

リスク分析とデータマイニング技術

NEC 中央研究所

共通基盤ソフトウェア研究所 マイニングRG

& ビジネスイノベーションセンター データマイニングセンター

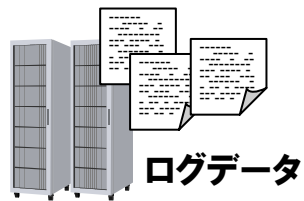
 U can change.

目次

- NECのマイニング技術
- ログ分析技術
 - 分析技術の紹介
 - 分析事例
- テキスト分析技術
 - 分析技術の紹介
 - 「eHyouban/マイニングサービス」について

データマイニングとは？

蓄積された膨大なデータから、意味のあるパターンや隠れた関係を
効率よく発見する技術



データマイニング

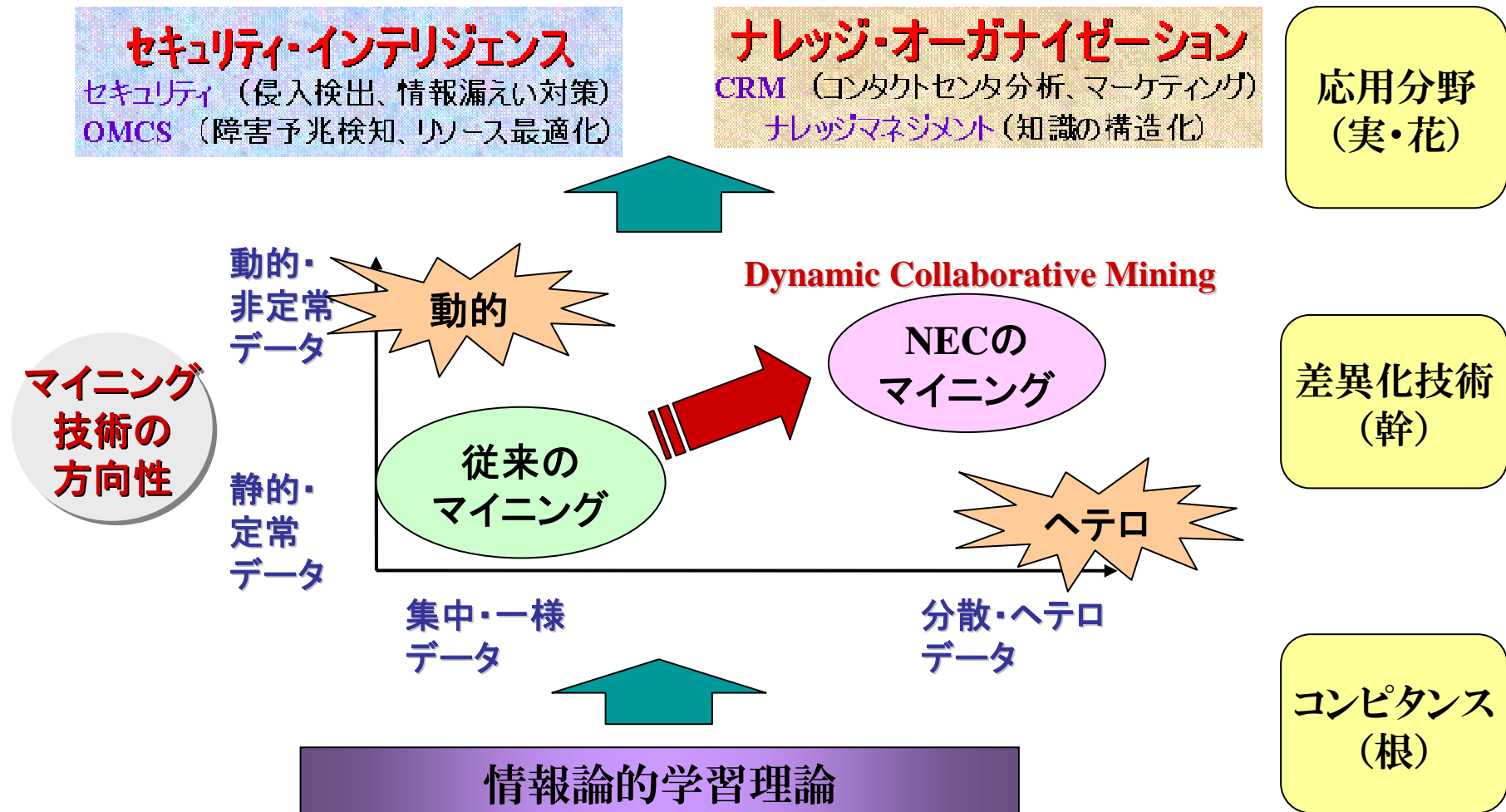
不正アクセスの検出

評判や関心の
発見

相関関係の発見

NECのデータマイニング技術の特徴

- **動的かつヘテロ**な分析に強みをもつマイニングエンジンを開発しSL・サービス事業を高付加価値化
- **セキュリティ性能障害マイニング**と**テキストマイニング**に特化した目的志向のエンジン群を構成



ログ分析技術

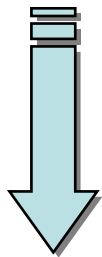


U can change.

セキュリティ・インテリジェンス技術

データマイニング技術の未知のウイルス検出、障害検出への応用
 -2005年先端技術大賞フジ・サンケイ・ビジネス・アイ賞受賞-

静的特性
の分析

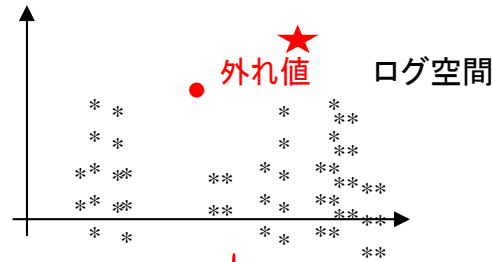


動的特性
の分析

SmartSifter

外れ値検出

大量データからの統計的
外れ値の検知

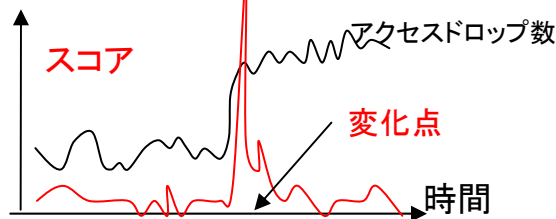


- ・不正検出
- ・侵入検出

ChangeFinder

変化点検出

時系列データからの急激に
変化する時点を検知

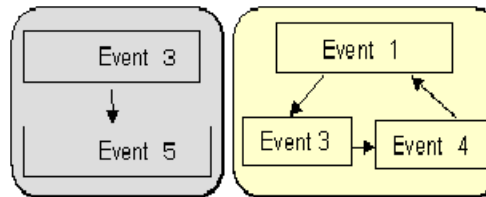


- ・未知ウイルス・SPAM検出
- ・DoS攻撃検出

AccessTracer

異常行動検出

行動履歴データから
異常行動パターンを検出



正常パターン

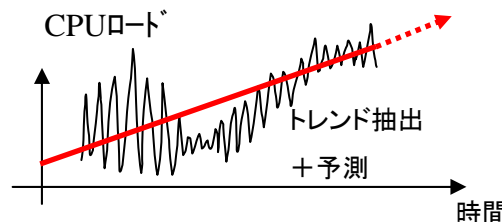
異常行動パターン

- ・なりすまし検出
- ・Syslog分析

TrendLiner

トレンド分析

時系列データからオンライン
で周期性・トレンドを抽出



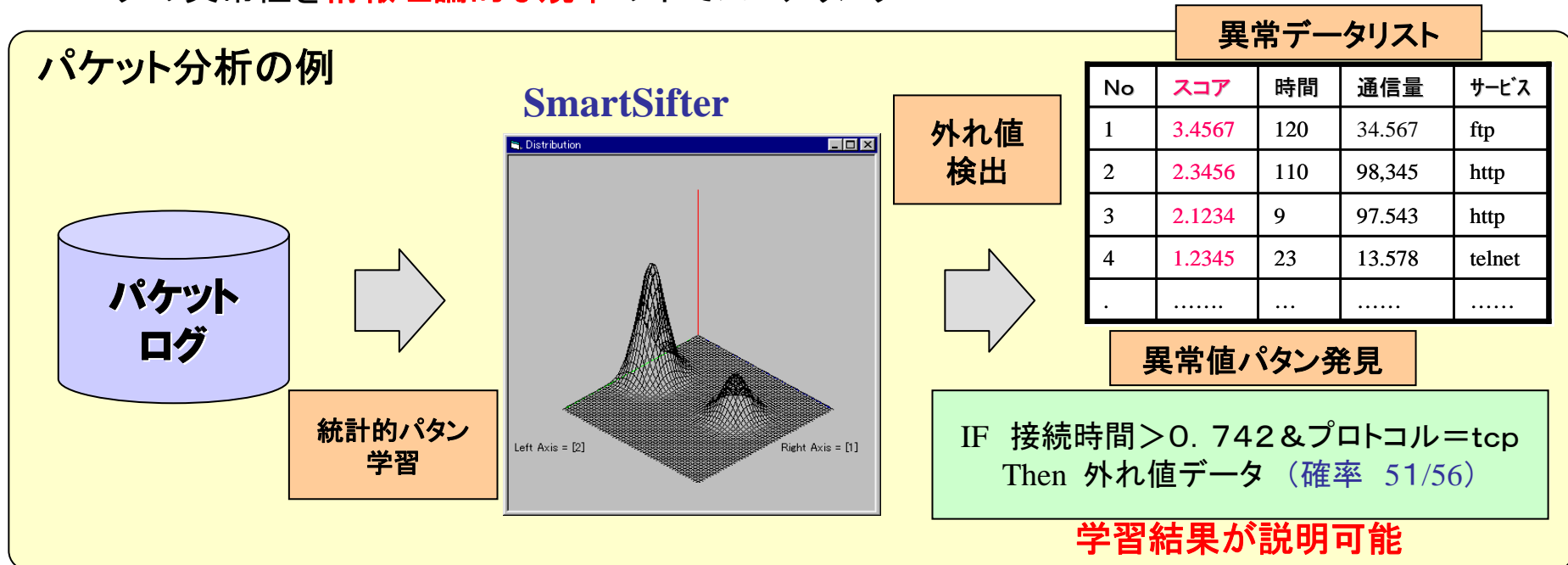
- ・障害予兆検出
- ・ベースライン監視

外れ値検出エンジン SmartSifter

- 機能** 大量データからの外れ値検出
応用 不正検出、侵入検出、突発型障害検出
特長 リアルタイムで適応的な外れ値検出

通常の閾値監視よりもパタンの多様性や時間変化への適応性が高い

- 膨大なパケット・ログの生成パターンを**確率モデルで表現**(ガウス混合分布)
- モデルを**教師なしのオンライン忘却型学習アルゴリズム**で適応的に学習
 - #過去のデータをほどよく忘れることにより、**正常なパターンが時間的に変化しても適応して学習**
- ログの異常性を**情報理論的な規準**の下でスコアリング



