機械学習による
通信データ生成へのアプローチ

長岡技術科学大学
大学院工学研究科
准教授 渡部 康平

自己紹介

渡部 康平
博士 (情報科学)

[所属]
長岡技術科学大学
大学院工学研究科

[職位]
准教授

[研究分野]
QoS計測，トラヒック・モビリティのモデリング etc.

2011年 3月	首都大学東京 大学院システムデザイン研究科 博士前期課程 修了． 修士 (工学) 取得.
2012年 4月	日本学術振興会 特別研究員 (DC2) に採用. (2014年3月まで)
2014年 3月	大阪大学 大学院情報科学研究科 博士後期課程 修了． 博士 (情報科学) 取得.
2014年 4月	長岡技術科学大学 大学院工学研究科 電気電子情報工学専攻 助教に着任． (2019年10月まで)
2019年 11月	長岡技術科学大学 大学院工学研究科 電気電子情報工学専攻 准教授に着任． 現在に至る.

本日の内容

- ◆通信データにおけるデータ生成
- ◆活用する機械学習の技術
- ◆ネットワークトポロジ生成技術
- ◆ネットワークトラヒックの生成技術
- ◆通信デバイスの移動軌跡の生成技術
- ◆本日のまとめと今後について

通信データにおけるデータ生成

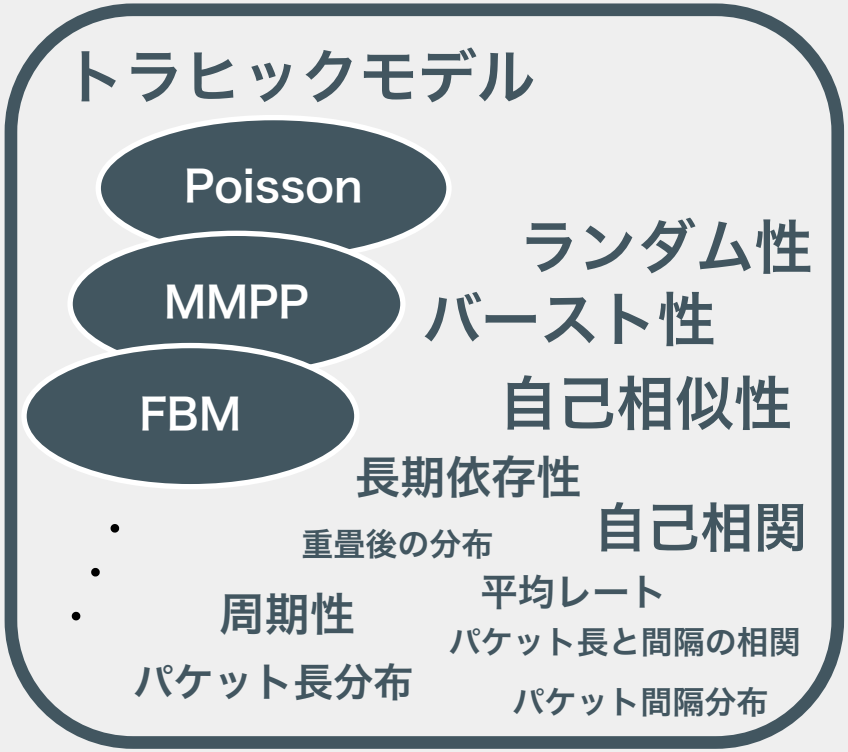
ネットワークにおける実データ問題

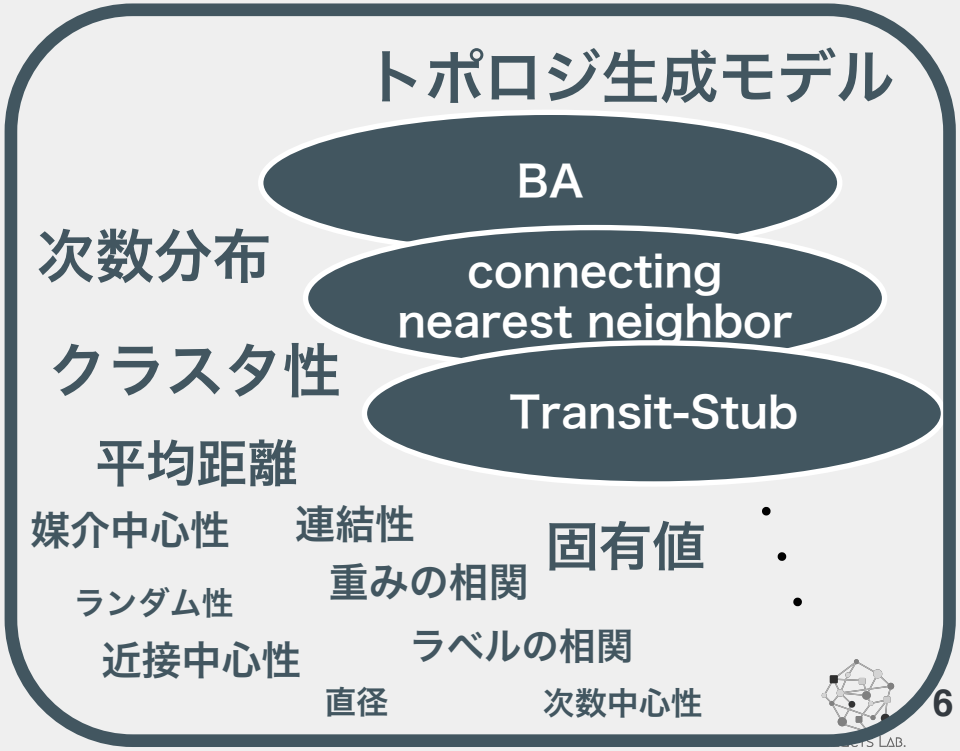
- ◆ネットワークに関する製品テストやシミュレーションは重要.
- ◆説得力のあるテストやシミュレーションには**実データが必要不可欠**.
  - ◆トポロジ，トラヒック，フロー，デバイスの移動軌跡，etc.
- ◆しかし，**多くの実データは非公開**.
  - ◆通信データはプライバシーの塊.
  - ◆一部の企業や組織内で収集され，独占的に利用される.
  - ◆公開されているのは一部のデータのみ.
- ◆GAFAによるデータ独占問題のネットワーク版.
- ◆独占により，多くの研究者が制約を受けている.

必要な特性を備え，丁度いい規模で，必要な組み合わせが揃った
実データが入手できることはほとんどない.

統計的モデルで生成することの問題点

- ◆実データを利用できない場合，統計的モデルでデータを生成する.
- ◆統計的モデルは，着目する特性を再現するが，**実データが持つ特徴の一面を再現しているに過ぎない**.
  - ◆一面を再現するにも専門性を要するチューニングが必要.
  - ◆実データの特性を多面的に表現できないため，実データを使った実験結果とは乖離がある.





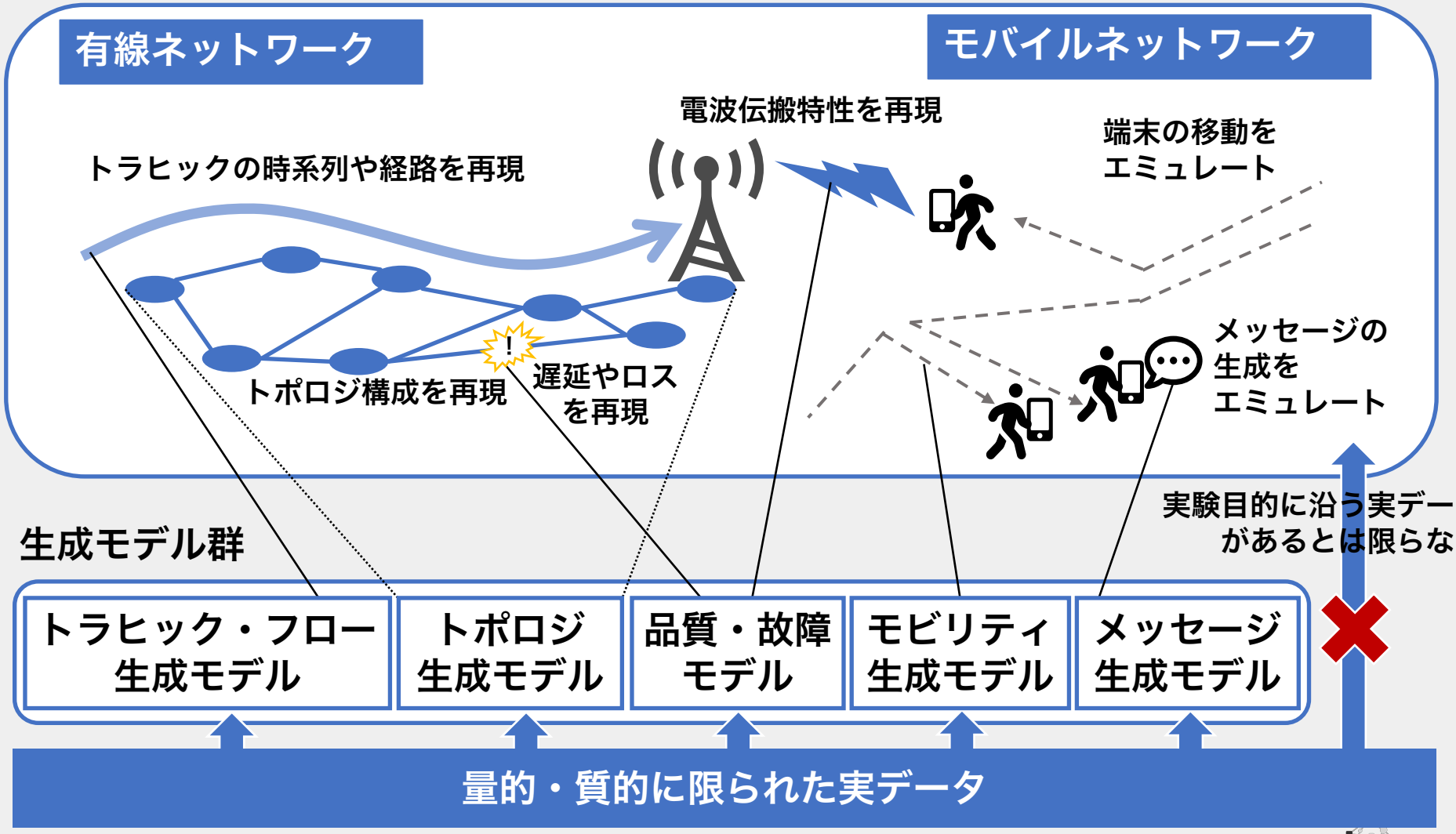
機械学習的なアプローチ

◆近年の機械学習技術の目覚ましい発展は，従来では明示的なモデル化が難しい複雑な生成問題にも，ブラックボックスモデリングの適用を可能にしている.

本プロジェクトでは，**実データの代替として利用可能**で，かつ**実験の利用目的に沿った任意の特性を持つ疑似データ**の生成を目指す.

