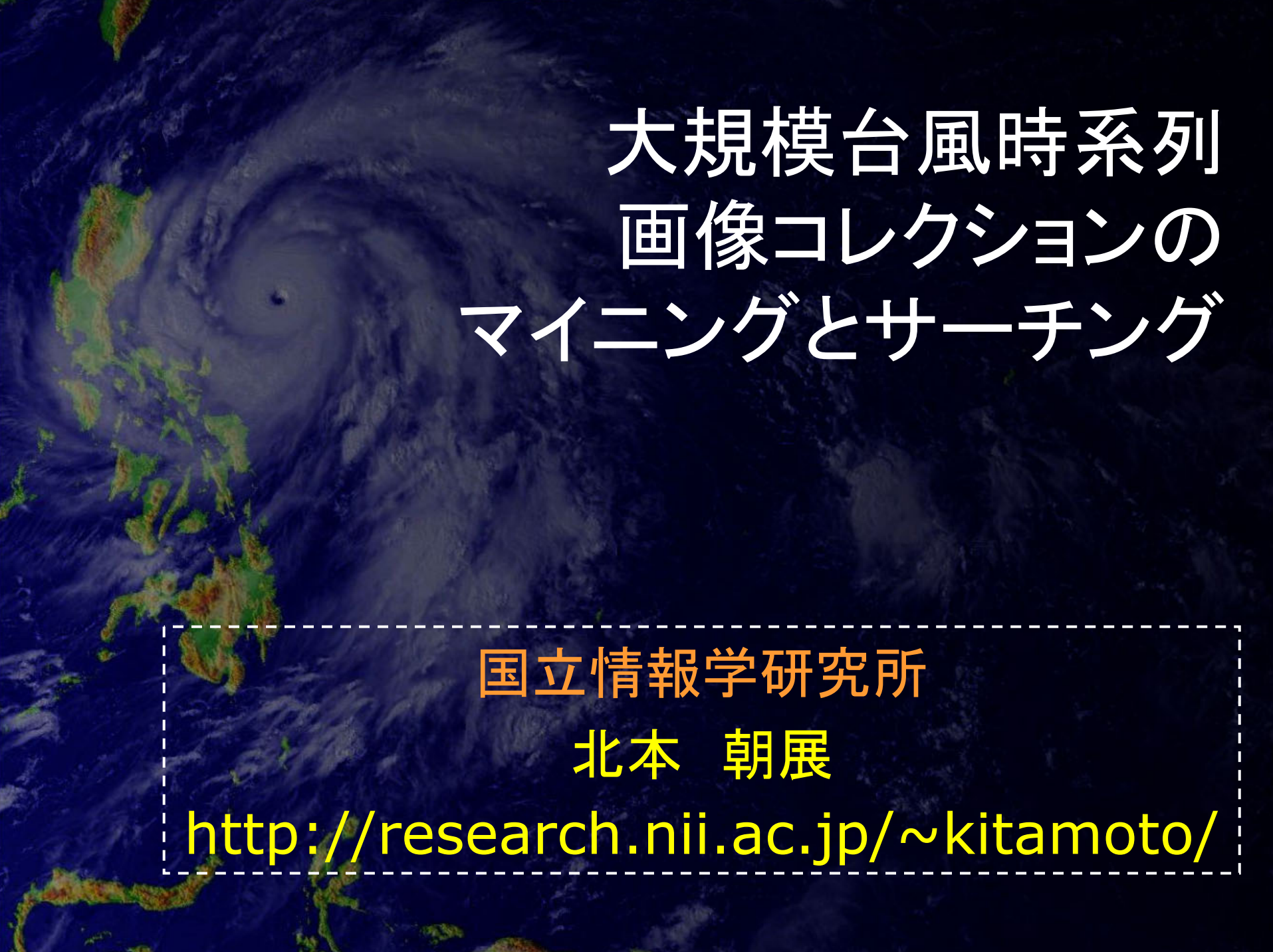


A satellite image of a typhoon with a clear eye, swirling over the Philippines. The text is overlaid on the right side of the image.

大規模台風時系列 画像コレクションの マイニングとサーチング

国立情報学研究所

北本 朝展

<http://research.nii.ac.jp/~kitamoto/>

発表の概要

1. はじめに

2. データコレクションを基盤とする研究スタイル

3. 台風時系列画像コレクション

4. サーチング

5. マイニング

6. おわりに

地球科学データの洪水

- 地球観測衛星による観測データが爆発的に増大している。
 1. 高分解能化（より細かく）
 2. 多波長化（よりいろいろな波長で）
 3. 高頻度化（より頻繁に）
- 地球科学データは、そのほとんどが使われないまま、テープの中に眠っている。
- そこには未知の知識が隠れているはず。

データベース基盤に基づく科学研究

- 生物情報学・バイオインフォマティクス
 - ーヒューマンゲノム計画
- 天文学・宇宙科学
 - ースカイサーベイ計画

科学の様々な分野において、**大規模データベース基盤の整備と活用**が、研究における必須の要素となりつつある。

背景

- センシング技術およびデータ蓄積技術の発達により、従来よりも格段に大量のデータが生成できるようになった。
 - 網羅的な測定に基づき、測定対象をセンサ上ではなくコンピュータ上で選択する考え方が一般的となった。
1. 大量データ管理のためのデータベース
 2. 大量データ分析のためのマイニング

