

衛星画像からのパターン学習と 時空間データマイニング

高知大学自然科学研究系理学部門
本田理恵

宇宙科学情報解析シンポジウム

2009.2.23

宇宙航空研究開発機構 宇宙科学研究本部

はじめに

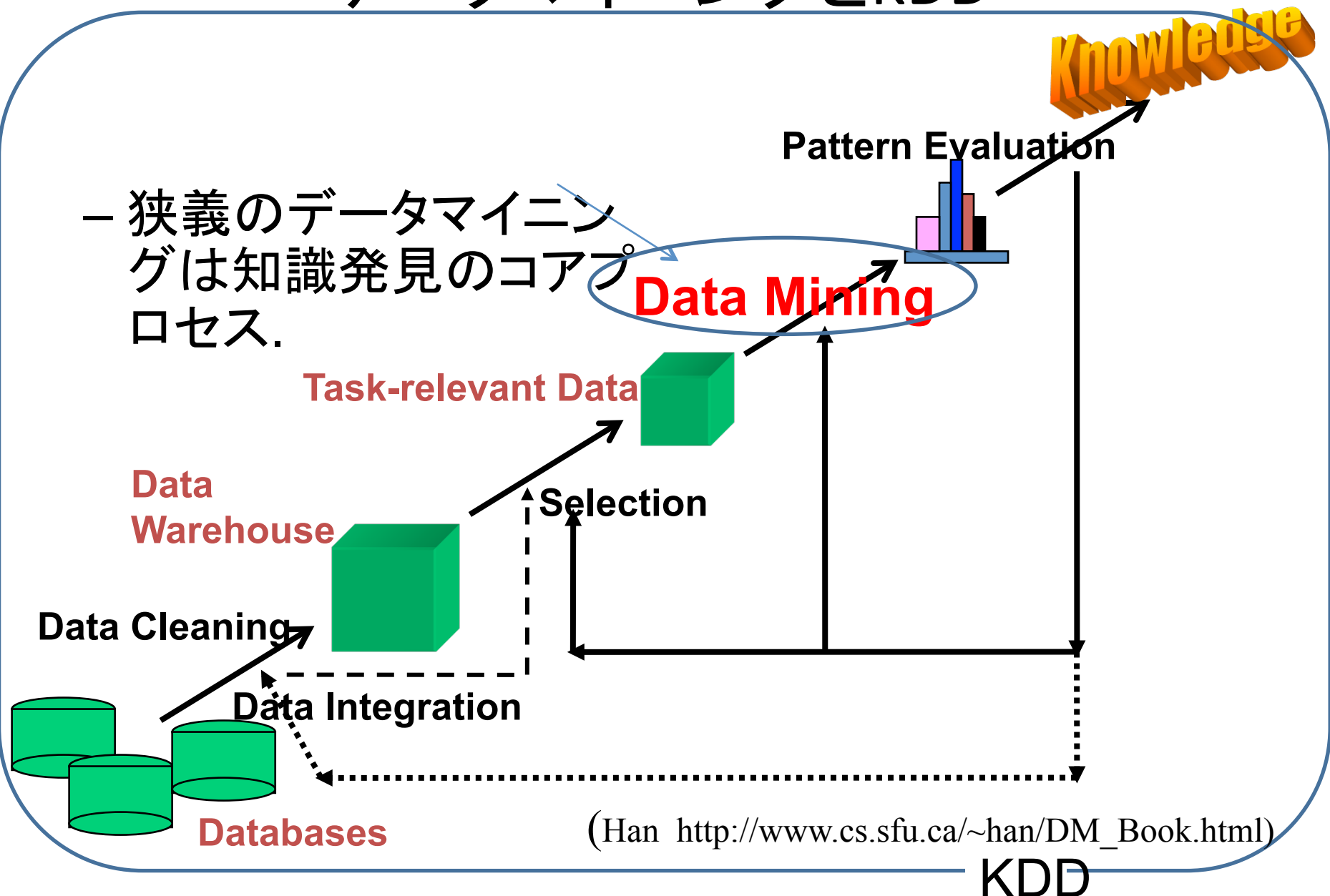
- 惑星科学の分野でも大規模なデータが取得され始めている。
 - KAGUYA/SELENE: 原データだけでテラバイトオーダー
- データの大規模性を生かした新たな知識発見が期待される。
 - データマイニング、機械学習の方法論を利用した、コンピュータによる発見支援環境の整備が重要に.
 - 地球観測衛星の分野ではすでに始まっている。

データマイニング(data mining)とは

- 多量のデータから有用な知識を発掘する技術の総称
 - 人工知能(A.I. 特に機械学習)、統計、情報検索などを総合的に活用した実用志向の研究分野
- KDDと呼ばれることもある
 - KDD (knowledge discovery from Database: データベースからの知識発見)
 - “KDD is the **non-trivial** process of identifying **valid(有効で)**, **novel(新規で)**, **potentially useful(利用性があるって)**, and **ultimately understandable patterns(最終的には理解可能なパターン)** in the data.” (by Usama Fayyad)

データマイニングとKDD

- 狭義のデータマイニングは知識発見のコアプロセス.



衛星画像・データ分野における 大規模化の問題

- 従来、専門家がデータ(画像、グラフなど)を(視覚的に)見て行っていた作業が破綻
 - 研究者は成果の上げやすい興味のある個所だけをピックアップして解析するようになる
 - 大部分のデータが死蔵される可能性がある
 - 従来研究の詳細化、洗練化は実施できるものの、新規事実を見落とす可能性がある

衛星データの特徴

- データそのものから校正に必要な情報を抽出しなければならないことが多い。
- ノイズ・欠落が頻発することが多い。
- 場(時間・空間)のデータであることが多い
 - 場のデータのままの分析
 - 特徴的な対象(オブジェクト)やイベントを抽出して分析

衛星画像からのデータマイニングについて の研究事例

パターン学習

(a)惑星画像からのクレーター抽出

(Honda et al. 2000¹, Honda et al. 2001³, Honda et al. 2002⁵, Burl et al., 2005¹⁰)

(b)マルチバンド画像による惑星の地質判定 (横田 2003⁷, Honda et al. 2003⁸)

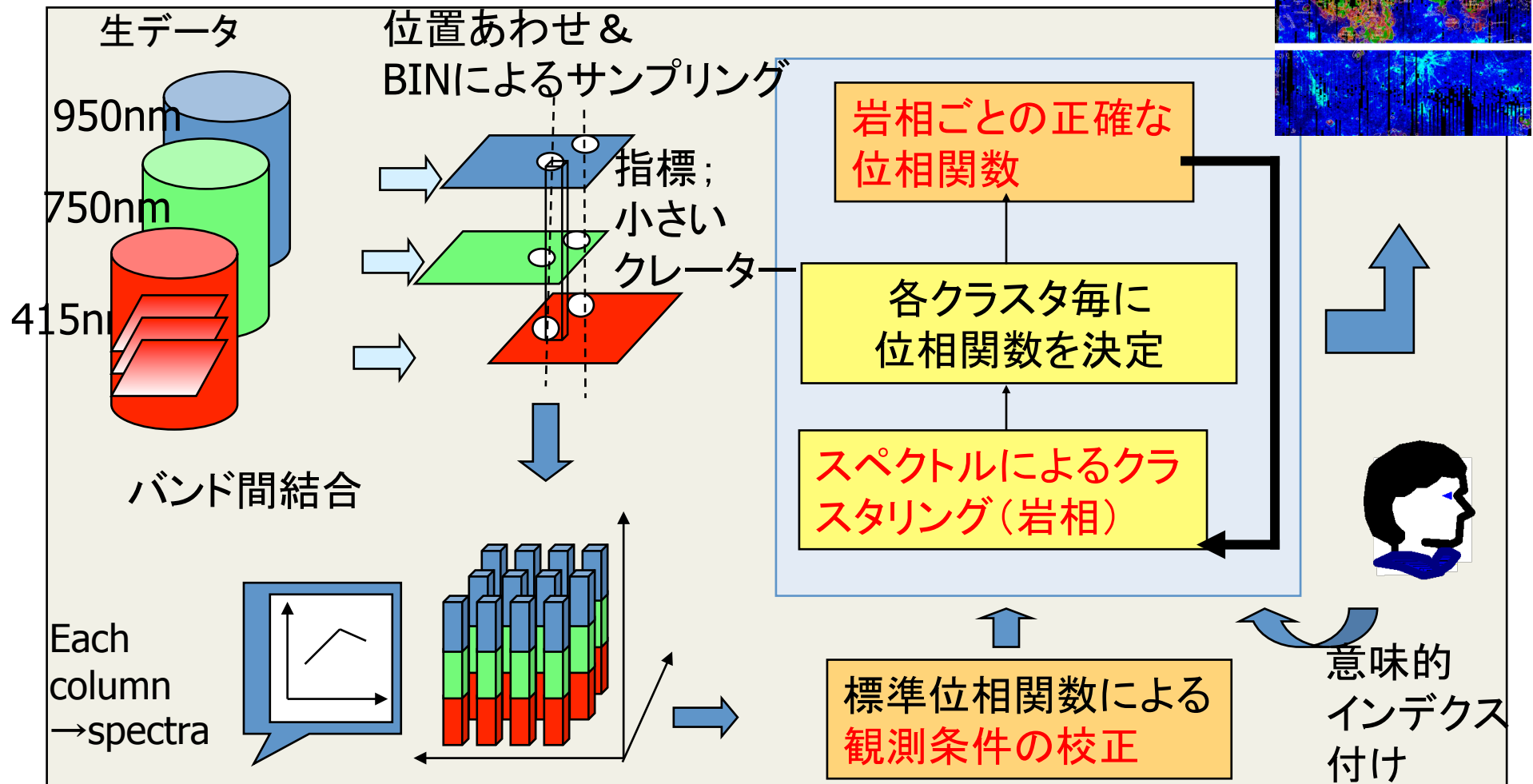
時空間データマイニング

(c)気象画像からの時空間パターンの検出 (Honda et al. 2000², Honda et al. 2001⁴, Honda et al. 2002⁶, 石津 2009¹³)

