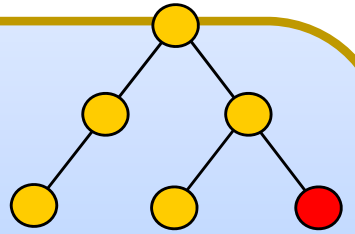
平成25年度JST 特別課題調査
「ビッグデータ統合利活用のための次世代基盤技術の創出・体系化」

位置情報マイニングの現状と展望
～ 実世界高速非構造マイニングの最前線 ～



有村 博紀
北海道大学大学院情報科学研究科
http://www-ikn.ist.hokudai.ac.jp/~arim/
e-mail: arim@ist.hokudai.ac.jp

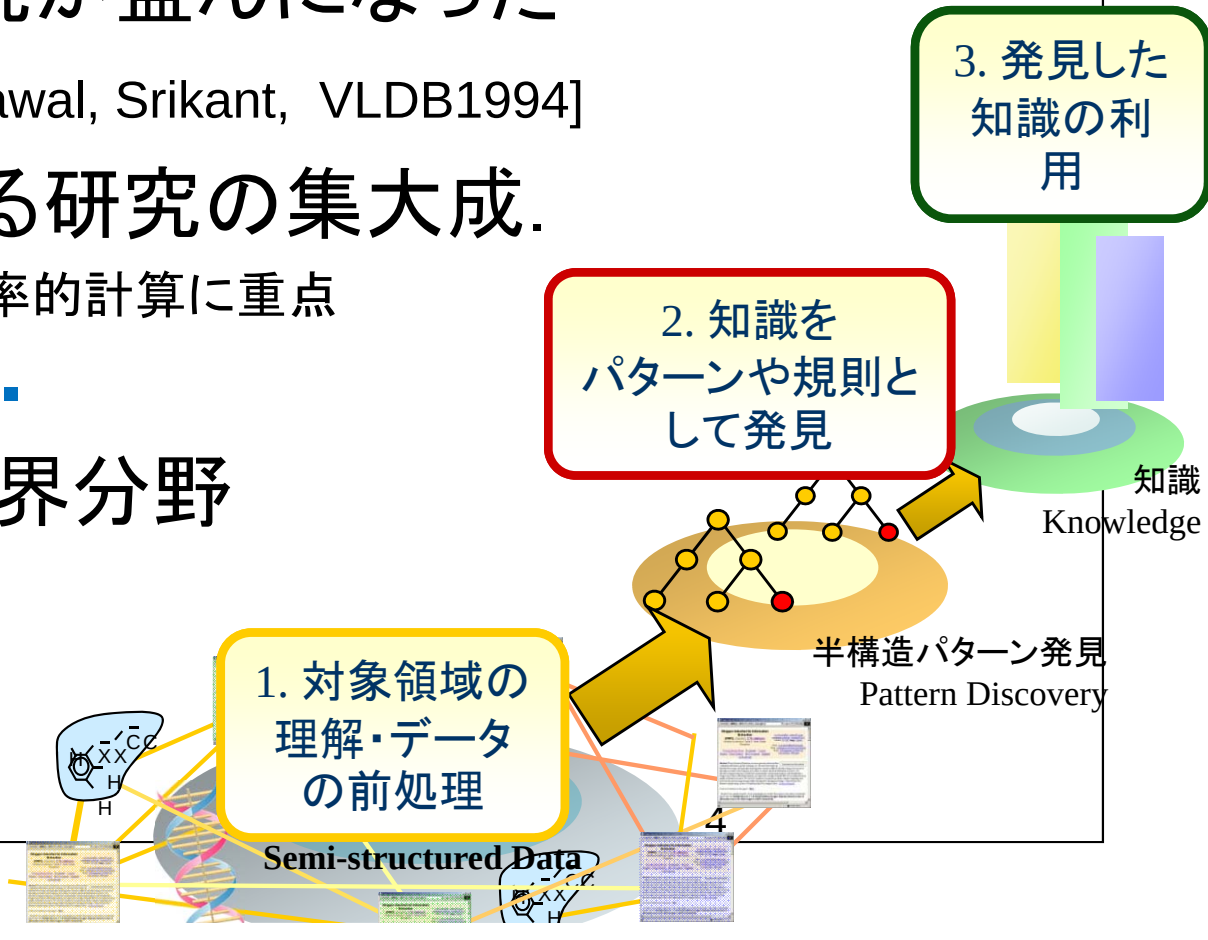

Asahi-dake, Daisetsu-zan, Hokkaido

実世界からのデータマイニングに向けて

実世界の大規模・非定型・時空間データの超高速処理に正面から取り組む！

データマイニングとは

- 大量のデータから人間にとって有用なパターンや規則を効率良くとりだす方法の研究
- 1990年代半ばから研究が盛んになった
  - Apriori algorithm [Agrawal, Srikant, VLDB1994]
- 潜在的には古くからある研究の集大成.
  - ただし，大量データに対する効率的計算に重点
- 機械学習・数理統計学・データベース技術の境界分野



データマイニングの動向

パターン発見

- トランザクションデータから共通して出現する規則性を発見する
- 頻出パターン発見 [Agrawal et a. ‘94]
- 最適化マイニング [森下 ’96, ’98, ‘00]

予測学習・自動分類

- 不完全なデータから，未知の規則を学習する
- SVM [Vapnik ‘96],
- Boosting [Shapire & Kearns ‘96]
- C4.5 [Quinlan ‘96]

構造マイニング

- 非定型構造データから特徴的な部分構造を規則性を発見する
- グラフマイニング[Washio & Motoda ‘00], [Zaki ’02], [Uno, Asai, Arimura, ’02, ‘03]

