見えログ: 人間による計算機
ログ解析を支援する
ログ情報ブラウザ

高田 哲司

Sony CSL, Interaction Lab.

DBWeb 2003

ログ情報とは?

こんなものである

kterm

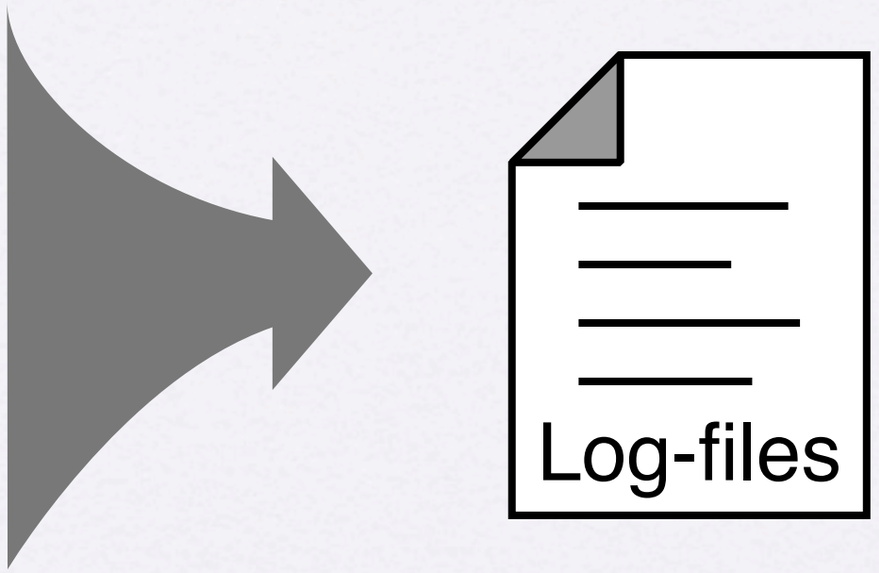
```
Jan 31 12:05:23 6E:hostA ftpd[66021]: FTP LOGIN FROM pisa.italia.is.foo.co.jp as
 taro
Jan 31 12:40:26 6D:hostA ftpd[63641]: connection from pisa.italia.is.foo.co.jp
Jan 31 12:40:26 6E:hostA ftpd[63641]: FTP LOGIN FROM pisa.italia.is.foo.co.jp as
 taro
Jan 31 13:10:55 6D:hostA ftpd[66014]: connection from pisa.italia.is.foo.co.jp
Jan 31 13:10:55 6E:hostA ftpd[66014]: FTP LOGIN FROM pisa.italia.is.foo.co.jp as
 taro
Jan 31 13:47:50 6D:hostA ftpd[60545]: connection from pisa.italia.is.foo.co.jp
Jan 31 13:47:50 6E:hostA ftpd[60545]: FTP LOGIN FROM pisa.italia.is.foo.co.jp as
 taro
Jan 31 14:40:21 6D:hostA ftpd[63796]: connection from pisa.italia.is.foo.co.jp
Jan 31 14:40:21 6E:hostA ftpd[63796]: FTP LOGIN FROM pisa.italia.is.foo.co.jp as
 taro
Jan 31 15:11:11 6D:hostA ftpd[66461]: connection from pisa.italia.is.foo.co.jp
Jan 31 15:11:11 6E:hostA ftpd[66461]: FTP LOGIN FROM pisa.italia.is.foo.co.jp as
 taro
Feb  1 08:33:09 5E:hostA su[66526]: failed: ttyq2 changing from taro to root
Feb  1 08:33:12 5E:hostA su[66496]: failed: ttyq2 changing from taro to root
Feb  1 08:33:15 5E:hostA su[66524]: failed: ttyq2 changing from taro to root
Feb  1 08:33:19 5E:hostA su[66467]: failed: ttyq2 changing from taro to root
Feb  1 08:33:23 5E:hostA su[66474]: failed: ttyq2 changing from taro to root
Jan 31 17:05:26 6D:hostA ftpd[66583]: connection from pisa.italia.is.foo.co.jp
Jan 31 17:05:26 6E:hostA ftpd[66583]: FTP LOGIN FROM pisa.italia.is.foo.co.jp as
 taro
Jan 31 19:30:54 6E:hostA login[66782]: ?@siena.italia.is.foo.co.jp as nishi
Jan 31 19:39:30 6E:hostA login[66877]: ?@siena.italia.is.foo.co.jp as nishi
Feb  1 13:28:09 6B:hostA runpriv[61891]: Running privilege ipld for user taro.
Feb  1 13:28:10 6B:hostA runpriv[40161]: Running privilege intlset for user taro
.
Feb  1 13:28:27 6B:hostA runpriv[67063]: Running privilege dtshutdown for user t
aro.
```

計算機ログ情報

各種サーバ

監視システム

検知システム



稼働状況把握　　**種々の異常検出**　　**不正侵入対策**

システム管理にとって重要かつ唯一の情報源

ログ情報の調査は必要不可欠

現状は?

- やってない (定期的に実施 44%)
- 地道に手作業
- 既成/独自のスクリプトで定期的に処理
- 既成のログ監査システムを利用

ログ調査の問題点(1)

- ログ情報認識負荷の大きさ

  文字情報/膨大な量

  = 単調かつ退屈な長時間作業

- 注目すべきログ情報の不明確さ

  どれが不正行為の痕跡かが不明確

  = 異常事象/ログ自体への知識が必要

ログ調査の問題点(2)

- 手作業

  人間は異常認識/判定能力は高いが、
  長時間の単調作業に正確性を求めるのは困難

- 自動化は困難?

  キーワード抽出等だけで大丈夫?
  現状で抽出できている情報だけで大丈夫?
  今はいいが、今後は?

人間側の問題

- 認識負荷の問題
- 長時間の単純作業
- 専門知識や経験

「決してほめられない作業」という実態も...

ログ側の問題

- 文字情報
- 膨大な量
- 多様な記録内容、データ形式と偏在性
- 注目すべき情報の不明確さ
  - すべての記録が重要ではない
  - あらゆるログメッセージを知る人なんていない

現状のログ生成は調査のことを考慮していない

何が必要か?

人間による調査を支援する

ログの望ましくない特性を補う

Human Interfaceが必要!

なぜブラウザなのか?

- 自動化は困難

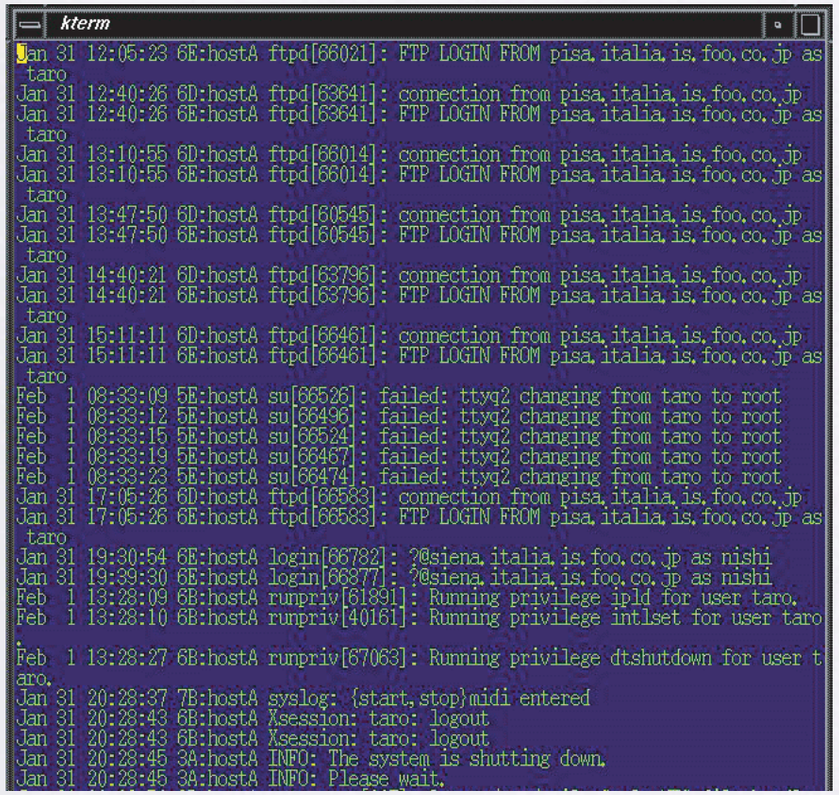
  人間が見ることが大事

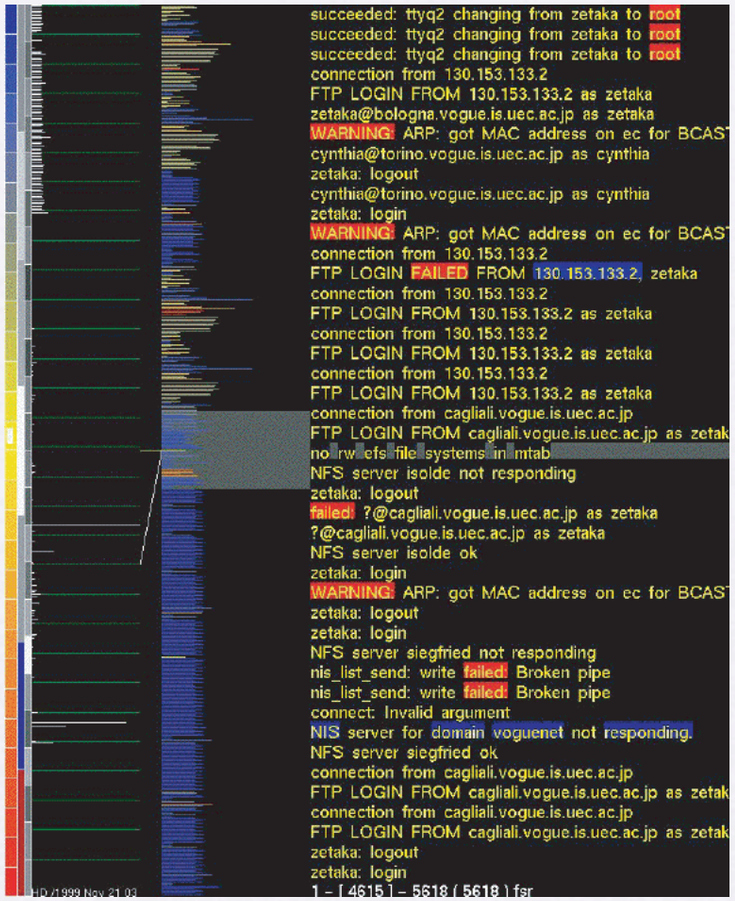
- 人間による調査が必要

  対話性が必要不可欠

見えログ

視覚的&対話的ログ情報ブラウザ





見えログの特徴

1. 情報視覚化

    情報認識の容易さ、対話的操作

2. 異常ログメッセージの抽出

    頻度解析と視覚的パターンを用いた抽出支援

3. ログの統合

    時間に基づくログの統合
    記録フォーマットの違いを吸収

情報視覚化とは?(1/3)