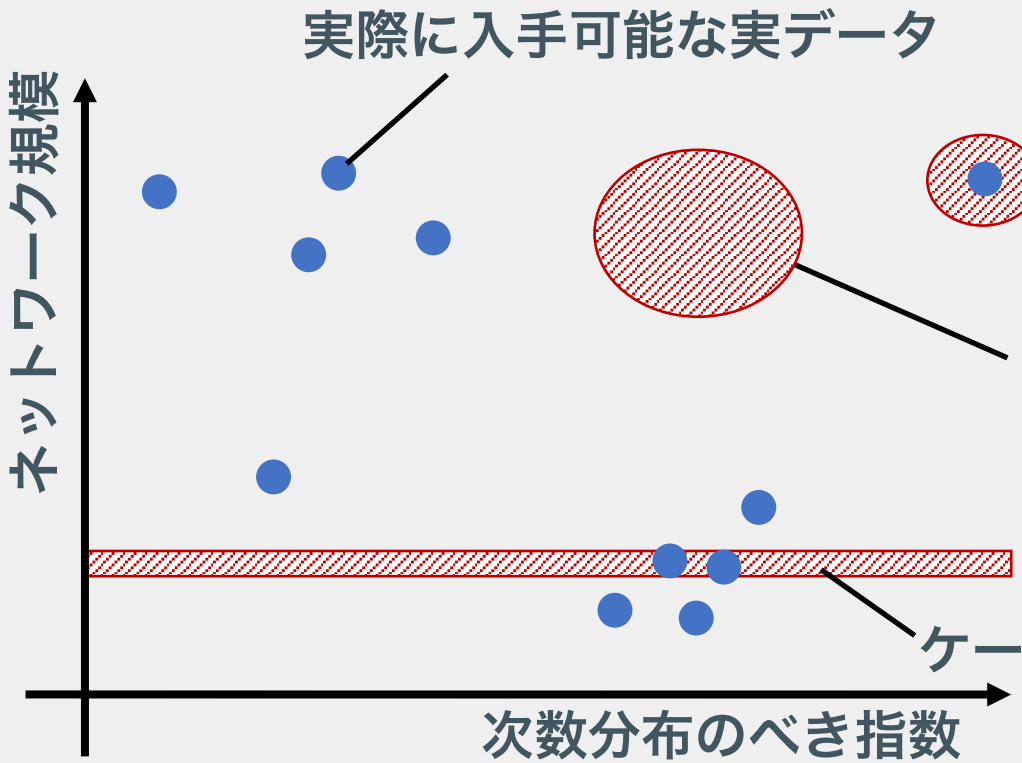
プロジェクトの概要

シミュレーション・実験環境



実データを模擬したデータの有用性

- 実データを模擬したデータは，結果の信頼性を保ちつつ，柔軟なテスト・シミュレーションを実現する.
  - 実データの多くの特性を維持.
  - 量的制約がない.
  - パラメータチューニング可能



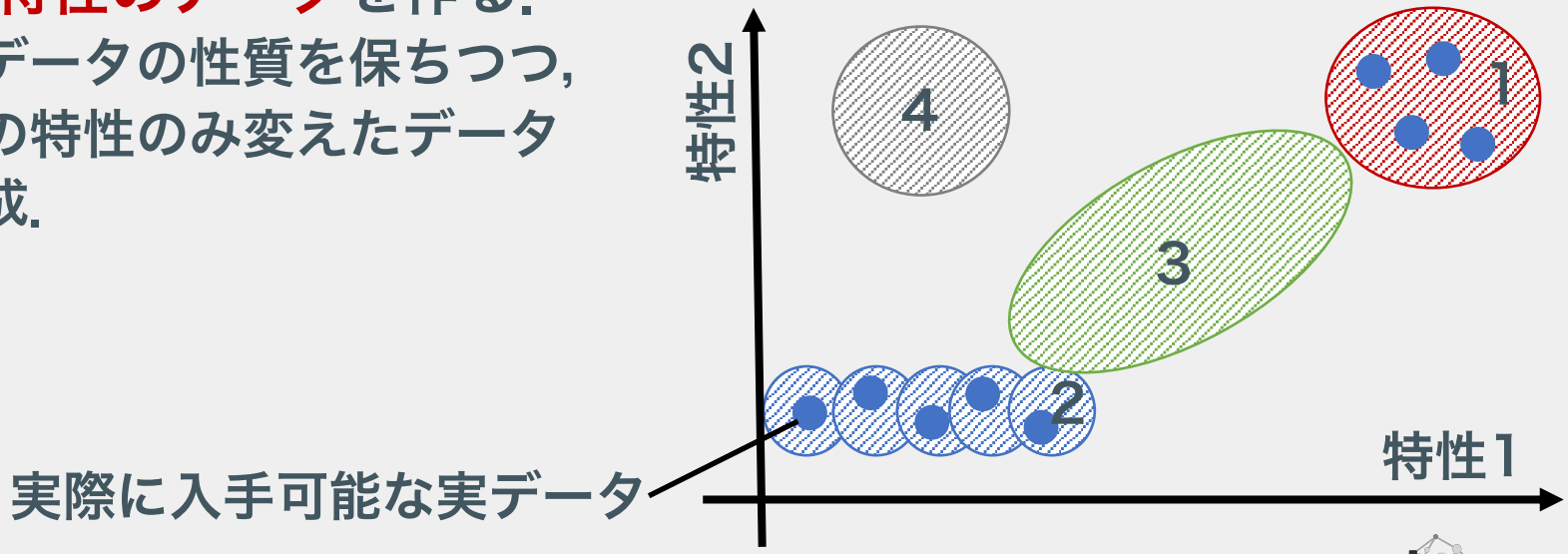
ケース1: 似たようなデータがたくさん欲しい

ケース2: 小規模なデータしか手に入らないが規模の大きなデータがほしい

ケース3: 特定のパラメータを変化させながら実験したい

目標到達までのステップ

1. 統計的に**同じ性質のデータ**を作る.
   - ◆データとしては異なるが，平均的特性は類似するものを生成.
2. **特定の性質を持つデータ**を狙って作る.
   - ◆与えられた複数のデータのうち，狙った性質を持つもののみを生成.
3. **中間的特性のデータ**を作る.
   - ◆与えられたデータの中間的特性のデータを生成.
4. **任意の特性のデータ**を作る.
   - ◆元のデータの性質を保ちつつ，任意の特性のみ変えたデータを生成.



ネットワーク関連データ生成での課題

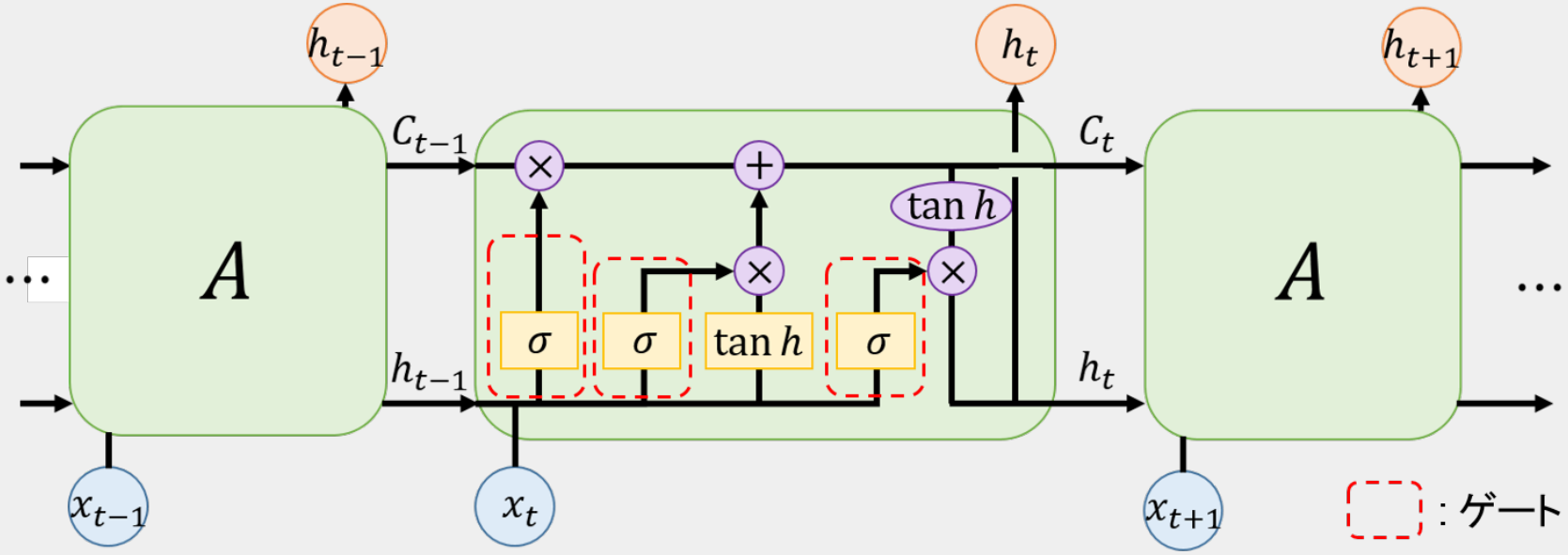
画像を中心に発展してきたニューラルネットワークを始めとする機械学習的なアプローチをネットワークに適用するには，いくつかの問題が予見される.

- ◆分布(の裾)に対する要求条件がシビア.
  - ◆トラヒックデータでは，パケット長や間隔の分布が重要.
  - ◆トポロジデータでは，次数分布の裾の挙動が重要.
  - ◆統計的なモデルの考え方で，機械学習に制約を掛ける？
- ◆生成結果の良し悪しの判別方法が難しい.
  - ◆実データと似ていることをどう証明するか.
  - ◆検定などを使って多面的に特性を再現していることを示す？
  - ◆既存の判別手法を騙せるレベルを目指す？

活用する
機械学習の技術

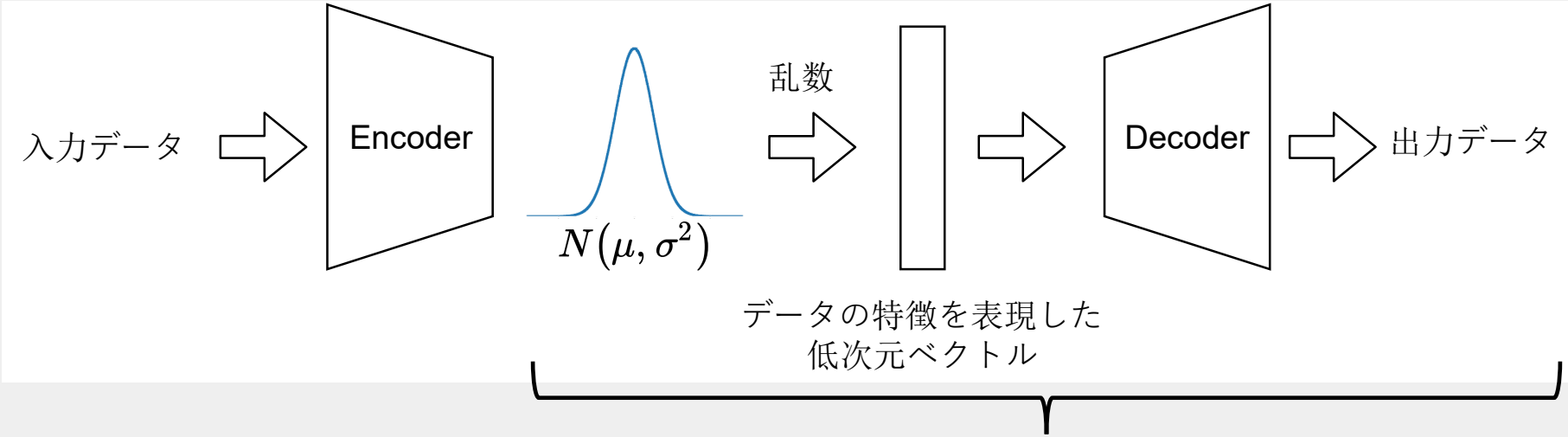
活用する機械学習の技術～LSTM

◆時系列をニューラルネットワークで処理するための枠組み.
- ◆x_t にデータを再帰的に入力していくことで，**時系列データを処理可能.**
- ◆長さ n のデータを入力して h_n を得れば，**埋め込み (データの特徴を表現する低次元ベクトル) を取得可能.**
- ◆h_t から x_{t+1} を推定するようにすれば，**時系列の生成に利用可能.**



