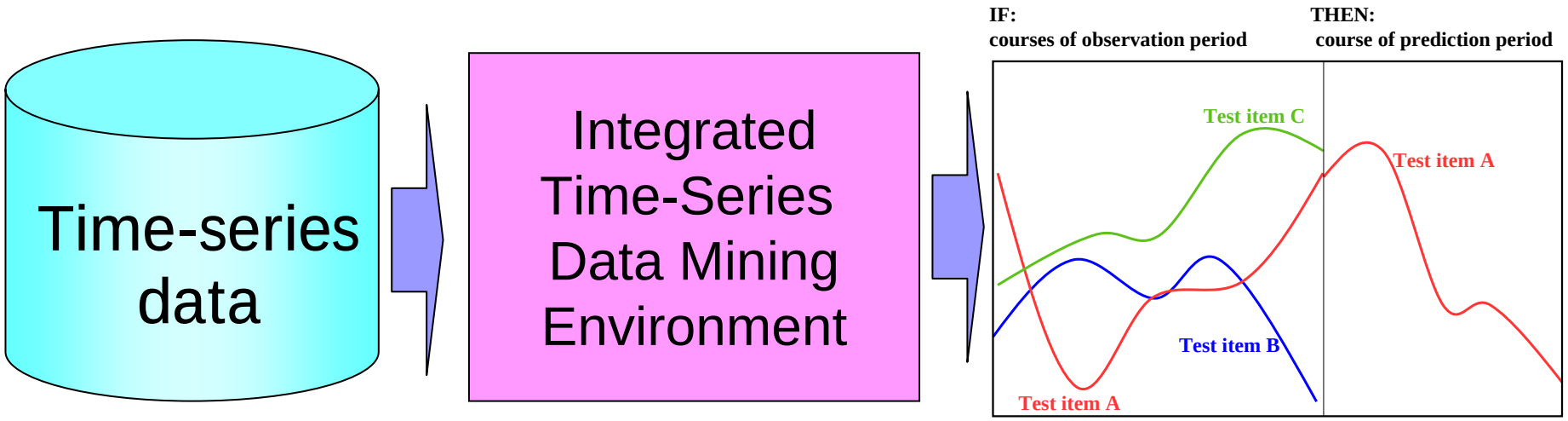
	Input time-series data	Mining approach	Post-processing
Our approach	ill-formed/ well formed	Rule induction based on time-series patterns	Visualizing patterns as graphs, Active interaction
Pattern extraction methods	ill-formed/ well formed	Particular pattern extraction algorithm	Visualizing patterns as graphs
Statistical methods, Signal processing methods	well formed	Particular time-series analysis method	(Visualizing as graphs)

Contents

- Background
- The Integrated Time-Series Data Mining Environment
- Case Study on Chronic Hepatitis Data Mining
- Conclusion