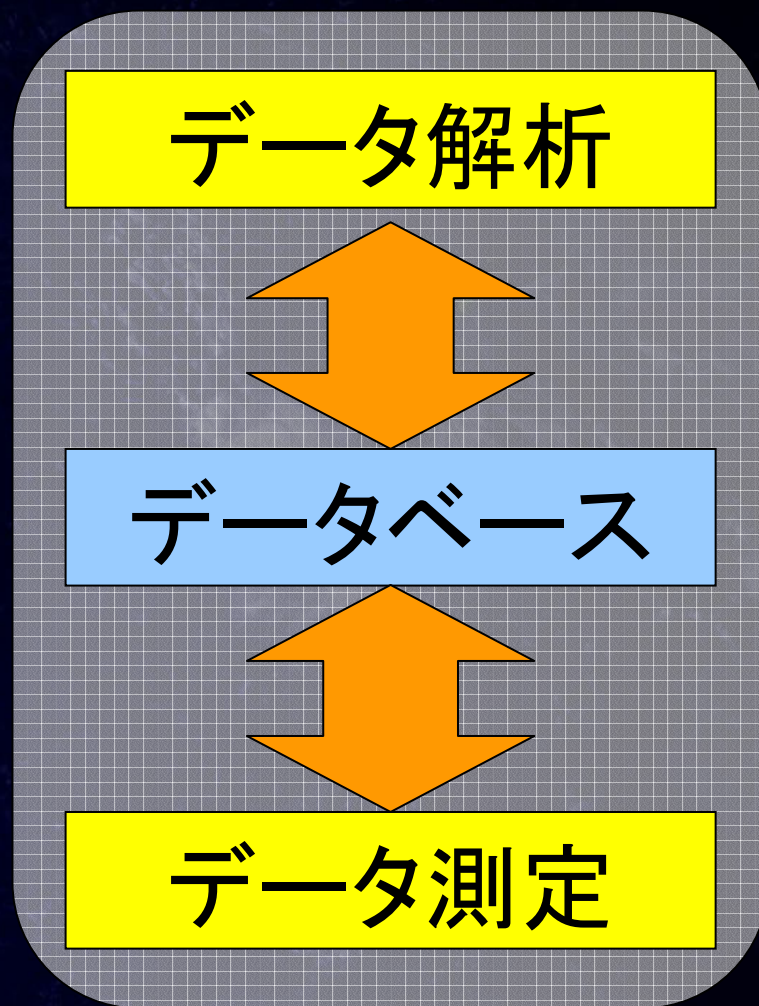
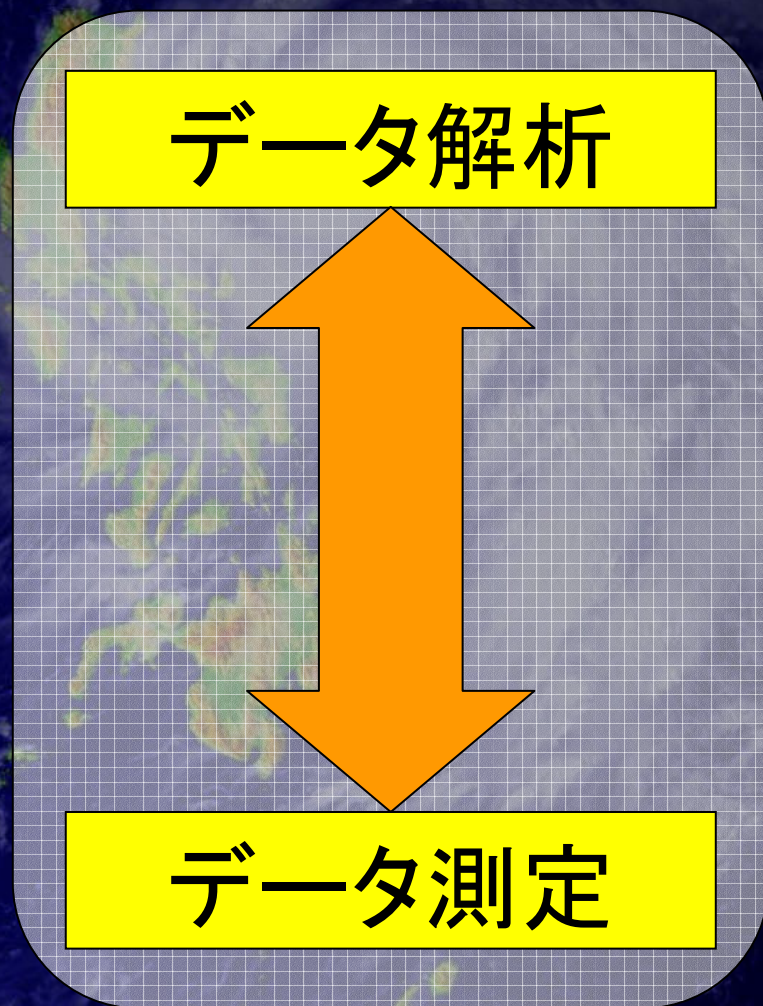
本研究の目的

