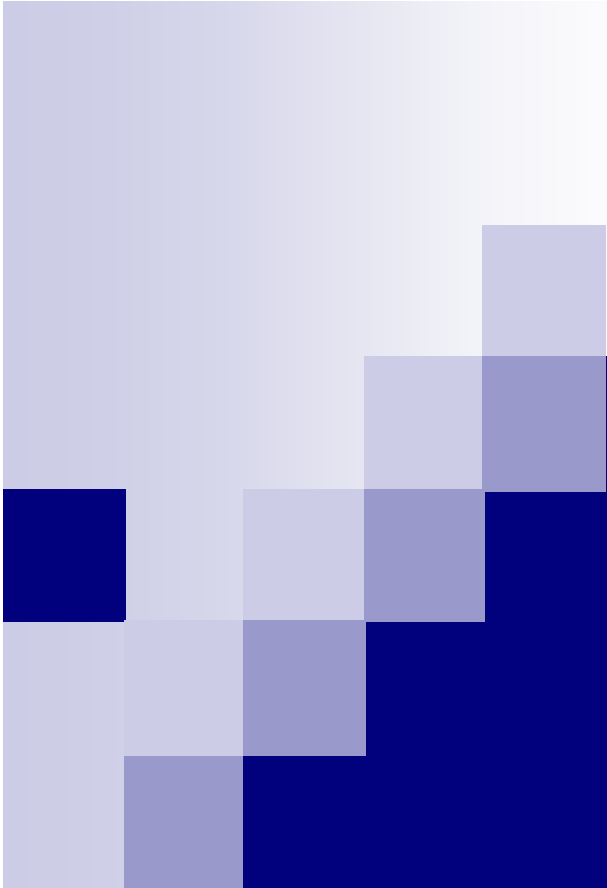




統合型時系列データマイニング 環境の開発と評価

阿部秀尚 (島根大学医学部医学科医療情報学講座)

山口高平 (慶應義塾大学理工学部管理工学科)

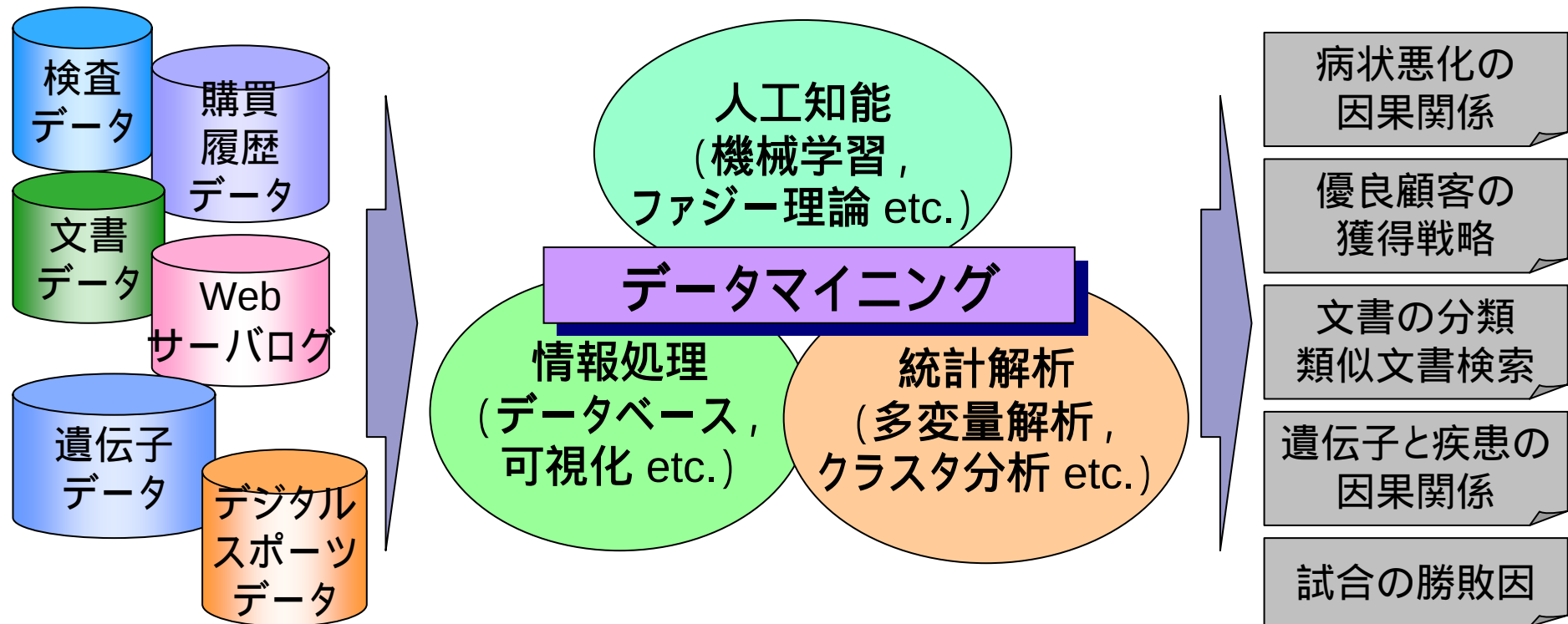


発表内容

- 研究背景と目的
- 統合型時系列データマイニング環境の設計と実装
- 慢性肝炎データマイニングにおける評価
- まとめ

データマイニング(知識発見)

データに存在する未知で有益なパターン(知識)を見出すこと



時系列データマイニング, テキストマイニング
Webマイニング, バイオインフォマティクス,
スポーツデータマイニング, センサデータマイニング etc.