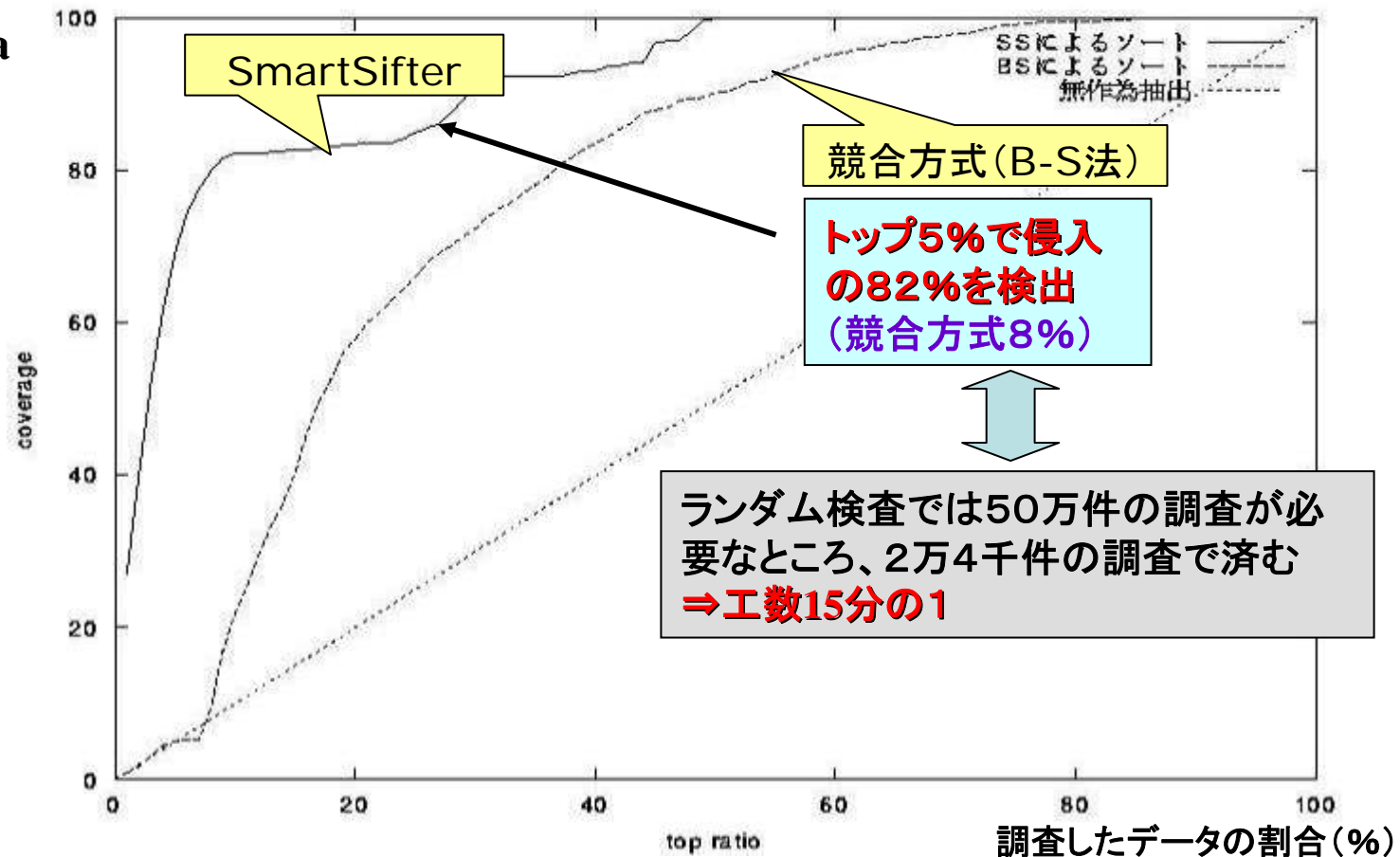
SmartSifter の性能評価

KDDCup1999データによる性能評価実験結果
 (http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html)

- 5つの属性(service, protocol_type, duration, src_bytes, dst_bytes)
- “log_in” が成功したデータのみ利用 ⇒ 侵入
- データ数.....500,000 侵入混合率.....1700(0.35%)

1700 data

検出された侵入(%)



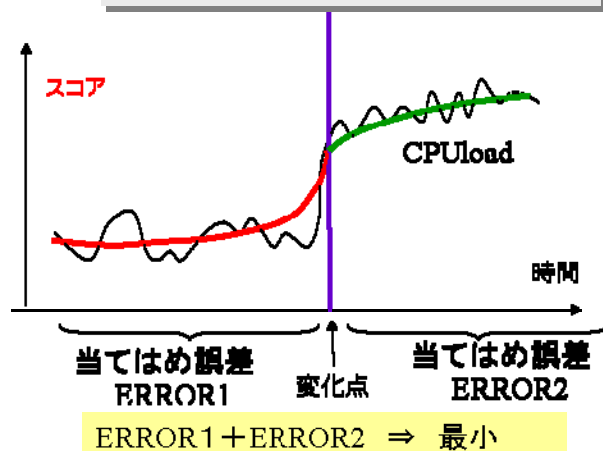
変化点検出エンジン *ChangeFinder*

目的: 時系列データ中に異常が集中発生した時点(変化点)を検出

機能: 時系列データをリアルタイムに学習して変化点スコアを算出

応用例: DoS 攻撃検出、ワーム検出、異常トラフィック検出

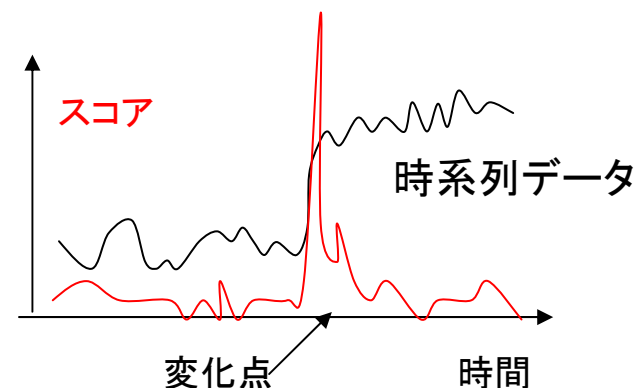
従来手法(統計的検定法)



[Guralnik & Srivastava 99] など

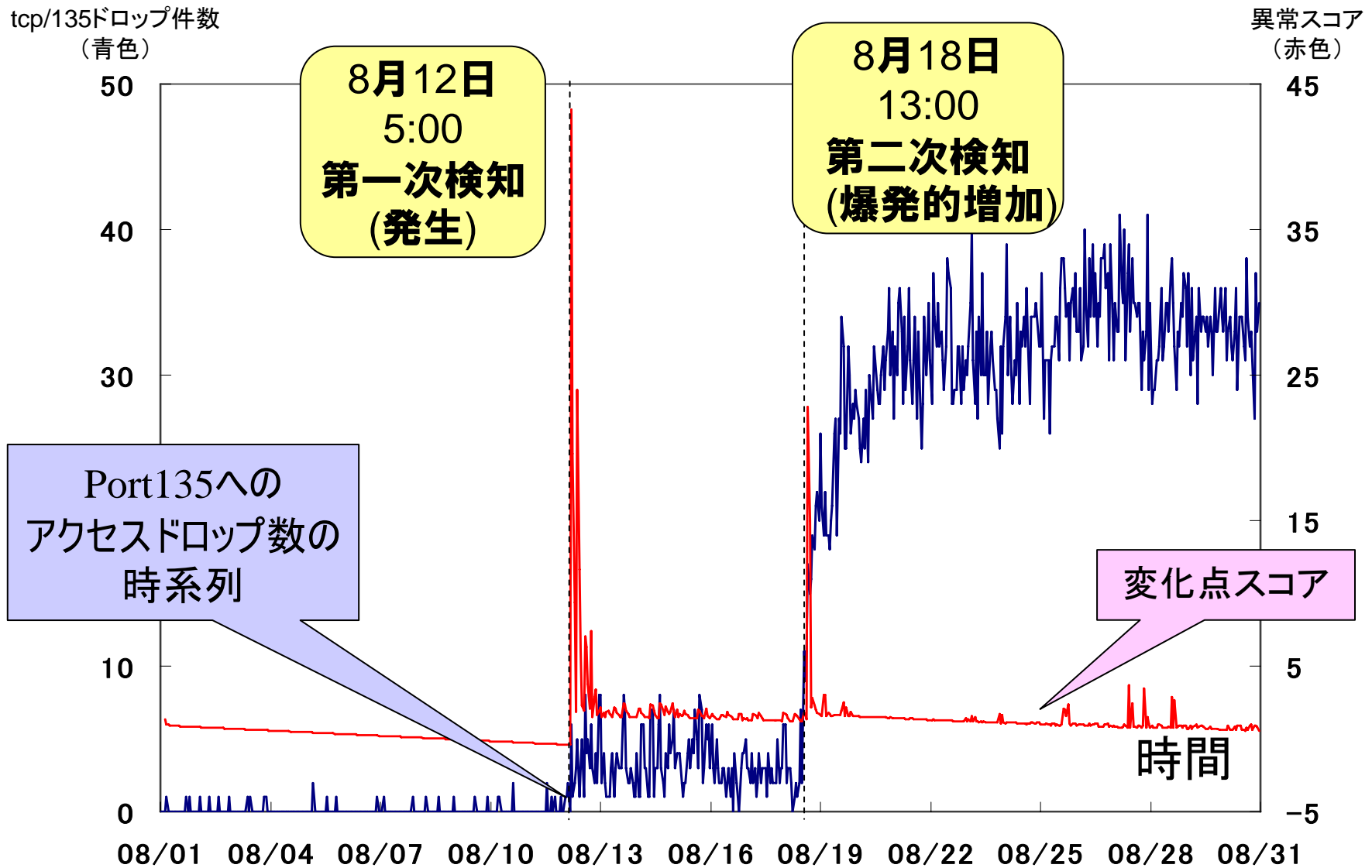
- バッチ処理
 - 全ての変化点候補に対して統計的検定を繰り返す
- ⇒ **計算コスト大**