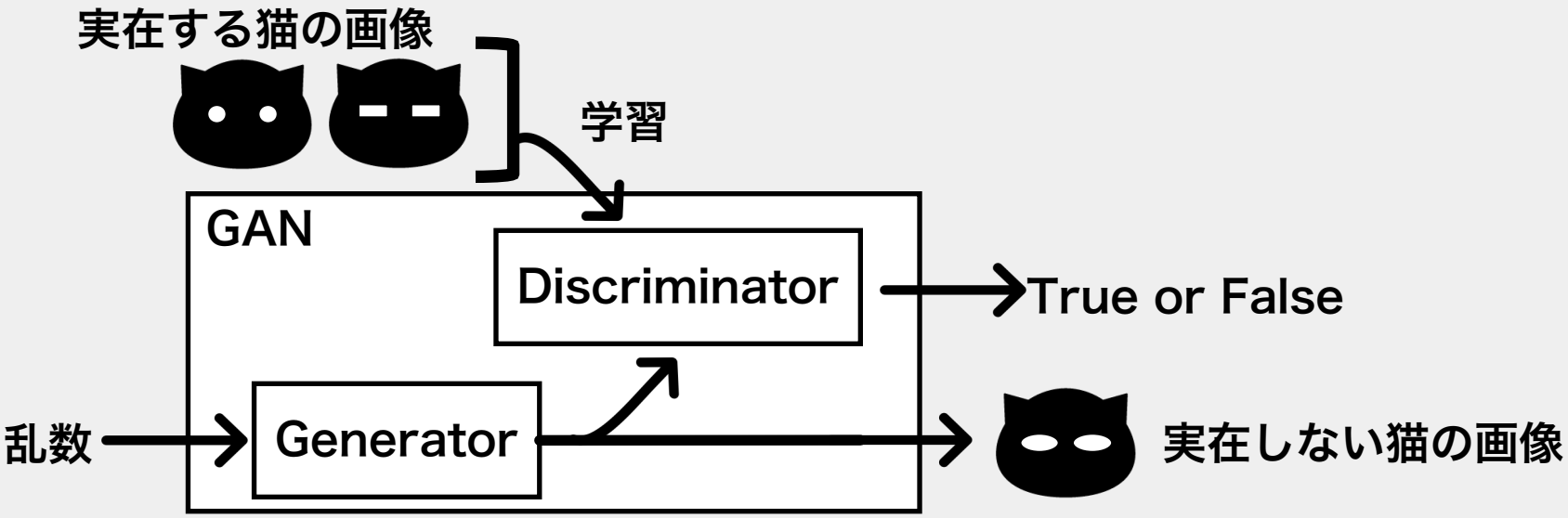
活用する機械学習の技術～VAE

- Variational Autoencoder (VAE) は，**入力データを低次元にベクトルとして表現して，乱数からデータを生成するための枠組み.**
  - ニューラルネットワークにより，データの特徴を表現する低次元ベクトルの分布を推定.
  - 低次元ベクトルの分布として正規分布を仮定.
  - 乱数から生成するため，**学習データに含まれないデータを生成可能.**



活用する機械学習の技術〜GAN

- ◆Generative Adversarial Networks (GAN) は，2つのニューラルネットワークからなる生成モデル.
  - ◆Generator: 乱数を元にDiscriminatorを騙せるデータを生成
  - ◆Discriminator: Generatorの偽のデータと本物のデータを識別.
- ◆2つのニューラルネットワークが互いに**競いながら精度を上げる**.



ネットワーク
トポロジの生成

ネットワークトポロジの生成

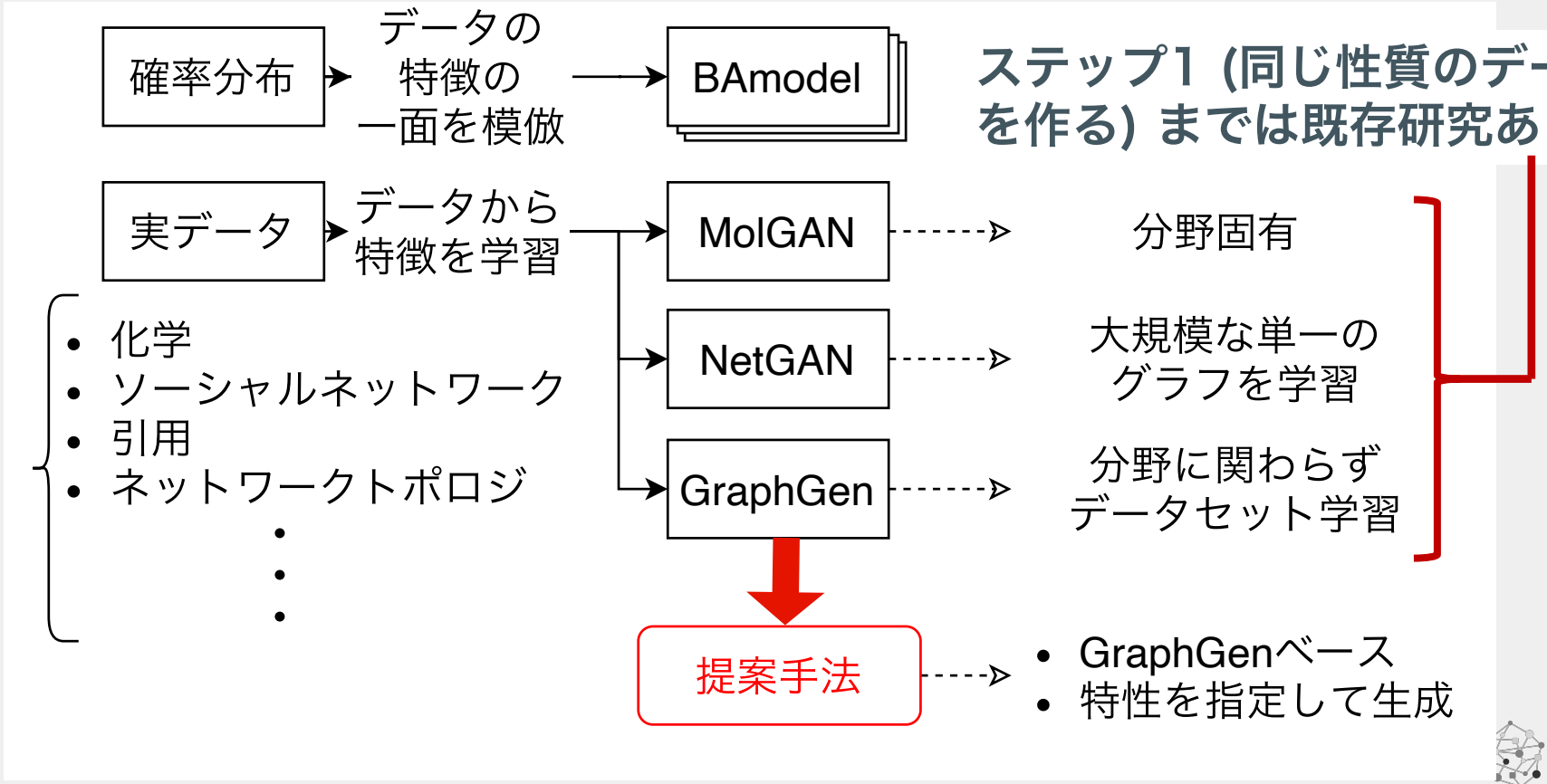
- ネットワークトポロジのデータは様々なテスト，シミュレーションで活用されている.
  - 通信プロトコルの検証，通信品質の評価，噂や感染症の伝搬，etc.
- ネットワークトポロジの実データに関しては，様々な特性が研究されてきた.
  - スケールフリー性: 次数分布のべき指数で定量化
  - クラスタ性: クラスタ係数で定量化
  - 他にも，ネットワークトポロジ (グラフ) に関する特徴量は多数.

ネットワークトポロジのデータを学習して，**任意の特性のトポロジデータを出力するモデル**を目指す.

既存の研究

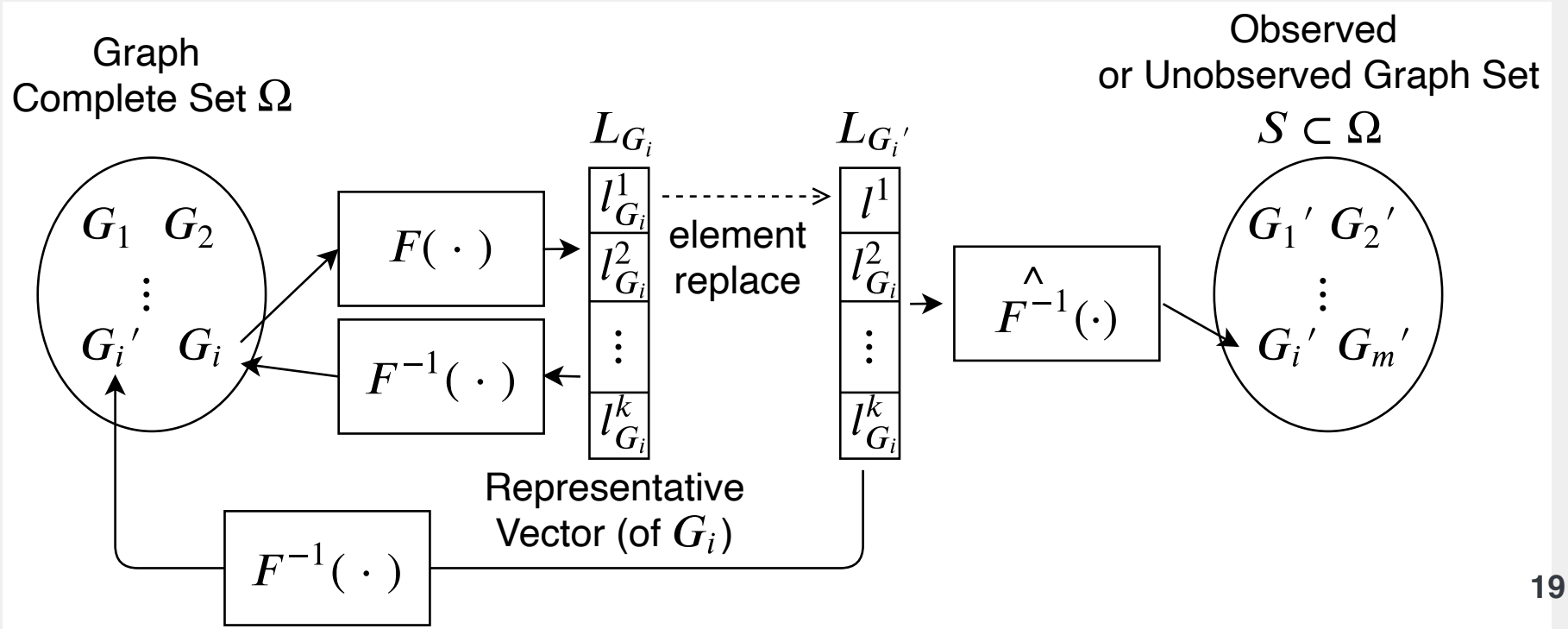
◆ネットワークトポロジの生成に関しては，**統計的生成モデルを中心に古くから研究**されてきた.