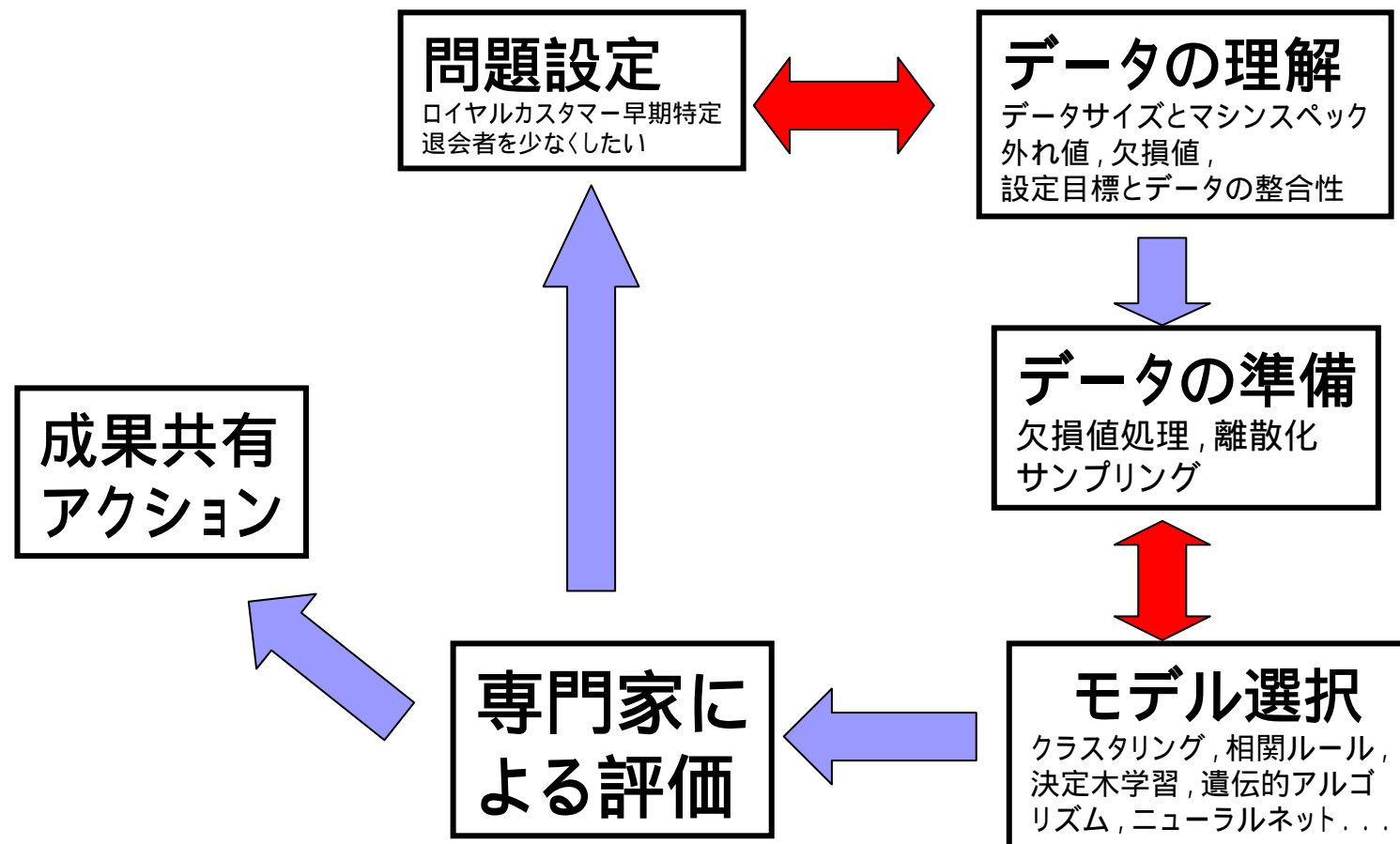
データに有用なパターン(構造的記述)を発見する



知識発見プロセス(2)

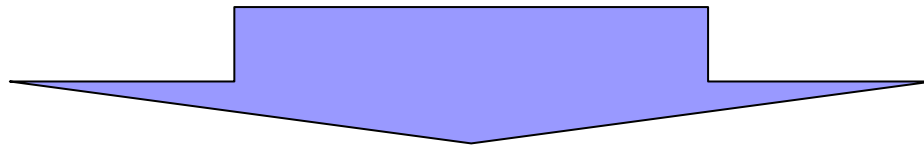
CRISP-DM Processes

→ データマイニングツール開発企業による標準プロセスの提唱



研究課題・目的

- 多くのデータマイニングツールは個々の処理の実行環境のみを提供
- 時系列データマイニングに関して体系的な支援が求められている



- 必要なデータマイニング処理を同定し、体系的な実行環境を提供
- ユーザによる評価・相互作用を重視

関連研究

	想定する入力 時系列データ	マイニング手法	後処理手法
本研究	非定型・定型	<u>時系列パターンに 基づくルール生成 (多くのルール生 成アルゴリズムが 利用可能)</u>	<u>ルールの可視化 ルール評価モデ ルによる評価作 業支援</u>
パターン抽出手法	非定型・定型	特定のパターン抽 出手法	パターンのグラ フ化
統計的手法 信号処理手法	定型	特定の時系列解 析手法	グラフ化(一部)



