人流・車流の軌跡類似度を用いたクラスター分析

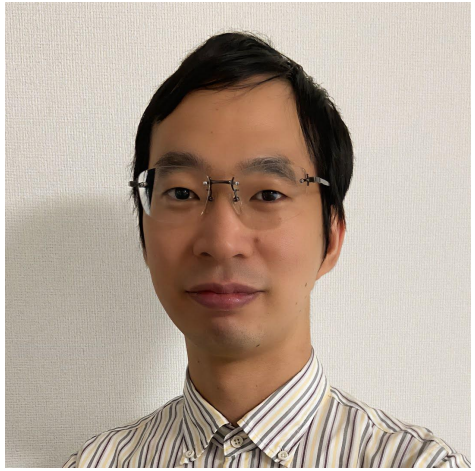
トヨタ自動車株式会社　鳥越貴智

takatomo_torigoe@mail.toyota.co.jp

Spatial Data Science Bootcamp 2023 Tokyo

空間データと私

- 2019~　　**シニアリサーチャー@トヨタ自動車**（大手町）
  - データ分析とシミュレーションの研究開発を担当
  - どちらも人や車の移動に関わるものなので、空間アルゴリズムや可視化の重要性を実感中……
- 2016~2019　**データ・機械学習エンジニア**@サイバーエージェント
  - 隣でジオターゲティング広告やっているのを眺めてたり……
- 2012~2016　**無線通信シミュレータエンジニア**@構造計画研究所
  - 測地系変換の実装をして、なかなか位置が合わなかったり……

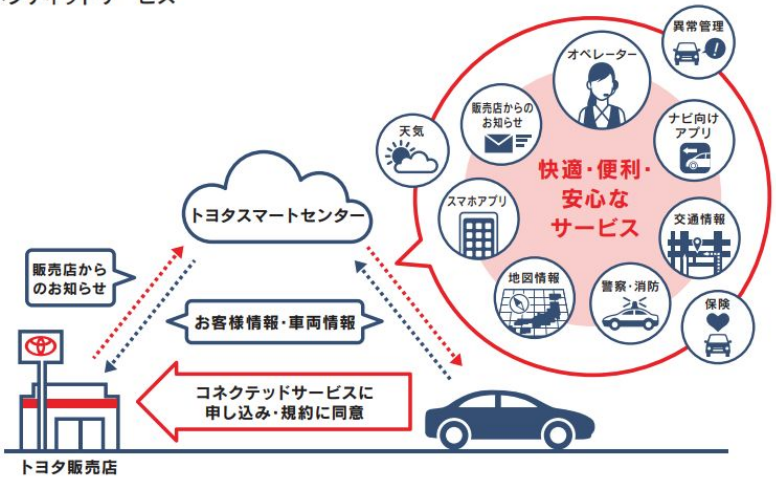


鳥越　貴智

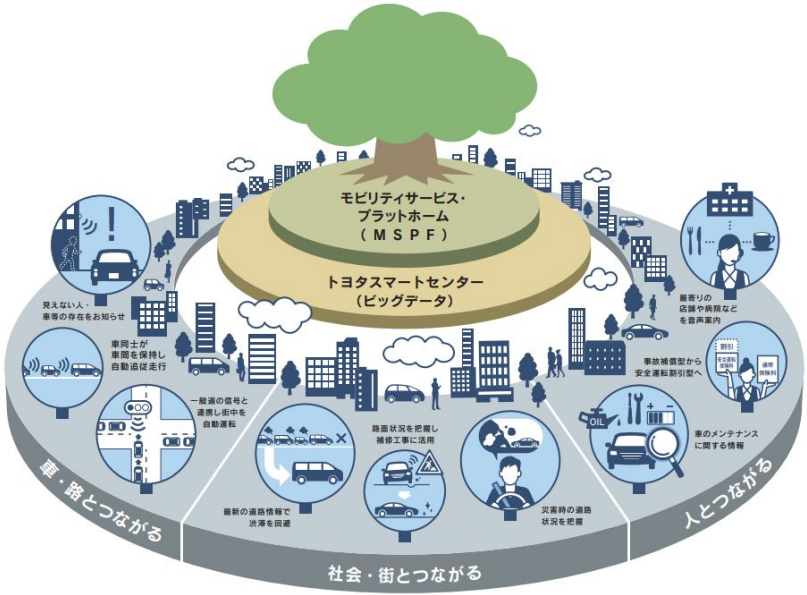
https://www.linkedin.com/in/takatomo-torigoe/