クラスタリング

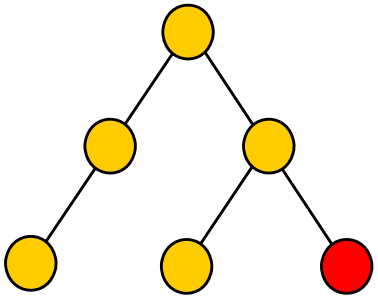
- データを類似したものどうしグルーピングする.
- 大規模・不完全なデータからの高速クラスタリング
- K-means, CLARANS, DBSCAN

確率モデリング

- 高次元大規模データから不確実な現象を予測・モデル化する
- ベイジアンネットワーク [Pearl ’90s]
- HMM [Asai], MCMC, ベイズ推定・MDL・AIC

新しいタイプのデータマイニング

- テキストマイニング
  - 自然言語テキスト
  - 情報抽出
  - 意味マイニング
- ストリームマイニング
  - センサー監視
  - 近似統計処理



ビッグデータ時代の
データマイニングとは？

ビッグデータ時代の大規模データ処理

情報世界と実世界の融合

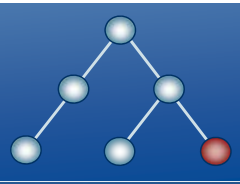
- 大量のデータ　クラウド
- 多数のCPU
- 高速なネットワーク
- 膨大な計算

「集中」

- さまざまなデバイス
- 多様な人間活動と応用
- 多様で非均一な時空間
- 不完全で複雑なデータと情報

「分散」

どこが新しい動きなのか？



- ひとり一人のミクロな解像度の世界的規模のマクロデータ
- リアルタイムに解析可能な時代になってきた
- データ、ハードウェア、知識発見技術の成熟

Facebook
(月間10億ユーザ)

Twitter
(7000件/秒)

Suica
(260万回利用/1日)

数十テラバイトの10億以上のトランザクションのデータを毎日処理.

環境からの情報。気象・交通・自然・社会

Collective Intelligence

Machine Learning

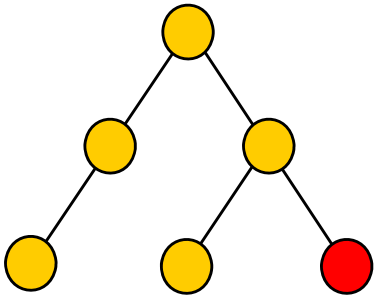
Hadoop

Data Mining

GFS/BigTable

NLP

点から線へ，面へ.
人と人、人とモノの関係性をインタラクションから探る時代



位置情報マイニングの現状

ビッグデータマイニングの鍵

動機：モビリティ，サイバーフィジカル

- CPS（サイバーフィジカルシステム）
- 人とモノのモビリティに関心
  - 移動を通じて人間の活動にアクセス
  - 社会活動の最適化（スマートXX）
  - 各種サービス・産業の基盤と媒体となる？
- 大量の移動体データ
  - プローブカー，歩行者，野生動物？
  - GPS, スマホ，WIFI，etc.
- どのような情報を取り出す？
  - 時空間における移動の解析・予測
  - 移動パターンの発見（「トラジェクトリパターン」）

GCPSプロジェクト(H23～H27予定)

- 「グリーン・サイバー・フィジカル・システム基盤技術開発」(代表:坂内先生／安達先生)
- NII，九大，北大，阪大の4拠点で
  - NII(安達淳)「IT統合基盤のCPS共通技術」
  - 九大(安浦寛人)「データ収集／解析技術と学研都市スマートシティ化への適用」
  - 北大(田中譲)「オープン・スマート・フェデレーション技術とスマート除排雪への適用実証実験」
  - 阪大(東野輝夫)「プラットフォーム技術と都市街区における行動」
- 北大は「スマート除排雪への適用実証実験」

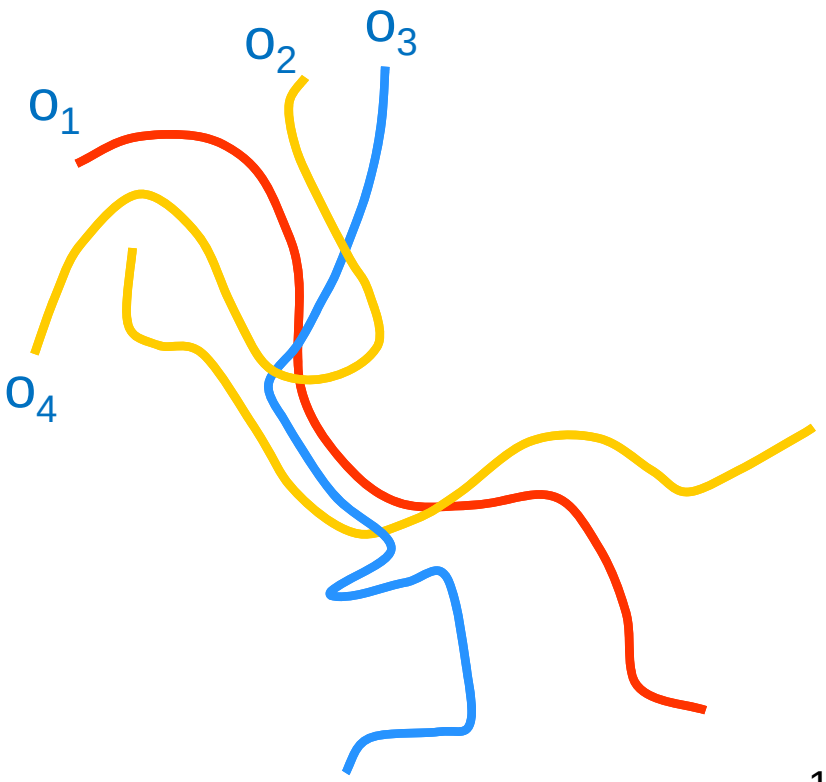
軌跡（トラジェクトリ）データ

- 時空間データはさまざまな定式化が可能

① 移動体（moving objects）の集合 $O = \{o_1, \ldots, o_n\}$
- 野生動物，歩行者，プローブカー
- 付加情報は仮定しない（属性ラベルなし）

