Digital Twin Travellers

Disaggregated travel demand from aggregated mobile phone data – a Privacy by Design approach

PhD Candidate

Cuauhtemoc Anda Castro

Chair

Prof. Bryan T. Adey

Committee

Prof. Kay W. Axhausen

Prof. Joachim M. Buhmann

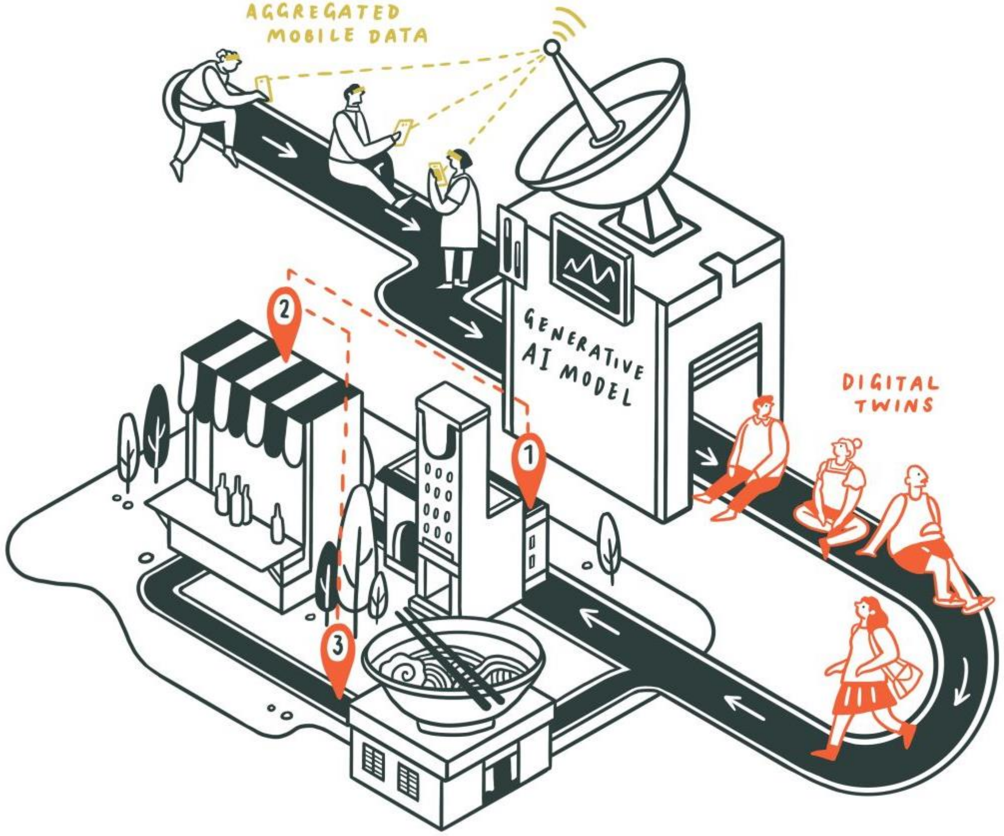
Prof. Marta C. Gonzalez

Prof. Peter Vortisch

Dr. Pieter J. Fourie

Date

July 7, 2022



IVT Institut für Verkehrsplanung und Transportsysteme
Institute for Transport Planning and Systems

(FCL) FUTURE CITIES LABORATORY

(SEC) SINGAPORE-ETH CENTRE

DataSpark

ETH zürich

Agenda

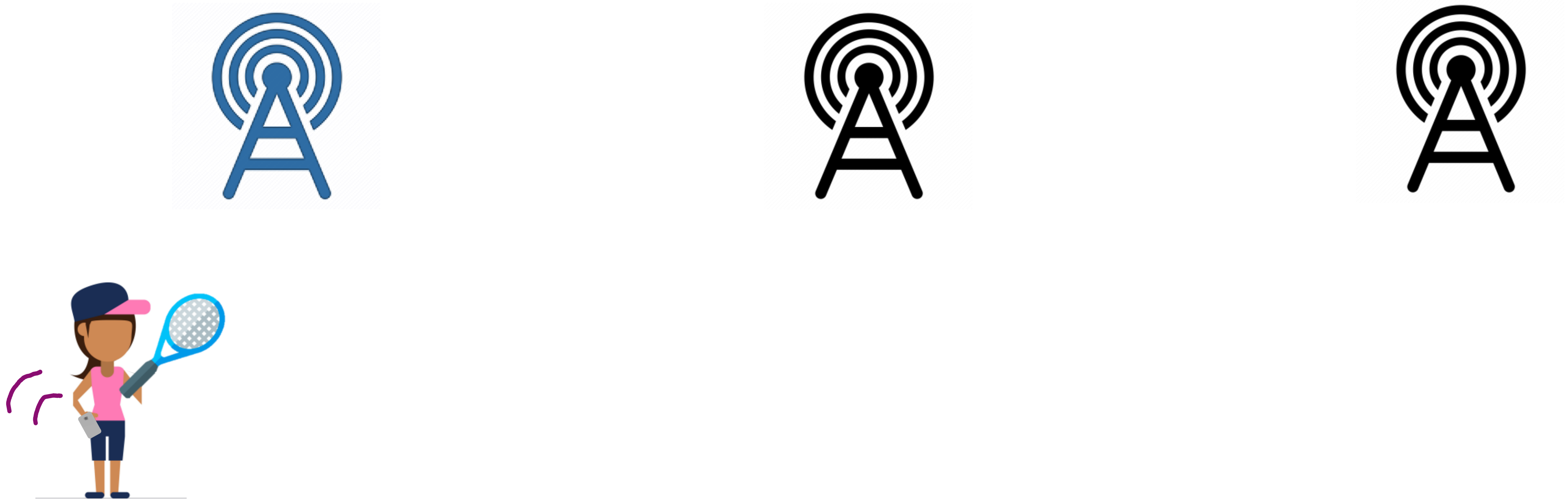
1. Mobile network signalling data
2. Motivation
3. Intuitive example
4. Digital Twin Travellers
5. Experiment and results
6. Application
7. Conclusion

Agenda

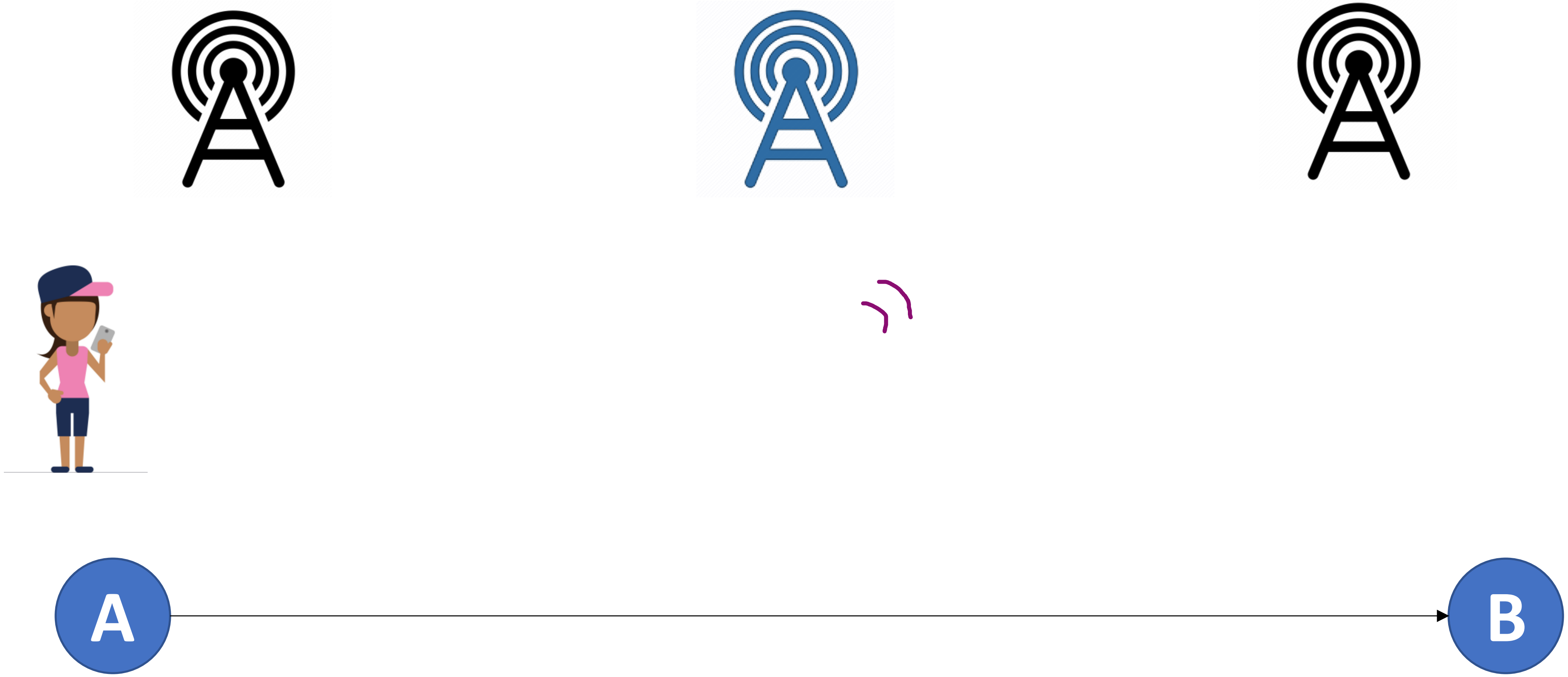
1. Mobile network signalling data
2. Motivation
3. Intuitive example
4. Digital Twin Travellers
5. Experiment and results
6. Application
7. Conclusion

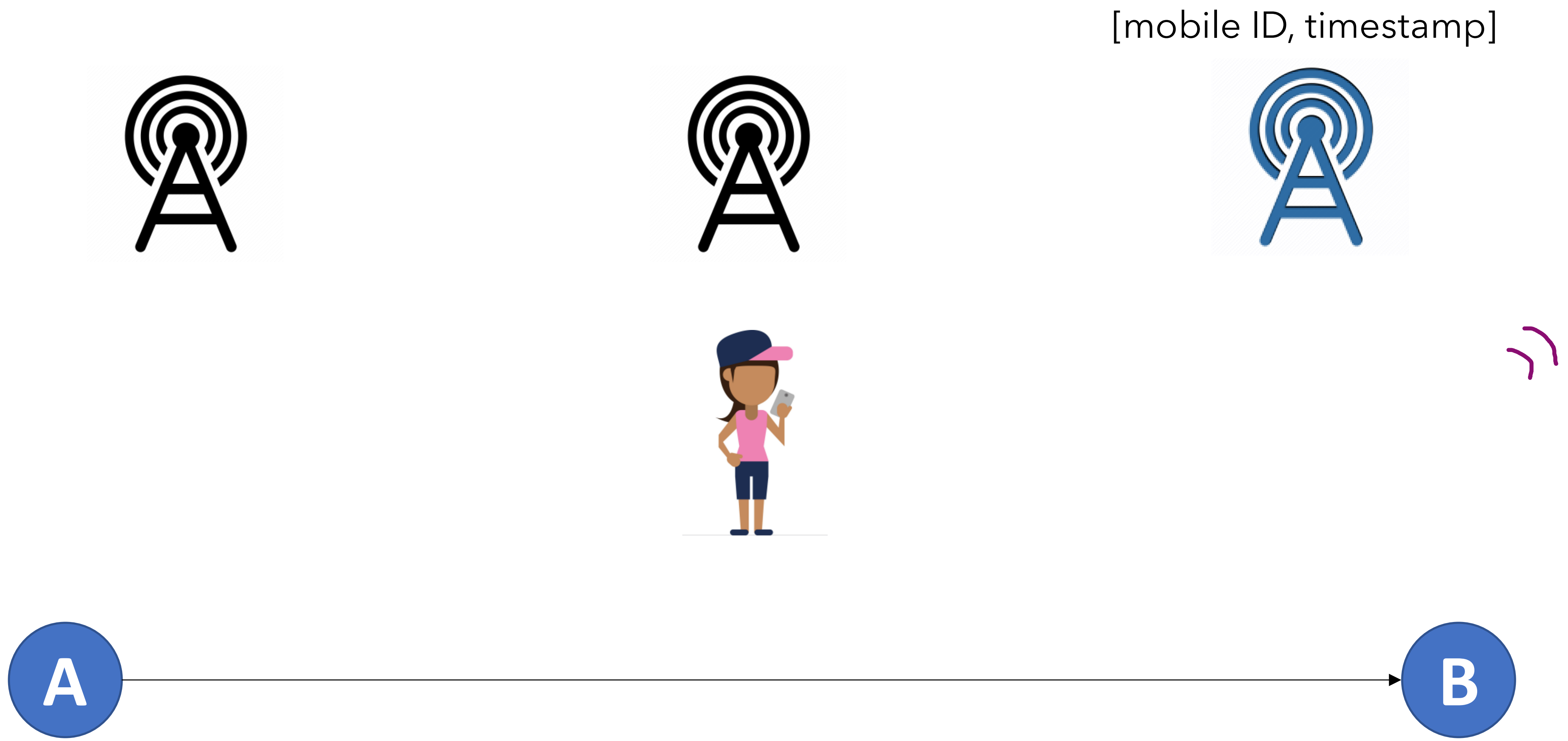
[mobile ID, timestamp]

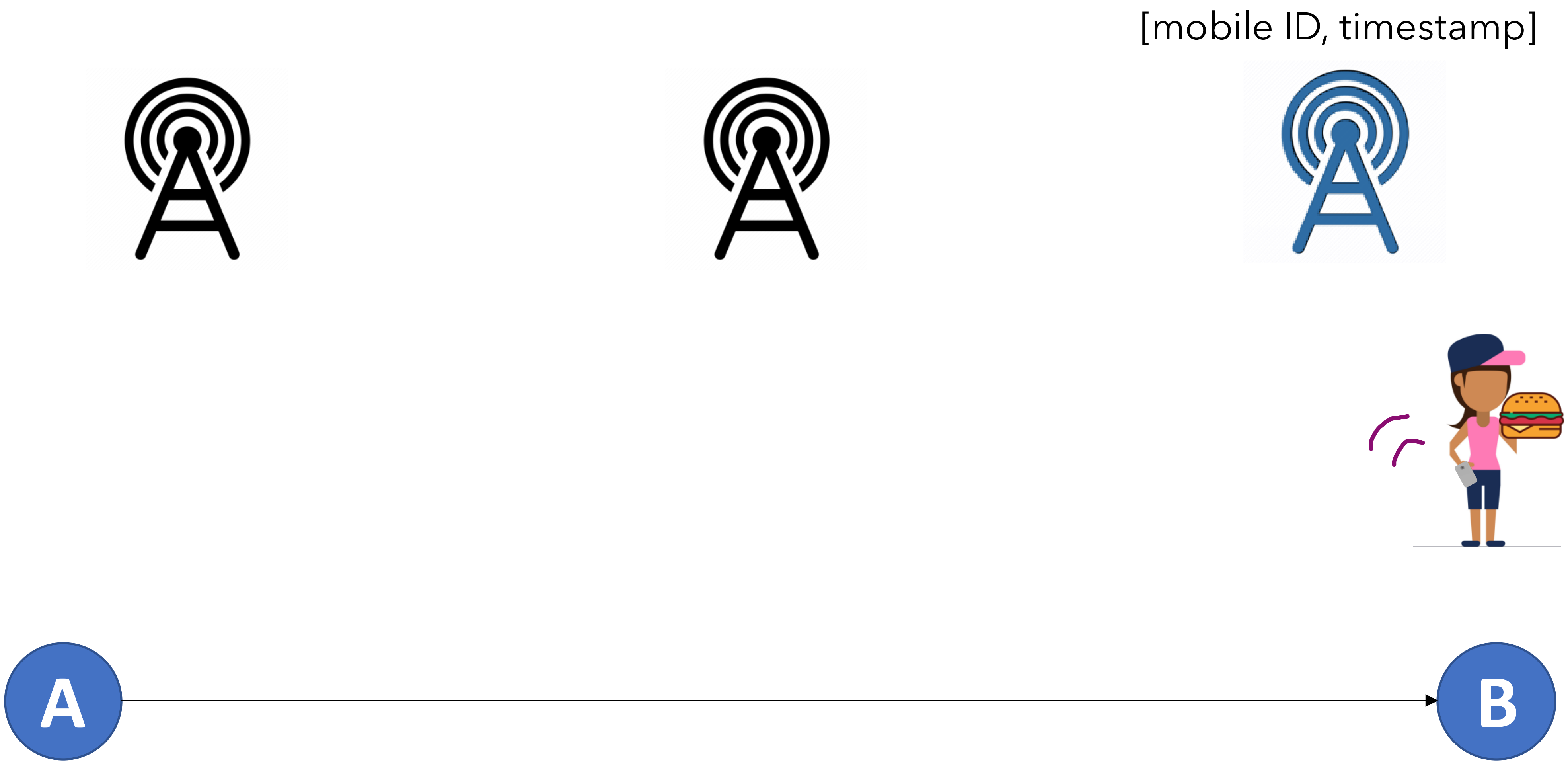
[mobile ID, timestamp]

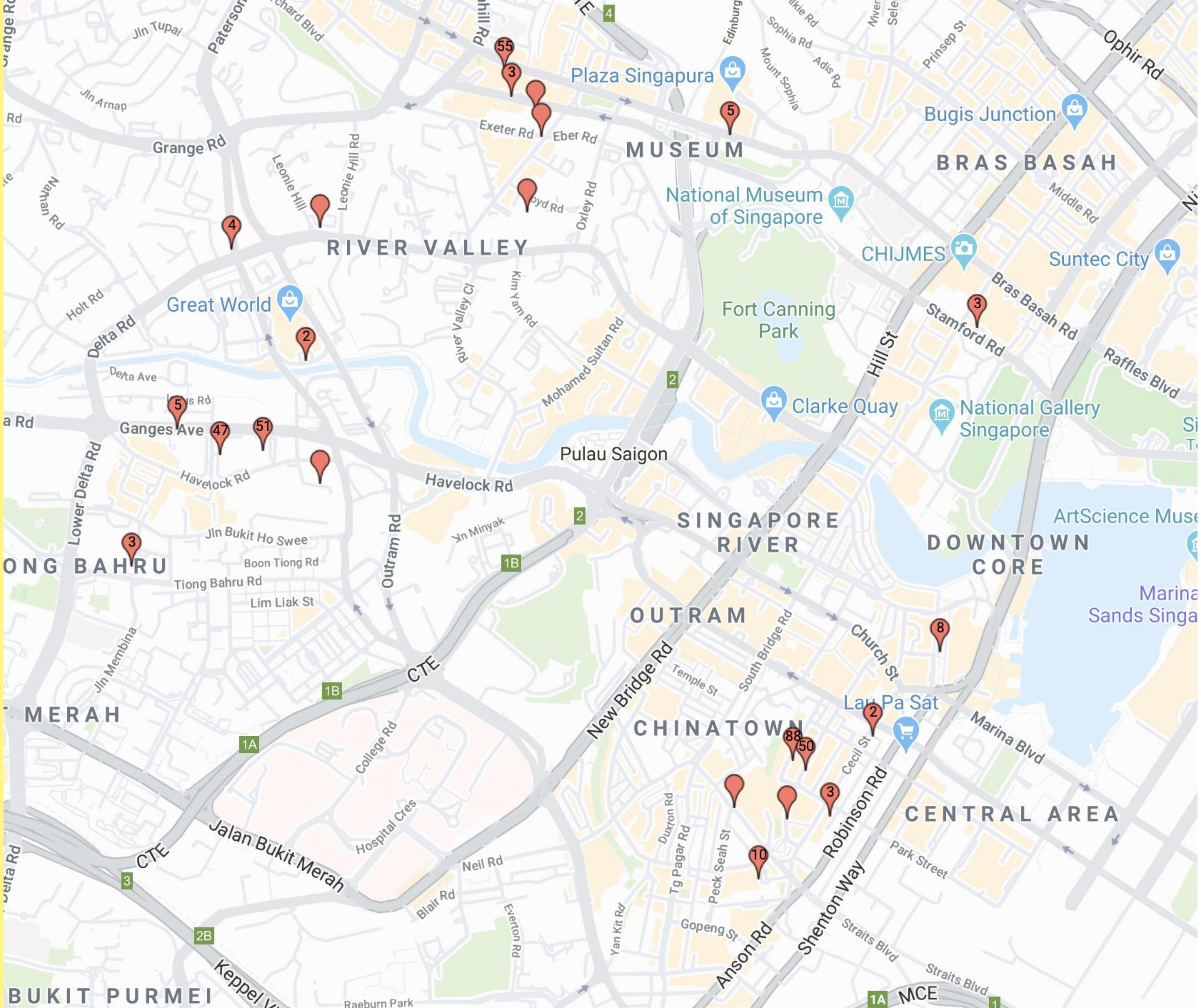


[mobile ID, timestamp]