- Scientific Visualization（**情報可視化**)
  - 専門家が使用
  - 科学的現象の解明
  - 物理データ/測定結果/シミュレーション結果が対象
- Information Visualization (**情報視覚化**)
  - 誰もが利用 (インタフェース)
  - 検索/関連の発見
  - 関係などの抽象的なデータが対象
    - プログラムの実行状態、Web構造、ファイルシステム
  - 対話性を重視
    - 図になってからが大事

視覚化の分類(2/3)

	データ	視覚化手法
Scientific Visualization	具体的	固定的
Data Visualization	具体的	自由
Information Visualization	任意	自由

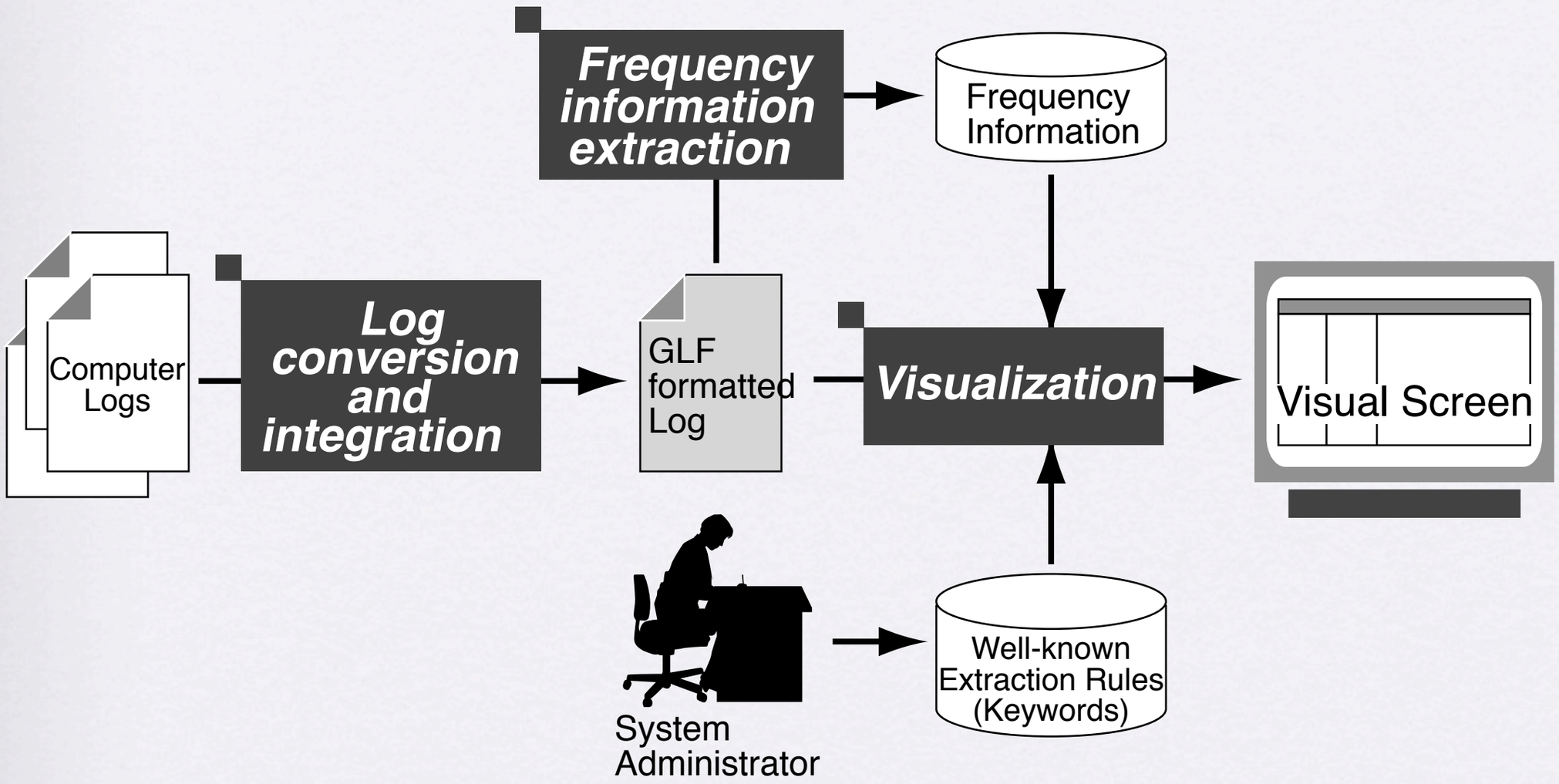
(増井俊之氏のWebより引用)

情報視覚化による利点 (3/3)

- 情報の把握を容易かつ高速に
- 多くの(量/種類)の情報を一度に提示可能
- 対話的作業による直感的なデータ処理

文字情報の認識は逐次処理、画像情報の認識は並列処理

見えログ: Module構成



三種の処理から構成

中間Log形式(1)

General Log Format

ログ変換例

Jan 10 18:42:02 foo.ac.jp in.telnetd[2424]: connect from someone.else.net

Extraction and Conversion

998972366	*foo.ac.jp*	*in.telnetd*	*connect from someone.else.net*
Time	**Tag 1**	**Tag 2**	**Message**

中間Log形式(2)

変換方法は自由

現状は...

Tag1: 計算機(ホスト)名

Tag2: ログを出力したプログラム名

UNIX系OSのsyslogの出力形式と同一

変換の目的

- 種々の計算機ログへの対応と時刻による統合
- 出現頻度解析のためのデータクリーニング

異常を表すログとは...

- 既知の異常を表すメッセージ

- めったに現れないメッセージ
  あれ? みたことのないログが出力されてる

- めったに現れないメッセージ出力状況
  ある特定の時間帯に大量のログ出力

仮説: 異常をあらわすログは、大量の正常をあらわすログの中に埋もれている!

疑わしいログの抽出方法

- 既知の異常を表すメッセージ

  キーワード定義によるパターン検索

- めったに現れないメッセージ

  出現頻度による抽出

- めったに現れないメッセージ出力状況

  視覚的パターンによる抽出

出現頻度解析

ログメッセージ

connect from tokyo
connect from tokyo
connect from osaka
connect from kyoto
connect from osaka
connect from tokyo
connect from osaka
connect from osaka
connect from kyoto
connect from hakata
connect from osaka
connect from tokyo
...

抽出

単語の出現頻度

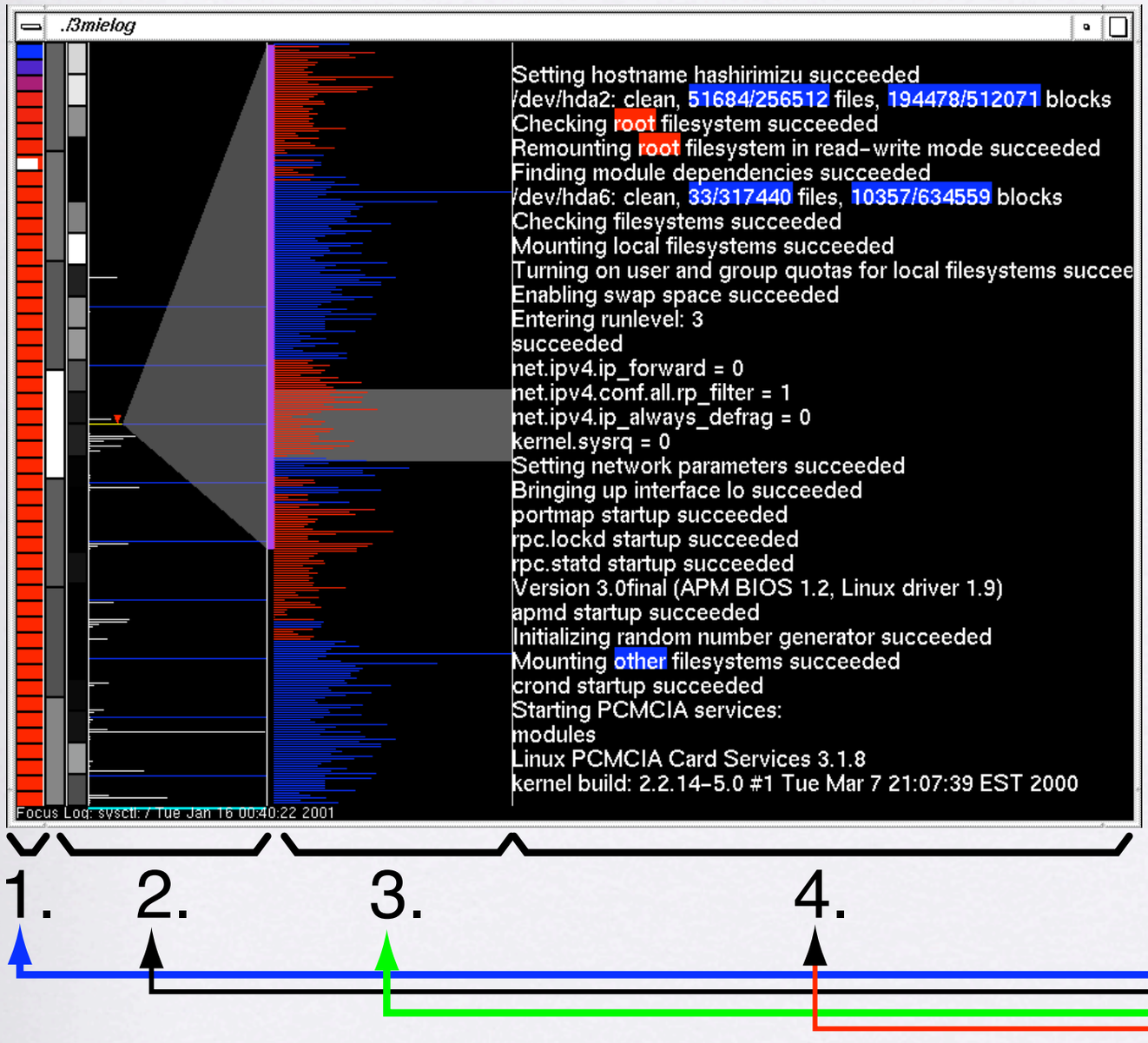
出現数	単語
12	connect
12	from
5	osaka
4	tokyo
2	kyoto
1	hakata

