

ELF/VLF 波動観測データからの 類似現象検索システム

笠原 禎也, 平野 晃朗, 河崎 健一郎, 高田 良宏
(金沢大学)

【背景】

科学衛星/地上装置による観測データ

－ 長期観測による観測データ量の増大

例：あけぼの衛星VLF観測データ(1989～2008)：約20TB

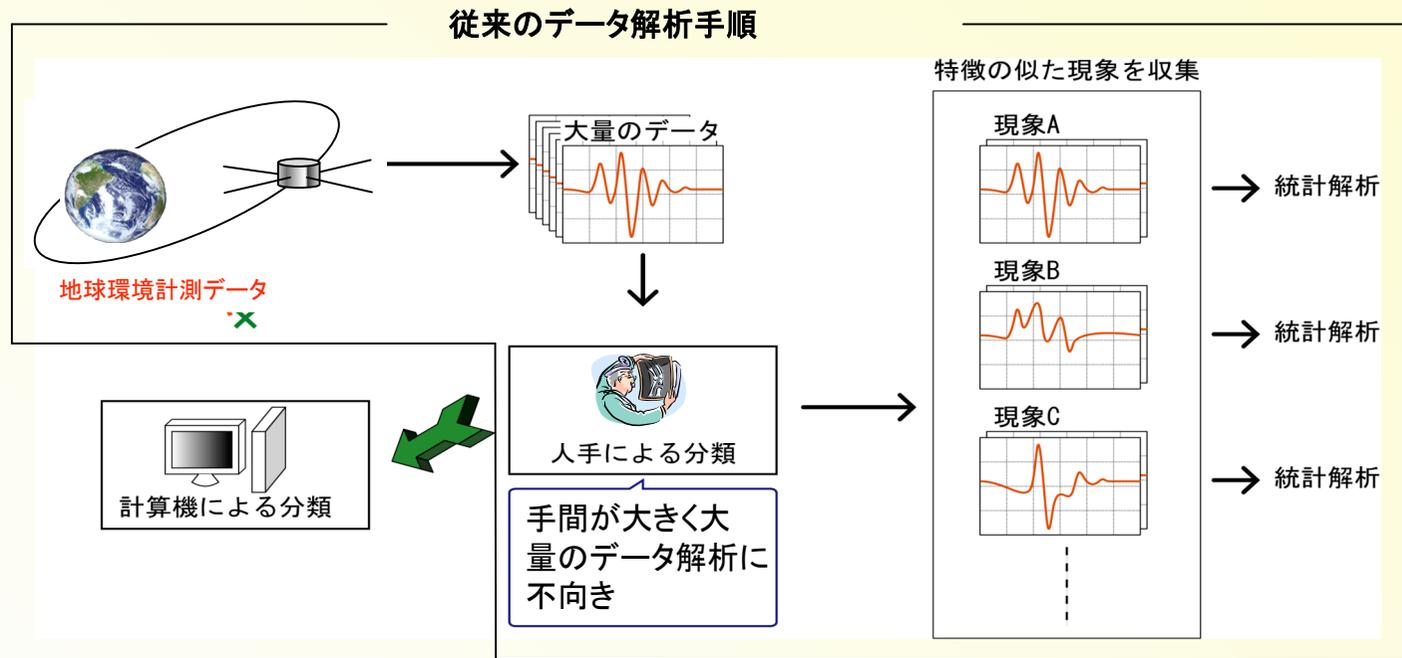
－ 観測性能の向上によるデータ生成量の増大

衛星	データ種類	データ転送量
あけぼの (1989年)	衛星全体	240MB / day
	うちVLF/MCA	2MB / day
かぐや (2007年)	衛星全体	10GB / day
	うちLRS/WFC	1GB / day

⇒大量データ中から興味ある現象の抽出には、計算機を用いたデータ処理が必須

観測データ中から、興味ある現象とそれに類似する現象を高速に、高精度に検索が行えるシステムの開発

【大規模データベースからの発見的情報の自動抽出】

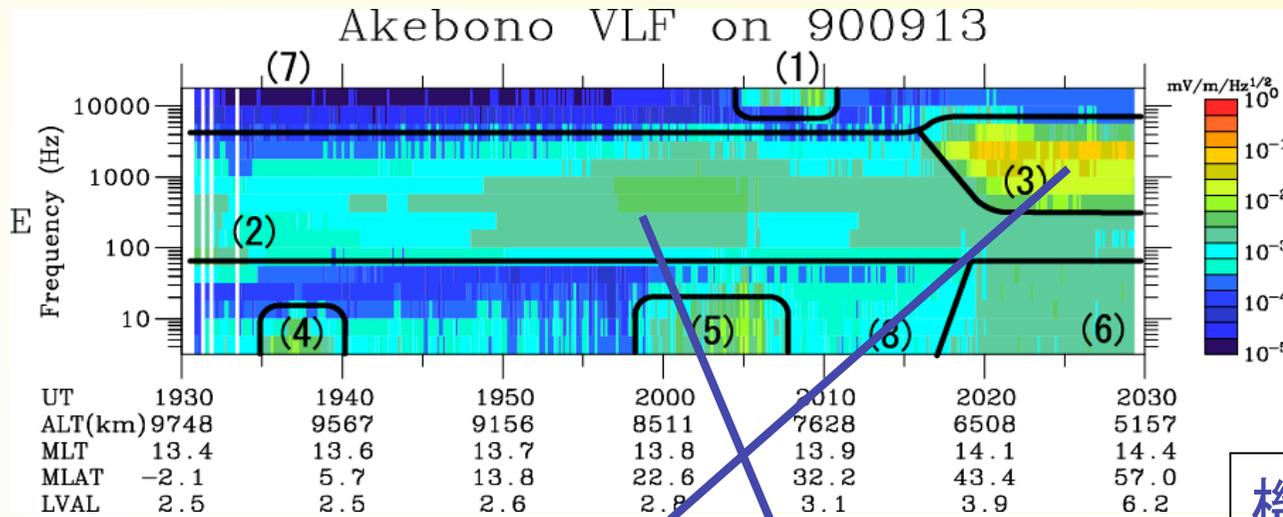


Tbyteオーダーのデータをすべて(専門知識を有する)人間がサーベイし、解析することは事実上不可能！

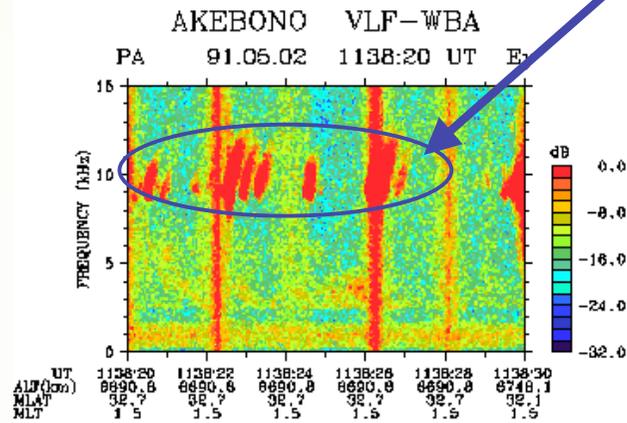
⇒ データを専門家と同じ抽象的概念に基づき、コンピュータが自動判別・分類するしくみが必要

蓄えるだけでは成果につながらない！

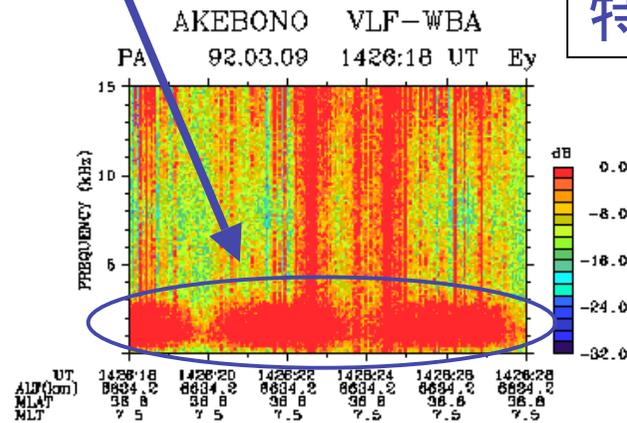
【データの特徴の自動認識・分類】



機械的にデータの特徴を識別したい！



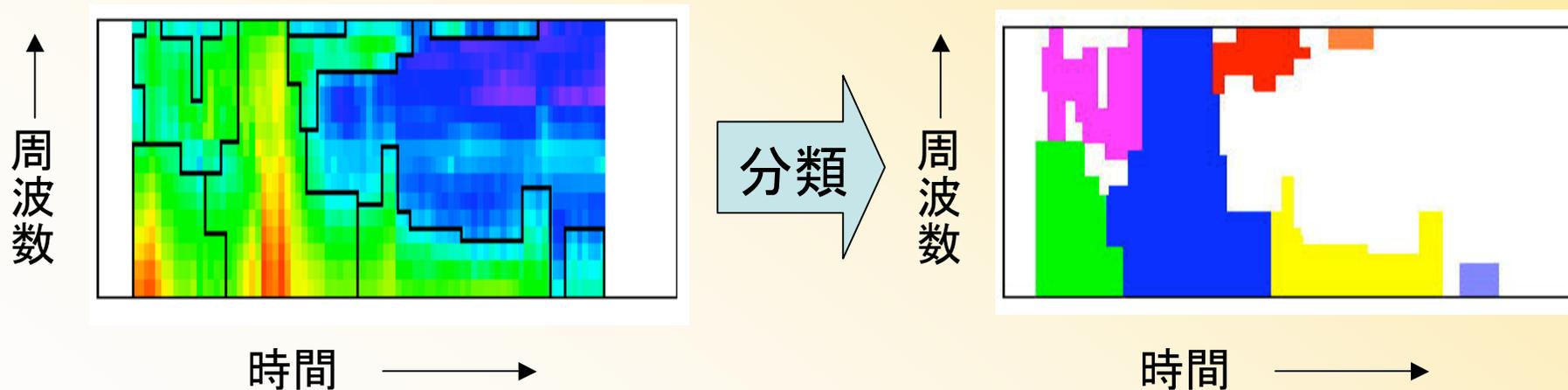
コーラス



プラズマ圏ヒス

【Event Finder の開発(従来法)】

- 研究者が用いる評価基準の定量化・正規化
- データの種類によらない汎用データ識別アルゴリズム
- ブラウザによる分類結果の検索・表示システム
- 興味深いデータの抽出アルゴリズム