[mobile ID, timestamp]
A
B


[mobile ID, timestamp]
A
B

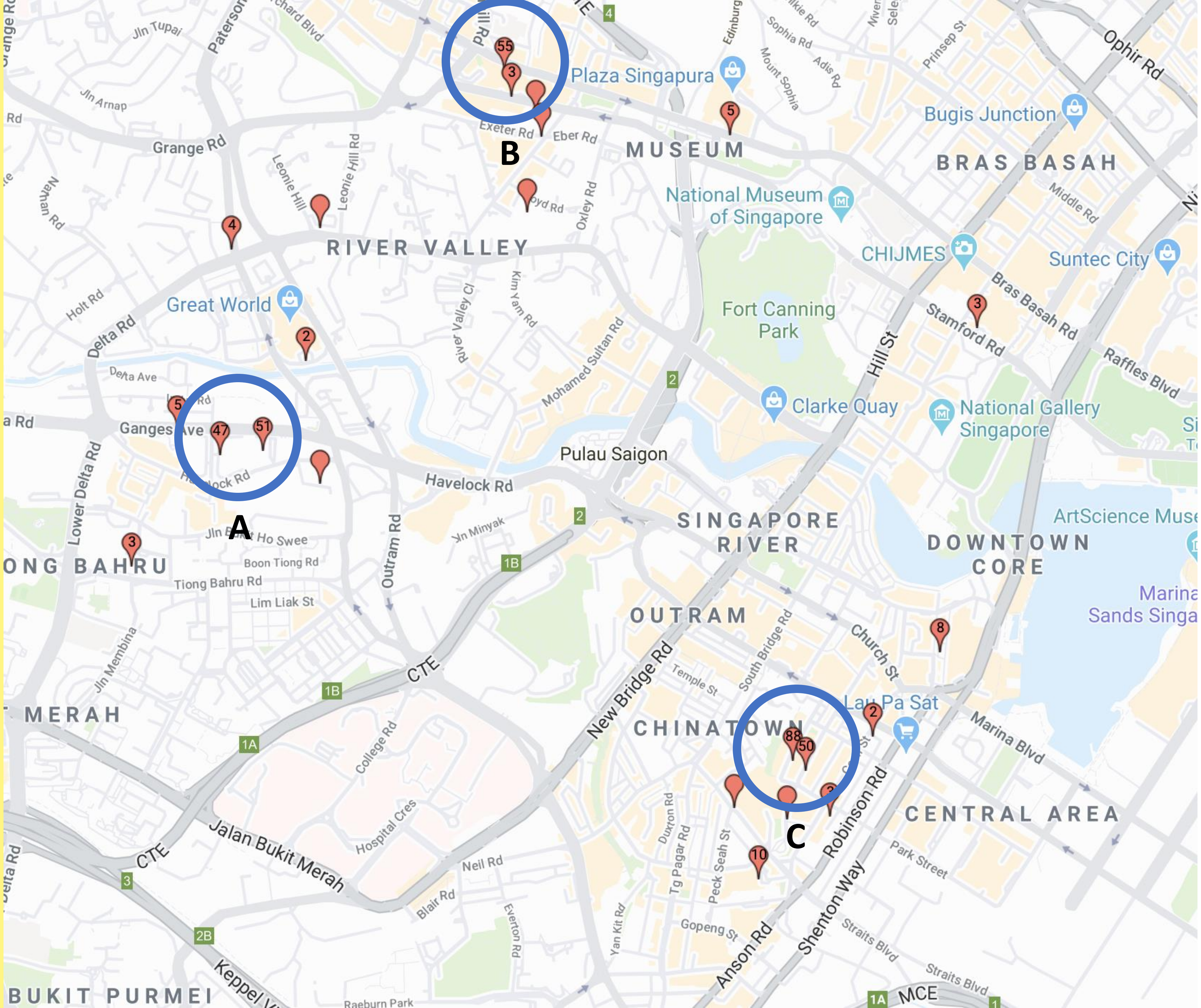


CX083914

7th October 2021

Red markers: cell towers

Number: number of connections



CX083914

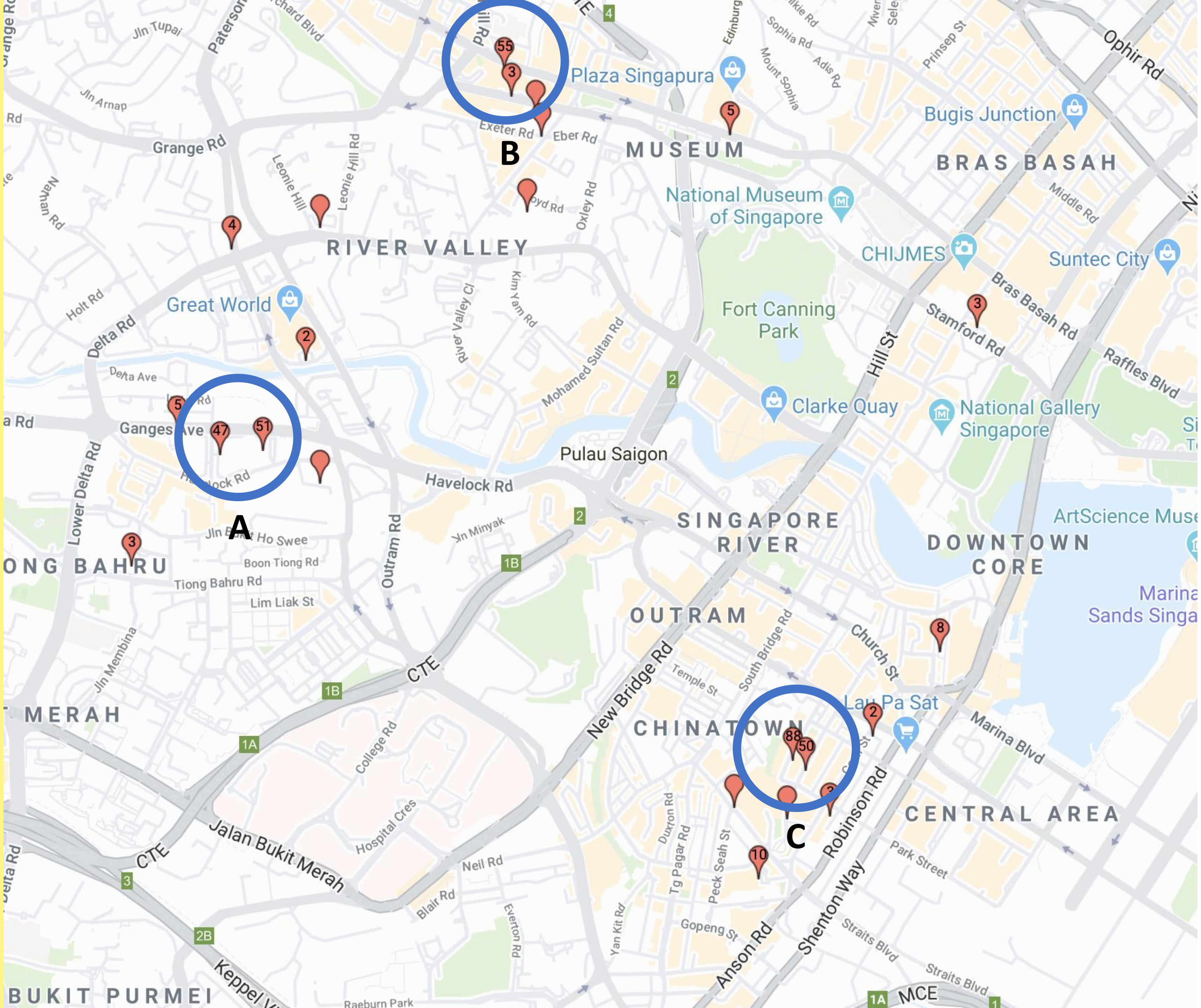
Start time	End time	Location
00:00	11:00	A: Tiong Bahru
11:00	12:00	B: Orchard
12:00	13:00	C: Chinatown
14:00	Next day	A: Tiong Bahru

Daily itinerary

Benefits for Transport Planning

- Proxy to **people's location** and **movement**
- **Extensive coverage** of the population (cross-sectional data)
- Data collected **continuously** (longitudinal data)
- **Infrastructure** already **in place**
- **High mobile phone penetration rates** in developed and developing countries
- **Understand travel demand** at an unprecedented level of detail

But …



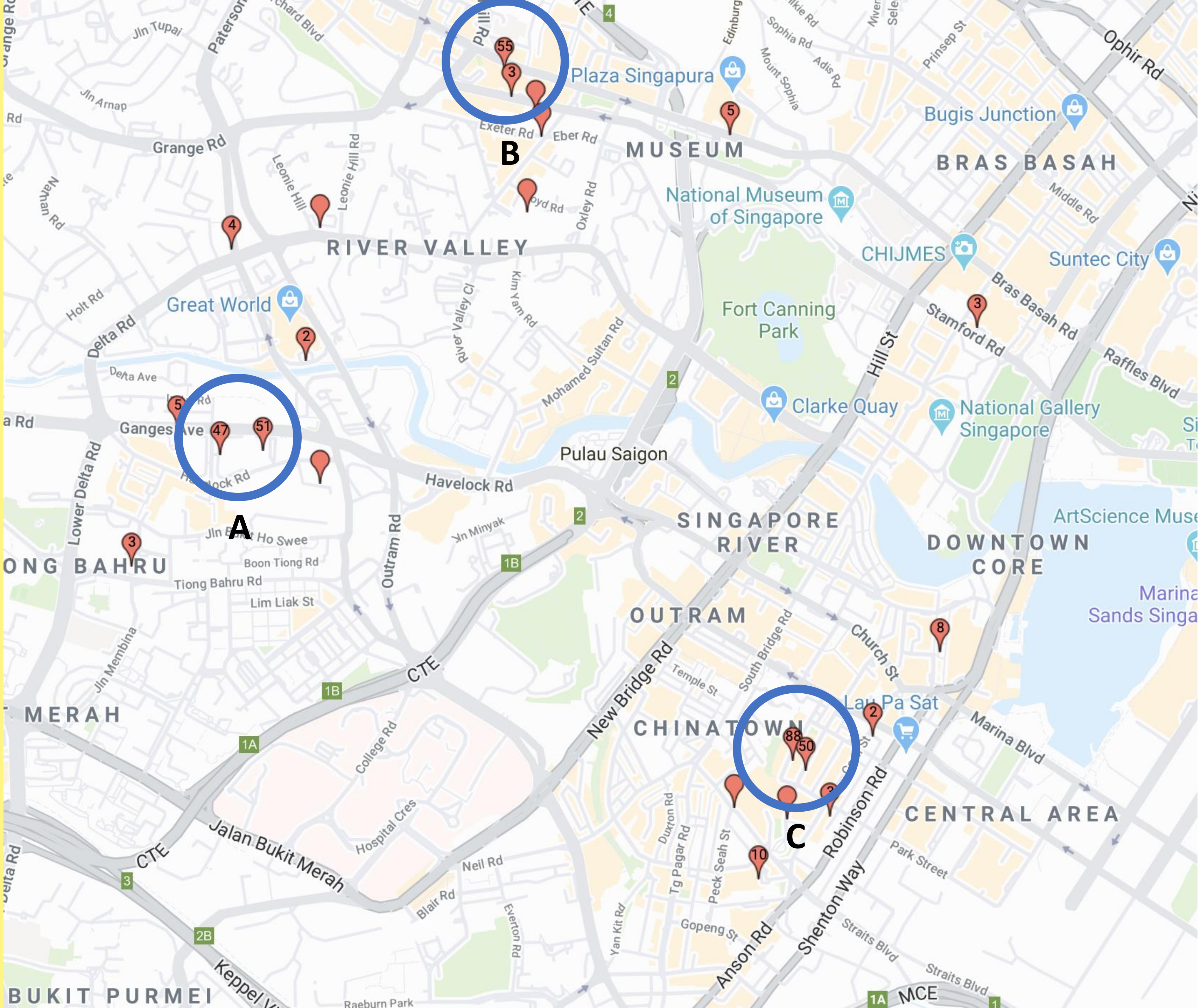
CX083914

Start time	End time	Location
00:00		
		...hinatown
14:00	Next day	A: Tiong Bahru

4 spatio-temporal points can reidentify a person!

Daily itinerary

De Montjoye, Y. A., Hidalgo, C. A., Verleysen, M., & Blondel, V. D. (2013). Unique in the crowd: The privacy bounds of human mobility. *Scientific reports*, *3*(1), 1-5.



Temo

Start time	End time	Location
00:00		
		…hinatown
14:00	Next day	A: Tiong Bahru

Privacy problem!

Daily itinerary

Agenda

1. Mobile network signalling data
2. Motivation
3. Intuitive example
4. Digital Twin Travellers
5. Experiment and results
6. Application
7. Conclusion

Can mobile phone data be used for

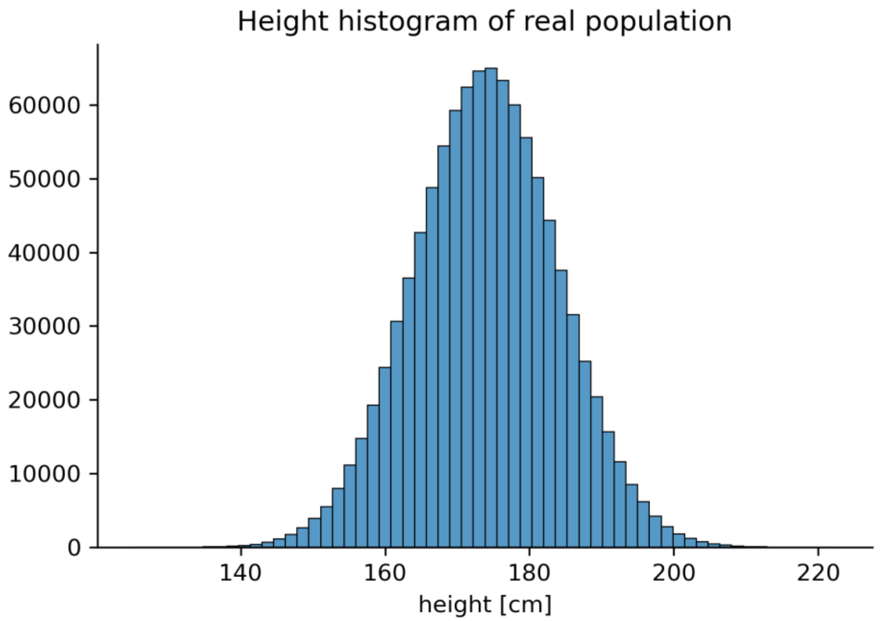
(disaggregated) **transport planning** without compromising users' **privacy**?

Agenda

1. Mobile network signalling data
2. Motivation
3. Intuitive example
4. Digital Twin Travellers
5. Experiment and results
6. Application
7. Conclusion

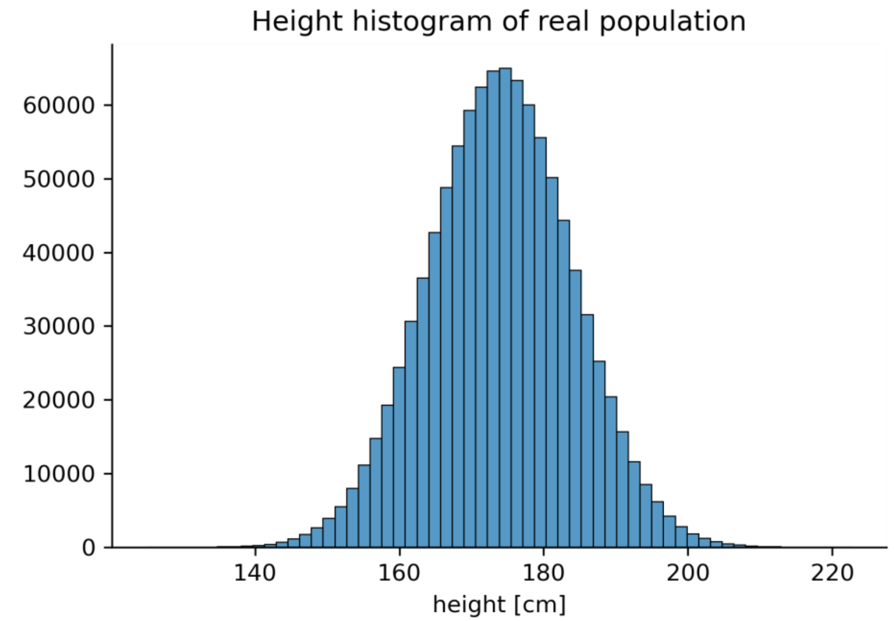
Real population height (cm)

176.4577
180.9375
161.8581
178.7459
...



Real population height (cm)

176.4577
180.9375
161.8581
178.7459
...



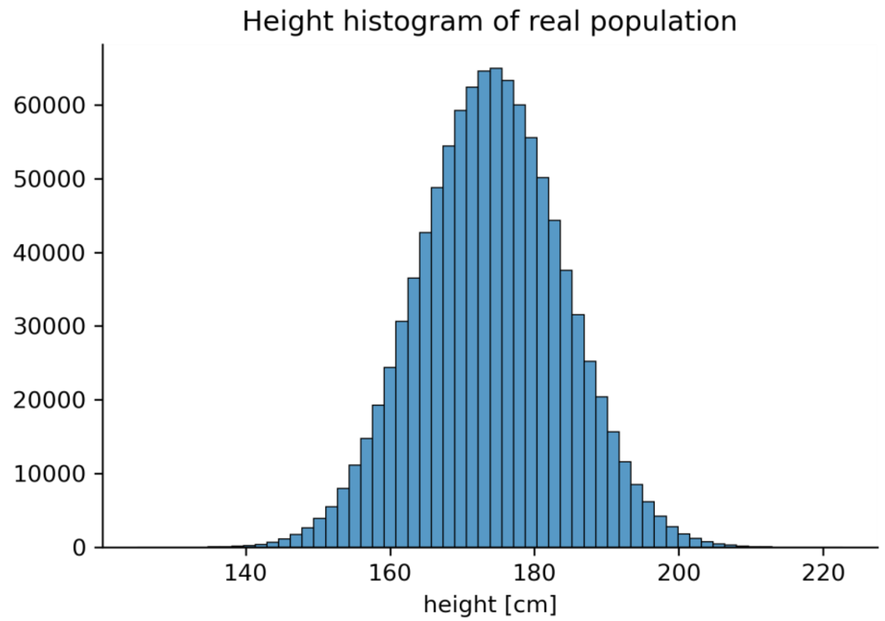

```
[ ]: np.random.normal()
```

Real population height (cm)

176.4577
180.9375
161.8581
178.7459
...

```
[2]: np.random.normal()
```

```
[2]: 182.7496
```

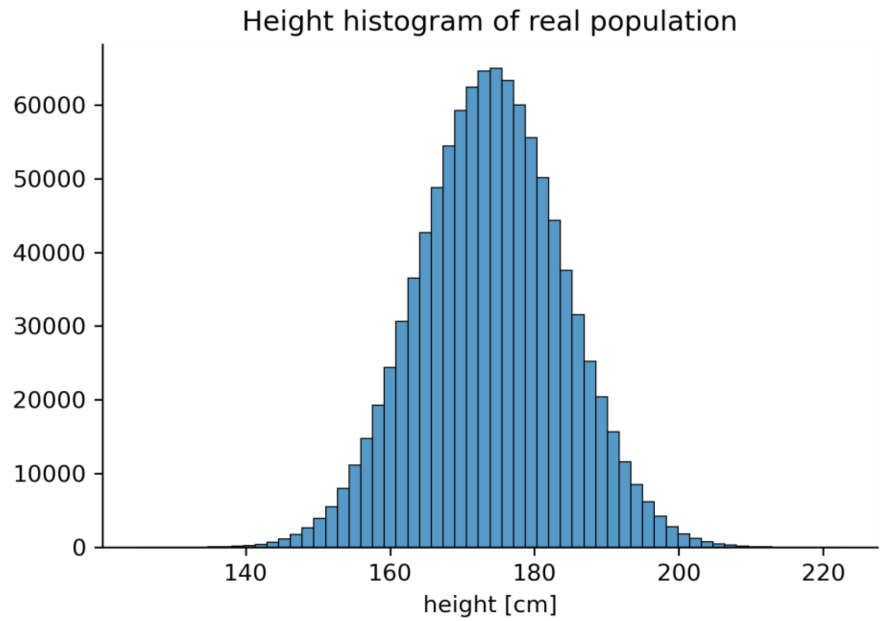


Real population height (cm)

176.4577
180.9375
161.8581
178.7459
...