(d)地球観測衛星による植生データからの時空間変動 (Honda 2006¹¹, 吉岡、本田2008¹², 山下2009¹⁴)

(b) 月のマルチスペクトル画像に対するマ イニング (横田2003⁷, Honda et al., 2003⁸

スペクトル
ユニット図



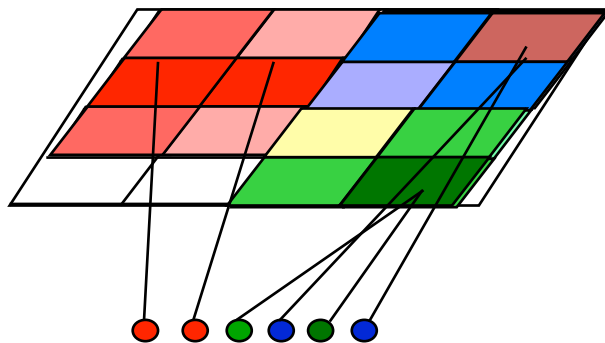
クラスタリングの手法

Kohonenの自己組織化マップ (SOM)

2層のニューラルネットワークからなる多次元ベクトルのための教師なし学習法. 入力ベクトルの特徴を2次元の出力層に投影する。結果はクラスタリングとその可視化にも利用できる。

競合層(特徴マップ)

$$M = \{m_i \mid m_i \in \mathbb{R}^n, i = 1, 2, \dots, k\}$$



入力層(input vectors)

$$X = \{x_j \mid x_j \in \mathbb{R}^n, j = 1, 2, \dots, d\}$$

For all x_j ,

$$\bullet c = \operatorname{argmin} |x_j - m_c(t)| \quad (c: \text{winner})$$

$$\bullet m_i(t+1) = m_i(t) + \alpha(t) (x_j - m_c(t))$$

when $i \in N_c$

N_c : c 's neighborhood

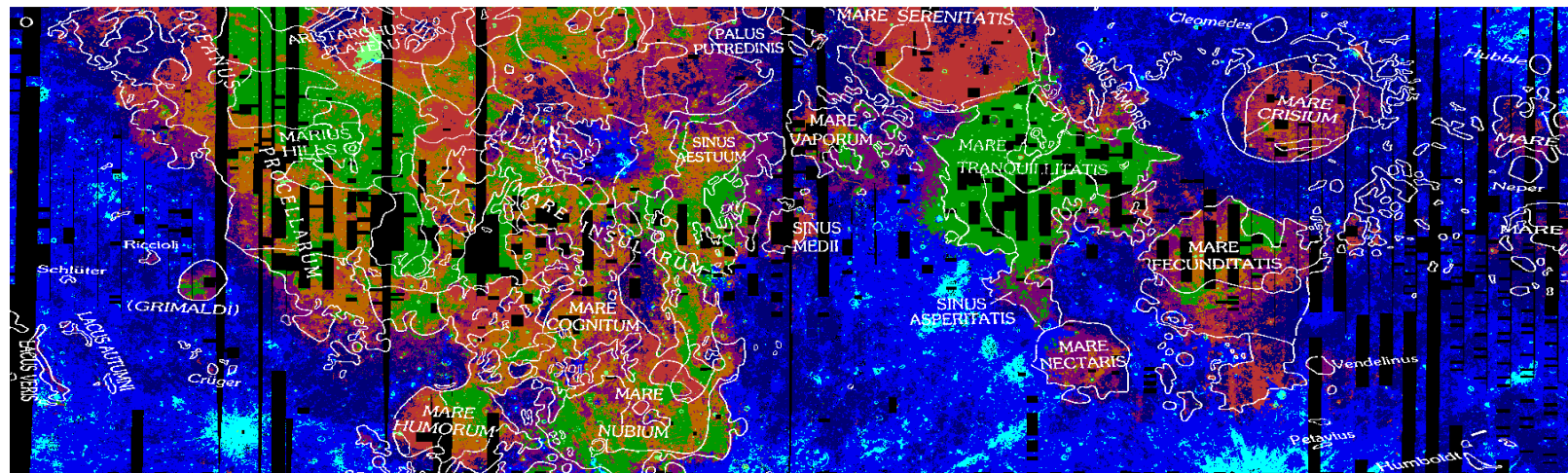
α : the learning rate

(0.3)

Iterate the above process for T cycles.

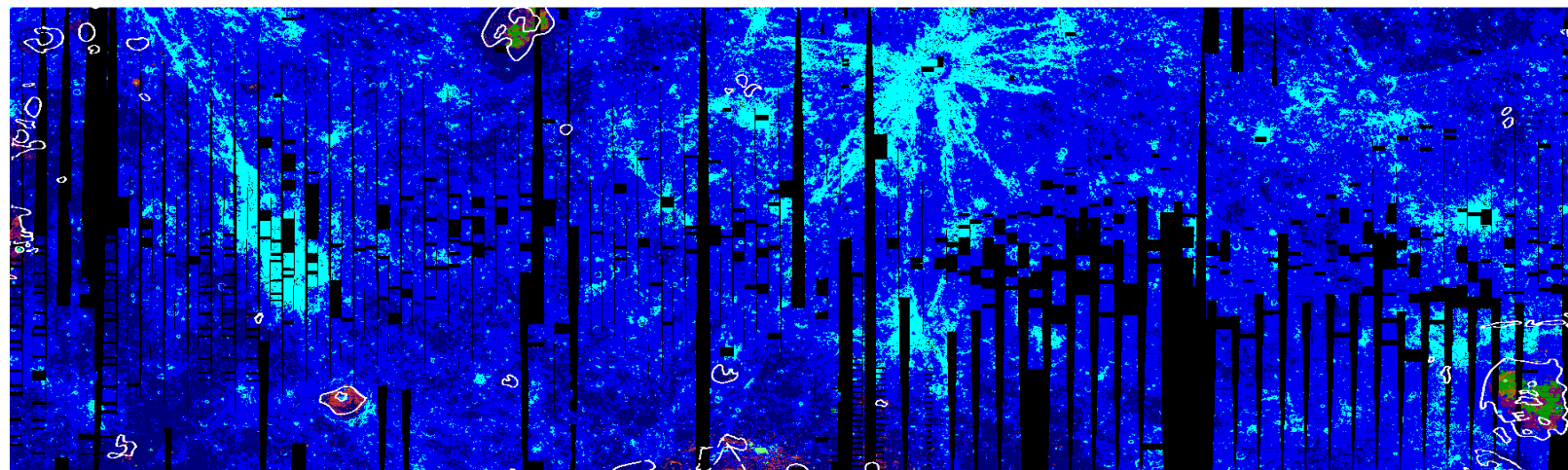
月マルチスペクトルデータベースから得られたス
ペクトルユニット図 (Clementine衛星)

SOM# 0 1 2 3 4 6 7 8



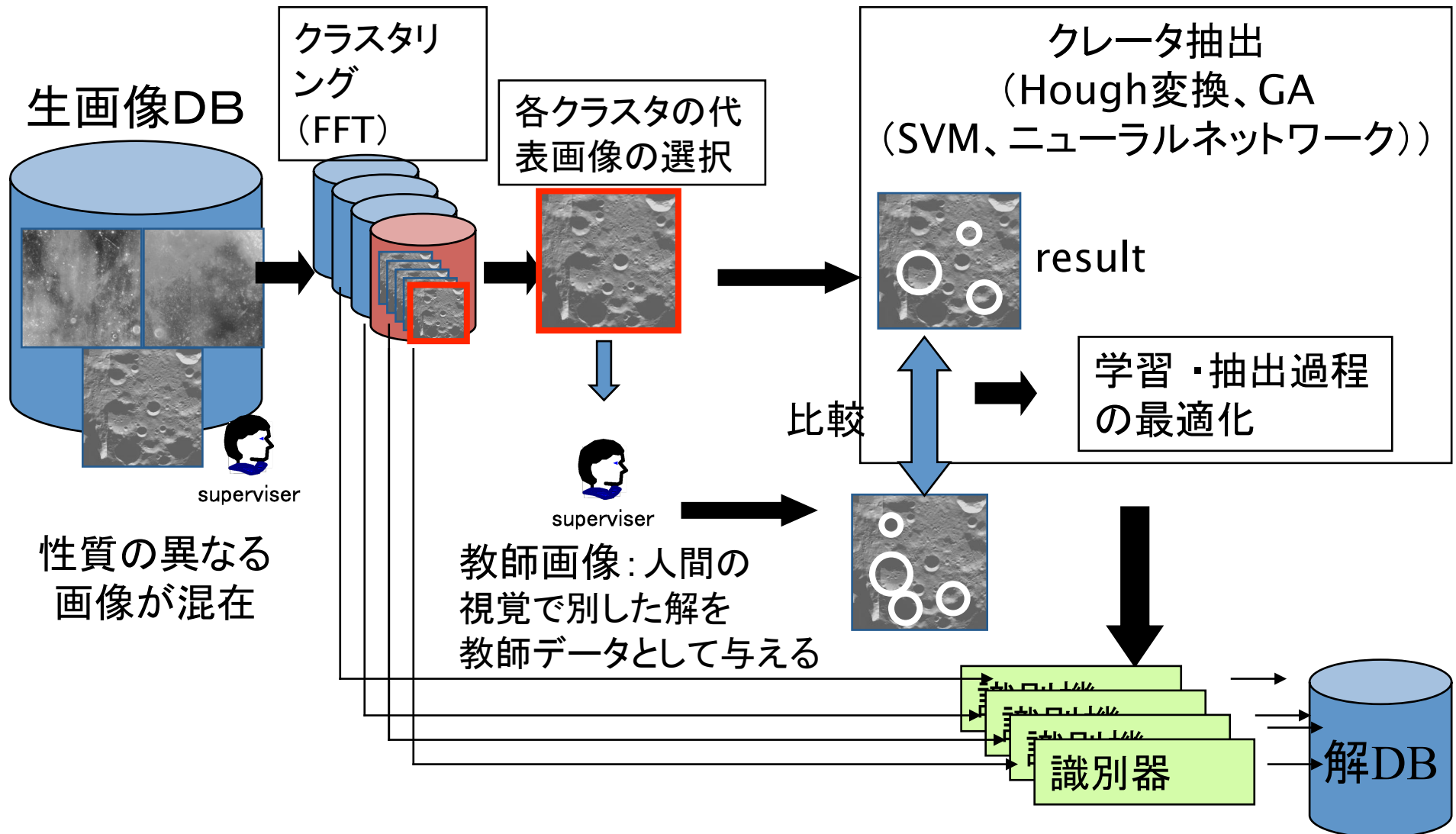
white
boundari
es

(Wilhelm
S,
1987)