1. データベース基盤に基づく科学研究の一例として、地球科学データ、特に気象学的に重要な現象である「台風」を選ぶ。
2. 台風に関する網羅的なデータベースに基づく研究はほとんど前例がないため、情報学の実世界問題として挑戦的課題である。
3. 得られる知見を気象学にフィードバックする。
4. 本研究で確立する方法論を他の分野(生物情報学など)に応用する。

発表の概要

1. はじめに
2. データコレクションを基盤とする研究スタイル
3. 台風時系列画像コレクション
4. サーチング
5. マイニング
6. おわりに

階層モデル



科学目的の場合、データベースは公共的。

公共データベースの利点

1. 「データ収集者」の媒介により、データ測定者とデータ解析者の分業が可能。
2. 複数人が同一の測定を無駄に反復するのを防ぐ一方で、複数人による繰り返し測定がデータの信頼性を高める。
3. データベーススキーマに基づくデータ表現・データ操作の形式化により、見通しよいアーキテクチャを実現できる。

公共データベースの要件

1. 網羅性

- 解析に必要なデータを事前に測定。

2. 一貫性

- 全データを同等の品質・方法で測定。

3. 汎用性

- データ解析法独立のデータ定義方法。

4. 多様性

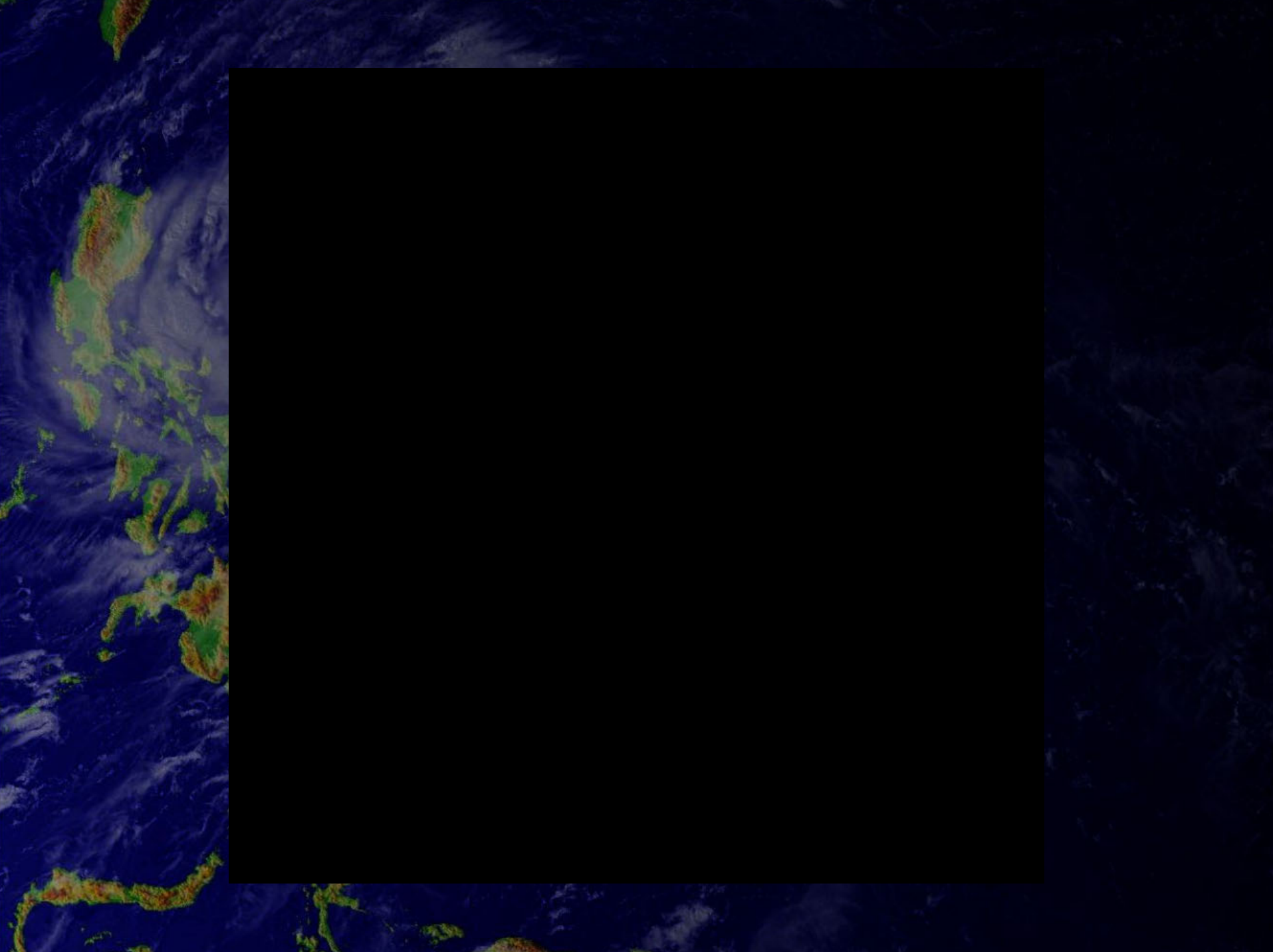
- データ検索法独立のデータ操作方法。

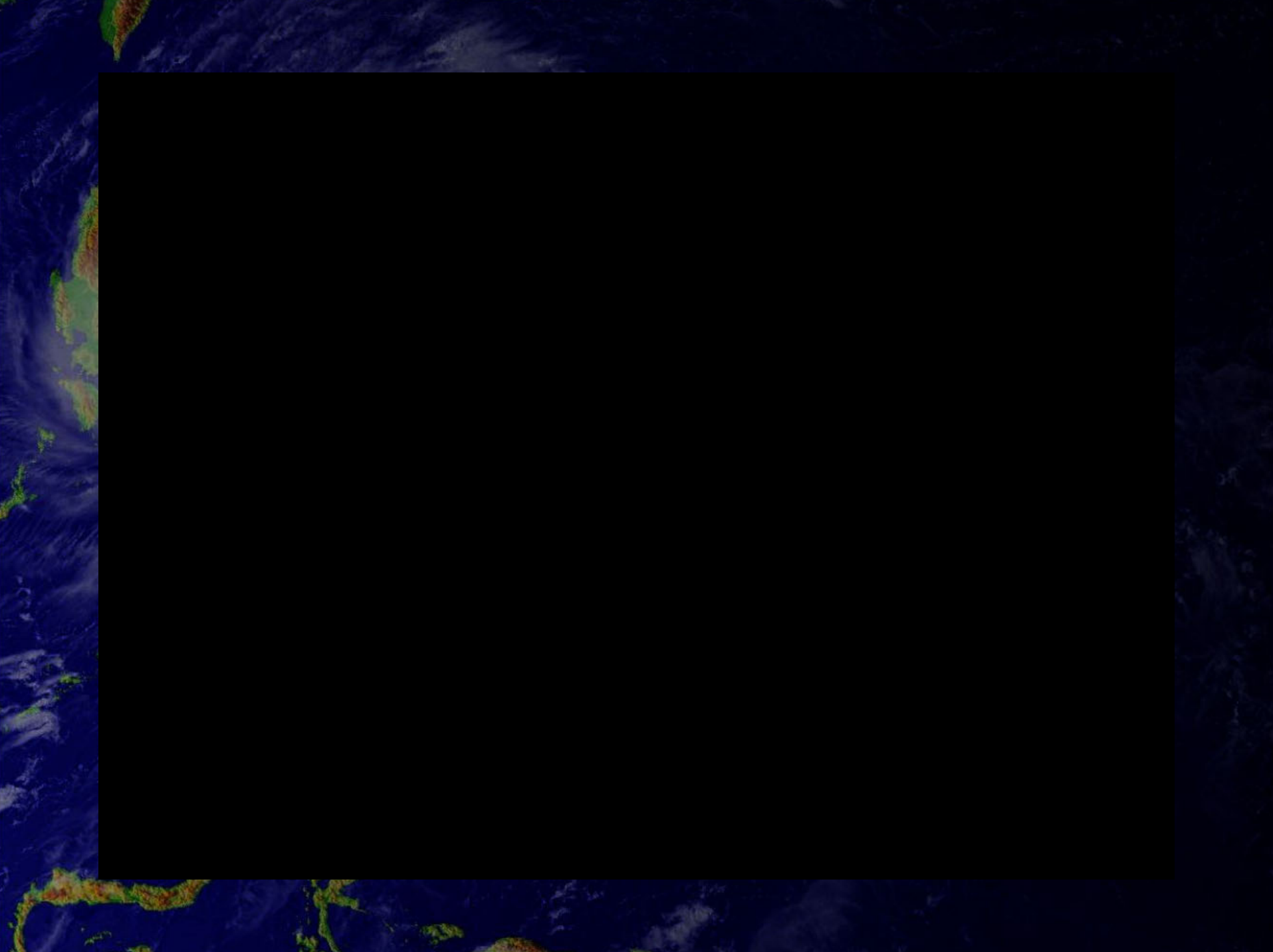
発表の概要

1. はじめに
2. データコレクションを基盤とする研究スタイル
3. 台風時系列画像コレクション
4. サーチング
5. マイニング
6. おわりに

台風時系列画像コレクション

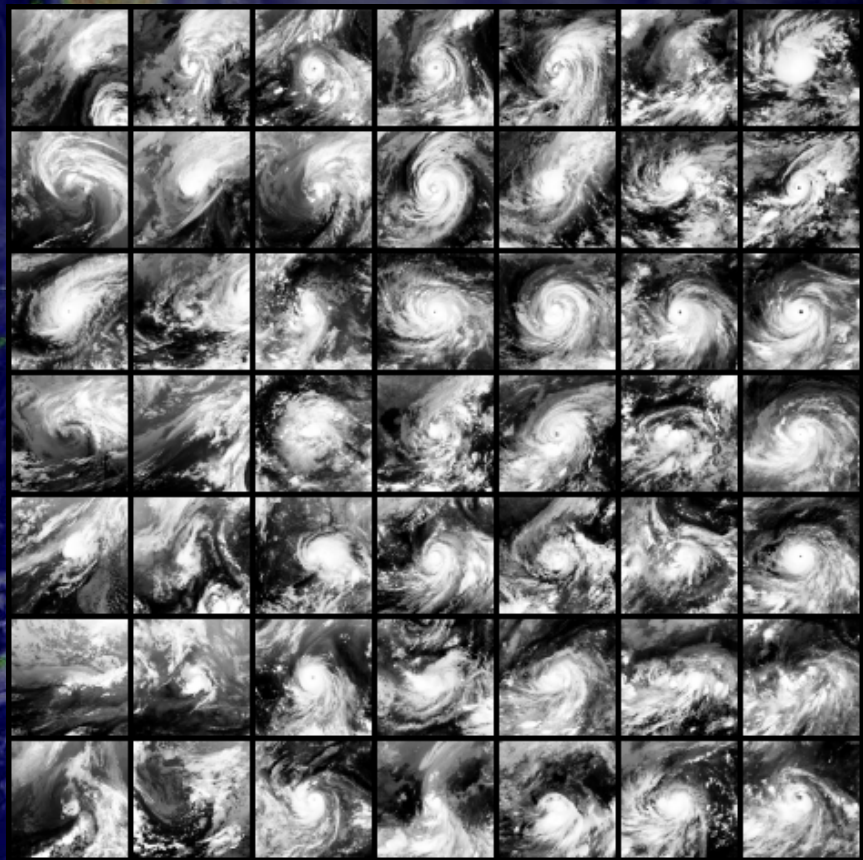
<i>As of Mar. 2003</i>	<i>W.N. Pacific</i>	<i>W.S. Pacific</i>
<i>Best Track Dataset</i>		
<i>Organization</i>	<i>Japan (JMA)</i>	<i>Australia (BOM)</i>
<i>Lat. range</i>	<i>N. of equator</i>	<i>S of equator</i>
<i>Lon. range (E)</i>	<i>100~180</i>	<i>90~170</i>
<i>Typhoon Image Collection (Total 44,136)</i>		