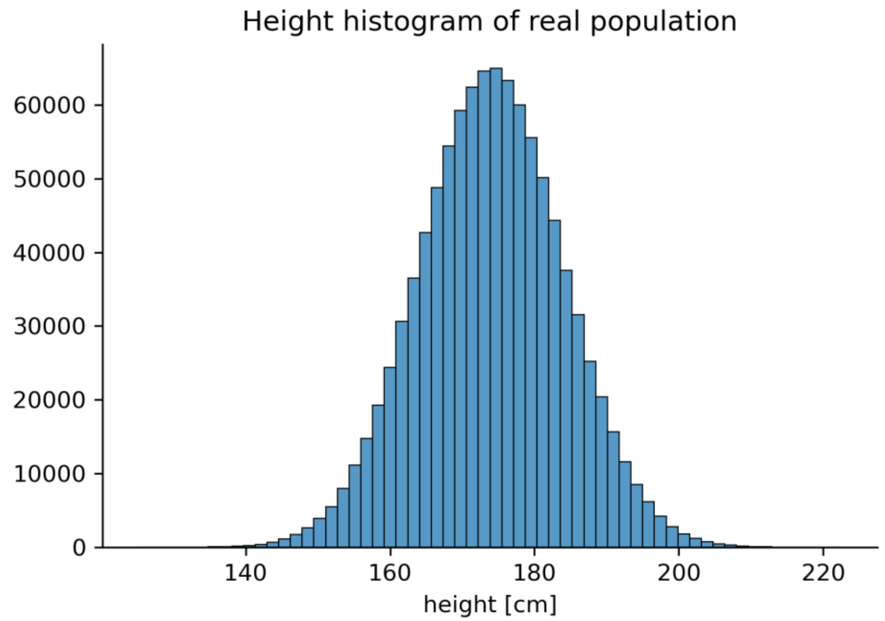
```
[3]: np.random.normal()
```

```
[3]: 176.8822
```



Real population height (cm)

176.4577
180.9375
161.8581
178.7459
...

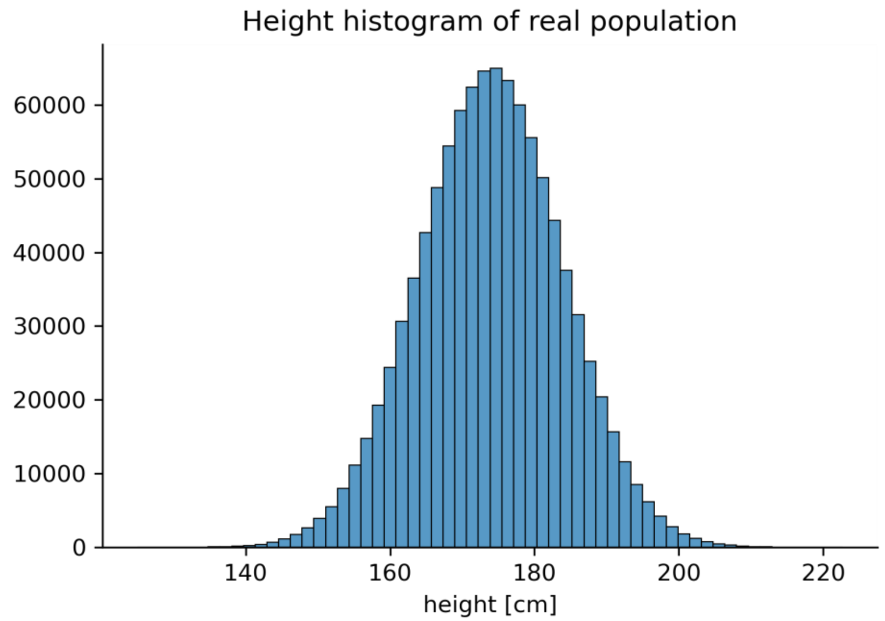



```
[4]: np.random.normal()
```

```
[4]: 171.0241
```

Real population height (cm)

176.4577
180.9375
161.8581
178.7459
...

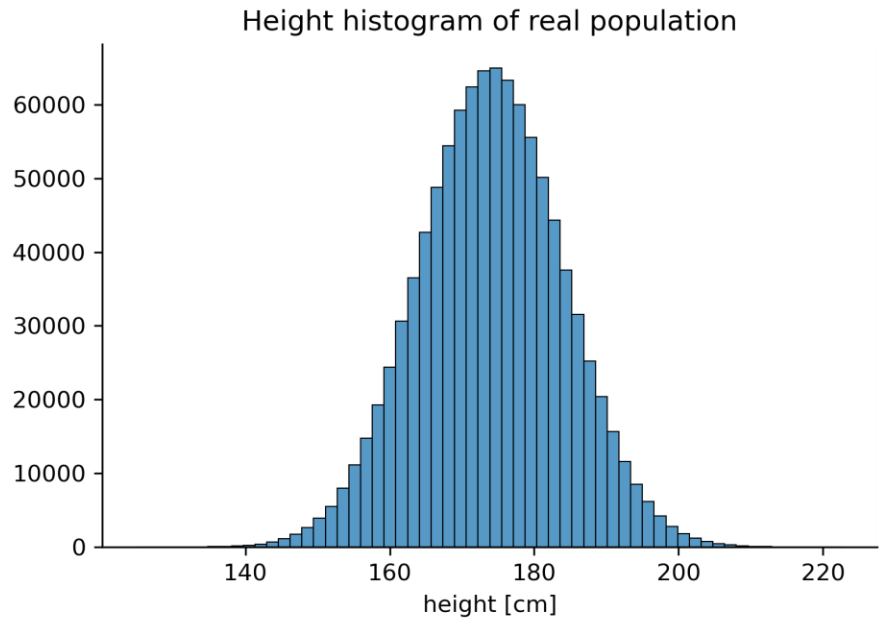



```
[5]: np.random.normal()
```

```
[5]:  167.5589
```

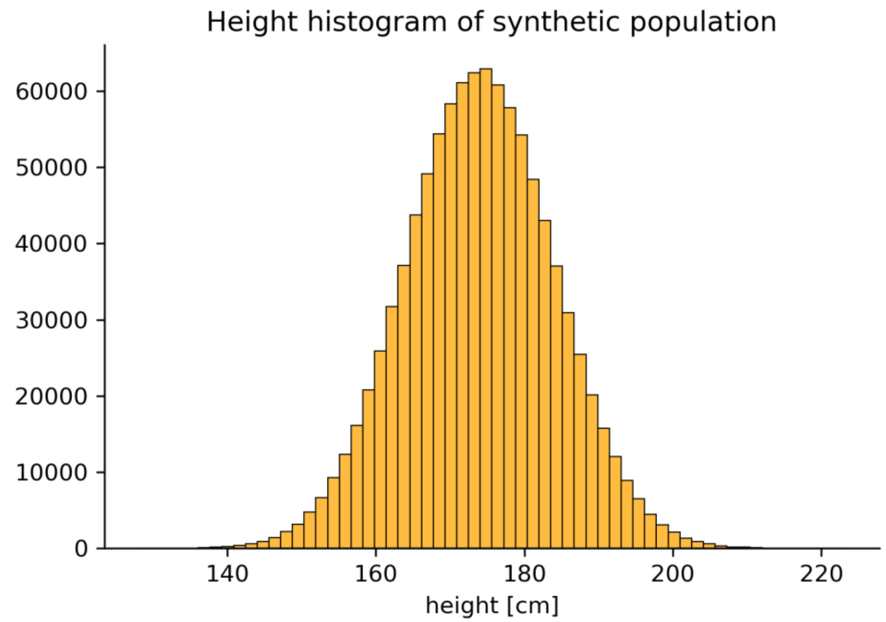
Real population height (cm)

176.4577
180.9375
161.8581
178.7459
...



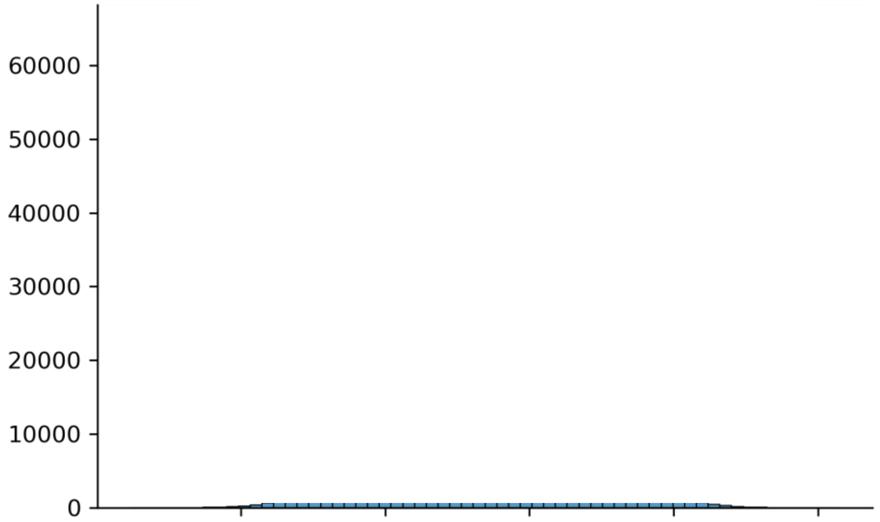
Synthetic population height (cm)

182.7496
176.8822
171.0241
167.5589
...



Real population itineraries

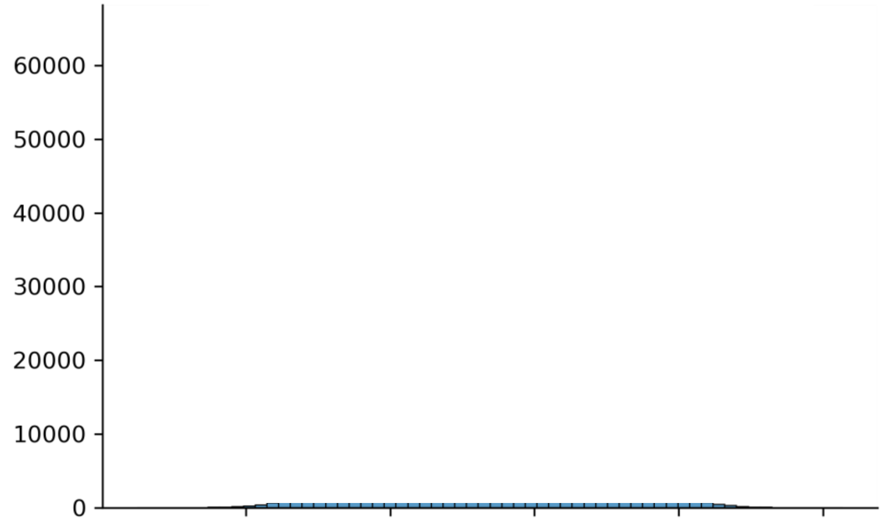
	user	startTime	subzone	duration	endTime
0	Temo	00:00	TIONG BAHRU	11	11:00
1	Temo	11:00	ORCHARD	1	12:00
2	Temo	12:00	CHINATOWN	1	13:00
3	Temo	14:00	TIONG BAHRU	15	29:00




```
[ ]: generateDigitalTwinTraveller()
```

Real population itineraries

	user	startTime	subzone	duration	endTime
0	Temo	00:00	TIONG BAHRU	11	11:00
1	Temo	11:00	ORCHARD	1	12:00
2	Temo	12:00	CHINATOWN	1	13:00
3	Temo	14:00	TIONG BAHRU	15	29:00



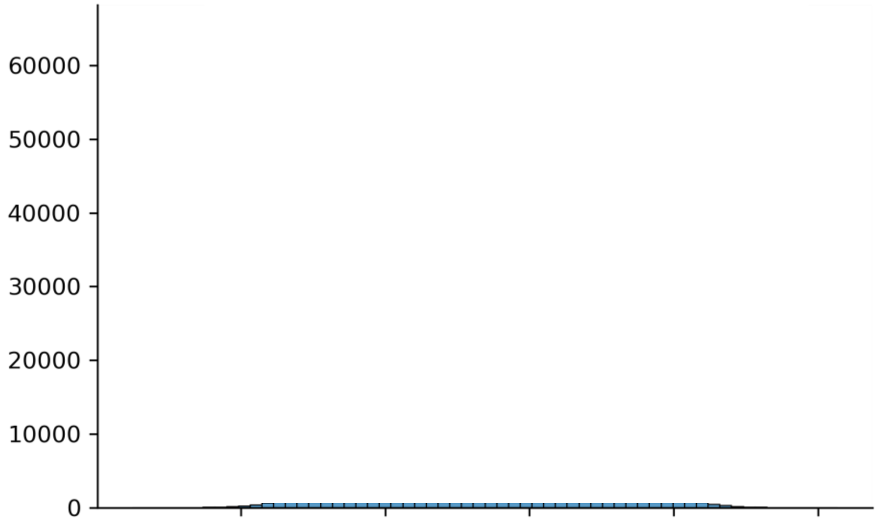
[2]:
```
generateDigitalTwinTraveller()
```

[2]:

agent	startTime	subzone	duration	endTime
A1	00:00:00	BOON LAY PLACE	8	08:00:00
A1	08:00:00	ANSON	9	17:00:00
A1	18:00:00	MACPHERSON	0	18:00:00
A1	20:00:00	BOON LAY PLACE	12	32:00:00

Real population itineraries

	user	startTime	subzone	duration	endTime
0	Temo	00:00	TIONG BAHRU	11	11:00
1	Temo	11:00	ORCHARD	1	12:00
2	Temo	12:00	CHINATOWN	1	13:00
3	Temo	14:00	TIONG BAHRU	15	29:00




```
[3]: generateDigitalTwinTraveller()
```

[3]:

agent	startTime	subzone	duration	endTime
A2	00:00:00	SERANGOON CENTRAL	7	07:00:00
A2	08:00:00	FRANKEL	3	11:00:00
A2	11:00:00	MARINE PARADE	2	13:00:00
A2	13:00:00	FRANKEL	3	16:00:00
A2	16:00:00	SERANGOON CENTRAL	16	32:00:00

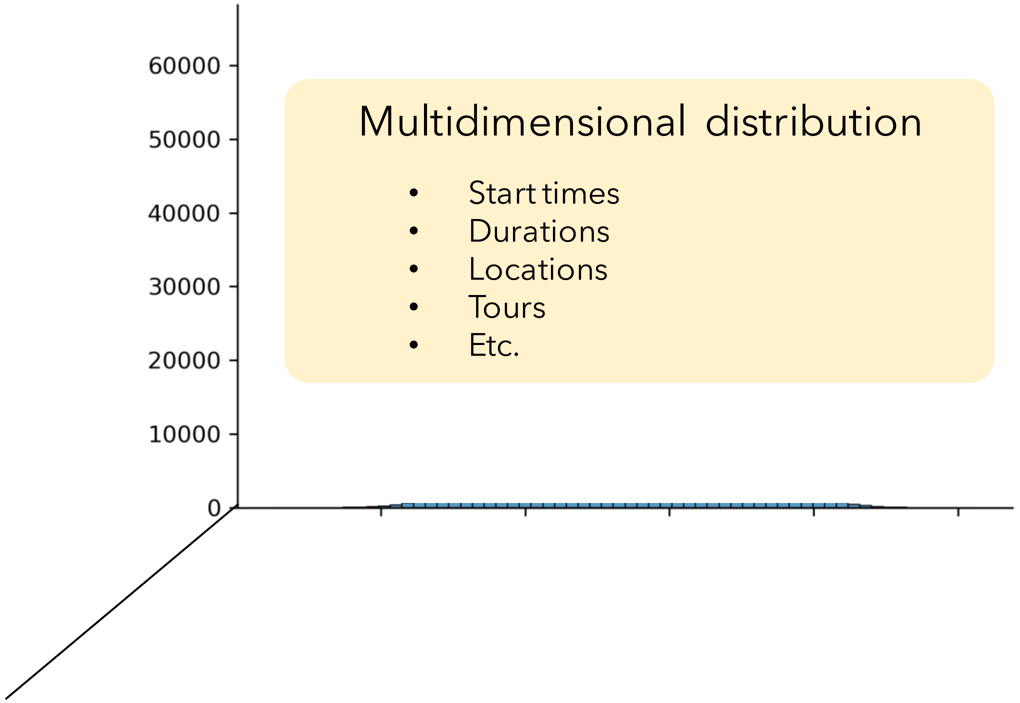
Real population itineraries

	user	startTime	subzone	duration	endTime
0	Temo	00:00	TIONG BAHRU	11	11:00
1	Temo	11:00	ORCHARD	1	12:00
2	Temo	12:00	CHINATOWN	1	13:00
3	Temo	14:00	TIONG BAHRU	15	29:00

```
[4]: generateDigitalTwinTraveller()
```

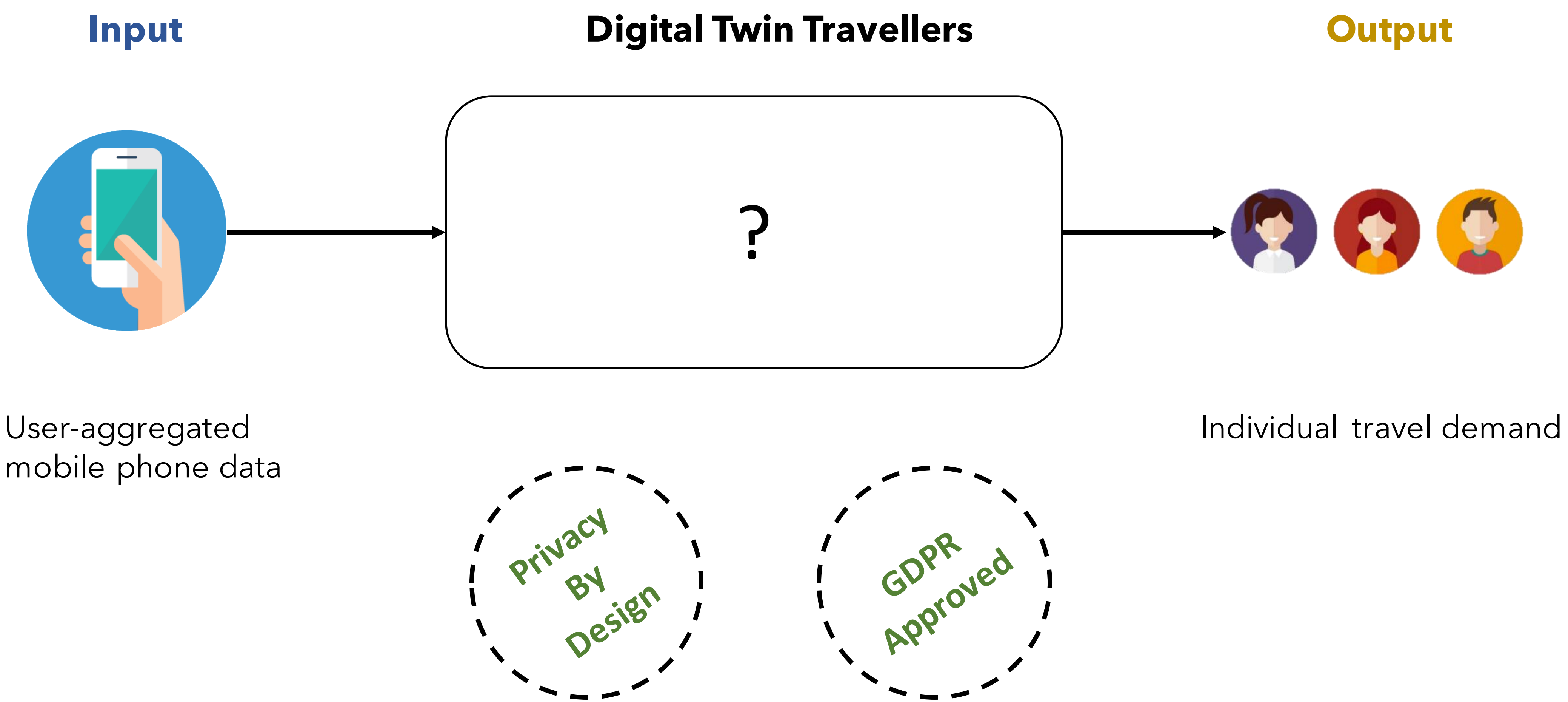
[4]:

agent	startTime	subzone	duration	endTime
A3	00:00:00	SELETAR HILLS	8	08:00:00
A3	08:00:00	RAFFLES PLACE	9	17:00:00
A3	19:00:00	BALESTIER	0	19:00:00
A3	20:00:00	SELETAR HILLS	11	31:00:00



Multidimensional distribution

- Start times
- Durations
- Locations
- Tours
- Etc.



Input
Digital Twin Travellers
Output
?
User-aggregated mobile phone data
Individual travel demand
Privacy By Design
GDPR Approved

Research questions

1. Can we **synthesise** realistic and **individual mobility demand** from **aggregates** of mobile phone data?

Research questions

2. How can we **validate** that the **synthetic** mobility **population** behaves **similarly** to the **real population**?