(a)惑星画像からのクレータ抽出システム

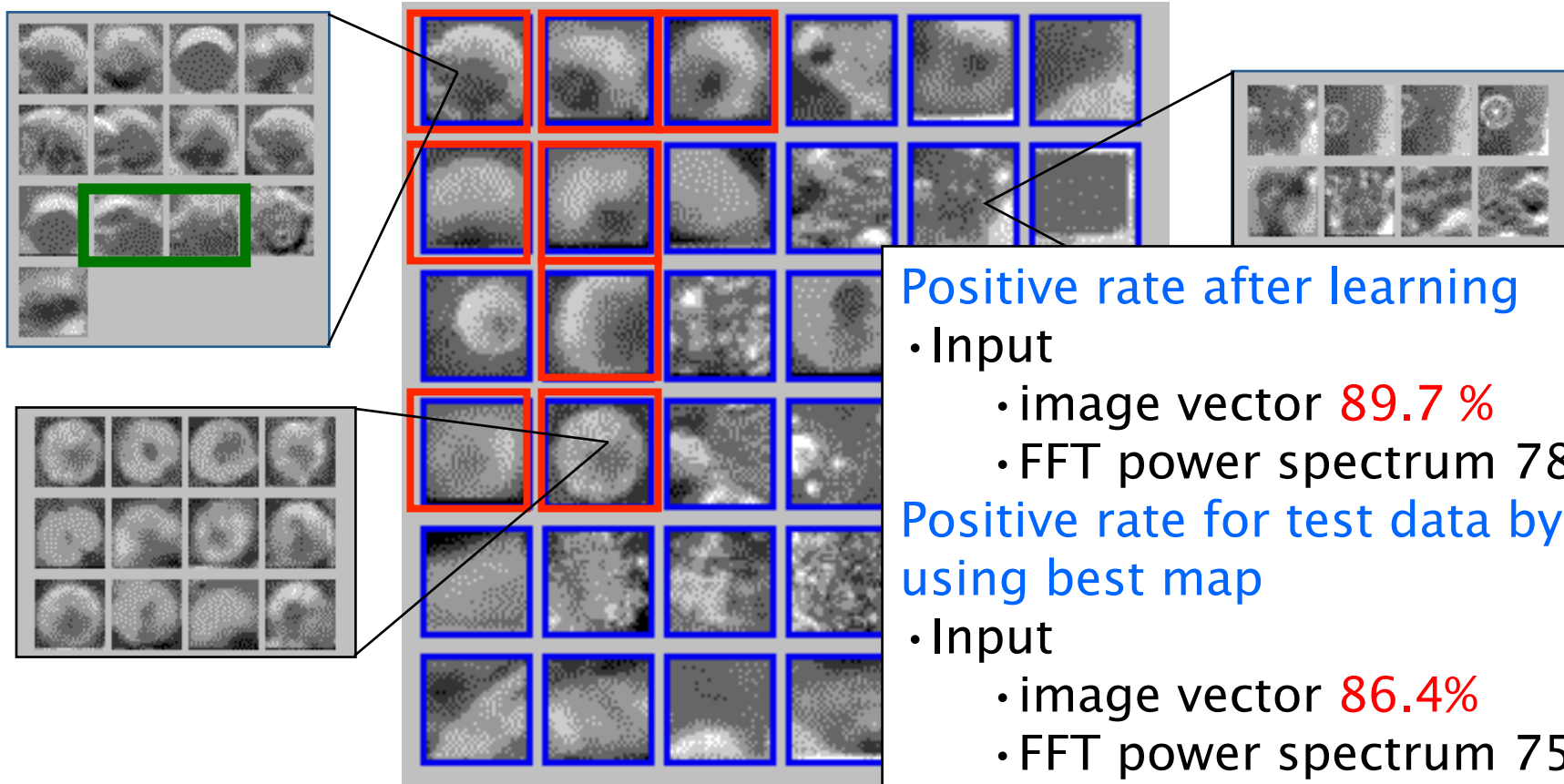
■ (Honda et al. 2000¹, Honda et al. 2001³, Honda et al. 2002⁵)



Result: Learning of true/false candidates by SOM

True > 50%

False > 50%



Positive rate after learning

• Input

- image vector **89.7 %**
- FFT power spectrum 78.4%

Positive rate for test data by using best map

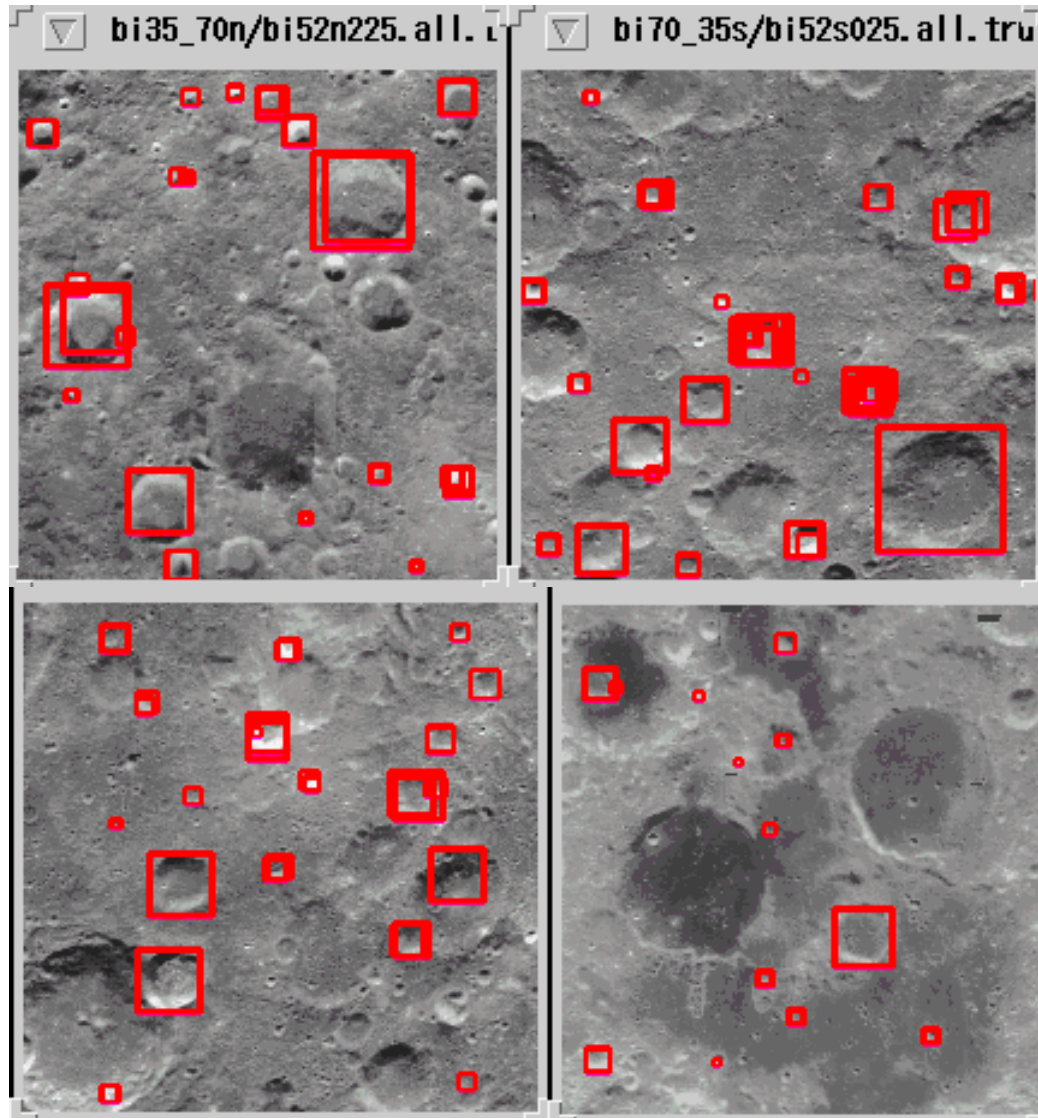
• Input

- image vector **86.4%**
- FFT power spectrum 75.5%

Neighborhood distance 3, iteration: 52500
Number of data: 323

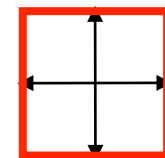
Example of results

Cluster ID
11 15
6 7



Three- level
hierarchical
Images

Radius 9-72
pixels



Maximum
target crater size

衛星画像からのデータマイニングについて の研究事例

パターン学習

(a)惑星画像からのクレーター抽出

(Honda et al. 2000¹, Honda et al. 2001³, Honda et al. 2002⁵, Burl et al., 2005¹⁰)

(b)マルチバンド画像による惑星の地質判定 (横田 2003⁷, Honda et al. 2003⁸)