句の出現頻度

出現数	句
12	connect from
5	from osaka
4	from tokyo
2	from kyoto
1	from hakata

調査者が注目すべき
ログメッセージ

画面構成



1. タグ表示領域
2. 時間表示領域
3. アウトライン表示領域
4. メッセージ表示領域

タグ情報表示領域

- 格子の数
  ログを出力したプログラムの種類
- 格子の色
  各プログラムからのログ出力数の概要

時刻情報表示

二種類の情報提示方法を採用

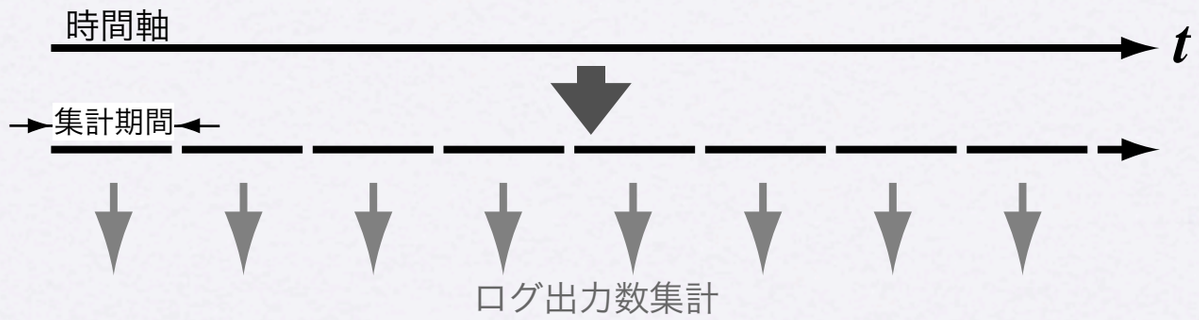
- 全時間帯における出力傾向

  記録された全時間帯を定期期間ごとに区切り、
  各期間ごとの出力数を提示

- 周期期間毎の出力傾向

  曜日と毎時毎にログの出力数を集計して提示

全時間帯における出力傾向



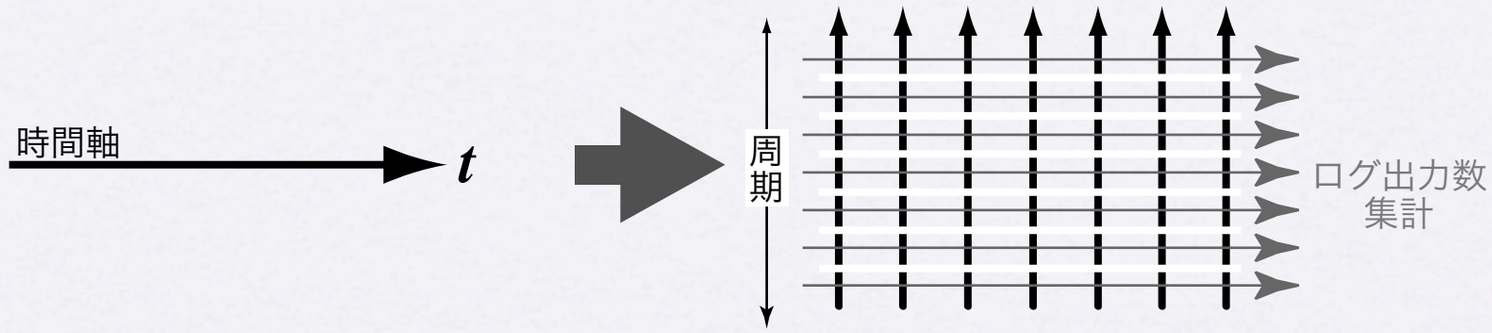
GLF形式のログ

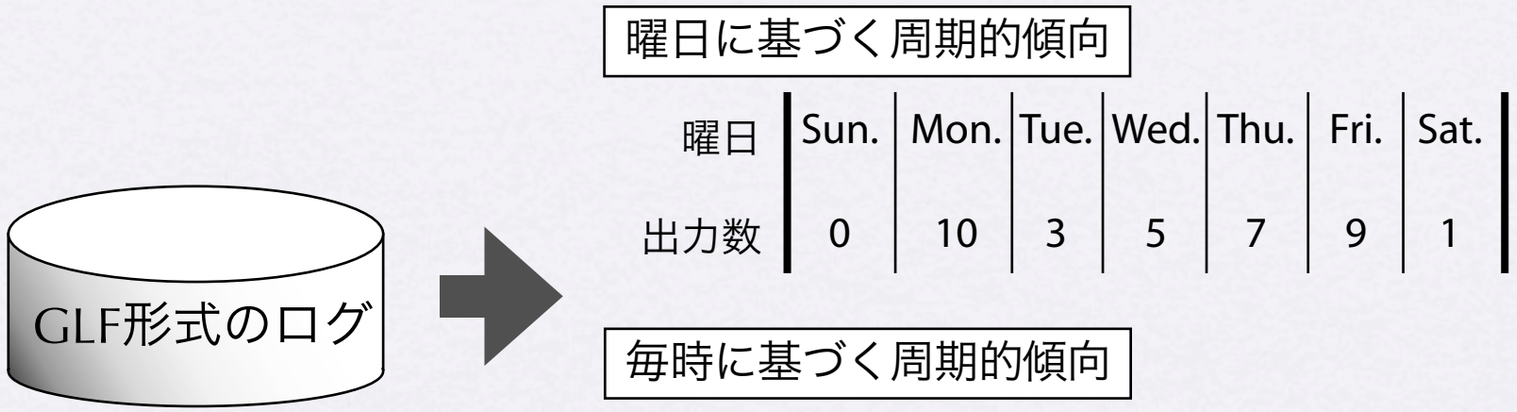
集計期間	ログ出力数
Oct 25 8時台	5
Oct 25 9時台	8
Oct 25 10	20
Oct 25 11	33
Oct 25 12	60
Oct 25 13	31
Oct 25 14	20
Oct 25 15	25
...	...

この例での集計期間は1時間

一定期間ごとのログ出力数を集計

周期的特徴による集約





周期における単位期間毎のログ出力数を集計

時刻情報表示領域

周期的特徴の表示

曜日別

月	9
火	10
水	2
木	3
金	12
土	1
日	0

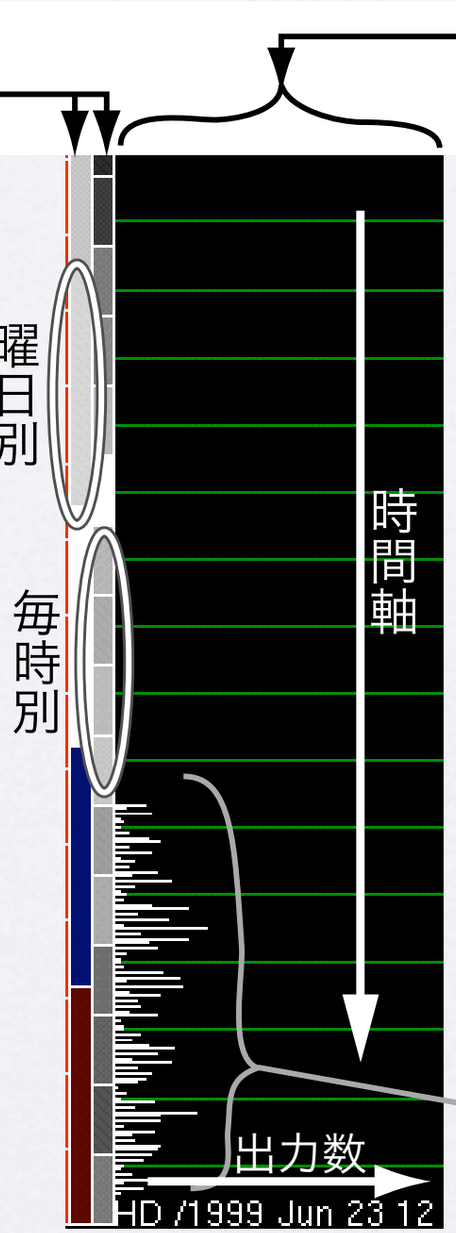
毎時別

0	0
1	3
2	5
•	•
•	•
•	•
19	9
20	10
21	18
22	12
23	5

各時間帯の出現数を格子状に図化

格子の色はログ出力数を表す

最大が白、最小が黒



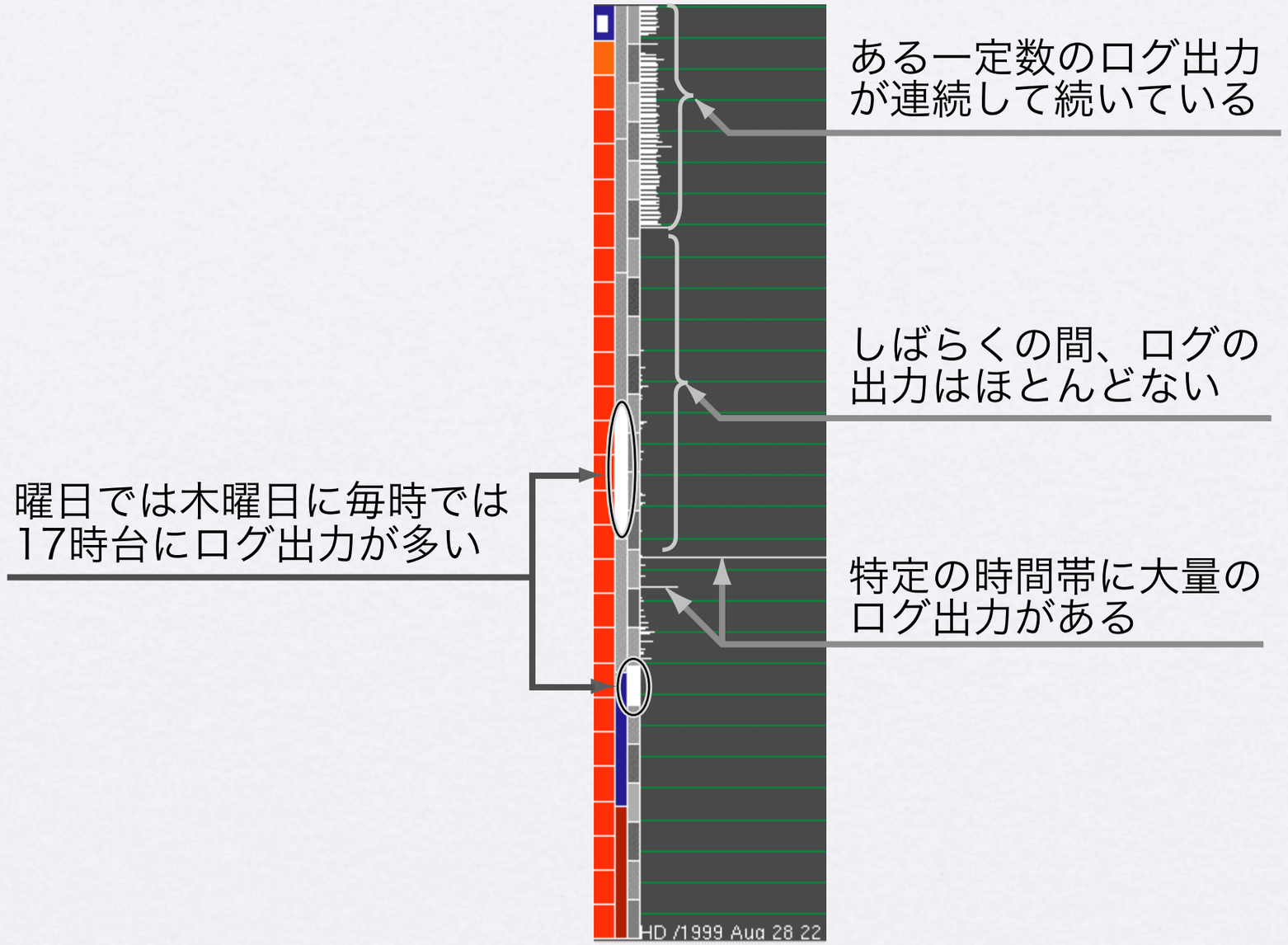
出力傾向の表示

時刻(1時間毎)	出力数
Mar 1 10	2
Mar 1 11	3
Mar 1 12	1
Mar 1 13	4
Mar 1 14	1
Mar 1 15	2
Mar 1 16	1
Mar 1 17	1
Mar 1 18	1