トヨタのコネクティッドサービス

トヨタのコネクティッドサービス



コネクティッドで広がるスマートモビリティ社会

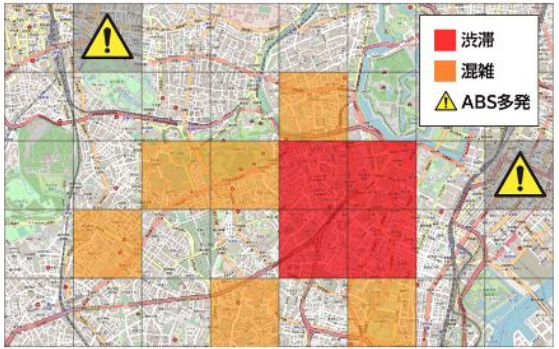


[コネクティッドカーから取得するデータの利活用・保護の取組みについて]

コネクティッドデータの統計化

移動データはプライバシー性が高いため、適切に統計化してから利活用。オーソドックスなのは、

メッシュ統計



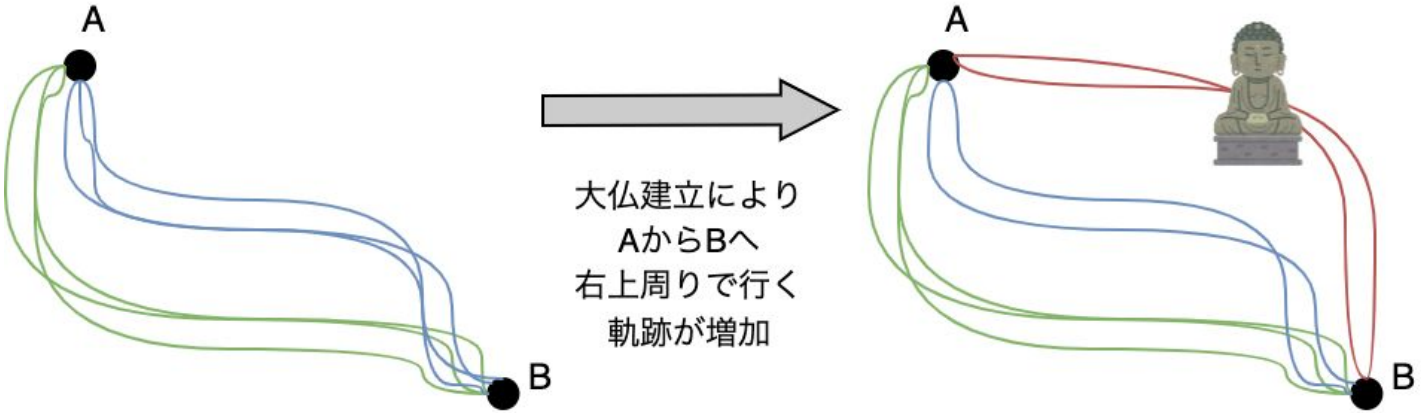
データの仕組み｜トヨタドライブ統計｜Toyota Biz Center

道路リンク統計

ソリューション「交通情報」｜トヨタドライブ統計｜Toyota Biz Center

これらは各車の移動について、通過したメッシュや道路リンクごとに統計化しているが、
どこから出発して・どこを通り・どこへ向かったのか、**移動軌跡として扱える統計手法**はないか……？

移動軌跡をクラスタリングできると……



- 複数の移動軌跡が全く同一の経路を辿ることは稀だが、大まかに同じものはたくさんある
- 上図のように、**類似した移動軌跡をクラスタリング**（＝分類）できると、

  クラスターごとに、移動時間・速度・停車回数・燃費……などの統計値を取って分析できる
- そのためには**軌跡の類似度**を定義する必要がある

軌跡の類似度とは……

軌跡の類似度は様々なものが提案されている。

GPSの測位誤差・間隔による影響が少ないものとして

右図の**CDTW** (Continuous Dynamic Time Warping)

