Tensorial Machine Learning for Discovering Patterns and Causality from Spatiotemporal Data

International Merchandise Trade & Urban Human Mobility

Xinyu Chen

M3S Project

May 20, 2024

Outline

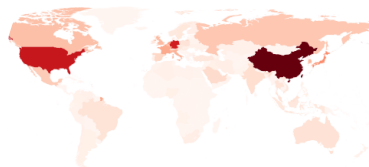
A quick look:

- Motivations
- Frameworks
  - Pattern discovery
  - Causal effect imputation
- Spatiotemporal data
  - International merchandise trade
  - Urban human mobility
- Application to M3S

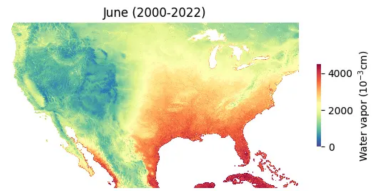
Motivation

- Spatiotemporal systems & data scenarios

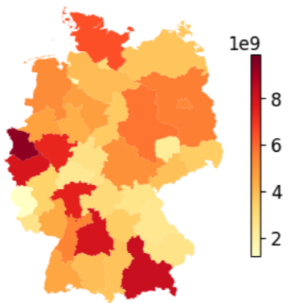
Transportation



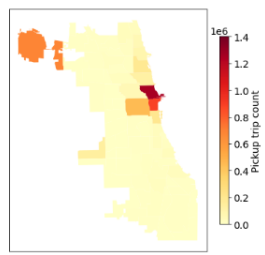
International trade



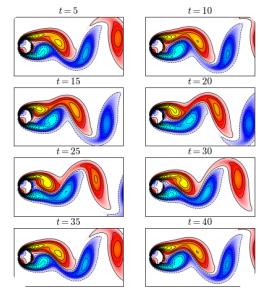
Climate



Energy



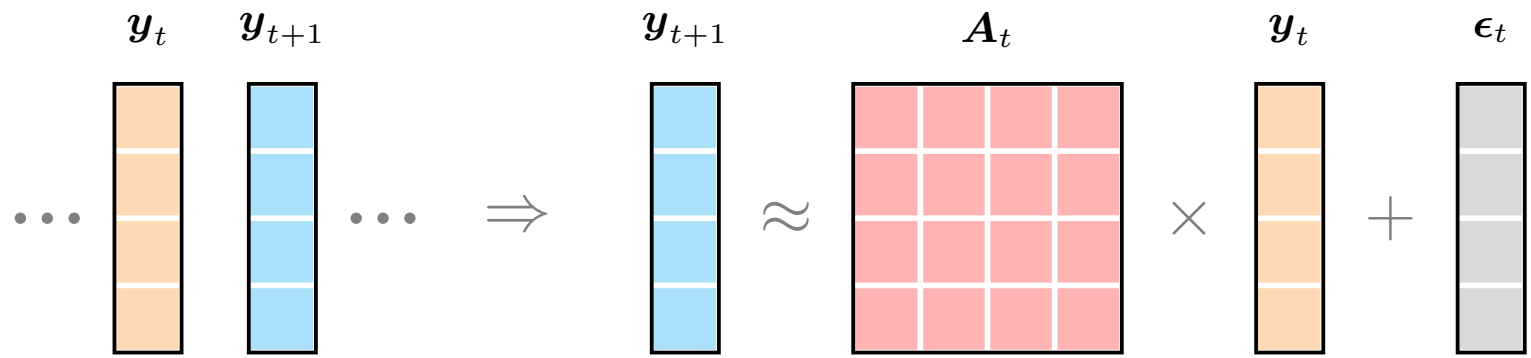
Mobility



Fluid flow

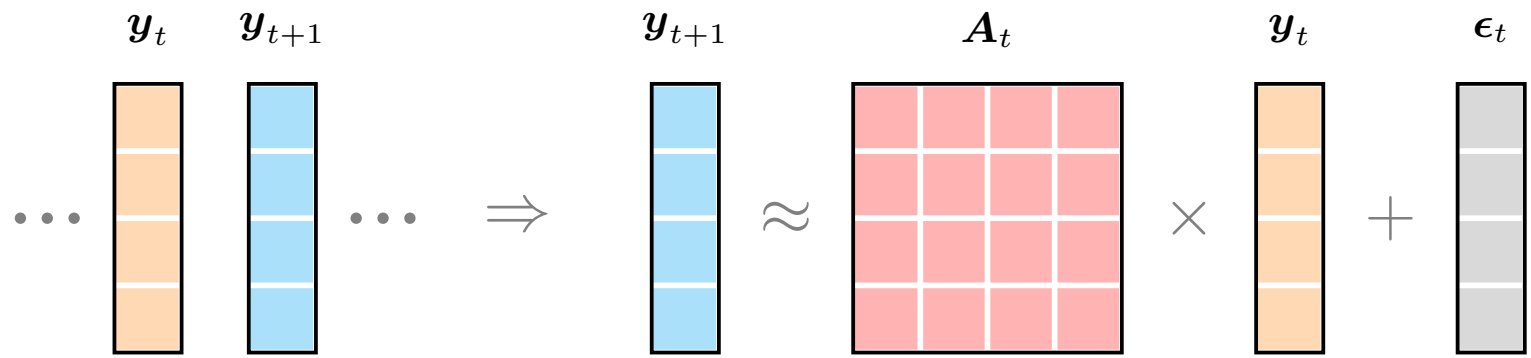
Pattern Discovery

- How to characterize dynamical systems?



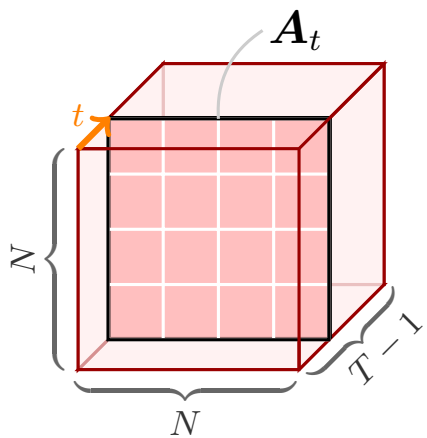
Pattern Discovery

- How to characterize dynamical systems?