② 時間 **T**
- 連続時間 **T** = R
- 離散時間 **T** = [0..T]. （等間隔）

③ 空間 **S**
- 2次元連続空間 **S** = R^2
- 2次元のメッシュ **S** = $[0..u]^2$
- 道路ネットワーク **S** = (V, E)



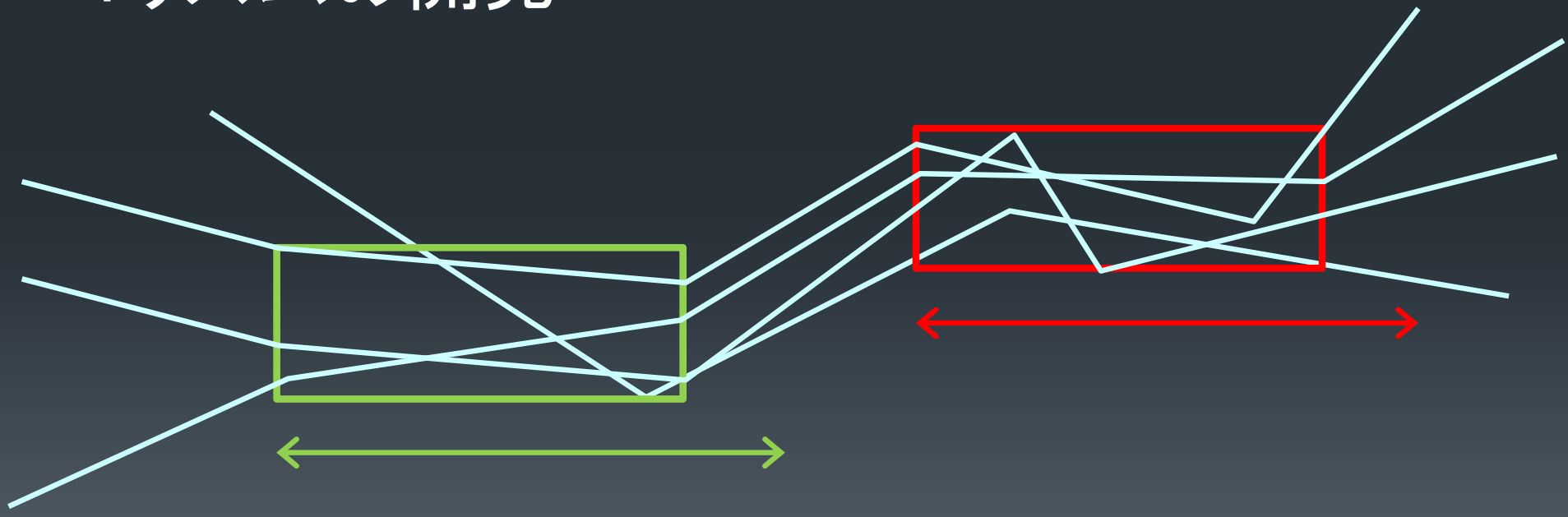
Mining Interesting Locations and Travel Sequences from GPS trajectories (Zheng, WWW'09)

- GeoLife Project
  - Microsoft Research Asia
  - Mobile phone with GPS
  - Purpose: Recommendation
    - Interesting location
    - Travel sequences
  - Using Tree-based index structure