◆近年では，機械学習的なアプローチを使い，**データを元に生成する研究も登場**してきている.



問題の定式化

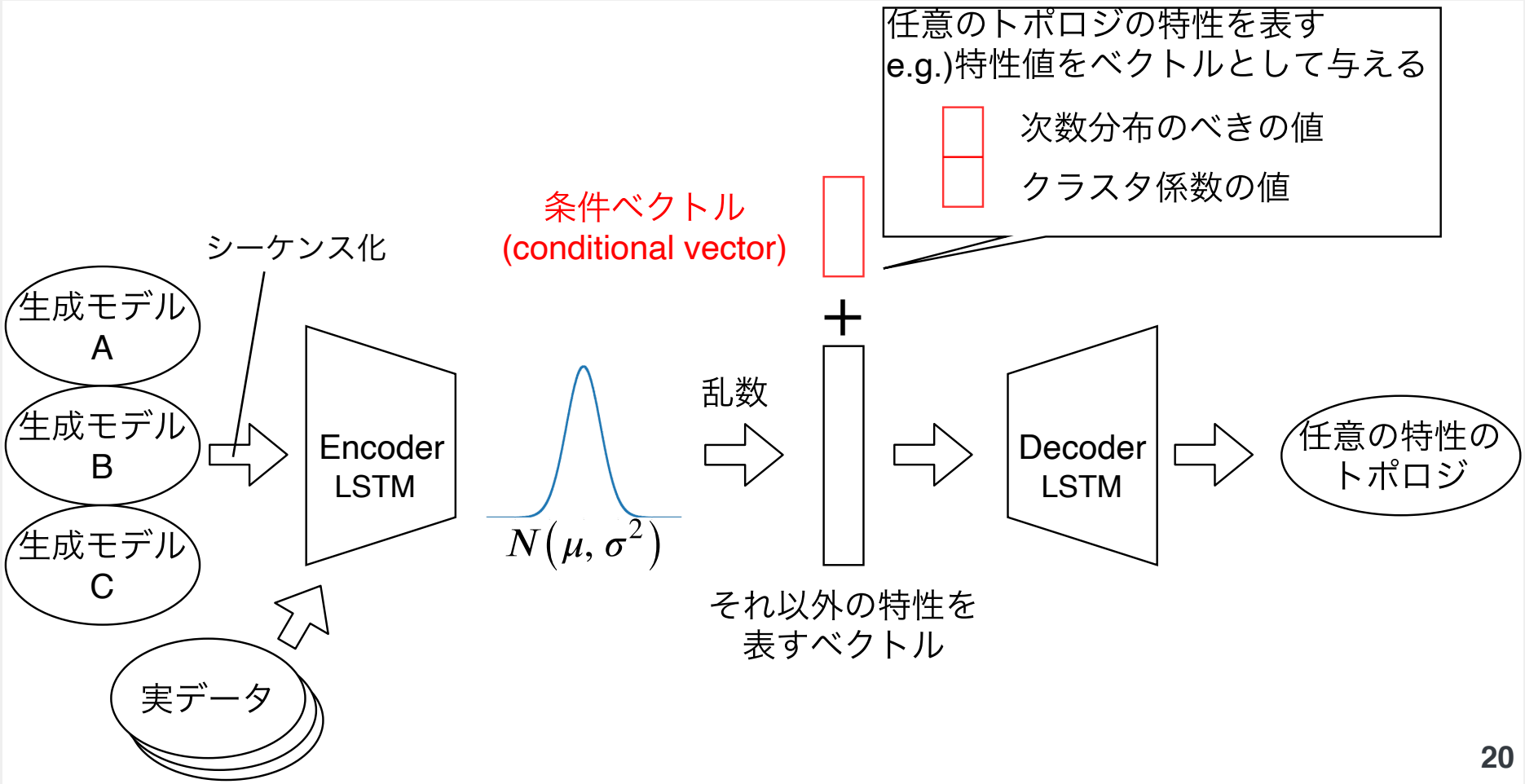
- ◆任意のグラフ G_i をグラフの特性を表現するベクトル L_{G_i} に変換する写像 F を考える.
  - ◆L_{G_i} の各成分はグラフ特性: 次数分布のべき指数，クラスタ係数など.
- ◆L_{G_i} の一部の成分(特性)を別の値に変更し，F の逆写像 F^{-1} で変換すると，他の特性を維持しつつ任意の特性を持つグラフになる.
- ◆本研究で解く問題は，一部の G_i から F^{-1} を推定する問題.



提案モデル

◆条件付きVAEを使って実装.

◆VAEでエンコードした**グラフの特徴を表すベクトルに，変更したい特性を結合**することで，出力するグラフを制御.



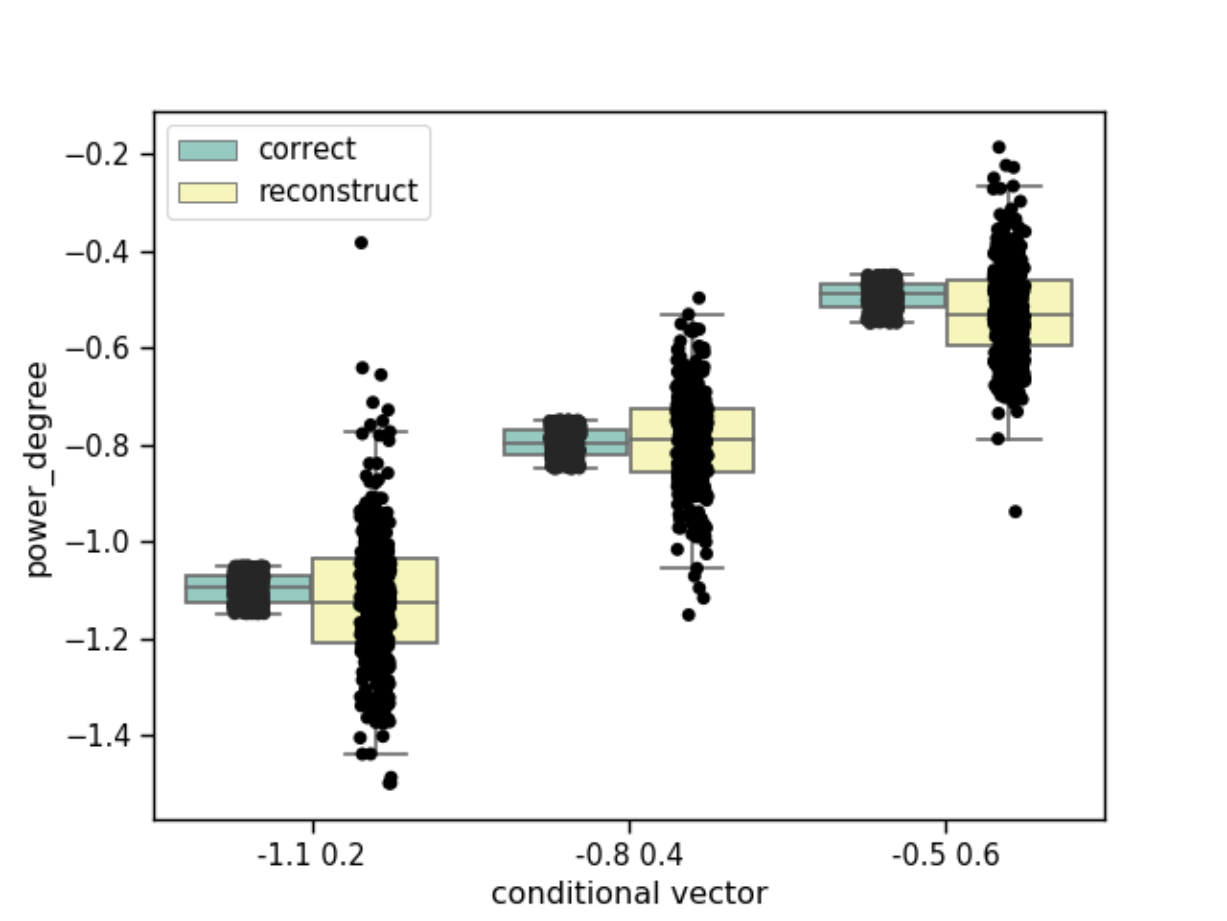
学習の条件

- LSTMの各種パラメータは以下の通り.
  - 隠れ層の次元: 256
  - 埋め込み層の次元: 128
  - 正規分布に変換を行う線形層の次元: 20
  - ミニバッチ数: 60
  - エポック数: 400
- Connecting Nearest Neighbor (NN) モデルで生成した25ノードのグラフから，特定の特性を持つデータのみ抽出して学習.
- 次数分布のべき指数とクラスタ係数を条件ベクトルとして指定.