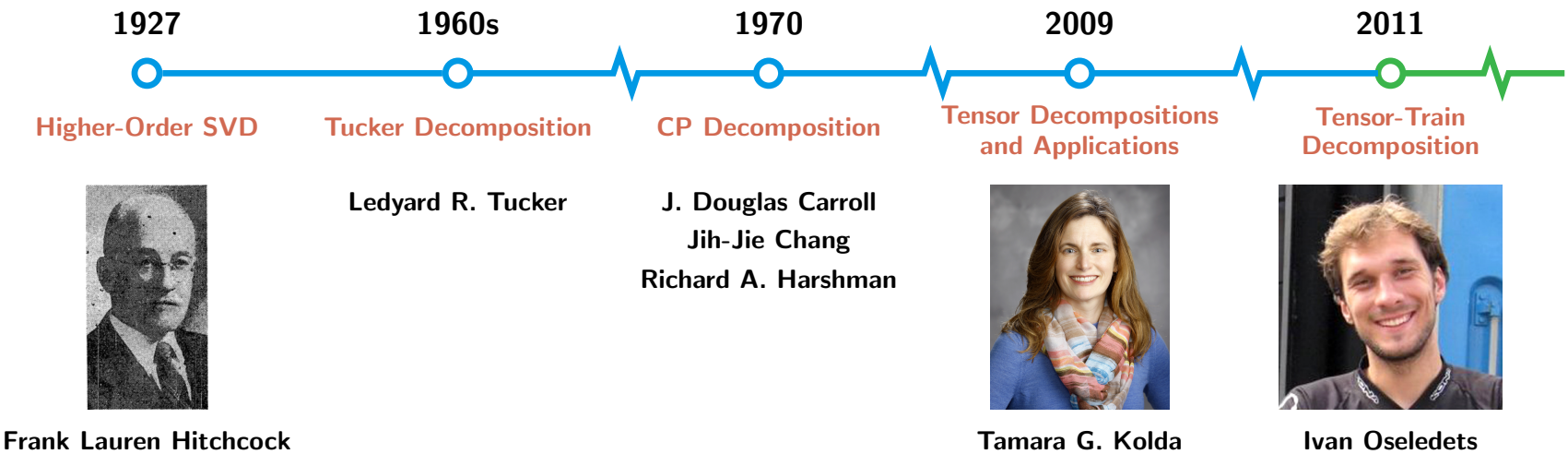


- On spatiotemporal systems $\boldsymbol{Y} \in \mathbb{R}^{N \times T}$:

$$\underbrace{\boldsymbol{y}_{t+1} = \boldsymbol{A}\boldsymbol{y}_t + \boldsymbol{\epsilon}_t}_{\text{time-invariant (e.g., DMD)}} \quad \text{v.s.} \quad \underbrace{\boldsymbol{y}_{t+1} = \boldsymbol{A}_t\boldsymbol{y}_t + \boldsymbol{\epsilon}_t}_{\text{fully time-varying (ours)}}$$



- How to discover spatial/temporal modes (patterns) from the tensor $\boldsymbol{\mathcal{A}} \triangleq \{\boldsymbol{A}_t\}_{t \in [T-1]}$?


1927
1960s
1970
2009
2011
Higher-Order SVD
Tucker Decomposition
CP Decomposition
Tensor Decompositions and Applications
Tensor-Train Decomposition
Ledyard R. Tucker
J. Douglas Carroll
Jih-Jie Chang
Richard A. Harshman
Frank Lauren Hitchcock
Tamara G. Kolda
Ivan Oseledets

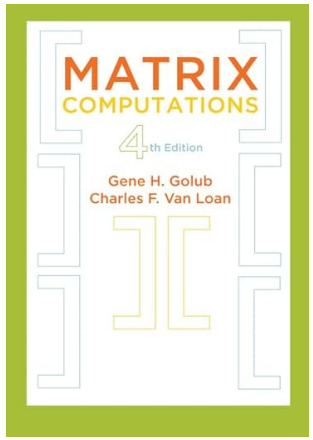
Pattern Discovery

- Tensor factorization:

$$\boldsymbol{\mathcal{A}} = \underbrace{\boldsymbol{\mathcal{G}} \times_1 \boldsymbol{W} \times_2 \boldsymbol{V} \times_3 \boldsymbol{X}}_{\text{Tucker decomposition}}$$

$$\Updownarrow$$

$$\boldsymbol{A}_t = \boldsymbol{\mathcal{G}} \times_1 \underbrace{\boldsymbol{W}}_{\text{spatial modes}} \times_2 \boldsymbol{V} \times_3 \underbrace{\boldsymbol{x}_t^\top}_{\text{temporal modes}}$$



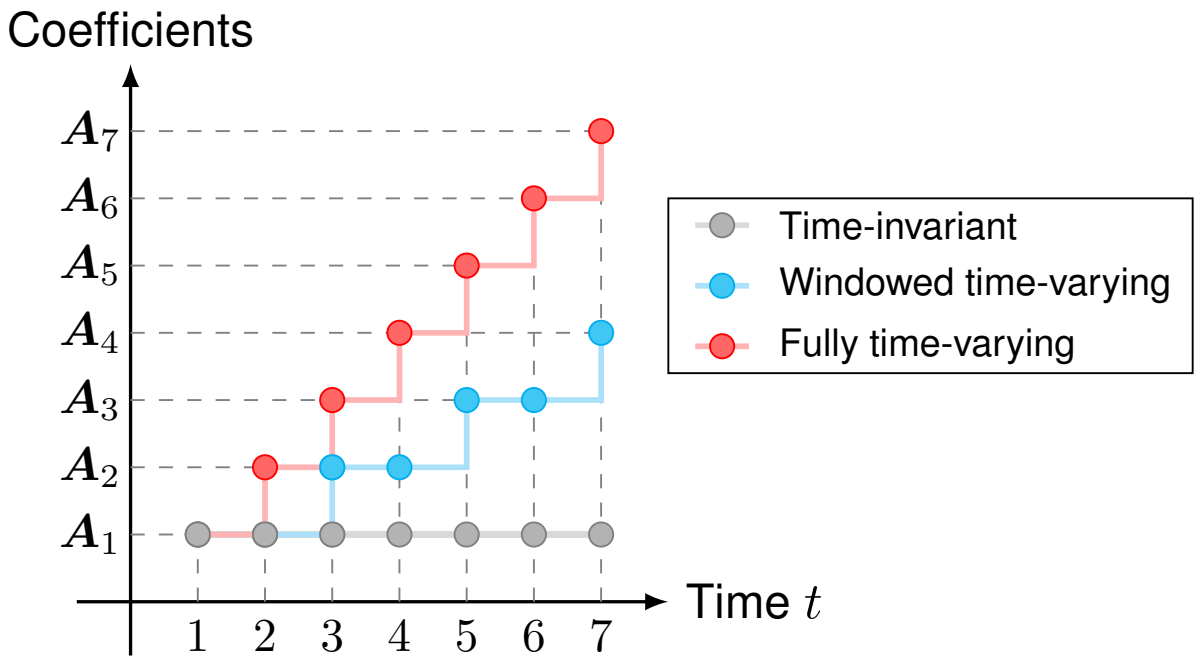
- **(Ours)** Dynamic autoregressive tensor factorization (DATF):