【主な特徴量】

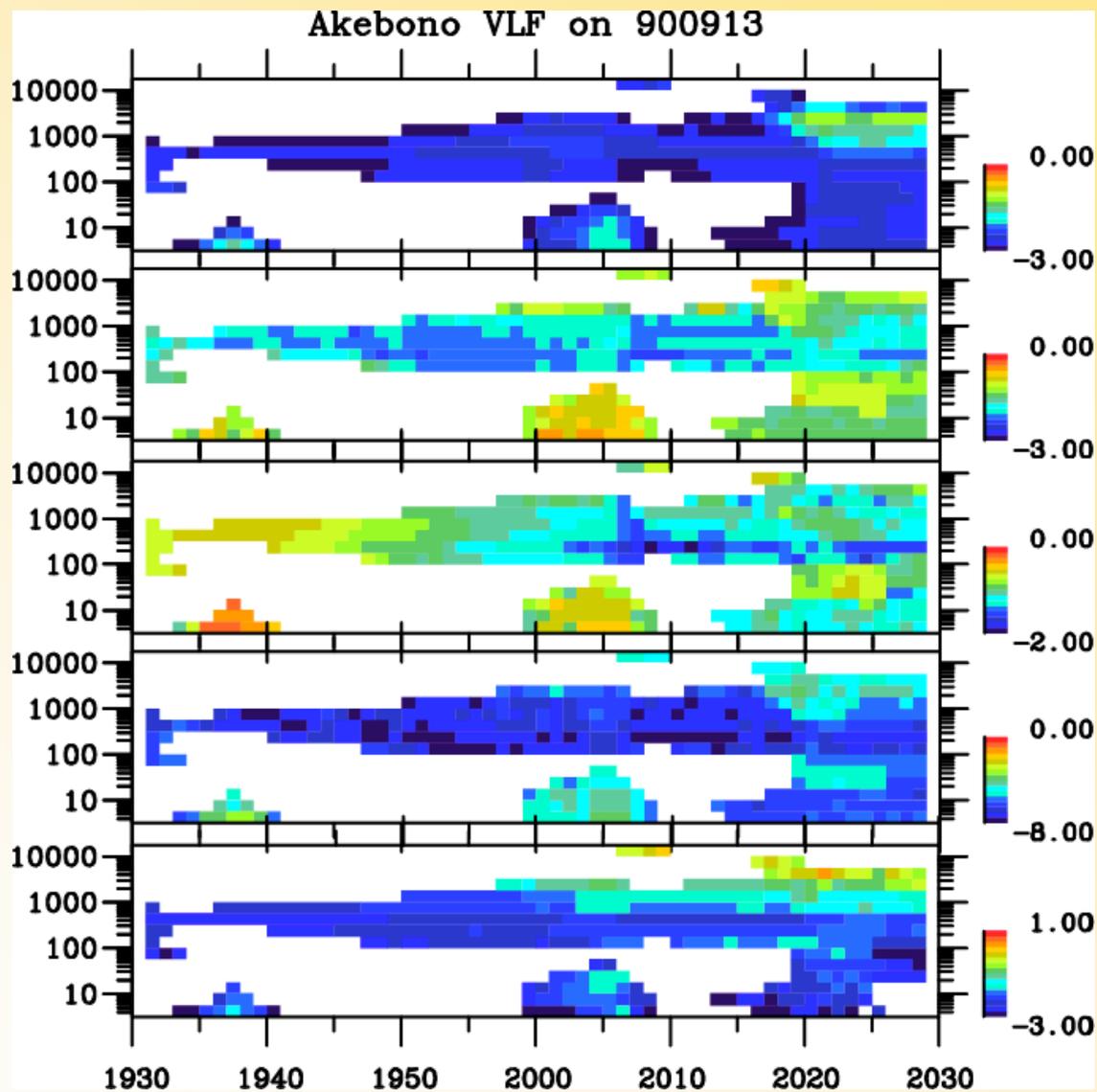
平均電界強度

電界強度の時間変動

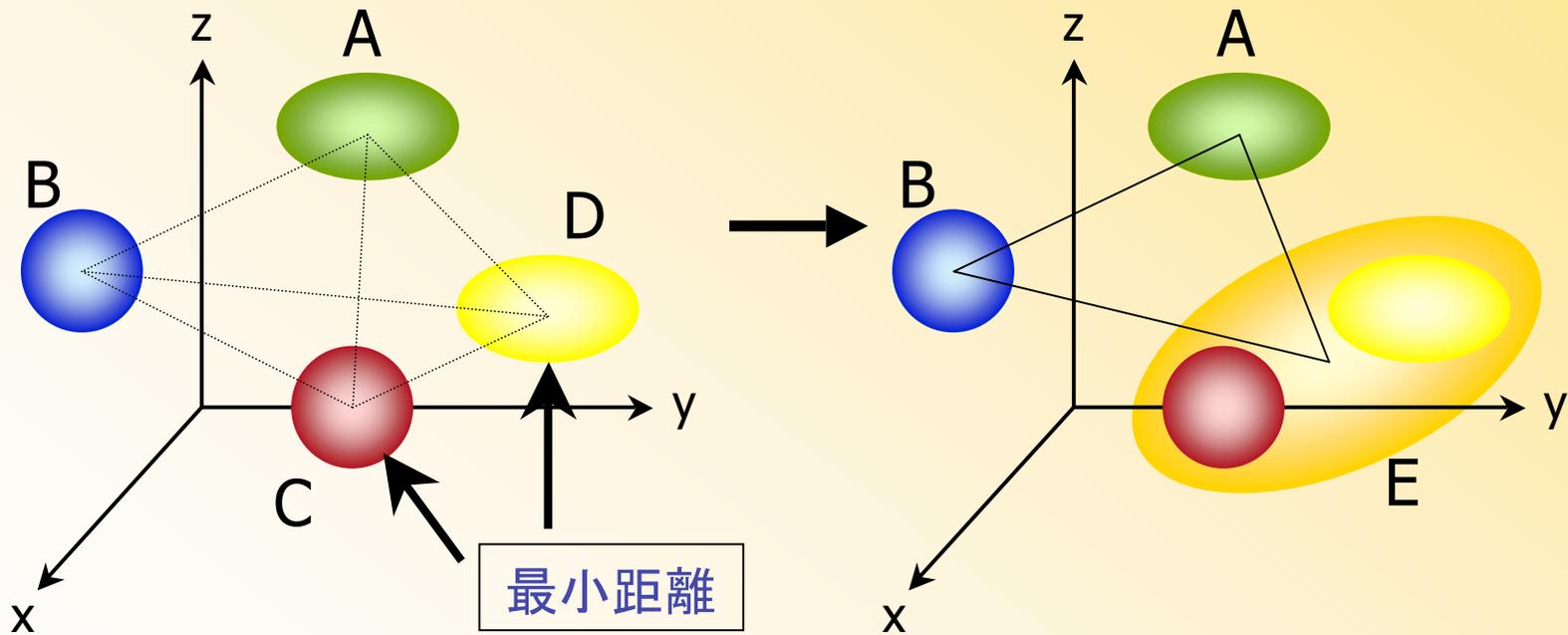
衛星スピン変動成分

電力分散

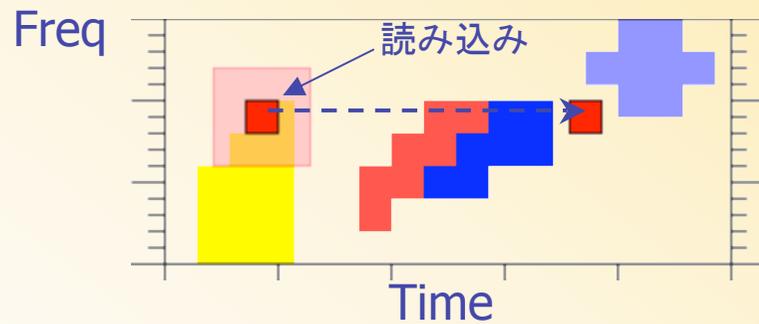
電磁界比



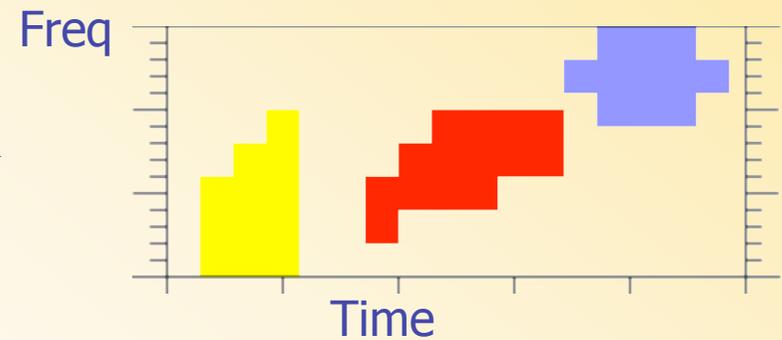
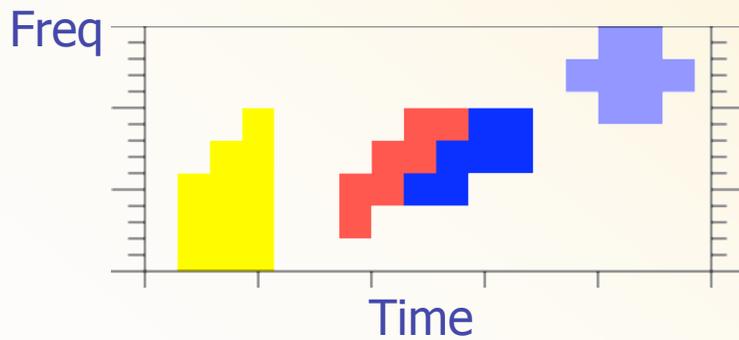
【クラスタ分析】



【クラスタリング処理】

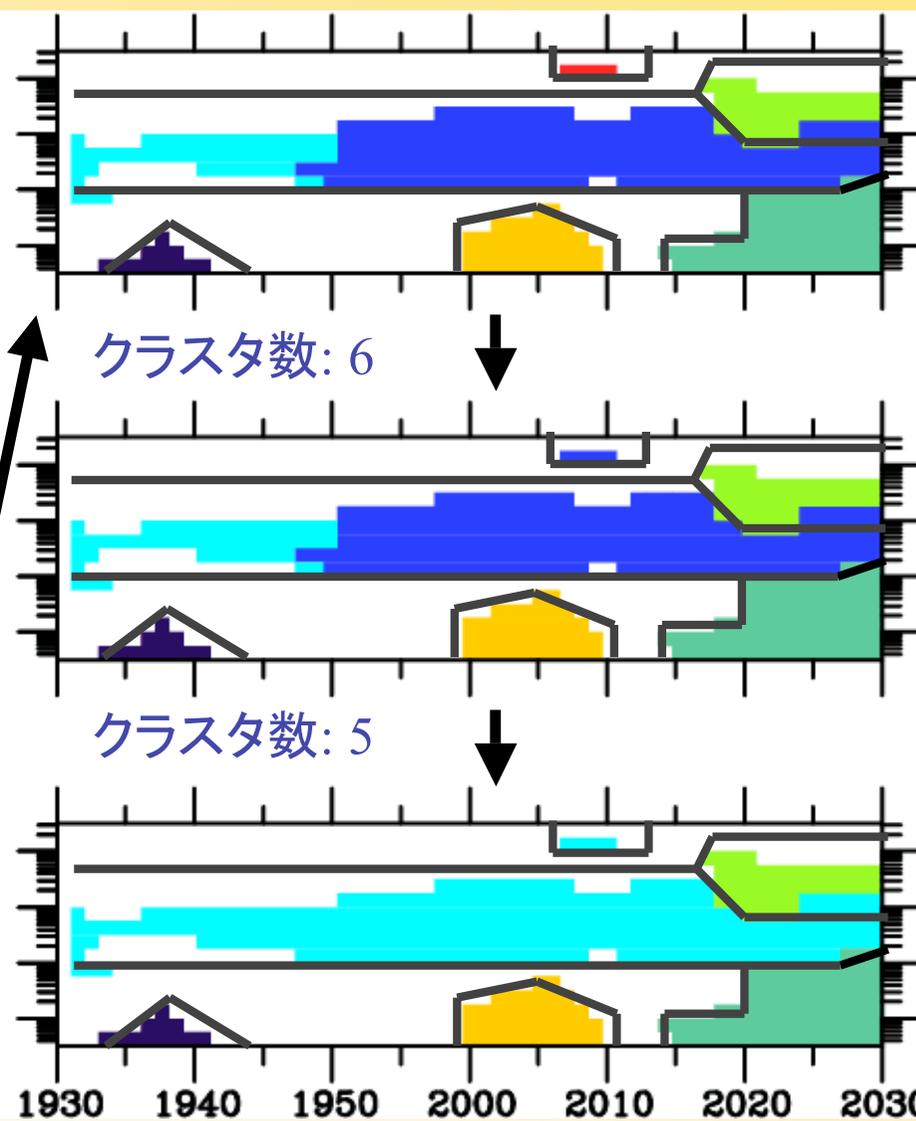
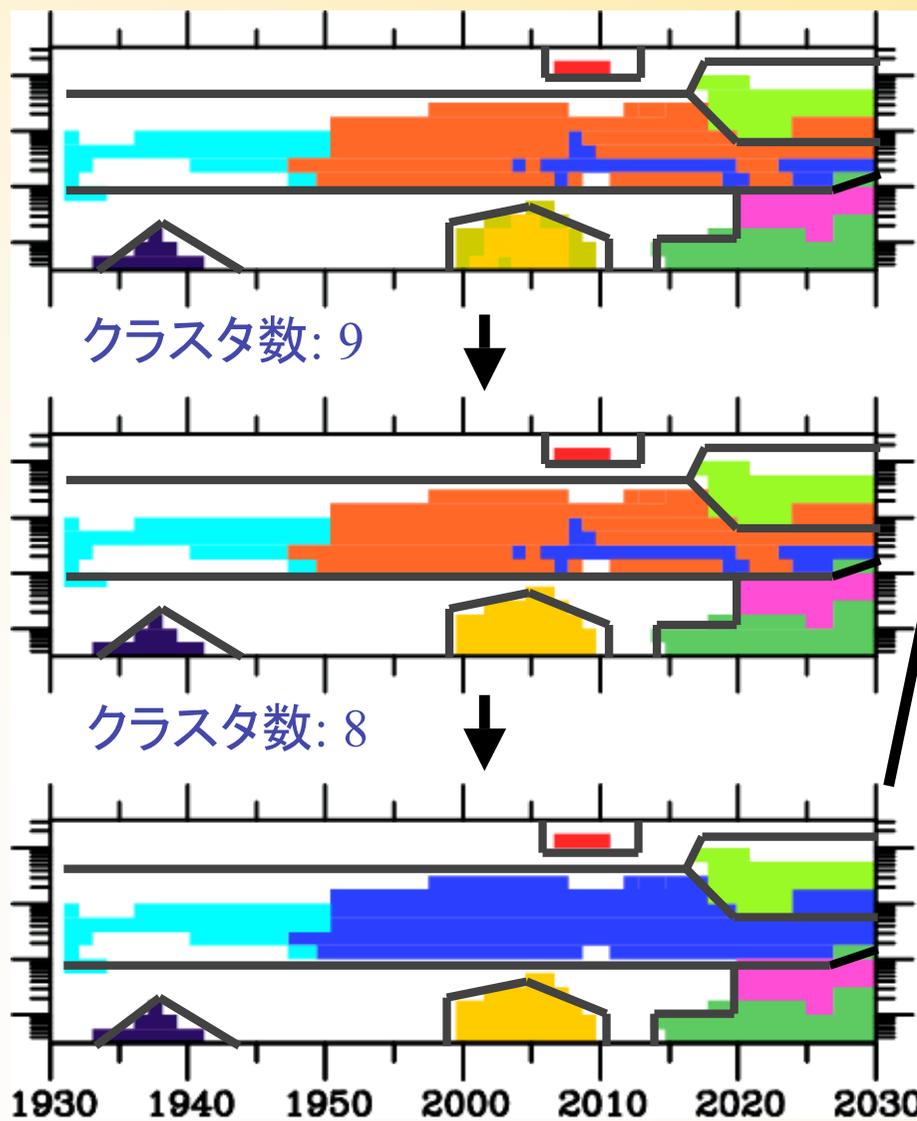