ID	次数分布のべき	クラスタ係数	データ数
-0.5_0.6	-0.5	0.6	400
-0.8_0.4	-0.8	0.4	400
-1.1_0.2	-1.1	0.2	400

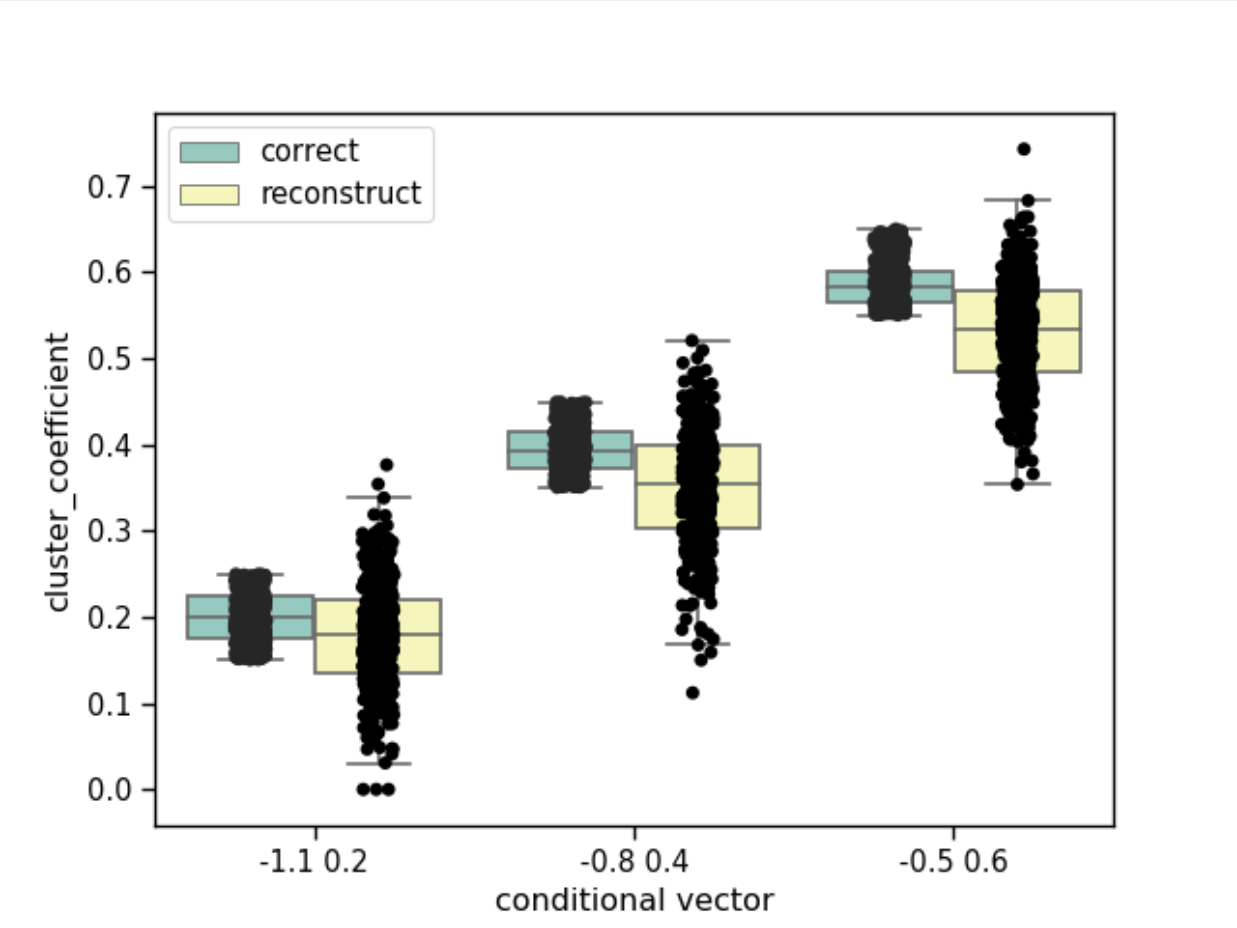
学習データの再現

◆条件ベクトルを変更した場合の生成結果の次数分布のべき指数.
  ◆ばらつきはあるものの，ある程度指定した値の周辺を生成している.



学習データの再現

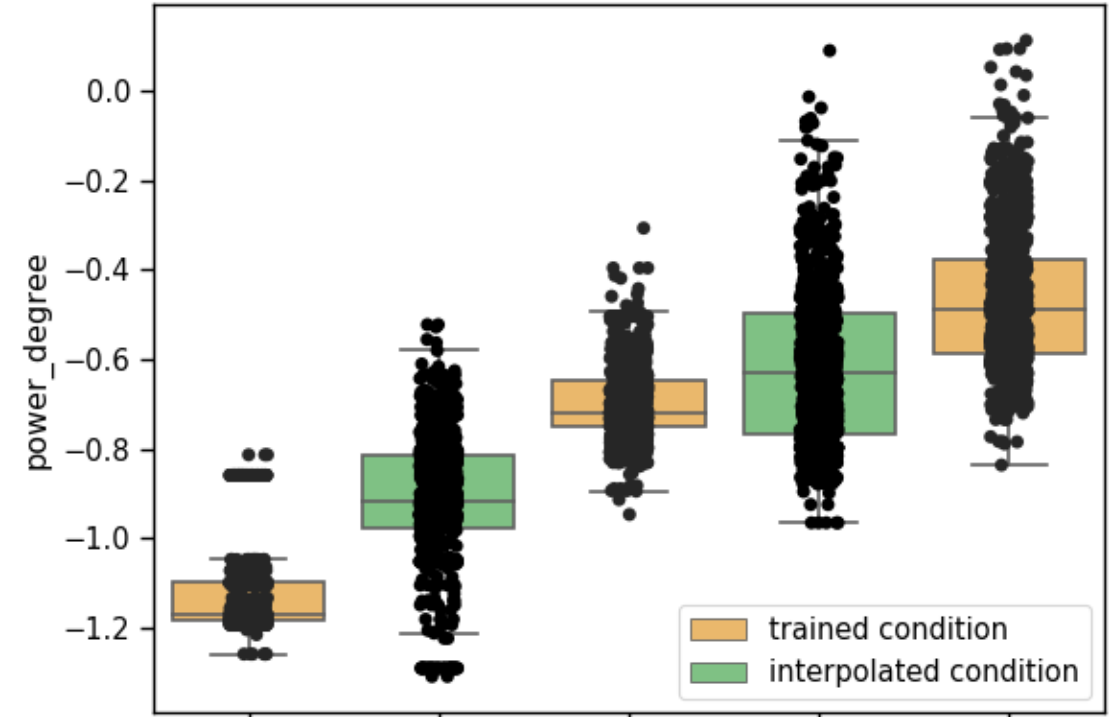
◆条件ベクトルを変更した場合の生成結果のクラスタ係数.
  ◆やや下にずれているが，同様の結果.



中間的特性のデータの生成

◆データセット中に含まれない中間的な値を条件ベクトルとして与えた場合の次数分布のべき指数.

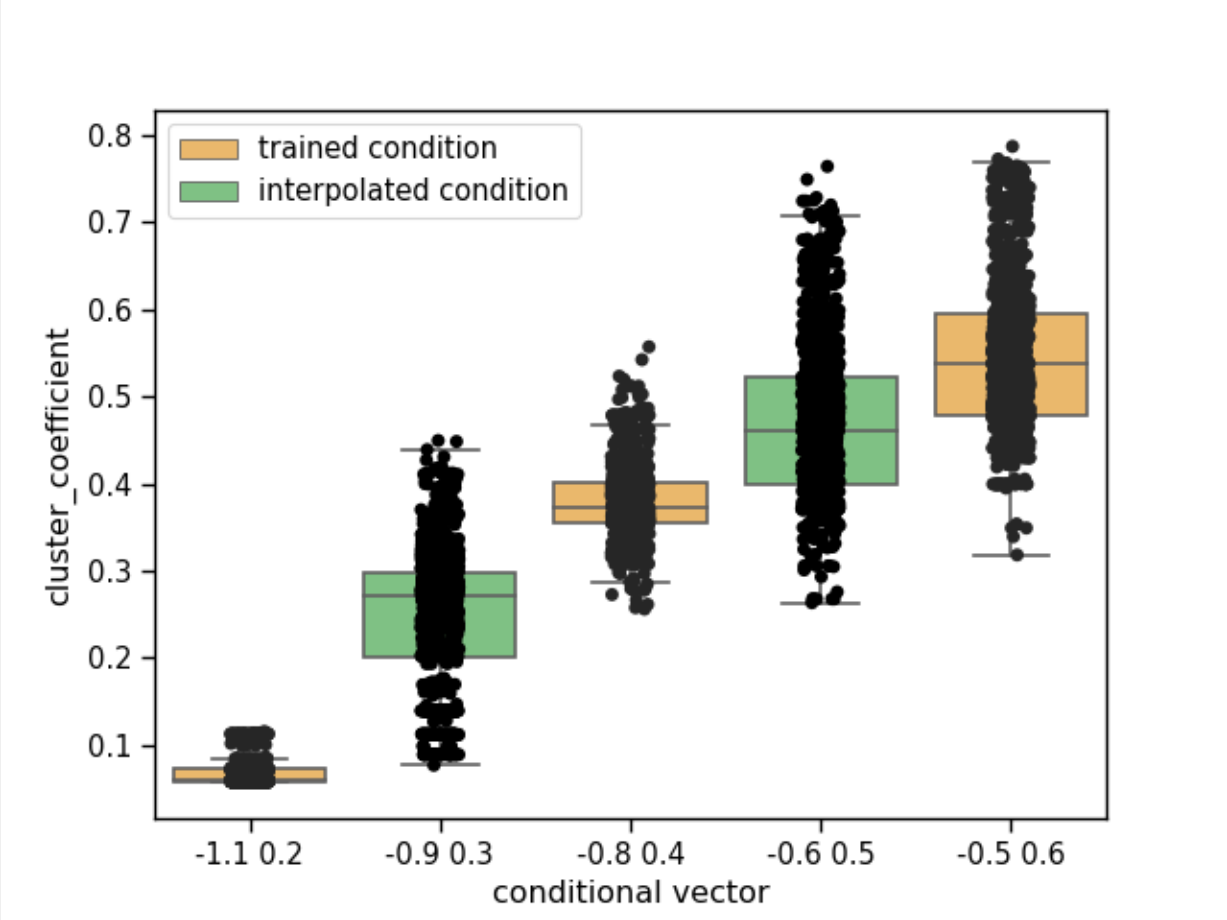
◆中間的な値を生成しており，条件ベクトルで特性の調整できている.



中間的特性のデータの生成

◆データセット中に含まれない中間的な値を条件ベクトルとして与えた場合のクラスタ係数.

◆次数分布同様，条件ベクトルにより特性が調整できている.



ネットワークトポロジ生成のまとめ

ネットワークトポロジのデータを学習して，**指定した値に近い特性を持つトポロジデータを生成するモデル**を開発.

◆次数分布のべき指数，クラスタ係数について指定した値のデータのみを生成可能.

◆学習に含まれていない中間的な特性を持つデータを生成可能.

◆未検証な項目.

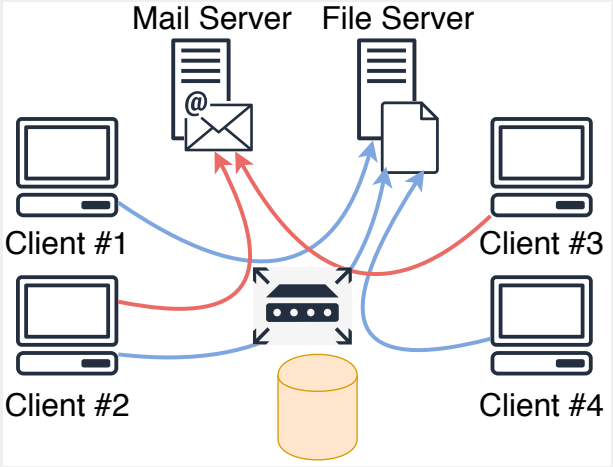
- ◆中間的ではなく，実データの境界領域を生成したときの挙動.
- ◆条件ベクトルとして与えていない特性の挙動が維持されているか.
- ◆次数分布のべき指数とクラスタ係数以外の任意の特性に適用可能か.