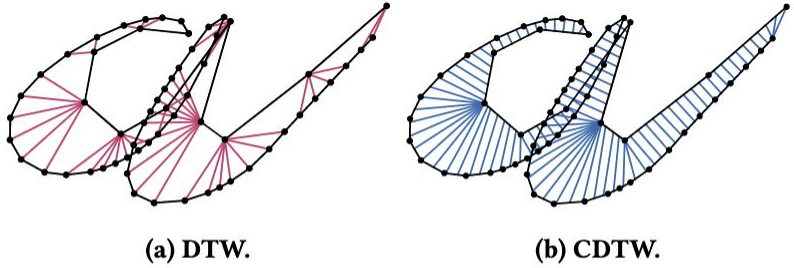
があり、これは2軌跡の良いマッチング関数を求め、

その**マッチング距離の平均**を計算するもの。

類似度が高い（＝CDTWが小さい）軌跡同士を

同じクラスターにまとめることができれば、

やりたいことができそう……！



(a) DTW. (b) CDTW.

Figure 2: An example of discrete and continuous alignments of points along trajectories of differing complexities by DTW and CDTW, respectively.

Milutin Brankovic et al., 2020
(k, l)-Medians Clustering of Trajectories Using
Continuous Dynamic Time Warping

$$\mathrm{CDTW}(P,Q) = \inf_{\sigma} \frac{\int_0^1 s\Big(P(t), Q(\sigma(t))\Big)\ \Big(L(P) + L(Q)\dot{\sigma}(t)\Big)dt}{L(P) + L(Q)}$$

軌跡クラスタリングの研究開発

- 移動軌跡のデータ規模はかなり大きい！
  - [人流] 携帯電話契約数：2億台（2021年、総務省）
  - [車流] コネクティッドカーの新車販売台数：推定370万台（2021年、富士経済）
- しかし軌跡の計算はとにかく重い……
  - 軌跡の類似度の計算量は、たいてい軌跡長の2乗以上
  - 軌跡のクラスタリングでは、その類似度計算（2軌跡の比較）を大量に行う

分散処理フレームワーク**Apache Sparkを用いた大規模な軌跡クラスタリング手法を研究開発**し、

日本データベース学会主催のDEIM2023にて発表

『大規模軌跡クラスタリングの類似度演算回数均一化による MapReduce モデルの分散処理効率化』

実験紹介：データセット

世界最大規模のオープンな人流データである

Datasets from the ATC shopping centerから

357万軌跡を取得。

範囲	900m^2
単位（XY）	ミリメートル
日数	延べ92日間
人数	21万
測位点数	36億

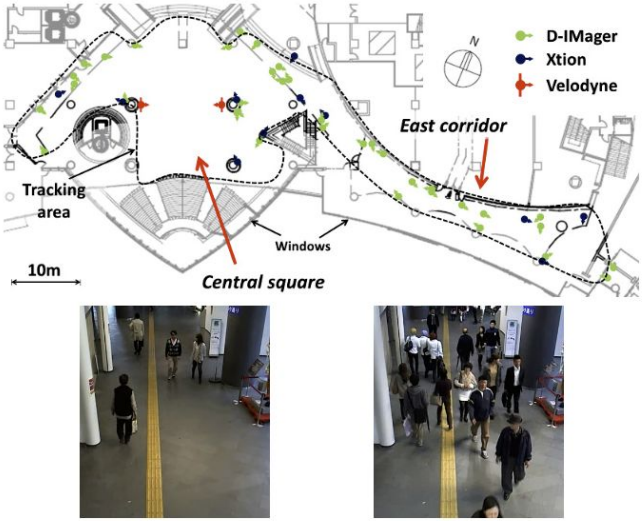


Fig. 7: Tracking area and sensor setup in ATC shopping mall. The dashed line shows the border of the area covered by the sensors. The photos below show the corridor area in the afternoon on a typical weekday (left) and weekend (right).