時空間データマイニング

(c)気象画像からの時空間パターンの検出 (Honda et al. 2000², Honda et al. 2001⁴, Honda et al. 2002⁶, 石津 2009¹³)

(d)地球観測衛星による植生データからの時空間変動 (Honda 2006¹¹, 吉岡、本田2008¹², 山下2009¹⁴)

気象画像からの時空間変動パターンの検出

(Honda et al. 2000², Honda et al. 2001⁴, Honda et al. 2002⁶, 石津2009¹³)

- 時系列気象画像の時空間の変動パターン

- 雲の空間分布

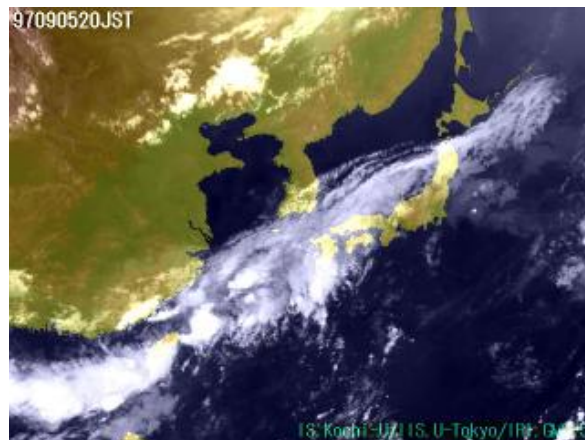
- 前線、台風、筋状雲などの典型的なパターンが存在

- 時間変動

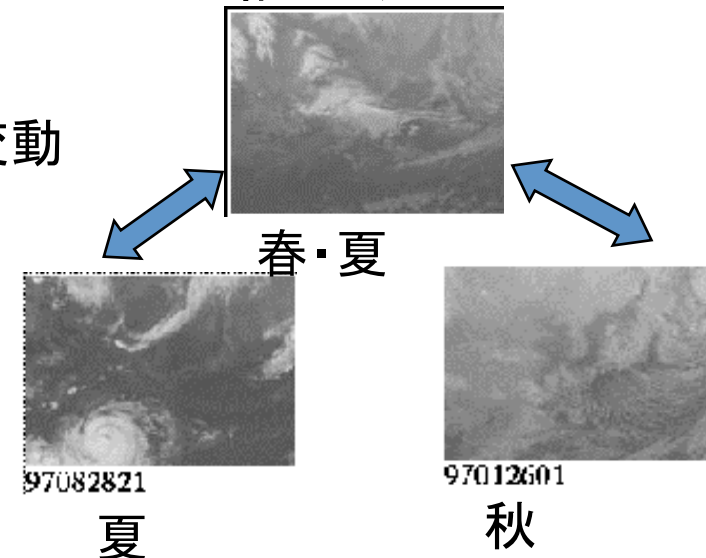
- 特定の雲パターンの発生時期、時間変動にも規則性

- このような時空間の変動パターンを自動要約・発見する要素技術を研究し、研究者の知識発見を支援システムを構築する

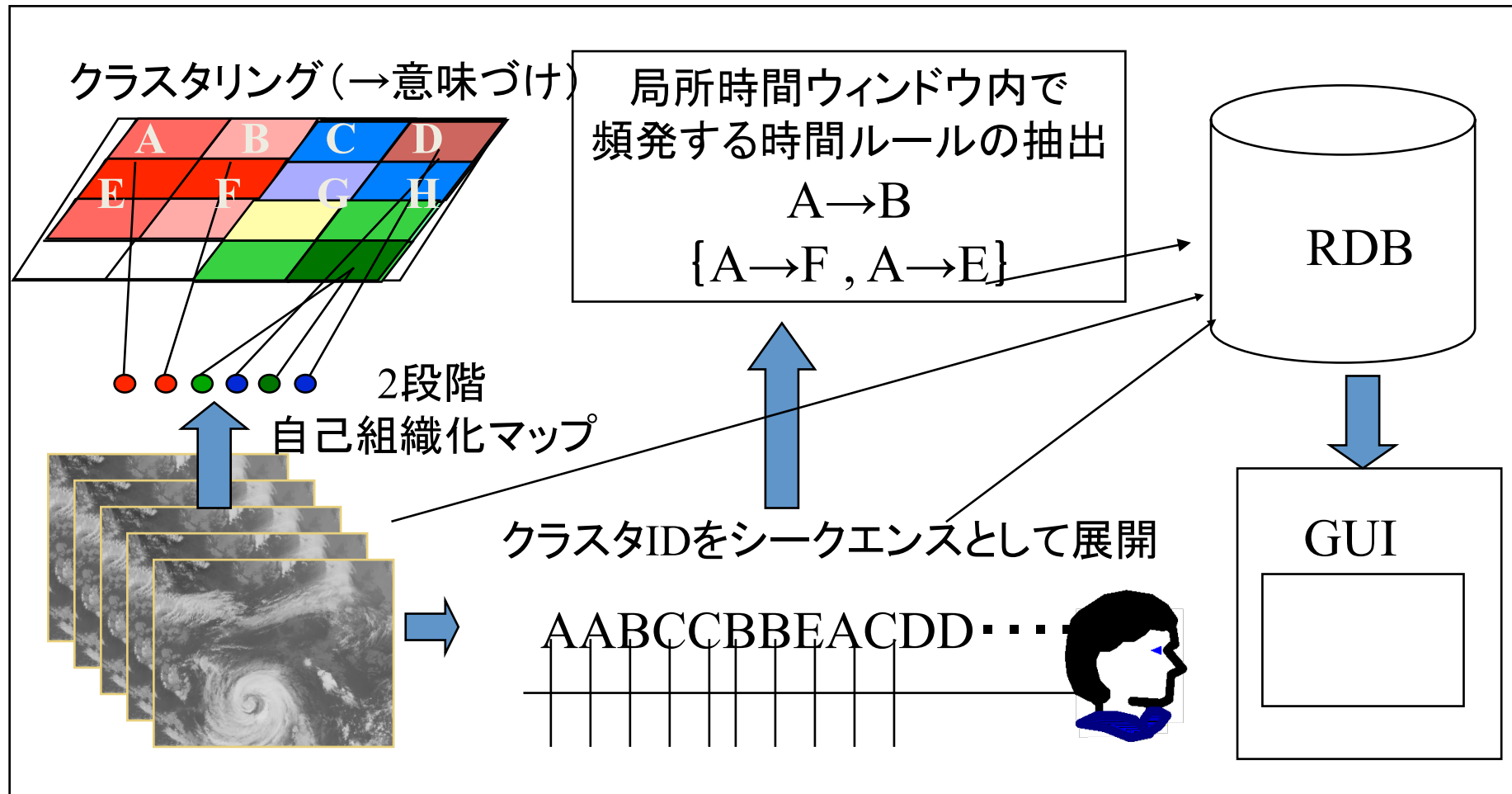
短期変動

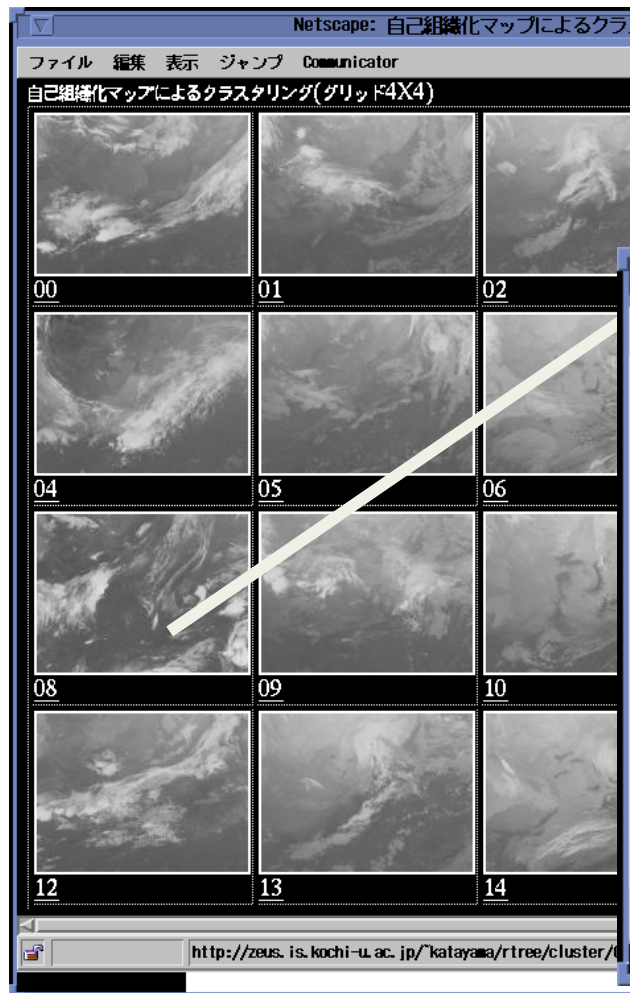
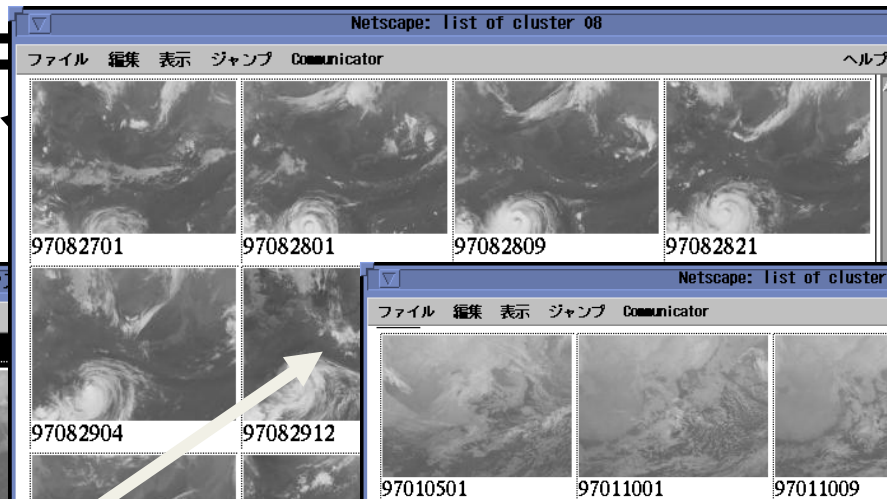