<i>Num of seasons</i>	<i>9 seasons</i>	<i>6 seasons</i>
<i>Num of seq.</i>	<i>188</i>	<i>70</i>
<i>Num of images</i>	<i>33,378</i>	<i>10,758</i>
<i>Images per seq.</i>	<i>53~433</i>	<i>25~513</i>



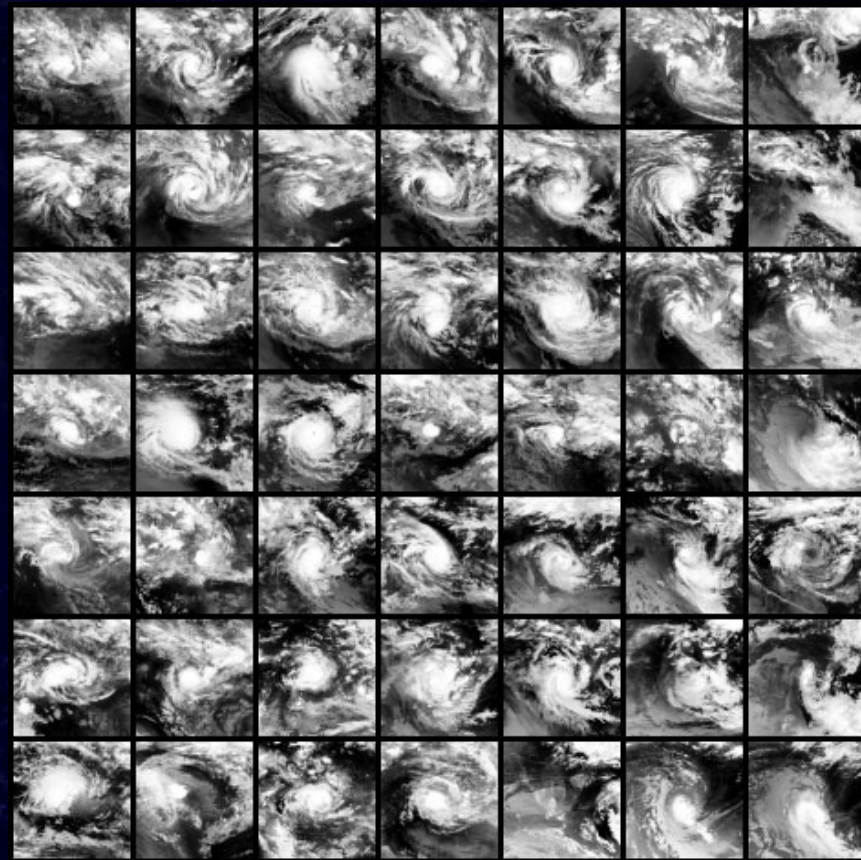


台風雲パターンの多様性

K-means clustering of image collection



Northern Hemisphere



Southern Hemisphere

台風時系列画像コレクションの特徴

- 台風中心を追跡し、画像中心と位置合わせした**枠取り画像**を生成。
- 自然界に存在する**複雑系パターン**の一例として、**カオス的な時間発展**を示す。
- **注釈つき画像データコレクション**として種々の**テストベッド**にも利用できる。
- <http://www.digital-typhoon.org/>