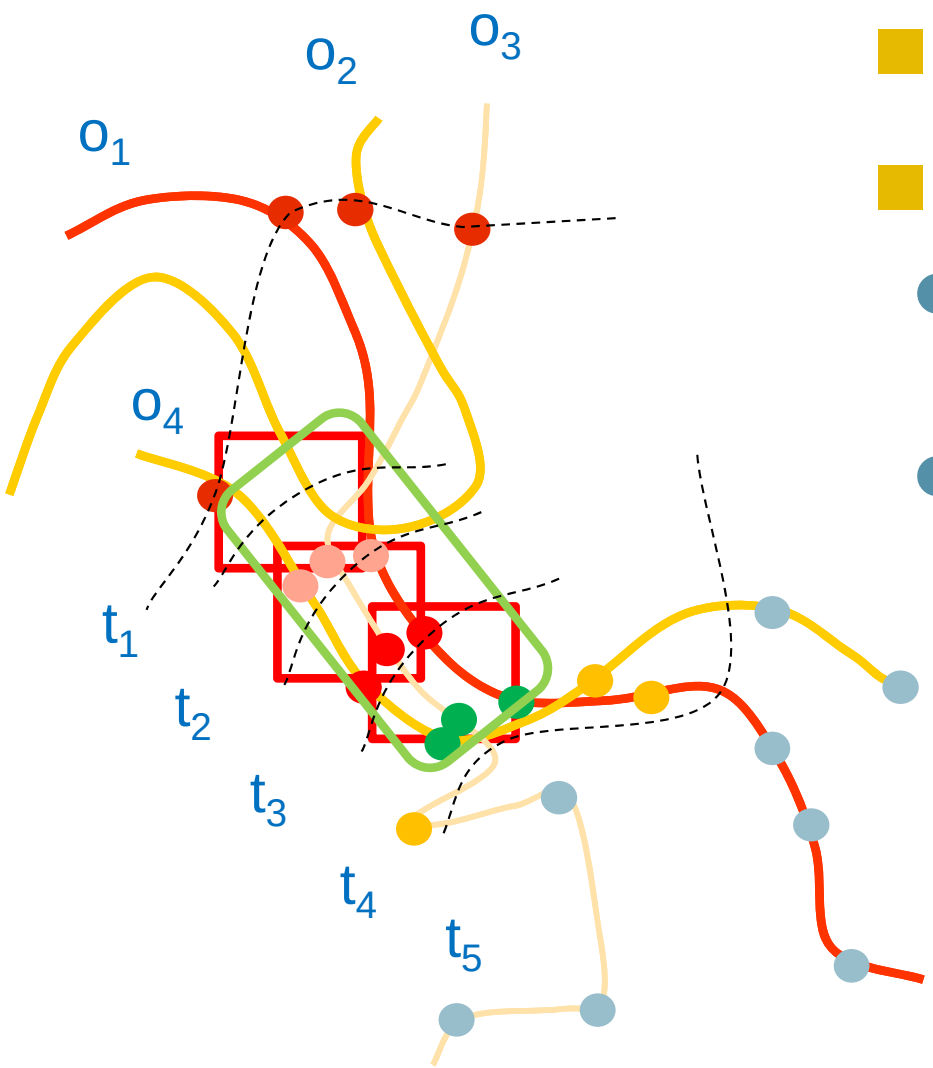
有村博紀, 北海道大学

時空間ストリームマイニング

- 極大群れパターン(flock pattern)発見:
  - 2次元／3次元の時空間データストリーム
  - 離散構造列挙技術に基づく高速マイニングアルゴリズムの開発



「群れ」パターンマイニング



- トラジェクトリマイニング
- 群れ」パターン P = (X, A)
  - Gudmundssonらによって導入 (AGIS2006)
  - 移動体の集団 X = {o_1, ⋯, o_m} が，ある長さ k 以上の時間区間 A = [beg, end] の間，距離 r 以内で一緒に移動することを表す

「一緒に移動する」とは，Aの各時点 t において，すべての移動体の位置が，一辺 r のある矩形に含まれること.

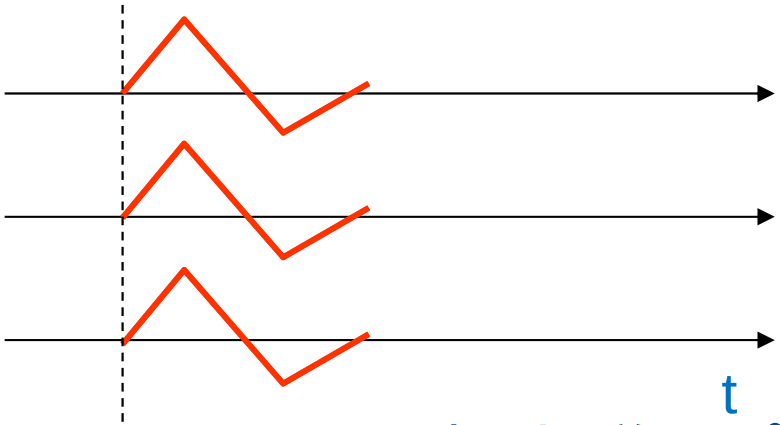
r = 10m, k = 3

P = (X={o_1,o_3,o_4}, A=[t_2..t_4])

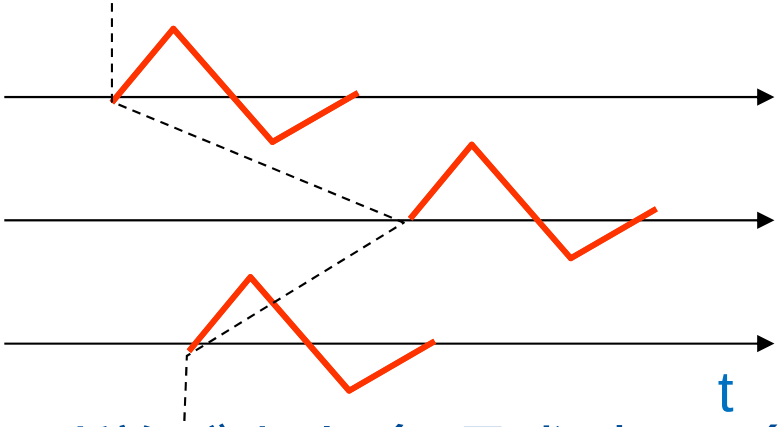
定義:「群れ」パターン

- 同期型と非同期型のパターン
  - 移動体の位置が，厳密に同じ時刻で同期するか（同期型）(Gudmundsson他'06），同期せず相対時刻の意味で近接するか（非同期型）の違い.
  - オリジナルの群れパターンは同期型.
  - 今回は，非同期型も導入する.