The Integrated Time-Series Data Mining Environment

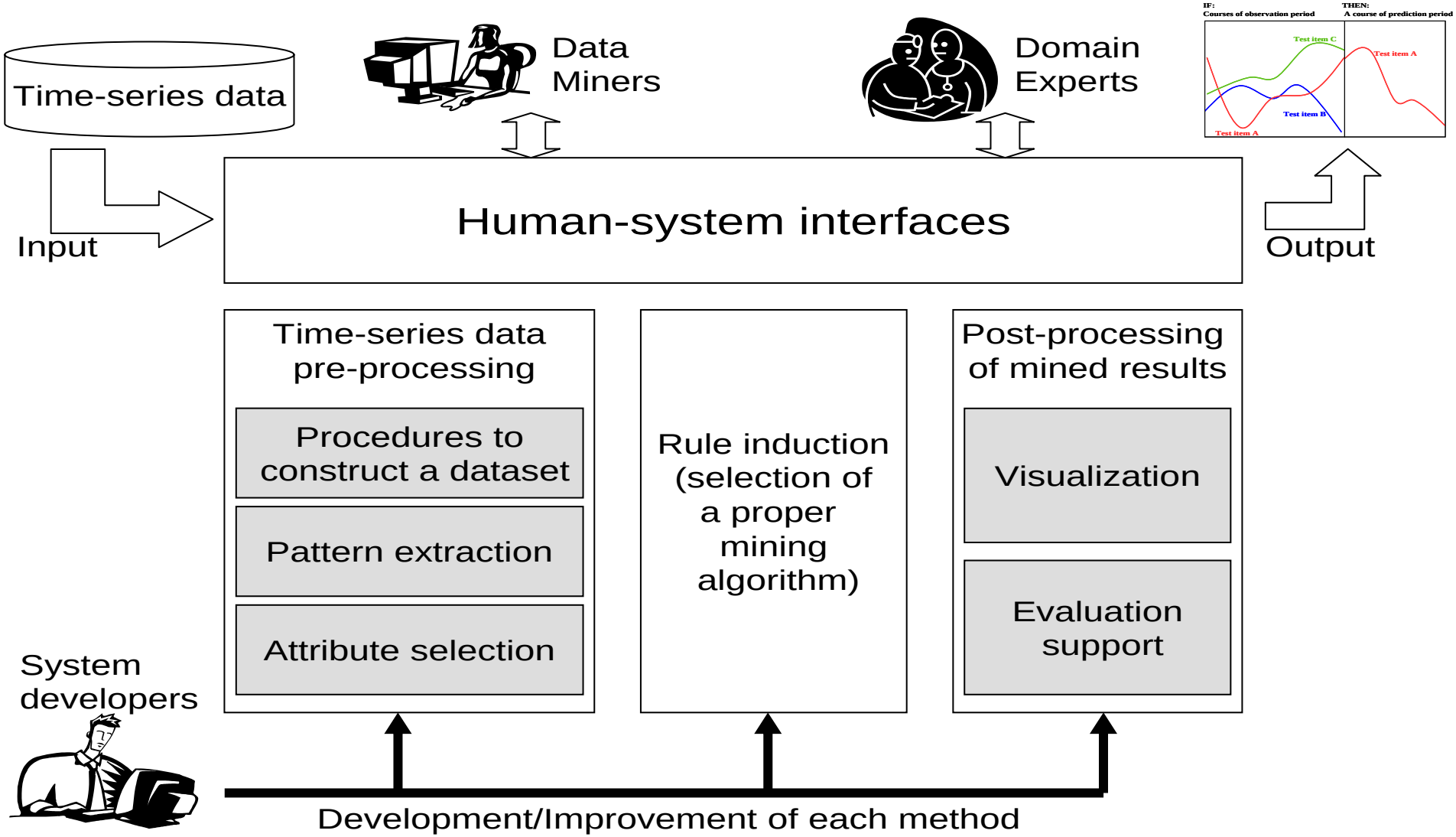
- Input
  - ill-formed/well-formed time-series data
- Output
  - IF-THEN rule based on time-series patterns