長期変動



時系列気象画像に対する知識発見支援システム (Honda et al. 2000², Honda et al. 2001⁴)



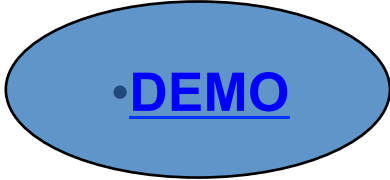


Netscape: クラスタの特徴の説明					
ファイル 編集 表示 ジャンプ Communicator					ヘルプ
クラスタ	季節	目立つ特徴	クラスタ	季節	目立つ特徴
00	春夏	前線、台風	08	夏	太平洋高気圧、前線
01	春秋	西高東低	09	春から夏	梅雨前線、台風
02	春秋	西高東低	10	冬	冬型、渦状の雲
03	春秋	帯状の高気圧	11	冬	冬型、渦状の雲
04	秋	移動性高気圧	12	夏	太平洋高気圧、台風
05	春秋	前線	13	春夏	高気圧、台風
06	秋から冬	筋状の雲	14	冬	寒冷前線
07	春秋	帯状の高気圧	15	春、秋	移動性高気圧

86.0(876/1022)	84.6(876/1044)
----------------	----------------

デモ

- 生画像、画像のクラスIDとその意味付け、抽出した時間ルールをRDBに統合することにより、下記のような意味的質問に答えるシステムが実現できる：
 - “秋に、台風が通過した後、前線が発達するような気象パターンは存在するか。またそれはどのようなものか?”



• DEMO

デモ

Time-sequential retrieval of weather images

Select the option for the image sequence you want to see.

spring

summer

autumn

winter

January

February

first event

front

typhoon

HP

LP

HP in W LP in W

band-like HP

time

front

typhoon

HP

LP

HP in W LP in E

band-like heigh pressure

next event

submit

clear

Retrieval result

97 09 01 03 – 97 09 11 03
97 10 19 01 – 97 10 26 01
97 10 19 01 – 97 10 28 01
3 cases are hit

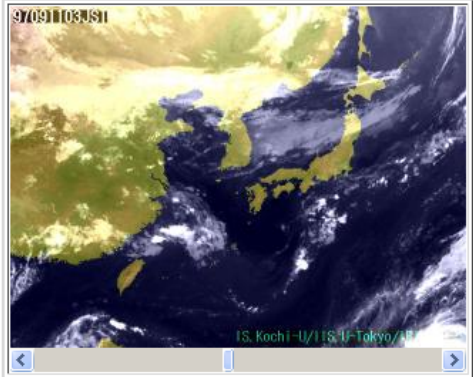
97 09 01 03- 97 09 11 03

submit

clear

Select the case you want to browse.

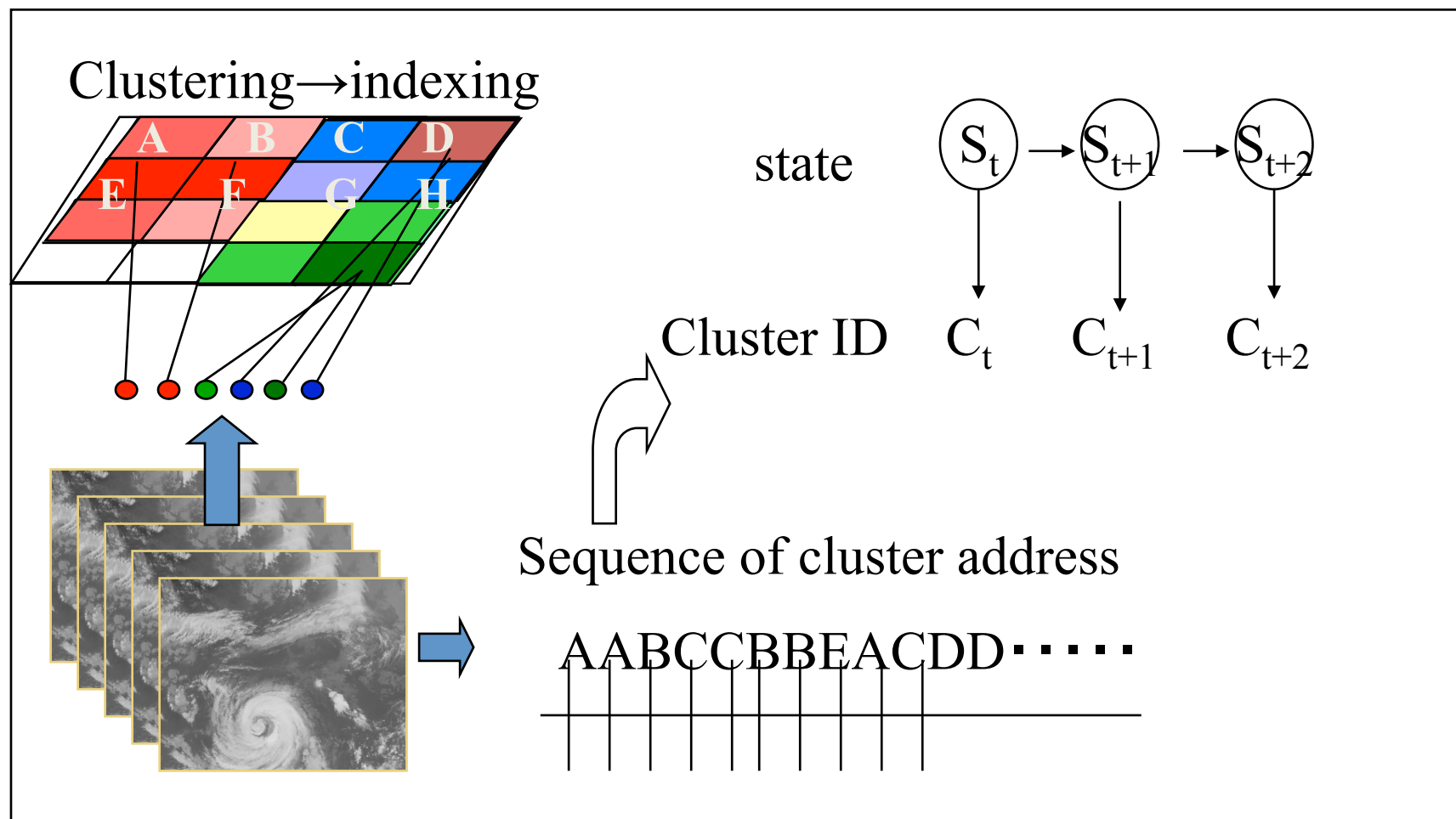
Weather images as an animation



You can control the speed of animation. Left: faster, Right slower

(Honda et al. 2000², Honda et al. 2001⁴)

隠れマルコフモデルによる長期変動の調査 (本田ほか 2005⁹)

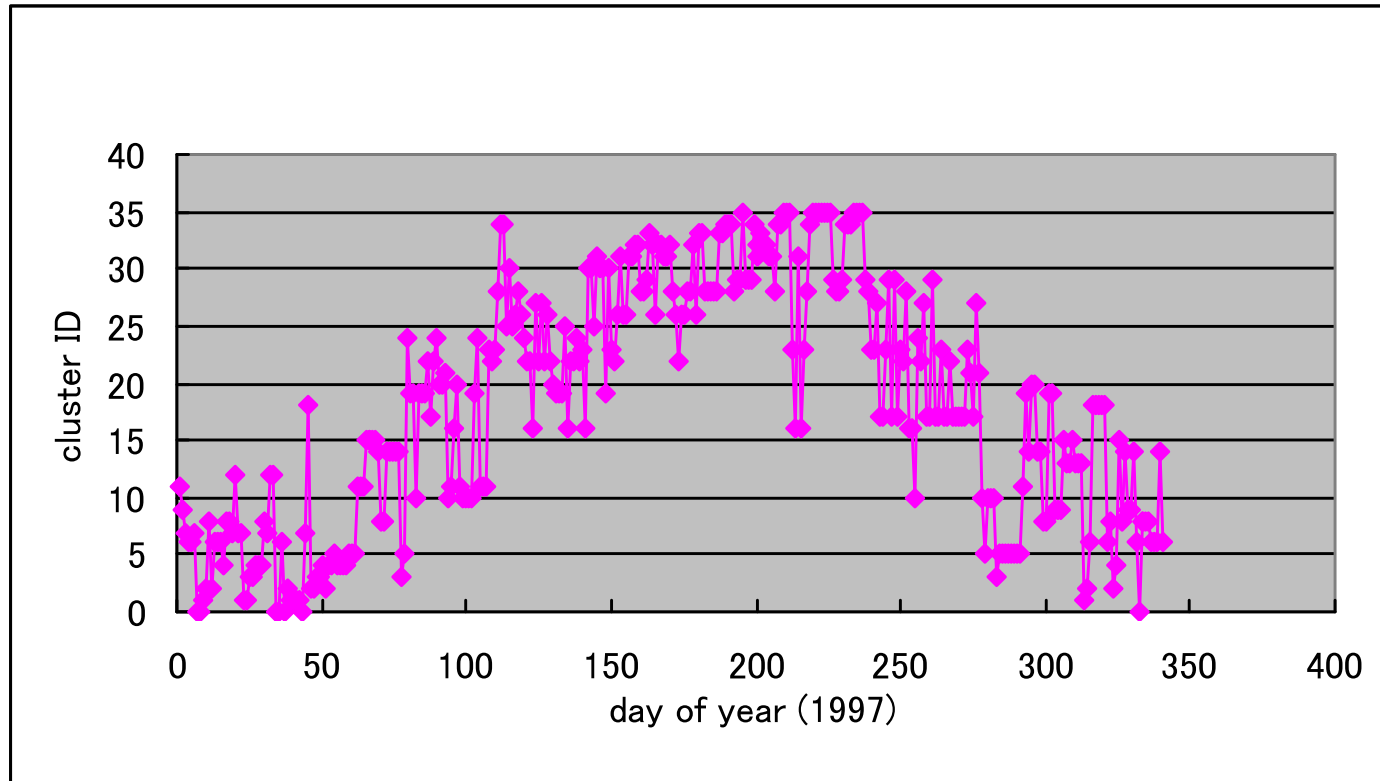


実験

- 1997－2003年のGMS5の日本周辺の雲画像¹⁵を隠れマルコフモデルによりモデル化
- Baum Welchアルゴリズム, EMアルゴリズムによってパラメータ推定
- 入力記号系列: クラスタの記号系列を 28日ごとにセグメント化[95個]
- 状態数は2～8の変動パラメータとする。