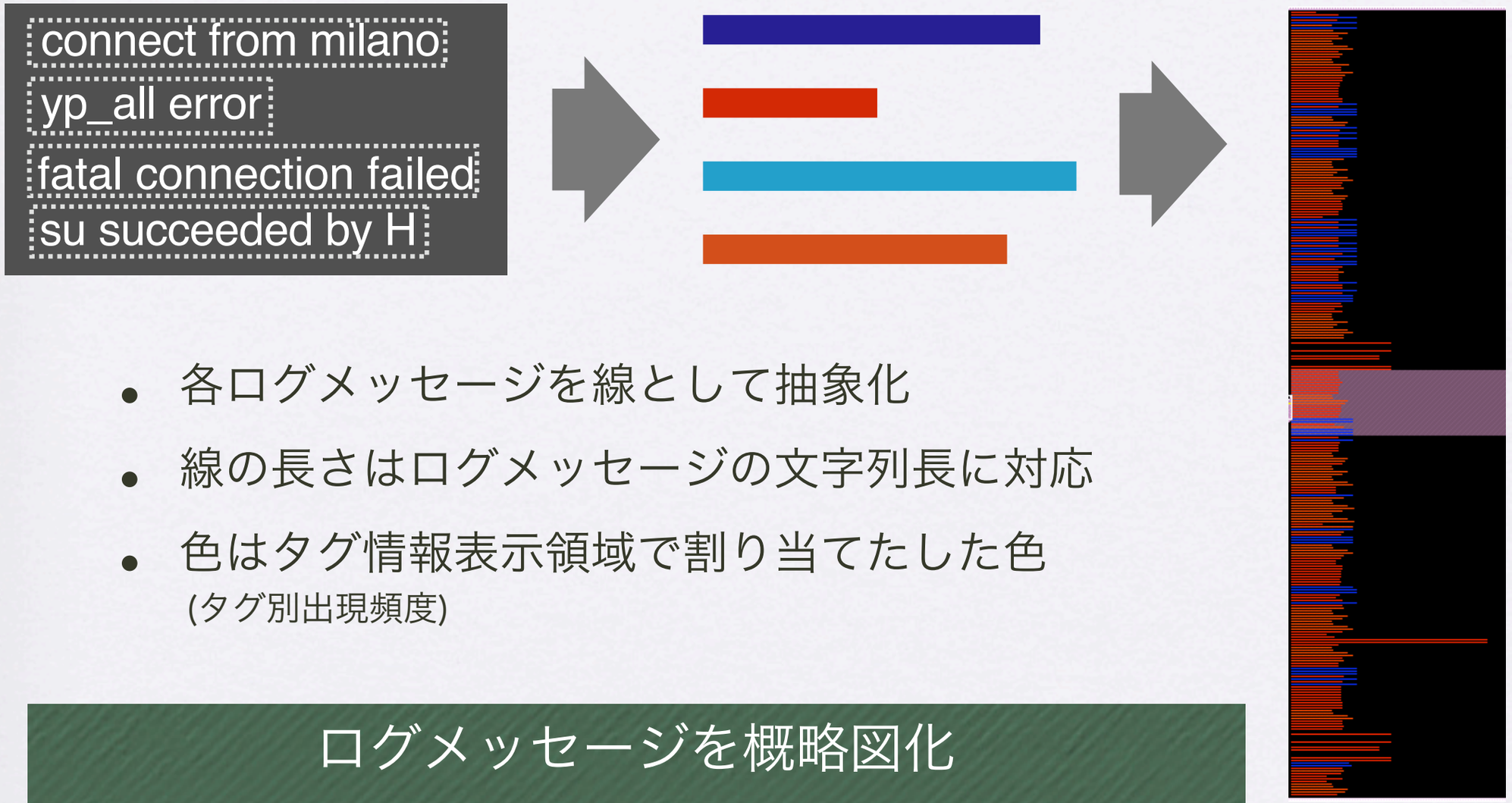
出力傾向はヒストグラムとして表示

調査事例1

時間情報の視覚化に注目した調査例



Outline表示領域



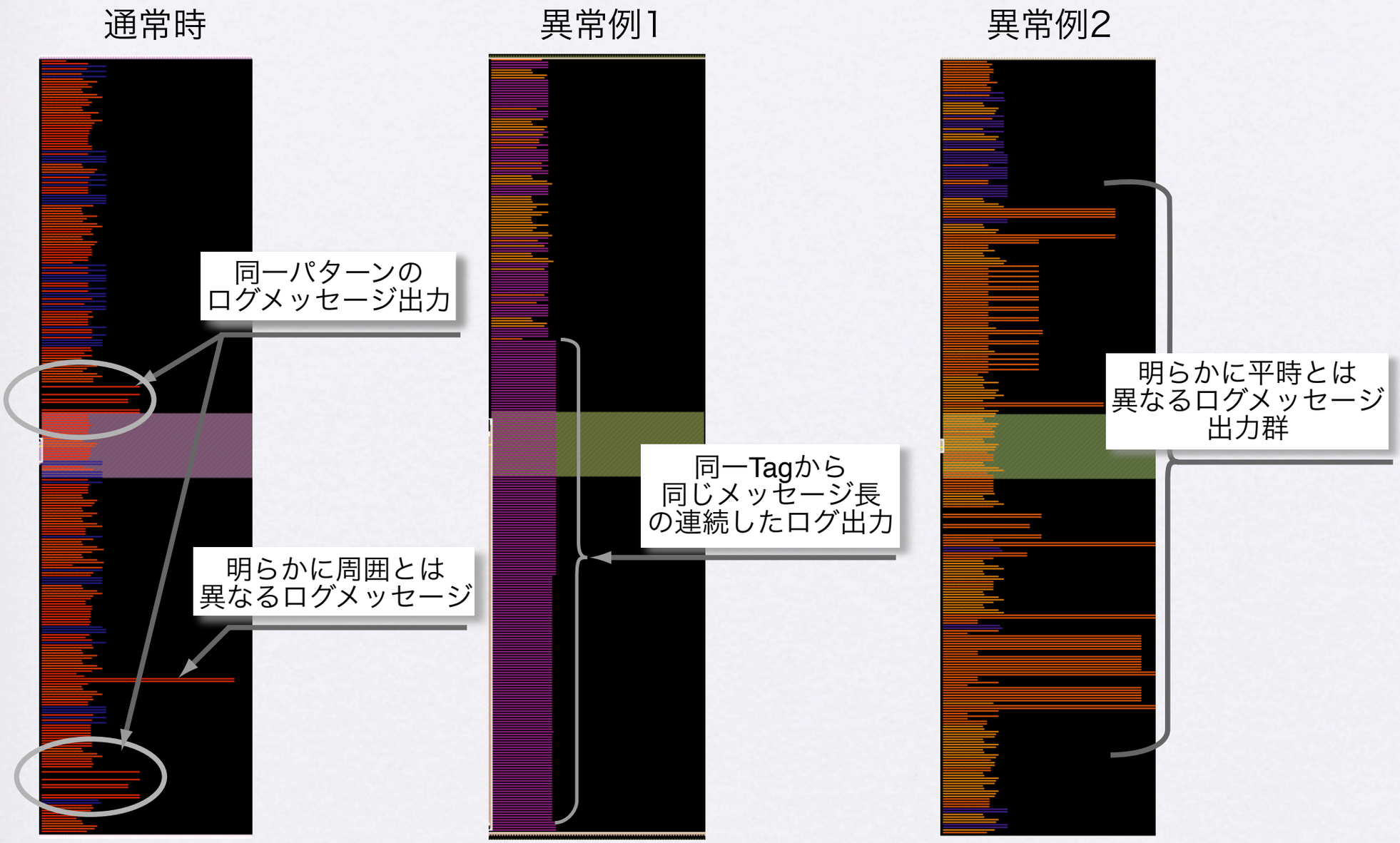
- 各ログメッセージを線として抽象化
- 線の長さはログメッセージの文字列長に対応
- 色はタグ情報表示領域で割り当てたした色
  (タグ別出現頻度)

