



**POLYTECHNIQUE  
MONTREAL**

UNIVERSITÉ  
D'INGÉNIERIE



# **Open-Source Projects: Machine Learning for Transportation Data Imputation and Prediction**

Reproducible Research Workshop

TRB 103rd Annual Meeting · Washington, D.C., USA

**Xinyu Chen**

January 11, 2024

## Open-source & reproducible research:

- ① GitHub: <https://github.com/xinychen>
- ② Slides: <https://xinychen.github.io/slides/transdim.pdf>
- ③ Project website: <https://spatiotemporal-data.github.io>

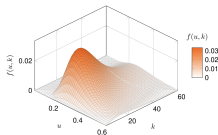
### ML algorithms



transdim

(1.1k stars)

### Visualization tools



awesome-latex-drawing

(1.2k stars)

## 1. Storytelling with Data

## 2. Spatiotemporal Traffic Data Modeling

- Reformulate traffic data imputation
- Reformulate traffic forecasting

## 3. Python Implementation

- Tools & Packages
- Traffic data processing
- Switch from CPU to GPU

## 4. “Sustainable” Research

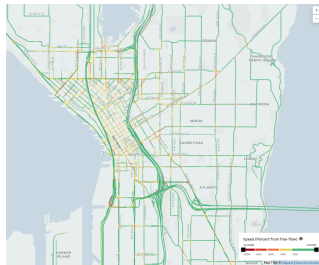
- Post something that matters

# Storytelling with Data

- Uber (hourly) movement speed data



NYC movement



Seattle movement

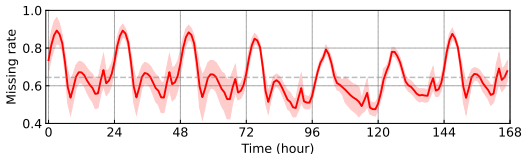
- $\{\text{road segment, time step (hour), average speed}\}$
- $\mathbf{Y} \in \mathbb{R}^{N \times T}$  with  $N$  spatial locations  $\times T$  time steps
- Computing hourly speed: Road segments have 5+ unique trips.

**Issue:** Insufficient sampling of ridesharing vehicles on the road network!

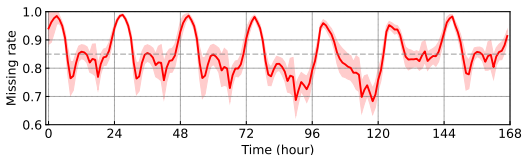
# Storytelling with Data

## High-dimensional & sparse

- **NYC** movement speed data (2019)
  - 98,210 road segments & 8,760 time steps (hours)
  - Overall missing rate: 64.43%



- **Seattle** movement speed data (2019)
  - 63,490 road segments & 8,760 time steps (hours)
  - Overall missing rate: 84.95%



# Storytelling with Data

---

- Data
- Quality
- Sparsity
- Estimation
- Imputation
- Interpolation
- Forecasting

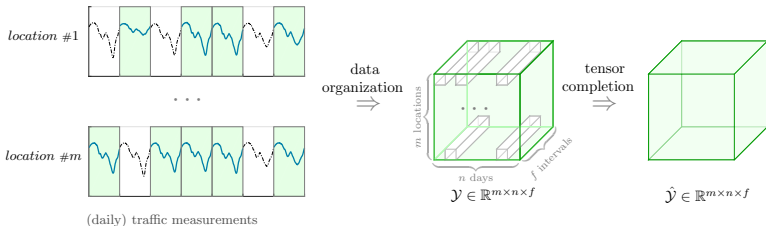
# Reformulate Traffic Data Imputation

Imputing missing traffic data

- Represent traffic data as tensors

$$\text{Tensorization: } \mathbf{Y} \in \mathbb{R}^{m \times t} \rightarrow \mathcal{Y} \in \mathbb{R}^{m \times n \times f}$$

w/  $m$  locations,  $n$  days, and  $f$  time intervals per day.



- Tensor completion (Observed index set  $\Omega$ )

$$\underbrace{\mathcal{P}_{\Omega}(\mathcal{Y})}_{\text{Partially observed}} \xrightarrow{\text{Estimate}} \underbrace{\mathcal{P}_{\Omega}^{\perp}(\mathcal{Y})}_{\text{Unobserved}}$$

# Reformulate Traffic Data Imputation

---

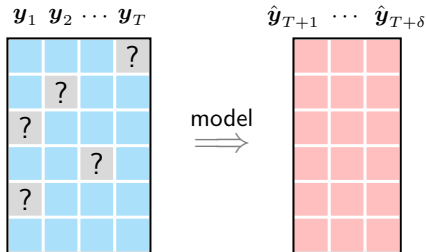


# Reformulate Traffic Forecasting

Forecasting urban traffic states with sparse data

- Problem definition ( $\delta$ -step ahead forecasting)

$$\underbrace{\{y_1, y_2, \dots, y_T\}}_{\text{Current traffic states}} \xrightarrow{\text{Estimate}} \underbrace{\{\hat{y}_{T+1}, \hat{y}_{T+2}, \dots, \hat{y}_{T+\delta}\}}_{\text{Future traffic states}}$$