発表の概要

1. はじめに
2. データコレクションを基盤とする研究スタイル
3. 台風時系列画像コレクション
4. サーチング
5. マイニング
6. おわりに

サーチングとマイニング

サーチング : 条件を満たすデータ(タプル)を選択

マイニング : 重要度の高い法則を選択



重要度を判定するためにはサーチングを多数回繰り返す必要がある場合が多い。

台風時系列画像のサーチャング

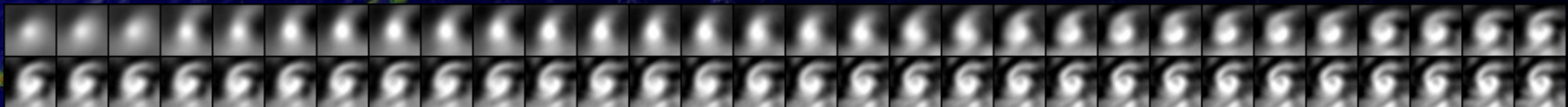
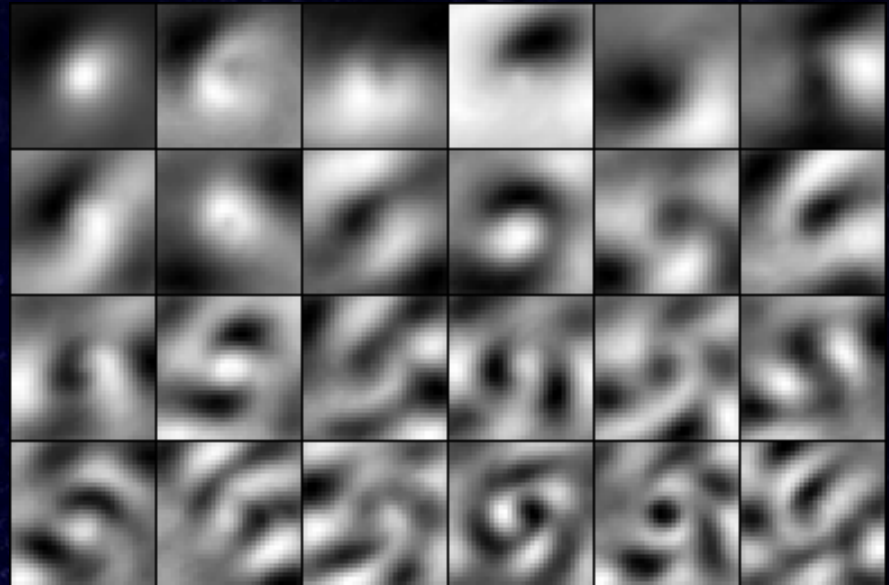
- 事例ベース推論における類似事例検索のためにサーチャングが重要。
- データマイニングプロセスにおける「部品」としてサーチャングが重要。

1. 個々の画像ごとのサーチャング
2. 画像系列フラグメントごとのサーチャング

固有台風表現

1. 枠取り台風画像の作成。
2. 雲分類アルゴリズムの適用。
3. 雲量ベクトルの計算。
4. 次元削減の適用。

北西太平洋データコレクションの固有台風表現。



類似画像検索

- 雲パターンの類似性に基づく検索機能.
- 過去の事例に基づく未来の予測のための仕組み.

<http://www.digital-typhoon.org/>

類似度による検索

Query 1	1	2	3	4
GMS596091711	GMS599100605	GMS502070112	GMS501082806	GMS597091810
199617 (WNP)	199920 (WNP)	200205 (WNP)	200112 (WNP)	199720 (WNP)
(N22.1,E128.8)	(N18.9,E118.4)	(N17.3,E132.3)	(N20.4,E143.6)	(N30.1,E141.8)
940 hPa / 85 kt	970 hPa / 65 kt	970 hPa / 65 kt	975 hPa / 60 kt	965 hPa / 70 kt
5	6	7	8	9
GMS595101209	GMS598121403	GMS595100213	GMS502030317	GMS599042905
199516 (WNP)	199815 (WNP)	199515 (WNP)	200202 (WNP)	199902 (WNP)
(N18.7,E108.3)	(N13.5,E108.3)	(N19.4,E113.0)	(N10.6,E134.4)	(N17.2,E113.5)
990 hPa / 50 kt	1002 hPa / 0 kt	985 hPa / 50 kt	955 hPa / 75 kt	990 hPa / 45 kt
10	11	12	13	14
GMS597081612	GMS502083012	GMS596051415	GMS597092408	GMS597082813
199713 (WNP)	200215 (WNP)	199603 (WNP)	199721 (WNP)	199716 (WNP)
(N24.0,E130.2)	(N30.9,E127.6)	(N16.9,E125.0)	(N15.7,E110.2)	(N22.4,E122.4)
945 hPa / 85 kt	950 hPa / 75 kt	930 hPa / 95 kt	988 hPa / 46 kt	955 hPa / 75 kt