ネットワーク
トラヒックの生成

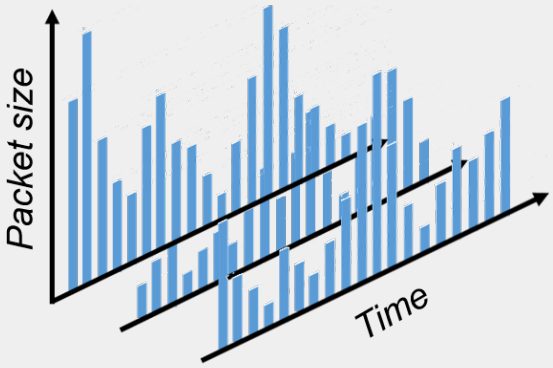
ネットワークトラヒックの生成

◆ネットワークのトラヒックデータは，通信システム実験で重要.
  ◆負荷テスト，品質評価，異常検知，etc.
◆トラヒックデータを生成することで，様々な実験が可能になる.
◆生成したいトラヒックデータには，様々な粒度が考えられる.

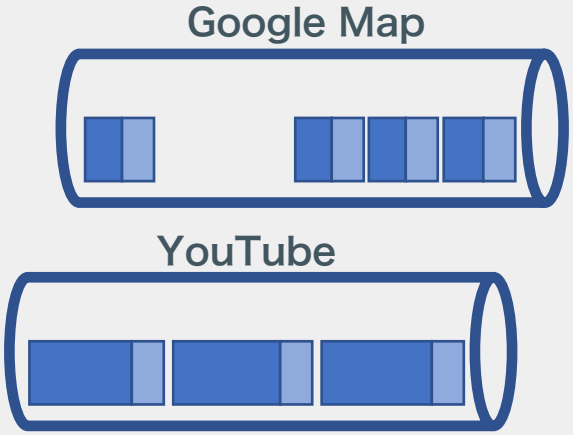
フローレベル



バイトレベル



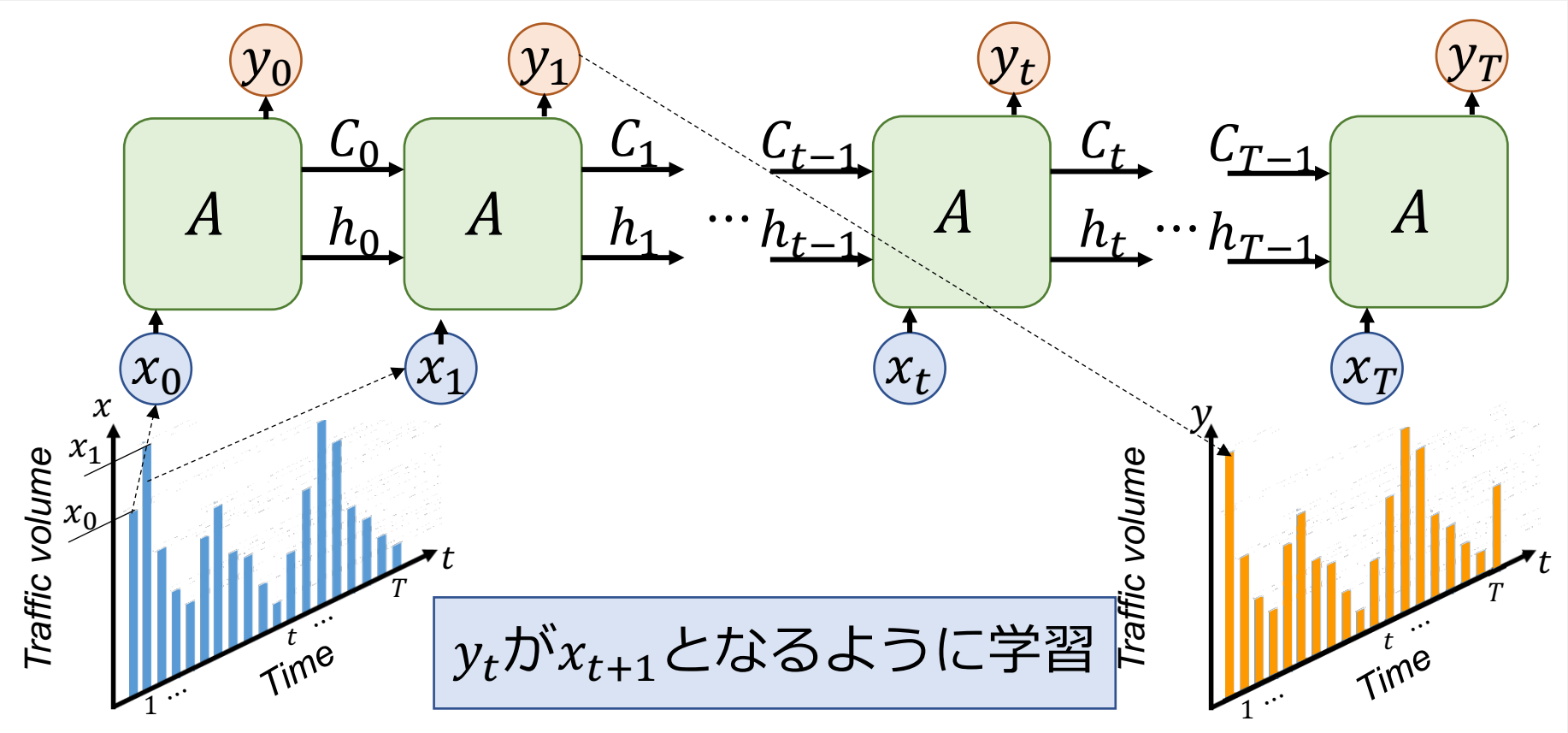
パケットレベル



バイトレベルのトラヒック生成

◆LSTMでバイトレベルのトラヒックデータを学習

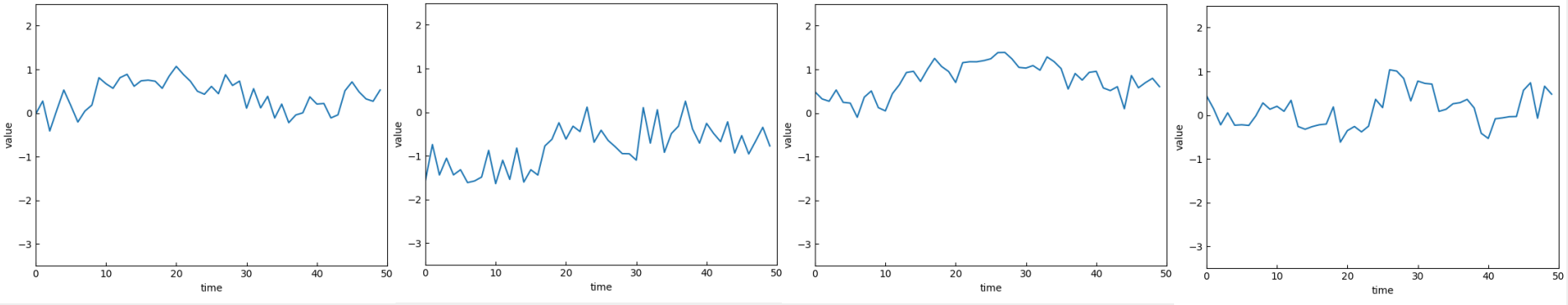
◆再帰的に未来の出力を生成することでデータ生成.



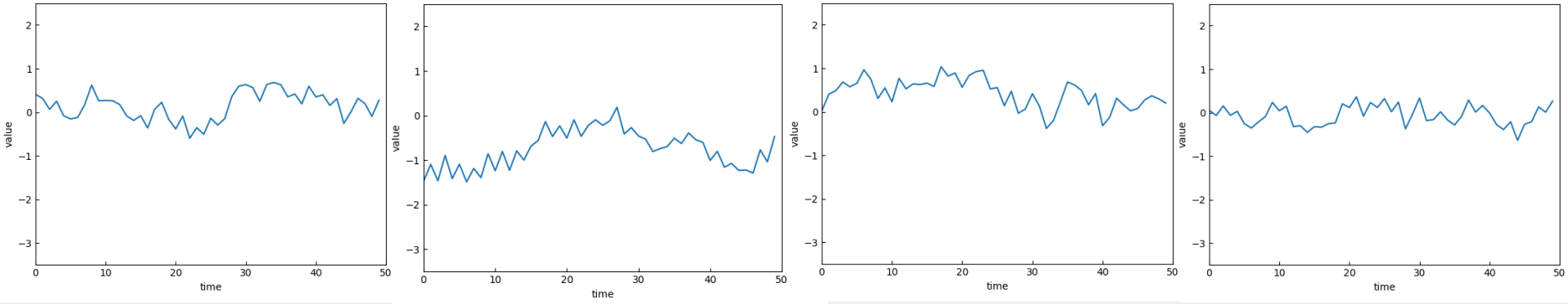
LSTMによるバイトレベルの生成

◆WIDEプロジェクト[1]のトラヒックデータを学習して生成.