Agenda

1. Mobile network signalling data
2. Motivation
3. Intuitive example
4. Digital Twin Travellers
5. Experiment and results
6. Application
7. Conclusion

Setting

$f_{\boldsymbol{X}}(\boldsymbol{x}) = P(\boldsymbol{X}_1 = \boldsymbol{x}_1,\ \boldsymbol{X}_2 = \boldsymbol{x}_2,\ \dots,\ \boldsymbol{X}_n = \boldsymbol{x}_N)$

True distribution of daily itineraries

$\boldsymbol{X}_k = [S_k,\ D_k,\ Z_k,\ \dots]$

Daily itinerary random variables:

Start time, Duration, Location, …

$g_{\boldsymbol{X}}(\boldsymbol{x}) \approx f_{\boldsymbol{X}}(\boldsymbol{x})$

Generative model as an approximation

$g_{\boldsymbol{X}}(\boldsymbol{x}) = P(\boldsymbol{X}_1) \prod_{k=2}^{N} P(\boldsymbol{X}_k | \boldsymbol{X}_{k-1})$

Individual mobility as a Markovian system

1st step

Extend a typical Markov model using Dynamic Bayesian Networks to model individual mobility

Why choosing a Markov model and Dynamic Bayesian Networks?

Learning a Dynamic Bayesian Network via Maximum Likelihood

Privacy By Design

GDPR Approved

1. Obtain **Likelihood Function**

   $Likelihood(\mathbf{\Theta}) = P(\boldsymbol{Data}|Model(\mathbf{\Theta}))$

2. **Minimize negative log-likelihood** subject to parameters of random variables sum to one

3. For **categorical** and fully **observable** random variables a **closed-form solution** is obtained.

4. Learning is counting the **occurrences** in the data (i.e. frequencies).

Privacy By Design

GDPR Approved

Input

User-aggregated
mobile phone data

Digital Twin Travellers

?

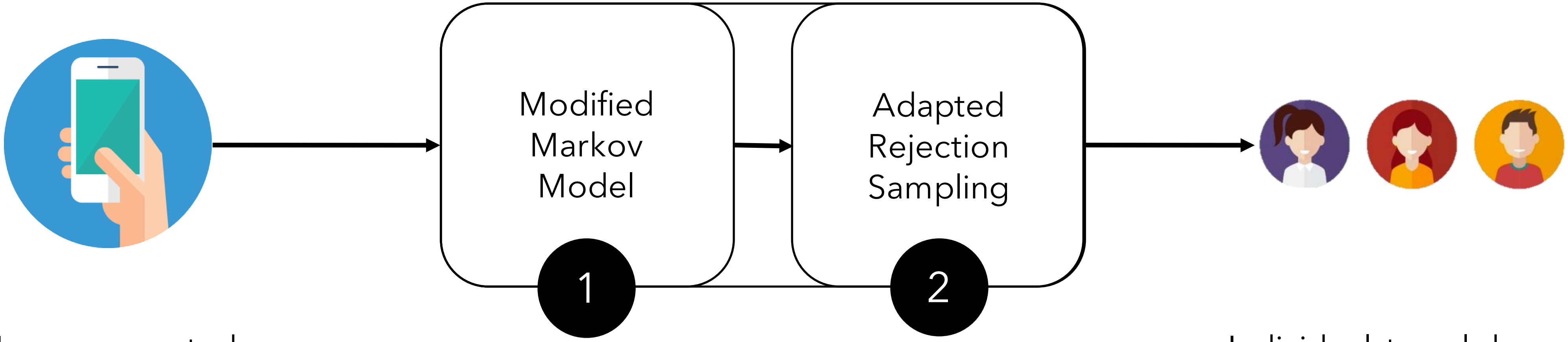
Output

Individual travel demand

Privacy By Design

GDPR Approved

Input | **Digital Twin Travellers** | **Output**



User-aggregated mobile phone data

Individual travel demand

Generative Model g(x)

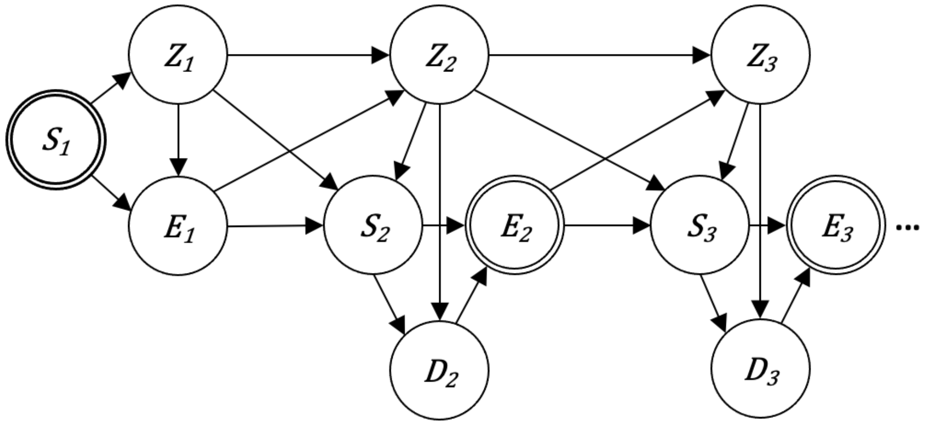
Strategy to get closer to true distribution f(x)

Digital Twin Travellers

Modified Markov Models of Urban Mobility

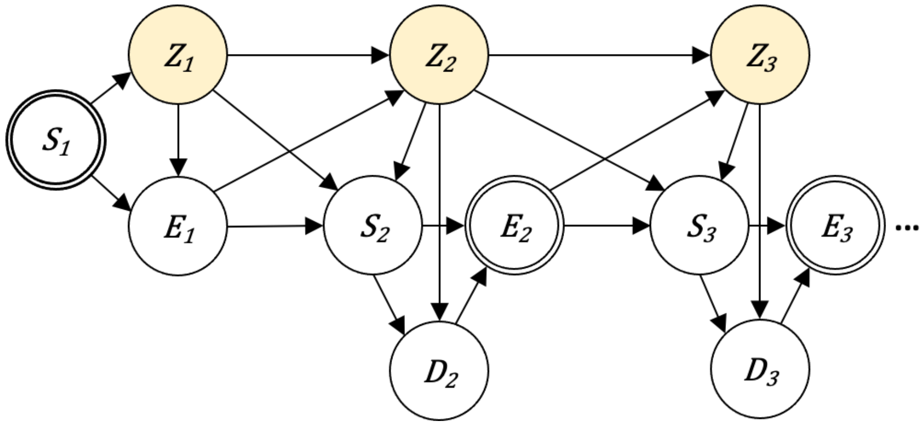
BM: Baseline Markov model

1. Graphical model



BM: Baseline Markov model

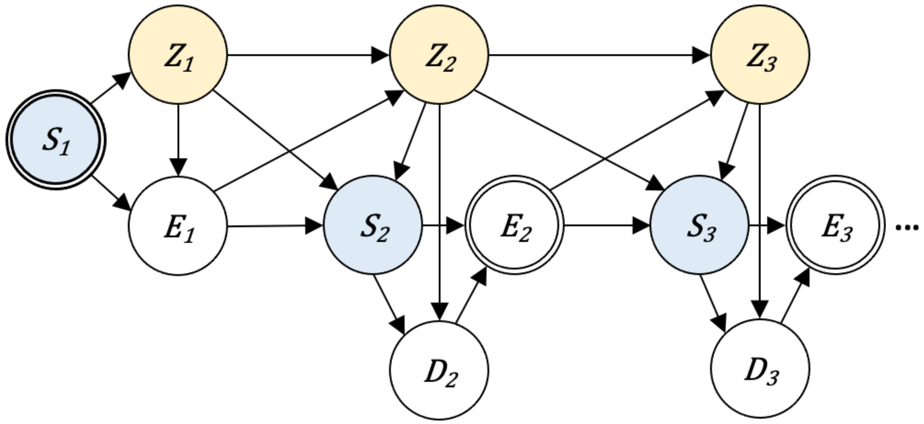
1. Graphical model



Zones

BM: Baseline Markov model

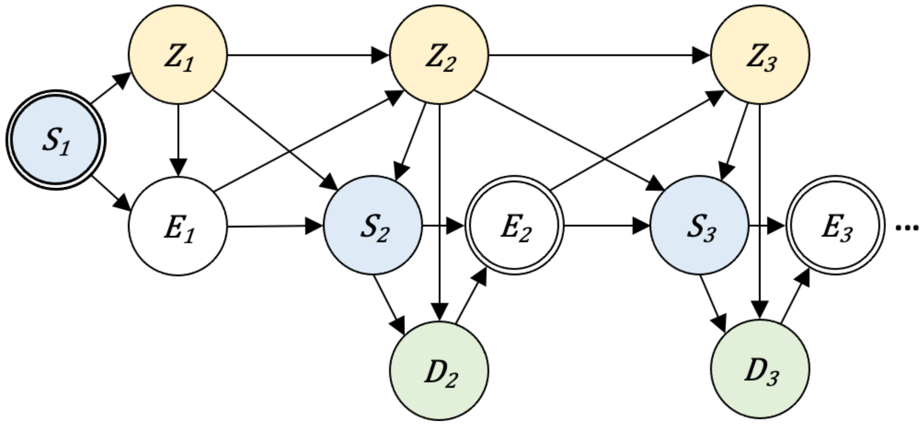
1. Graphical model



Start times

BM: Baseline Markov model

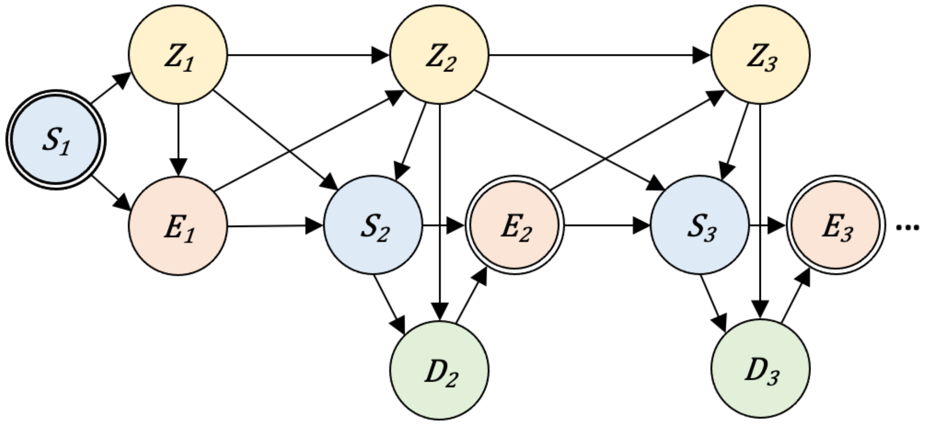
1. Graphical model



Durations

BM: Baseline Markov model

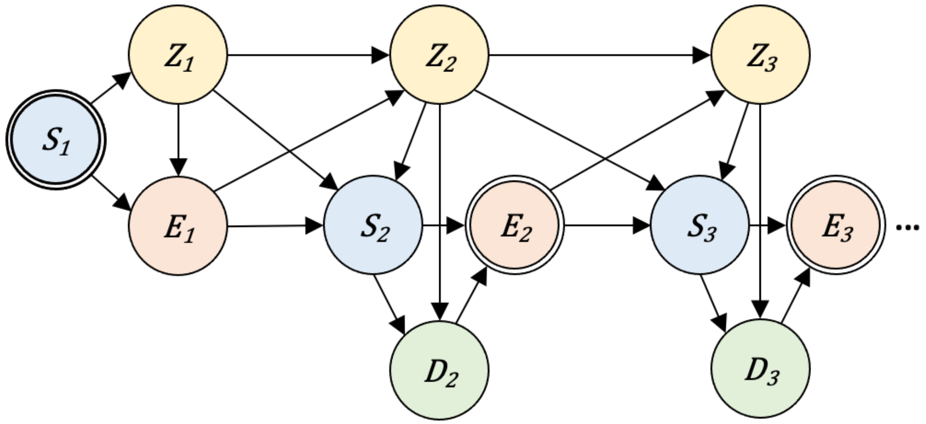
1. Graphical model



End times

BM: Baseline Markov model

1. Graphical model



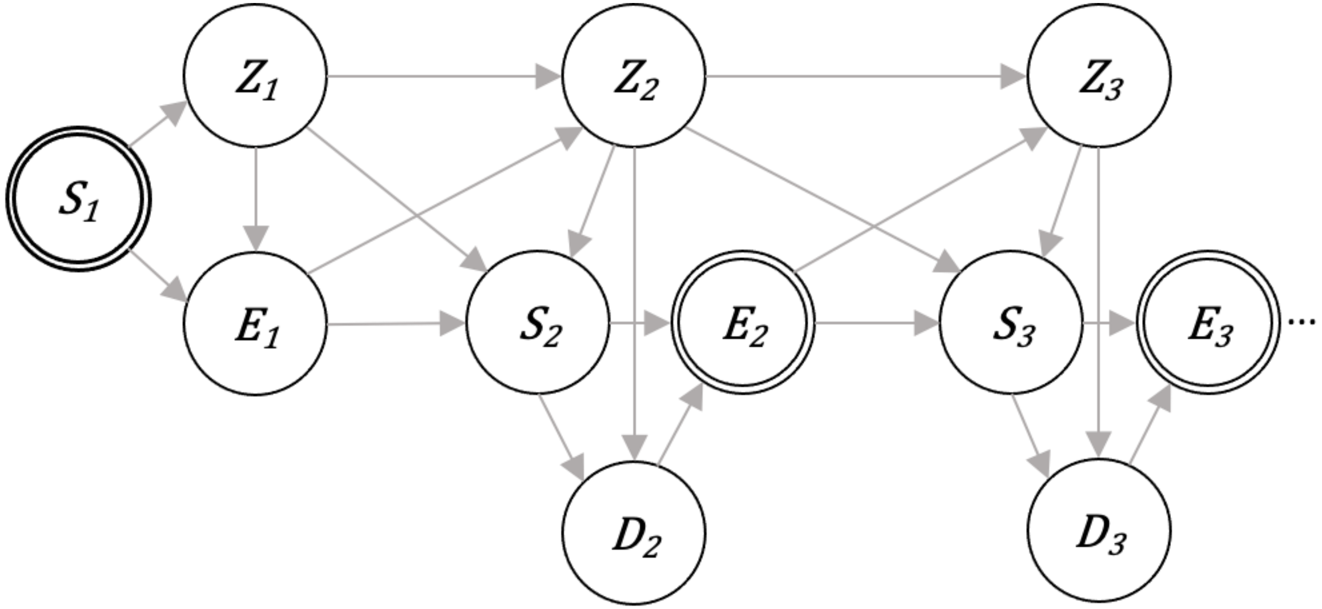
3. Histograms required

	CPD	Histogram
1	$P(Z_1 \mid S_1)$	Initial location
2	$P(E_1 \mid Z_1, S_1)$	Initial departure time given zone
3	$P(Z_k \mid Z_{k-1}, E_{k-1})$	Dynamic OD matrix
4	$P(S_k \mid Z_k, Z_{k-1}, E_{k-1})$	Start time given OD pair and origin end time
5	$P(D_k \mid Z_k, S_k)$	Duration given zone and start time

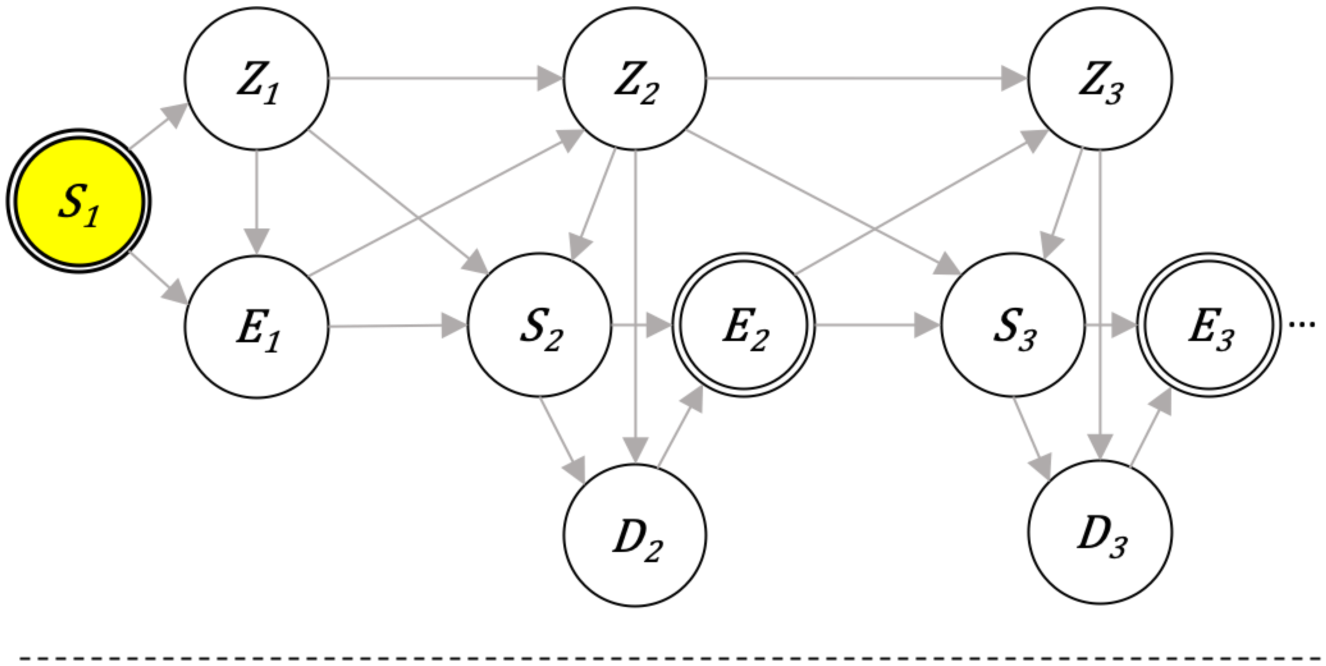
2. Joint Probability Distribution

$$\boldsymbol{P}(\boldsymbol{Z}_{\mathbf{1}:\boldsymbol{N}}, \boldsymbol{S}_{\mathbf{1}:\boldsymbol{N}}, \boldsymbol{E}_{\mathbf{1}}, \boldsymbol{D}_{\mathbf{2}:\boldsymbol{N}}) = P(S_1)P(Z_1 \mid S_1)P(E_1 \mid S_1, Z_1)\prod_{k=2}^{N} P(Z_k \mid E_{k-1}, Z_{k-1})P(D_k \mid Z_k, S_k)P(S_k \mid Z_{k-1}, E_{k-1})$$