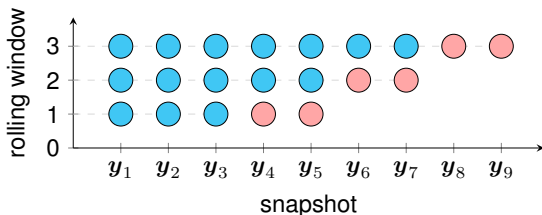
# Reformulate Traffic Forecasting

(Rolling) Forecasting urban traffic states with sparse data

1st rolling step:  $\{y_1, y_2, y_3\} \rightarrow \{y_4, y_5\}$

2nd rolling step:  $\{y_1, y_2, y_3, y_4, y_5\} \rightarrow \{y_6, y_7\}$

3rd rolling step:  $\underbrace{\{y_1, y_2, y_3, y_4, y_5, y_6, y_7\}}_{\text{Current traffic states}} \rightarrow \underbrace{\{y_8, y_9\}}_{\text{Future traffic states}}$



TMF<sup>1,2</sup>

Jupyter Notebook

---

<sup>1</sup>tracebase: <https://github.com/xinychen/tracebase>

<sup>2</sup>tpami

# Python Implementation

---

## Tools & Packages

Python



colab

CPU computing



GPU computing



CuPy

\*NumPy for GPU

## Traffic Data Processing

- Data format: `.npz` (compressed format)
- Easy to use
  - Connect with `numpy` (for CPU)
  - Connect with `cupy` (for GPU)

## NYC Uber movement dataset:

- `hourly_speed_mat_2019_1.npz` (91 MB)
  - $98210 \times 744$  matrix
  - 23,228,581 observations
- `hourly_speed_mat_2019_2.npz` (85.2 MB)
  - $98210 \times 672$  matrix
  - 21,912,460 observations
- `hourly_speed_mat_2019_3.npz` (38.1 MB)
  - $98210 \times 264$  matrix
  - 10,026,045 observations

## Switch from CPU to GPU

---



CuPy

```
import numpy as np
```

```
import cupy as np
```

Python implementation of algorithms with the `numpy` package (Using less packages can improve the reproducibility)  
Easy to convert the codes from CPU to GPU

# Post Something That Matters

Post well-documented **data processing files** (e.g., processing Chicago taxi data)

- Beginners to build coding skills
- Researchers to build research ideas

## Matching Taxi Trips with Community Areas

There are three basic steps to follow for processing taxi trip data:

- Download taxi trips in 2022 in the .csv format, e.g., `taxi_trips_2022.csv`.
- Use the `pandas` package in Python to process the raw trip data.
- Match trip pickup/dropoff locations with boundaries of the community area.

```
import pandas as pd

data = pd.read_csv('taxi_trips_2022.csv')
data.head()
```

For each taxi trip, one can select some important information:

- **Trip Start Timestamp:** When the trip started, rounded to the nearest 15 minutes.
- **Trip Seconds:** Time of the trip in seconds.
- **Trip Miles:** Distance of the trip in miles.
- **Pickup Community Area:** The Community Area where the trip began. This column will be blank for locations outside Chicago.
- **Dropoff Community Area:** The Community Area where the trip ended. This column will be blank for locations outside Chicago.

```
df = pd.DataFrame()
df['Trip Start Timestamp'] = data['Trip Start Timestamp']
df['Trip Seconds'] = data['Trip Seconds']
df['Trip Miles'] = data['Trip Miles']
df['Pickup Community Area'] = data['Pickup Community Area']
df['Dropoff Community Area'] = data['Dropoff Community Area']
data
```

Figure 2 shows taxi pickup and dropoff trips (2022) on 77 community areas in the City of Chicago. Note that the average trip duration is **1207.75 seconds** and the average trip distance is **6.16 miles**.

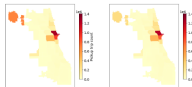


Figure 2. Taxi pickup and dropoff trips (2022) in the City of Chicago, USA. There are 4,763,061 remaining trips after the data processing.

For comparison, Figure 3 shows taxi pickup and dropoff trips (2019) on 77 community areas in the City of Chicago. Note that the average trip duration is **915.62 seconds** and the average trip distance is **3.93 miles**.

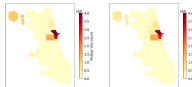


Figure 3. Taxi pickup and dropoff trips (2019) in the City of Chicago, USA. There are 12,684,572 remaining trips after the data processing. See the data processing codes.

