トラジェクトリーマイニングの統計学

竹内　一郎

名古屋工業大学/理研/物質材料研究機構

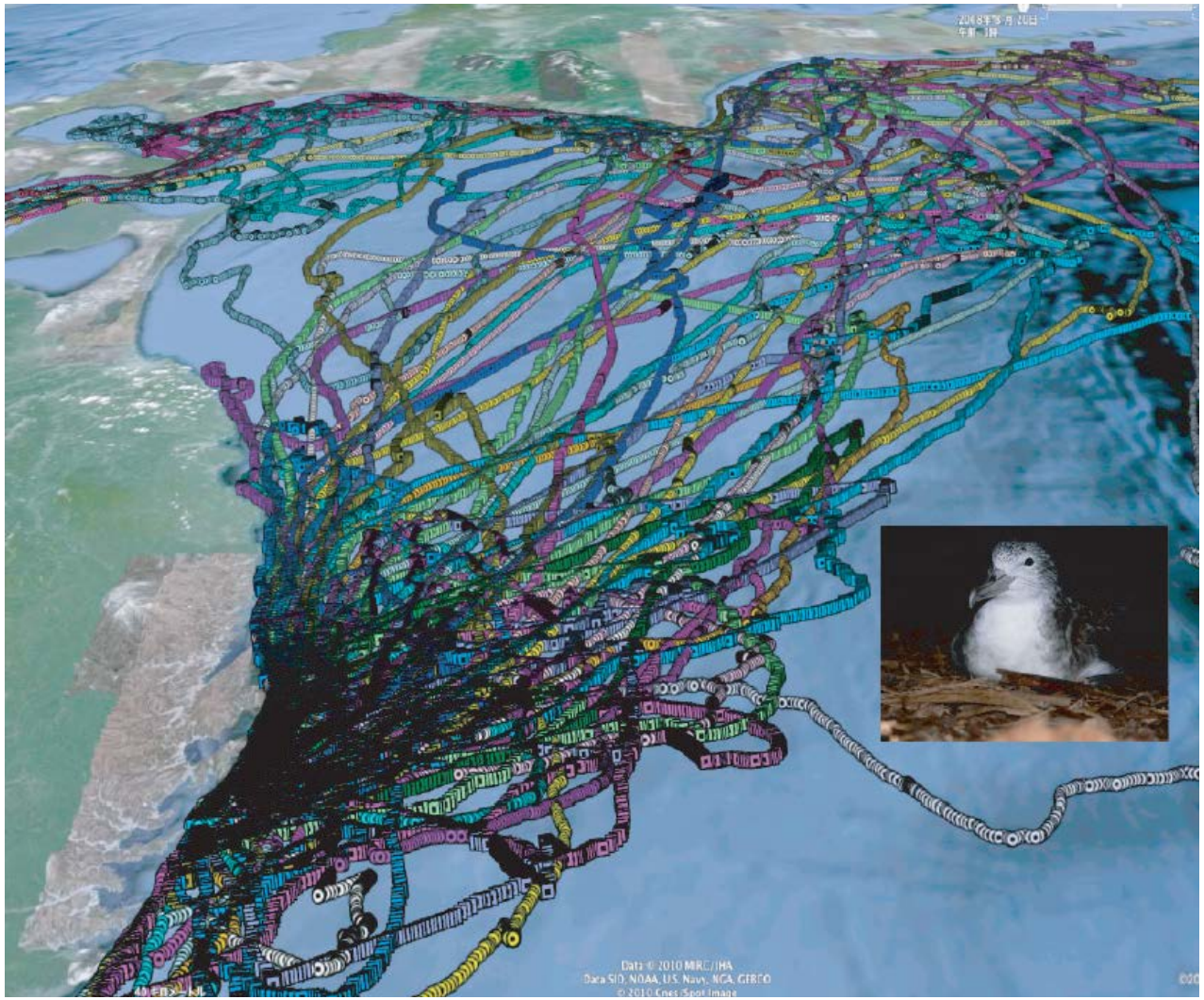
謝辞: 本発表は以下の研究室との共同研究によるものです

北大小川研，阪大木村研, 広大玉木研，東大津田研，名大依田研, 名工大竹内研（順不同）

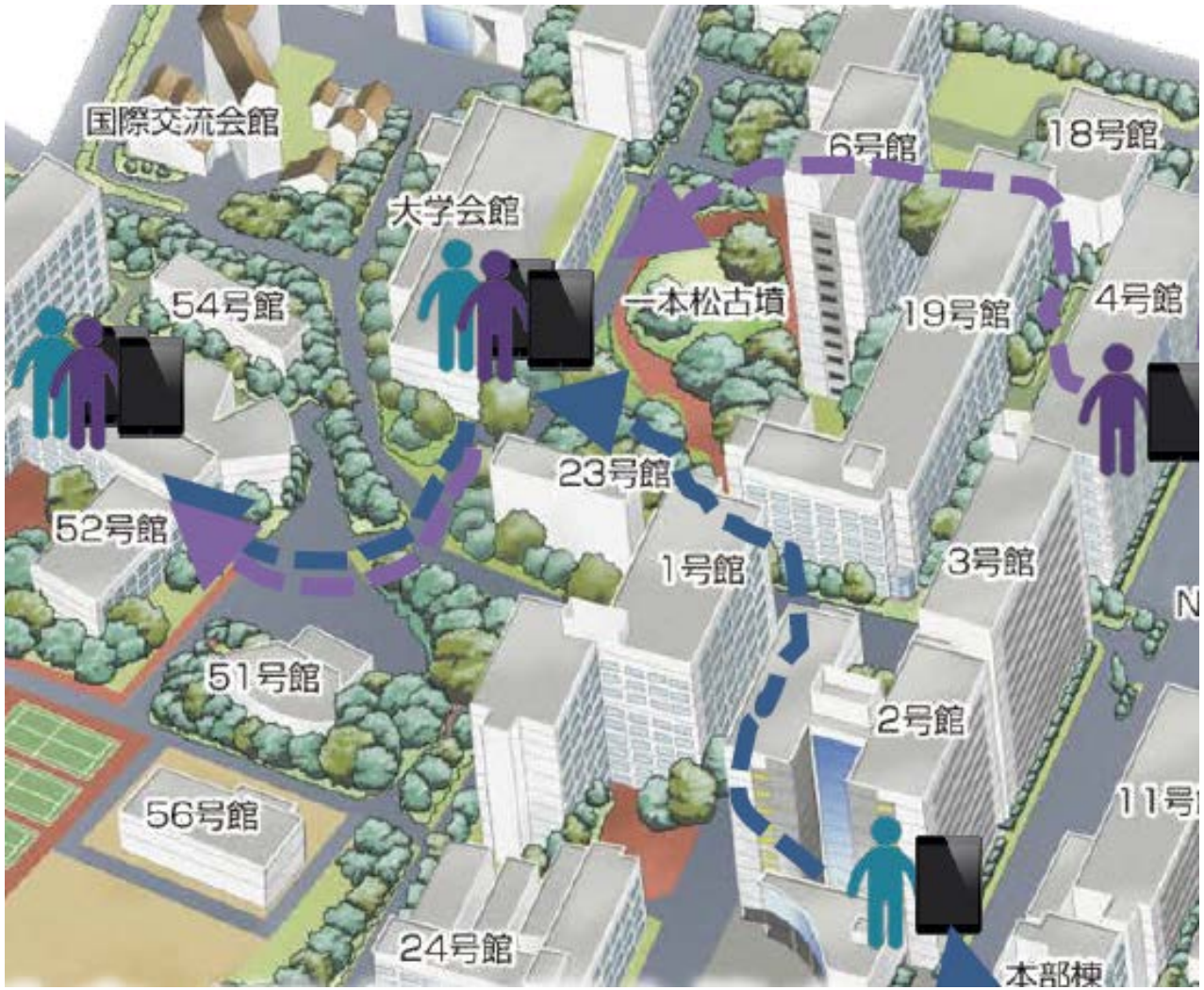
Trajectory Mining

$=$ Trajectory Data $+$ Data Mining

トリのトラジェクトリーマイニングの例



ヒトのトラジェクトリーマイニングの例



Definition of Trajectory Data

- A **simple trajectory data** is formulated as

$$\{(x_t, y_t)\}_{t=1}^{T} := (x_1, y_1), (x_2, y_2), \ldots, (x_T, y_T),$$

where (x_t, y_t) is the x (e.g., altitude) and y (e.g., latitude) coordinates of the location at time t.