ChangeFinder



- リアルタイムに変化点スコアを計算
- 適応性
- **高速性** $O(n^2)$ から $O(n)$ へ

ワーム検出への応用例1(MS.Blast)



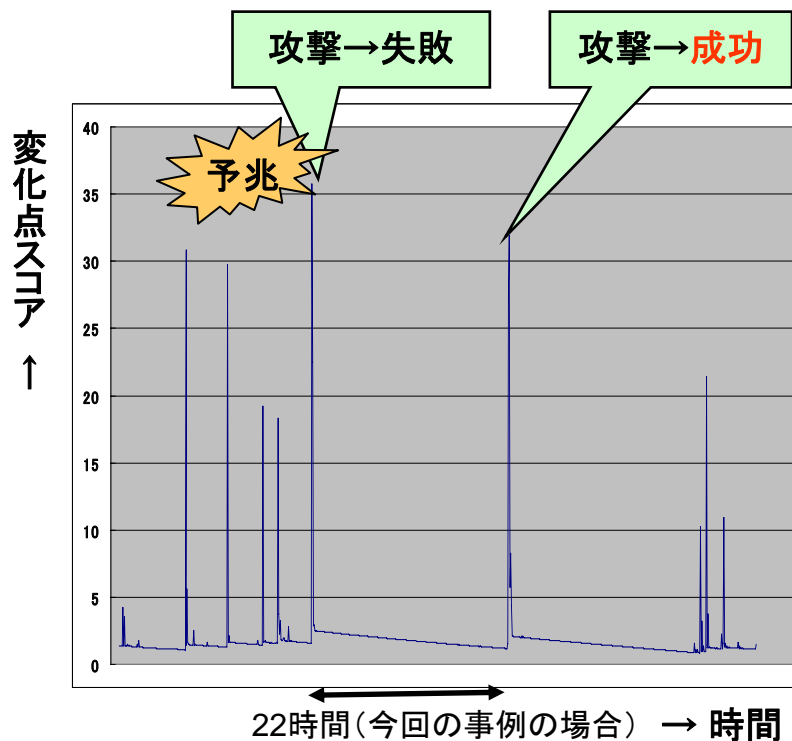
Webアクセスログの分析

検出率100%・誤検出率0.3%でSQLインジェクション攻撃の開始をリアルタイムに検出

分析対象: 実際に攻撃のあった事例の3日間344万行のログ

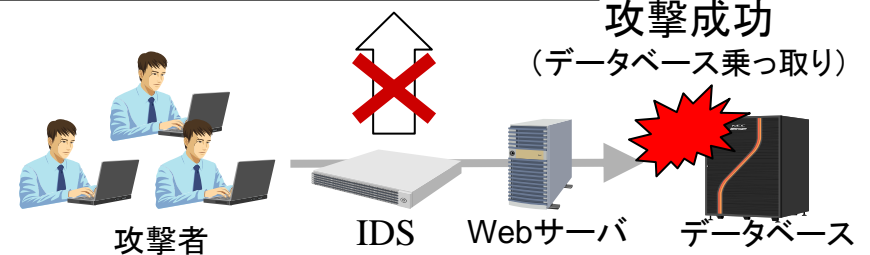
分析結果: 12点の変化点を検出。検出率100%。誤報率0.3%

トラフィックの急激な増加を捉えることでSQLインジェクション攻撃の開始を検出



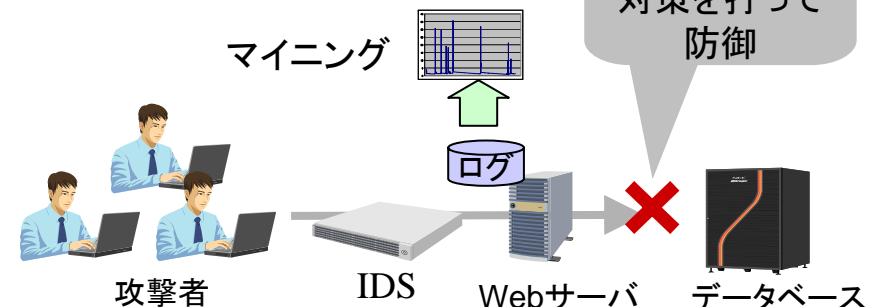
適用前

シグネチャに合えばIDSが攻撃を検知するが、それ以外はすり抜ける



適用後

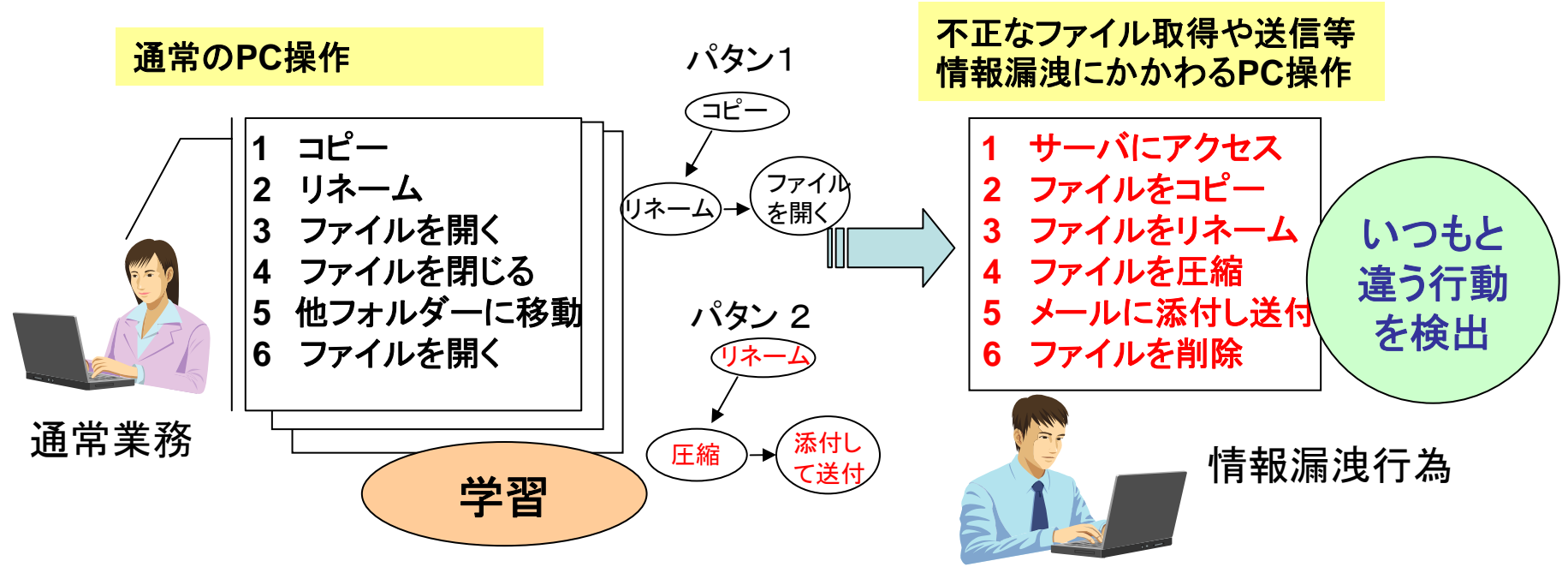
シグネチャベースでは検知できない未知の攻撃の予兆を検出



異常行動検出エンジン AccessTracer

シンボルの系列からパターンを学習、異常な系列をパターンからのずれとして検出

PCの操作履歴を用いた「情報漏洩対策の例」



適用候補

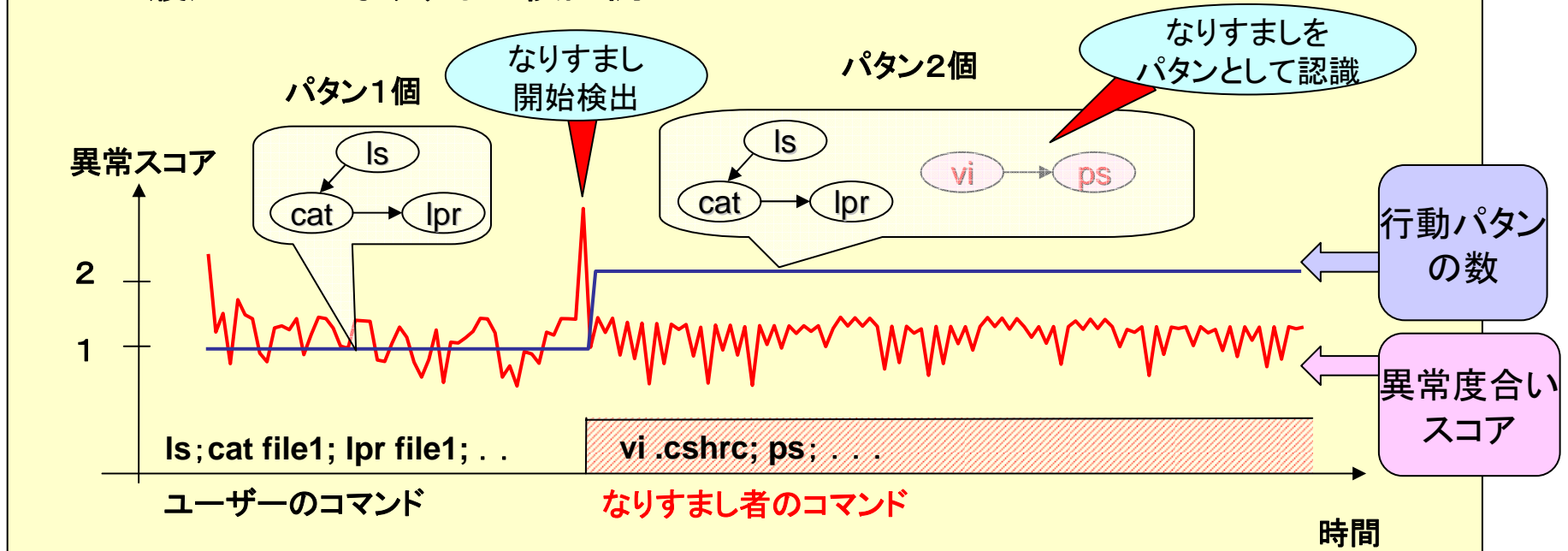
<p>セキュリティ</p>	<p>PCの操作履歴からの情報漏洩対策 トランザクションログからの不正業務検出 Webログからの不正アクセス検出、ログからの未知ワーム検出</p>
<p>OMCS</p>	<p>Syslogからの未知パタンの障害検出、CPUロードからの異常検出</p>