発表内容

- 研究背景と目的
- 統合型時系列データマイニング環境の設計と実装
- 慢性肝炎データマイニングにおける評価
- まとめ

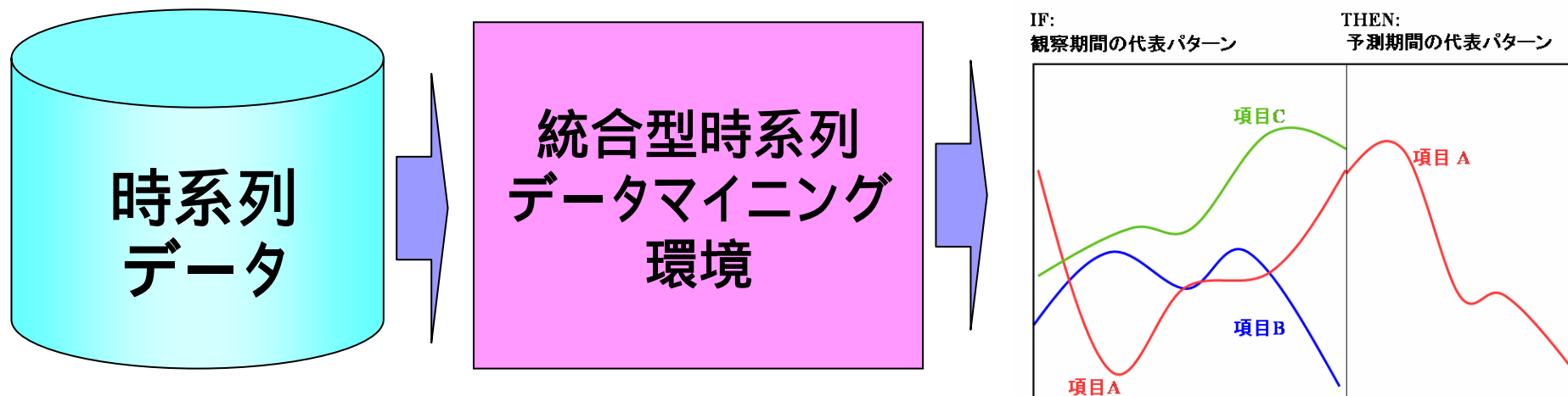
統合型時系列データマイニング環境

■ 入力

- 定型・非定型の時系列データ

■ 出力

- 時系列パターンを基にしたIF-THENルール

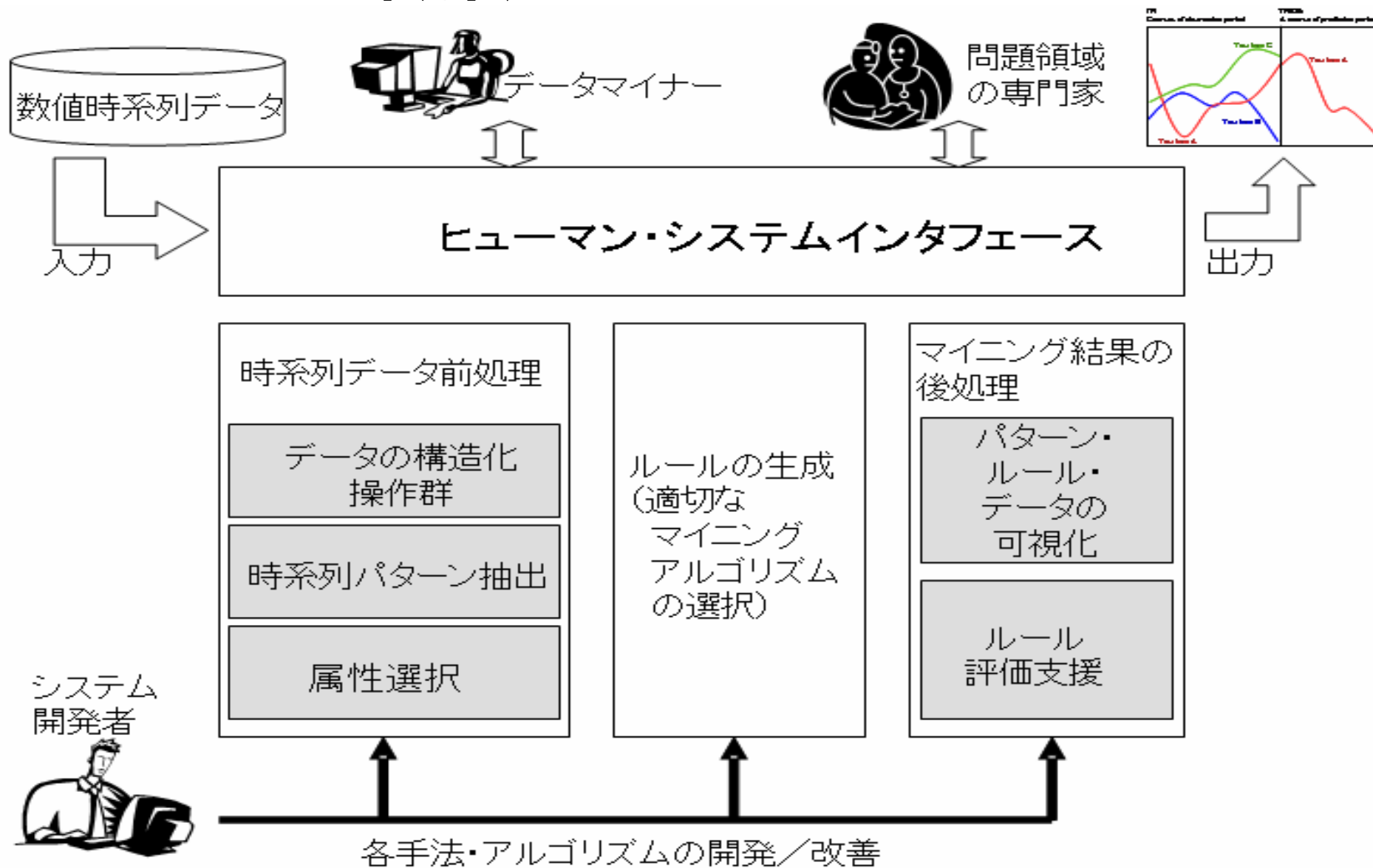




時系列ルールマイニングに必要な 操作の同定

- データの前処理
 - データ構造化処理
 - 時系列パターンの抽出
 - 属性選択
- マイニング
 - ルール生成
- マイニング結果の後処理
 - マイニングされたルールの可視化
 - ルール選択
 - ルール評価支援
- その他のデータマイニング手続き
 - 条件付きデータ選択
 - データ結合(join)

システム概観



システムフロー

データの预处理

データの構造化

データ洗浄

データ統合

観測周期の
均一化

補間

パターンの抽出

サブシーケンスの
切り出し

クラスタリング

属性選択