- A **feature-associated trajectory data** is formulated as

$$\{(x_t, y_t, \boldsymbol{z}_t)\}_{t=1}^{T} := (x_1, y_1, \boldsymbol{z}_1), (x_2, y_2, \boldsymbol{z}_2), \ldots, (x_T, y_T, \boldsymbol{z}_T),$$

where $\boldsymbol{z}_t$ is a vector of associated (e.g., environmental and/or biological) features at time t.

Note that **LOGBOT** *is developed for obtaining rich features!*

Trajectory Data, Mining, and Knowledge

Tasks in Trajectory Mining (according to Zhang paper)

- Trajectory data preprocessing

  noise filtering, stay point detection, trajectory compression, trajectory segmentation, map matching
- Trajectory data management

  trajectory indexing and retrieval, distance/similarity of trajectories
- Uncertainty in a trajectory

  reducing uncertainty from trajectory data, privacy of trajectory data,
- Trajectory pattern mining

  moving together patterns, trajectory clustering, mining sequential patterns from trajectories, mining periodical patterns from trajectories
- Trajectory classification

  whole-trajectory classification, *sub-trajectory classification*
- Anomalies detection from trajectories

  detecting outlier trajectories, identifying anomalous events by trajectories

black: irrelevant, blue: relevant, considered, red: relevant, unconsidered

トラジェクトリーマイニングの統計学

トラジェクトリーマイニングの統計学

はまだ整備されていない

トラジェクトリーマイニングの統計学

はまだ整備されていない

=Trajectory Knowledge の統計的信頼性を評価できない

これまでのこころみ

1. p 値付き系列マイニング

系列マイニング　＋　Tarone 多重検定補正

2. p 値付きイベント検出

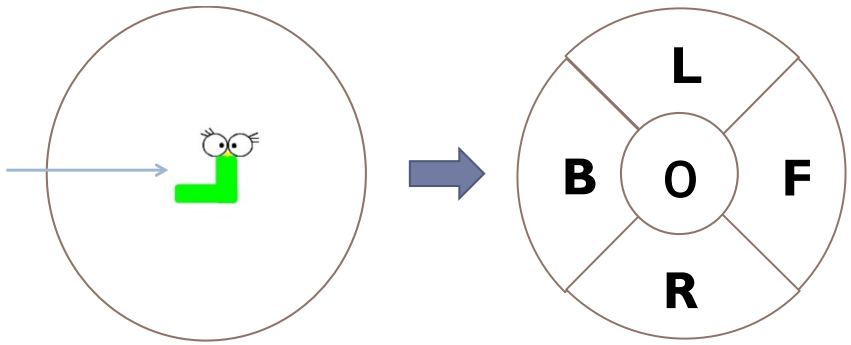
イベント検出　＋　Selective Inference

トピック 1

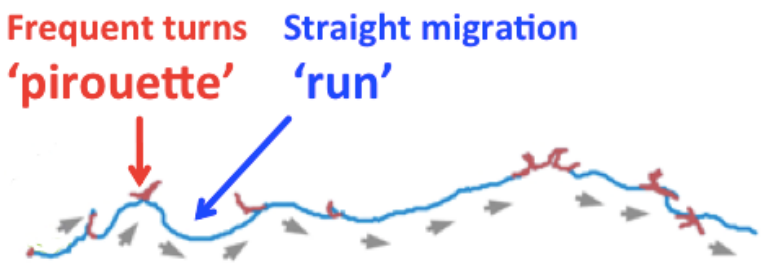
ナビゲーション系列マイニングとその統計的評価

ナビゲーション行動の系列表現

- 線虫の移動を 3 次元のシンボル系列で表現（$5 \times 2 \times 3 = 30$ 状態）
  - 移動方向：{**F**, **B**, **L**, **R**, **O**}