Dražen Brščić et al., 2013
Person tracking in large public spaces using 3-D range sensors

実験紹介：クラスタリング結果

右図は、**357万軌跡**を**16クラスタ**に分類し、

クラスターごとに100軌跡をサンプリング可視化したもの。

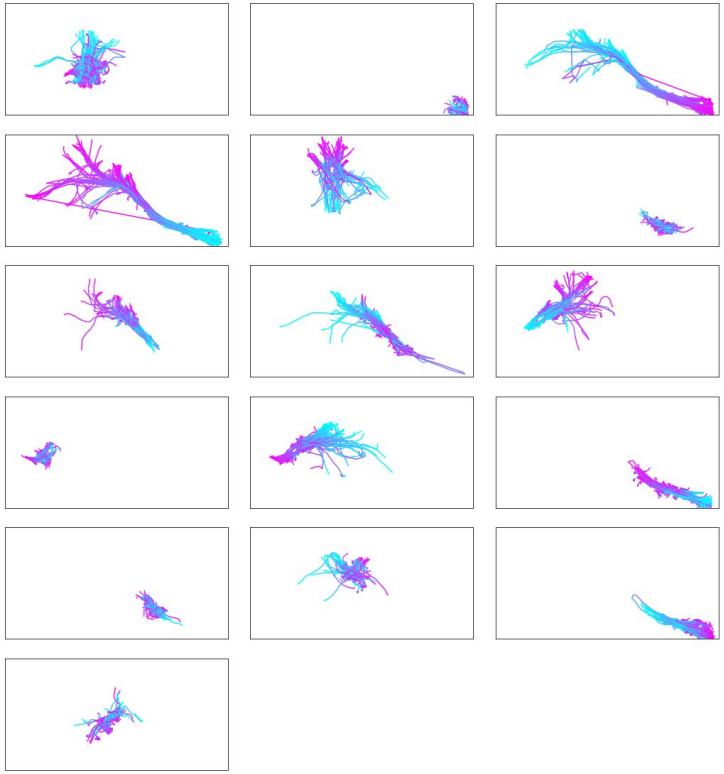
軌跡ごとに、始点（**水色**）から終点（**紫色**）へ

グラデーションの線として描いている。

ショッピングセンターを利用する人流として、

入口・経路・出口の近しい軌跡群が

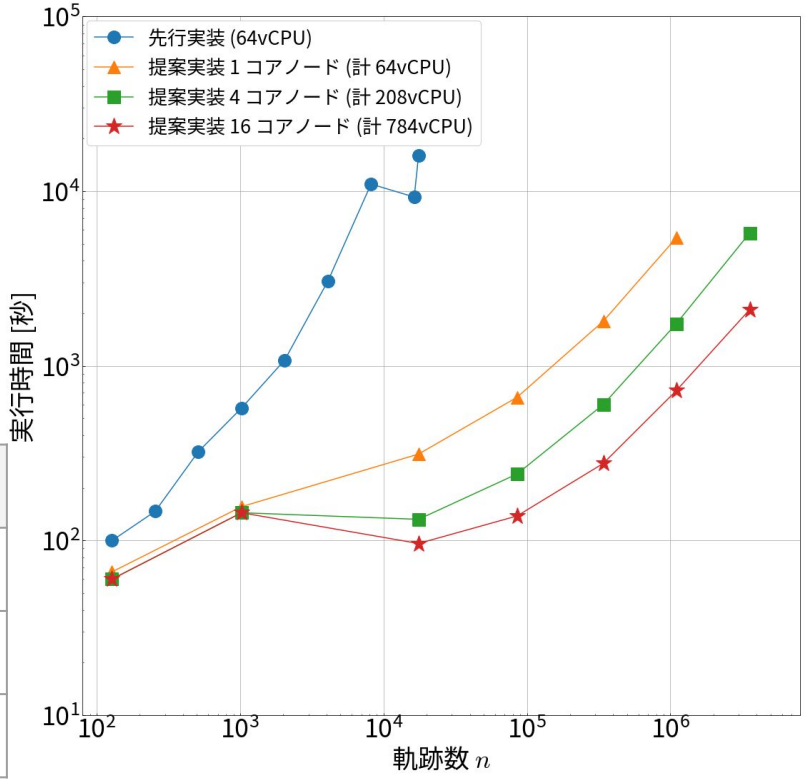
同じクラスターに所属している様が見て取れる。



実験紹介：スケーラビリティ

- 先行実装に対して、実行時間を桁レベルで短縮
- 計算機クラスターのコアノード数（ワーカー数）を増やすことで、実行時間を一定に抑えたまま扱う軌跡数を増やせるように