■ 学習データの一部



■ 生成データの一部

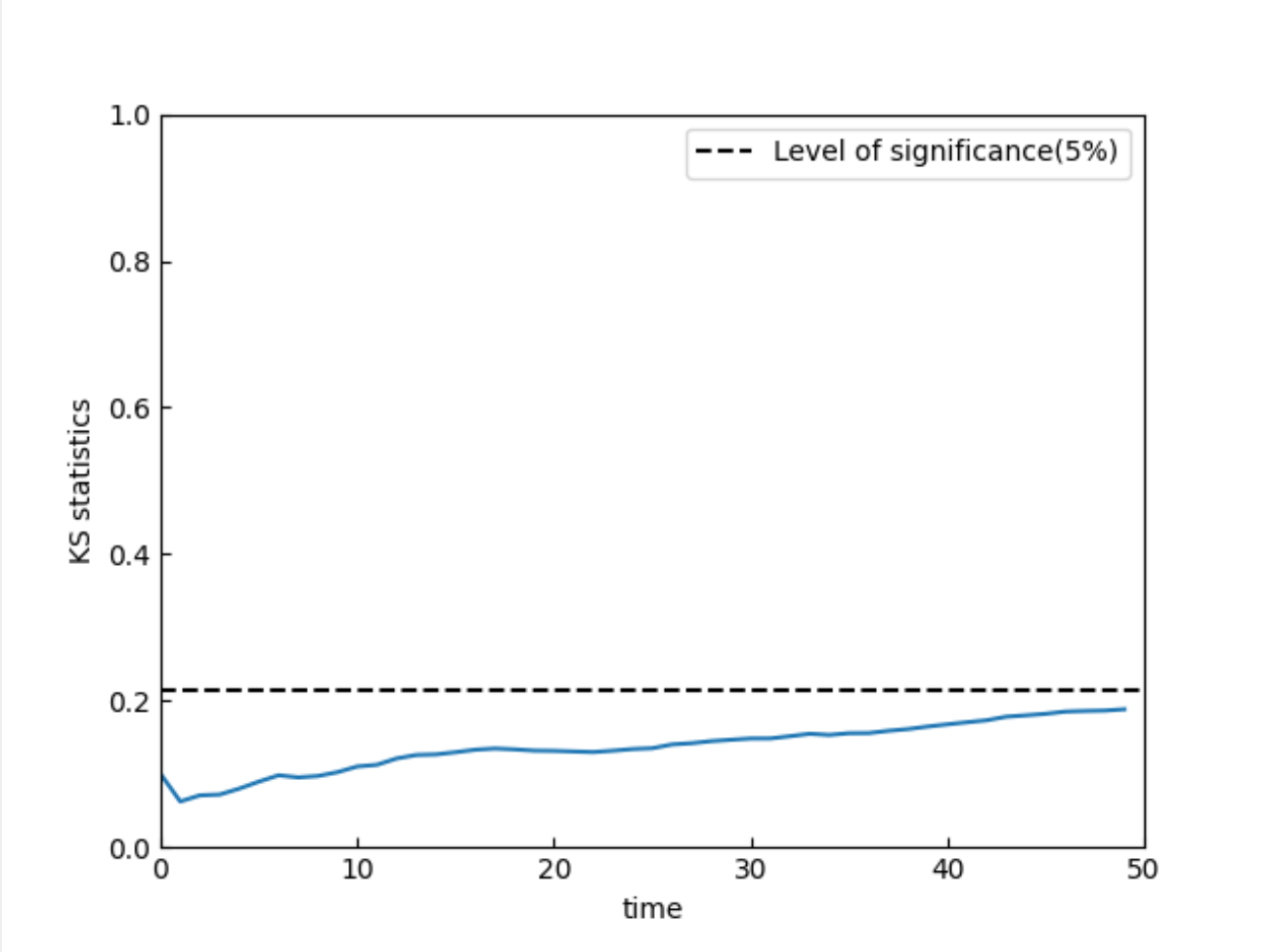


[1] Kenjiro Cho, Koushirou Mitsuya, and Akira Kato. Traffic data repository at the wide project, USENIX Association, pp. 51, 2000.

各時刻ごとのバイト数の分布

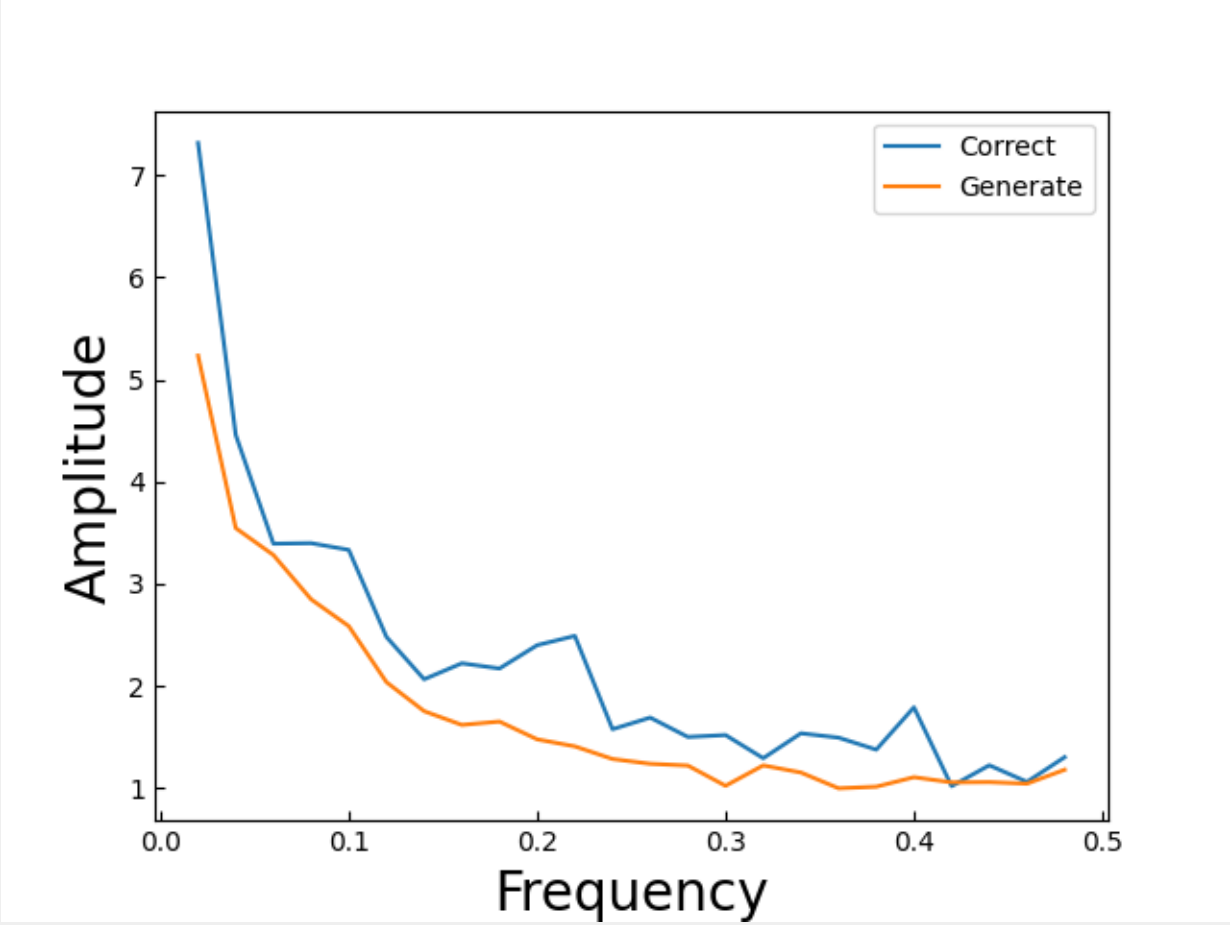
◆学習データと生成データで各時刻ごとのバイト数の分布を，サンプル数80としてKS検定で検証.

◆棄却されないので，それなりに近い分布になっている.



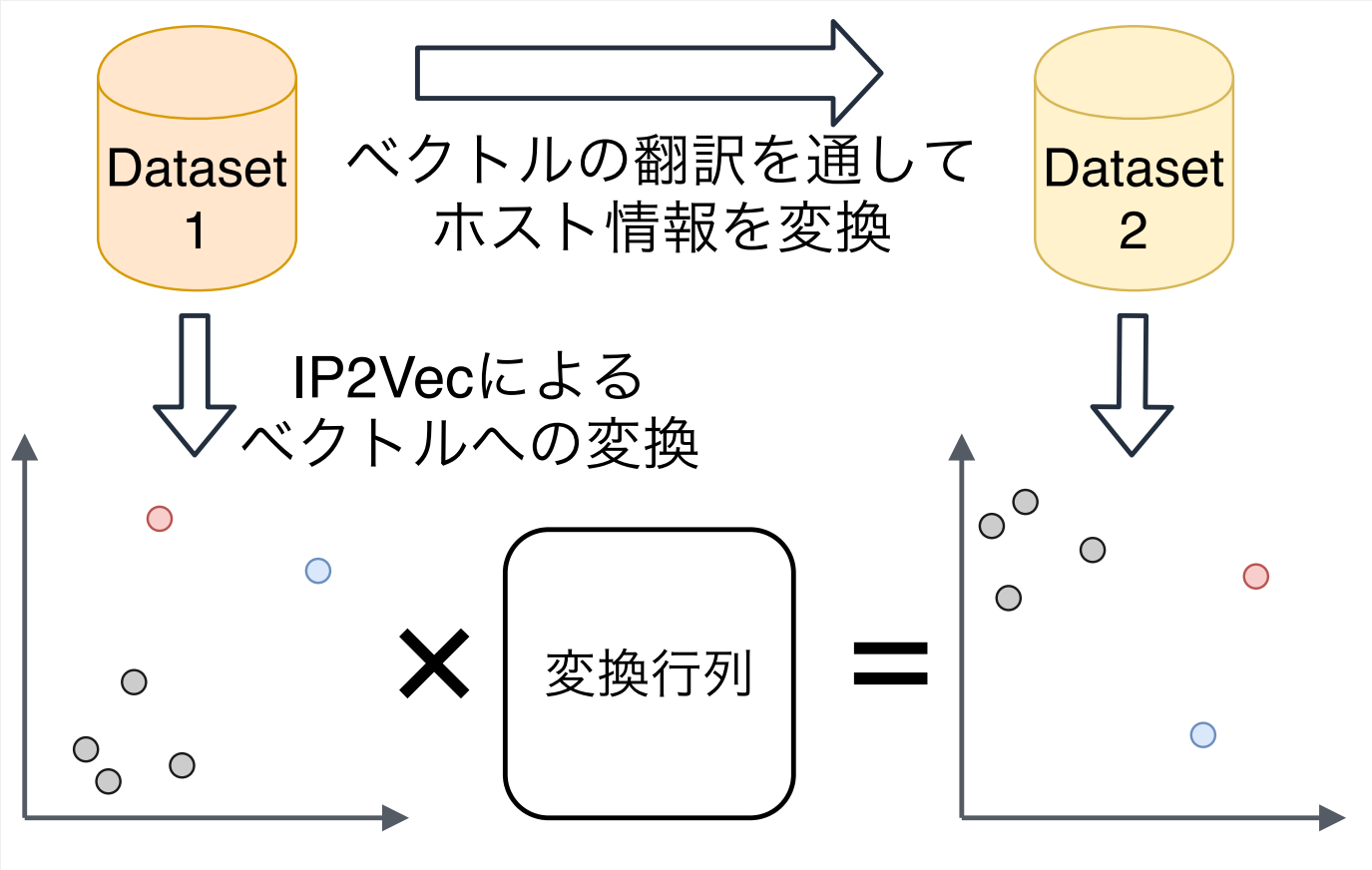
バイトレベル生成データの周波数特性

◆周波数特性を見ると比較的似た挙動が出ている.



フローレベルの生成

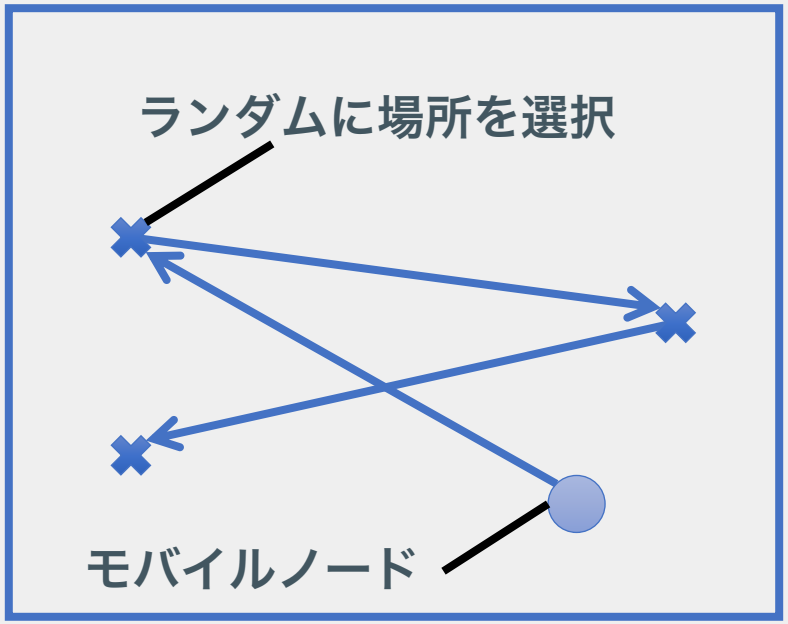
- フローレベルについては，複数のデータセット間のデータの翻訳にトライしている.
  - 言語間の翻訳に類似した手法で，フローを埋め込んだベクトルを変換.