  - Run or Pirouette： {**N**, **P**}



  - 匂い勾配：{**U**, **M**, **D**}

線虫のナビゲーション行動の系列マイニング

- 線虫移動の 3 次元シンボル系列:

$$\text{(例)} \begin{bmatrix} \mathrm{O} \\ \mathrm{P} \\ \mathrm{U} \end{bmatrix} \begin{bmatrix} \mathrm{R} \\ \mathrm{N} \\ \mathrm{M} \end{bmatrix} \begin{bmatrix} \mathrm{R} \\ \mathrm{P} \\ \mathrm{D} \end{bmatrix} \begin{bmatrix} \mathrm{O} \\ \mathrm{P} \\ \mathrm{U} \end{bmatrix} \begin{bmatrix} \mathrm{O} \\ \mathrm{P} \\ \mathrm{U} \end{bmatrix} \begin{bmatrix} \mathrm{L} \\ \mathrm{N} \\ \mathrm{M} \end{bmatrix} \begin{bmatrix} \mathrm{O} \\ \mathrm{P} \\ \mathrm{U} \end{bmatrix} \begin{bmatrix} \mathrm{R} \\ \mathrm{N} \\ \mathrm{M} \end{bmatrix} \cdots$$

- 野生型と変異型で頻度の異なるシンボル系列を発見したい

$$\text{(例)} \begin{bmatrix} \mathrm{F} \\ \mathrm{N} \\ \mathrm{D} \end{bmatrix}^2 \begin{bmatrix} \mathrm{B} \\ \mathrm{P} \\ \mathrm{D} \end{bmatrix} \begin{bmatrix} \mathrm{F} \\ \mathrm{P} \\ \mathrm{D} \end{bmatrix}^2 \text{が野生型で高頻度に起こる}$$

- パターン総数

$$\text{長さ 20 までの系列総数} = (5 \times 2 \times 3)^{20} > 3.4 \times 10^{29}$$

- ボンフェローニ法による多重検定補正をすると...

$$p < \frac{0.05}{3.4 \times 10^{29}} \text{の場合に統計的有意と判定}$$

Sakuma et al. (under preparation) *c.f., Terada et al., 2013*

- ボンフェローニ補正は過度に保守的
- Westfall-Young 法（並べ替え法）により系列相関を考慮
- タローネによる多重検定補正（for Fisher 正確検定）

（例）頻度が 5 のパターン

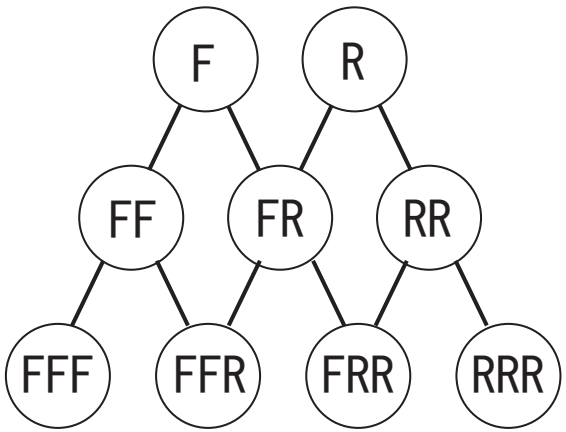
	野生型	変異型
系列アリ	k	$5-k$
系列ナシ	N_0-k	$N_1-(5-k)$

$\Rightarrow$

	野生型	変異型
系列アリ	5	0
系列ナシ	N_0-5	N_1

低頻度パターンは極端な場合でも p 値が小さくならない

- 分岐限定法による効率的な探索



線虫ナビゲーションの結果

DOP3

パターン	補正P値	総サポート	サポート(DOP3)	サポート(N2)
F:P:D14	0.000366	25	24	1
F:P:D15	0.000782	20	20	0
F:P:D8 F:P:U4	0.004194	18	18	0
F:P:D10	0.018554	97	62	35
B:P:D F:P:D14	0.021539	20	19	1
F:P:D16	0.021666	16	16	0
B:P:D F:P:D15	0.021666	16	16	0
F:P:U F:P:D8	0.033188	26	23	3
F:P:D8 F:P:U3	0.04694	19	18	1