ログメッセージを概略図化

広域の概要把握を容易に/パターン認識が可能に

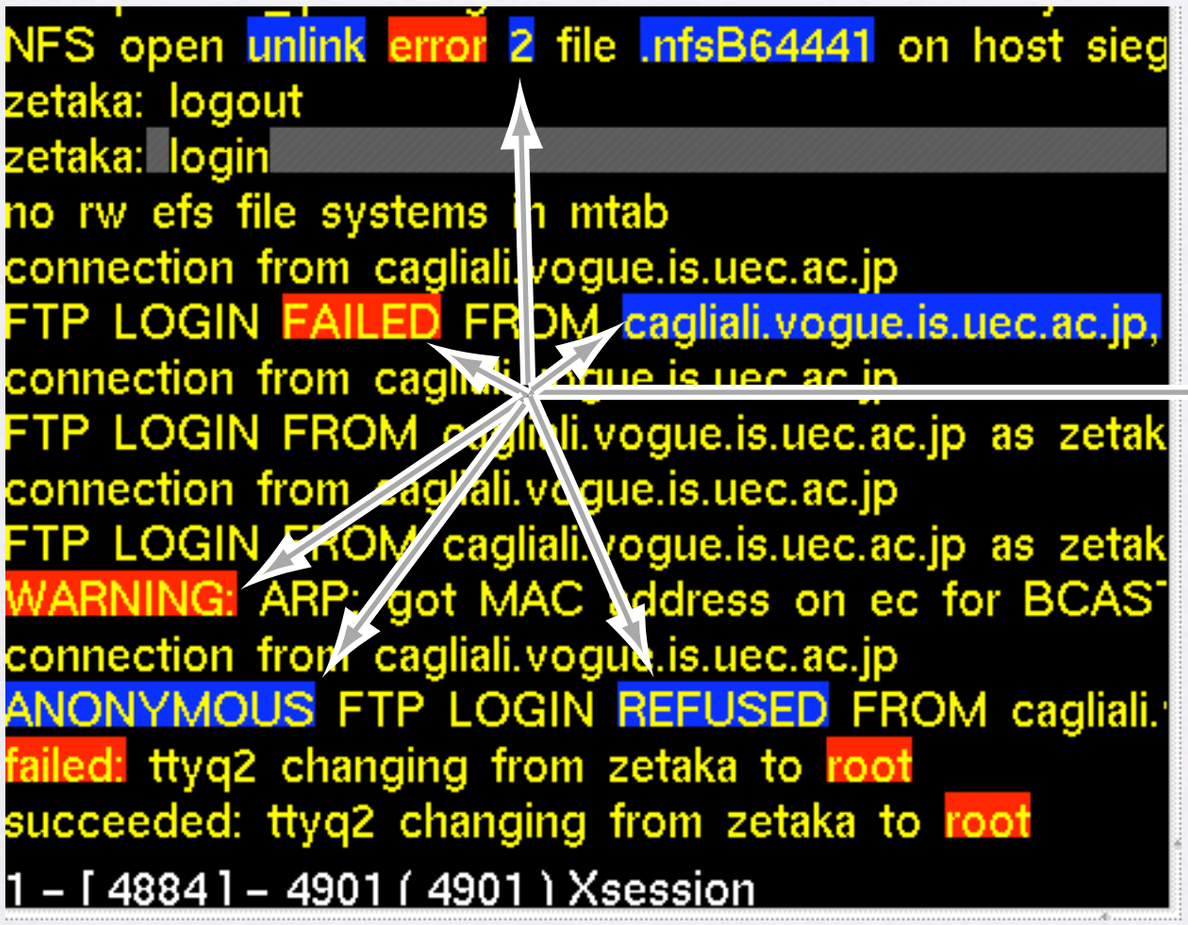
調査事例2

アウトライン表示に注目した調査例



Message表示領域

文字による表示 + 注目すべきログの提示



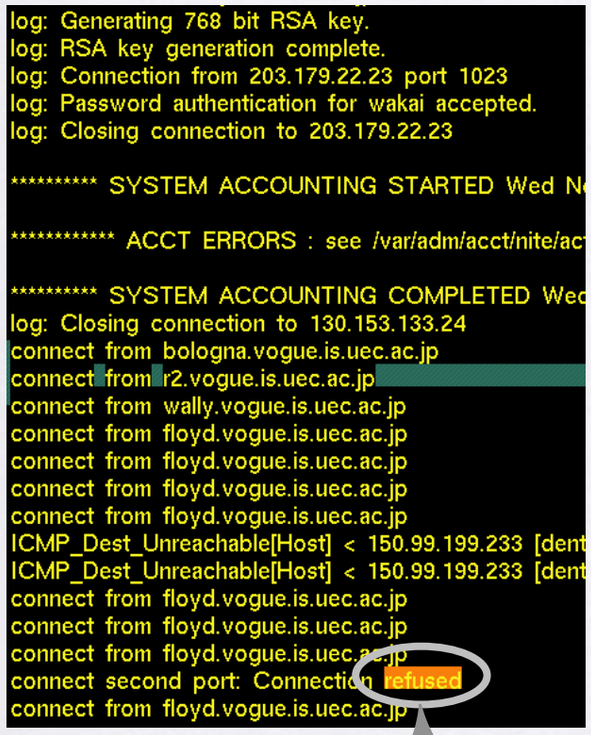
注目すべきログの
ハイライト表示

赤背景色
パターン検索によるキーワード抽出

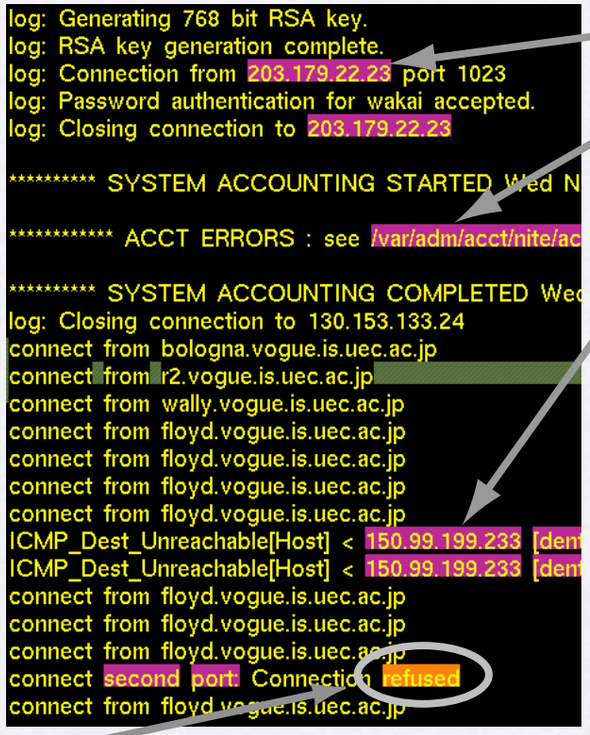
青背景色
頻度解析によるキーワード抽出

調査事例3

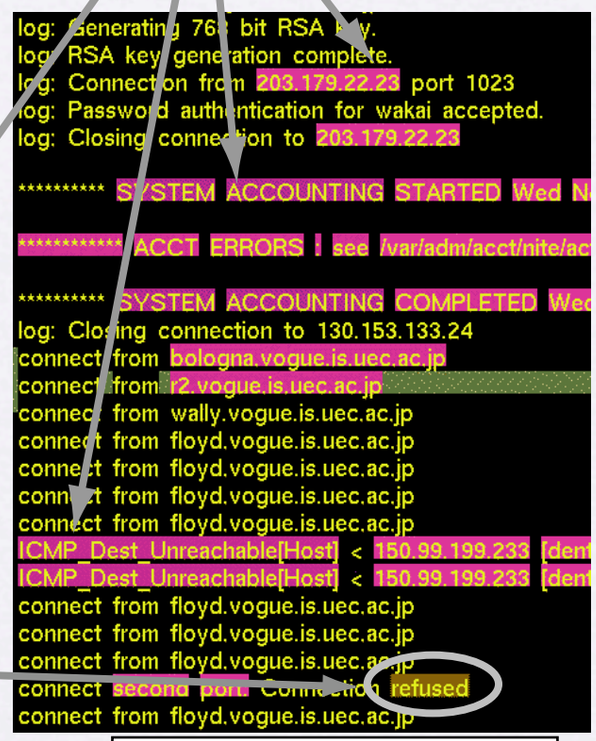
メッセージ表示領域に注目した調査例



出現頻度による抽出なし



低閾値による
出現頻度抽出を実行



中程度の閾値による
出現頻度抽出を実行

青色のハイライトは
出現頻度に基づく
低出現頻度の単語や句

赤色のハイライトは
既定のキーワード等に
一致した単語

対話機能

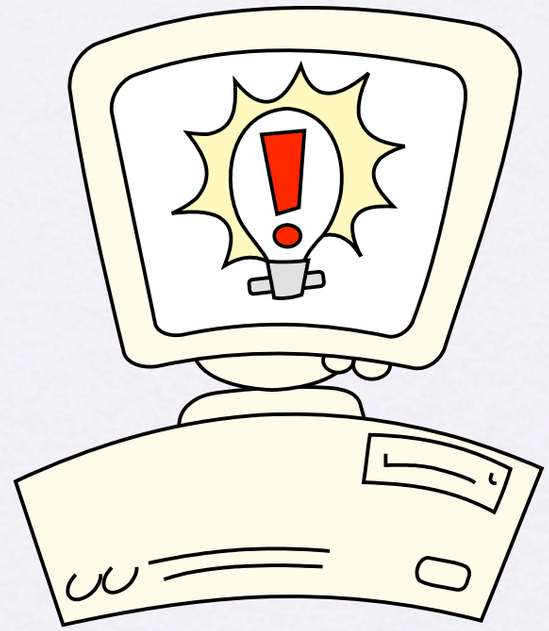
人間による調査作業を支援

- 閲覧支援
  - 要約(縮退)
- 発見支援
  - フィルタリング
  - 低出現頻度情報の抽出

対話機能(1/3)

フィルタリング機能

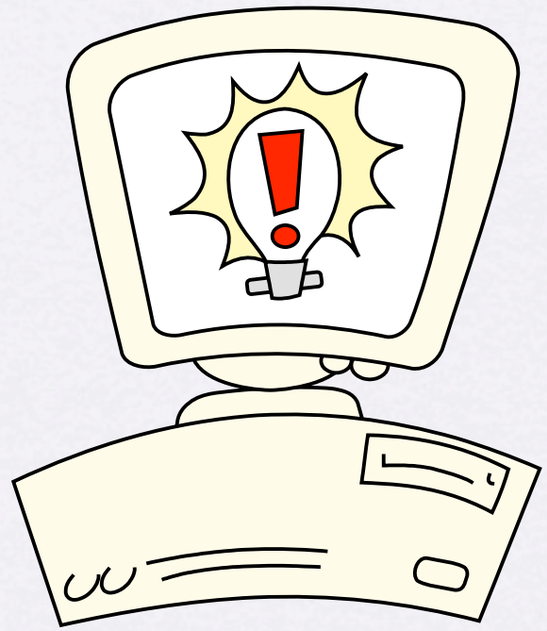
- タグ
- 時刻
- アウトライン
- 単語



demonstration

対話機能(2/3)

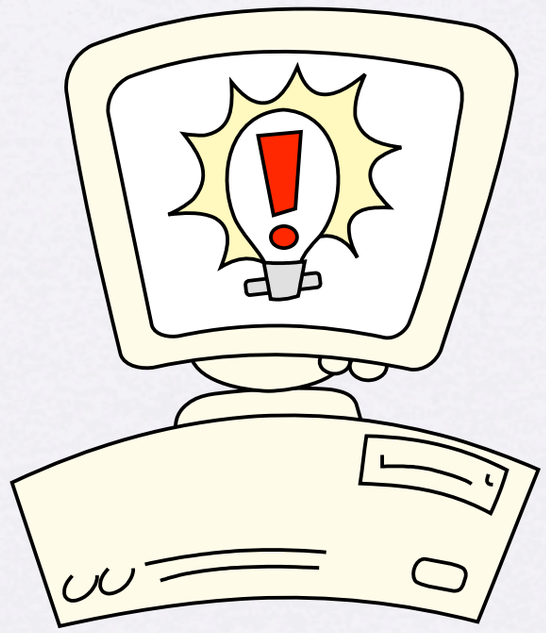
- 低出現頻度と推測されるログの取得
- タグ(アウトライン)
- 単語/句



demonstration

対話機能(3/3)