BM

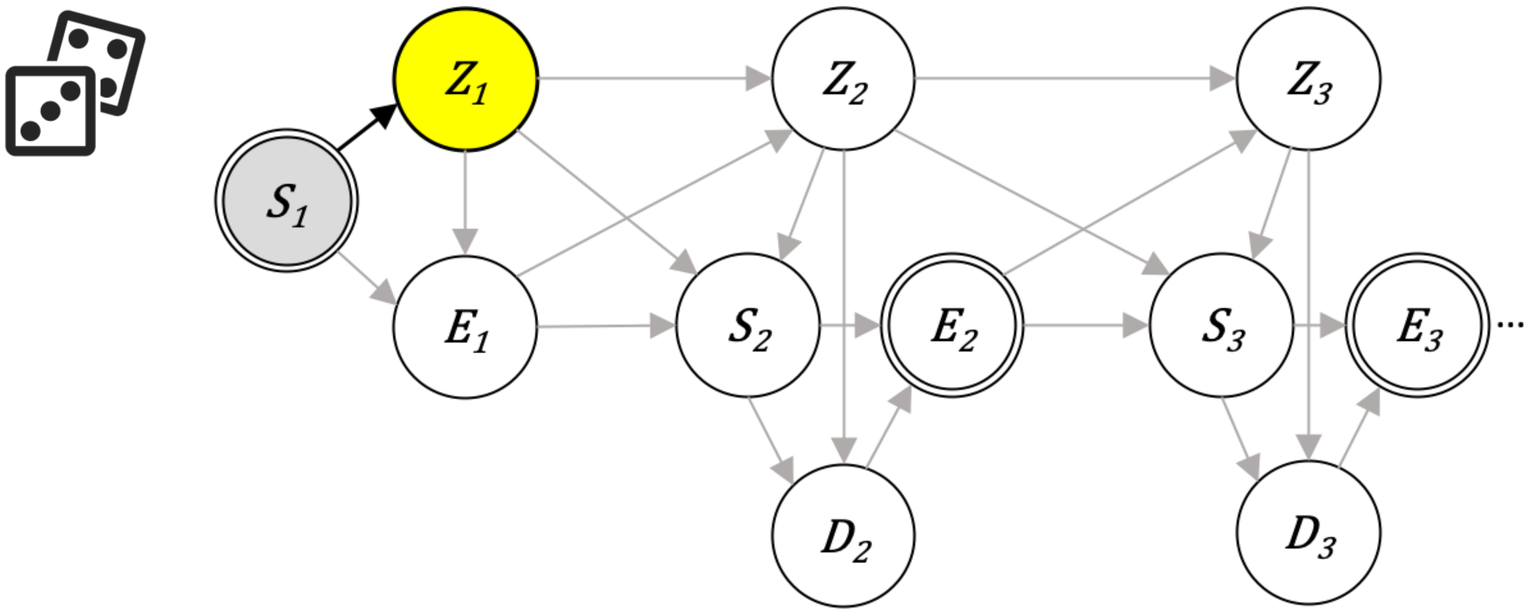


BM



S_1 = 00:00

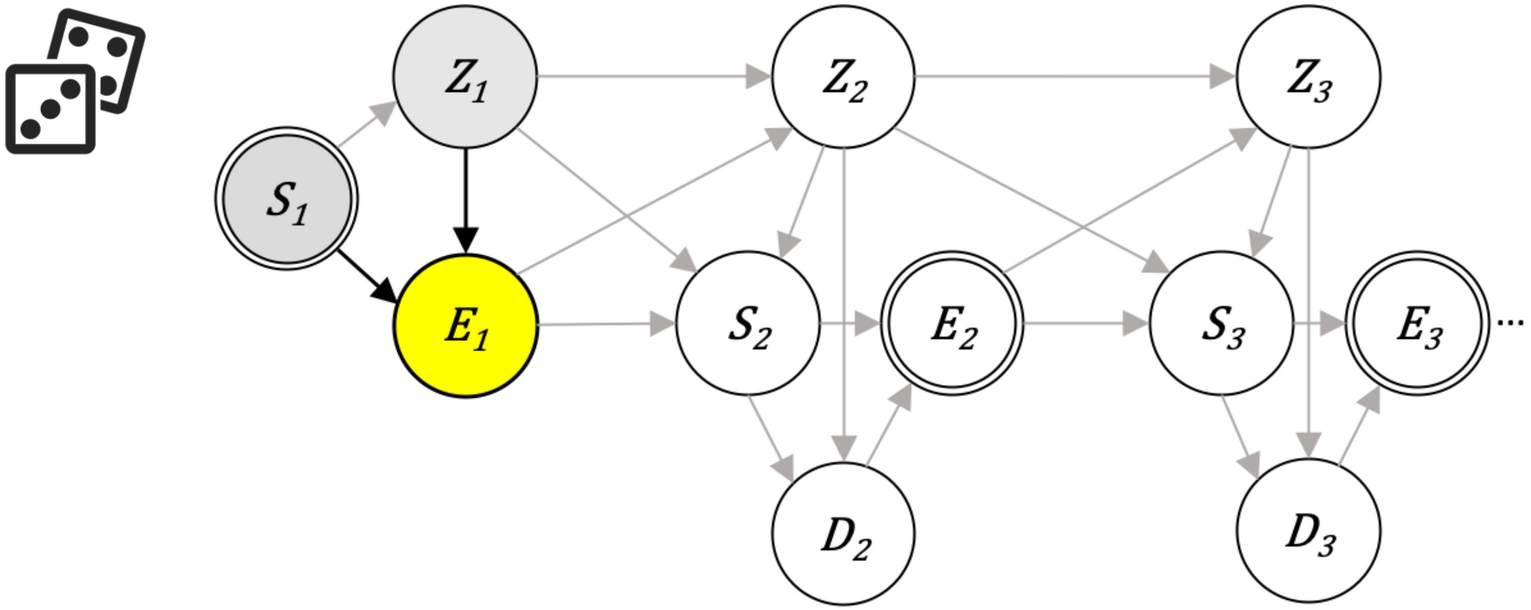
BM



Z_1 = Tiong Bahru

S_1 = 00:00

BM

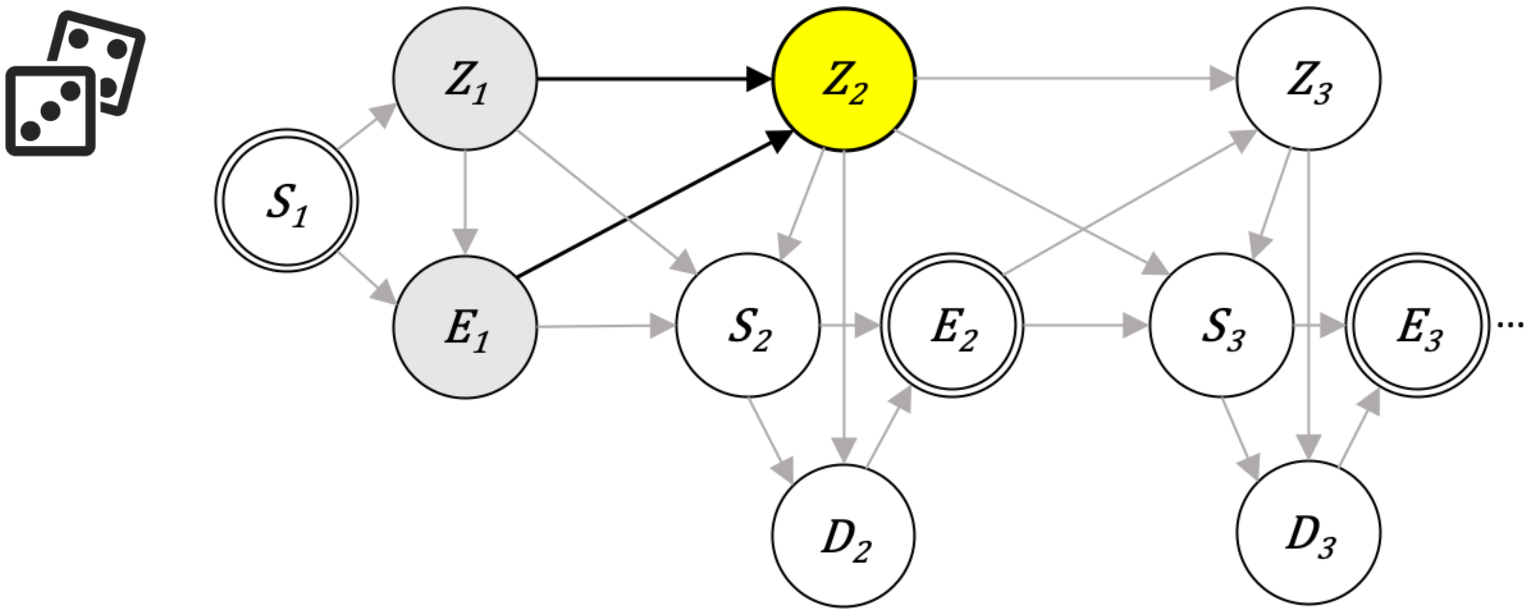


Z_1 = Tiong Bahru

S_1 = 00:00

E_1 = 08:00

BM



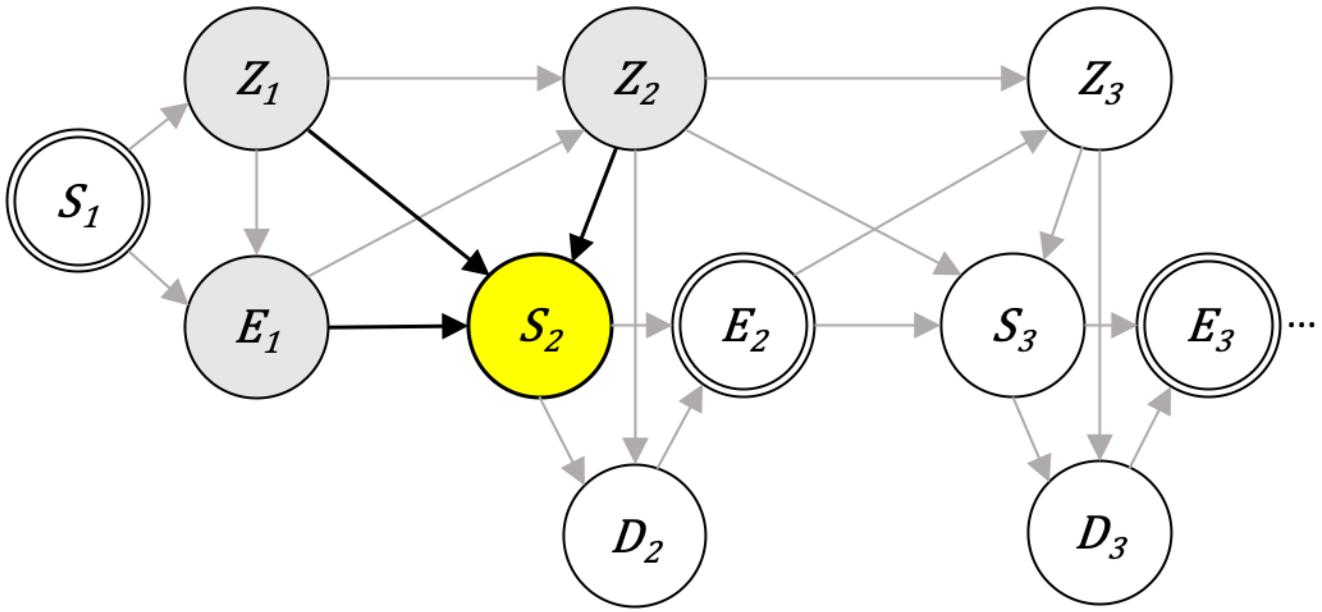
Z_1 = Tiong Bahru

S_1 = 00:00

E_1 = 08:00

Z_2 = Raffles Place

BM



Z_1 = Tiong Bahru

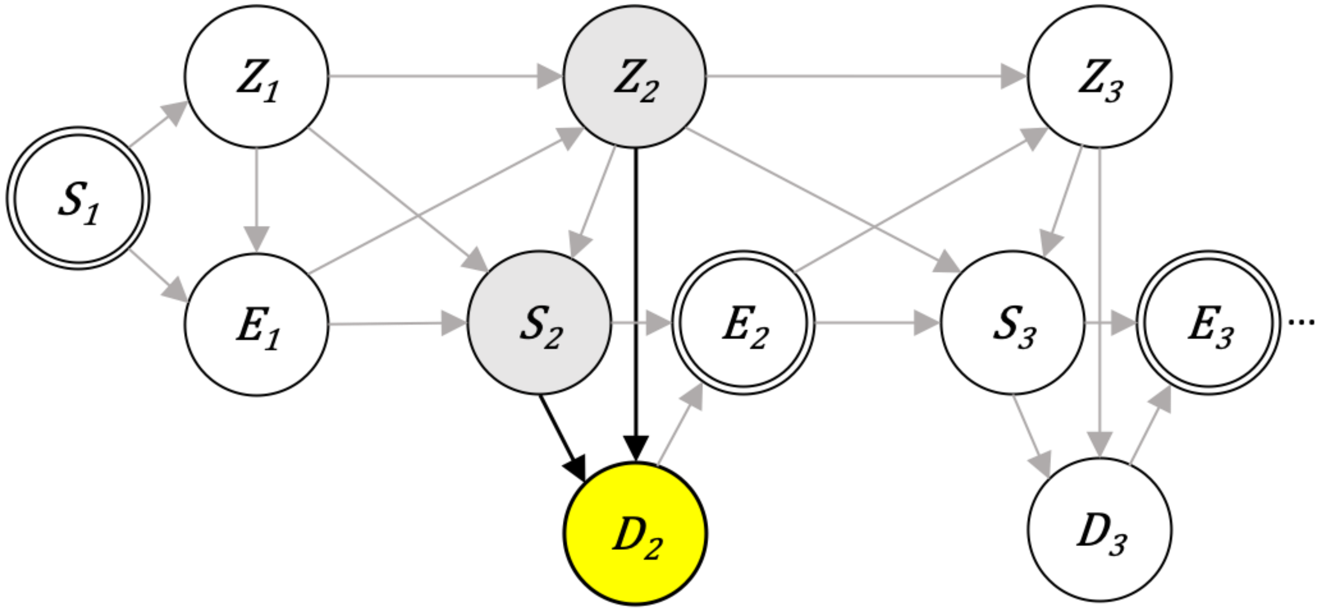
S_1 = 00:00

E_1 = 08:00

Z_2 = Raffles Place

S_2 = 09:00

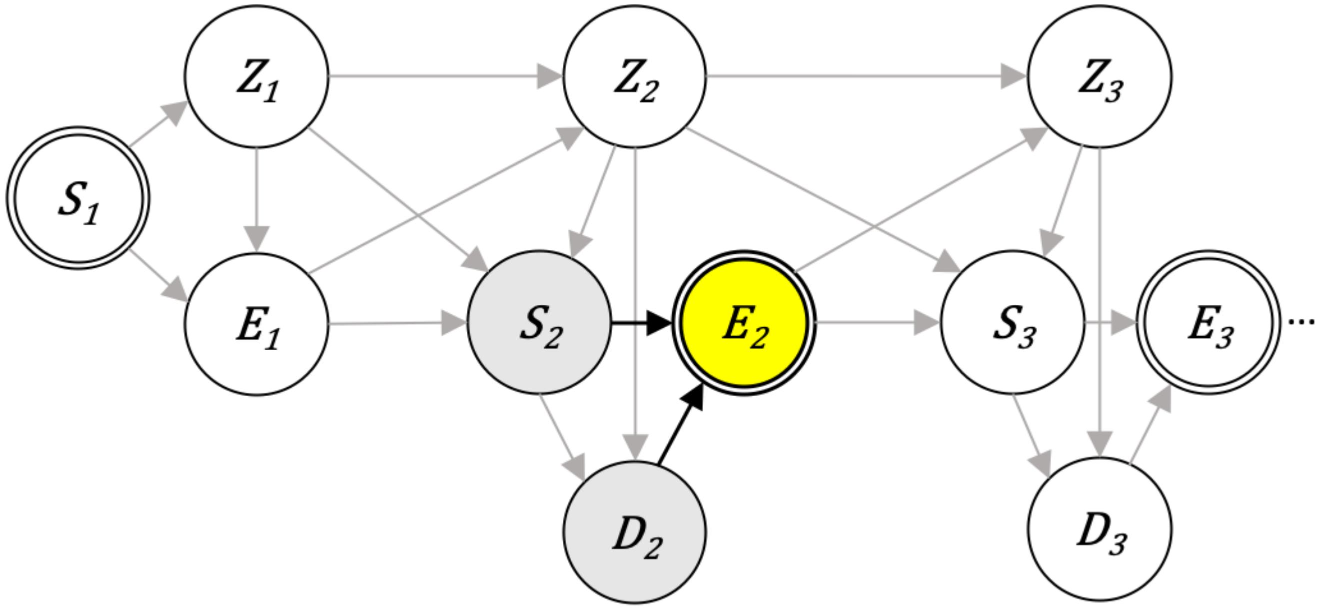
BM



Z_1 = Tiong Bahru
S_1 = 00:00
E_1 = 08:00

Z_2 = Raffles Place
S_2 = 09:00
D_2 = 4 hours

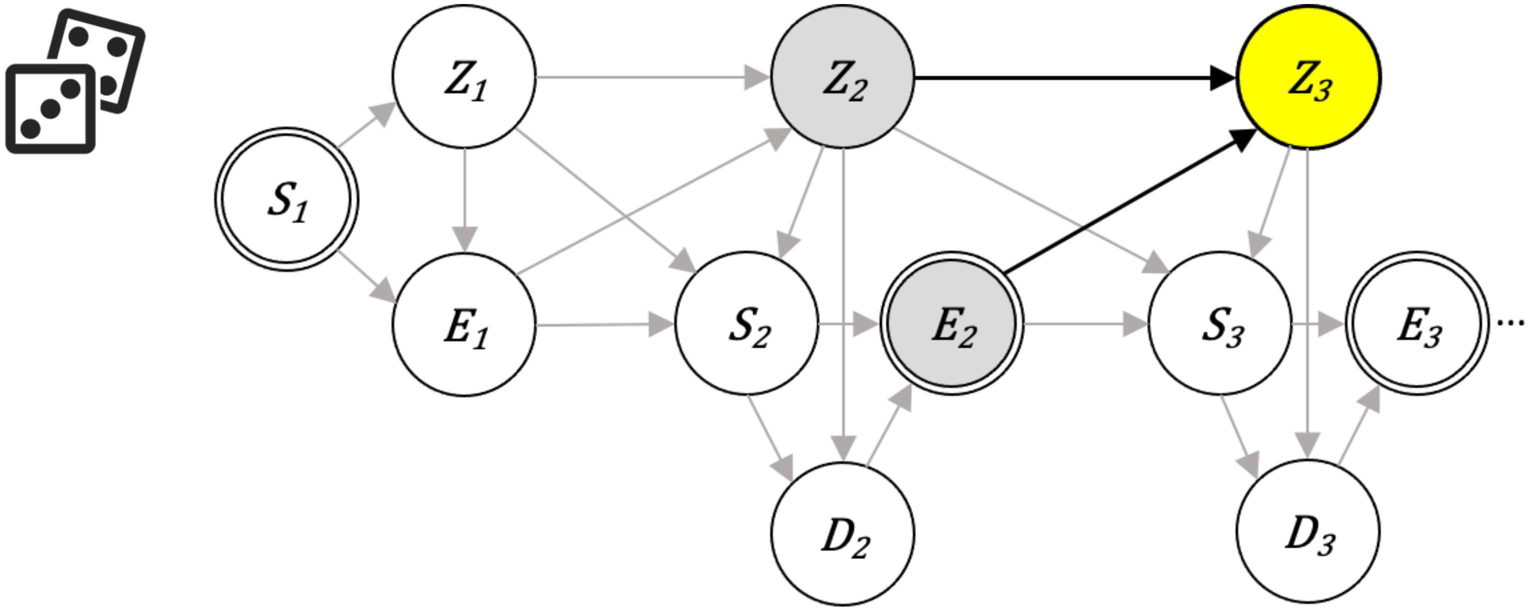
BM



Z_1 = Tiong Bahru
S_1 = 00:00
E_1 = 08:00

Z_2 = Raffles Place
S_2 = 09:00
D_2 = 4 hours
E_2 = 13:00

BM

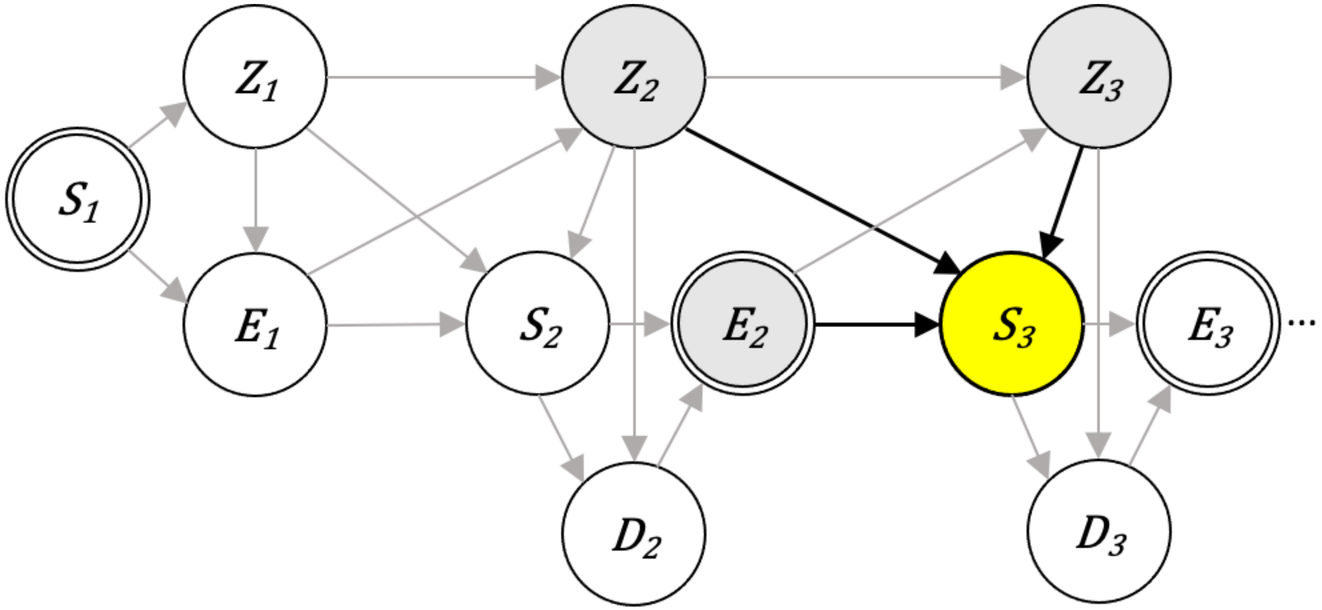


Z_1 = Tiong Bahru
S_1 = 00:00
E_1 = 08:00

Z_2 = Raffles Place
S_2 = 09:00
D_2 = 4 hours
E_2 = 13:00

Z_3 = Maxwell

BM

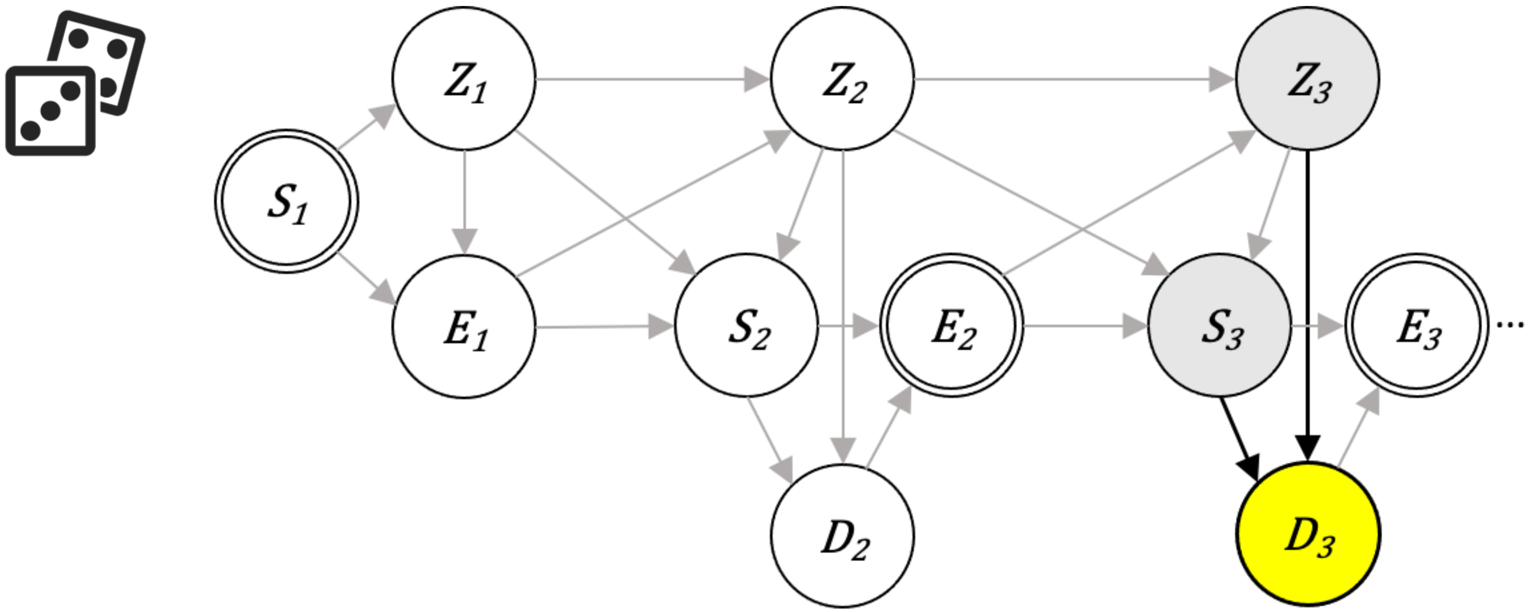


Z_1 = Tiong Bahru
S_1 = 00:00
E_1 = 08:00

Z_2 = Raffles Place
S_2 = 09:00
D_2 = 4 hours
E_2 = 13:00

Z_3 = Maxwell
S_3 = 13:00

BM

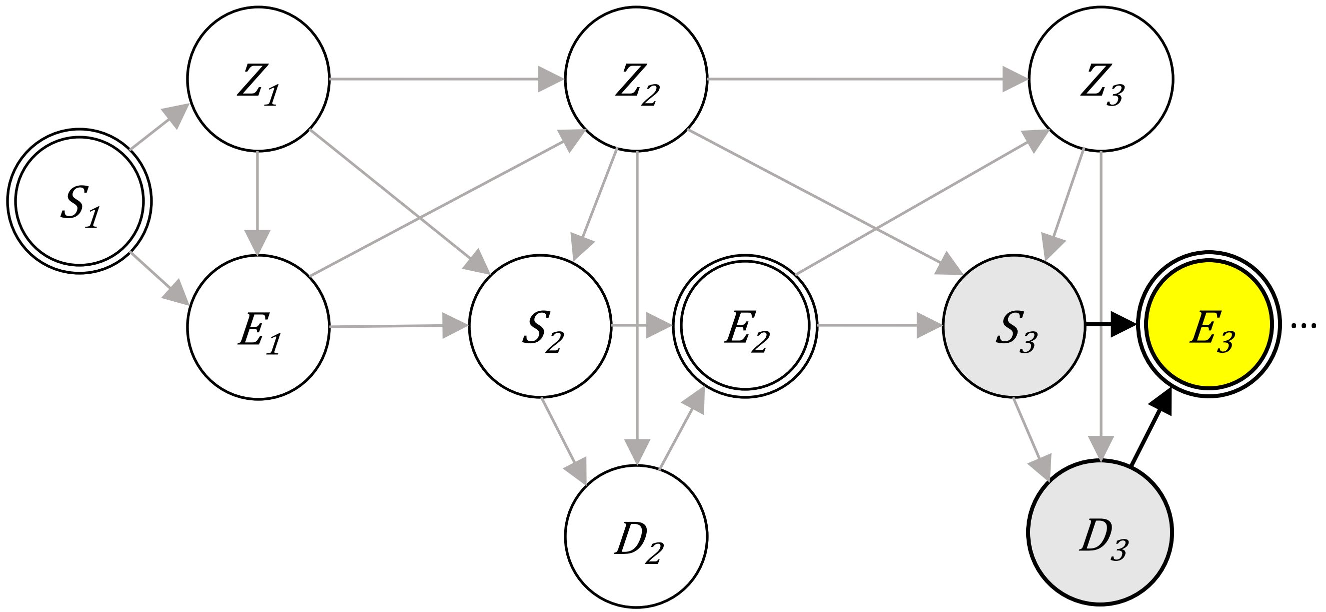


Z_1 = Tiong Bahru
S_1 = 00:00
E_1 = 08:00

Z_2 = Raffles Place
S_2 = 09:00
D_2 = 4 hours
E_2 = 13:00