N2

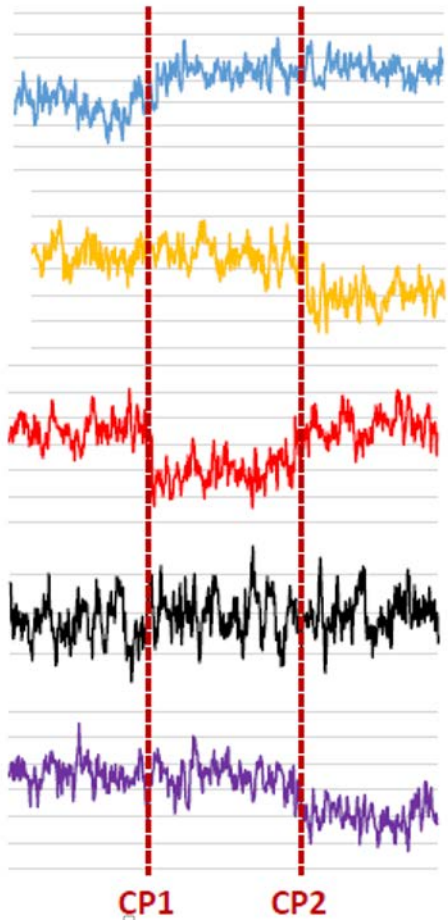
パターン	補正P値	総サポート	サポート(DOP3)	サポート(N2)
F:N:D48	1.88E−06	105	35	70
F:N:D60	1.53E−05	74	19	55
F:N:D2 B:P:D	0.000111	76	21	55
F:N:D38	0.000127	121	47	74
F:P:D F:N:D49	0.001245	49	10	39
B:P:D F:N:D30	0.00151	71	20	51
B:P:D2 F:N:D31	0.006026	42	8	34
F:P:D F:N:D54	0.008336	39	7	32
F:N:D72	0.012137	46	10	36
F:N:D2 B:P:D F:P:D2	0.015668	27	3	24