通信デバイスの移動軌跡の生成

デバイスの移動軌跡の生成

- ◆通信デバイスの移動は，モバイル通信のシミュレーションで重要.
- ◆従来は統計的アプローチが利用されてきた.
- ◆データドリブンなシミュレーションもあるが，地理的条件などが限定されてしまうため，一般的には使えない.

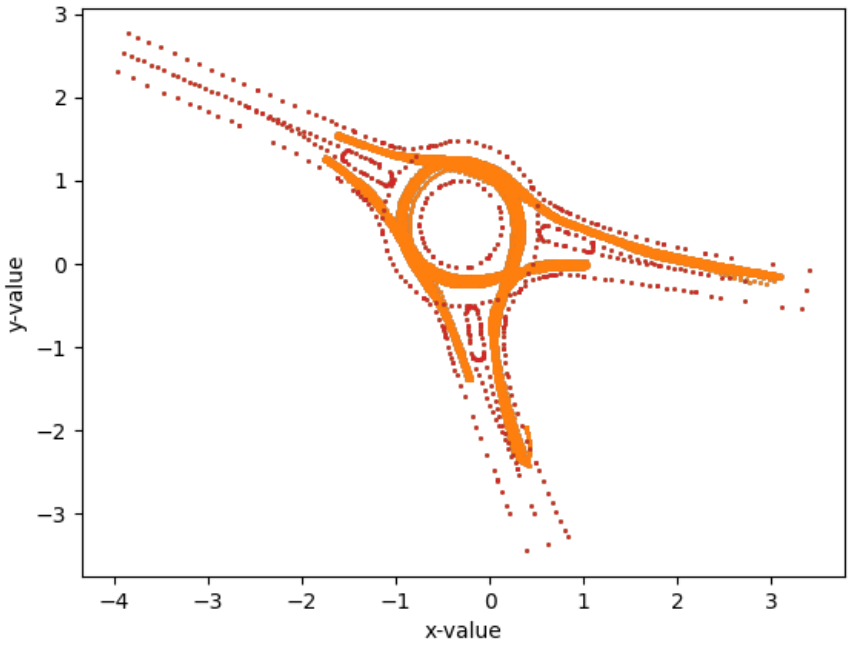


Random Waypoint mobility model

デバイスの移動軌跡データの地理条件の変換

◆実データを元に生成しつつも，地理的条件が自由に設定可能な生成モデルの構築を目指す．

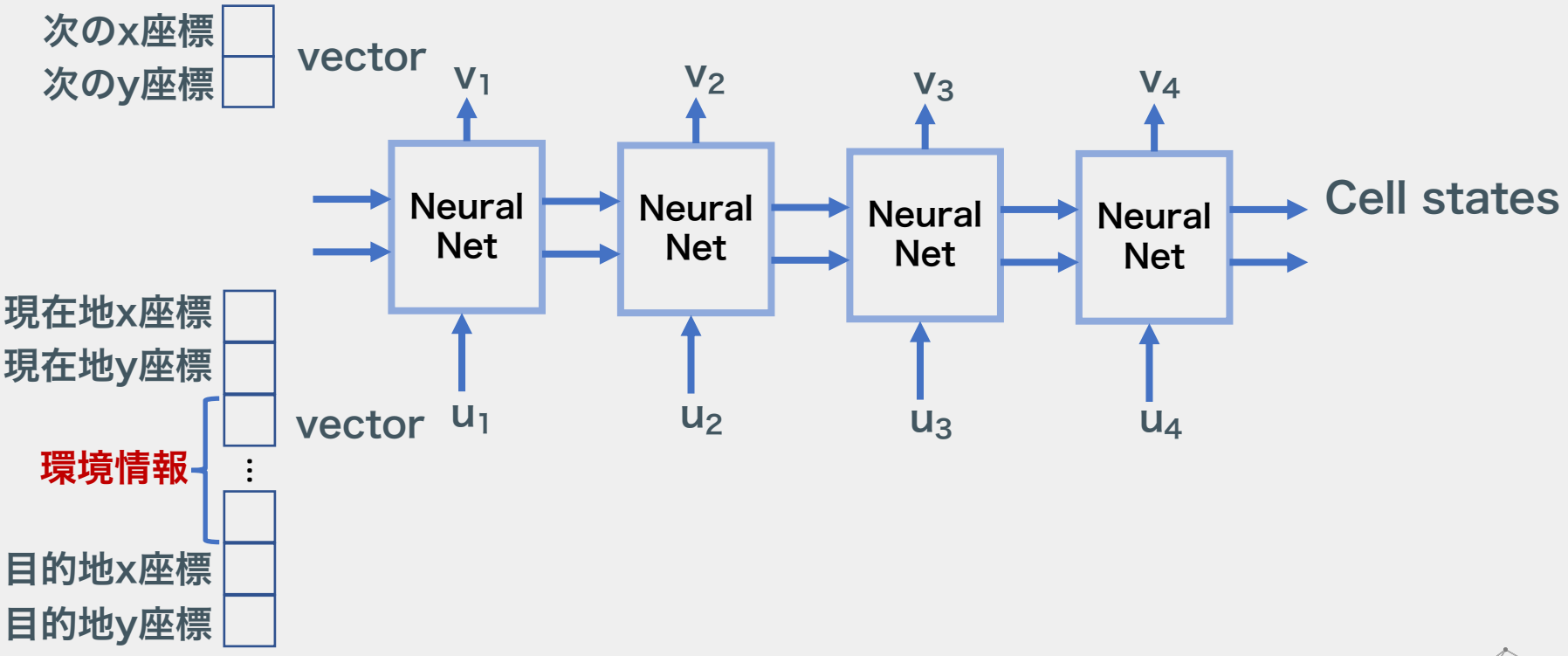
入手可能な実データ



別の道路なら
どうなるか

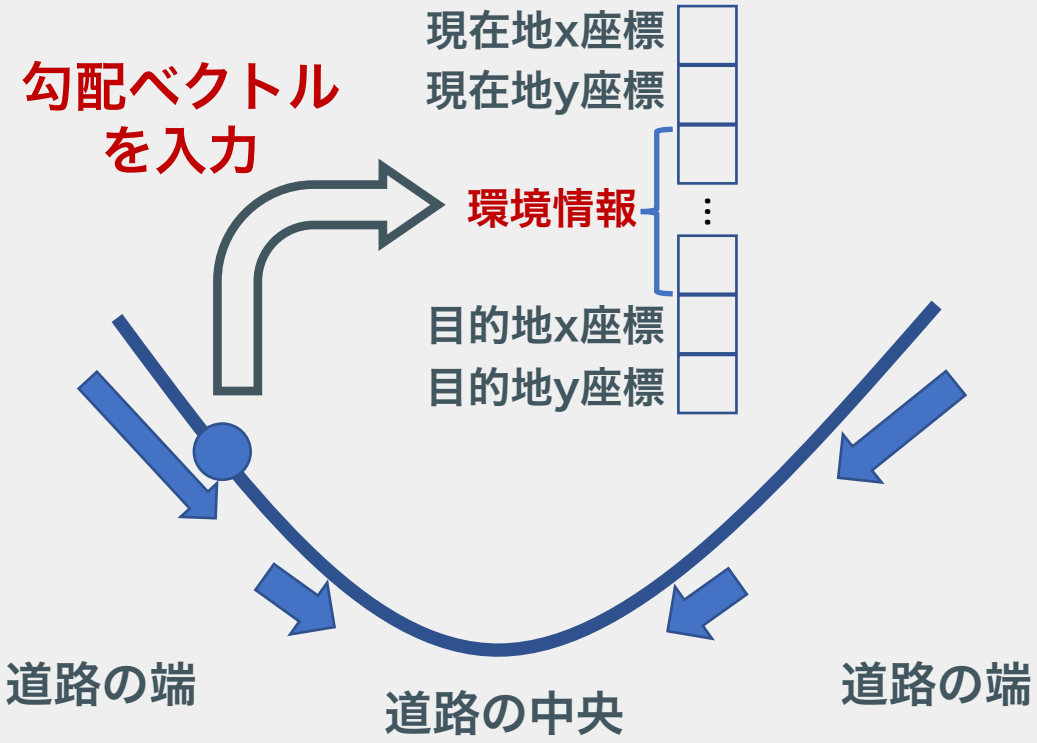
デバイスの移動軌跡をLSTMで学習

◆LSTMに環境情報(=地図情報)を入力して，再帰的に生成.

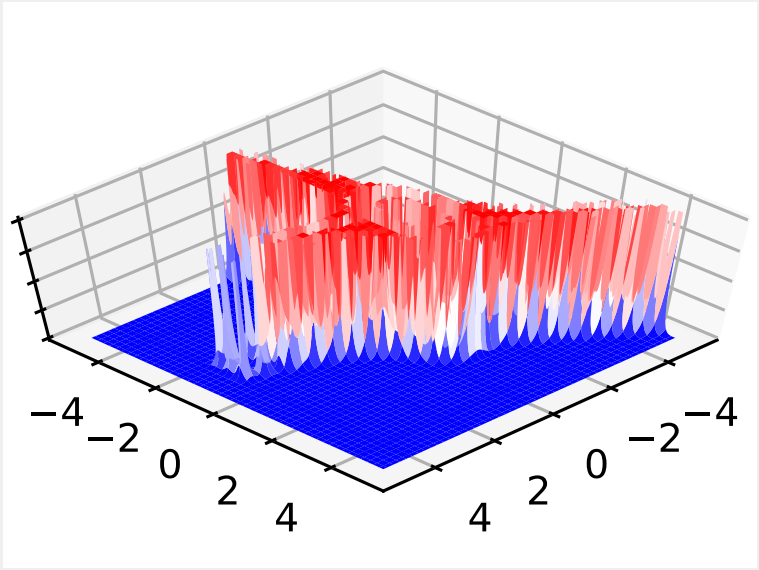


デバイス移動軌跡とポテンシャル場

- ◆道路を表すポテンシャル場を形成し，勾配を環境情報として入力.
- ◆パスプランニングの古典的な手法であるポテンシャル場を使う方法と機械学習的なアプローチを組み合わせ.



ポテンシャル場



本日のまとめと今後について

まとめと今後

- 我々の研究グループでは，通信データの生成をキーワードに，データ生成問題にチャレンジしています.
  - ネットワークトポロジ: 少しの成果
  - 通信トラヒック: 道半ば
  - デバイスの移動軌跡: 道半ば
- 他にも，様々なデータ生成に取り組む予定.
  - 通信品質，電波特性，メッセージ生成，etc.
- 機械学習と統計的モデルの融合的なアプローチの検討.

再現するデータのアイデアがある方，実際にデータをお持ちの方，機械学習に詳しい方，面白そうと思った方，
我々と一緒に共同研究しませんか？

ご連絡をお待ちしております.

ご清聴ありがとうございました.