類似画像検索

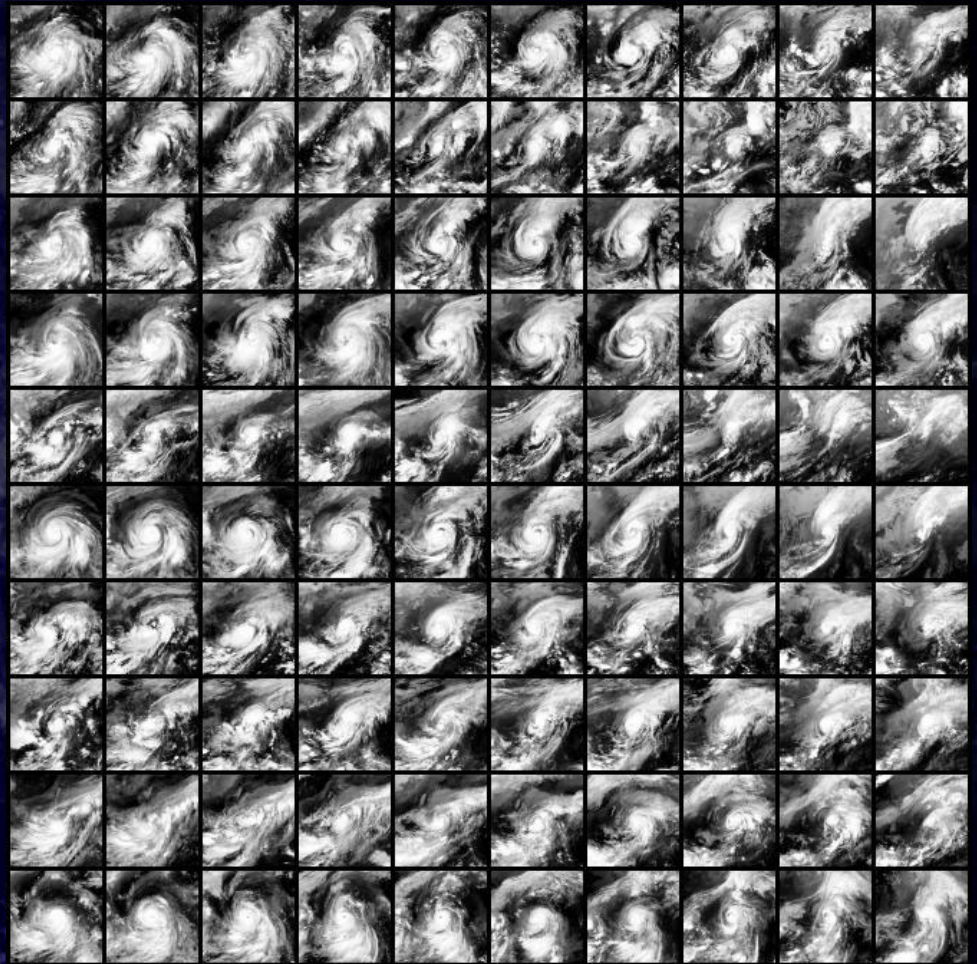
- 画像特徴の選択、および類似距離尺度の定義などに関する多くの研究がある。
- 台風雲パターンの類似性判定は、専門家による目視判読に頼っている。
- 類似性判定に関する手順(ドボラック法)は定められているが、定量的ではない。
- ユークリッド空間などの単純な表現が気象学的に有効かは検証が必要である。

類似系列フラグメント検索

1. 基準となる系列と他の系列との間で大域的整列 (**global alignment**) を計算し、**マッチングスコア**を算出する。
2. 整列を複数の系列に対して適用する (**multiple alignment**) ことで、**類似した時間変化**をサーチングする。
3. **動的計画法**と**クラスタリング**が重要な構成要素である。

類似系列フラグメント

- 台風のライフサイクルに対応するような類似系列フラグメントを検索。
- 複数系列が整列されているので、時間を正規化した上で系列間を比較できる。



データ操作言語

- 欲しい機能を直接的に記述できる**高レベルデータ操作言語**が必要である。
- 実装のパフォーマンスを考えると、独自の**高レベルデータ操作言語はSQLにマッピング**できれば望ましい。しかしそれが可能ではない場合もある。
- **問題領域に即したデータ操作言語**をどのようにして簡単に実現するか？