AccessTracer の主要機能

- 機能**
1. 異常系列の出現を早期に検出してアラームを出す
 2. 新しい行動パタンの出現を特定

なりすまし検出実験で従来法 (Naïve Bayes) よりも誤警報を半減

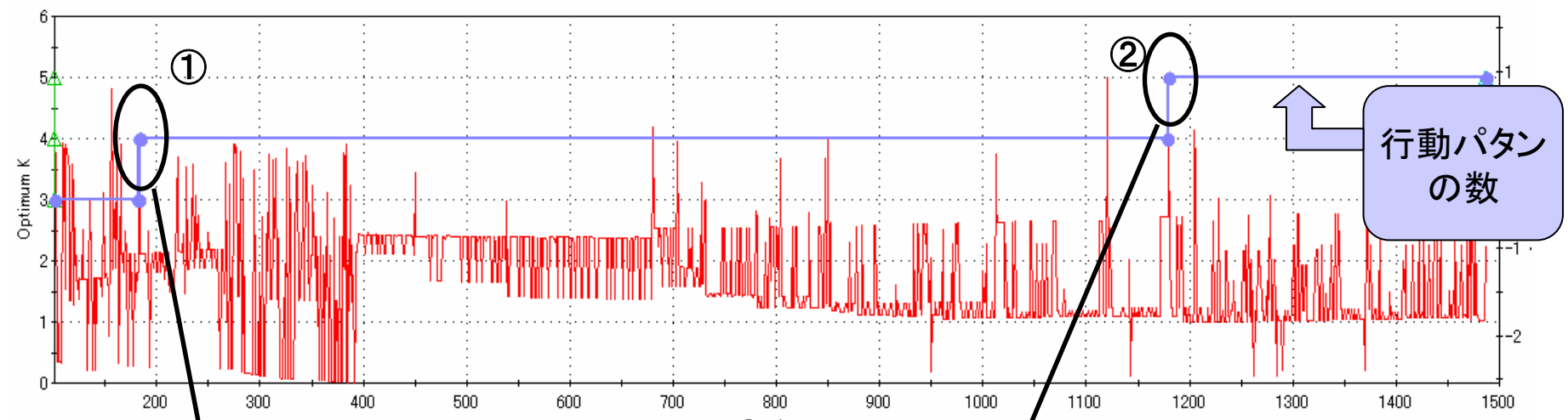
コマンド履歴からのなりすまし検出例



- 行動パターンをリアルタイムに適応的に学習 (オンライン忘却型学習アルゴリズム)
- 新しい行動パタンの出現をダイナミックにトラッキング (動的モデル選択理論)

世界初の機能

未知のワーム検出への応用1



cluster probability (session# 184)			
C. 1	HTTP_Server_ID	-> HTTP_Server_ID	0.342
C. 2	FTP_Pasv	-> FTP_Pasv	0.329
C. 3	SNMP_Community	-> TCP_Probe_Other	0.169
C. 4	SMB_Filename	-> SMB_Filename	0.161

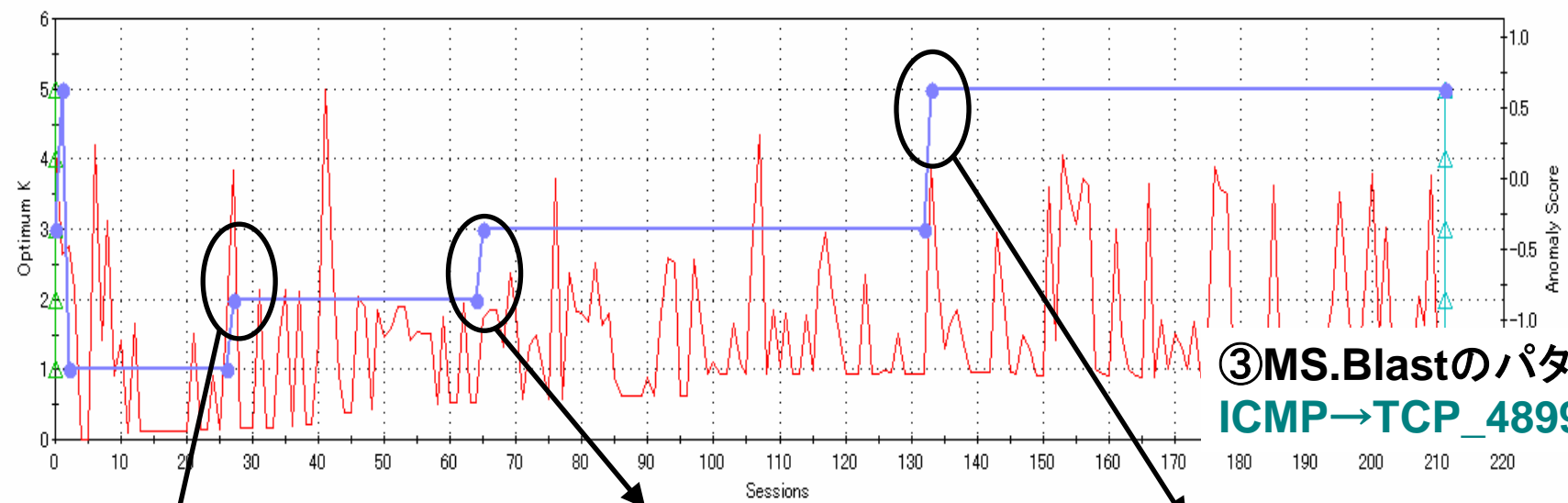
①未知パタンのワーム
 Bat.Mumu.A.Wormのパターンを特定
 パスワード認証の繰り返し行為

cluster probability (session# 1179)			
C. 1	SMB_Filename	-> SMB_Filename	
C. 2	SMB_Filename	-> SMB_Filename	
C. 3	HTTP_Server_ID	-> HTTP_Server_ID	
C. 4	Audit_TFTP_Get_Filename	-> BackOrifice_Ping	
C. 5	FTP_Pasv	-> FTP_Pasv	

②擬似攻撃(Windows Enumeration
 Check by SuperScanV4)のパターンを特定
 脆弱性スキャンの繰り返し行為

【効果】 未知パタンのワームの行動の特徴を発見

未知のワーム検出への応用2



③MS.Blastのパターン
ICMP→TCP_4899

cluster probability (session# 27)		
C. 1	FW/TCP_MS-RPC → FW/TCP_4444	0.502
C. 2	FW/TCP_nbsession → FW/TCP_445	0.498

①MS.Blastのパターン
MS-RPC とポート4444使用

cluster probability (session# 65)		
C. 1	FW/TCP_5000 → FW/TCP_445	0.3
C. 2	FW/TCP_nbsession → FW/TCP_445	0.3
C. 3	FW/TCP_nbsession → FW/TCP_nbsession	0.3

②TCP_445を使用するワームのパターン (Lovgate or Gaobot?)
アクセス可能なファイル共有を探す行為

cluster probability (session# 133)		
C. 1	ICMP → FW/TCP_4899	
C. 2	FW/TCP_5000 → FW/TCP_445	
C. 3	IDS/Netbios_Session_Request → IDS/Netbios_S	
C. 4	FW/TCP_445 → FW/TCP_nbse	
C. 5	FW/TCP_MS-RPC → FW/TCP_4444	

④DOS攻撃のパターン
ポート139(NetBIOS Session Service)を使用

【効果】 IDSとFWのログの相関分析により、未知の行動パターンの特徴を発見

WebSAM LogCollector (製品)

① ログ一括収集機能 (エージェント/マネージャ)

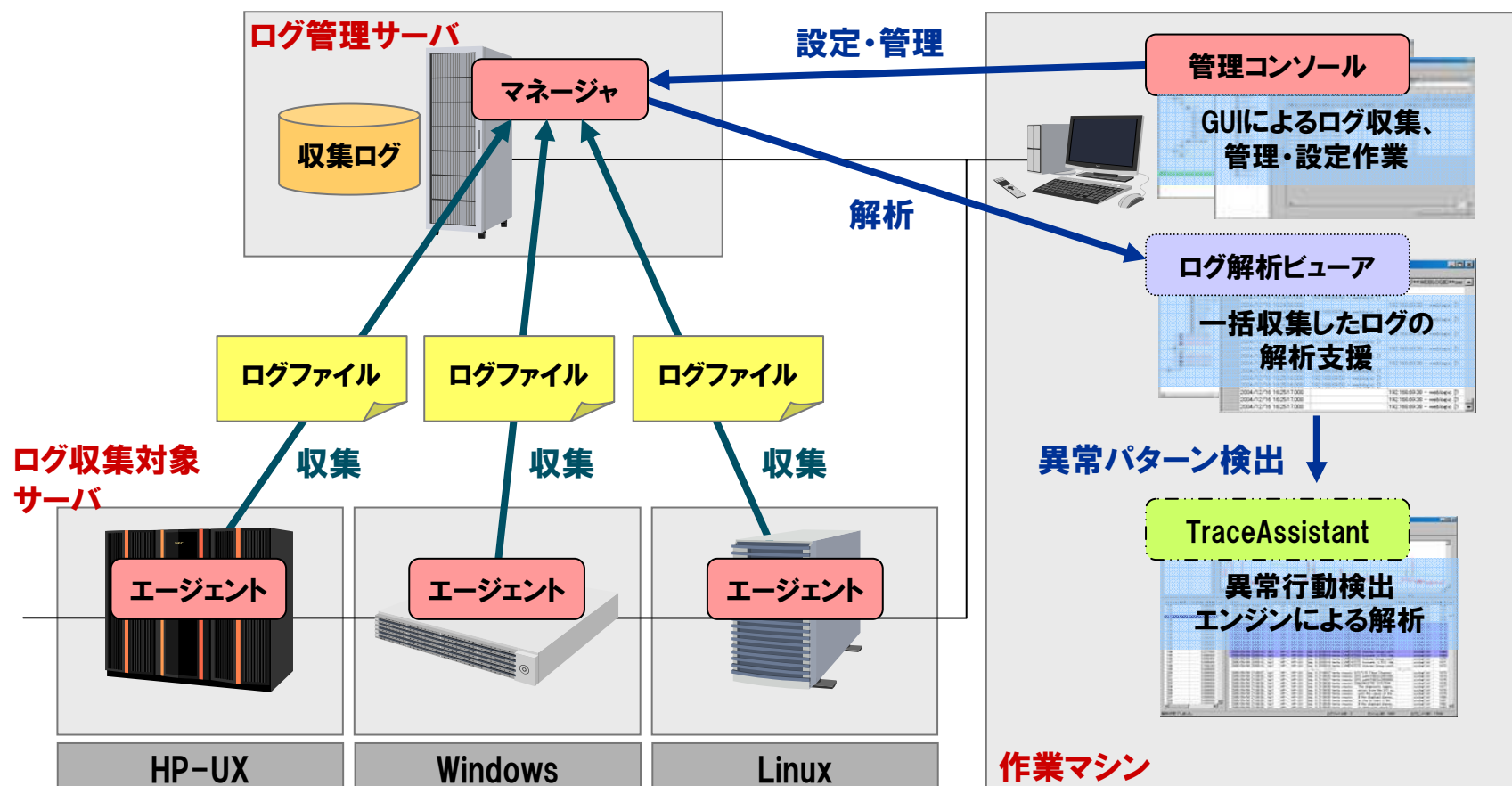
- 複数のサーバに散在する各種ログを一括収集・一元管理

② ログ解析支援機能 (ログ解析ビューア)