- 閲覧支援
  - ログの要約(縮退)



demonstration

利点

- ログ情報認識負荷の軽減
  - 図化によるログ情報認識の容易化
  - 概要から詳細までシームレスに認識可
- 注目すべき(疑わしき)ログの提示
  - キーワードによるパターン検索
  - 低出現頻度(単語/句)
  - 図的パターン
- 対話的調査
  - 人間による作業を支援
  - 見えログ内で作業を完結可能

今後の課題

- 低出現頻度 = 疑わしいログか?
  - 抽出された情報 = 調査すべき情報か?
  - 他の抽出方法の模索
- 作業における自由度の高さ
  - 疑わしいログ発見のための定型処理は?
- さらなる調査負担の軽減
  - “見る”作業に対する負担軽減策は?
  - 処理の自動化、多様なログへの対応
- 膨大な量のログへの対応
- 調査だけでなく監視へ

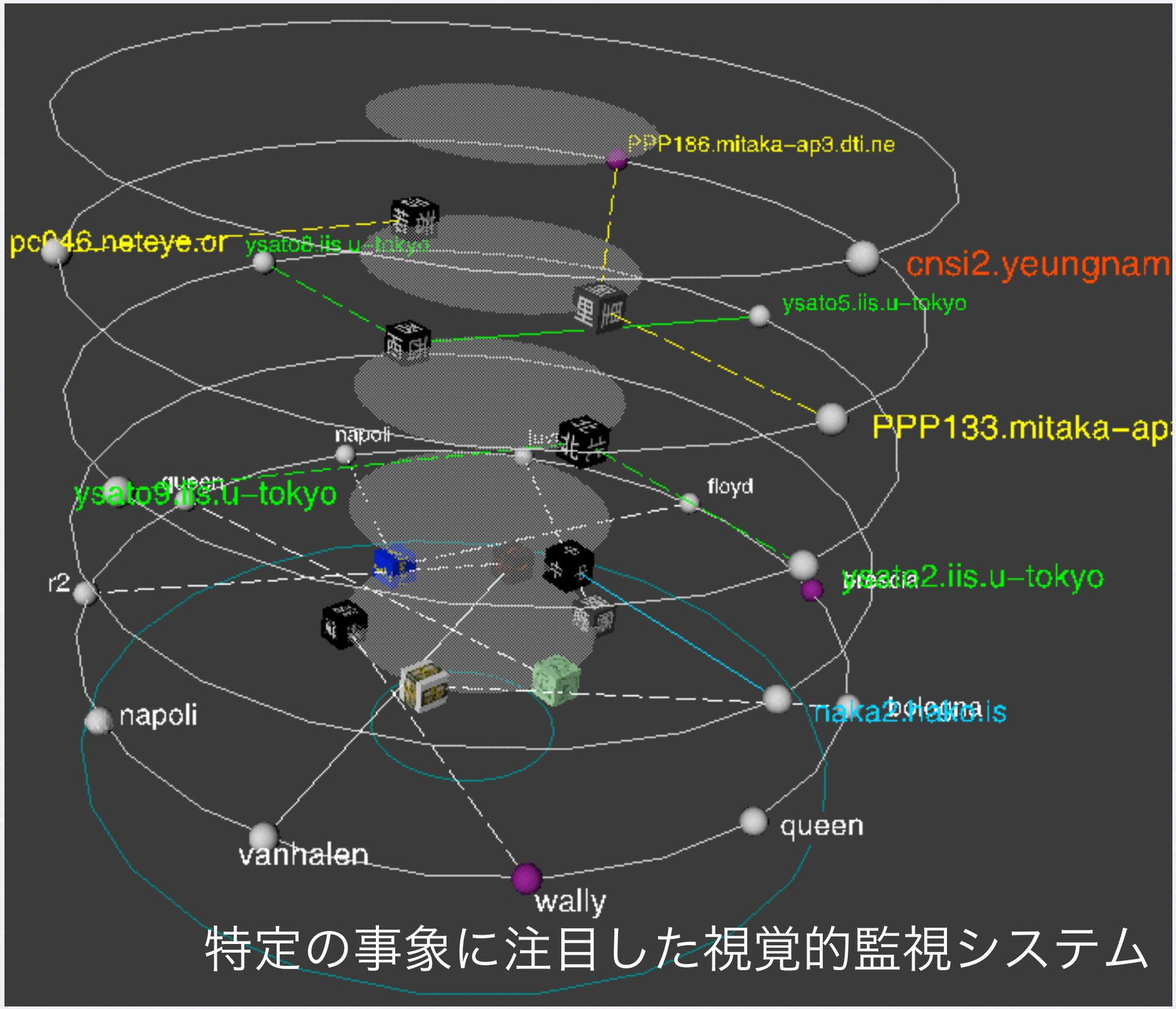
http://www.csrs.is.uec.ac.jp/ にて論文公開

不正侵入対策と視覚化システム

他の研究事例

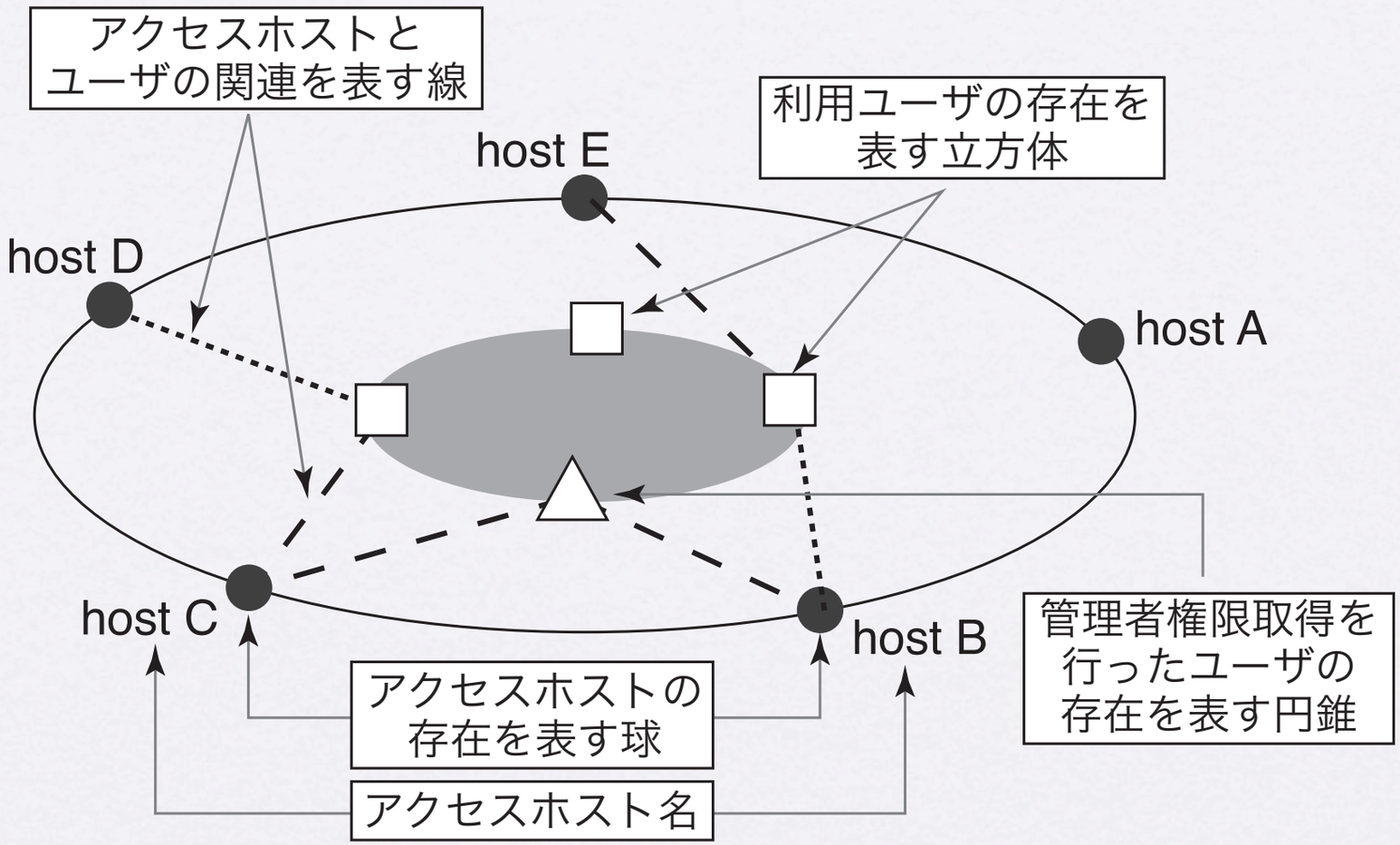
- 鼓
- ログデータ分析用ビジュアルブラウザ
- Intrusion and Misuse Detection in Large-Scale Systems