同期型



非同期型

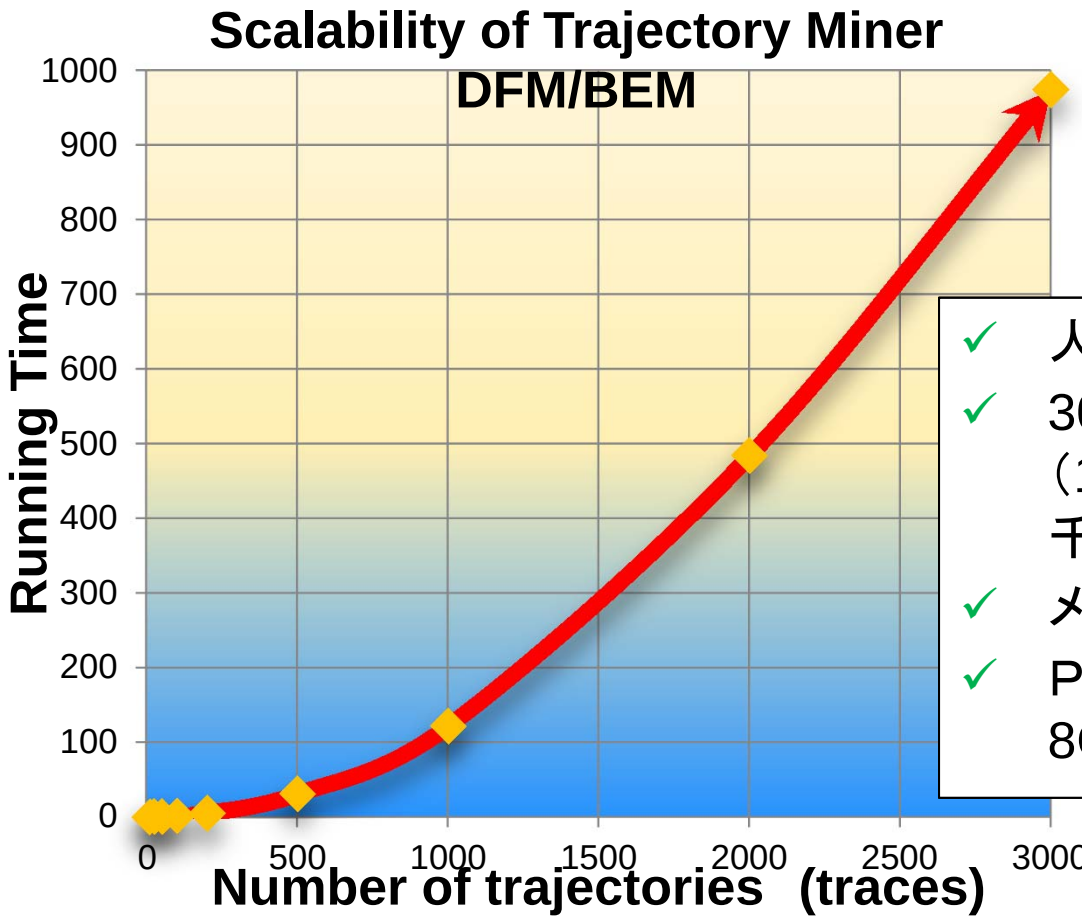


Q: すべての極大群れパターンの列挙が出力多項式時間（多項式遅延）でできるか？(最大はNP完全)

高速「群れ」パターンマイニング手法

- GCPS情報学研グループとの共同開発
- 頻出＆極大トラジェクトリパターン発見
- 同期と非同期のパターンを両方マイニング可能
- C++言語。今後の展開：ビッグデータ向けの大規模化

Hiroki Arimura (HU)



- ✓ 人工軌跡データ
- ✓ 30点×3000トレース（10秒サンプリングで5分×3千台の自動車に対応）
- ✓ メモリ1.28MB使用
- ✓ PC（Intel Core i7, 2.9GHz, 8GBメモリ）

Takeaki Uno (NII)

EXP1: Comparison of Sync & Async

- Comparing the #patterns and cputime for **FPMsync** and **FPMasync** algorithms

#patterns found	sync patterns	async patterns
true patterns	40	40
FPMsync	40	0
FPMasync	41	43

時間がズレたパターンが正しく見つかっている

total time	sync patterns	async patterns
FPMsync	0.640 sec	0.640 sec
FPMasync	0.733 sec	0.686 sec