設定クラス数になるまで繰り返し



クラスタ数:10

クラスタ数:7



19:30 -- 20:00 UT September 13, 1990

【検索結果表示(従来法)】

検索条件

MCA Class Search

Beginning: Year: 1989 / Month: 3 / Date: 9 / Hour: 0 : 0
End: 1999 / Month: 12 / Date: 31 / Hour: 23 : 59

Class: class1 class2 class3 class4 class5
 class6 class7 class8 class9 class10

clear search

Results

Class1 1990-09-13 11:50:24+09
Class1 1990-09-13 11:58:56+09
Class1 1990-09-13 12:08:32+09
Class1 1990-09-13 12:16:00+09
Class1 1990-09-13 12:22:24+09
Class1 1990-09-13 12:25:36+09
Class1 1990-09-13 12:36:16+09
Class1 1990-09-13 12:40:32+09
Class2 1990-09-13 11:54:40+09
Class2 1990-09-13 11:56:48+09
Class2 1990-09-13 12:00:00+09
Class2 1990-09-13 12:03:12+09
Class3 1990-09-13 12:36:16+09
Class3 1990-09-13 12:40:32+09
Class4 1990-09-13 11:48:16+09
Class4 1990-09-13 12:11:44+09
Class5 1990-09-13 11:48:16+09
Class6 1990-09-13 12:21:20+09
Class6 1990-09-13 12:24:32+09
Class6 1990-09-13 12:39:28+09
Class7 1990-09-13 11:48:16+09
Class7 1990-09-13 12:18:08+09
Class7 1990-09-13 12:24:32+09

Matches 23

1990-09-13 11:48:00.0
MCA Spectrum

Akebono VLF on 1990/09/13

UT 11:48 12:04 12:20 12:36 12:52 13:08 13:24 13:40
ALT 6167 7950 9121 9693 9675 9068 7862 6047
MLT 10.3 11.7 12.4 12.9 13.3 13.7 14.2 14.9

Classes

1990/09/13-11:48

11:48 12:02 12:16 12:30 12:44 12:59 13:13 13:27 13:41 13:55

検索結果

波動スペクトル

分類結果

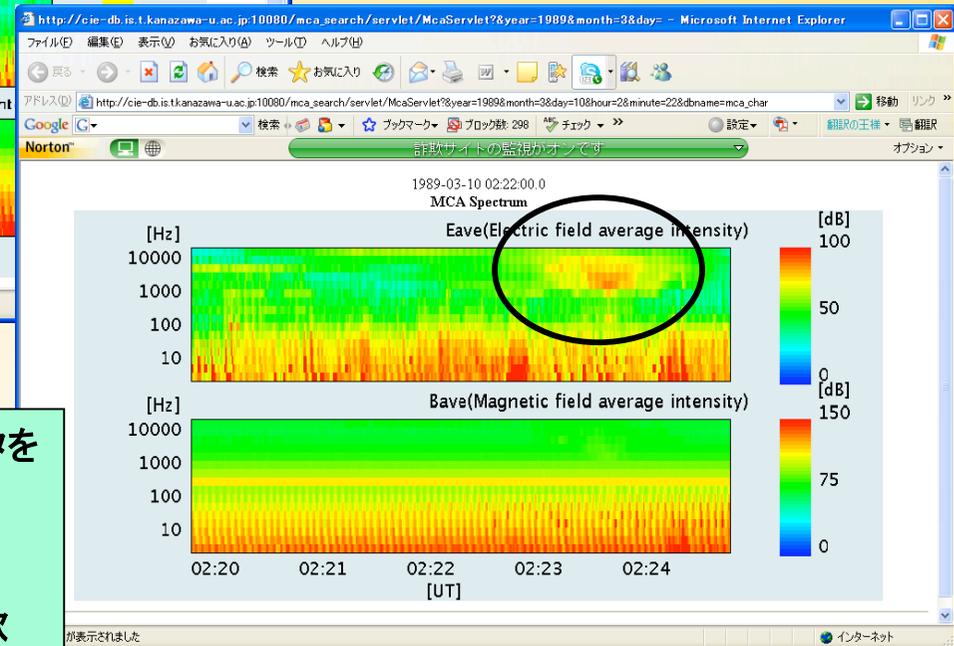
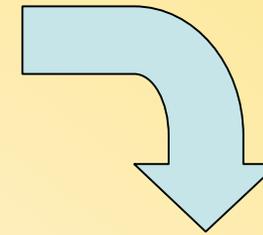
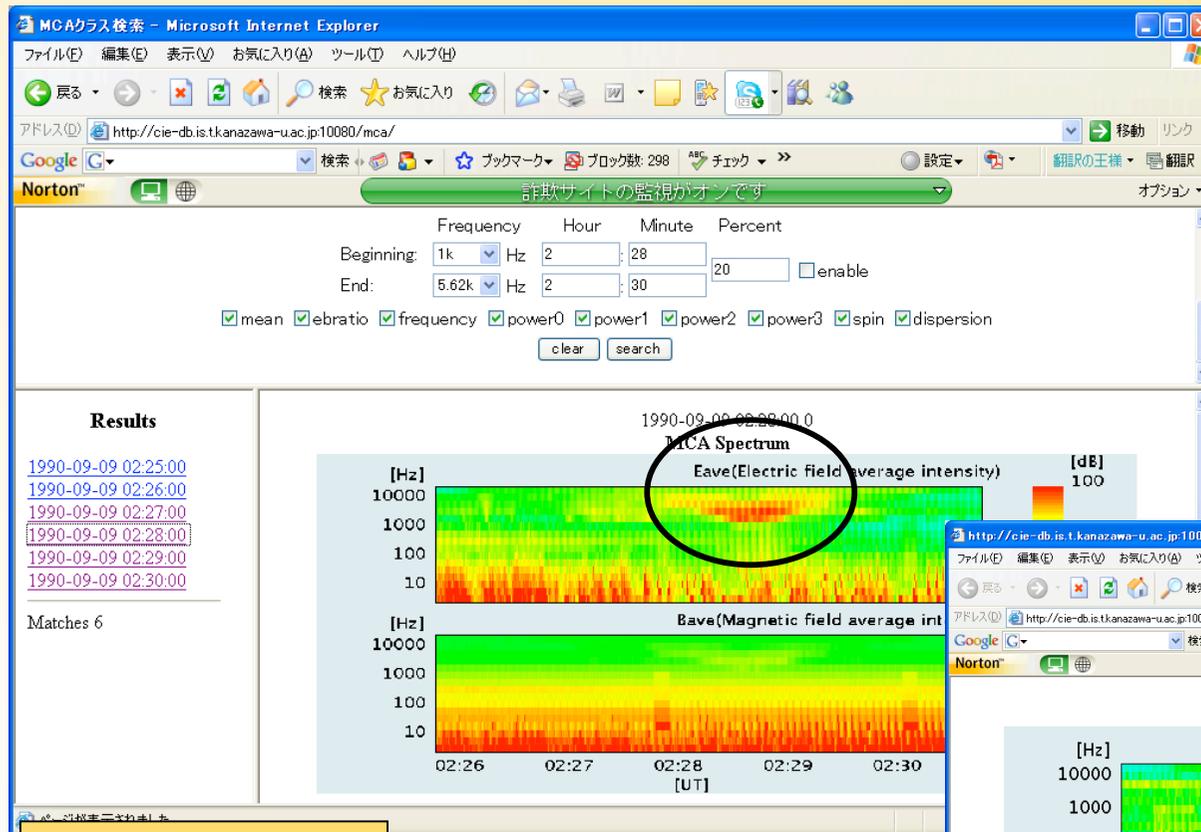
【従来法の問題点】

- 分類結果のあいまいさ
(特に複数現象が同時観測される場合など)
- 希少な現象に対する誤分類
(高頻度の現象は比較的精度よく分類可)

?? 本来我々がやりたいのは何か??

→ 既知の現象を数多く集めるのではなく
見つけたものと同類のデータを集めること!