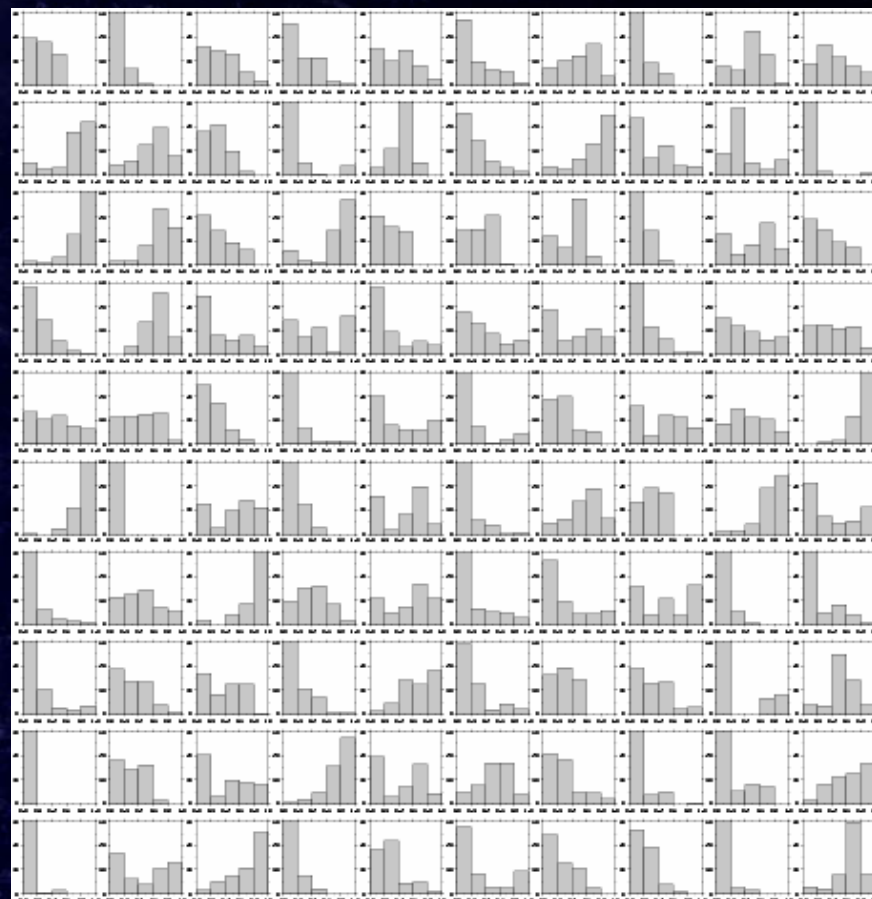
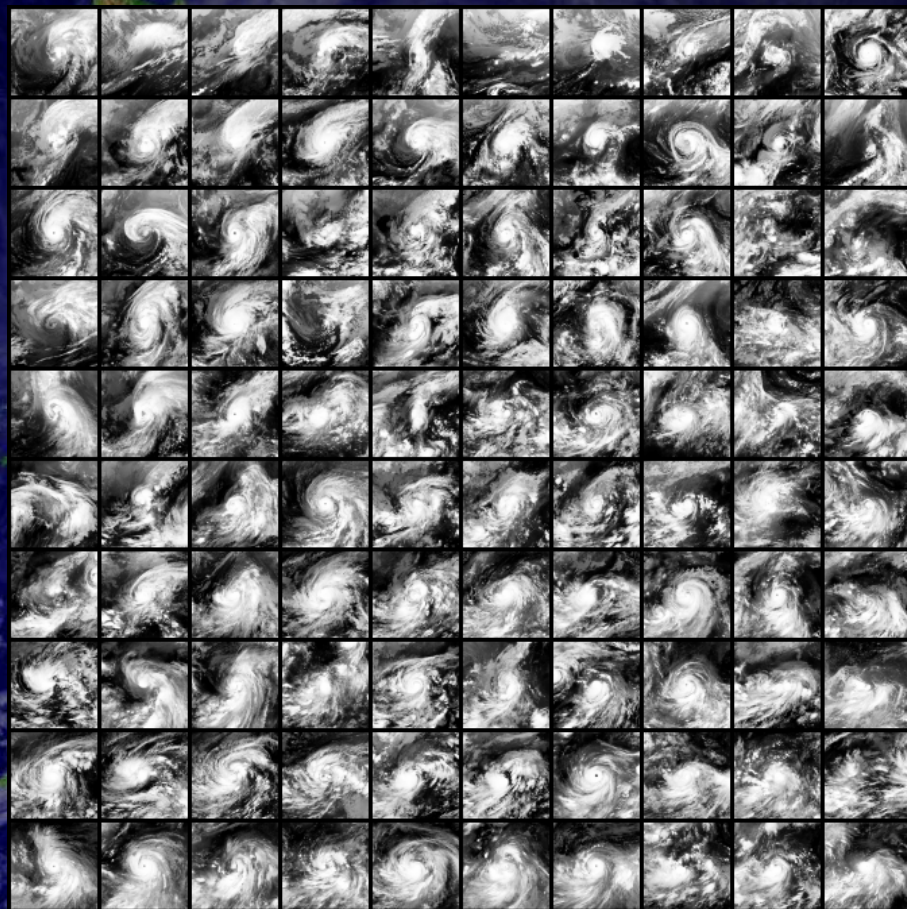
発表の概要

1. はじめに
2. データコレクションを基盤とする研究スタイル
3. 台風時系列画像コレクション
4. サーチング
5. マイニング
6. おわりに

台風画像マイニング

- 大規模データコレクションを「眺める」視点を提供し、そこから自動的(半自動的)に隠れた規則性・不規則性を発見するための方法。
- 現在のところ、低次元空間への射影や画像分類、クラスタリングなどの手法を種々試している段階。
- 予兆の発見は重要な研究課題の一つ。

クラスタリングと台風齢の関係



Clustering by K-means

Age Distribution

画像分類

- 発達しつつある台風と衰弱しつつある台風とを分類する問題を設定。
- サポートベクトル分類器とK-NN分類器を用いて画像分類問題の性能を比較。
- サポートベクトル分類器(ガウシアンカーネル)が最低の誤分類率(24.8%)を達成。
- 画像特徴およびカーネル設計を改善することにより、台風の発達において特徴的な法則を情報学的立場から追究する。

おわりに

- 「データベースを基盤とする科学研究」は今後も継続して発展すると予想できる。
- その一例として、「デジタル台風」プロジェクトに関する課題と現状をまとめた。
- サーチング：事例ベース推論のための類似事例検索機能として有用である。
- マイニング：気象学的にも興味深い法則の発見に有用である。

最後に

- 気象衛星「ひまわり」画像を研究目的にご提供くださる、東京大学生産技術研究所の安岡善文教授、喜連川優教授、根本利弘助手に感謝いたします。
- 台風中心位置等のデータは気象庁の観測に基づくものです。
- 本発表に関連する情報や画像などに関しましては、以下のウェブサイトをご覧ください。

<http://www.digital-typhoon.org/>