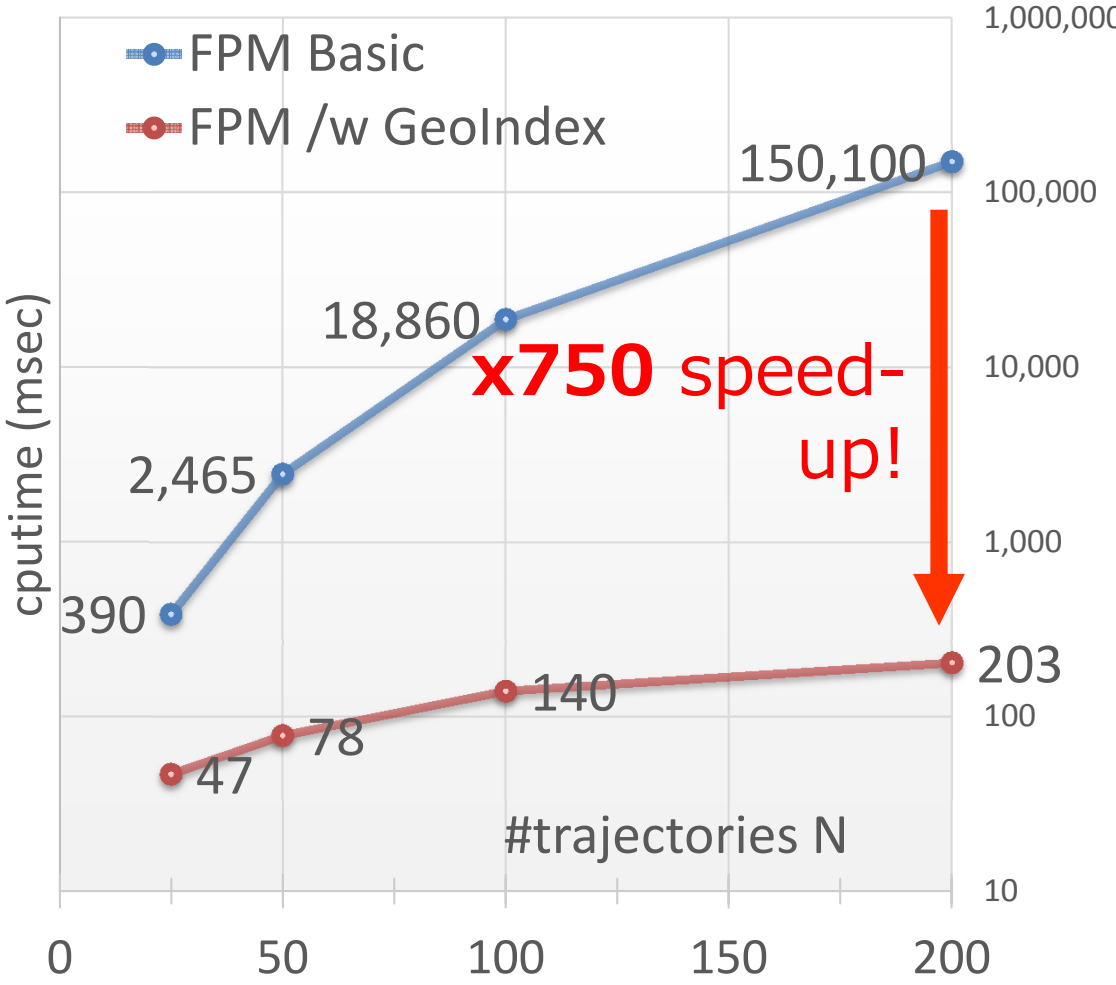
with geo-index

Setting

- Area 100.0 x 100.0
- 400 trajectories of length 100 generated by random walk with step 1.0 and angle ±90deg
- 5 inplanted copies of each of 40 random patterns of length 10 within width 1.0x1.0
- Mining with max width 1.0x1.0, min length 10, and frequency at least 5.

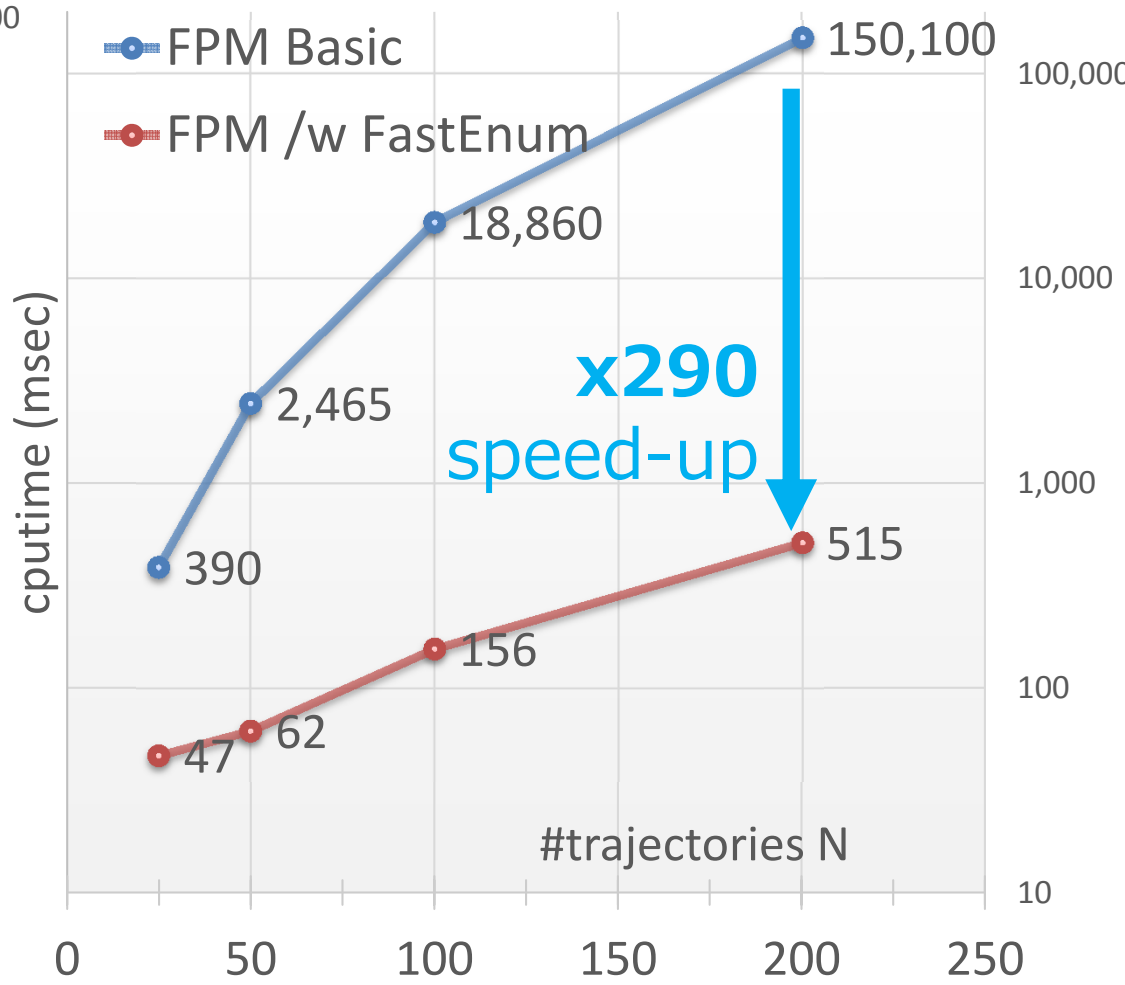
Experimental Results

EXP2 Speed-up by Geo-index



Setting: N = 25 to 200 trajectories of length 40 in which N/10 patterns x 5 copies are implanted

EXP3 Faster Enumeration

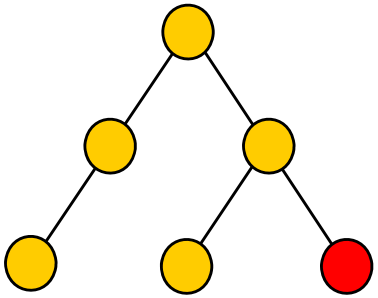


Setting: N trajectories of length 40. Other parameters are same to EXP2:

応用

「群れ」パターンマイニングで
何ができるか？

当日の発表で. . .