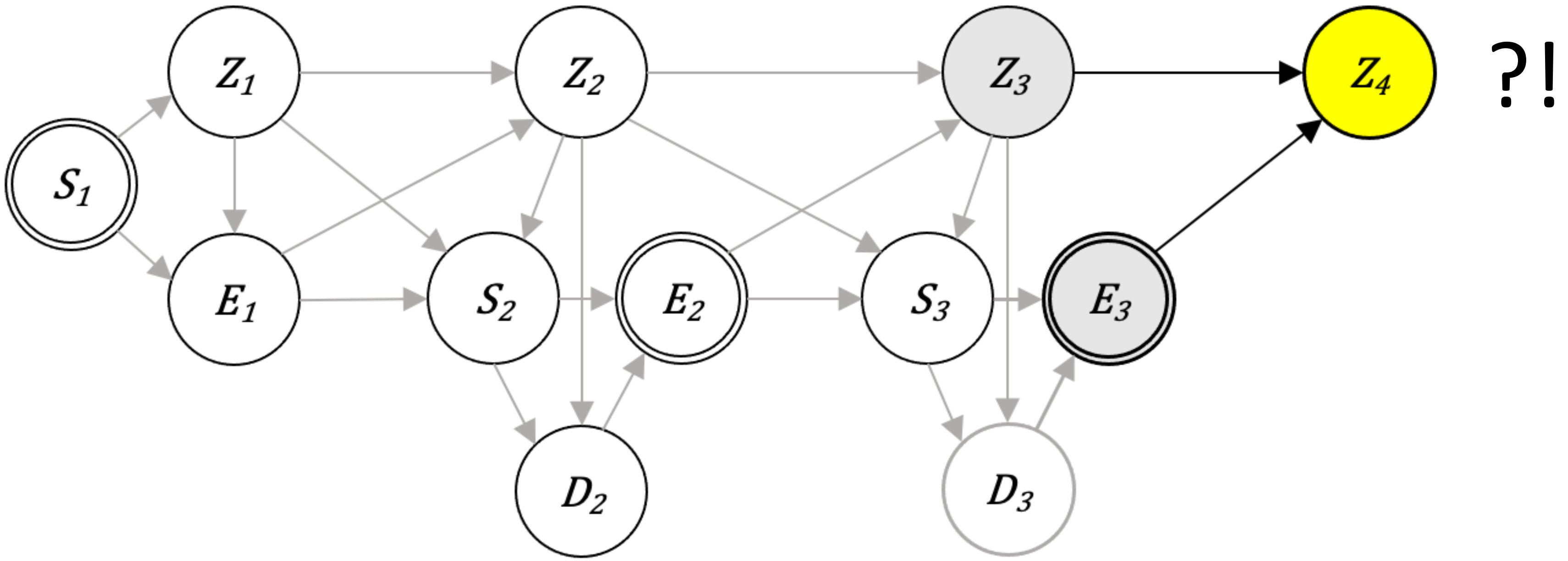
Z_3 = Maxwell
S_3 = 13:00
D_3 = 1 hour

BM



Z_1 = Tiong Bahru	Z_2 = Raffles Place	Z_3 = Maxwell
S_1 = 00:00	S_2 = 09:00	S_3 = 13:00
E_1 = 08:00	D_2 = 4 hours	D_3 = 1 hour
	E_2 = 13:00	**E_3 = 14:00**

BM



Z_1 = Tiong Bahru
S_1 = 00:00
E_1 = 08:00

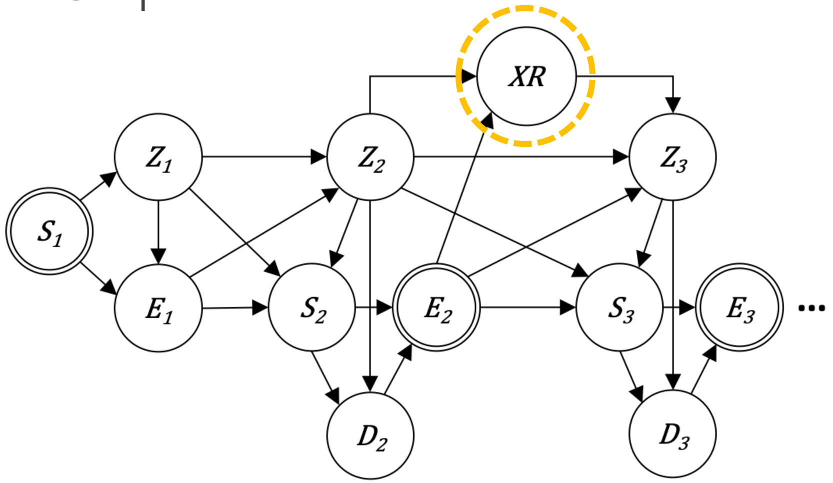
Z_2 = Raffles Place
S_2 = 09:00
D_2 = 4 hours
E_2 = 13:00

Z_3 = Maxwell
S_3 = 13:00
D_3 = 1 hour
E_3 = 14:00

Z_4 = ?

XR: Explore and Return model

1. Graphical model



3. Histograms required

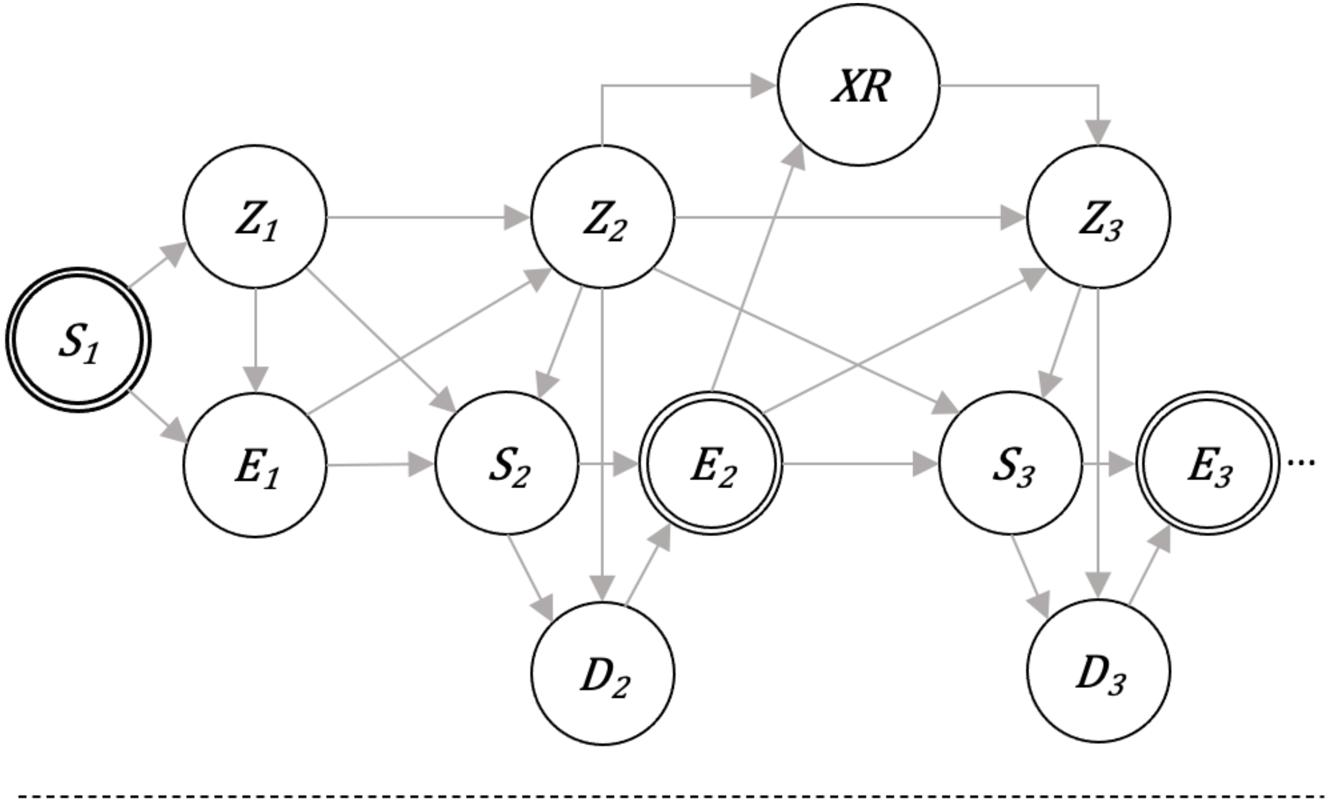
	CPD	Histogram
1	$P(Z_1 \mid S_1)$	Initial location
2	$P(E_1 \mid Z_1, S_1)$	Initial departure time given zone
3	$P(Z_k \mid Z_{k-1}, E_{k-1})$	Dynamic OD matrix
4	$P(S_k \mid Z_k, Z_{k-1}, E_{k-1})$	Start time given OD pair and origin end time
5	$P(D_k \mid Z_k, S_k)$	Duration given zone and start time
6	$P(XR_k \mid Z_{k-1}, E_{k-1})$	Explore or return

2. Joint Probability Distribution

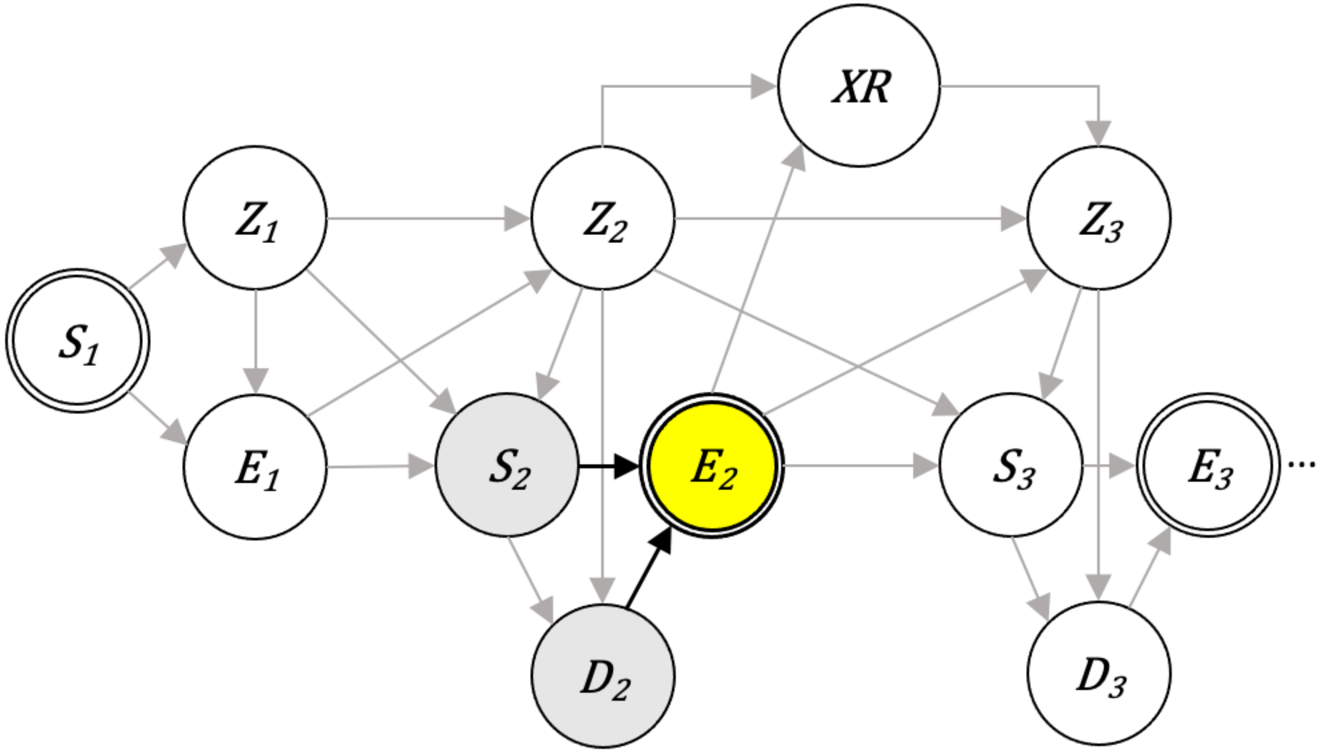
$$\boldsymbol{P(Z_{1:N}, S_{1:N}, E_1, D_{2:N}, XR_{3:N})}$$

$$= P(S_1)P(Z_1|S_1)P(E_1|S_1,Z_1)P(Z_2|Z_1,E_1)\prod_{k=2}^{N} P(S_k|Z_k,Z_{k-1},E_{k-1})P(D_k|Z_k,S_k)\boxed{\prod_{k=3}^{N} P(XR_k|Z_{k-1},E_{k-1})}P(Z_k|E_{k-1},Z_{k-1},XR_k)$$