- サーバ別・機能別に出力される各種ログを、単一の画面で、マージやソート、絞り込み検索が可能

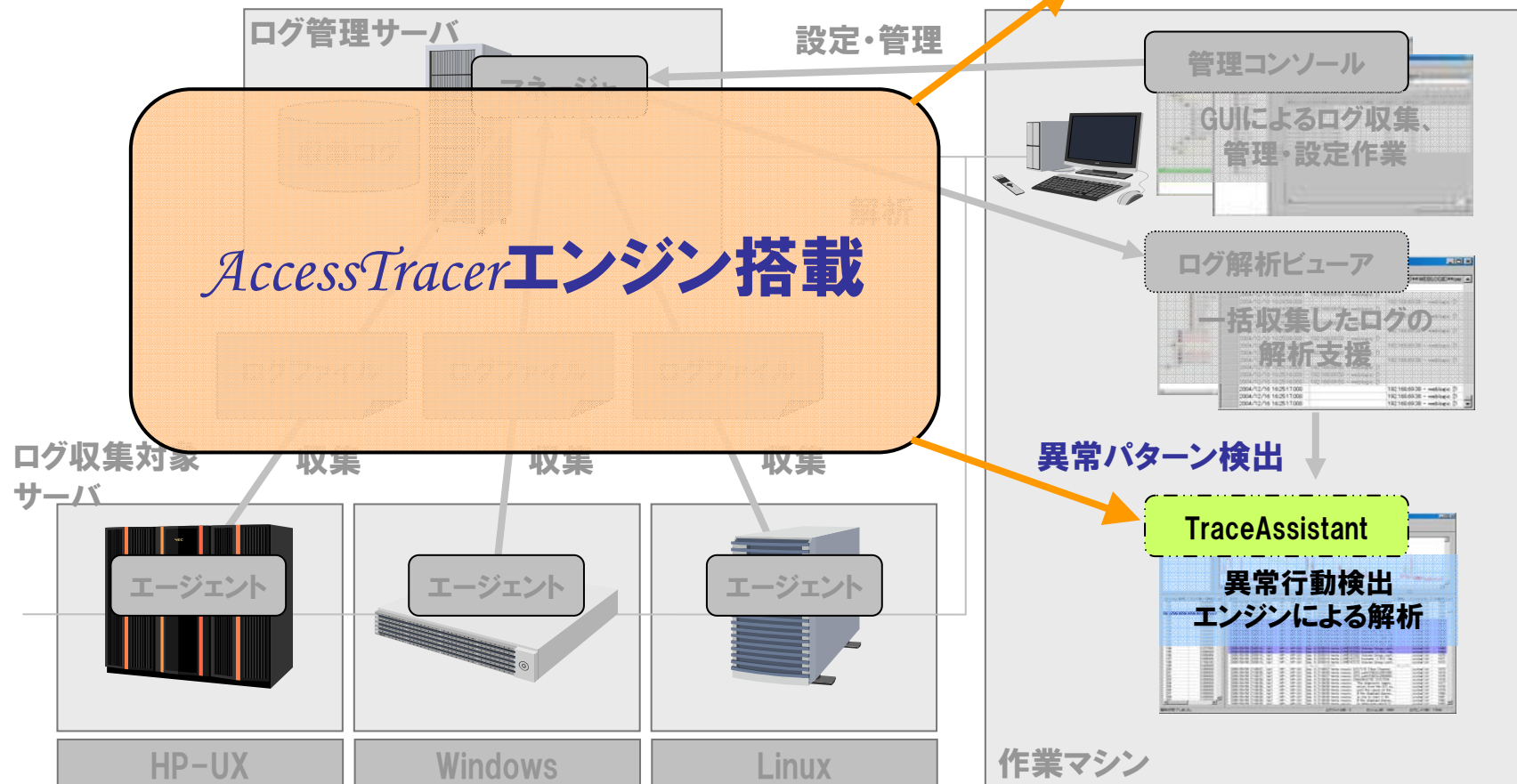
③ ログ解析支援機能 (TraceAssistant)

- 収集ログをパターン分析し、異常性の高い(通常と異なる)パターンを検出

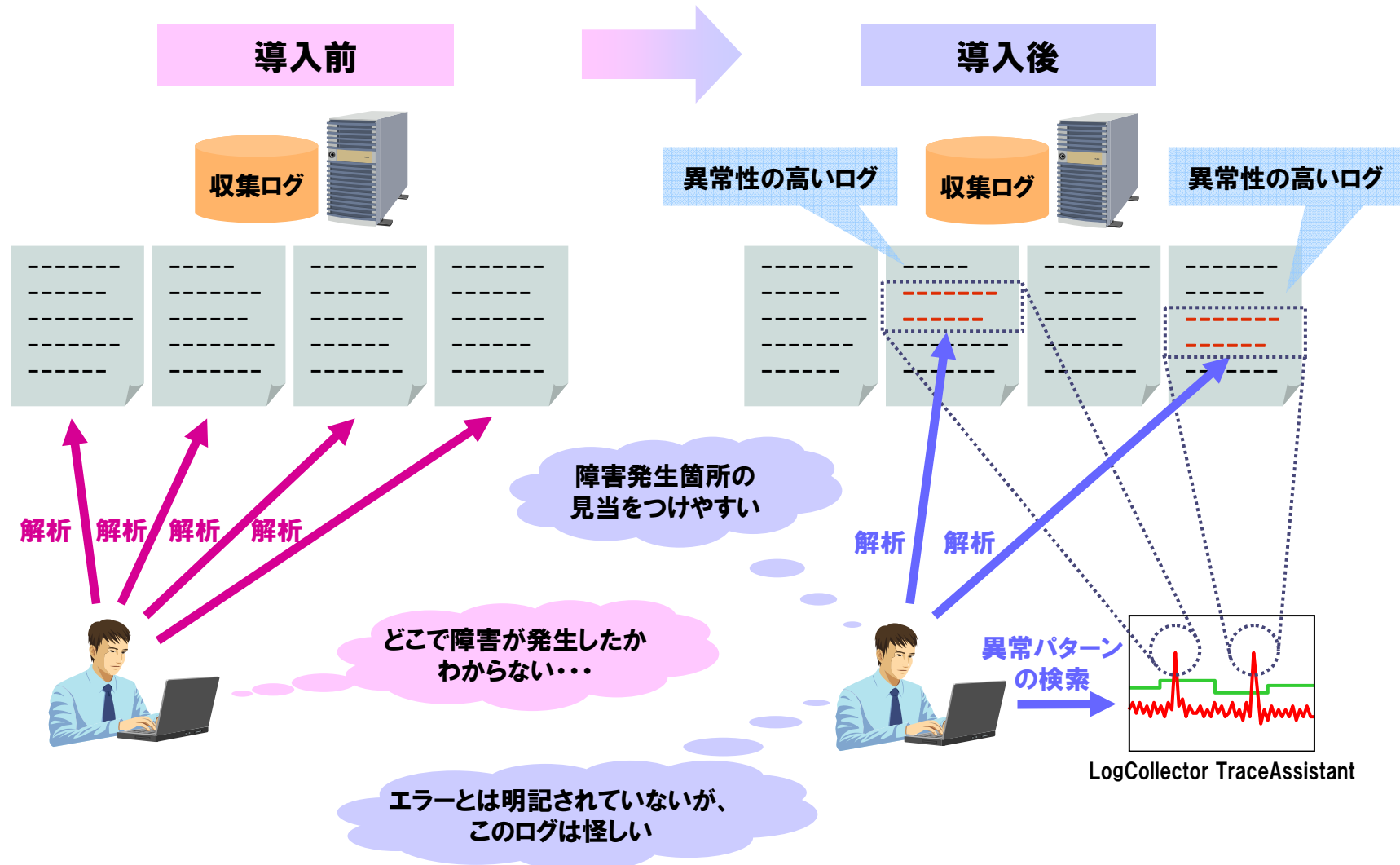


WebSAM LogCollector (製品)

- ① ログ一括収集機能 (エージェント/マネージャ)
 - 複数のサーバに散在する各種ログを一括収集・一元管理
- ② ログ解析支援機能 (ログ解析ビューア)
 - サーバ別・機能別に出力される各種ログを、単一の画面で、マージやソート、絞り込み検索が可能
- ③ ログ解析支援機能 (TraceAssistant)
 - 収集ログをパターン分析し、異常性の高い(通常と異なる)パターンを検出