視覚化例1: 鼓



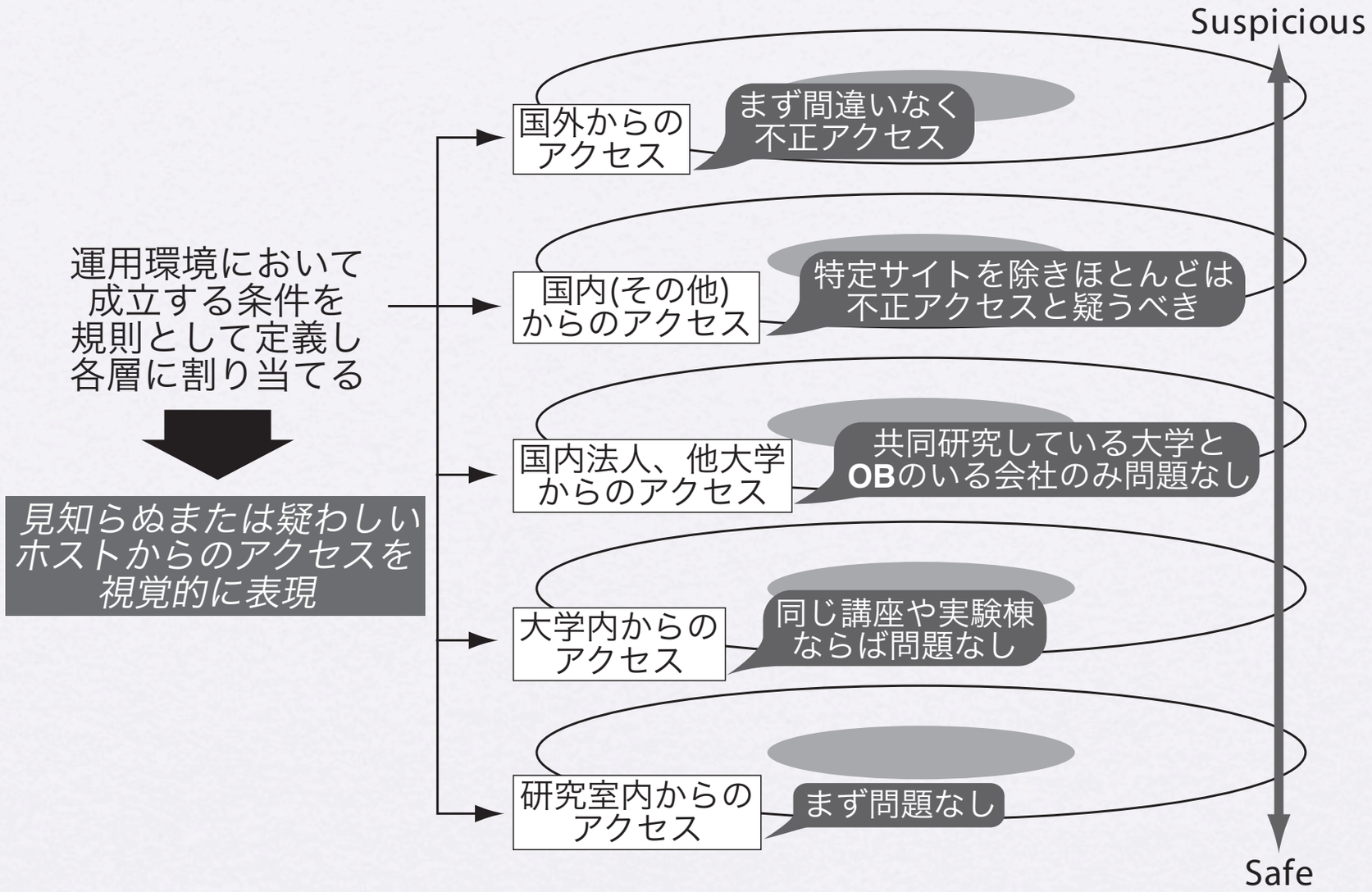
特定の事象に注目した視覚的監視システム

鼓: 視覚化手法(1)



アクセス - ログイン - 管理者権限取得の
状態遷移を視覚化

鼓: 視覚化手法(2)



層はアクセスホストを管理者のポリシに応じて分類し、視覚化

視覚化例2:

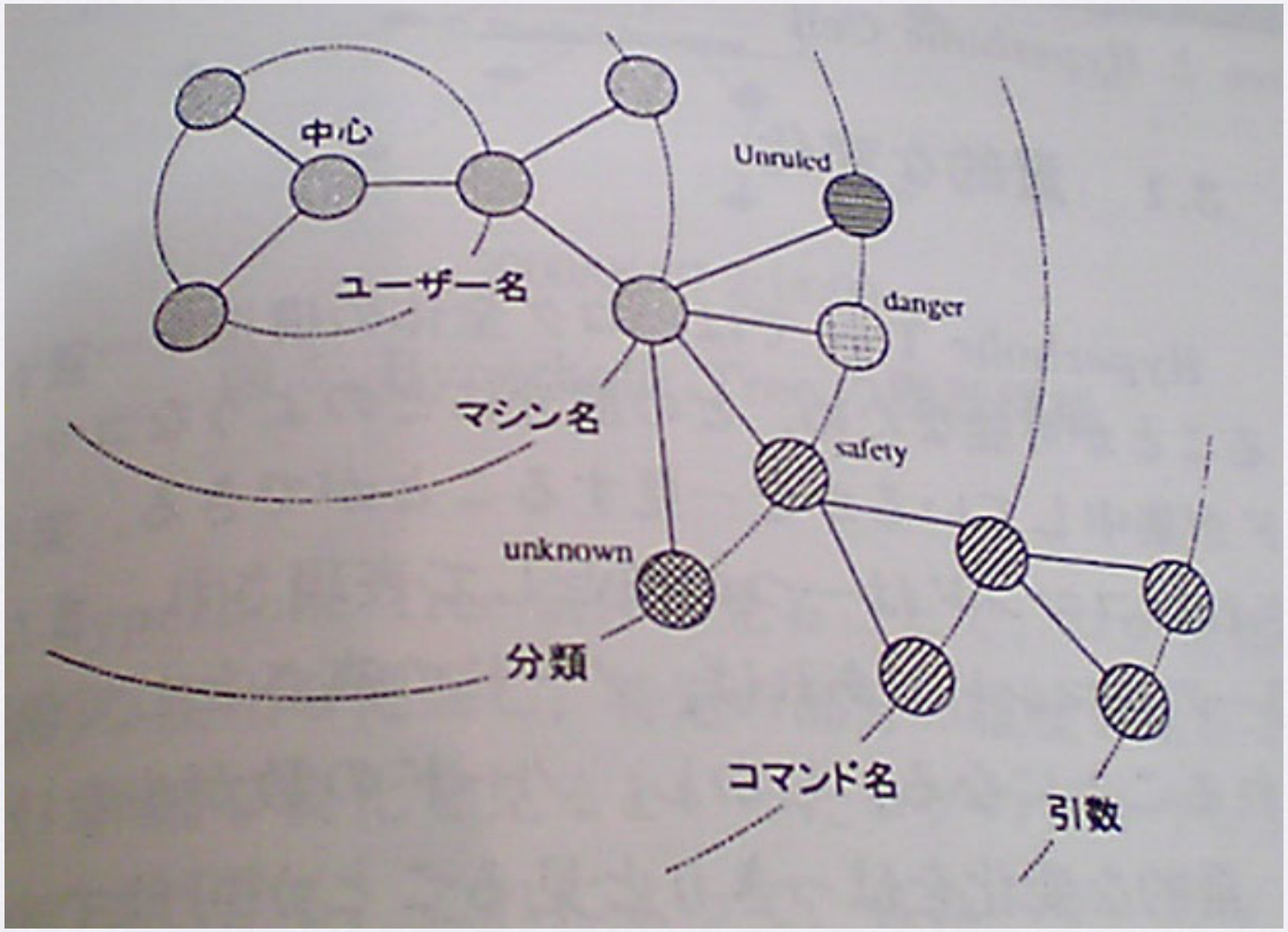
ログデータ分析用ビジュアルブラウザの設計
平石 広典、溝口 文雄
コンピュータセキュリティシンポジウム2000、
情報処理学会、pp.277--282, Oct 2000