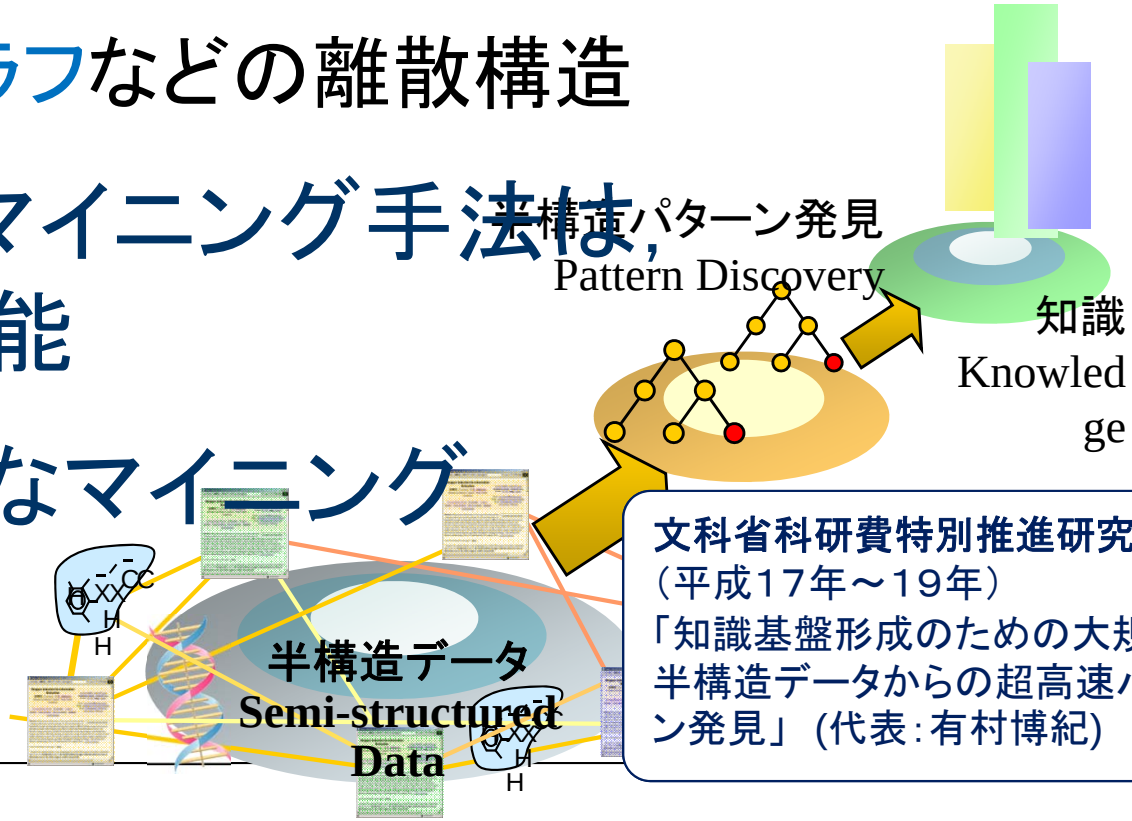


非構造＆半構造マイニング

ビッグデータマイニングの鍵

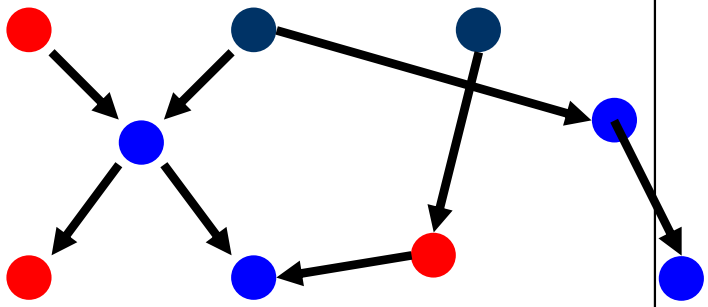
半構造マイニング

- 1990年代後半～の大規模半構造データの出現
  - 大規模で，多様な，非定型データ
  - 時間変化も(ビッグデータ)
  - 系列・木・グラフなどの離散構造
- 従来のデータマイニング手法は，直接適用不可能
- 高速かつ頑健なマイニング技術が鍵

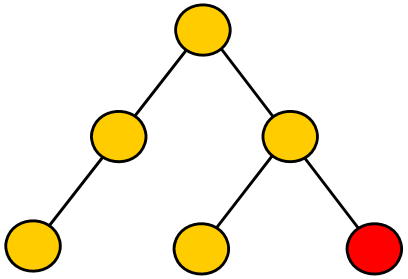


文科省科研費特別推進研究
（平成17年～19年）
「知識基盤形成のための大規模半構造データからの超高速パターン発見」 (代表：有村博紀)

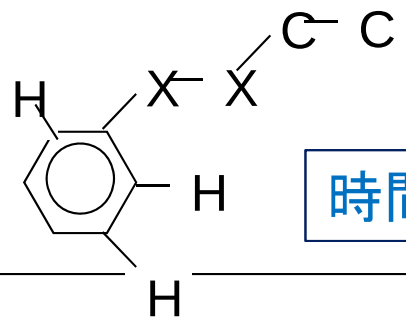
さまざまな半構造マイニング



- 系列（シーケンス）　時間と位置
  - ゲノム配列解析による個別医療
  - イベントストリームマイニング
- 木（木構造）　組み合わせ
  - 自然言語テキスト解析



構造



時間変化

- グラフ（関係）
  - 顧客と商品の購買や推薦関係
  - Twitter やSNSのユーザ同士のネットワーク　モノとモノの関係
  - 薬物・化合物の構造からの機能予測

群集

半構造データマイニングの歴史

～1995
Algorithm for finding subgraphs by MDL principle
Subdue [Holder *et al*. (KDD'94)]

1996
Finding frequent paths [Wang and Liu (KDD'97)]

1997
Finding Semi-structured Schema
[Nestrov, Abiteboul *et al*. (SIGMOD'98)]

1998

1999

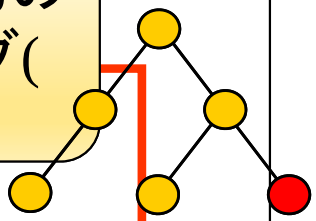
2000
Finding frequent subgraphs
AGM [Inokuchi, Wahio, Motoda (PKDD'00, MLJ 2003)]
FSG [Kuramochi *et al*. (ICDM'01)