$$\min_{\boldsymbol{\mathcal{G}}, \boldsymbol{W}, \boldsymbol{V}, \boldsymbol{X}} \ \frac{1}{2} \sum_{t \in [T-1]} \left\| \boldsymbol{y}_{t+1} - (\boldsymbol{\mathcal{G}} \times_1 \boldsymbol{W} \times_2 \boldsymbol{V} \times_3 \boldsymbol{x}_t^\top) \boldsymbol{y}_t \right\|_2^2$$

$$\text{s.t.} \quad \underbrace{\boldsymbol{W}^\top \boldsymbol{W} = \boldsymbol{I}_R}_{\text{orthogonal spatial modes}}$$

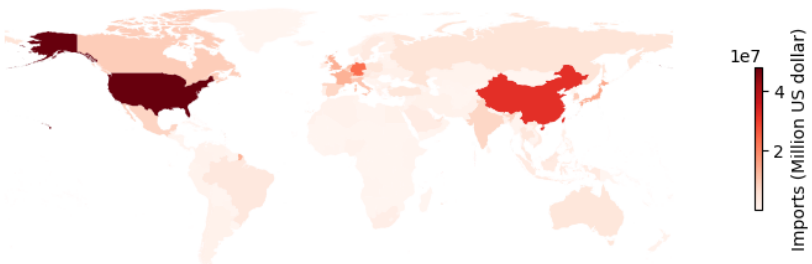
- On the data $\boldsymbol{Y} \in \mathbb{R}^{N \times T}$:

$$\underbrace{\boldsymbol{y}_{t+1} = \boldsymbol{A}\boldsymbol{y}_t + \boldsymbol{\epsilon}_t}_{\text{time-invariant (e.g., DMD)}} \quad \text{v.s.} \quad \underbrace{\boldsymbol{y}_{t+1} = \boldsymbol{A}_t\boldsymbol{y}_t + \boldsymbol{\epsilon}_t}_{\text{fully time-varying (ours)}}$$

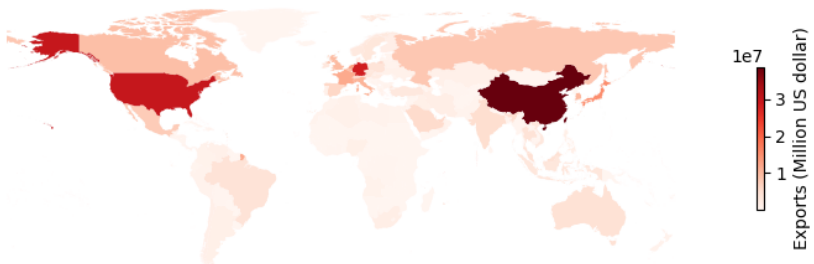


Pattern Discovery

- **Import/Export merchandise trade values (annual)**[1] (215 countries/regions & period of 2000–2022)
  - Total merchandise trade values
  - Represent import/export trade data as a 215-by-23 matrix

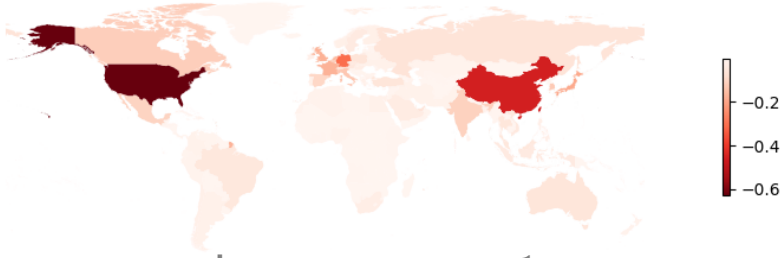


Imports from 2000 to 2022

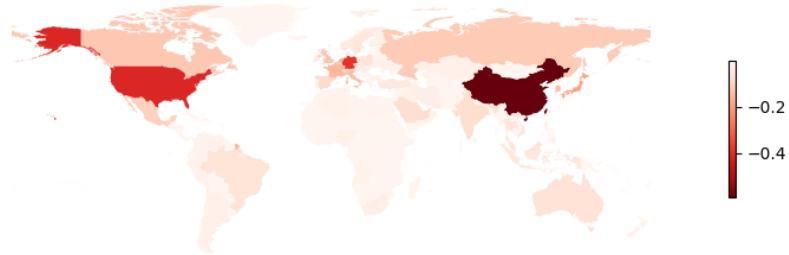


Exports from 2000 to 2022

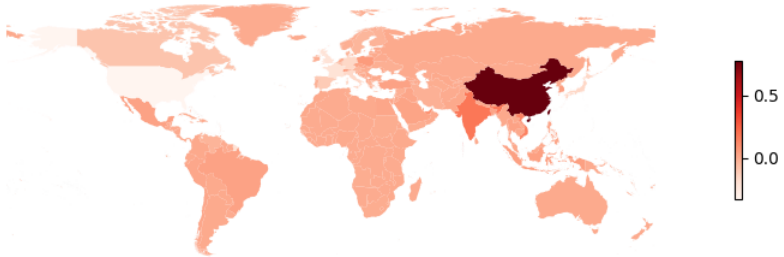
[1] The dataset is available at https://stats.wto.org.