類似データ検索システム



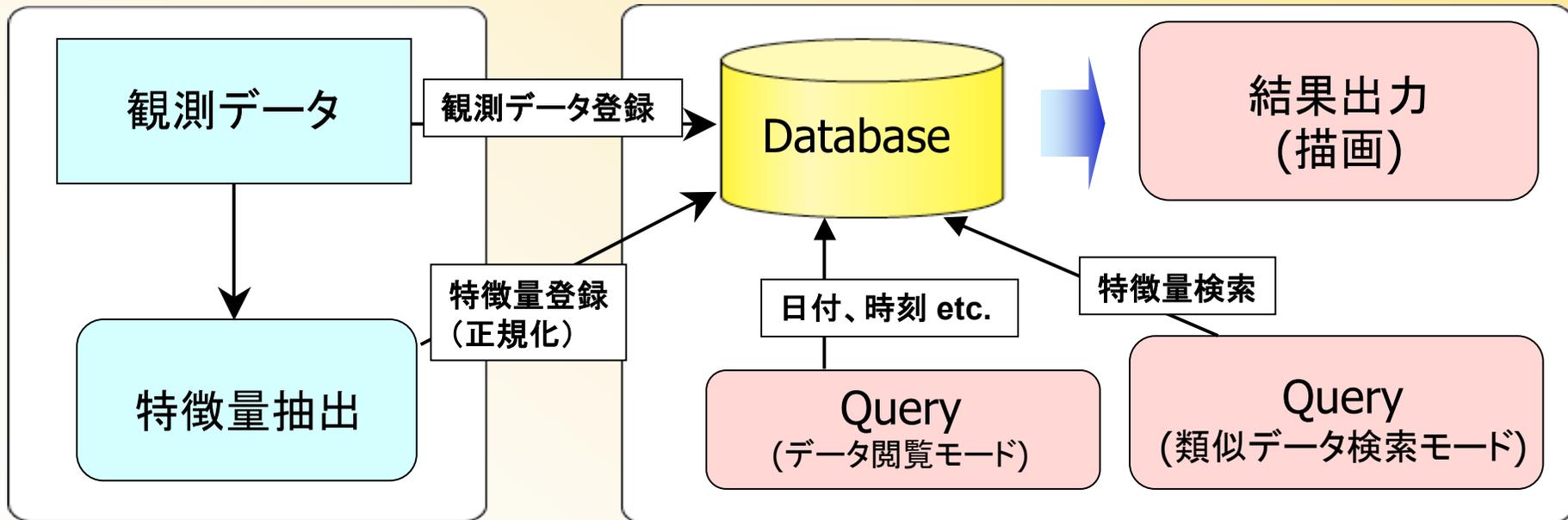
研究開発諸元

- データの意味・特徴を客観的指標で表現し、大量データを計算機の手で分類・体系化したデータベースの実現
(Automatic indexing)
- 体系化したデータベースから、あいまいな検索語を柔軟に解釈し、特徴的な未知・発見的データを検索・抽出
(Event finder system)

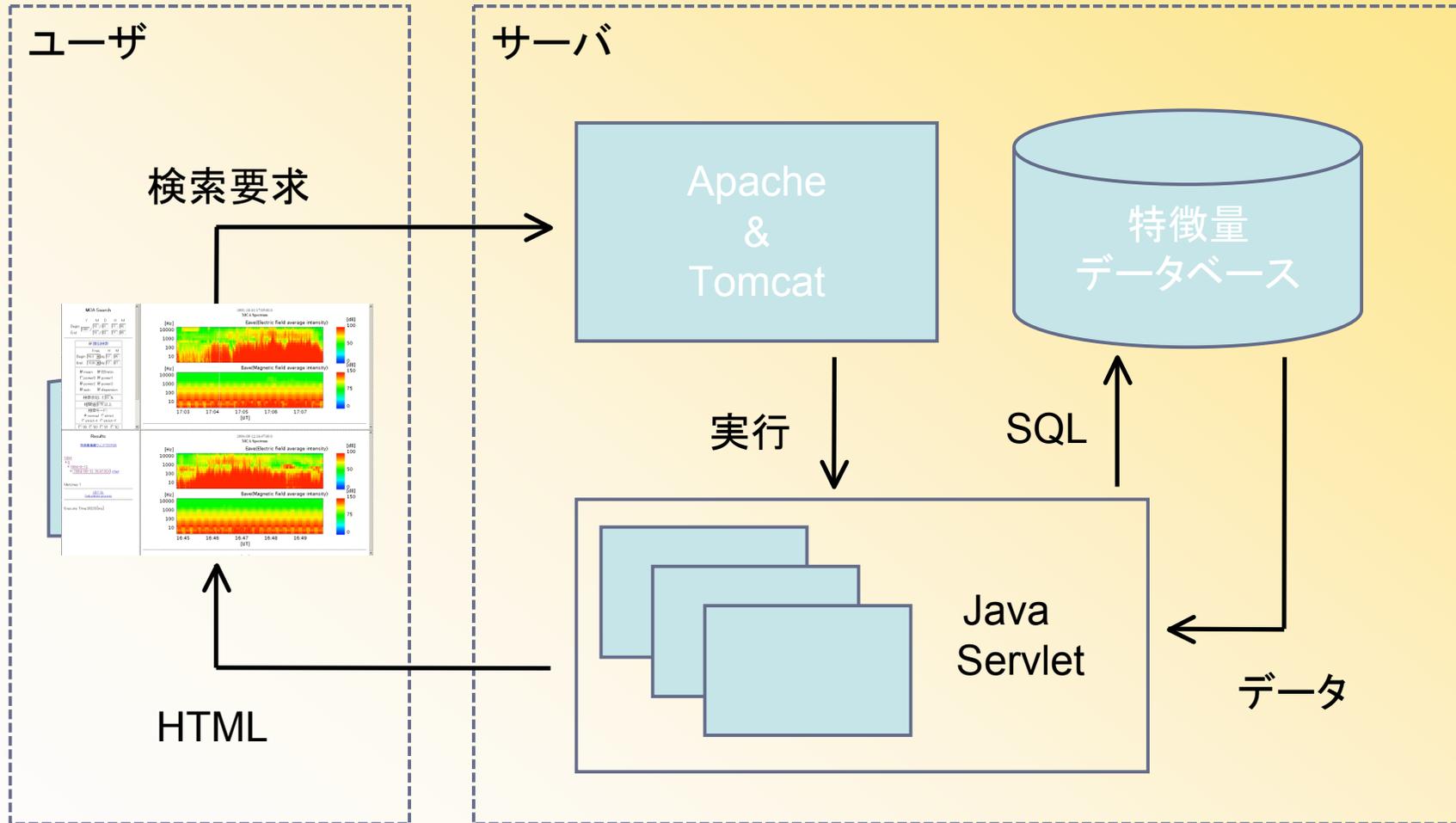
【新システム概要】

データ登録部
(データ依存)

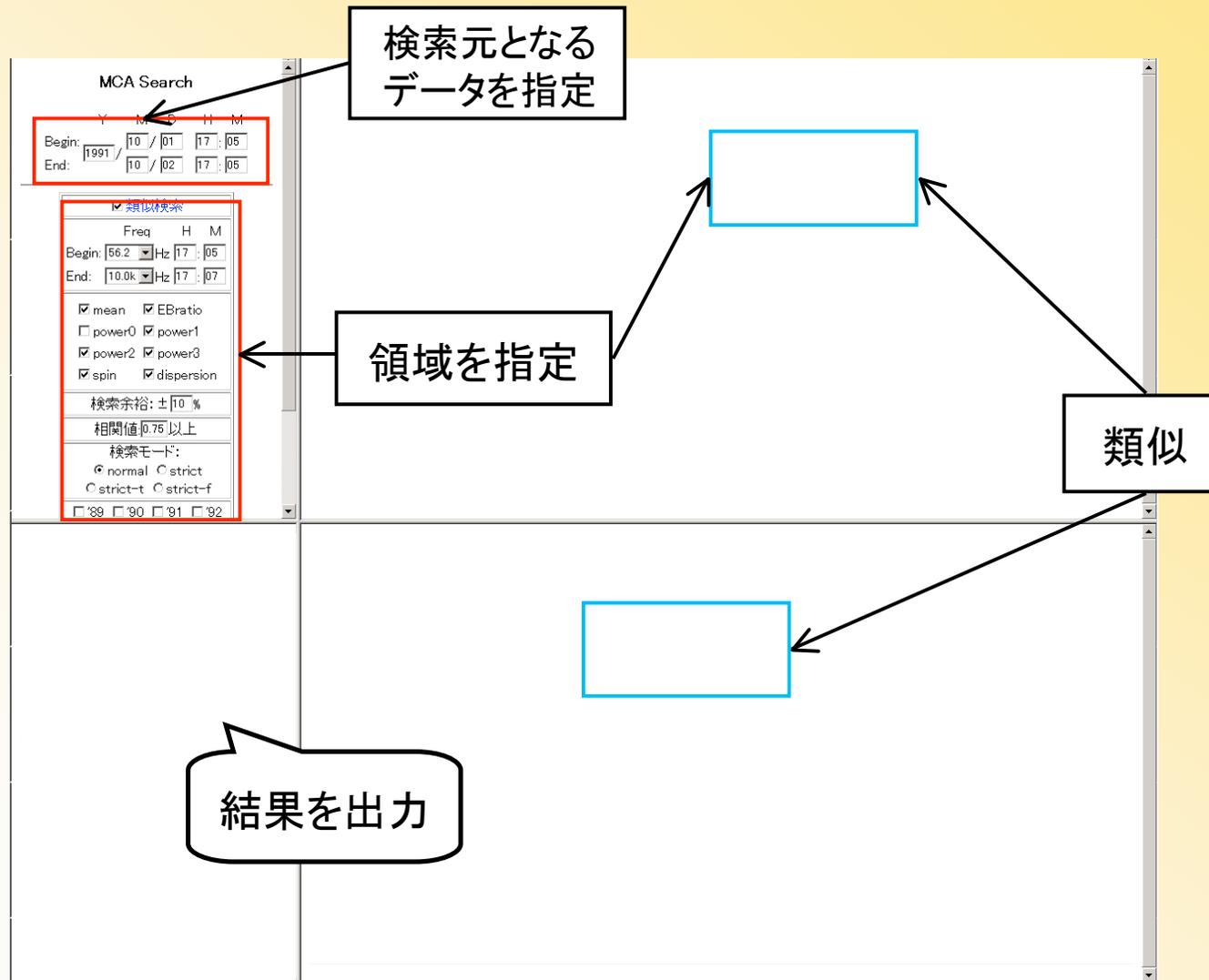
Event Finder
(データ非依存)