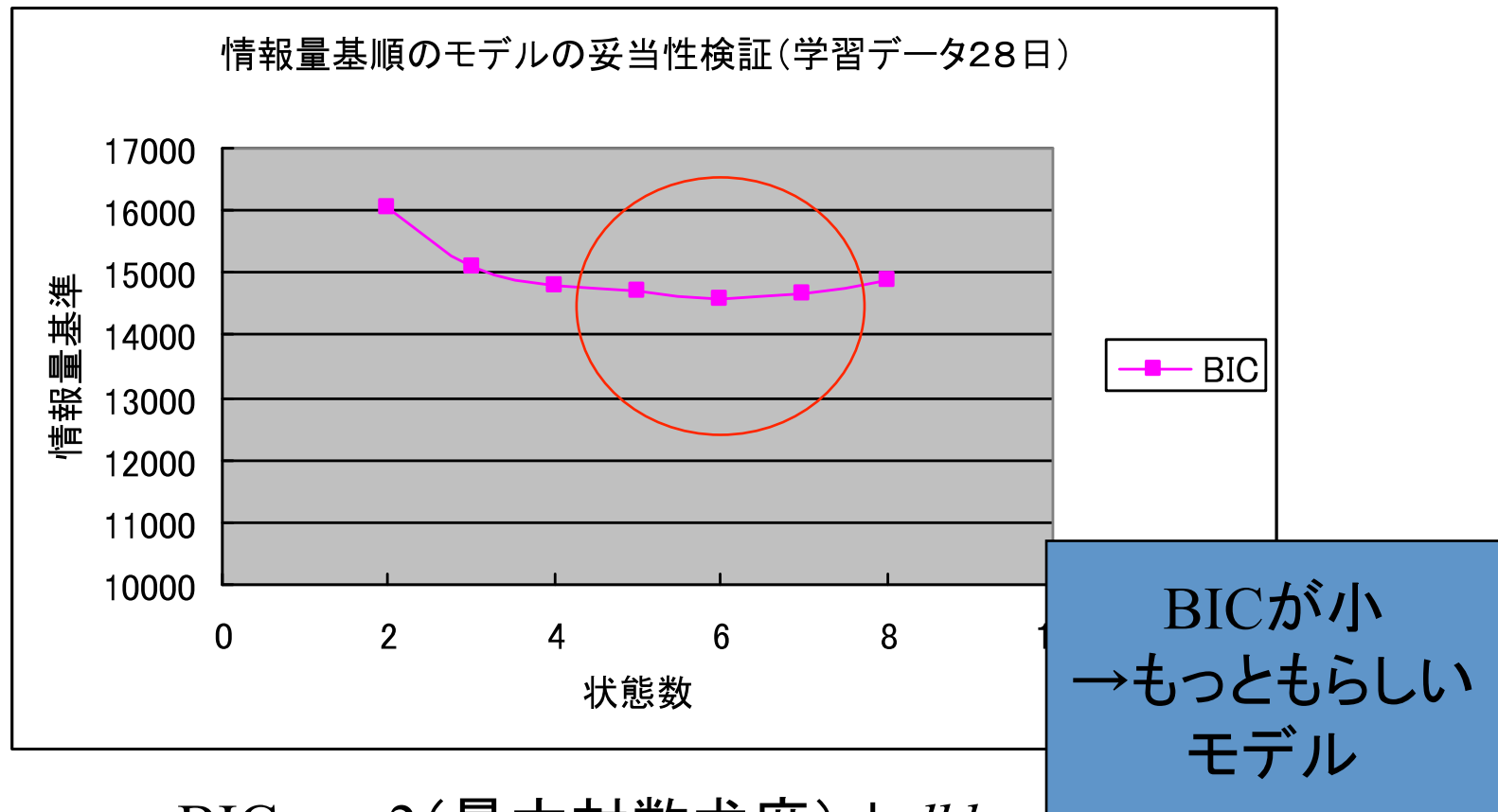
↓
モデルのパラメータの推定(状態遷移確率, 記号出力確率, 初期状態確率) → BICを利用

Example of time variation of cluster ID



Dataset for 4 years(1997-2000) used for leaning

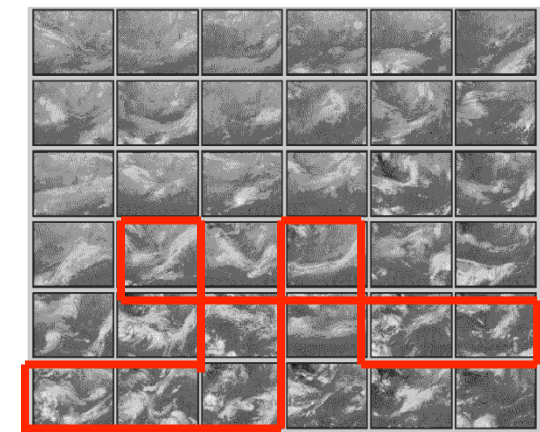
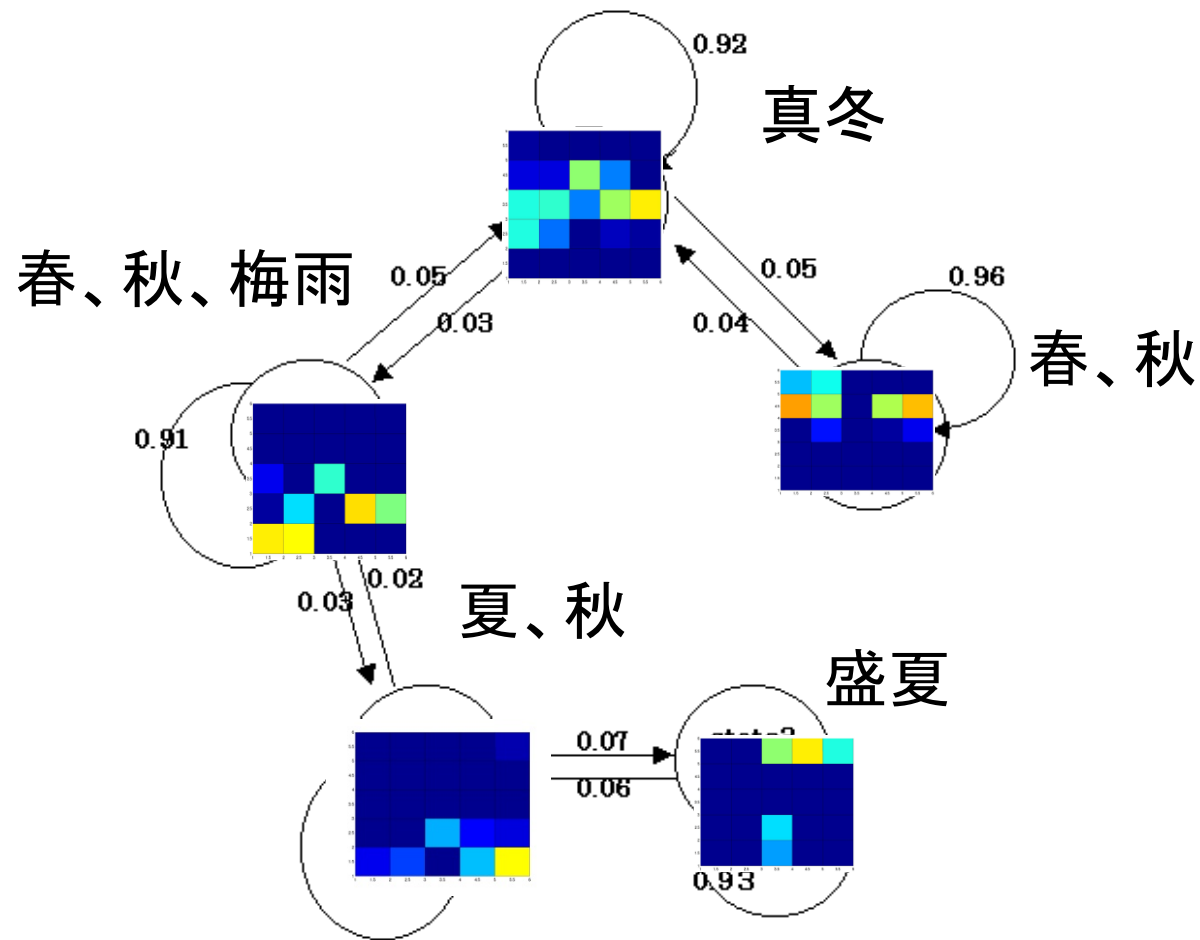
BIC(ベイズの情報量基準)による状態数選択