XR



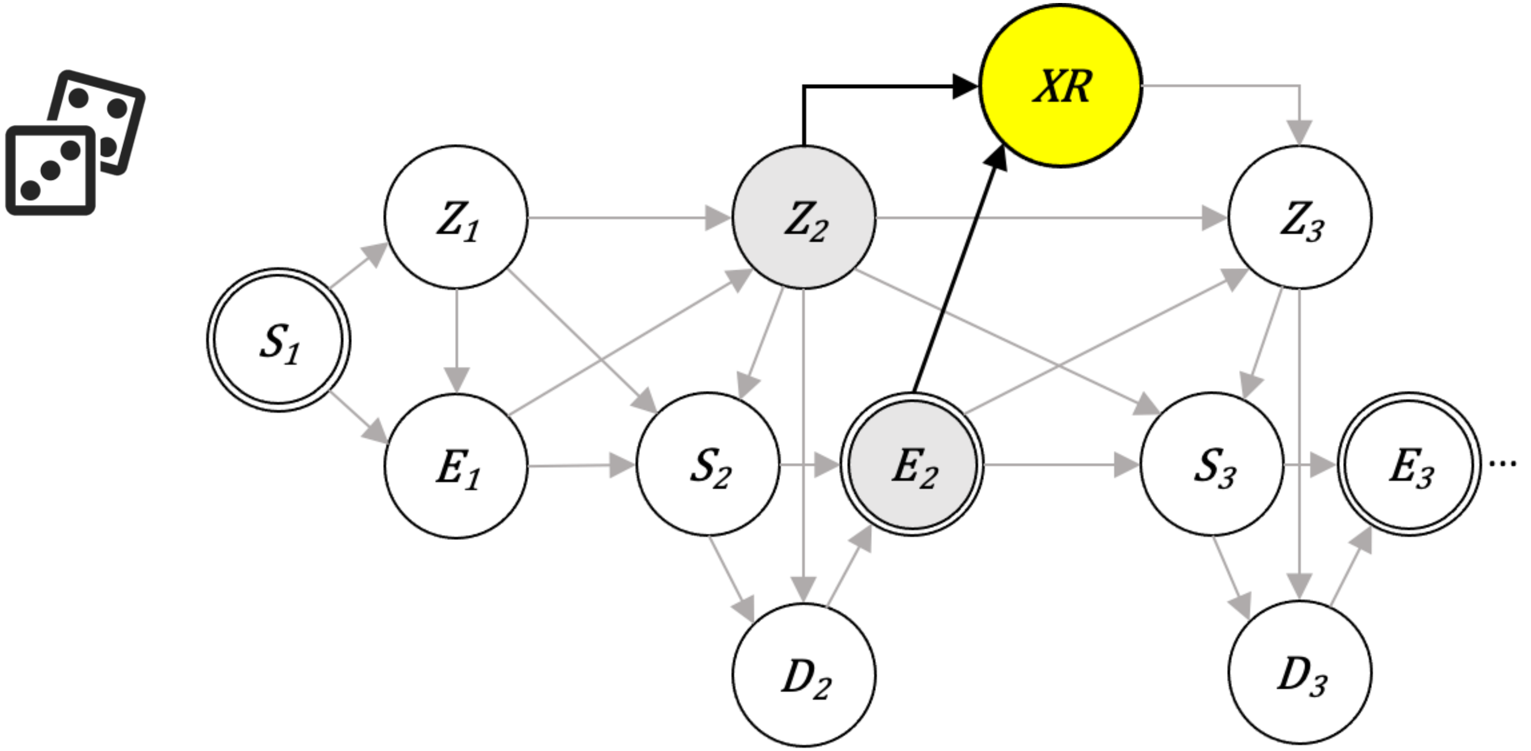
XR



Z_1 = Tiong Bahru
S_1 = 00:00
E_1 = 08:00

Z_2 = Raffles Place
S_2 = 09:00
D_2 = 4 hours
E_2 = 13:00

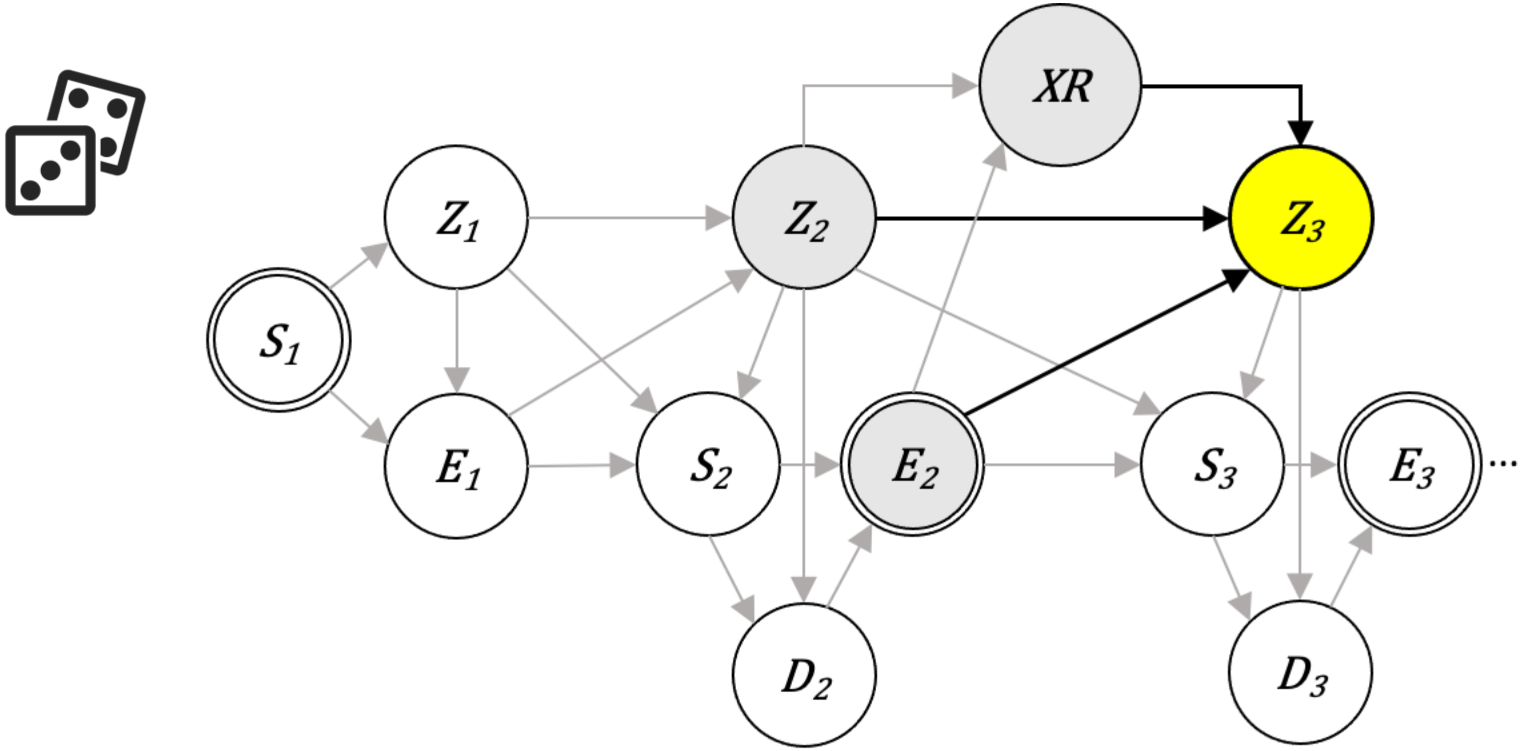
XR



Z_1 = Tiong Bahru
S_1 = 00:00
E_1 = 08:00

Z_2 = Raffles Place
S_2 = 09:00
D_2 = 4 hours
E_2 = 13:00
XR = Explore

XR

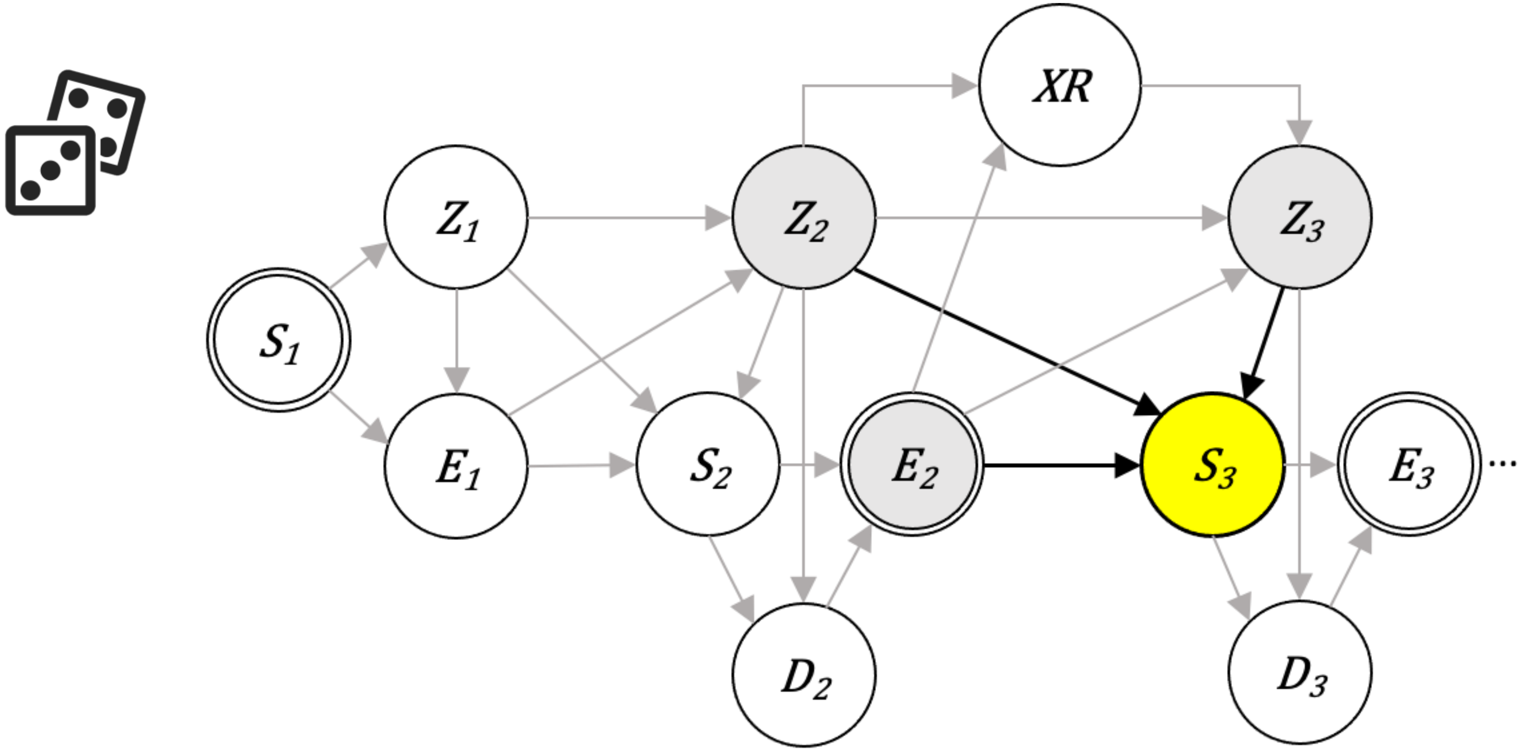


Z_1 = Tiong Bahru
S_1 = 00:00
E_1 = 08:00

Z_2 = Raffles Place
S_2 = 09:00
D_2 = 4 hours
E_2 = 13:00
XR = Explore

Z_3 = Maxwell

XR

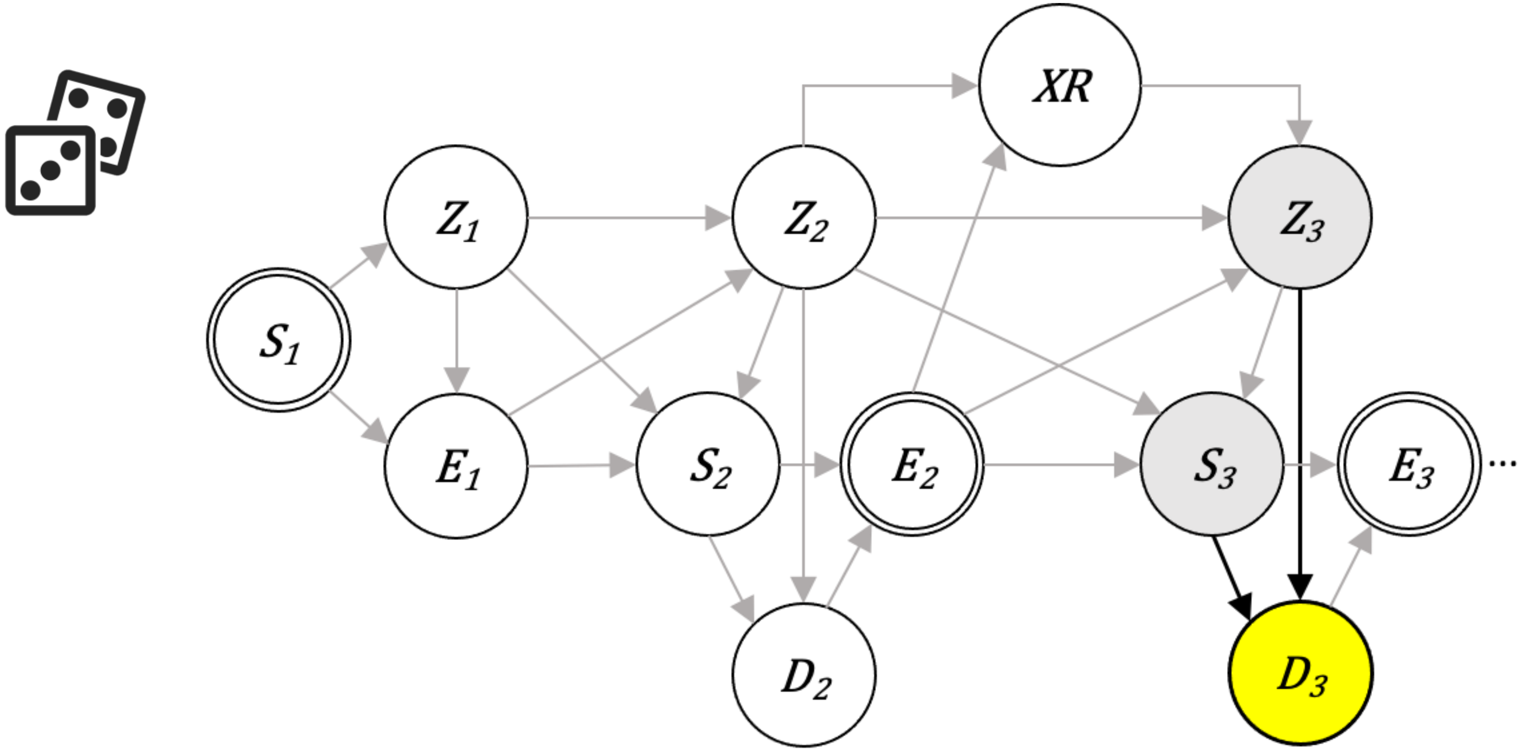


Z_1 = Tiong Bahru
S_1 = 00:00
E_1 = 08:00

Z_2 = Raffles Place
S_2 = 09:00
D_2 = 4 hours
E_2 = 13:00
XR = Explore

Z_3 = Maxwell
S_3 = 13:00

XR

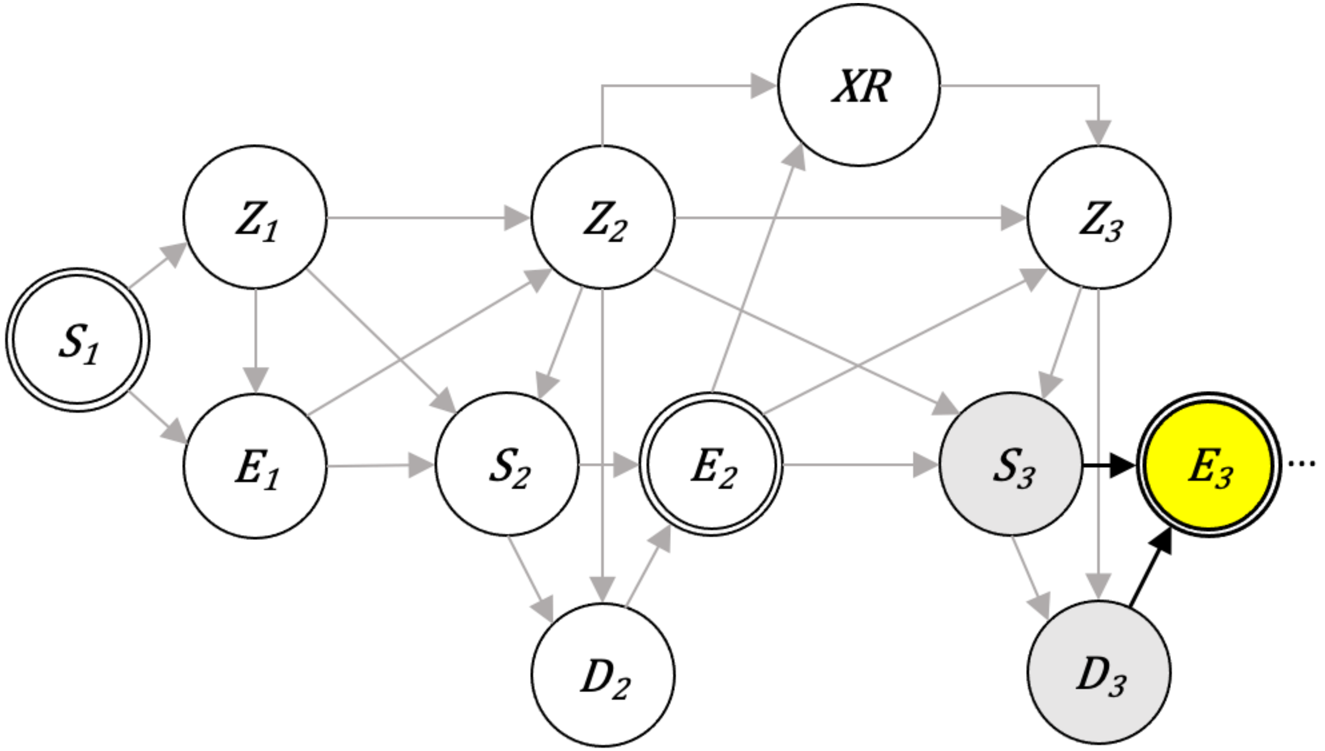


Z_1 = Tiong Bahru
S_1 = 00:00
E_1 = 08:00

Z_2 = Raffles Place
S_2 = 09:00
D_2 = 4 hours
E_2 = 13:00
XR = Explore

Z_3 = Maxwell
S_3 = 13:00
D_3 = 1 hour

XR

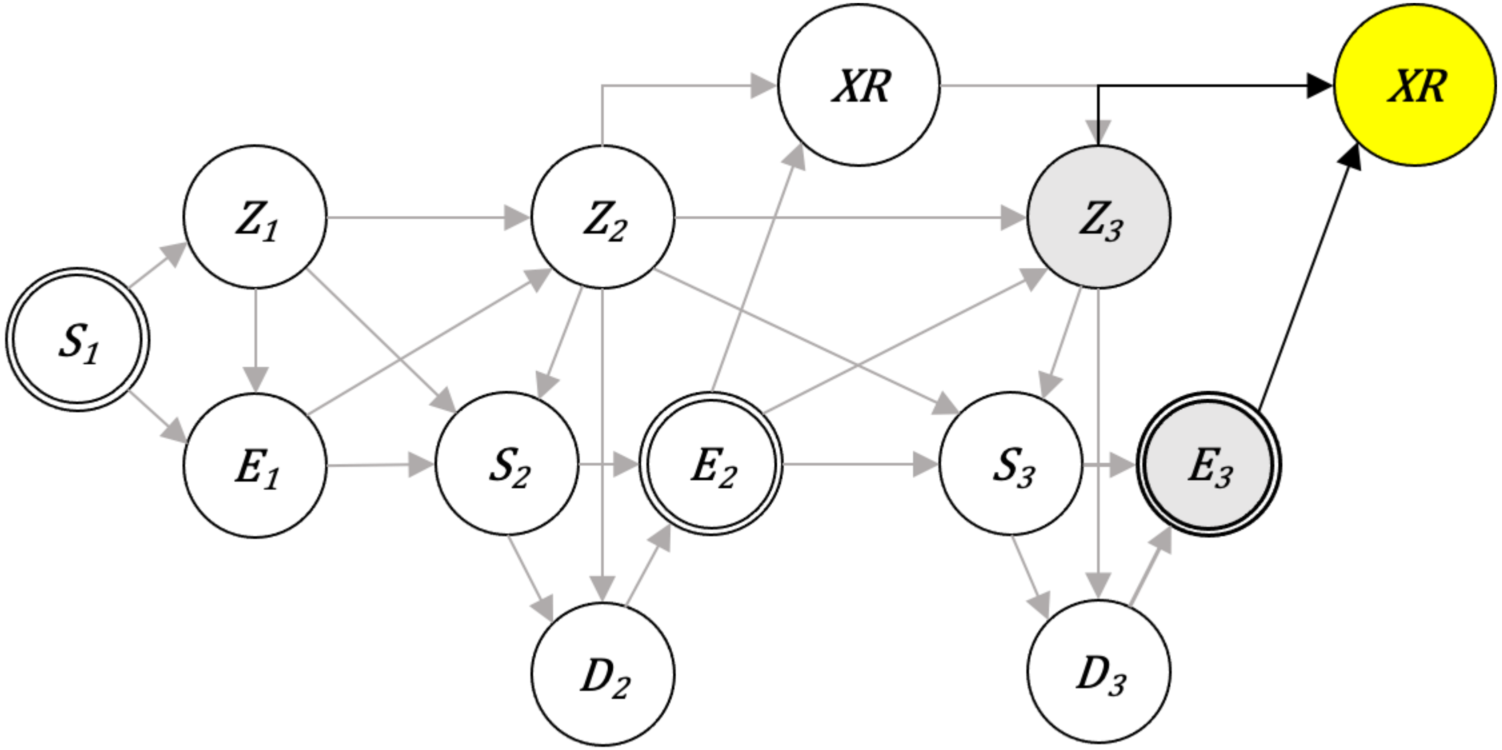


Z_1 = Tiong Bahru
S_1 = 00:00
E_1 = 08:00

Z_2 = Raffles Place
S_2 = 09:00
D_2 = 4 hours
E_2 = 13:00
XR = Explore

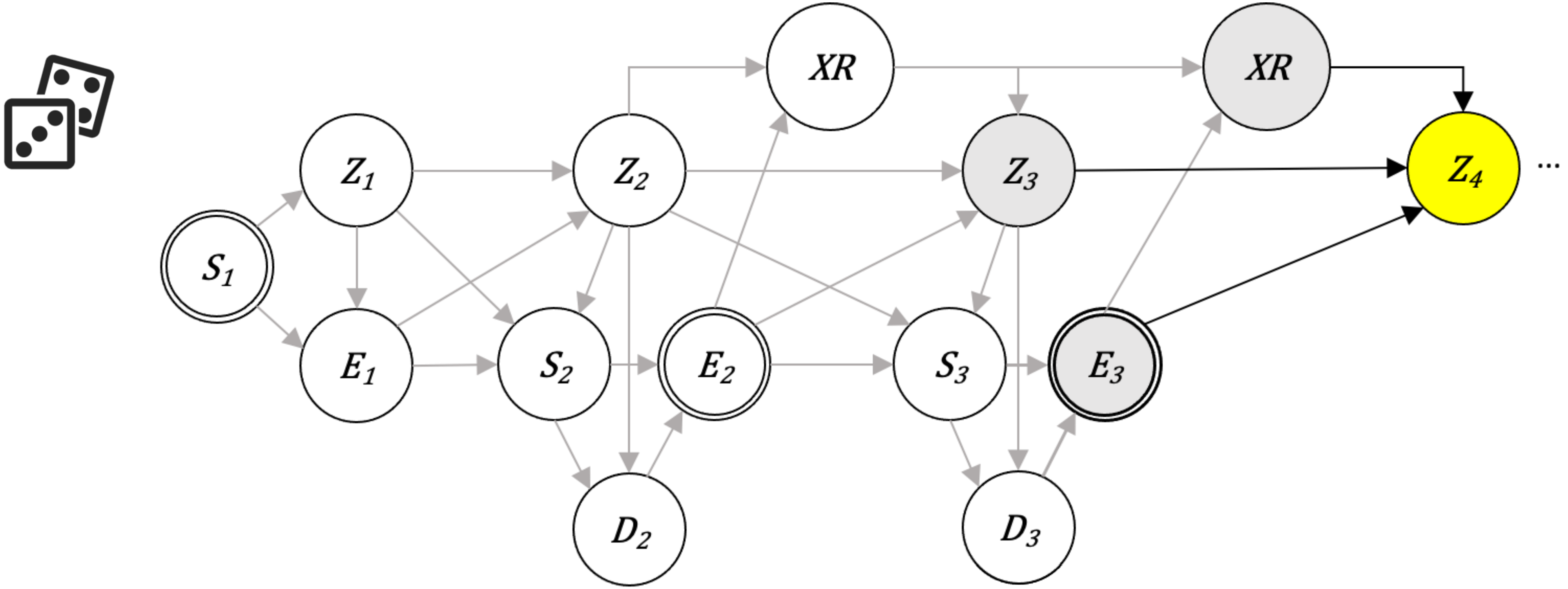
Z_3 = Maxwell
S_3 = 13:00
D_3 = 1 hour
E_3 = 14:00