マイニング

ルール生成

数値時系列データ

生データの
選択・可視化

観察期間・予測期間
の設定,
パターンの評価

抽出されたパターンの可視化

並び替え基準の設定,
ルール評価の入力

結果の後処理

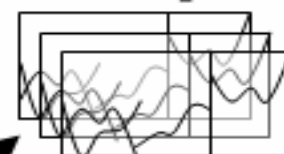
客観的指標に
よる並び替え

専門家の
評価基準
の推定

ルールの可視化・提示

操作に対応するコマンド群

- ・時系列データ前処理
- ・マイニングアルゴリズムの
選定
- ・ルールとデータの可視化
- ・ルールの評価
- ・データベース操作



グラフとして
可視化した
ルール群

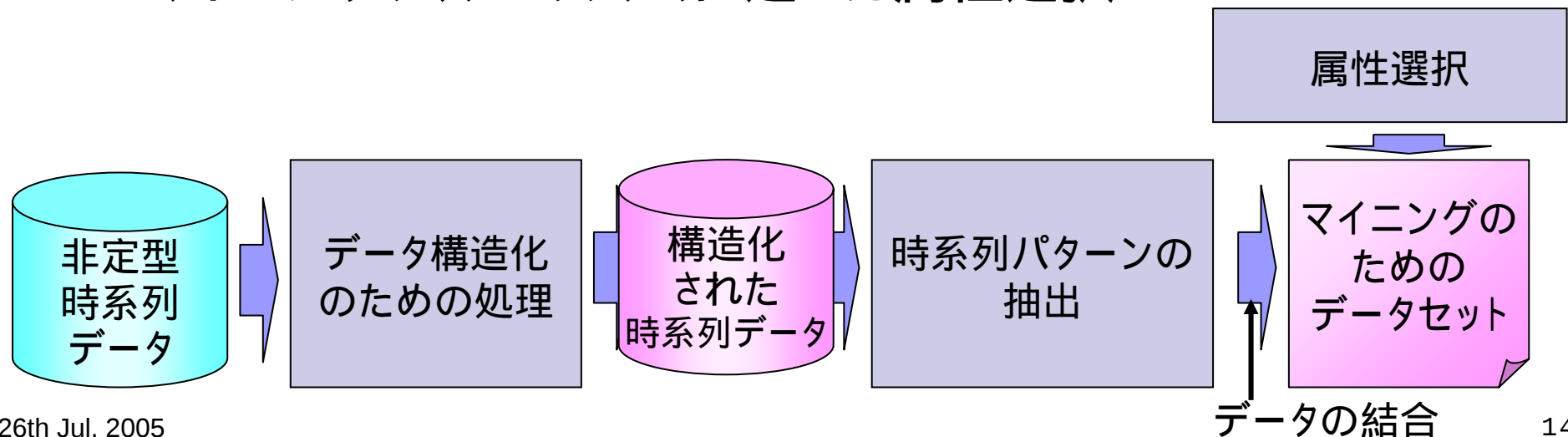


時系列ルールマイニングに必要な 操作の同定

- データの前処理
 - データ構造化処理
 - 時系列パターンの抽出
 - 属性選択
- マイニング
 - ルール生成
- マイニング結果の後処理
 - マイニングされたルールの可視化
 - ルール選択
 - ルール評価支援
- その他のデータマイニング手続き
 - 条件付きデータ選択
 - データ結合(join)