$$\text{BIC} = -2(\text{最大対数尤度}) + dk \log n$$

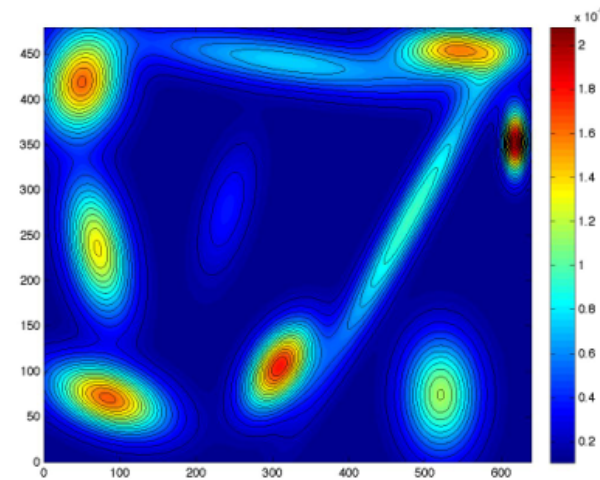
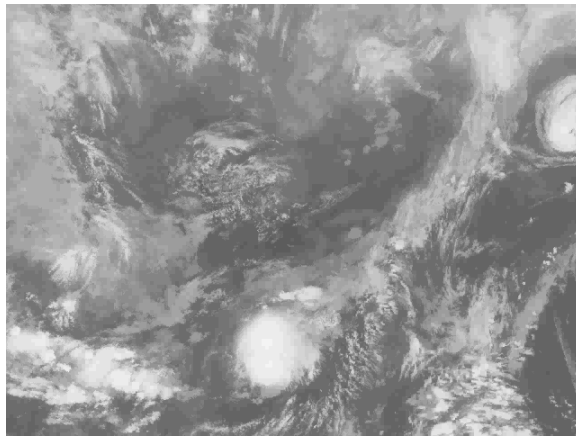
(dk :モデルの自由度 n :出力記号数)

実験結果 1



時間連続性を考慮した混合分布推定による オブジェクト追跡 — 一時系列気象画像への適用 — (石津2009¹³)

- 時空間データの変動の分析は、時空間の中の**オブジェクト** (対象) を抽出・追跡する問題ととらえる事ができる
 - 要約したオブジェクト情報を時空間変動についての知識発見に役立てる

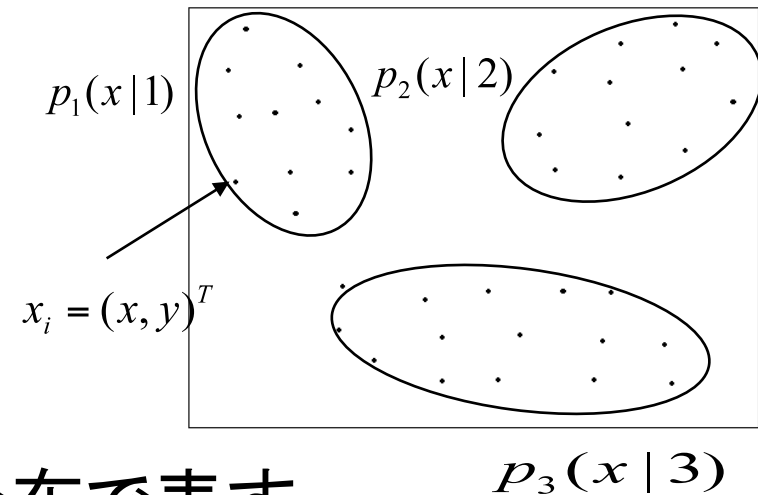


研究の目的

- 不特定数・不特定形状のオブジェクトを時間連続性を考慮しながら自動的に抽出、追跡する手法を開発し、時空間変動パターンなどの2次的な知識発見に活用する。
- この際、単なる連続的な変化だけでなくオブジェクトの分裂，消滅（吸収）も考慮に入れる
- 求められた解の可視化によって時空間変動のビジュアルデータマイニングを試みる

モデル

- オブジェクトが存在する点の座標集合を $\{(x_i, y_i) \mid i = 1, \dots, N\}$ とする



- 各オブジェクトを多変量正規分布で表す

$$p(x|j) = \frac{1}{(\sqrt{2\pi})^2 \sqrt{|\Sigma_j|}} \exp \left\{ -\frac{1}{2} (x - \mu)^T \Sigma_j^{-1} (x - \mu) \right\}$$

$$x = \begin{bmatrix} x \\ y \end{bmatrix}, \mu_j = \begin{bmatrix} \mu_{xj} \\ \mu_{yj} \end{bmatrix}, \Sigma_j = \begin{bmatrix} \sigma_{xj} & \sigma_{xyj} \\ \sigma_{yxj} & \sigma_{yj} \end{bmatrix}$$

μ : オブジェクトの中心
 Σ : オブジェクトの広がり

- オブジェクト集合を多変量正規分布の混合分布で表現する

$$P(x) = \sum_{j=1}^O \omega_j p(x|j)$$

ω_j : 重み係数
 O : 成分数

オブジェクト分布を求める
 = モデルパラメータを求める

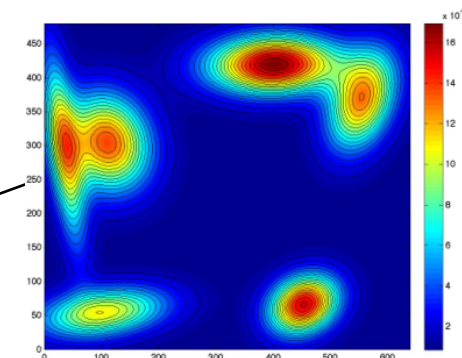
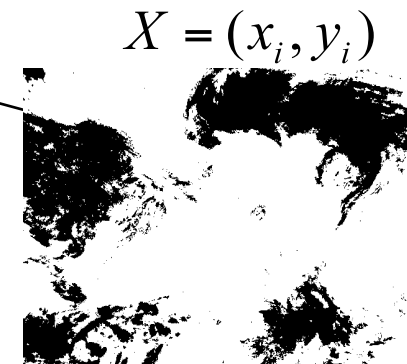
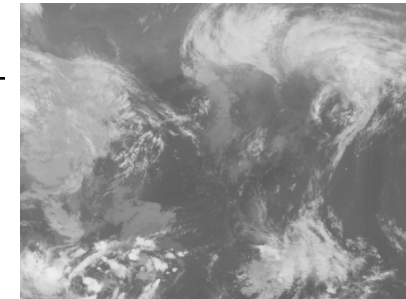
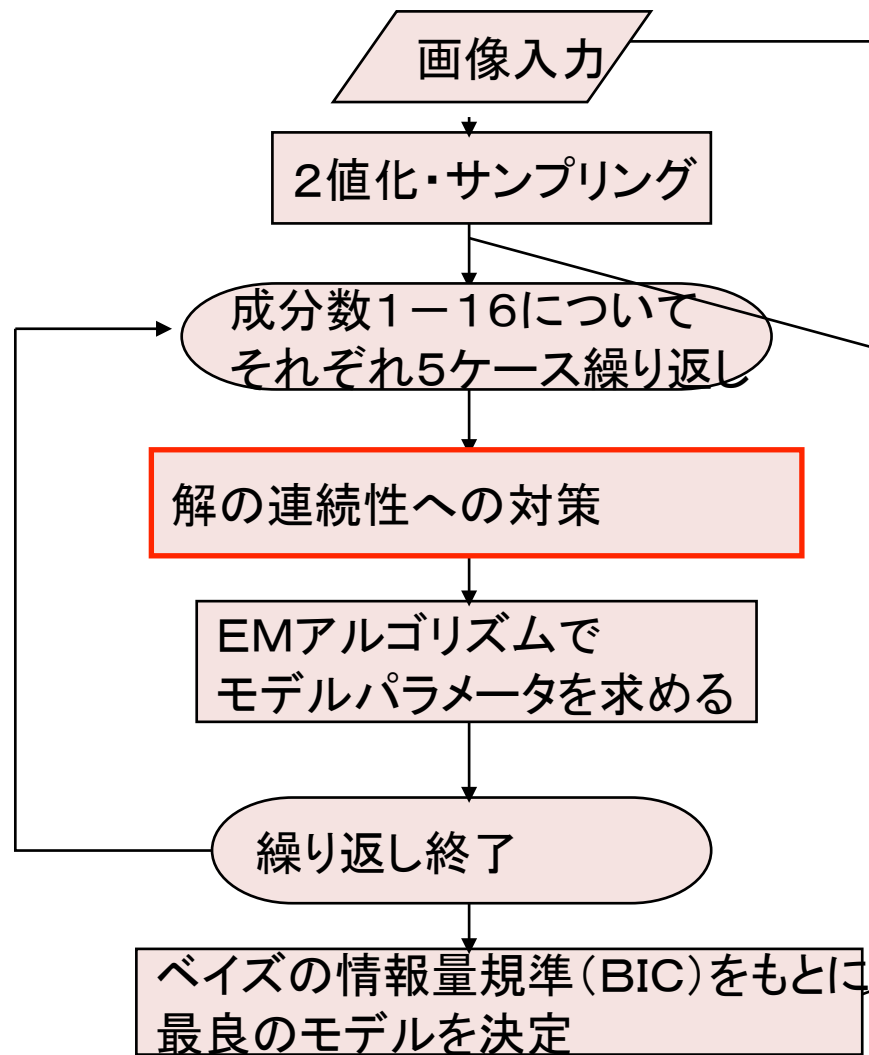
具体的な解法