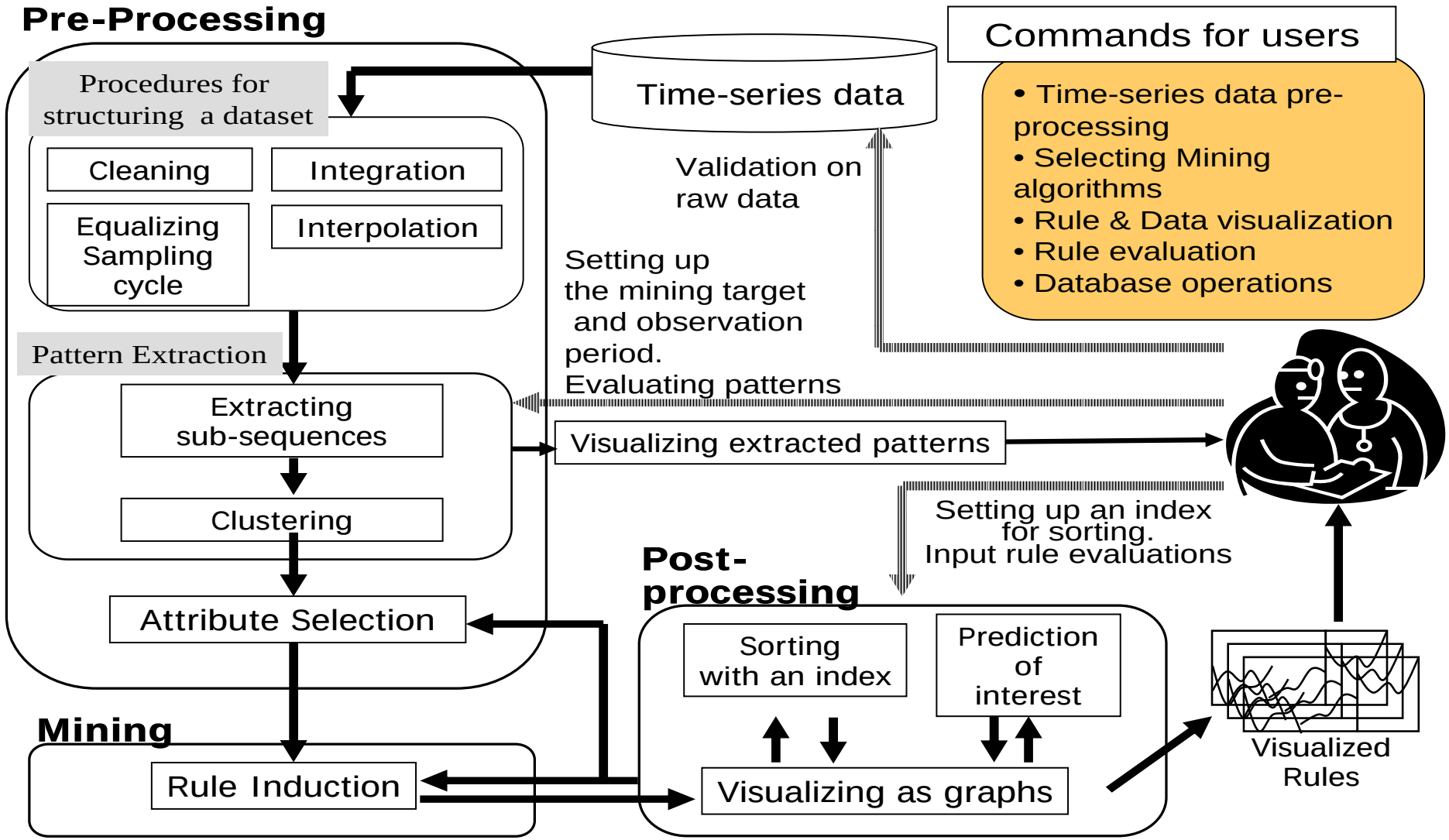
Procedures to Mine Time-Series Rules

- Data pre-processing
  - Pre-processing for data construction
  - Time-series pattern extraction
  - Attribute selection
- Mining
  - Rule induction
- Post-processing of mined results
  - Visualizing mined rules
  - Rule selection
  - Rule evaluation support
- Other database procedures
  - Selection with conditions
  - Join