トピック 2

多変量時系列からのイベント検出とその信頼性評価

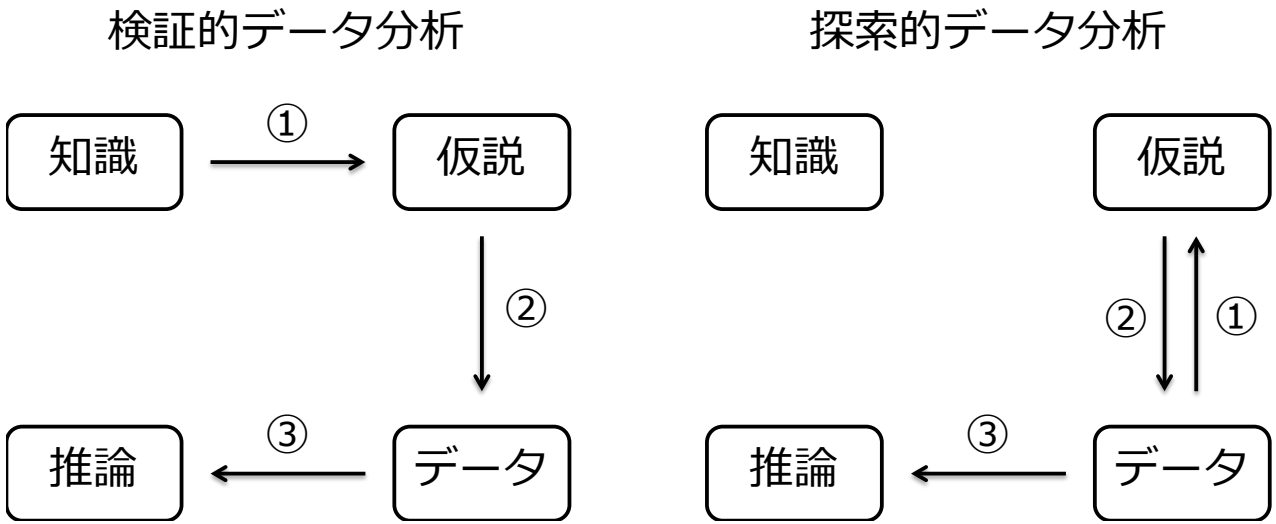
多変量時系列からのイベント検出

イベント検出は探索的



検証的データ解析から探索的データ解析へ

- 検証的データ解析から探索的データ解析へ

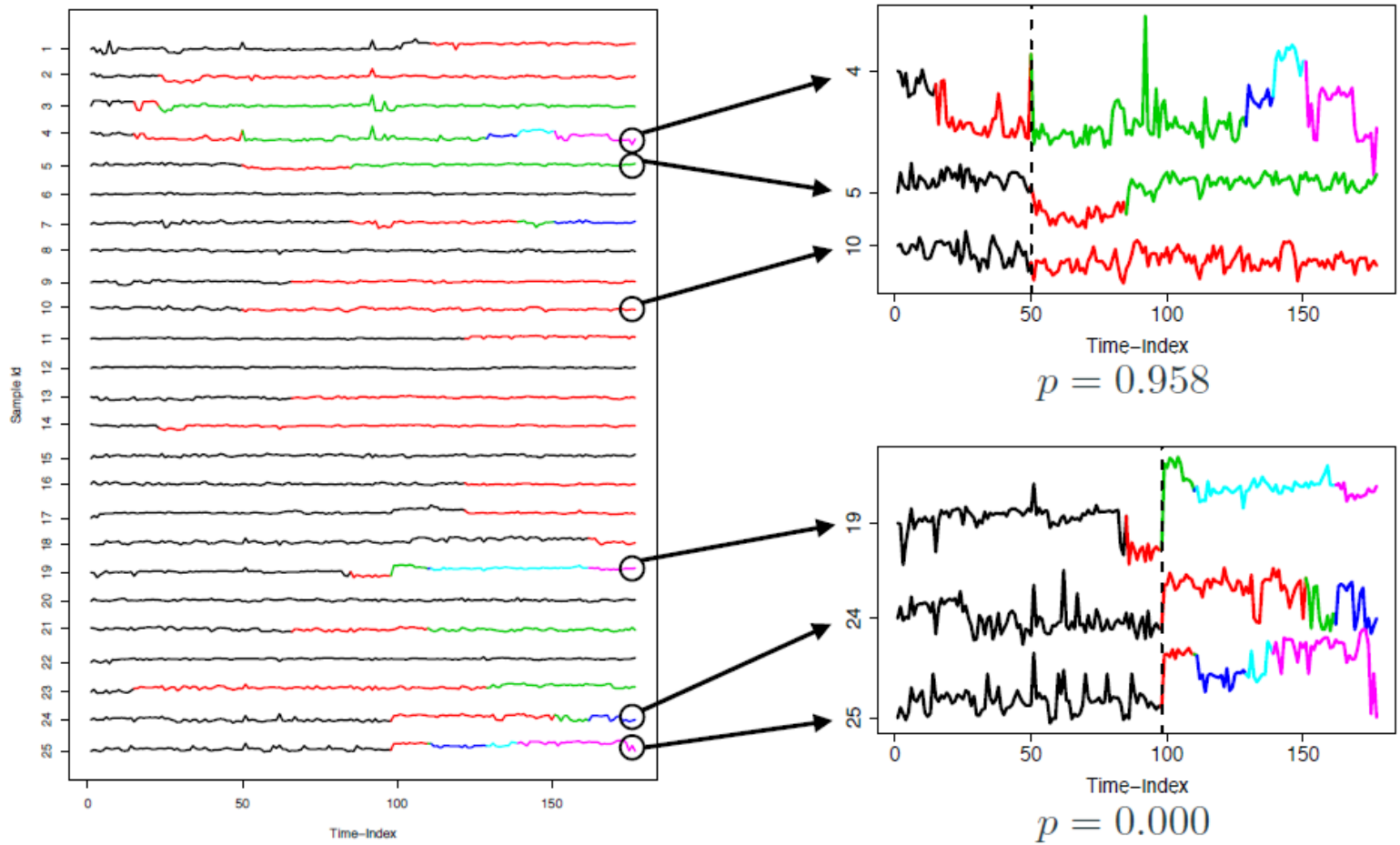


- Selective Inference

$$\mathrm{P}_{\text{帰無仮説}}(\text{統計量} > \text{閾値} \mid \text{仮説を選択}) < \text{有意水準}$$

ゲノム異常検出の結果

- array CGH データへの適用 (Takeuchi et al., 2009)
  - ✓ $T = 177, N = 46, h = 10, K = 3$



まとめ

- トラジェクトリーマイニング（TM）は探索的
- TM の結果の統計的信頼性を評価する方法論は整備されていない
- 多重検定補正，選択バイアス補正が必須
- 現状：タローネ補正，Selective Inference