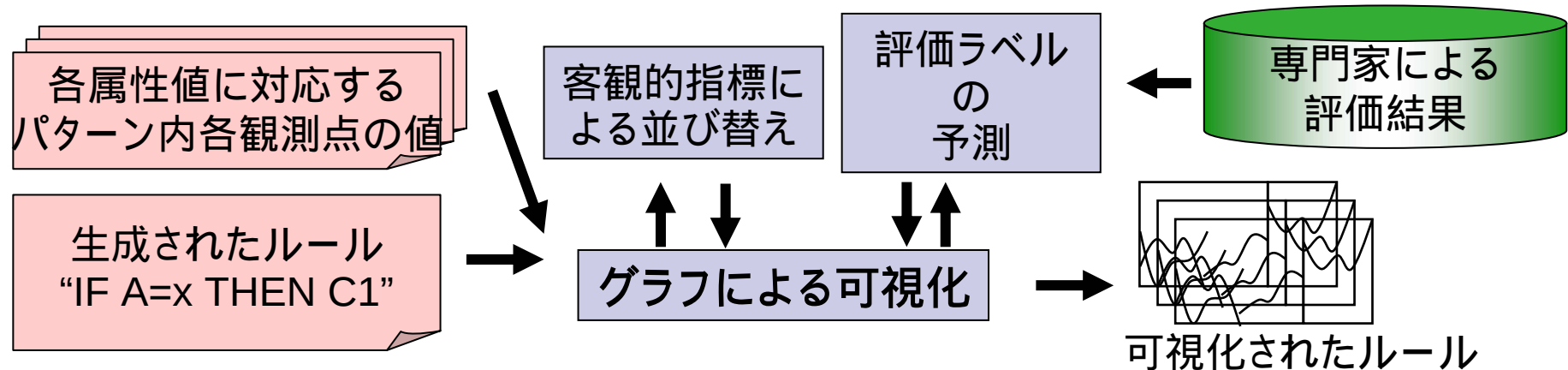
非定型時系列データに対するデータの前処理

- データ構造化のための処理
 - データの洗浄, データの統合, 観測間隔の均一化, 補間
- 時系列パターンの抽出
 - サブシーケンスの切り出し, クラスタリング (K-means, EM, 不等間隔な標本化・量子化による時系列クラスタリング)
- マイニングアルゴリズムに適した属性選択



積極的なルール評価支援による マイニング結果の後処理

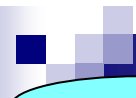
- 時系列ルールの可視化
 - テキスト表記の時系列パターンに基づくルールの可視化
- ルール選択
 - 客観的指標に基づくルールの並び替え提示
- ルール評価モデルによる評価作業支援
 - 過去の評価基準のモデル化によるルール評価の予測





統合型時系列データマイニング環境 の実装

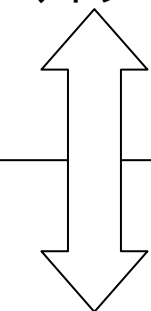
- インターフェース
 - CGIスクリプト(Perl)
- データ操作
 - リレーショナルデータベース, SQL
- クラスタリング・ルール生成
 - Weka(Java) + ラッパー(Perl)
 - 不等間隔な標本化・量子化に基づく時系列パターン抽出(Perl)
- 可視化
 - グラフ(チャート)作成(Perl+GD)
- データフォーマットの標準化
 - 入力・中間 時系列データ(CSV=ID, time-stamp, item, item value)
 - 中間データ(CSV)
 - マイニング用データセット(CSV → ARFF, C4.5)



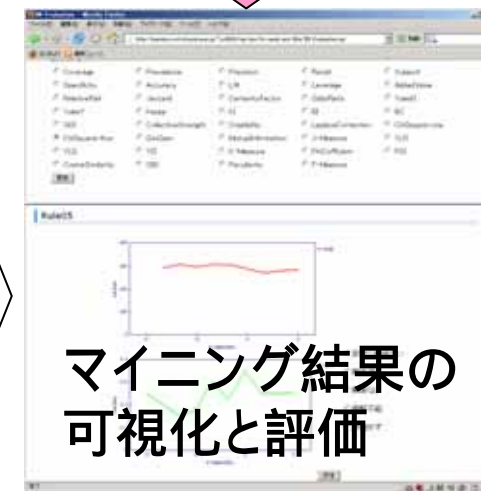
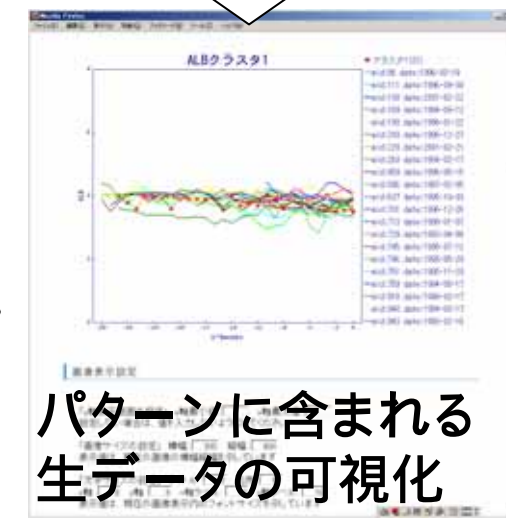
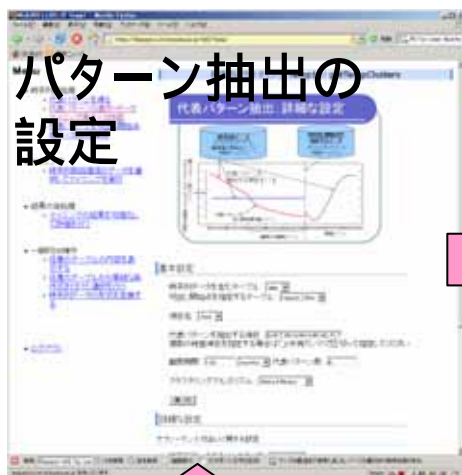
統合型時系列データマイニング環境における各ユーザ・インターフェースの役割