TraceAssistant活用イメージ



トレンド分析エンジン TrendLiner

時系列データから高速に学習、高精度に将来を予測するエンジン

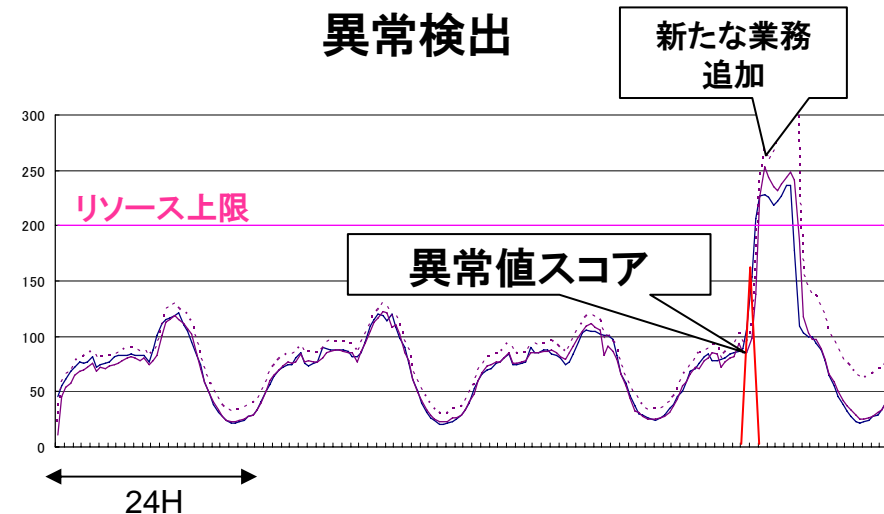
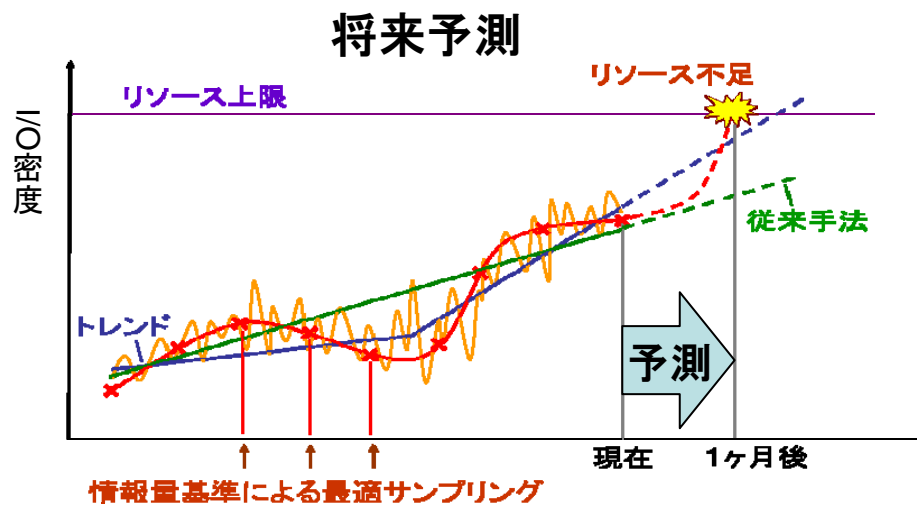
■変動する数値データの高精度長期的予測が可能

従来手法(状態空間モデル予測法)に比べて**誤差1割程度減**
線形回帰法に比べれば**誤差3割減**

■多数の対象を高速に予測

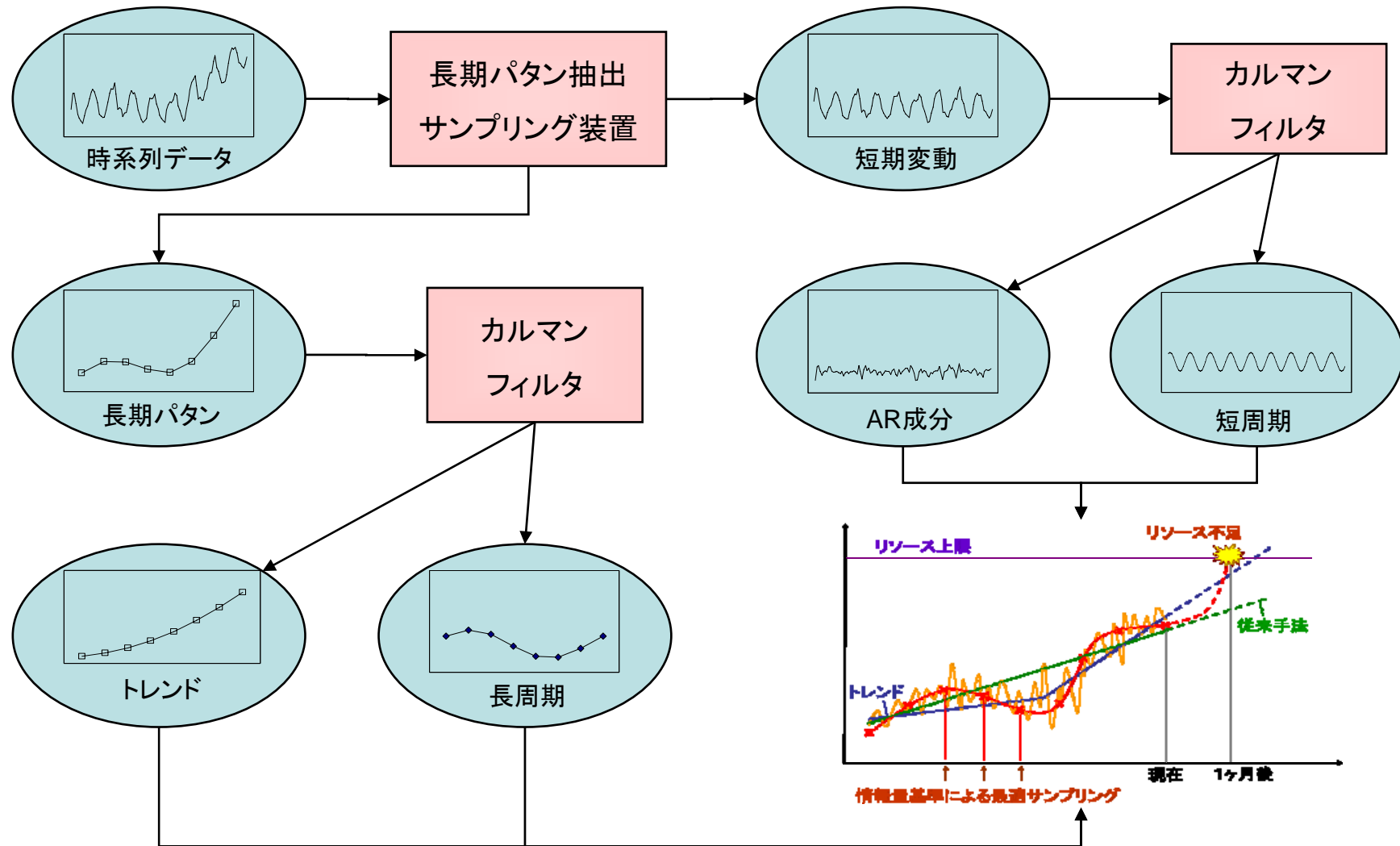
従来手法(状態空間モデル予測法)に比べて**100倍の高速化**
→1000台の機器の3ヶ月先を1分以下で予測