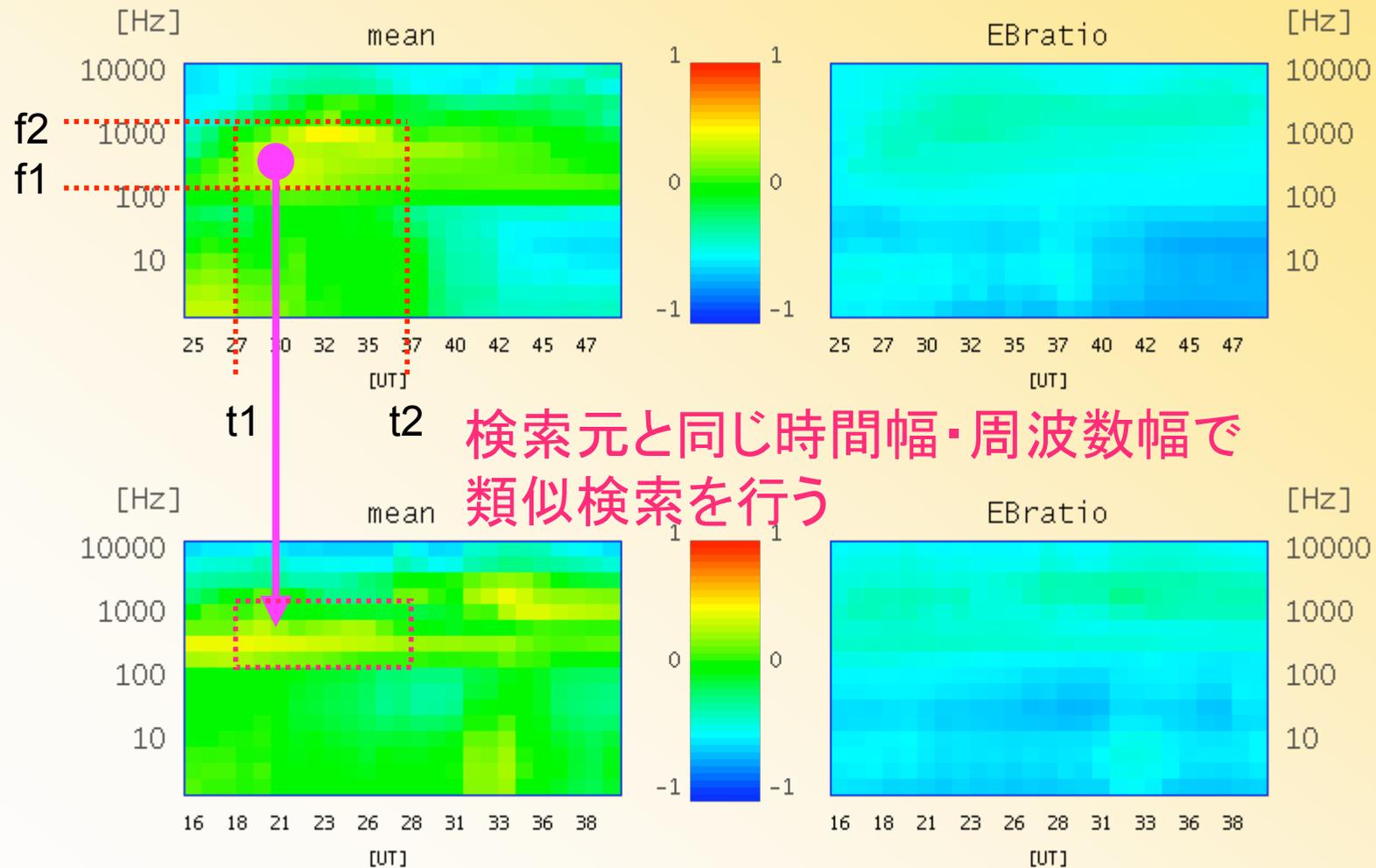
【システム構成】



【Event Finder 概観】



【検索アルゴリズム】



検索元と同じ時間幅・周波数幅で
類似検索を行う

類似度の指標として相互相関を用いる

【高速検索アルゴリズムの検討】

[問題点]

- 特徴量はRDB内で、表形式で格納
→幅のある検索が困難
- レコード数が非常に多い
(あけぼのMCAの場合:約4000万件)
→闇雲に相互相関をとるのは不可能

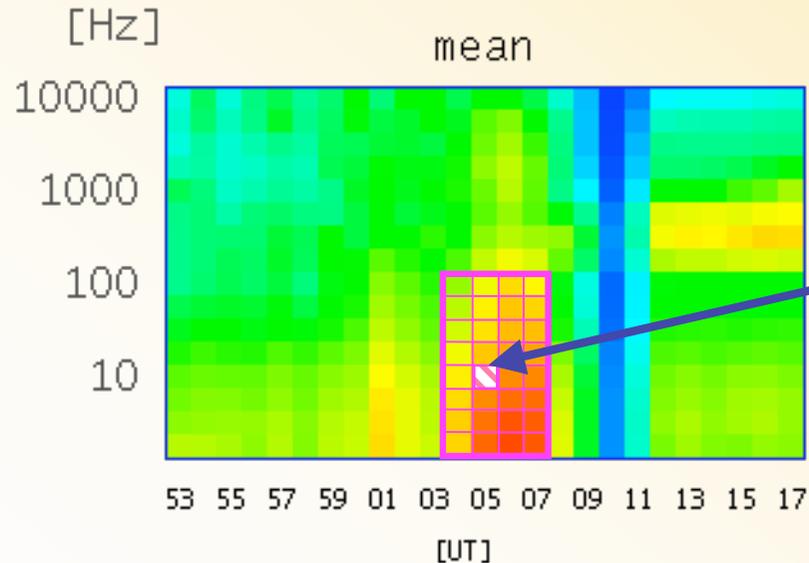
[対策]

- 2ステップ検索法の採用

【2ステップ検索】

Step1.

検索元領域の中心における特徴量と類似する特徴量をもつ日時をデータベース中から取り出す

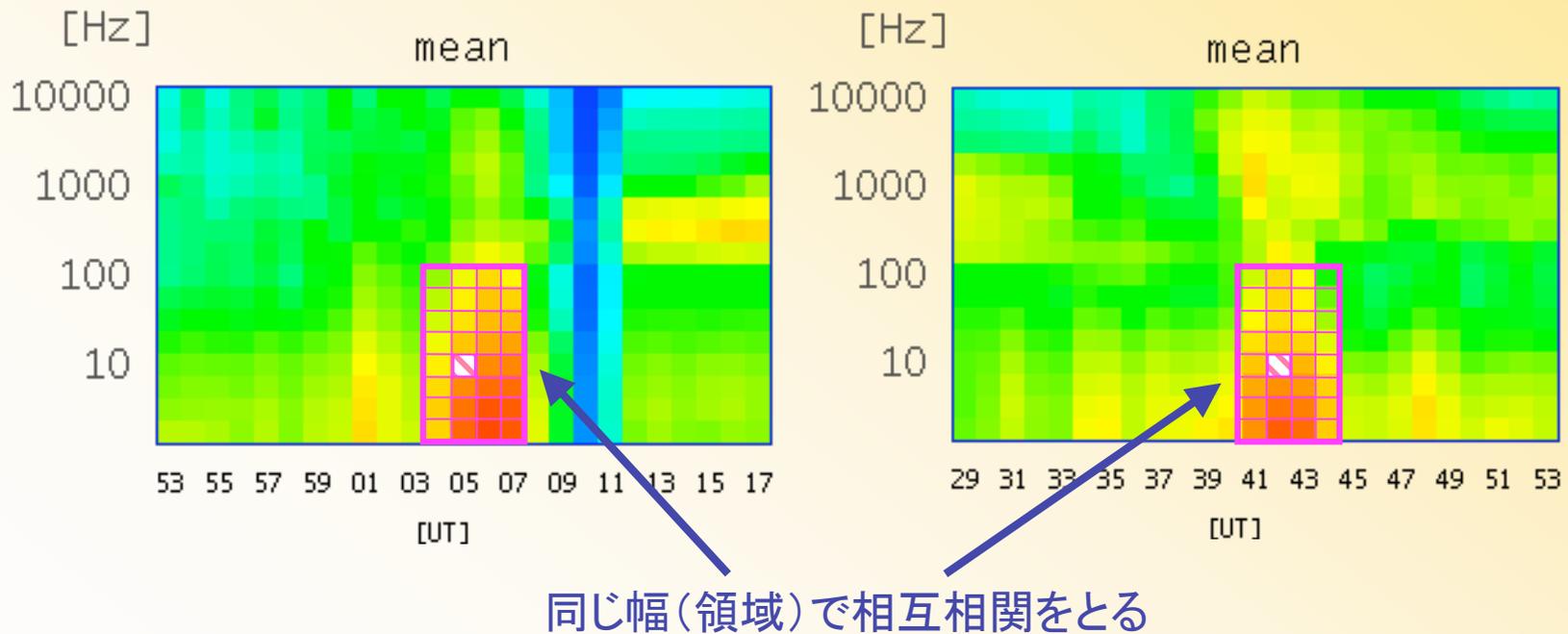


この時刻・周波数における特徴量の値と類似する値をもつ日時を、データベースから探す

【2ステップ検索】

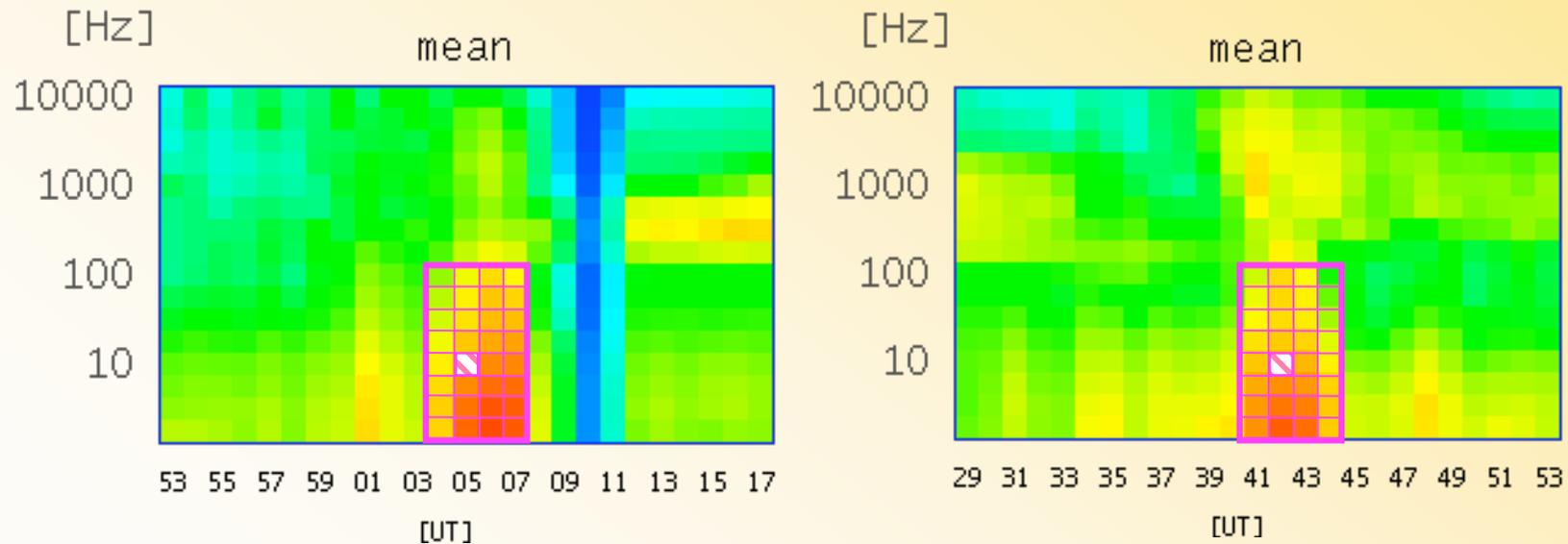
Step 2.