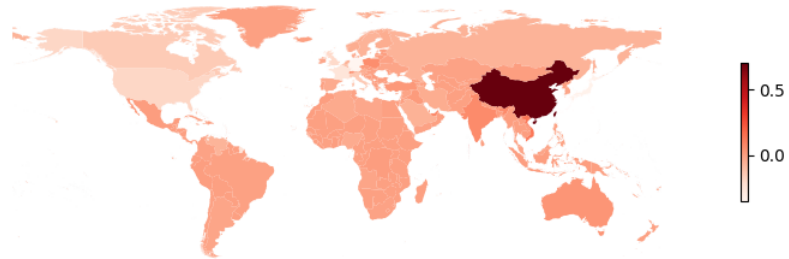
Import pattern 1



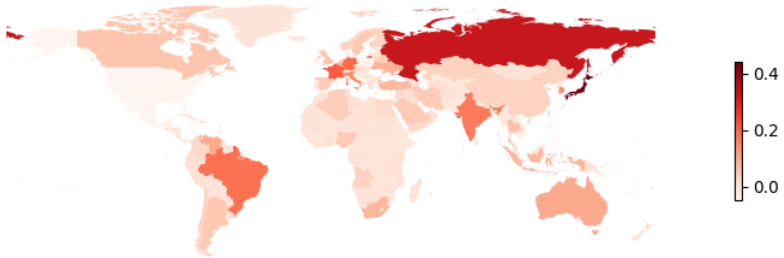
Export pattern 1



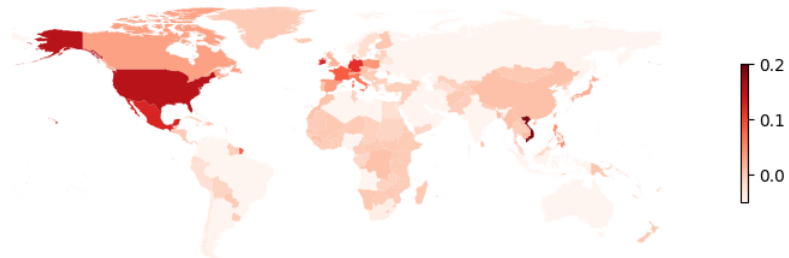
Import pattern 2



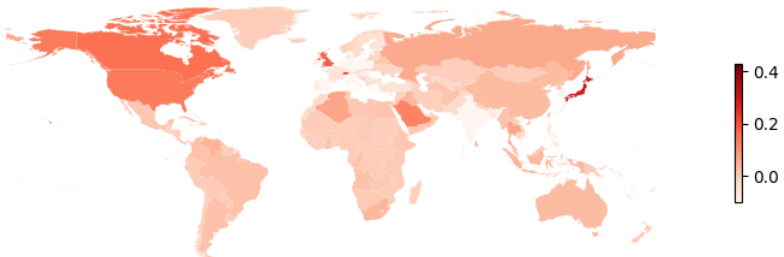
Export pattern 2



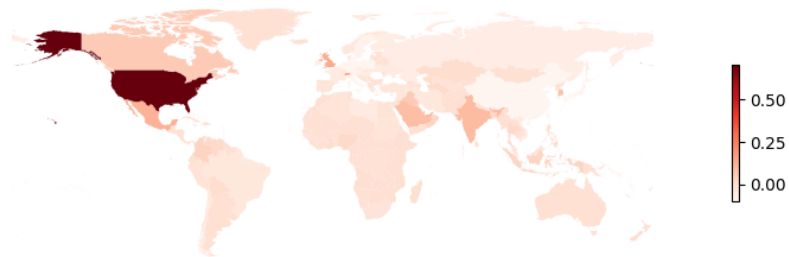
Import pattern 3



Export pattern 3



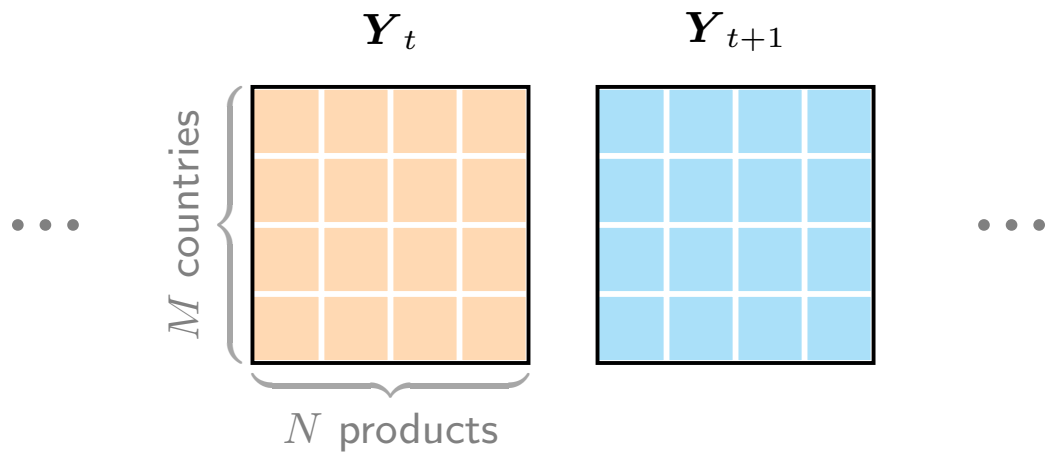
Import pattern 4



Export pattern 4

Pattern Discovery

- Three-dimensional trade (Country, Product, Year)