データ
マイナー



領域
専門家





発表内容

- 研究背景と目的
- 統合型時系列データマイニング環境の設計と実装
- 慢性肝炎データマイニングにおける評価
- まとめ

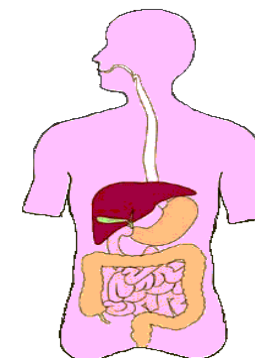
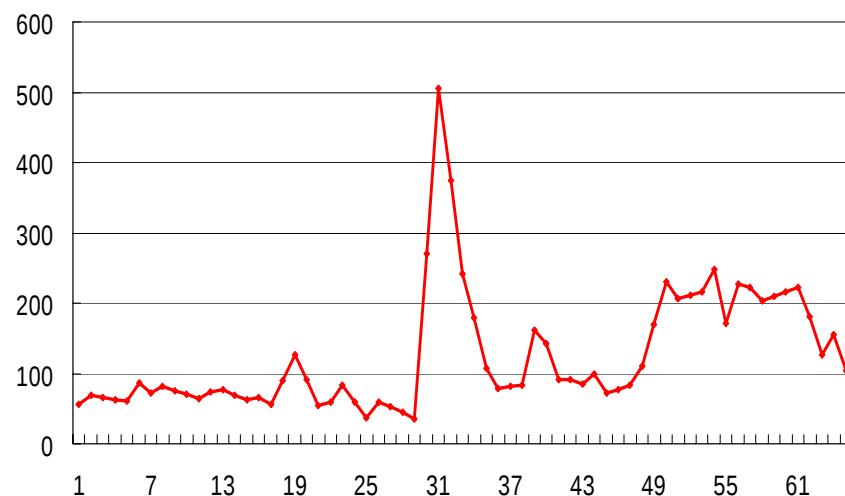


慢性肝炎データと目的データ

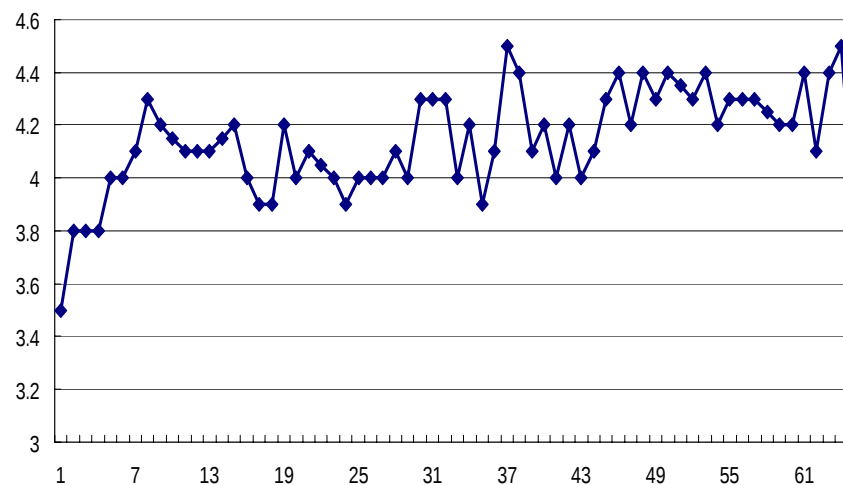
- 血液と尿の生化学検査データ・血算データ
 - 約190万レコード
 - 965検査項目
 - 771患者(B型・C型肝炎)
 - 最長で20年間の治療期間
- インターフェロン(IFN)治療効果判定データ
 - 195患者
 - インターフェロン治療終了後のGPT(ALT)の値から判定
 - インターフェロンの治療効果に関する治療期間中の特徴的な経過の発見