Step1で取り出した候補データと、元データの特徴量の相互相関をとる



【結果出力】

相関値がユーザーが指定した値以上ならば、
その日時を検索結果として出力する

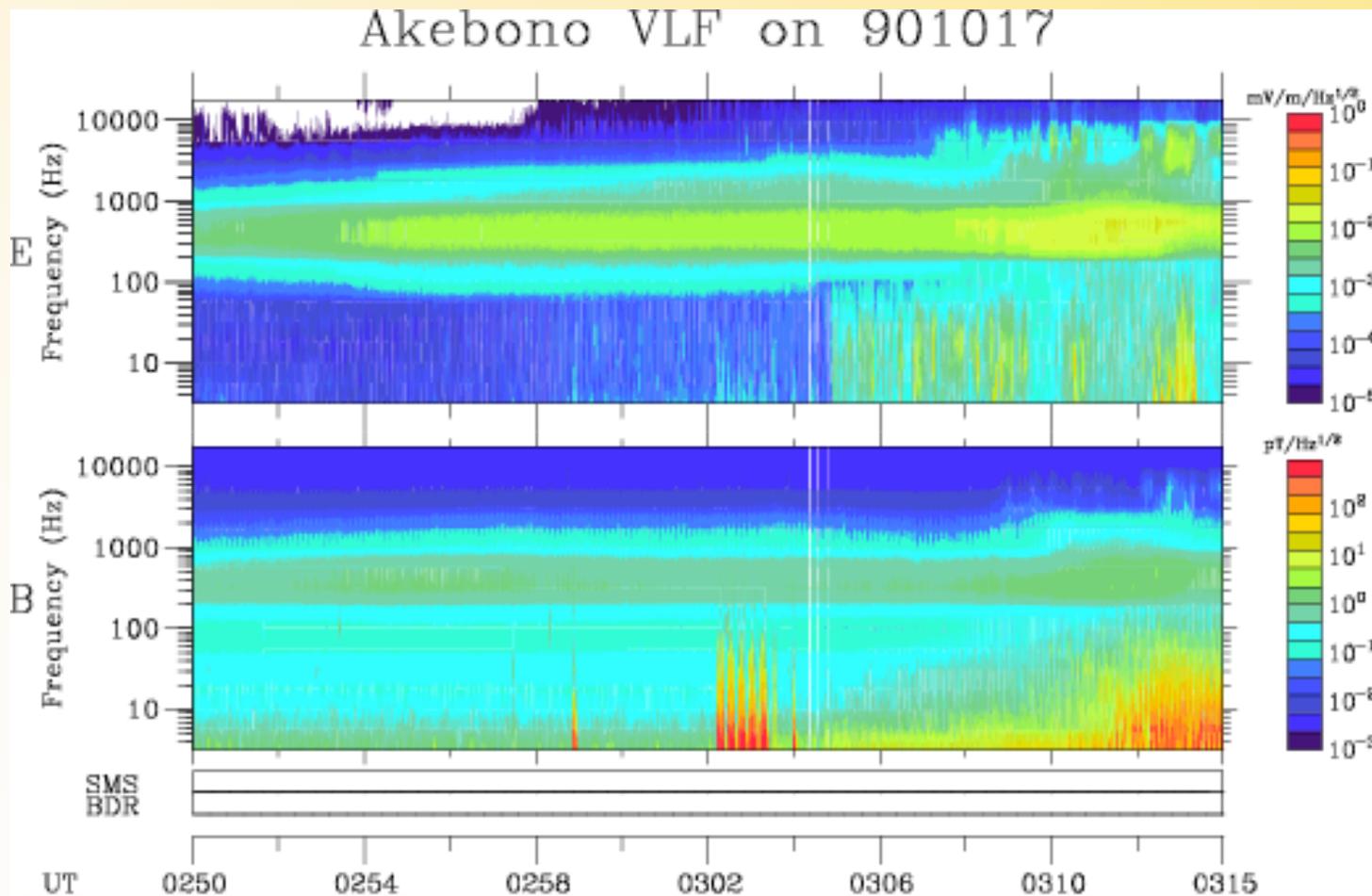


相関値 $R \leq$ 計算された相関値 \Rightarrow 類似
相関値 $R >$ 計算された相関値 \Rightarrow 非類似

※相関値Rはユーザーが指定

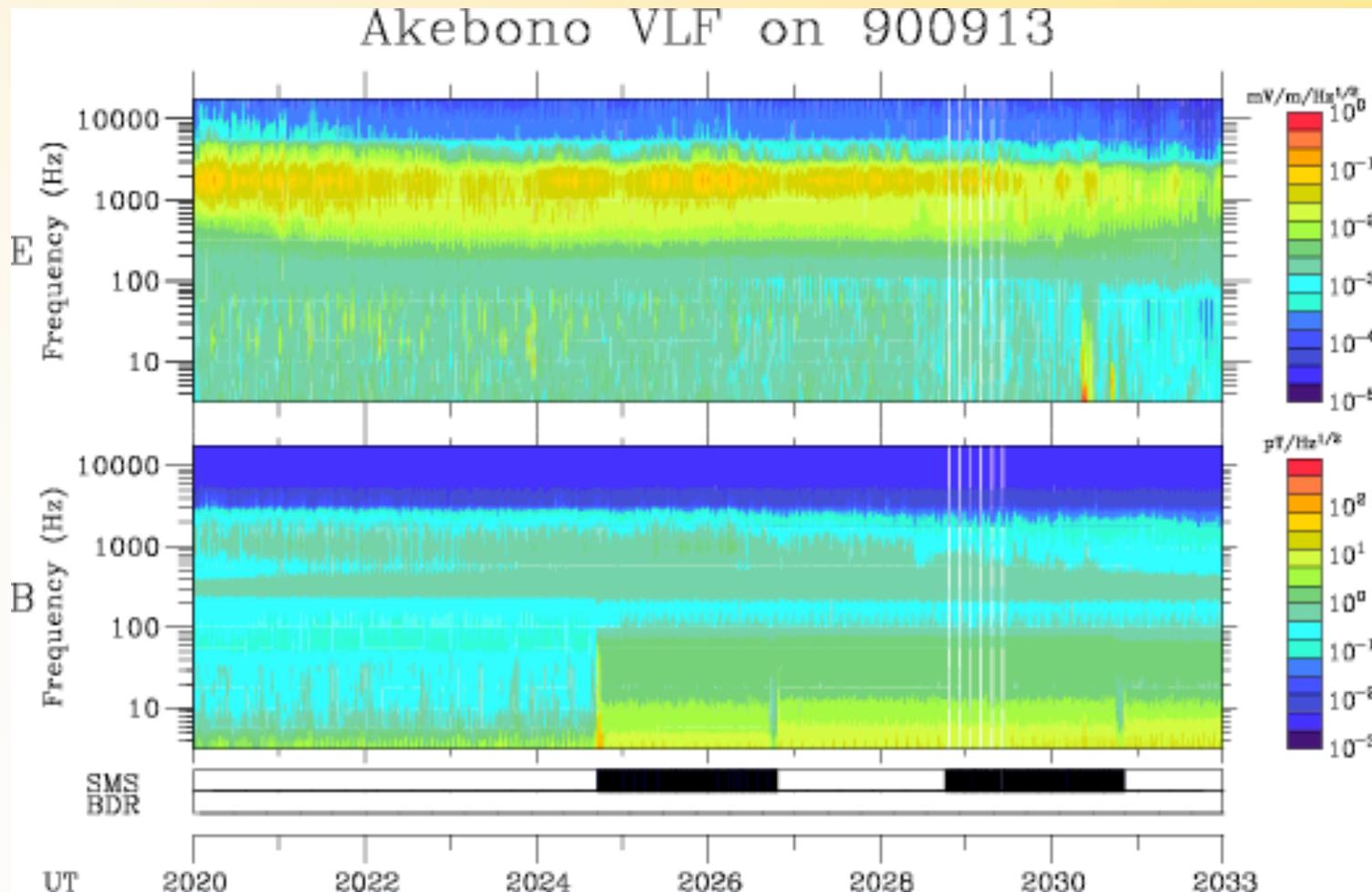
【プラズマ圏ヒス】

- 100Hz~1kHz の周波数帯で、時間変動がほとんどない電波雑音



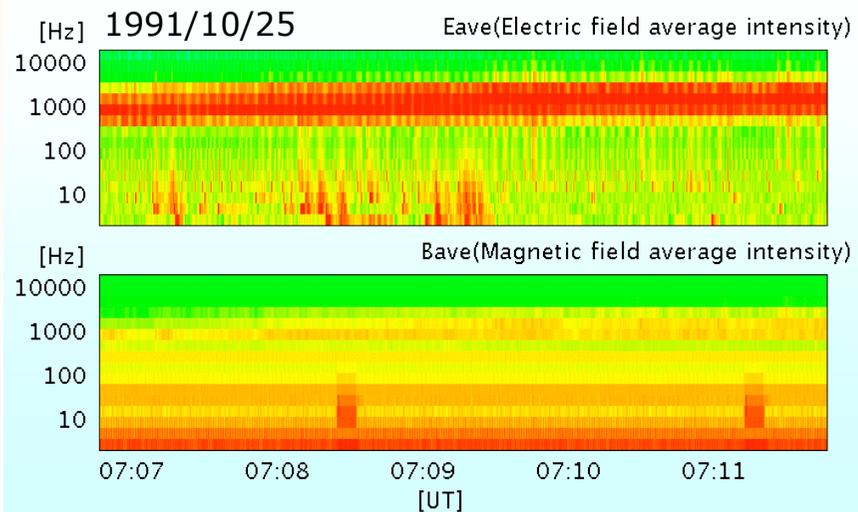
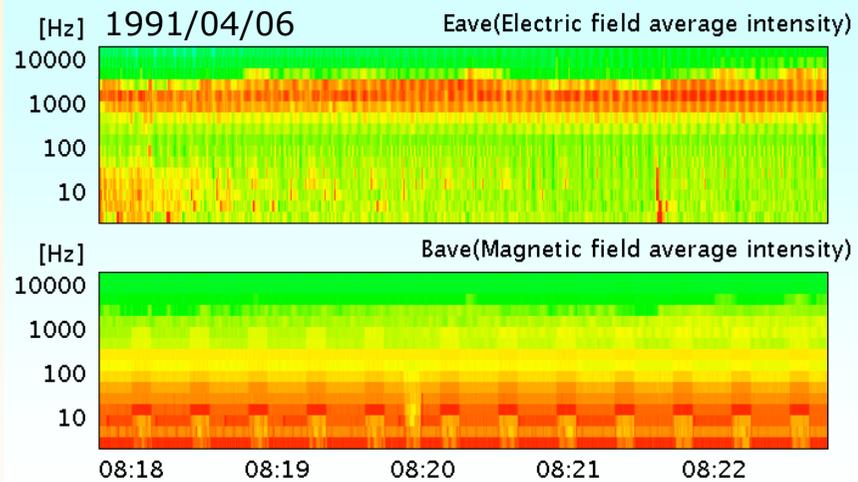
【コーラス】

- 500~10kHz の周波数帯でコヒーレントな波動現象で、観測周波数に緯度依存性がある

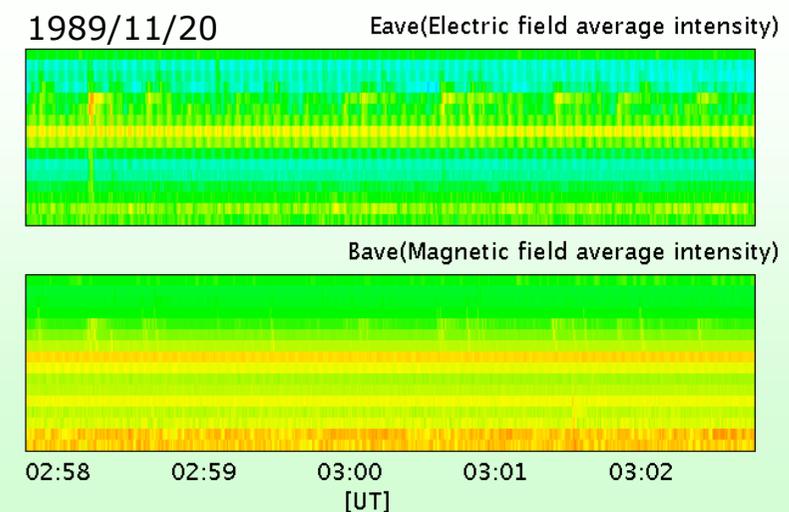
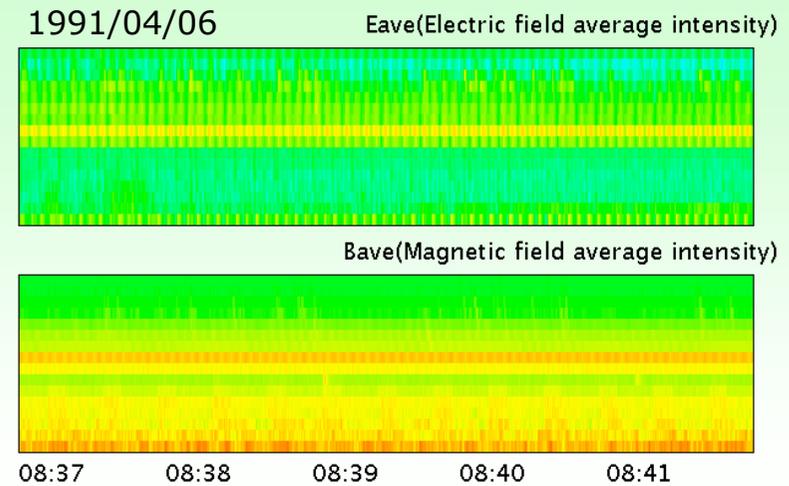