System Overview



System Flow

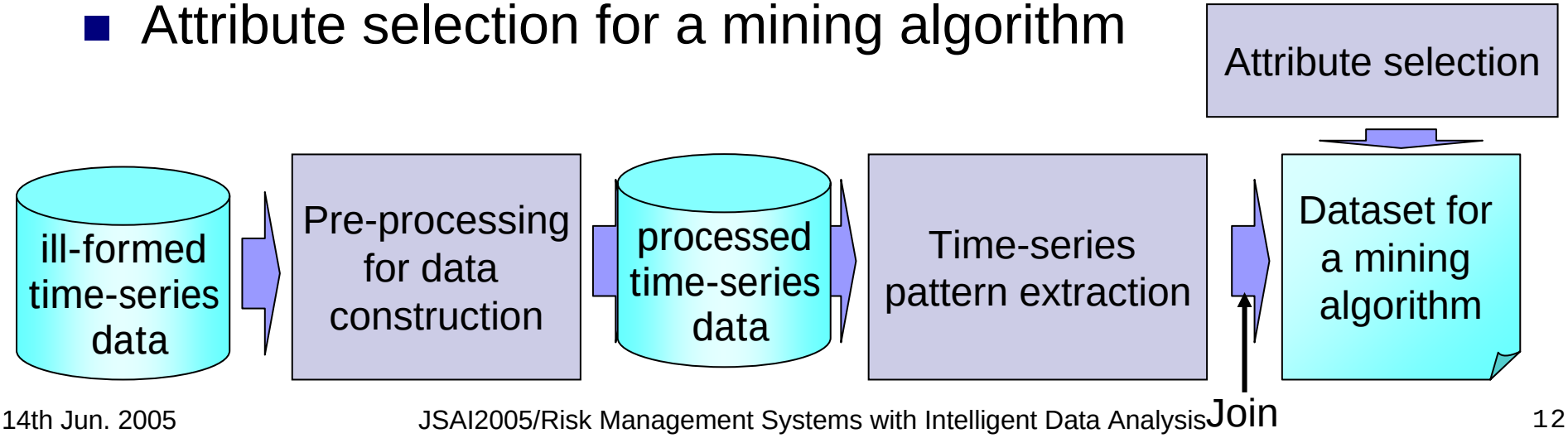

Pre-Processing
Procedures for structuring a dataset
Cleaning
Integration
Equalizing Sampling cycle
Interpolation
Pattern Extraction
Extracting sub-sequences
Clustering
Attribute Selection
Mining
Rule Induction
Time-series data
Validation on raw data
Commands for users
• Time-series data pre-processing
• Selecting Mining algorithms
• Rule & Data visualization
• Rule evaluation
• Database operations
Setting up the mining target and observation period.
Evaluating patterns
Visualizing extracted patterns
Setting up an index for sorting.
Input rule evaluations
Post-processing
Sorting with an index
Prediction of interest
Visualizing as graphs
Visualized Rules

Procedures to Mine Time-Series Rules

- Data pre-processing
  - Pre-processing for data construction
  - Time-series pattern extraction
  - Attribute selection
- Mining
  - Rule induction
- Post-processing of mined results
  - Visualizing mined rules
  - Rule selection
  - Rule evaluation support
- Other database procedures
  - Selection with conditions
  - Join

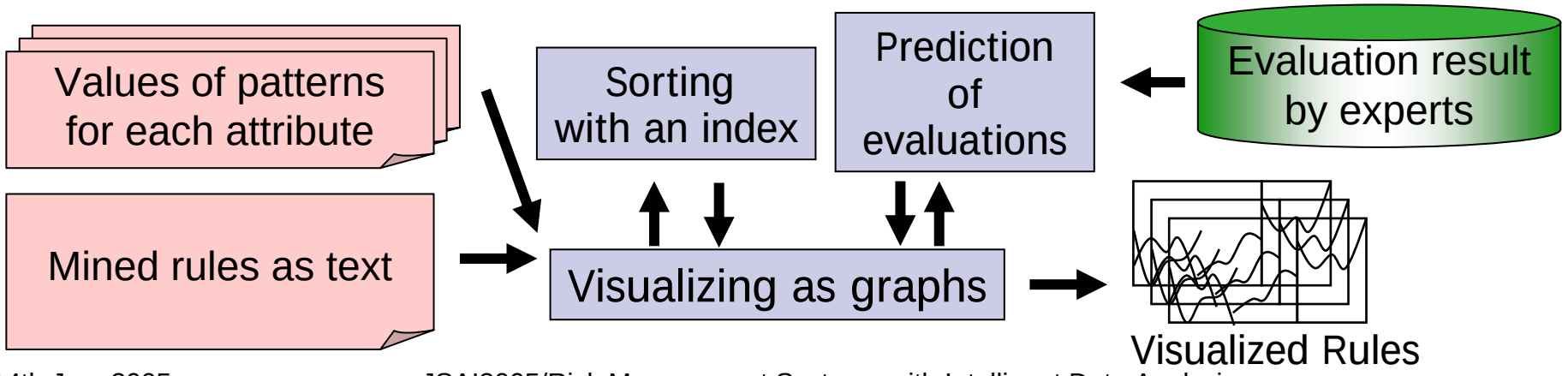
Data pre-processing for extracting time-series patterns from ill-formed data

- Pre-processing for data construction
  - Data cleaning, Integration of values, Equalizing sampling cycle, Interpolation
- Time-series pattern extraction
  - Extracting sub-sequences, Clustering (K-means, EM, our original pattern extraction algorithm)
- Attribute selection for a mining algorithm

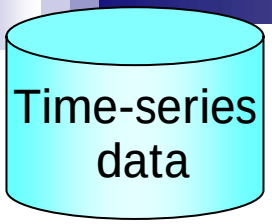


Post-processing with active human-system interaction

- Visualize mined rules
  - Visualizing text rules as graphical rules based on patterns
- Rule selection
  - Sorting graphical rules with indexes called objective measurement values
- Rule evaluation support
  - Predicting users' interest with re-using evaluated results