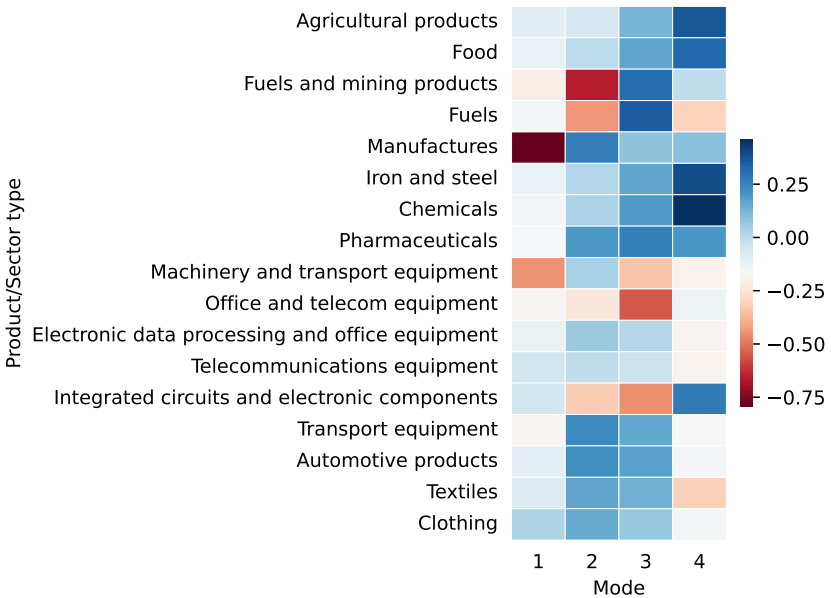


- On spatiotemporal systems $\boldsymbol{\mathcal{Y}} \in \mathbb{R}^{M \times N \times T}$:

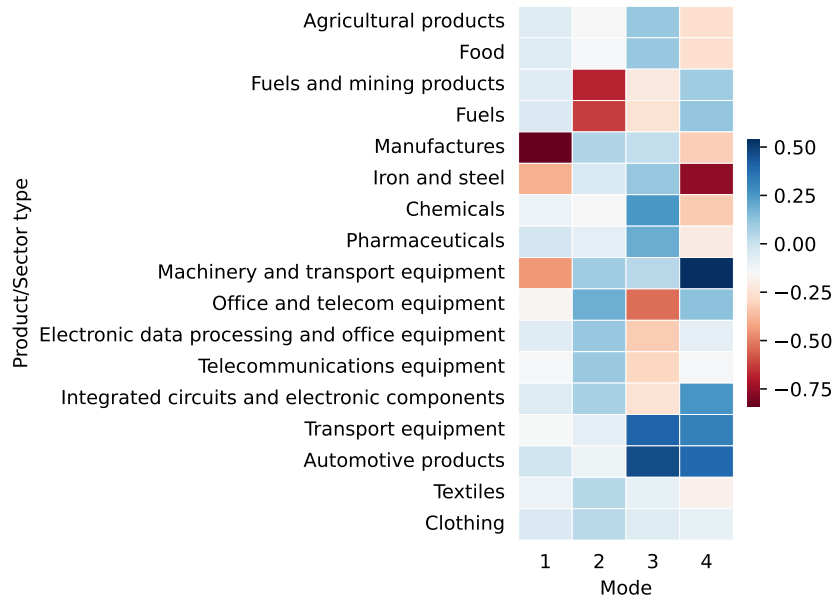
$$\underbrace{\boldsymbol{y}_{n,t+1} = \boldsymbol{A}_{n,t}\boldsymbol{y}_{n,t} + \boldsymbol{\epsilon}_{n,t}}_{\text{time-varying \& product-varying}}$$

Pattern Discovery

- **Product patterns** on 17 merchandise types



Imports



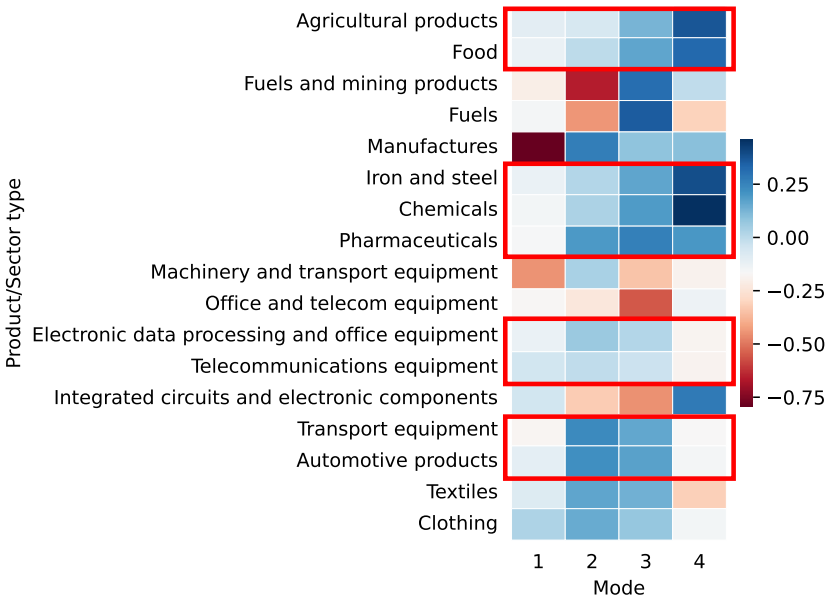
Exports

- Classify import/export merchandise according to product patterns
- Basic principle:

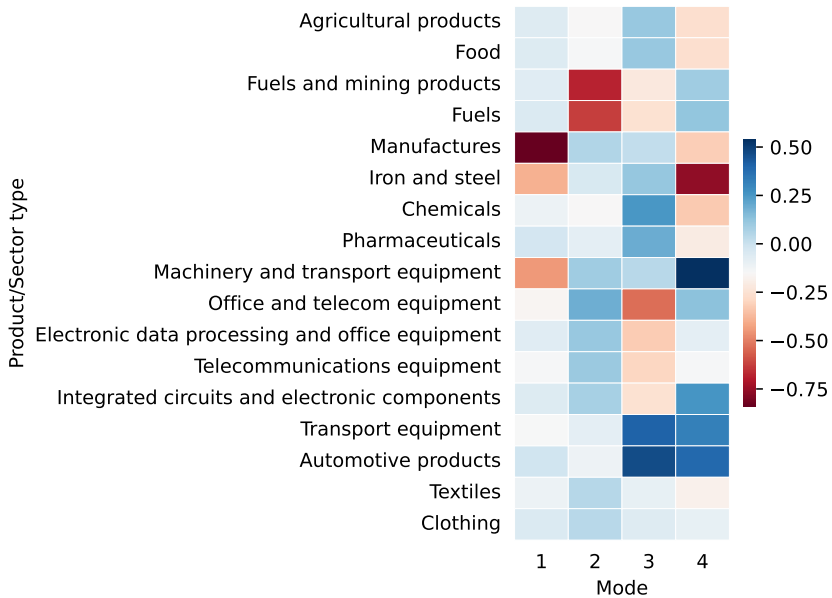
Import: What we buy? (demand) vs. Export: What we sell? (supply)