【検索結果例】

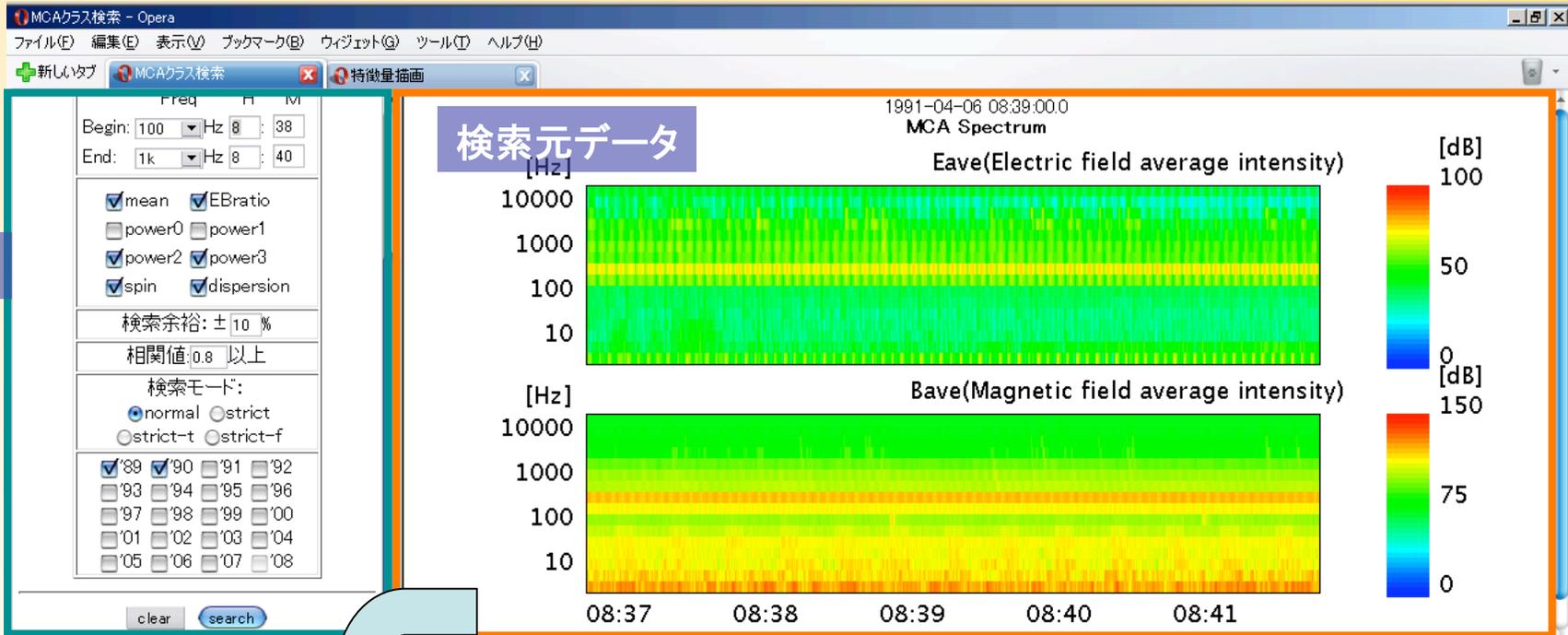
コーラス



プラズマ圏ヒス



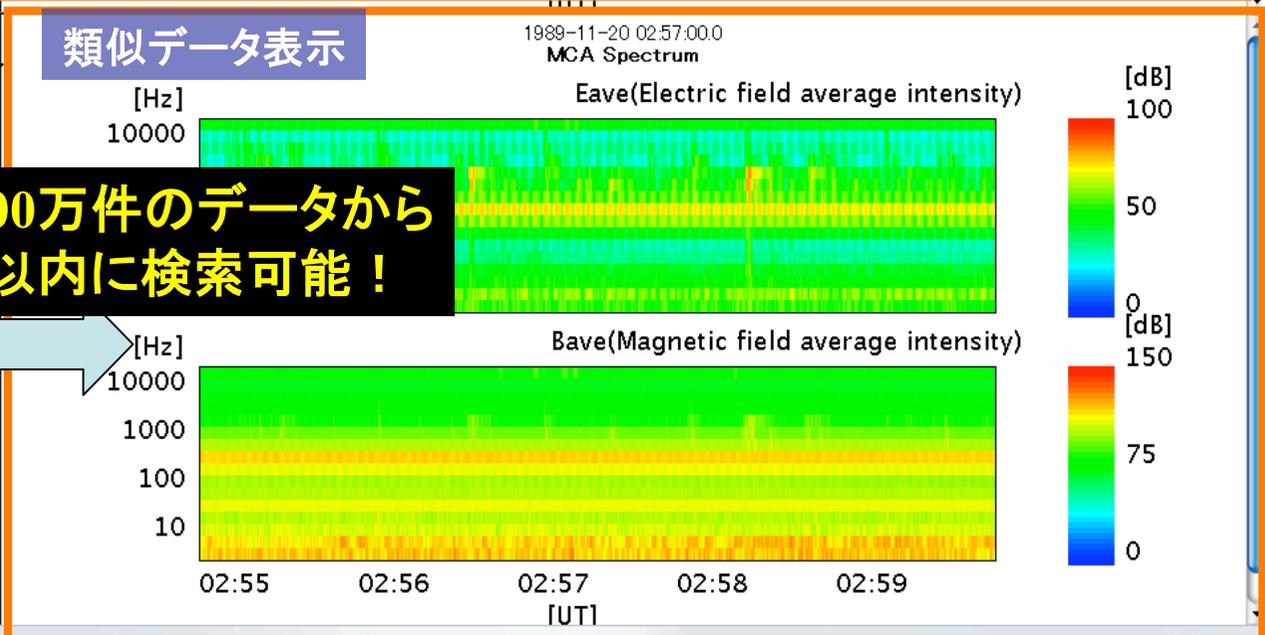
検索条件



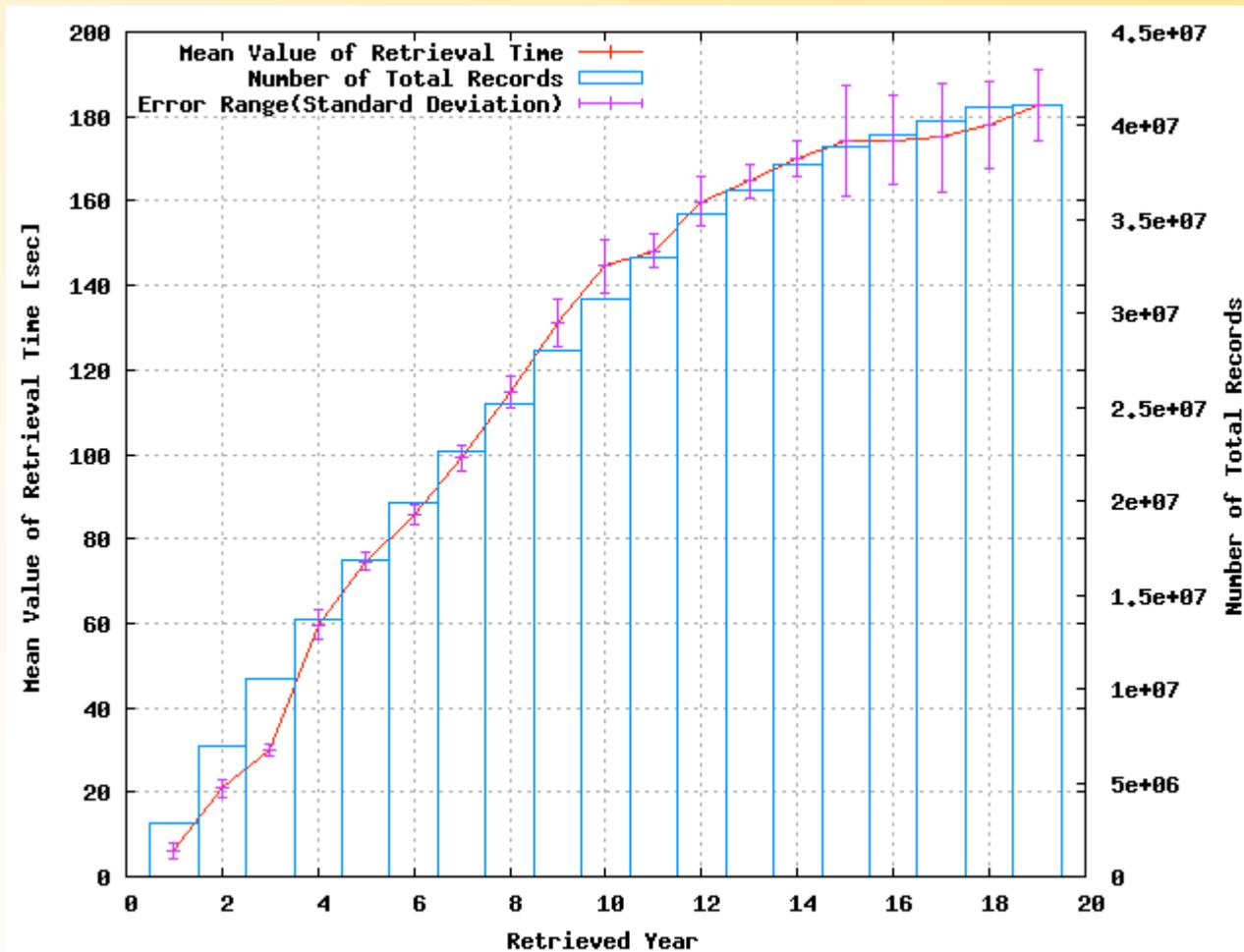
抽出



**約1000万件のデータから
30秒以内に検索可能！**



【検索年数と検索時間】



【利便性の向上】

- 検索オプション
 - ✓ 検索余裕・検索モード・同時検索年
 - ユーザの要望にあわせて柔軟な検索が可能
- 類似の判別に使用する特徴量の視覚化
 - ✓ 特徴量を画像として表示
 - 特徴量同士の比較が画像で可能に
- 検索過程・結果リスト保存機能
 - ✓ テキストファイル形式で保存
 - 検索結果リストや検索過程を保存・再利用可能に

【検索オプション】

- 柔軟な検索が行えるよう、検索オプションを追加
 - ✓ 検索に余裕幅をもたせて検索が可能
 - 検索元の特徴量の±数パーセントまで類似とみなす
 - ✓ 検索の厳しさを選択可能
 - 強度が同じ現象が長時間続くものと短時間のものとを区別
 - ✓ 同時に検索できる年を選択可能
 - ユーザが検索を行いたい年を複数選択可能

The screenshot shows a search options dialog box with the following elements:

- End: 56.2 Hz 8 : 36
- Checked options: mean, EBratio, power2, power3, spin, dispersion
- Unchecked options: power0, power1
- ① 検索余裕: ± 10 %
- 相関値: 0.9 以上
- ② 検索モード:
 - normal
 - strict
 - strict-t
 - strict-f
- ③ Year selection grid:

<input type="checkbox"/>	'89	<input checked="" type="checkbox"/>	'90	<input checked="" type="checkbox"/>	'91	<input type="checkbox"/>	'92
<input type="checkbox"/>	'93	<input type="checkbox"/>	'94	<input type="checkbox"/>	'95	<input type="checkbox"/>	'96
<input type="checkbox"/>	'97	<input type="checkbox"/>	'98	<input type="checkbox"/>	'99	<input type="checkbox"/>	'00
<input type="checkbox"/>	'01	<input type="checkbox"/>	'02	<input type="checkbox"/>	'03	<input type="checkbox"/>	'04
<input type="checkbox"/>	'05	<input type="checkbox"/>	'06	<input type="checkbox"/>	'07	<input type="checkbox"/>	'08
- Buttons: clear, search