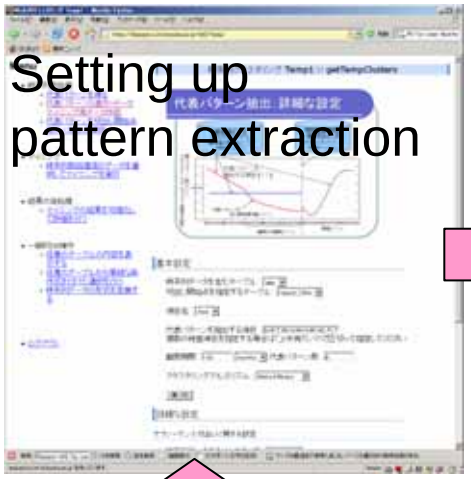


Graphical interfaces of the integrated time-series data mining environment

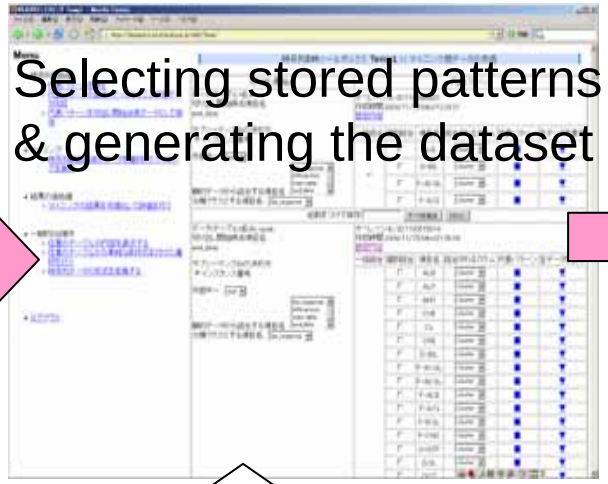


Data Miners

Domain Experts

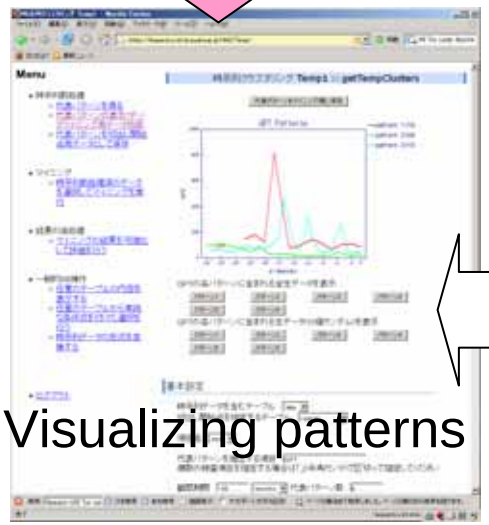


Setting up pattern extraction

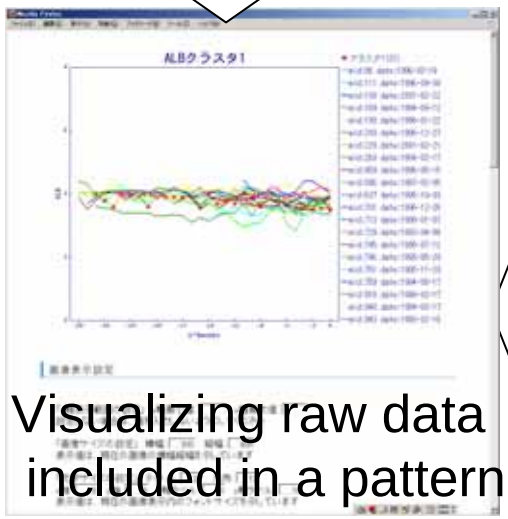


Selecting stored patterns & generating the dataset

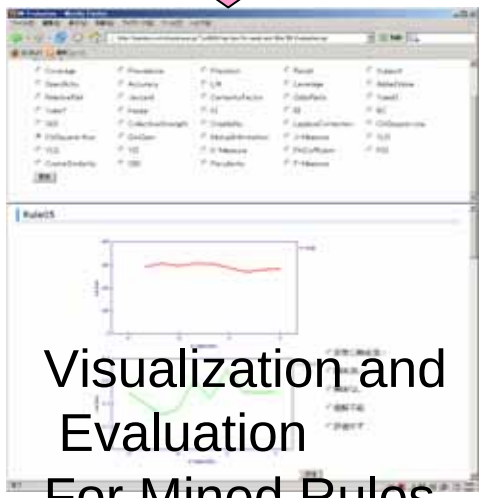
Attribute Selection & Rule Induction



Visualizing patterns



Visualizing raw data included in a pattern



Visualization and Evaluation For Mined Rules

Contents

- Background
- The Integrated Time-Series Data Mining Environment
- Case Study on Chronic Hepatitis Data Mining
- Conclusion