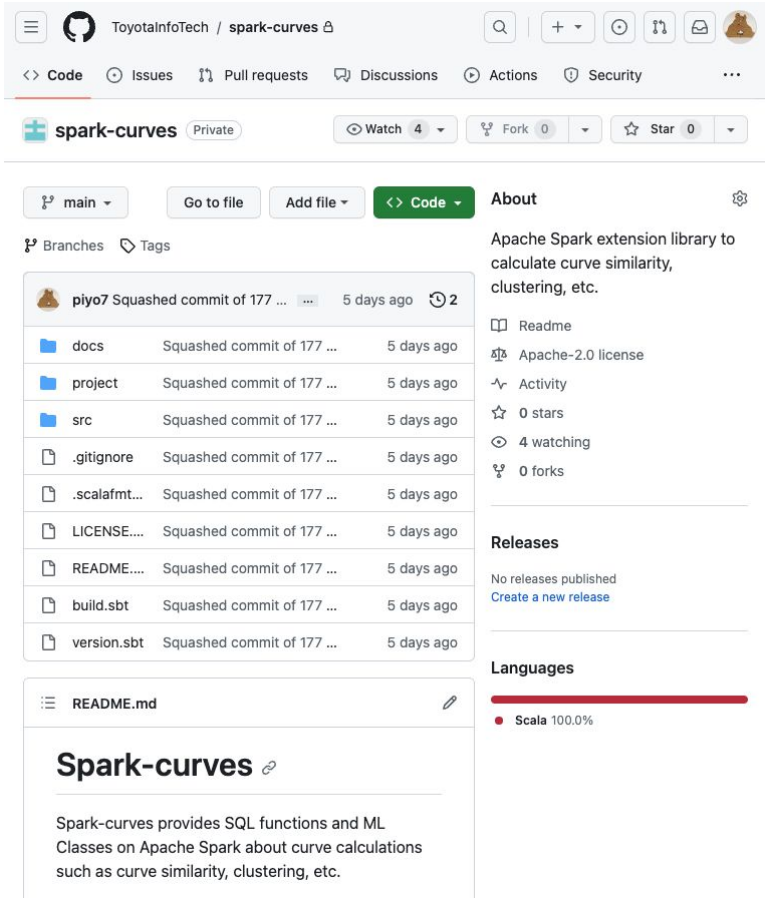
コアノード数	軌跡数	実行時間
1コアノード	340,960軌跡	30分
4コアノード	1,099,538軌跡	29分
16コアノード	3,578,281軌跡	25分



ライブラリ紹介

- 本研究で開発したSpark拡張ライブラリをOSS化！
  - 軌跡クラスタリングを行うMLクラス
  - 軌跡の類似度計算を含む20以上のSQL関数
- SQL関数についてはApache Sparkがサポートしている
  SQL・Scala・Python・Rどの言語からも呼び出し可
- AWSのEMR、GCPのDataproc、Databricksなど各種
  Apache Sparkマネージサービスにインストール可

https://github.com/ToyotaInfoTech/spark-curves



ライブラリ利用コード例

- CDTWのSQL関数利用クエリ

```
> SELECT continuous_dynamic_time_warping(array(array(-1d, 1d), array(0d, 1d), ...

2.722
```

- 軌跡クラスタリングのMLクラス利用コード

```
val klClustering = new KLClustering().setK(16).setL(32).setIter(4)
  .setCurveSimilarityFunc("continuous_dynamic_time_warping")
...

val klClusterModel = klClustering.fit(trajectories)
val result = klClusterModel.transform(trajectories)
```

まとめ

- 移動データの統計化として、軌跡の類似度でクラスタリングする手法がある
- 軌跡クラスタリングを用いた人流・車流分析の実用については探索中
- 分散処理性能に優れたアルゴリズム実装をOSSとして公開
  - まだまだ未整備ですが、フィードバック大歓迎です！
  - https://github.com/ToyotaInfoTech/spark-curves