* 状態空間モデル予測法～精度上のベンチマーク



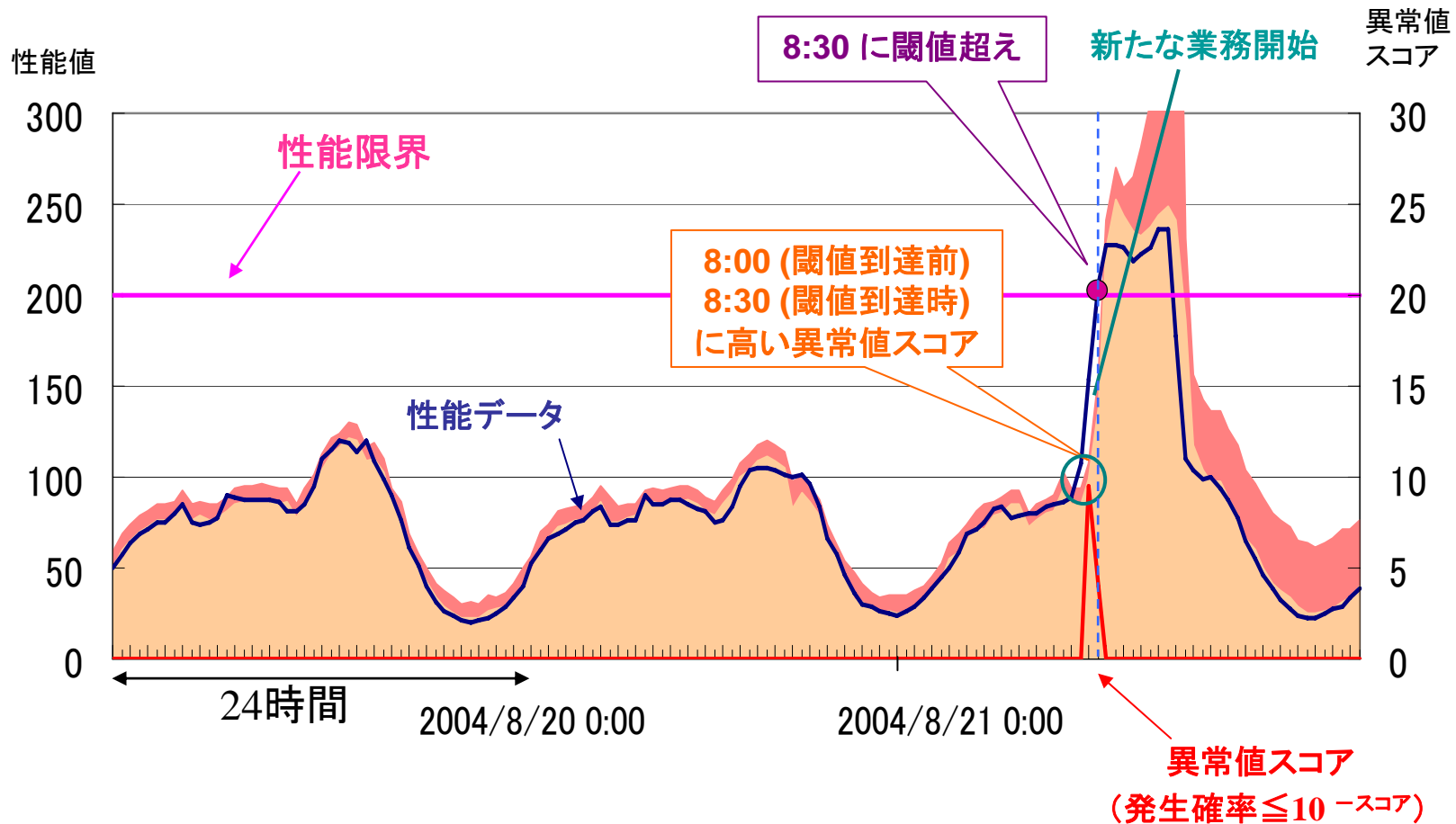
ポイント：階層的モデリング

長期的パターンと短期的パターンを分離することで高速化

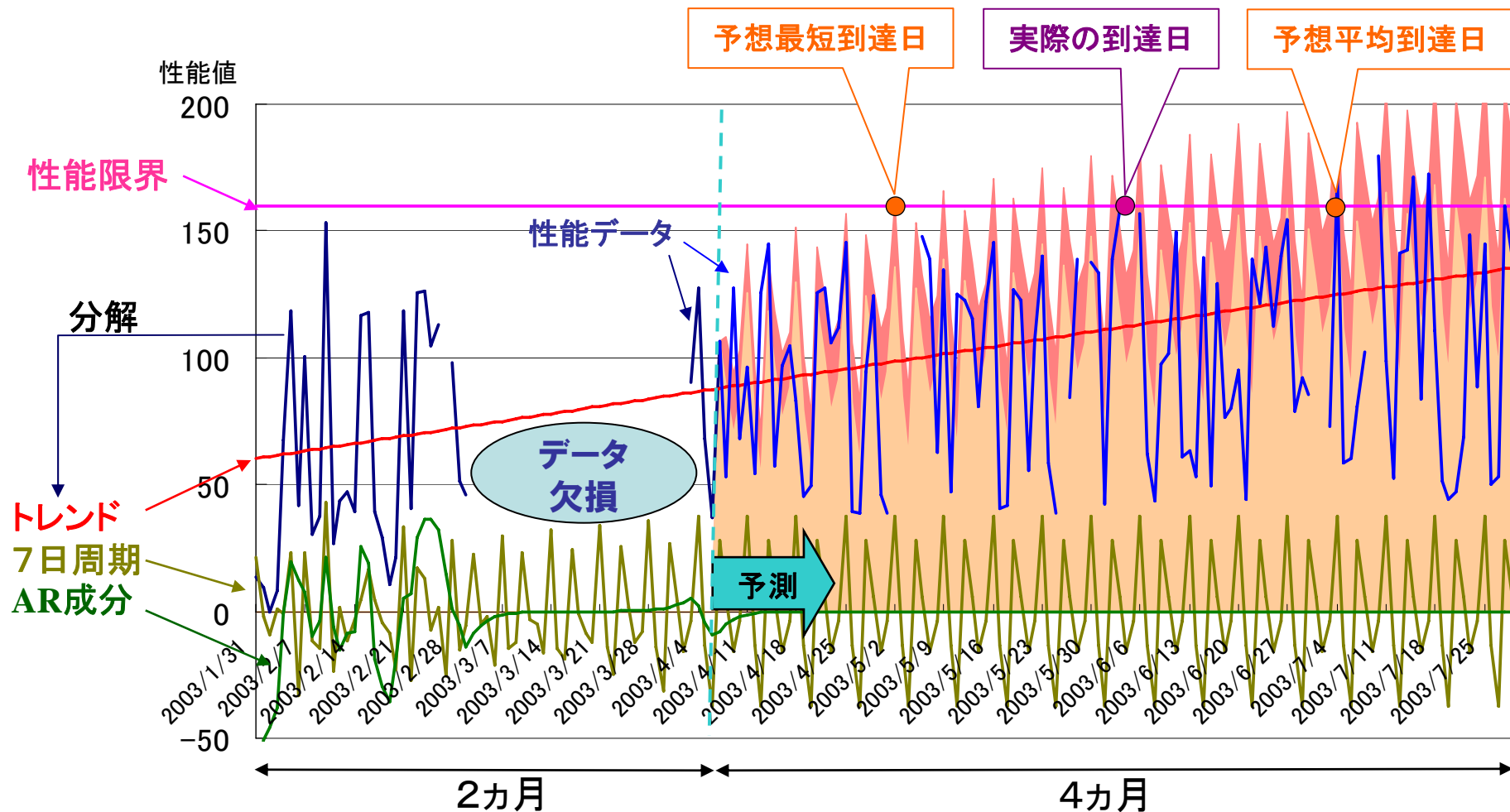


ネットワークトラフィック異常検出への応用

周期性を考慮しながら、いつもと違う異常の発生をリアルタイムに検知



BIGLOBEストレージ・転送レート予測への応用



欠損データを補完しつつリアルタイムにトレンドと周期成分を抽出、転送レートのピークが性能限界に達する時期を的確に予測

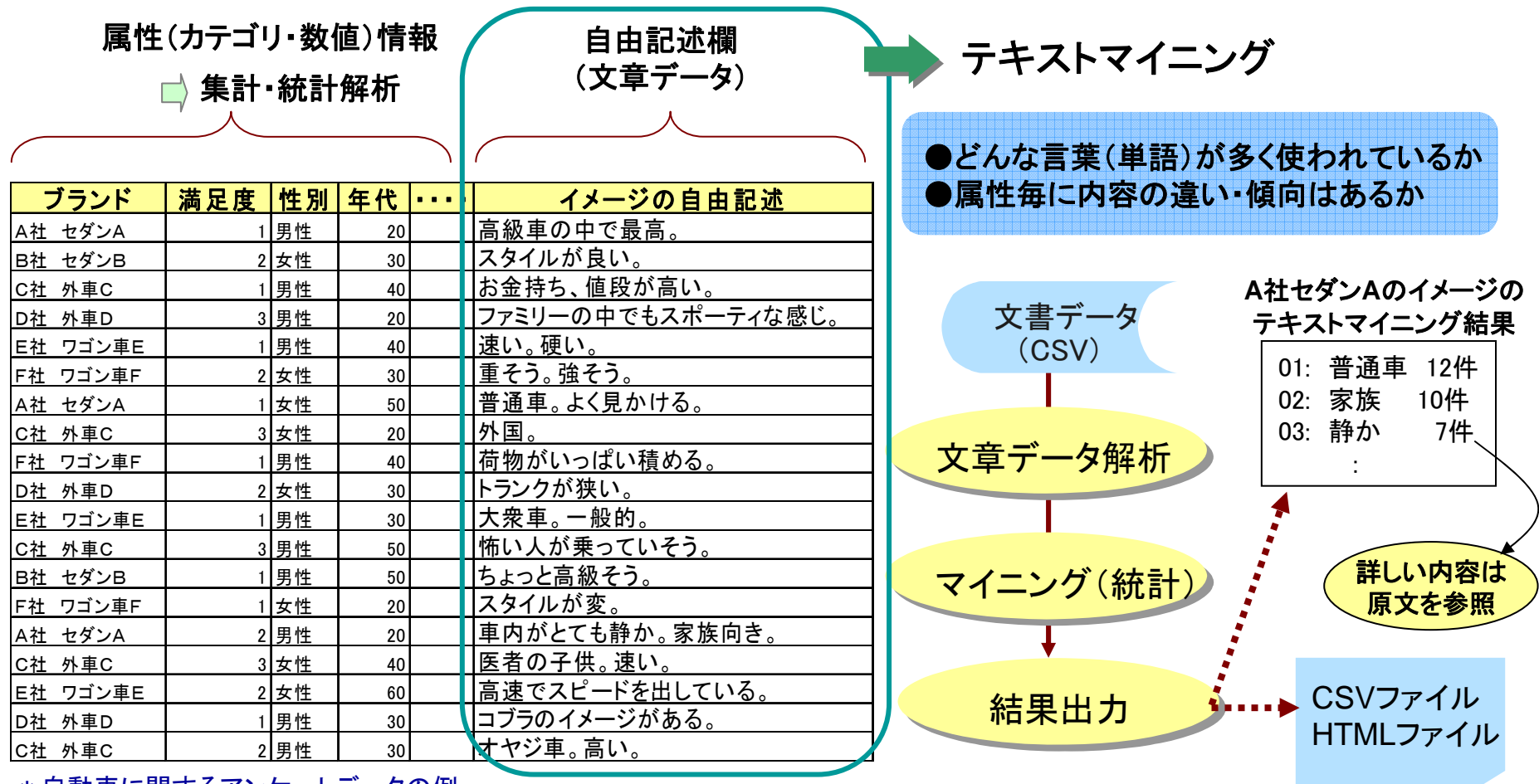
テキストマイニング技術



U can change.

テキストマイニングとは

テキストマイニングとは、アンケートの自由記述欄のような文章データを対象に、その内容（キーワード・傾向等）を定量的に分析するための技術です。

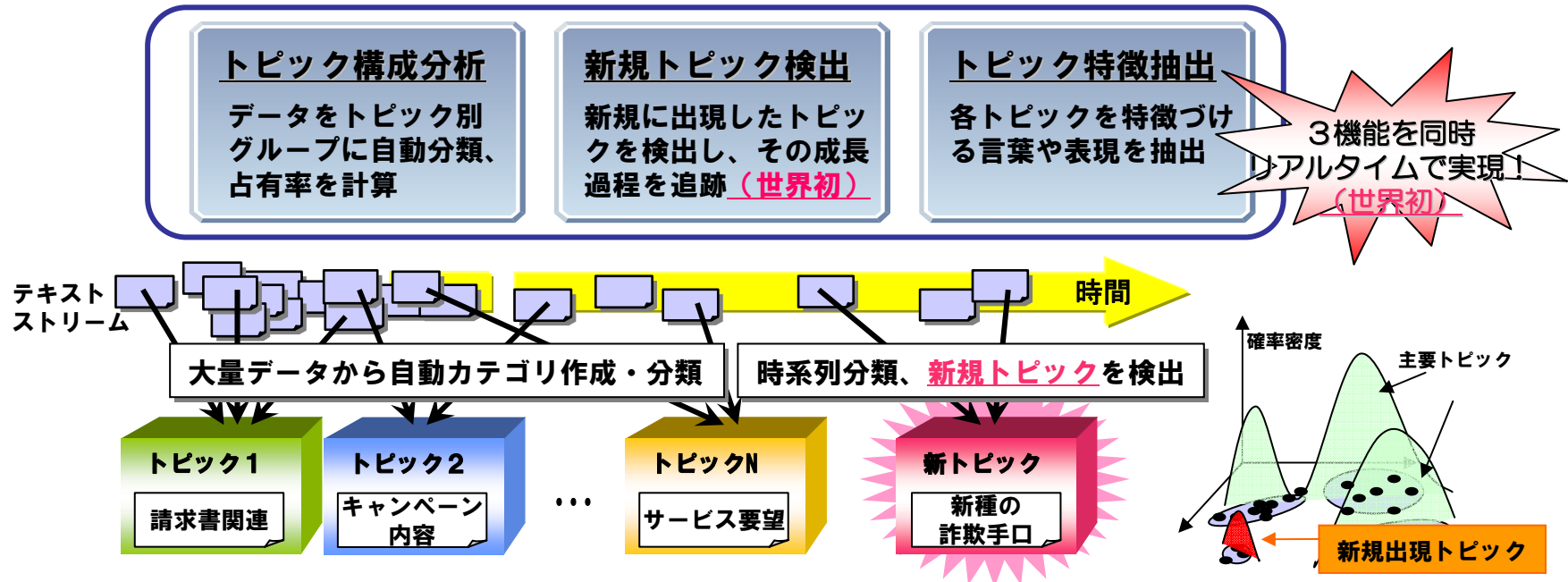


* 自動車に関するアンケートデータの例

TopicAnalyzer

- 逐次流れてくるテキスト情報を動的にクラスタリングし、トピックの時間的変化を捉える、新規トピックを検出することが可能。
- リアルタイムにトピック(意味)レベルでの分析が可能。
 - 従来: キーワードレベル または スタティック実行(蓄積系データ対象)
- 学習・忘却(トピック消滅)の仕組みを採用することで、高精度に実現。