Description of the chronic hepatitis data mining

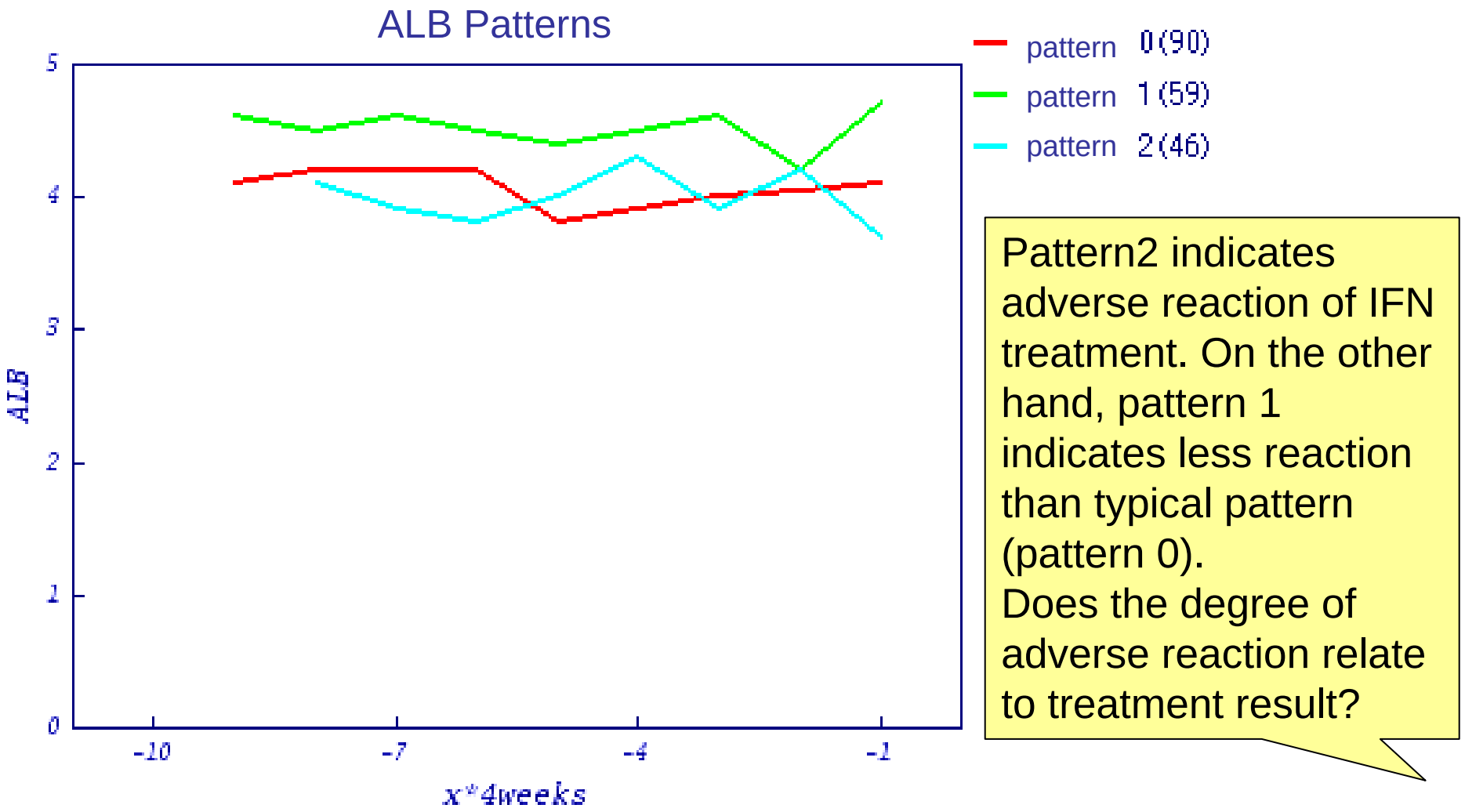
- Blood and urine laboratory test data
  - 1.9 million records
  - 965 test items
  - 771 patients (Hepatitis type B and C)
    - Up to 20 years for each sequence
- To find out risks related to IFN treatment results
  - 195 patients
  - Decided with GPT(ALT) values after finishing his/her IFN treatment
  - The risk means failures of his/her IFN treatment