慢性肝炎データの一例

検査項目:GPT



検査項目:ALB



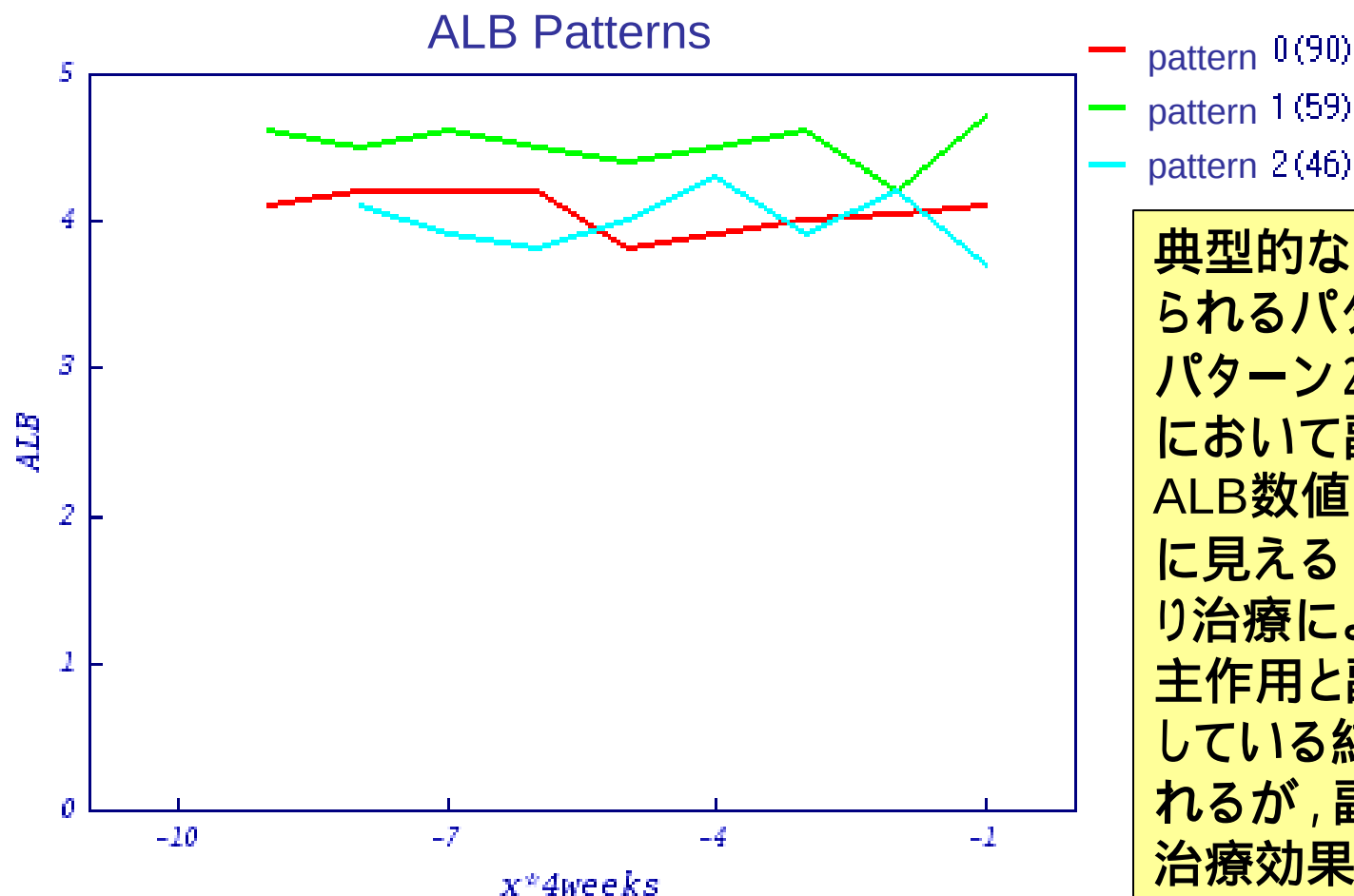


Phase1:

新たな仮説生成の手がかりの発見

- インターフェロン治療期間中の時系列パターンの抽出
- データマイナーの役割：
 - 観察期間の決定, パターン抽出アルゴリズムとそのパラメータの設定
 - 不等間隔な標本化に基づく時系列パターン抽出
- 専門家(内科医)の役割：
 - 抽出した時系列パターンと各パターンに含まれる生データを可視化することによる時系列パターンの評価

インターフェロン治療期間中の ALBに関する時系列パターン



典型的なIFN治療経過と見られるパターン0と比較して、パターン2は治療期間終盤において副作用によってALB数値が悪化しているように見える。パターン1はあまり治療による影響が無い。主作用と副作用は薬が作用している結果であると考えられるが、副作用の度合いは、治療効果にどの程度影響するのか？



Phase2:

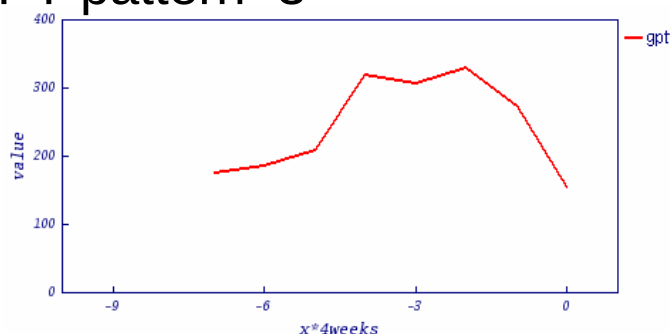
新たな仮説の生成

- IFN治療期間の時系列パターンに基づくルール生成
- データマイナーの役割：
 - 観察期間の設定, 時系列パターンアルゴリズムとそのパラメータの設定
 - 不等間隔な標本化と量子化による時系列パターンの抽出
 - PARTによるルール生成
- 専門家の役割：
 - 可視化された時系列ルール, 時系列パターン, 各パターンに含まれる生データを用いた評価作業

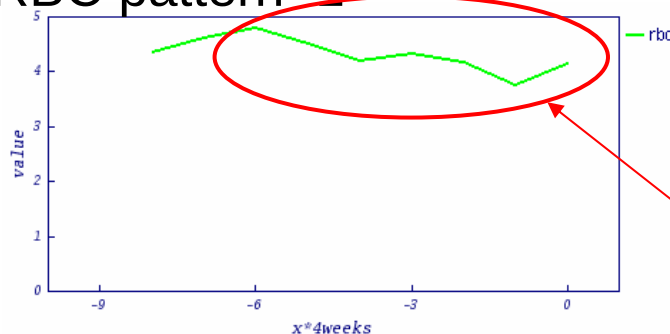
異なる治療効果判定となる代表的なルール

Rule 1: RBCパターン有

IF: GPT pattern=3



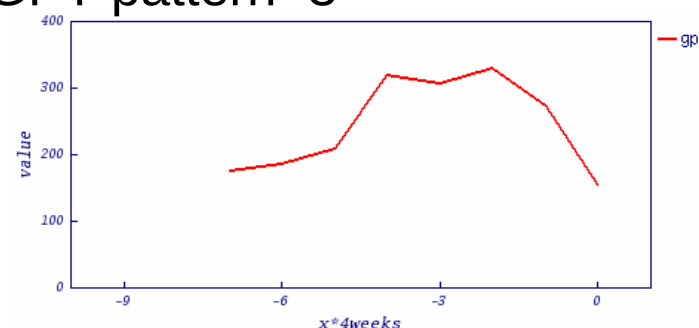
AND RBC pattern=1



THEN: 治療効果 = “有”