【記録機能】

検索結果のリストや検索過程を保存・再利用しやすくするための機能
テキスト形式でファイル保存することが可能

結果リスト記録機能

検索過程記録機能

Results
特徴量描画ウィンドウOPEN

1991
+3
+ 1991-3-5
+ 1991-03-05 13:30:00.0 char

+7
+8
+9
+10
+12

Matches 20

Execute Time:25561[ms]

LIST DL(右クリックで保存)
Calculation process

----- Record Time:2008/02/12 17:46:14 -----
<< Retrieval Condition >>
1991/07/11 5:37-5:47
3.16Hz - 316Hz
Check: mean EBraio power1 power2 power3 s
Retrieval Margin: 10%
Correlation Coefficient: 0.95 over
Retrieval Mode: normal
Retrieval Year: 1990,1991,
<< Hit Date-Time List >>
1991-03-05 13:30:00.0
1991-07-11 05:42:00.0
1991-07-13 03:36:00.0
1991-08-01 22:01:00.0
1991-08-04 15:13:00.0
1991-08-05 00:45:00.0
1991-08-05 00:53:00.0
1991-08-08 10:31:00.0
1991-08-12 19:48:00.0
1991-08-16 21:20:00.0
1991-08-17 23:19:00.0
1991-08-20 23:14:00.0
1991-08-20 07:11:00.0
1991-09-08 18:11:00.0
1991-09-14 14:22:00.0
1991-09-21 18:54:00.0
1991-09-22 17:48:00.0
1991-10-15 05:20:00.0
1991-10-15 05:43:00.0

検索結果
リスト

Results
特徴量描画ウィンドウOPEN

1991
+3
+ 1991-3-5
+ 1991-03-05 13:30:00.0 char

+7
+8
+9
+10
+12

Matches 20

Execute Time:25561[ms]

LIST DL(右クリックで保存)
Calculation process

----- Record Time:2008/02/12 17:46:14 -----
■ 1991-07-11 05:42:00.0
<< Retrieval Condition >>
1991/07/11 5:37-5:47
3.16Hz - 316Hz
Check: mean EBraio power1 power2 power3
Retrieval Margin: 10%
Correlation Coefficient: 0.95 over
Retrieval Mode: normal
Retrieval Year: 1990,1991,
<< Hit Date-Time List >>
× 1990-01-05 02:15:00.0
• mean: 0.83275753
• ebraio: 0.97797567
• power1: 0.842355
• power2: 0.9540498
• power3: 0.9708012
• spin: 0.9677288
• disp: 0.99588114
• power2: 0.95363246
• power3: 0.9640546

特徴量の相互相関

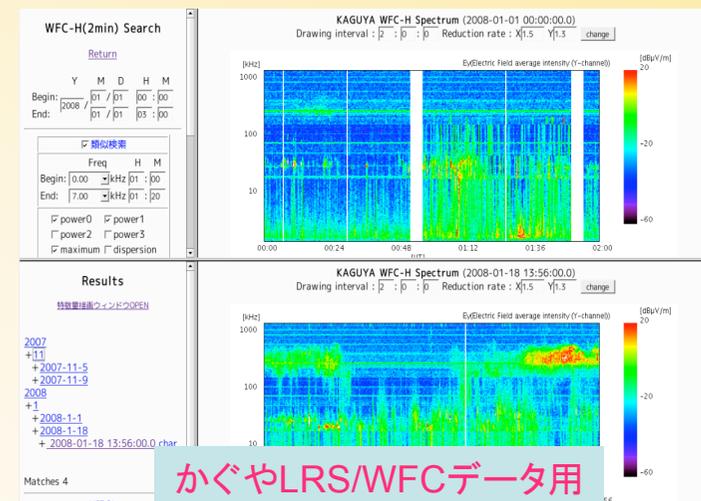
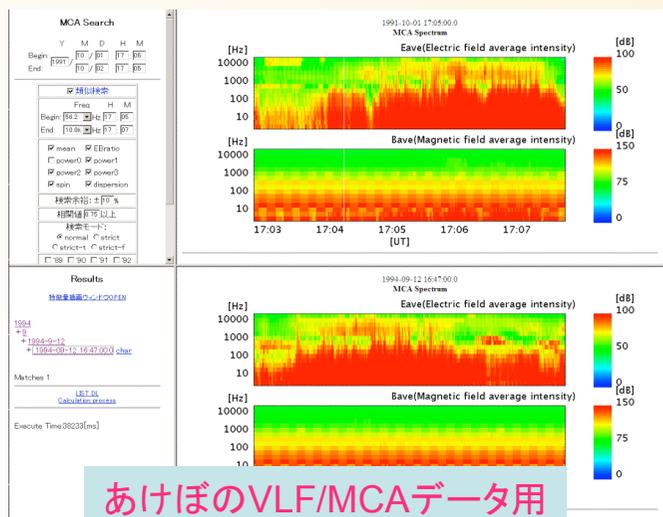
【システムの一般化】

◆プログラムの変更箇所を集約

- 外部の設定ファイルに記述
- 設定ファイルは変更しても再コンパイル・再起動が不要
 - ✓ 動作中に設定変更が可能

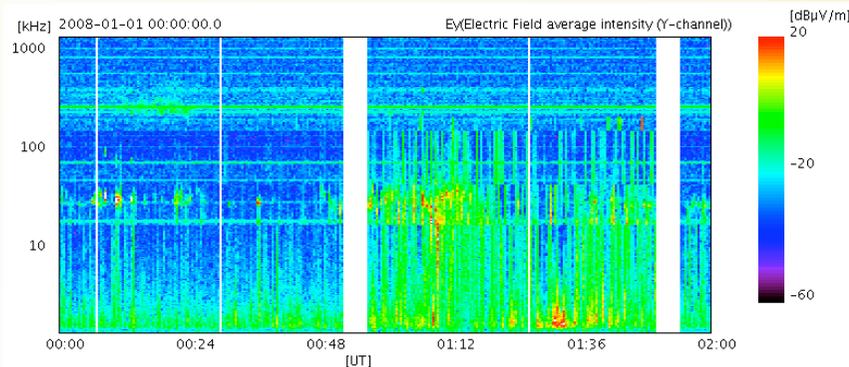
◆設定ファイルを要求に応じて切り替え

- データ種ごとに設定を用意
- ユーザが検索したいデータを選択

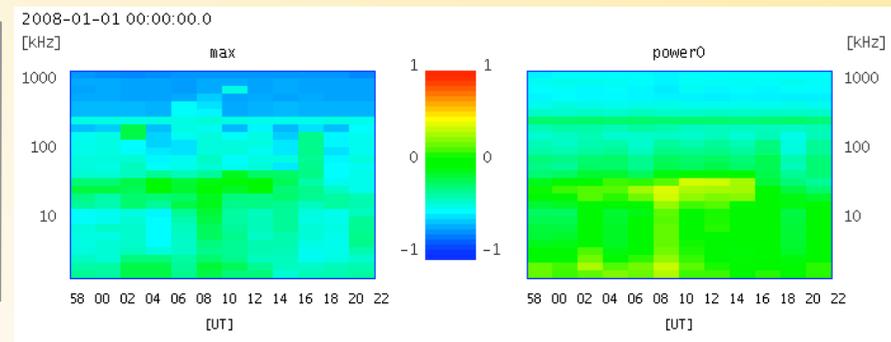


【かぐやLRS/WFCデータへの応用】

- 検索に使用する特徴量データベースが必要
 - WFC-Hデータの特徴量を算出し格納
 - 特徴量の変更に合わせて検索用Javaサーブレットを変更
- 描画システムの調整(スペクトル・特徴量)
 - MCAデータとはデータ点数が異なる

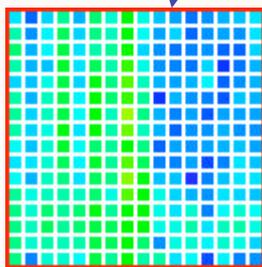
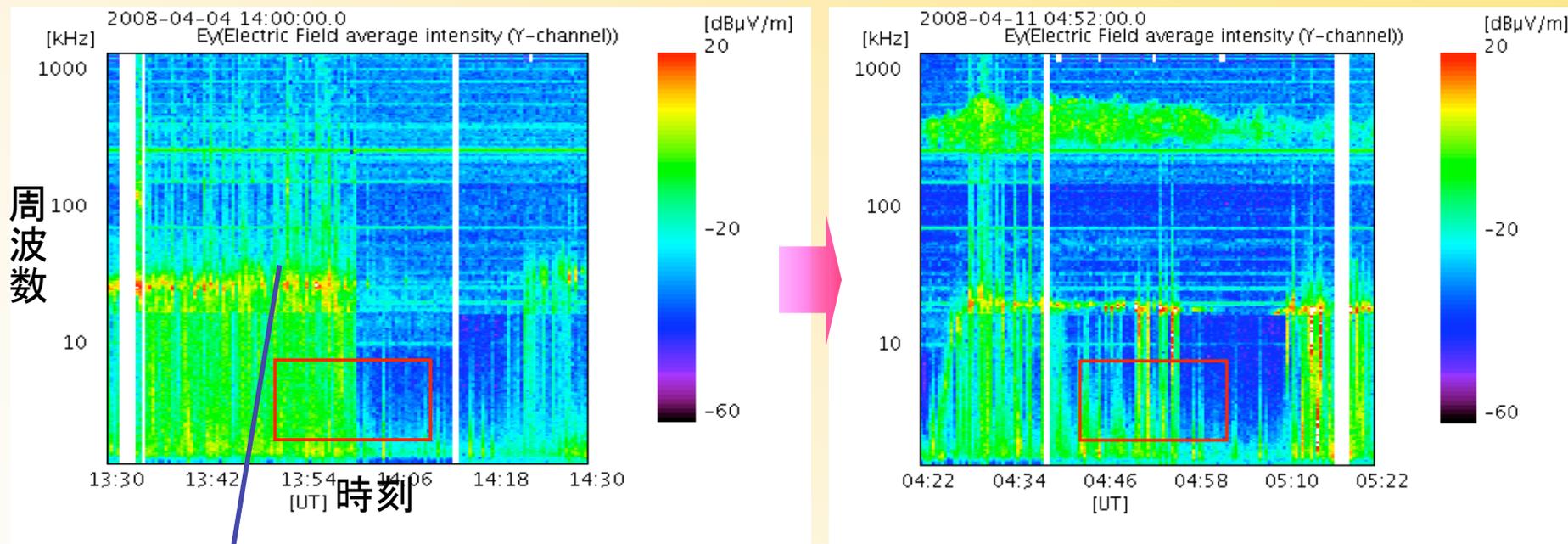


スペクトル



特徴量

【かぐやLRS/WFCへの応用(開発中)】



DCT

$$X_k = \sum_{n=0}^{N-1} x_n \cos \left\{ \frac{\pi}{N} \left(n + \frac{1}{2} \right) k \right\}$$

Power0	3328	0	-1517	0	317	0	-108	0	Power2
	0	0	0	0	0	0	0	0	
	-1517	0	605	0	-89	0	0	0	
	0	0	0	0	0	0	0	0	
	317	0	-89	0	-64	0	37	0	
	0	0	0	0	0	0	0	0	
Power1	-108	0	0	0	37	0	-29	0	Power3
	0	0	0	0	0	0	0	0	

【まとめ】

- あけぼのVLFスペクトルDBに対し、類似現象を効率よく検索、画像表示するシステムを開発
 - ⇒ 類似検索機能自体はデータ種に非依存！
(かぐやLRS/WFCスペクトルデータに拡張中)
- 従来型の自動分類ではなく、ユーザー指定のデータと類似するデータを高速かつ高精度に検索
- 検索パラメータや検索結果の保存機能など、ユーザーの利便性向上
- 異種データ(波形・粒子 etc.)への拡張法
データ量増に応じた検索機能の高速化を検討中