2000年 最初の頻出グラフマイニングの論文

2001

Finding frequent ordered trees
FREQT [Ours (SDM'02)], **Treeminer** [Zaki (KDD'02)]

2000年 最初の木マイニング(われわれ)



2002
DFS Graph mining
gSpan [Yan and Han (ICDM'02)]

2002年 グラフマイニング決定版

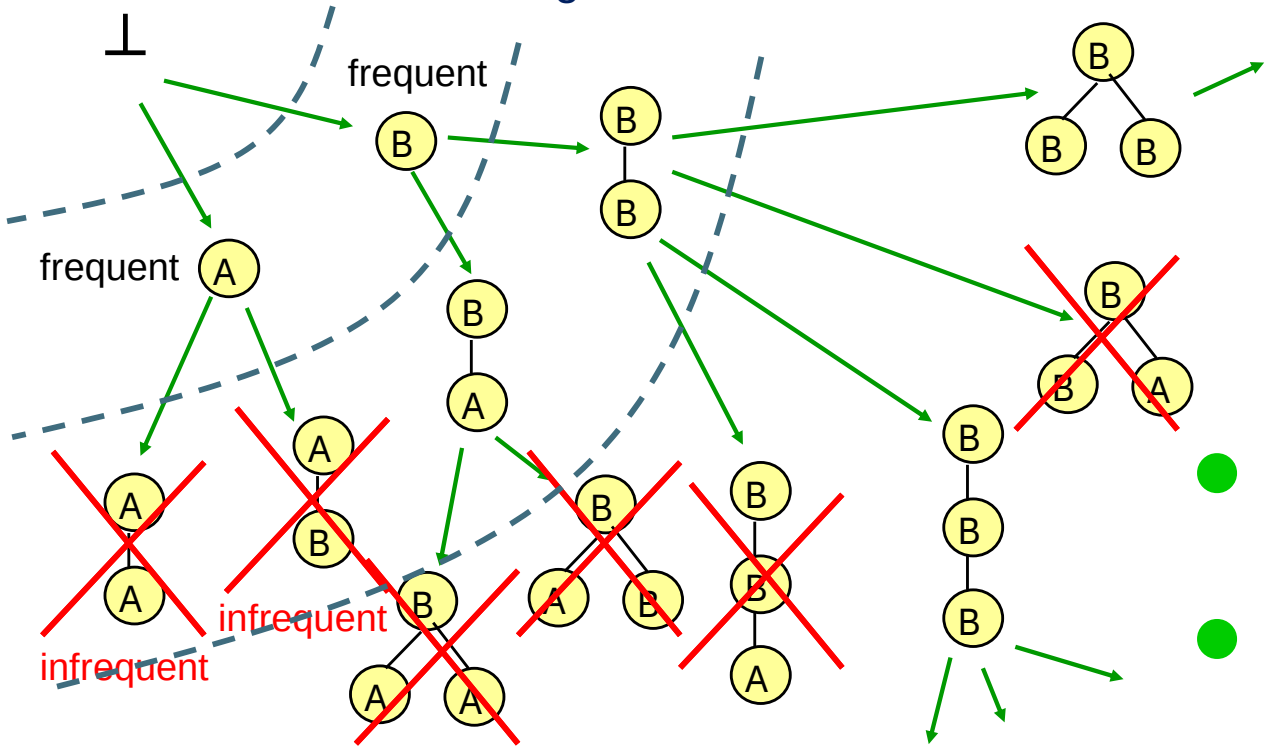
2003
Finding frequent un-ordered trees
UNOT [Ours (SDM'03)], **NK** [Nijssen, Kok (MGTS'03)]

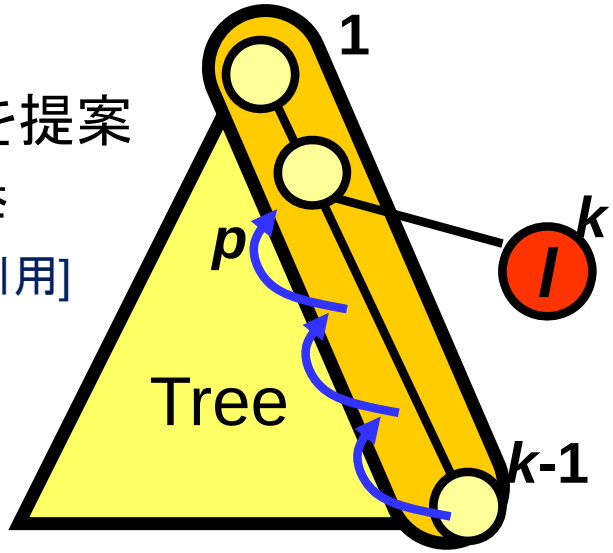
Closed Graphs mining [**closeGraph**: Yan&Han '03; Termier et al.'04] and many algorithms in 2000s

超高速半構造パターン発見技術

最右拡張技法 (Rightmost expansion)

- 発表者等が2002年に開発 [SIAM DM'02]
- 世界初の解あたり多項式時間アルゴリズムを提案
- 現在，さまざまな組合せ構造の効率よい列挙に用いられる [当該分野の国際会議論文の多数に引用]
  （引用435件（Google Scholar 2013調べ）





- Freqt [Asai, Arimura et al., SIAM DM'02]
- TreeMiner [Zaki, KDD'02]

25

まとめ

- ビッグデータ時代のデータマイニング
- 位置情報マイニング
- 超高速な非構造パターンマイニング
- 移動軌跡からのマイニング
- 今後の展望