Pattern Discovery

- **Product patterns** on 17 merchandise types



Imports



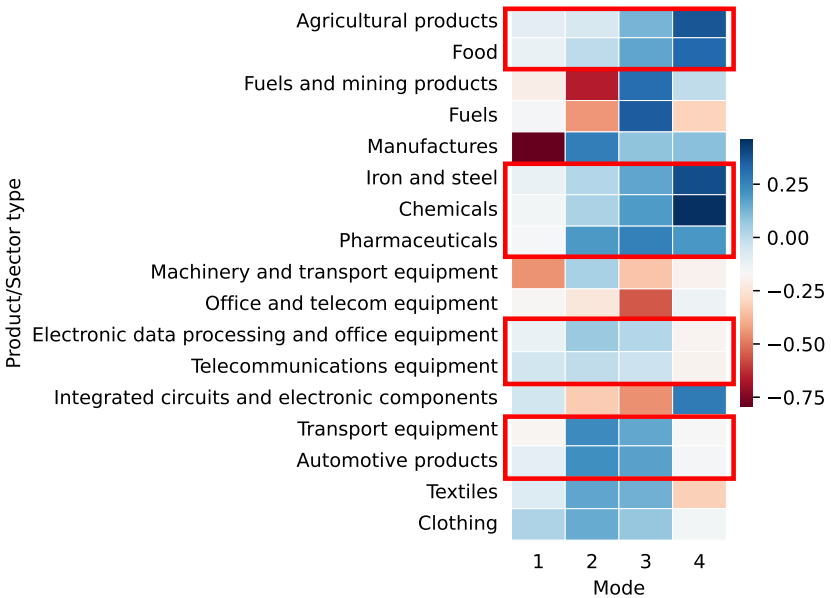
Exports

- Classify import/export merchandise according to product patterns
- Basic principle:

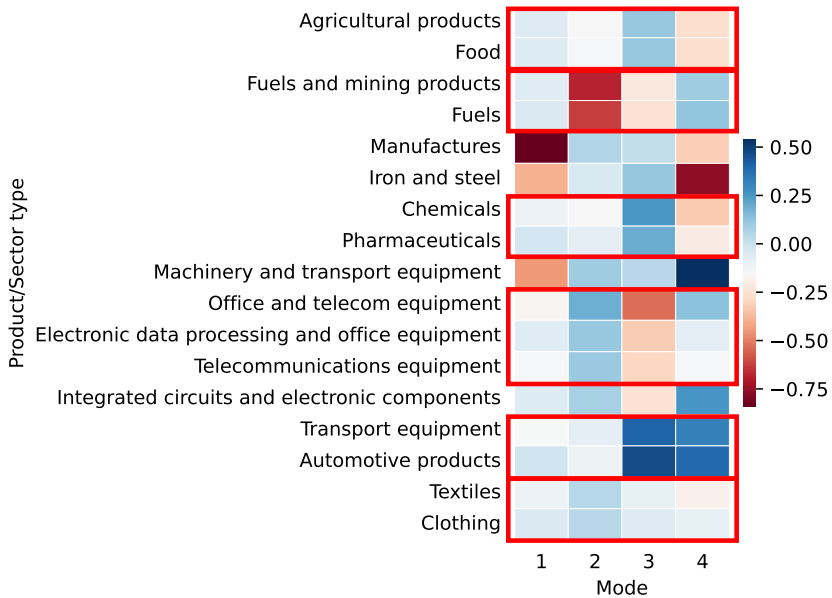
Import: What we buy? (demand) vs. Export: What we sell? (supply)

Pattern Discovery

- **Product patterns** on 17 merchandise types



Imports



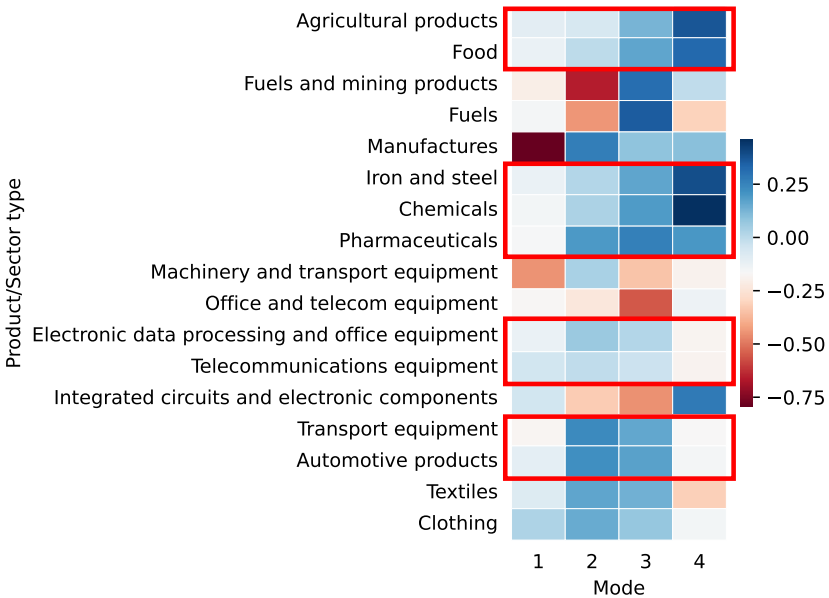
Exports

- Classify import/export merchandise according to product patterns
- Basic principle:

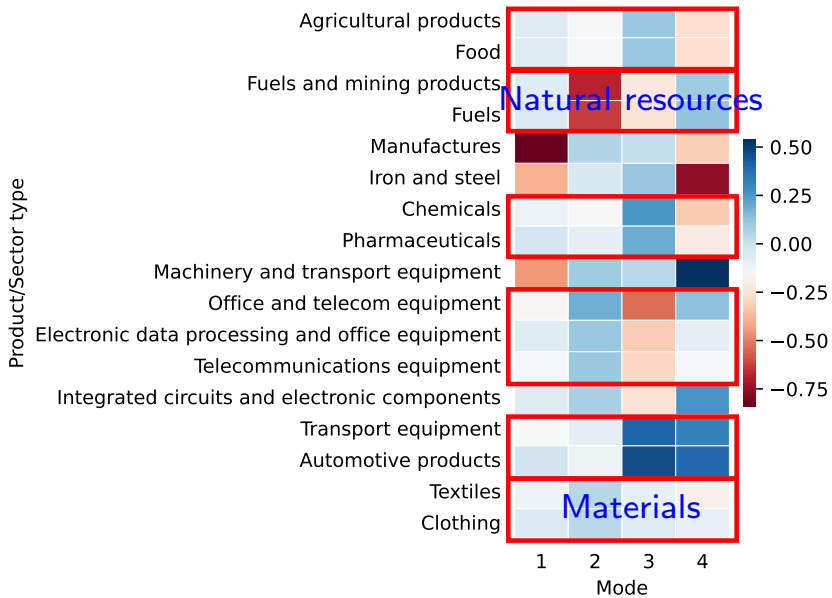
Import: What we buy? (demand) vs. Export: What we sell? (supply)

Pattern Discovery

- **Product patterns** on 17 merchandise types



Imports



Exports

- Classify import/export merchandise according to product patterns
- Basic principle:

Import: What we buy? (demand) vs. Export: What we sell? (supply)

Big International Trade Dataset

- Publicly available on Dropbox[2]
- Complicated dimensions[3]

Dimensions	Attributes
Exporter ID	231
Importer ID	231
Harmonized System (HS) code	5,018
Year	1995–2022 (28 years)
Value	Dollar ($)
Quantity	Metric Tons (MT)

Note: 247,647,741 rows in total.

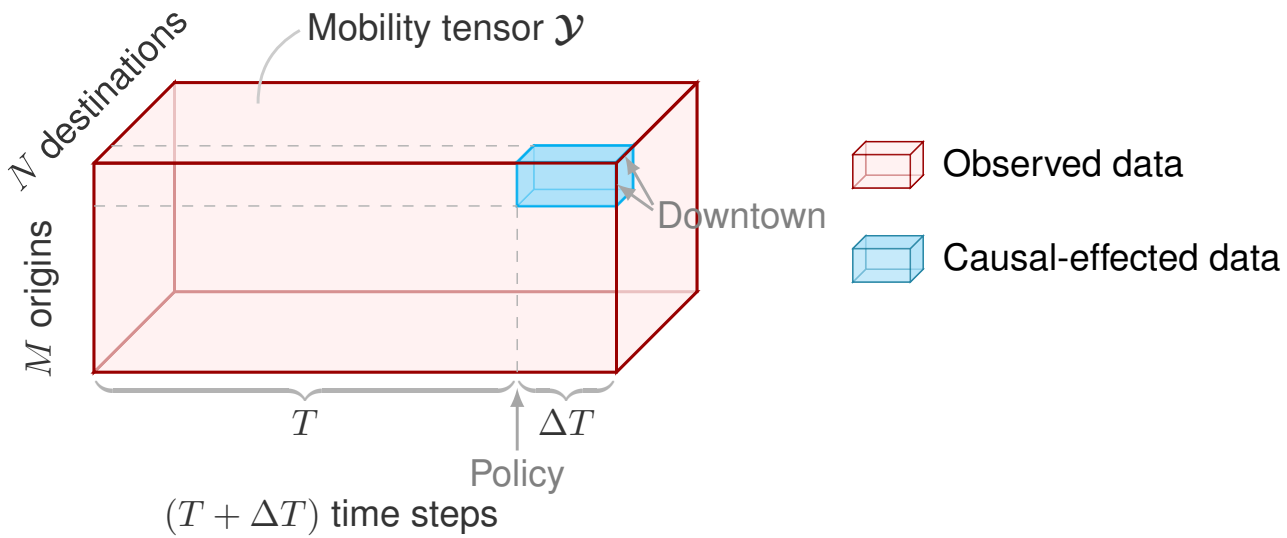
- Large data size: **3.5GB** zipped vs. **28GB** unzipped
- A lot of challenges unsolved:
  - Complicated dimensions, system evolution, non-negativity, long tails, high-order graphs/networks, causal effects, etc.

[2] Download the dataset at
https://www.dropbox.com/s/eodmonwvvo25jkm/trade_i_baci_a_92.tsv.bz2?dl=0

[3] The HS is a standardized numerical method of classifying traded products.

Causal Effect Imputation

- Urban human mobility w/ ODT
- Causal effect imputation framework

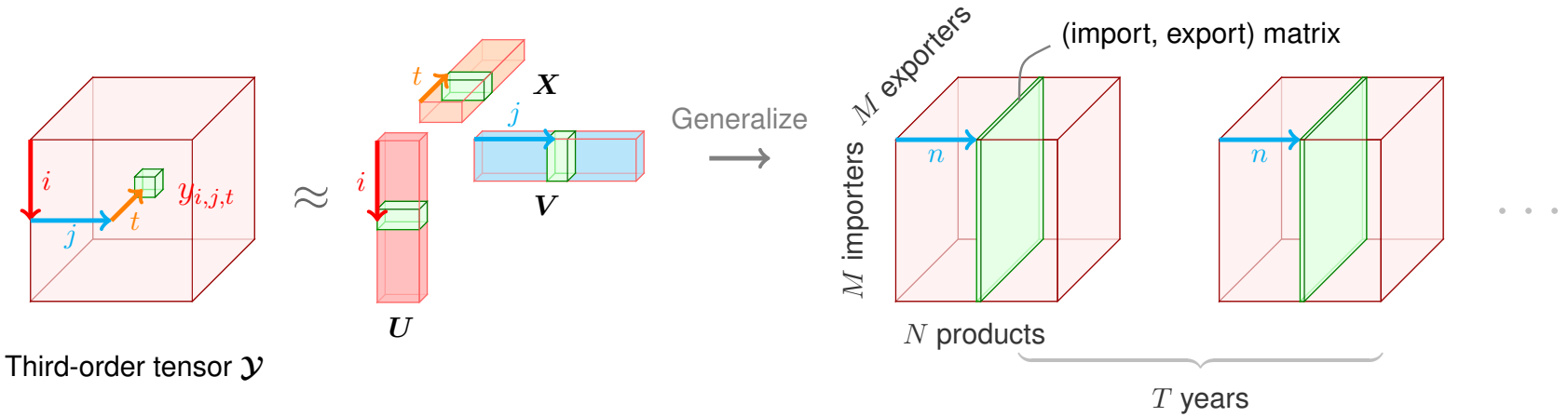


(Figure) Policy intervention (e.g., congestion pricing) in Downtown areas.