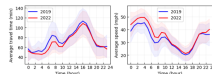


Figure 6. Average travel time and speed from area 32 (i.e., Downtown) to area 76 (i.e., Airport) in both 2019 and 2022.

```
import numpy as np
import matplotlib.pyplot as plt

fig = plt.figure(figsize = (8, 2.5))
ax = fig.add_subplot(1, 2, 1)

# Average travel time in 2019
al = df.groupby(['hour'])['Trip Seconds'].mean().values / 30
al = df.groupby(['hour'])['Trip Seconds'].std().values / 30
plt.plot(al, color = 'blue', linewidth = 1.5, label = '2019')
upper = al + al
lower = al - al
ax_bound = np.append(upper, upper, axisy=[0, 1]), np.arange(24, 24))
np.array([24 - 1, 24 - 1]), np.arange(24 - 1, -1))
y_bound = np.append(upper, upper, axisy=[0, 1]), lower()
np.arange(lower[1], upper[1]), np.arange(upper[1], lower[1]))
plt.fill_between(y_bound, color = 'blue', alpha = 0.5)

# Average travel time in 2022
al = df.groupby(['hour'])['Trip Seconds'].mean().values / 30
al = df.groupby(['hour'])['Trip Seconds'].std().values / 30
plt.plot(al, color = 'red', linewidth = 1.5, label = '2022')
upper = al + al
lower = al - al
```

Source: <https://spatiotemporal-data.github.io/Chicago-mobility/taxi-data>

# Post Something That Matters

---

Post **scientific problems** (e.g., spatiotemporal data modeling)

## Optimizing Interpretable Time-Varying Autoregression with Orthogonal Constraints

Generally speaking, any spatiotemporal data in the form of a matrix can be written as  $\mathbf{Y} \in \mathbb{R}^{N \times T}$  with  $N$  spatial areas/locations and  $T$  time steps. To discover interpretable spatial/temporal patterns, one can build a time-varying autoregression on the time snapshots  $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_T \in \mathbb{R}^N$  (Chen et al., 2023). The time-varying coefficients in the autoregression allow one to characterize the time-varying system behavior, but the challenges still remain.

To capture interpretable modes/patterns, one can use tensor factorization formulas to parameterize the coefficients and the optimization problem can be easily built. However, a great challenge would be how to make the modes "more interpretable", specifically, e.g., how to learn orthogonal modes in the modeling process. In this post, we present an optimization problem of the time-varying autoregression with orthogonal constraints as follows,

$$\begin{aligned} \min_{\mathbf{W}, \mathbf{G}, \mathbf{V}, \mathbf{X}} \quad & \frac{1}{2} \sum_{t=2}^T \|\mathbf{y}_t - \mathbf{W} \mathbf{G} (\mathbf{x}_t^\top \otimes \mathbf{V})^\top \mathbf{y}_{t-1}\|_2^2 \\ \text{s.t.} \quad & \begin{cases} \mathbf{W}^\top \mathbf{W} = \mathbf{I}_R \\ \mathbf{V}^\top \mathbf{V} = \mathbf{I}_R \\ \mathbf{X}^\top \mathbf{X} = \mathbf{I}_R \end{cases} \end{aligned}$$

where  $\mathbf{W} \in \mathbb{R}^{N \times R}$  and  $\mathbf{X} \in \mathbb{R}^{(T-1) \times R}$  refer to as the spatial modes and the temporal modes, respectively. This model can discover urban mobility transition patterns.

Source: <https://spatiotemporal-data.github.io/probs/orth-var>



# Why?

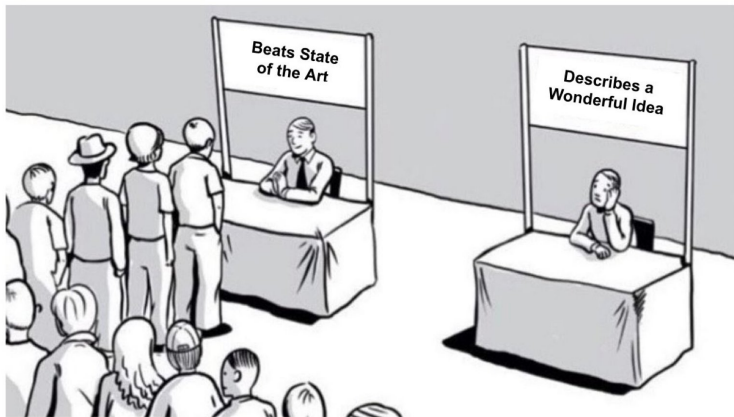
---

## Academic:

- Sustainable research environment (w.r.t. our team & followers)
- Interact with researchers from different fields
- Provide platform and benchmark for comparison
- Stimulate new algorithmic ideas

## Industry:

- Solution to ...



Source: Twitter



**POLYTECHNIQUE  
MONTREAL**

UNIVERSITÉ  
D'INGÉNIERIE



**IVADO**


# Thanks for your attention!

## Any Questions?

### About me:

 Homepage: <https://xinychen.github.io>

 GitHub: <https://github.com/xinychen>

 How to reach me: [chenxy346@gmail.com](mailto:chenxy346@gmail.com)