- モデルパラメータの決定
 - EMアルゴリズム
- 考慮すべき点
 - 成分数が未知
 - 可能性のある成分数の範囲を総当り試行
 - 大局的最適解が保障されない
 - K-平均法で求めた初期値を複数回与える
 - BICを用いて最適なモデルを決定
 - $$\text{BIC} = -2(\text{最大対数尤度}) + (6K - 1)\log N$$

k: 成分数 N: データ数

解法

MTSAT-IR1
2007年9月の画像¹⁵



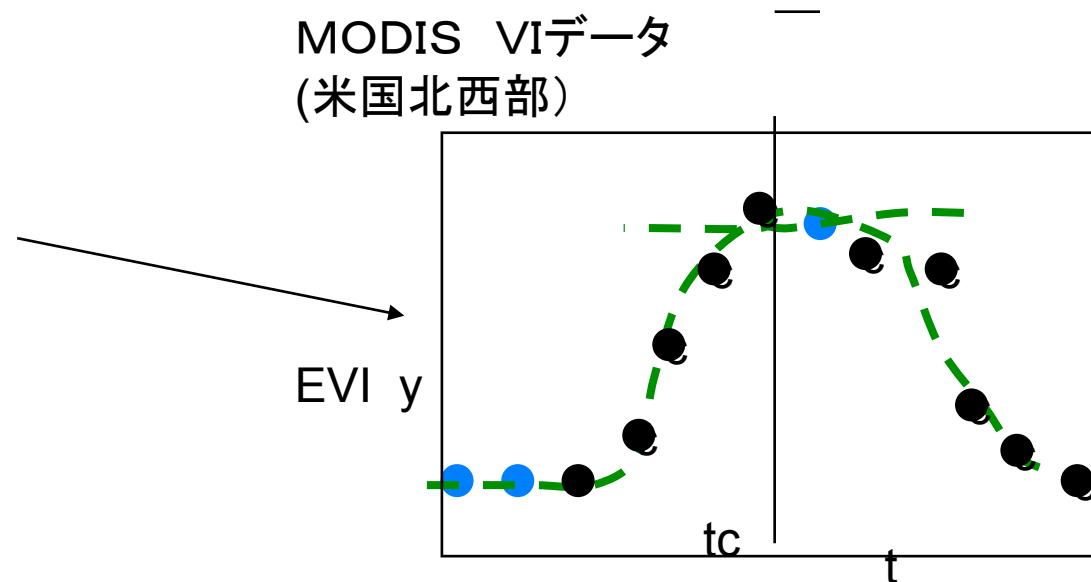
結果

- 解の引継ぎとラベル付けを導入することによって
時間的に連続した混合分布推定が可能になった
 - オブジェクト追跡、ならびに精度を保ったまま計算時間を短縮することが可能に

今後の展望

- 雲の状態変化の新規発生に対応する
- オブジェクトのテクスチャなどによるラベルづけ
- 高次の知識発見への利用

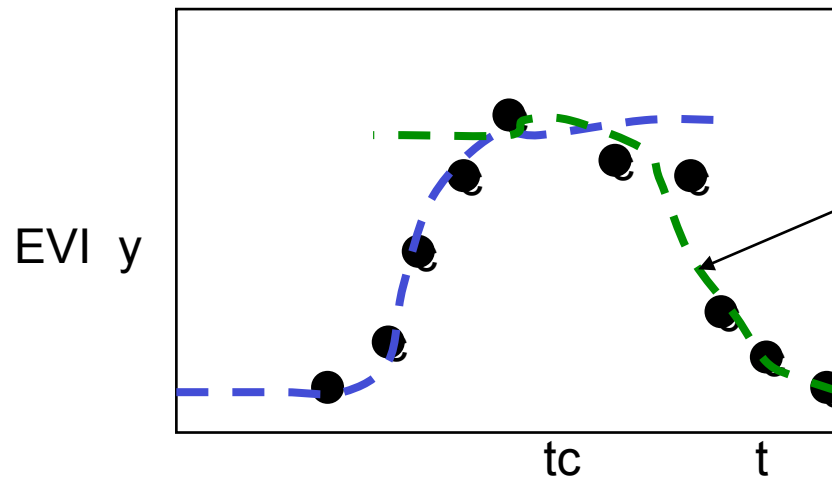
地球観測衛星による植生データからの時空間変動 (Honda 2006¹¹, UCI, ボストン大共同研究, 吉岡, 本田2008¹², 山下2009¹⁴)



- VI(植生指標)の時空間の変動パターンをモデリング
 - 変動パターンの自動要約、予測、変動発見に活用できる
 - 問題
 - データの欠落(上左図で青)、ノイズの影響で、事前の領域分割、補完などの恣意的なデータ操作なしではモデル化が難しかった。
 - データマイニング的アプローチは困難

地球観測衛星による植生データからの 時空間変動モデリング

- 確率密度分布を用いたアプローチ
 - 観測値を区分化ロジスティック関数の周りのガウス分布でモデル化



$$F(t_i | \theta) = \begin{cases} f_1(t_i | \theta_1) & t < t_c \\ f_2(t_i | \theta_2) & t \geq t_c \end{cases}$$

$$f_j(t_i | \theta_j) = \frac{c_j}{1 + e^{a_j + b_j t_i}} + d_j$$

- 事後確率を最大化するMaximum a Posterior (MAP) 法によるモデルパラメータ推定
 - 恣意的な操作なしでも安定なフィッティングを行うことが可能に

Method - Maximum a Posterior (MAP)

Estimation - $D = [y(t_1), y(t_2), \dots, y(t_m)]$

- Equations.

$$P(D | \theta) = \prod_{i=1}^m N(F(t_i | \theta), \sigma),$$

$$P(\theta) = \prod_{i=1}^n N(\mu_{\theta_i}, \tau_i) \quad \text{Prior}$$

$$S(\theta) = \log \underline{P(D | \theta)P(\theta)} \leftarrow \propto P(\theta | D)$$

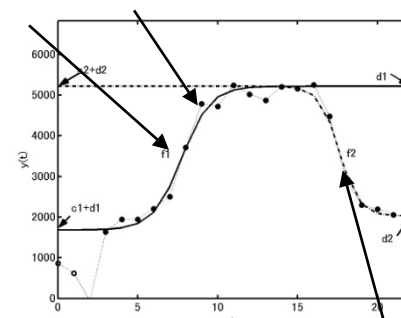
$$\hat{\theta} = \arg \max_{\theta} S(\theta)$$

- Prior parameters θ are given explicitly.

μ_{θ_j} : Mean of ML solutions for many pixels.

τ_i : Standard deviation of ML solutions for many pixels.

観測



モデル $F(t | \theta)$

Posterior Probability

まとめ

- 大規模データに対する知識発見支援システムの構築例、コアプロセスの検討例を示した
 - 欠落、ノイズへの耐性、モデルの不定性
 - 統計学的モデリングが有効(情報量基準によるモデル選択)
 - 時空間変動
 - 時間相関ルール、隠れマルコフモデル、オブジェクト追跡などの多角的なアプローチ
- 衛星画像に対する本格的なデータマイニングは個人レベルでは現実的には実施困難
 - アーカイブを持つセンターなどでの系統的な取り組みが重要

Reference

1. Honda R., O. Konishi, R. Azuma, H. Yokogawa, S. Yamanaka, Y. Iijima, Data mining system for planetary images - crater detection and categorization - , Proceedings of the International Workshop on Machine Learning of Spatial Knowledge in conjunction with ICML, Stanford, CA, 103-108, July, 2000.
2. Honda, R., H. Takimoto, and O. Konishi, Semantic indexing and temporal rule discovery for time-series satellite images, Proceedings of the International Workshop on Multimedia Data Mining in conjunction with ACM-SIGKDD Conference, Boston, MA, 82-90, 2000.
3. Honda R., Y. Iijima, and O. Konishi, Mining of topographic features from large scale planetary imagery, Proceedings of Fourth Workshop on Mining Scientific Datasets in conjunction with 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco CA, 21-27, August 2001.
4. Honda R., and O. Konishi, Temporal rule discovery for time-series satellite images and integration with RDB, Principles of Data Mining and Knowledge Discovery: Fourth European Symposium, 204-215, September 2001.

Reference

5. Honda R., Y Iijima, and O. Konishi, Mining of topographic feature from heterogeneous imagery and its application to lunar craters, Progress of Discovery Science, LNAI 2281, Springer Verlag, 395-407, Berlin Heiderberg 2002
6. Honda R., S. Wang, T. Kikuchi, and O. Konishi, Mining of objects from time-series images and its application to satellite weather imagery, Journal of Intelligent Information Science, 19:1, 79-93, 2002
7. 横田康弘, 可視・近赤外波長域における月面の光反射特性, 東京大学博士論文, 2003
8. Honda R., Y. Yokota., Y. Iijima, H. Mizutani, Multi-spectral lunar imagery mining including precise photometric correction and spectral spatial clustering, Proceedings of Workshop on Mining Scientific and Engineering Datasets in conjunction with SIAM international conference on Data Mining, 29-38, 2003

Reference

9. 本田理恵, 勝吉進一, 小西修, 隠れマルコフモデルによる気象画像の時空間変動パターンモデリング, 地球惑星関連学会合同学会(アブストラクト), 2005
10. Burl M., C. Chapman, B. Enke, R. Honda, W. Merline, P. Wetzler, Learning to detect small impact craters, Seventh IEEE Workshop on Applications of Computer Vision, Breckenridge CO, January 2005
11. Honda R., Temporal modeling and missing data estimation for MODIS vegetation data, The 2nd NASA data mining workshop., 4pp., 2006
12. 吉岡和浩、本田理恵, 地球観測衛星による植生指標データからの長期年間変動のモデリング-ML法、MAP法を用いて-, 地球惑星関連学会合同学会(アブストラクト), 2008
13. 石津光洋, 時間連続性を考慮した混合分布推定によるオブジェクト追跡一時系列気象画像への適用一, 高知大学理学部数理情報科学科卒業論文, 2008

Reference

- 14 山下智也、植生指標の時空間変動の統計学的モデリングー GIMMS NDVIデータによる検証 ー, 高知大学理学部数理情報科学科卒業論文, 2008
- 15 高知大学気象情報頁, <http://weather.is.kochi-u.ac.jp/>