Phase1:
Focusing expert's interest

- Pattern extraction about ALB values during IFN treatment

- Data miners:
  - Setting up observation period, pattern extraction algorithm and its parameters
  - Taking the original pattern extraction algorithm based on irregular sampling to calculate similarities between two sub-sequences
- Physician:
  - Evaluating patterns with visualized patterns and raw data included in interesting pattern as graphs on the interfaces

ALB patterns during IFN treatment



Pattern2 indicates adverse reaction of IFN treatment. On the other hand, pattern 1 indicates less reaction than typical pattern (pattern 0).
Does the degree of adverse reaction relate to treatment result?

Phase2:
Ensuring expert's hypothesis

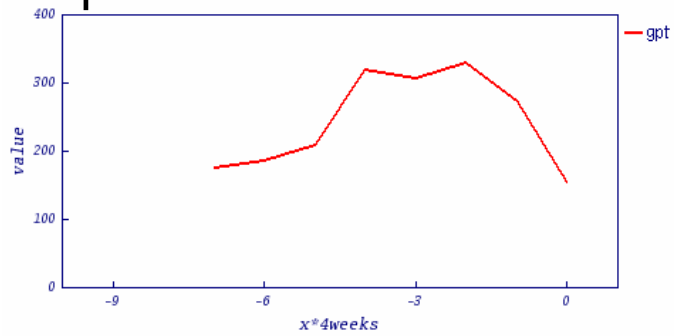
- Inducing rules to predict IFN treatment results from patterns of treatment periods

- Data miners:
  - Setting up observation period, pattern extraction algorithm and its parameters
  - Taking the original pattern extraction algorithm based on irregular sampling to calculate similarities between two sub-sequences
  - Selecting rule induction algorithm -> PART
- Physician:
  - Evaluating rules with visualized rules, patterns and raw data included in interesting pattern as graphs on the interfaces

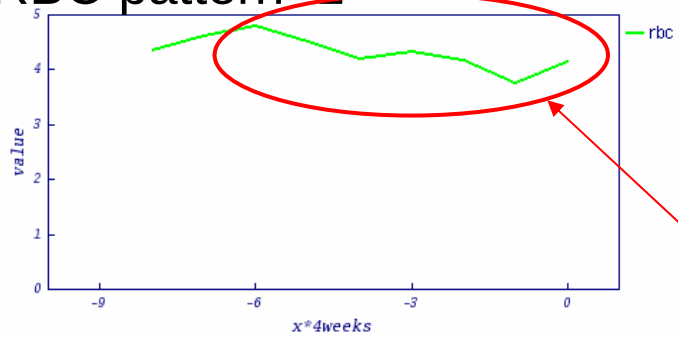
Representative rules having opposite IFN treatment results

Rule 1: with RBC pattern

IF: GPT pattern=3



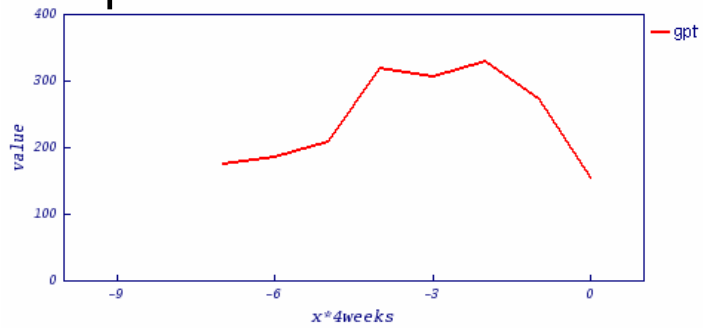
AND RBC pattern=1



THEN: bio_response = “Response”