Rule 2: RBCパターン無

IF: GPT pattern=3



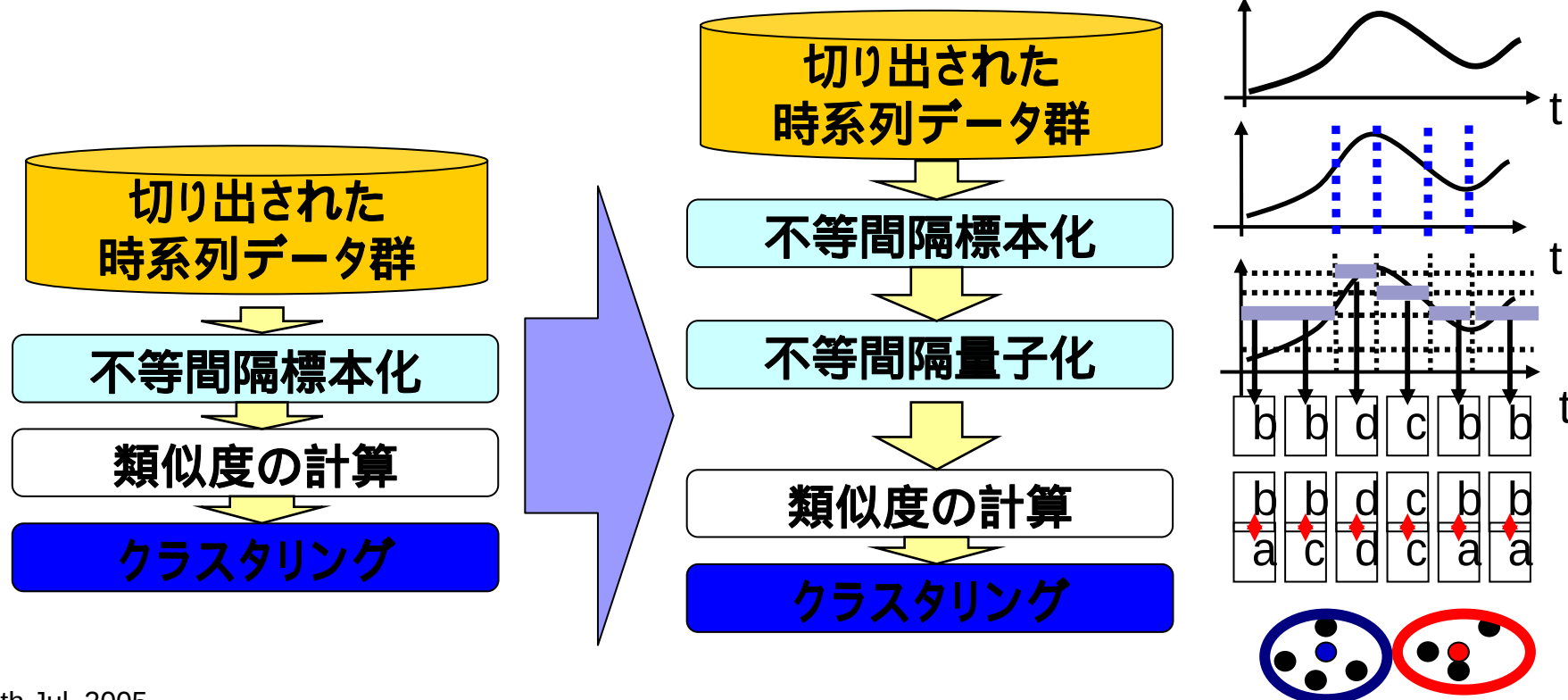
THEN: 治療効果 = “無”

IFN治療期間において貧血の傾向
(貧血はIFN治療の典型的な副作用の1つである)

パターン抽出アルゴリズムの改善

■ システム開発者の役割:

- 2つのサブシーケンスの類似度計算のための不等間隔標本化・量子化手法の開発





パターン抽出アルゴリズムの改善による専門家のルール評価結果の変化

評価ラベル	改善前	改善後
とても興味がある	4	15
興味がある	7	5
妥当	15	11
解釈困難	5	1
合計	31	32



発表内容

- 研究背景と目的
- 統合型時系列データマイニング環境の設計と実装
- 慢性肝炎データマイニングにおける評価
- まとめ



まとめ・今後の課題

- 時系列パターン抽出・IF-THENルール導出・積極的なルール評価支援を統合した時系列データマイニング環境の実現
- IFN治療効果に関係する新たな仮説の生成
- 専門家の評価を再利用したルール評価予測手法の開発
- データマイナーを支援するためのアルゴリズム選定手法の導入
- 他の問題領域データへの適用