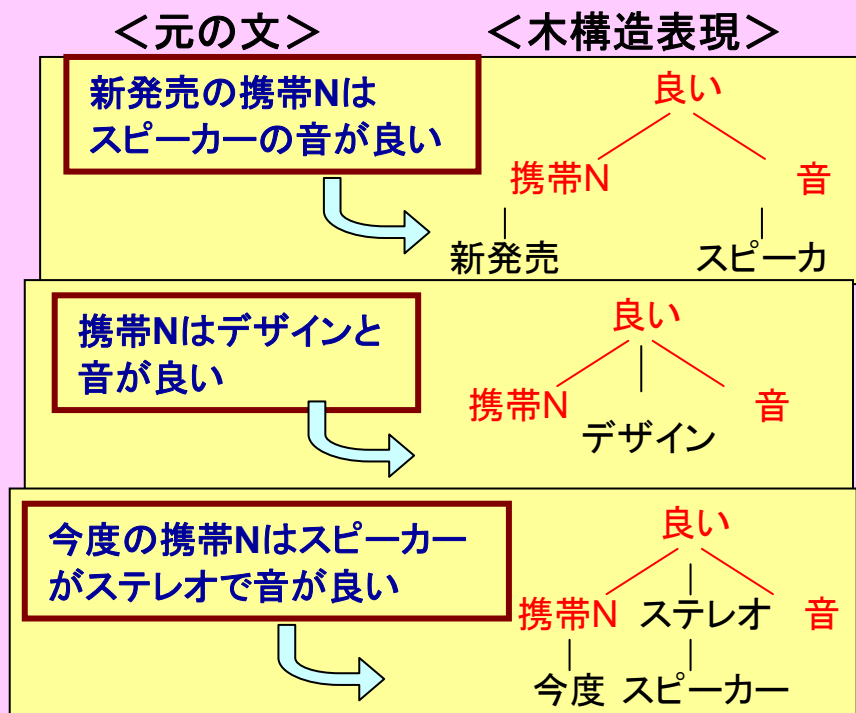
トピックの傾向の時間的変化をリアルタイムに分析



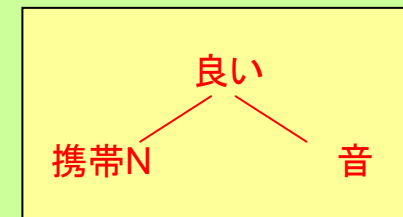
文脈マイニング(1/2)

- 文の意味を表現する木構造(構文木)をマイニングし、入力テキスト共通の意味を抽出
- 互いに包含・重複関係にある構文木をまとめあげ、記述内容として特徴的なものを抽出

① 文を木構造で表現



② 共通する特徴的な部分木を抽出



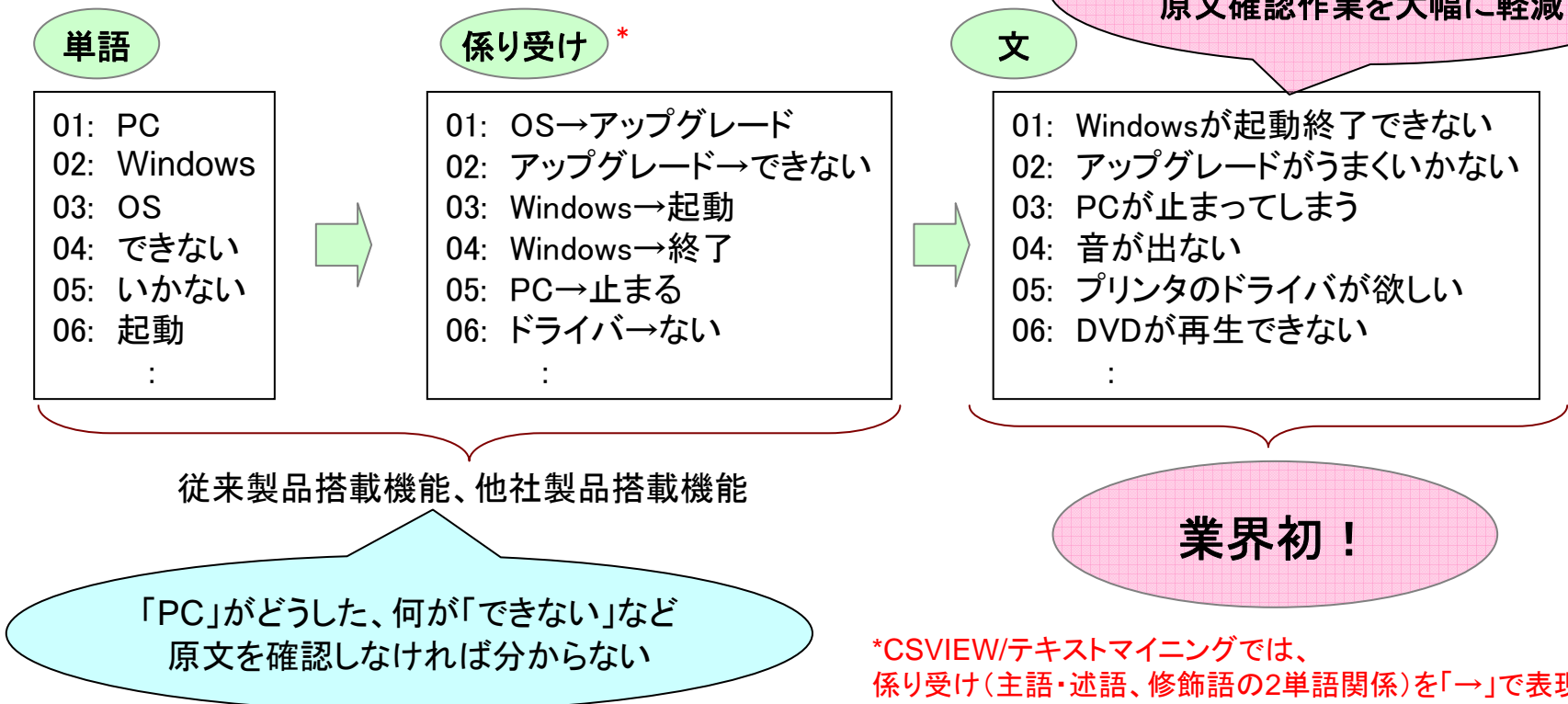
③ 部分木から文(再生成文)を生成

「携帯Nは音が良い。」

文脈マイニング(2/2)

結果を読みやすい文にまとめあげる、業界初の「文脈分析」として
製品(CSVIEW/テキストマイニング)に搭載
分析結果を従来の「単語」や「係り受け」に代わり、読みやすい「文」で提示

■「OSアップグレード」に関する問い合わせ分析をした結果表示例:



ブログをマイニング～企業活動に活用 (eHyouban／マイニングサービス)



U can change.

外部情報をマネージすることへの ニーズの高まり

企業活動の上で、次のようなニーズが大きくなっている？…

- 自分たちのプロモーションや施策は、当たっているのだろうか
- うちのブランドイメージ どうなってる？
- 企業説明会の情報を、学生同士がやりとりしてるようだ
- 情報漏えいを、外部から指摘された。まずい…

リーチできるチャンネルが限定されているので、小刻みに反応をみることや、問題の早期発見が難しい。



自社のデータと外部情報、かけあわせて見る事ができると…

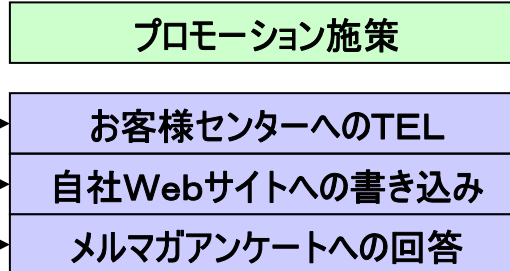
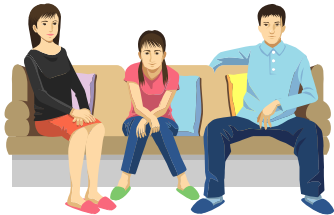
- 広報ニュース発表の波及が、記者BlogやキーマンWebでメッセージが伝播していないぞ … 波状攻勢をかけよう！
- 掲示板で気になる話題が … コールセンターにも同様な内容が来ていないか、即調査だ！

すばやく考え、
次の手を打てる

ブログを対象にデータマイニング～企業活動に活用

データマイニングとは： 大量に蓄積されたデータを解析し、項目の相関関係やパターンなどを発見する技術。
発見されたナレッジは、各種 経営戦略などに利用される。

◆消費者からの情報(従来)



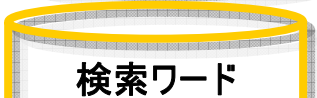
問題点

プロモーションの効果が分からない

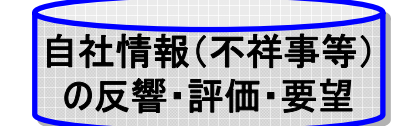
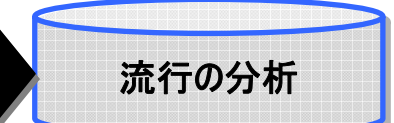
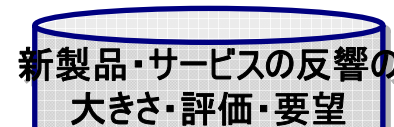
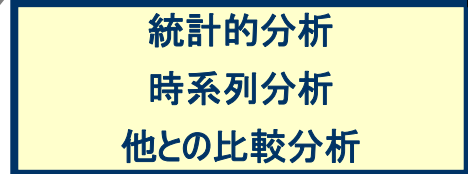
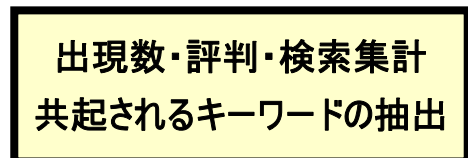
問題点

1. 偏った意見(クレマー、ファン)が多く、一般的な評価が分かりづらい
2. 商品・サービス企画等の戦略データとするには意見の量が十分ではない

◆データマイニング



「一般的な意見」の
「膨大なデータ」が存在



- ・プロモ効果測定
- ・自社ブランディング
- ・新規ビジネス企画
- ・危機管理

従来の顧客接点では捕捉できなかった情報の抽出・可視化を可能に

問い合わせ先

- **各エンジンの理論**

マイニングRG

mailto: mining@labs.jp.nec.com

- **エンジンの適用**

データマイニングセンター

mailto: dmc@labs.jp.nec.com

Empowered by Innovation

NEC