Rule 2: without RBC pattern

IF: GPT pattern=3

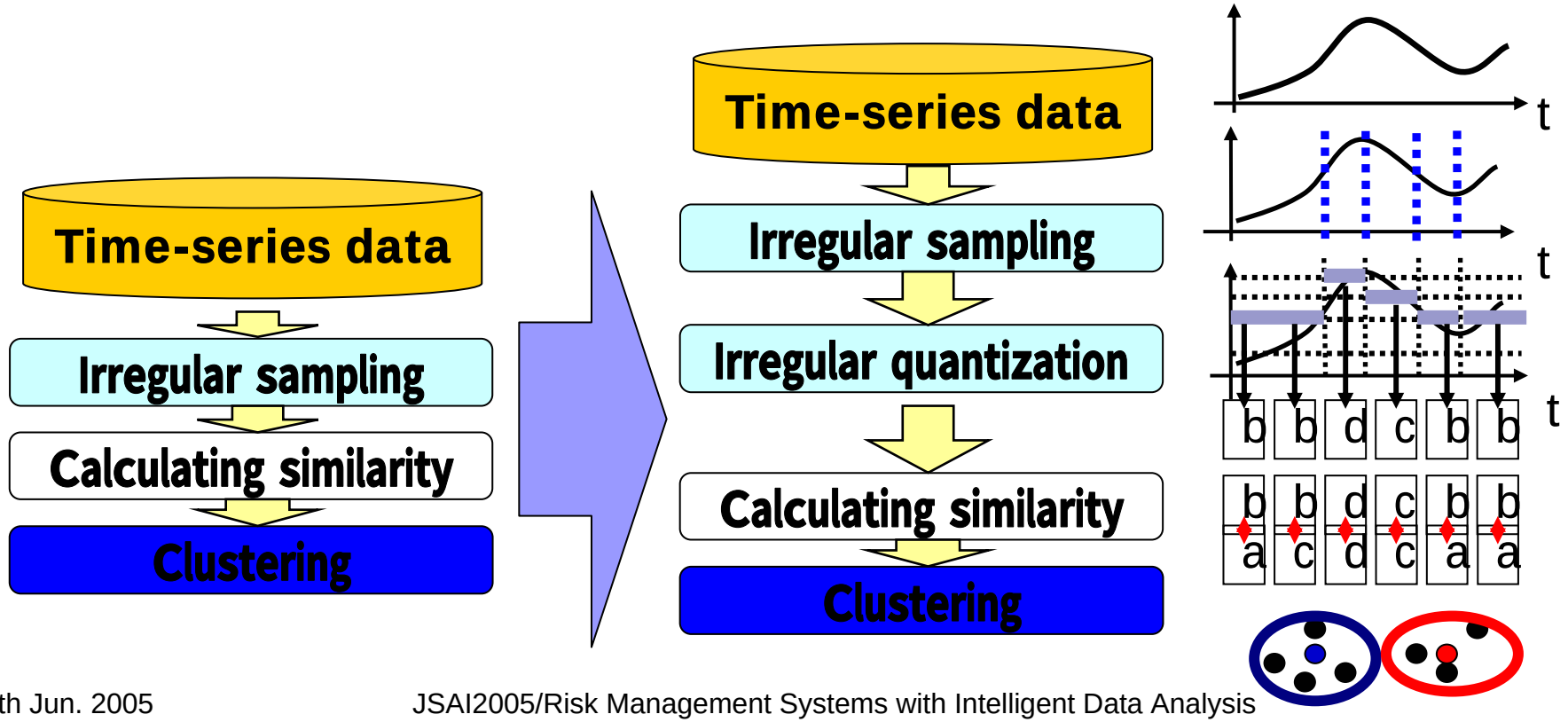


THEN: bio_response = “Non-Response”

Trend of anemia while IFN treatment
(anemia is a major adverse reaction)

Improvement of pattern extraction algorithm

- System Developers:
  - Improving the algorithm to calculate similarities between two sub-sequences



Evaluation results of the improvement of the pattern extraction algorithm

Evaluation Labels	Before Improvement	After Improvement
Very Interesting	4	15
Interesting	7	5
Fair	15	11
Difficult to understand	5	1
TOTAL	31	32

Contents

- Background
- The Integrated Time-Series Data Mining Environment
- Case Study on Chronic Hepatitis Data Mining
- Conclusion

Conclusion

- Implemented a time-series data mining environment, integrating time-series pattern extraction, rule induction, and rule evaluation support with active human-system interaction
- Succeeded in finding out a new hypothesis related to risks of IFN treatment result

- Developing active evaluation support re-using evaluations of domain experts
- Introducing algorithm selection sub-systems for each procedure to support data miners
- Applying this environment to other domains