- A lot of challenges unsolved:
  - Structural missing patterns & cross validation
  - How to reduce data biases (e.g., underlying counterfactual factors) and model biases (e.g., utilizing spatial correlations & temporal dynamics)?
  - How to measure the difference between imputation results and realistic data?

Causal Effect Imputation

- Not only for **mobility**, but also for **international trade**



- Technical challenges:
  - Nonconvex optimization
  - How to integrate spatiotemporal context (e.g., graph learning)?
  - Non-negativity ...

Human Mobility Dataset

- Chicago taxi/ridesharing data

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 ourside 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']
del data
df
```

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**.

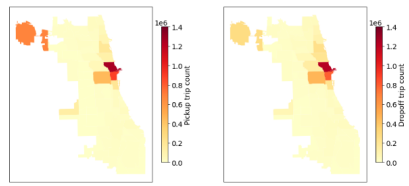


Figure 2. Taxi pickup and dropoff trips (2022) in the City of Chicago, USA. There are 4,763,961 remaining trips after the data processing.

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**.

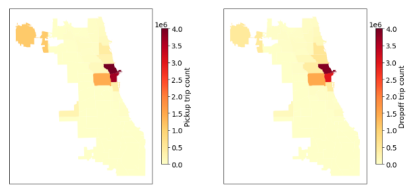


Figure 3. Taxi pickup and dropoff trips (2019) in the City of Chicago, USA. There are 12,484,572 remaining trips after the data processing. See the data processing codes.

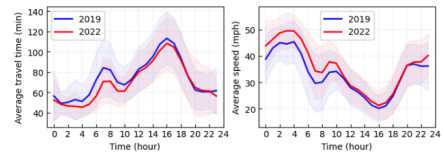


Figure 6. Average travel time and speed from area 32 (i.e., Downtown) to area 76 (i.e., Airport) in both 2019 and 2022.

```
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
m1 = df1.groupby(['hour'])['Trip Seconds'].mean().values / 30
s1 = df1.groupby(['hour'])['Trip Seconds'].std().values / 30
plt.plot(m1, color = 'blue', linewidth = 1.8, label = '2019')
upper = m1 + s1
lower = m1 - s1
x_bound = np.append(np.append(np.append(np.array([0, 0]), np.arange(0, 24)),
                              np.array([24 - 1, 24 - 1])), np.arange(24 - 1, -1,
y_bound = np.append(np.append(np.append(np.array([upper[0], lower[0]]), lower),
                              np.array([lower[-1], upper[-1]])), np.flip(upper))
plt.fill(x_bound, y_bound, color = 'blue', alpha = 0.05)

# Average travel time in 2022
m1 = df2.groupby(['hour'])['Trip Seconds'].mean().values / 30
s1 = df2.groupby(['hour'])['Trip Seconds'].std().values / 30
plt.plot(m1, color = 'red', linewidth = 1.8, label = '2022')
upper = m1 + s1
lower = m1 - s1
```

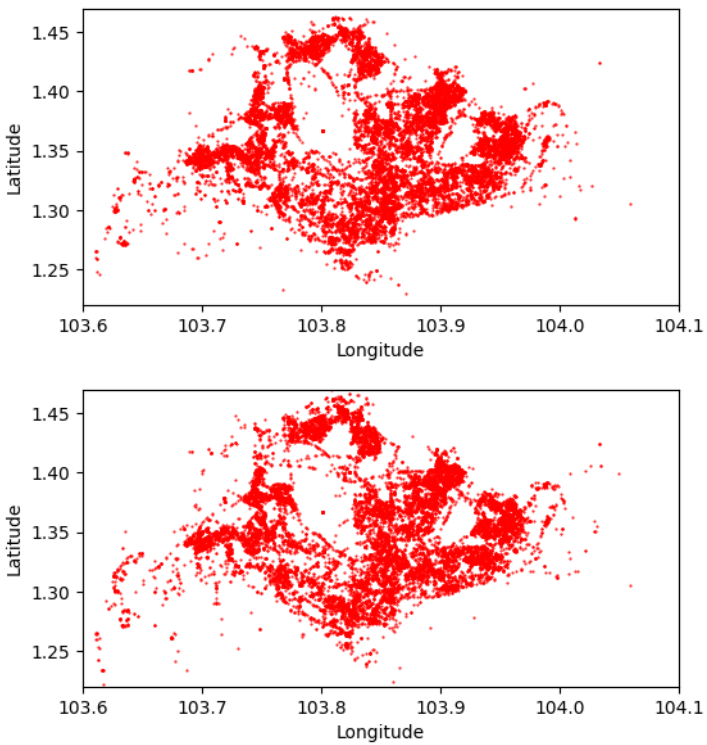
Source: https://spatiotemporal-data.github.io/Chicago-mobility/taxi-data

- Causal effect estimation of congestion pricing in Chicago

Application to M3S

Veraset dataset in Singapore[4]

- Urban trajectory (samples)



- Trajectory inference (e.g., trip purpose identification)
- Pattern discovery
  - (On 2D activity data) Uncover spatial patterns (e.g., POI patterns)
  - (On 3D mobility data) Uncover temporal patterns (e.g., long-term pattern transition impacted by special events/policies)
- Causal effect imputation
  - Infer the causal effects of policy intervention

[4] https://spatiotemporal-data.github.io/trajectory/veraset/

Thanks for your attention!

Any Questions?

About me:

- Homepage: https://xinychen.github.io
- How to reach me: chenxy346@gmail.com
- Or send to: xinychen@mit.edu