ログデータの効率的な閲覧を可能とするログブラウザ

- 平時のユーザの利用状況の把握
- 不正侵入時の痕跡発見支援

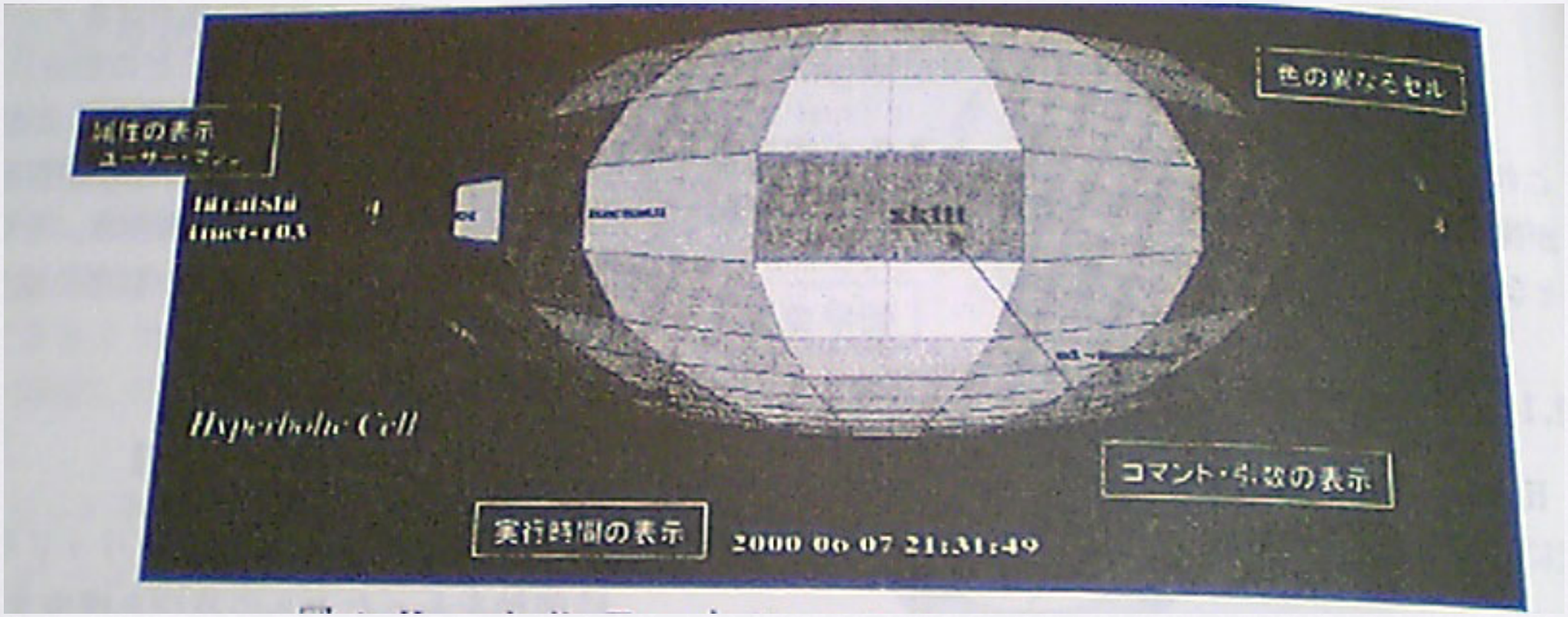
Hyperbolic Treeによる視覚化

Hyperbolic Treeによる視覚化



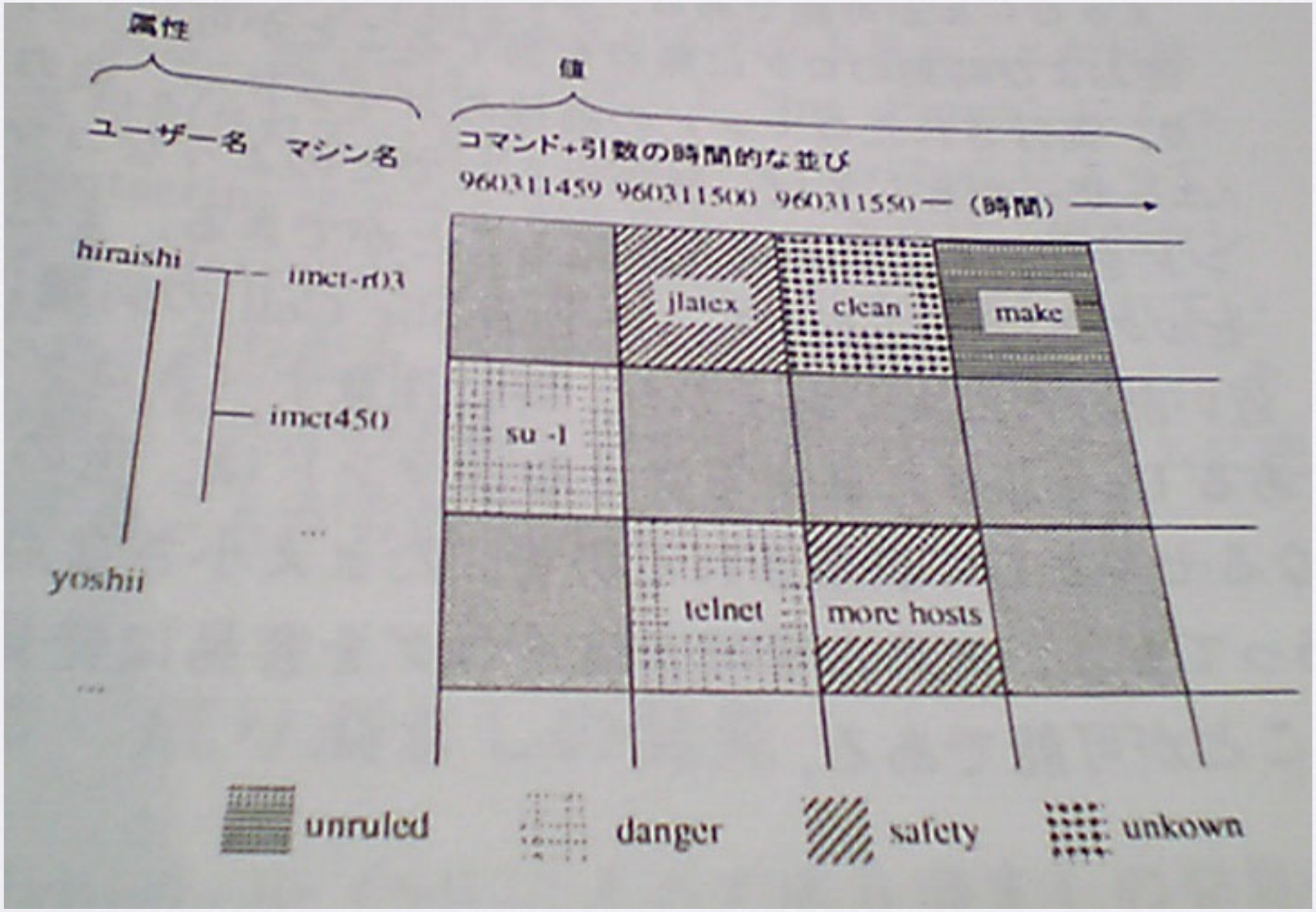
各円周上に意味付け、ノードの高さは子ノードの数

Hyperbolic Cellによる視覚化

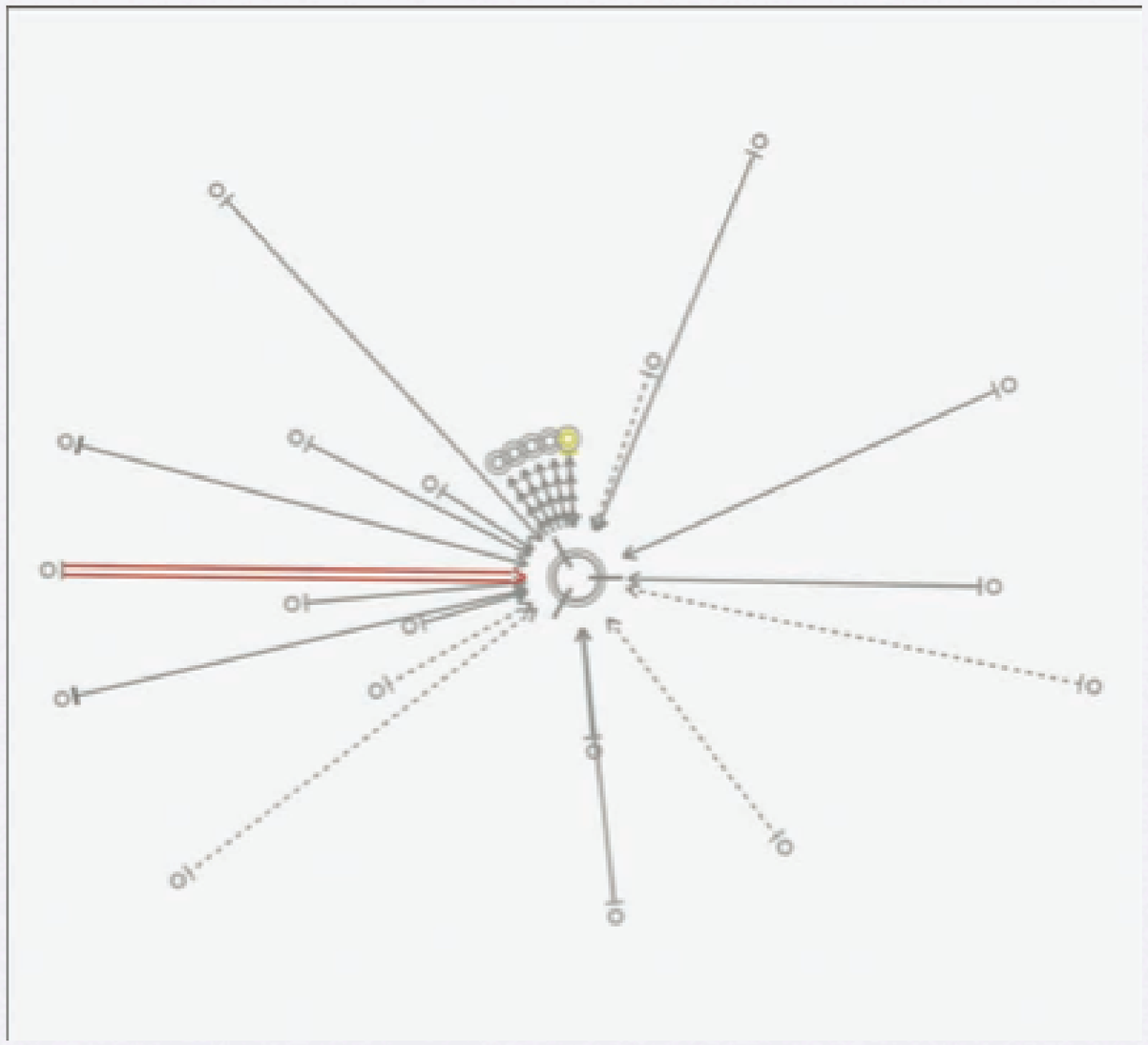


一定時間間隔ごとのコマンド実行履歴

Hyperbolic Cellによる視覚化

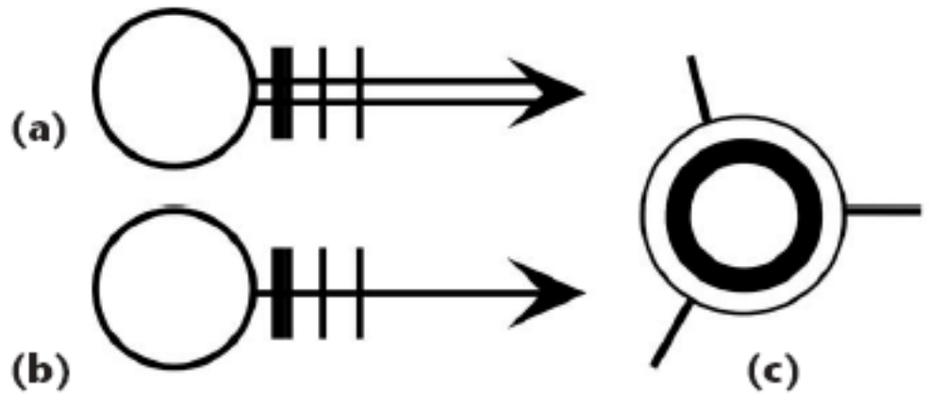


視覚化例 3:



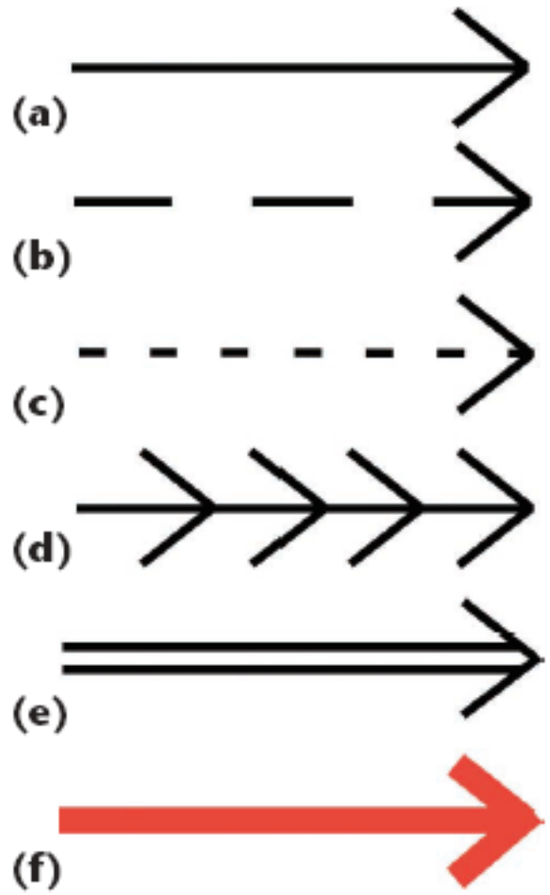
Intrusion and Misuse Detection in Large-Scale Systems
R.F.Erbacher, K.L.Walker and D.A.Frincke
IEEE CG&A, Vol.22, No.1, pp.38-48, 2002

視覚化手法(1)



2 Basic glyph organization. (a) The initial inetd connection to the system. (b) The resulting connection after authentication. (a) and (b) also represent the number of users with connections from the given remote host and the number of connections by said users through the use of the cross hatches. The monitored system, (c) showing number of users and load.

視覚化手法(2)



3 Line appearances and their relationships. (a) Telnet and rlogin connections as solid lines, (b) privileged FTPs as long dashed lines, (c) anonymous FTPs as short dashed lines, (d) Network file system (NFS) accesses as solid lines with many arrows, (e) initial inetd port connection, and (f) port scan.

ログ解析手法

- 定量分析(閾値モデル)
- 統計的手法
  - 閾値学習モデル
  - 多変量モデル
  - 状態遷移モデル
  - 確率分布モデル
- クラスター分析
- Neural Network
- Genetic Algorithm
- Data Mining

参考文献

不正侵入検知の教科書となる書籍

- ネットワーク侵入検知
  武田 圭史/磯崎 宏著、ソフトバンク
  ISBN4-7973-1253-X

- INTRUSION DETECTION

  Rebecca Gurley Bace、Macmillan Technical Publishing
  ISBN1-57870-185-6

おわり