XR



Z_1 = Tiong Bahru	Z_2 = Raffles Place	Z_3 = Maxwell
S_1 = 00:00	S_2 = 09:00	S_3 = 13:00
E_1 = 08:00	D_2 = 4 hours	D_3 = 1 hour
	E_2 = 13:00	E_3 = 14:00
	XR = Explore	**XR = Return**

XR



Z_1 = Tiong Bahru
S_1 = 00:00
E_1 = 08:00

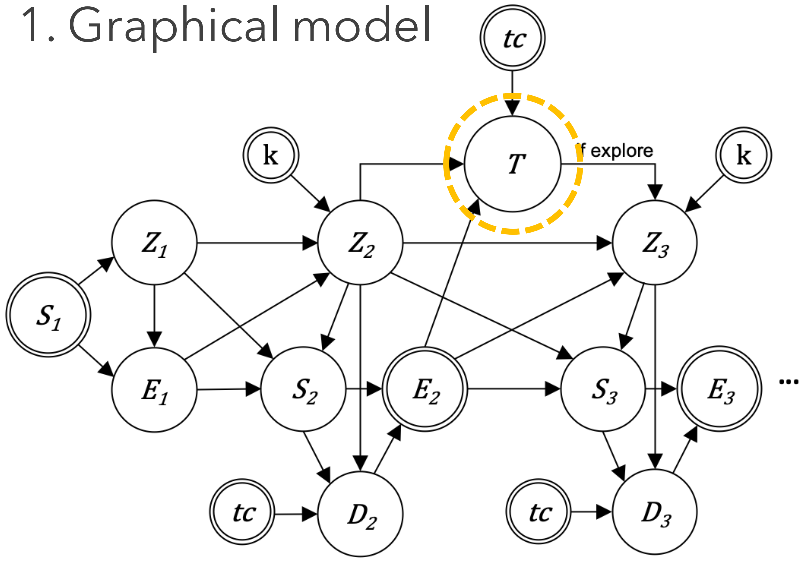
Z_2 = Raffles Place
S_2 = 09:00
D_2 = 4 hours
E_2 = 13:00
XR = Explore

Z_3 = Maxwell
S_3 = 13:00
D_3 = 1 hour
E_3 = 14:00
XR = Return

Z_4 = Raffles Place

TX: Tour Explicit model

1. Graphical model



3. Histograms required

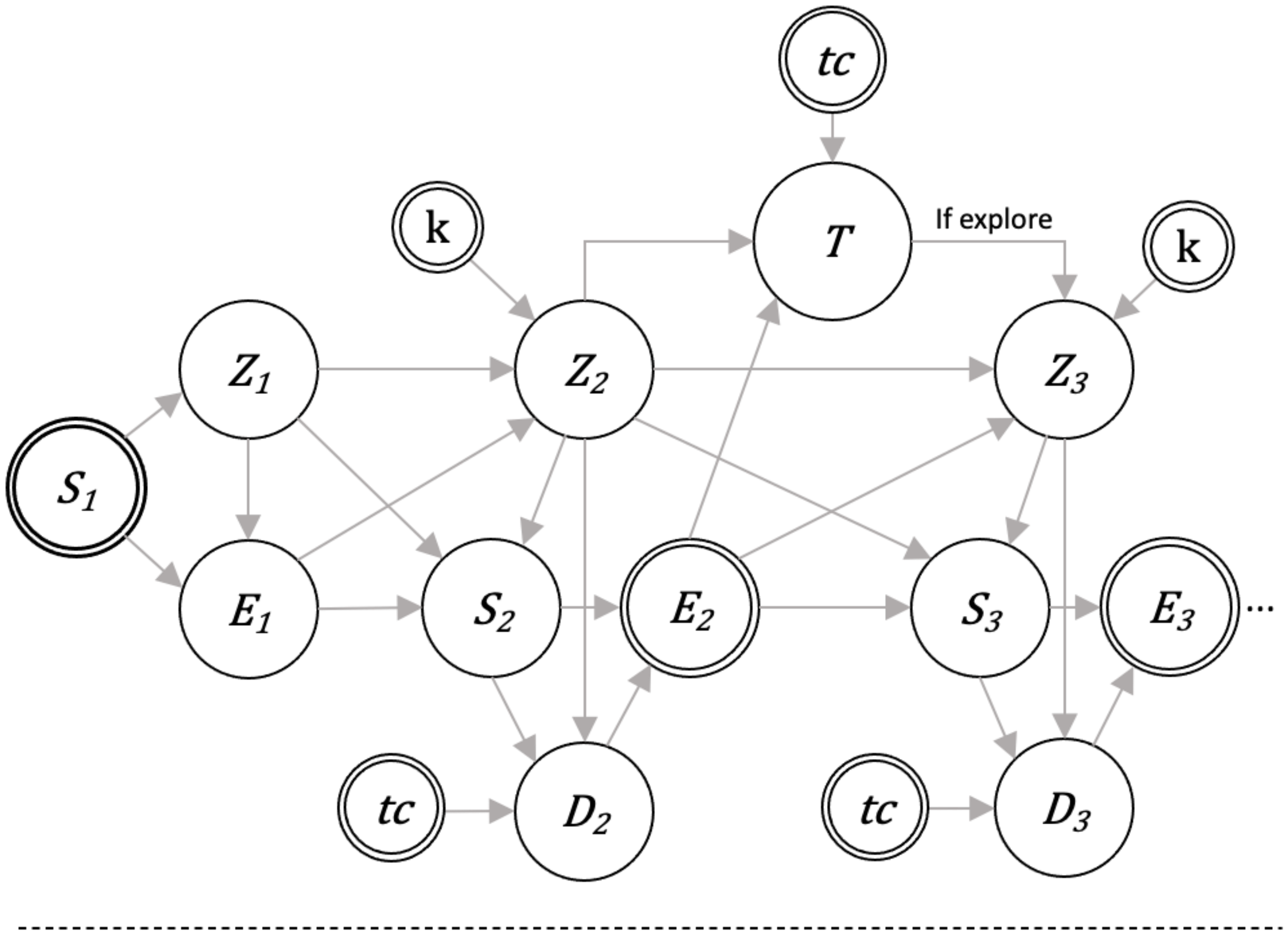
	CPD	Histogram
1	$P(Z_1 \mid S_1)$	Initial location
2	$P(E_1 \mid Z_1, S_1)$	Initial departure time given zone
3	$P(Z_k \mid Z_{k-1}, E_{k-1}, \mathrm{k})$	Dynamic OD matrix per number of staypoint
4	$P(S_k \mid Z_k, Z_{k-1}, E_{k-1})$	Start time given OD pair and origin end time
5	$P(D_k \mid Z_k, S_k, \mathrm{tc})$	Duration given zone, start time, tour code
6	$P(T_k \mid Z_{k-1}, E_{k-1}, tc)$	Tour: next digit given tour code, zone, time

2. Joint Probability Distribution

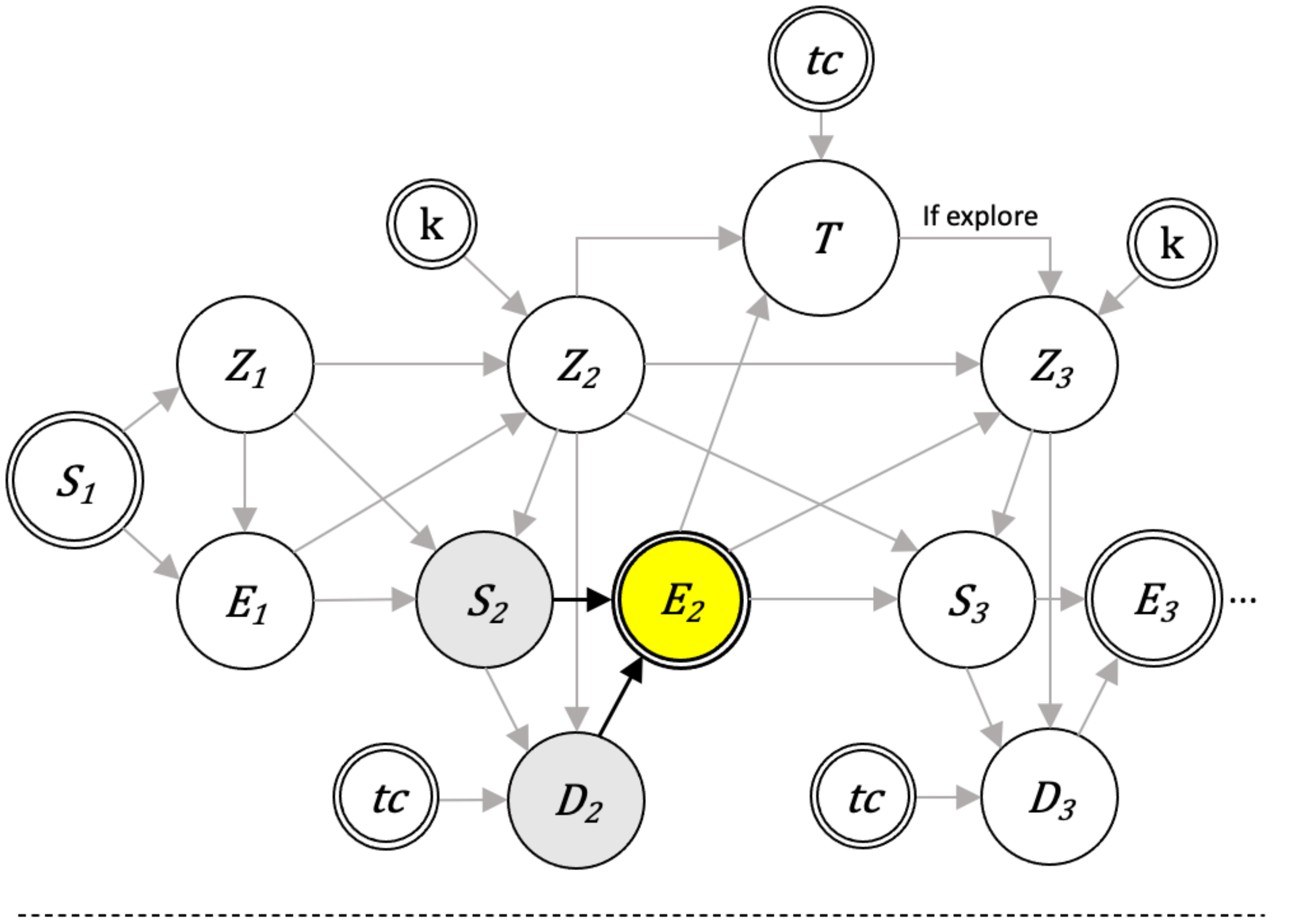
$$\boldsymbol{P(Z_{1:N}, S_{1:N}, E_1, D_{2:N}, T_{3:N})}$$

$$= P(S_1)P(Z_1 \mid S_1)P(E_1 \mid S_1, Z_1)P(Z_2 \mid Z_1, E_1, k) \prod_{k=2}^{N} P(S_k \mid Z_k, Z_{k-1}, E_{k-1}) P(D_k \mid Z_k, S_k, tc) \boxed{\prod_{k=3}^{N} P(T_k \mid Z_{k-1}, E_{k-1}, tc)} P(Z_k \mid E_{k-1}, Z_{k-1}, T_k, k)$$

TX



TX



Z_1 = Tiong Bahru
S_1 = 00:00
E_1 = 08:00

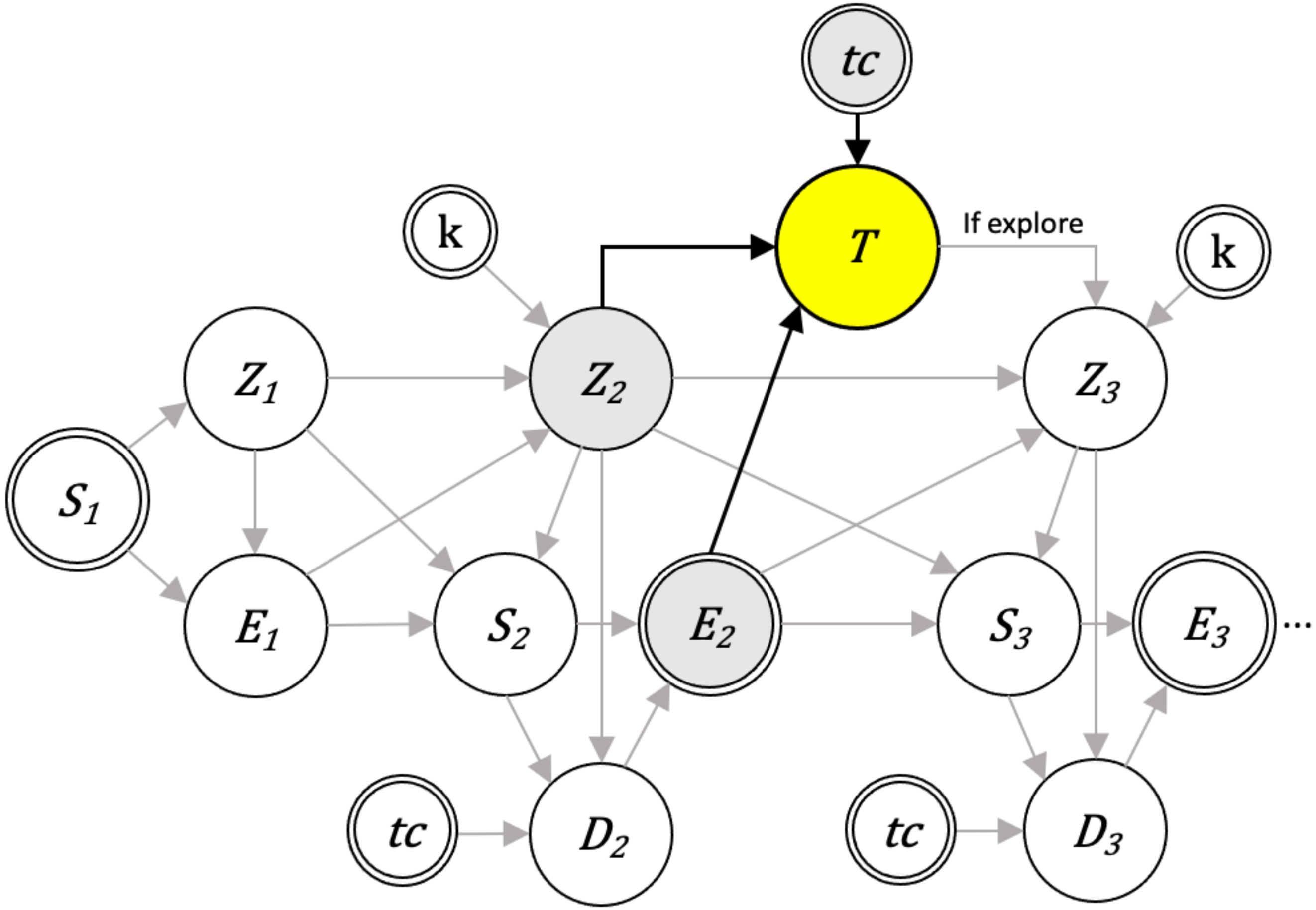
Z_2 = Raffles Place
S_2 = 09:00
D_2 = 4 hours
E_2 = 13:00

Tour network: **01**

Options:

- **0** Tiong Bahru
- **1** Raffles Place
- **2** Any other

TX



Z_1 = Tiong Bahru

S_1 = 00:00

E_1 = 08:00

Z_2 = Raffles Place

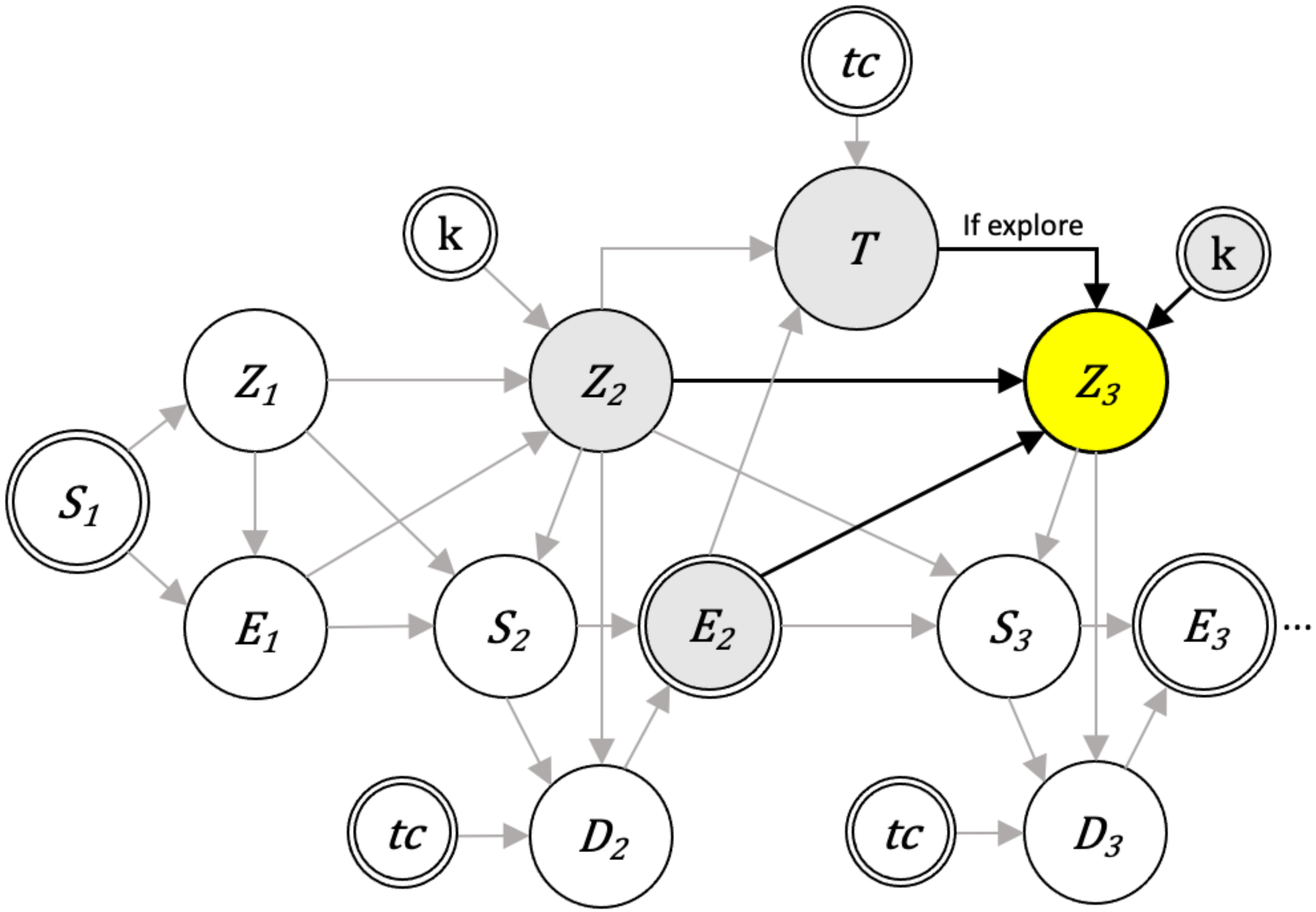
S_2 = 09:00

D_2 = 4 hours

E_2 = 13:00

T = 2 | tc = "01"

TX

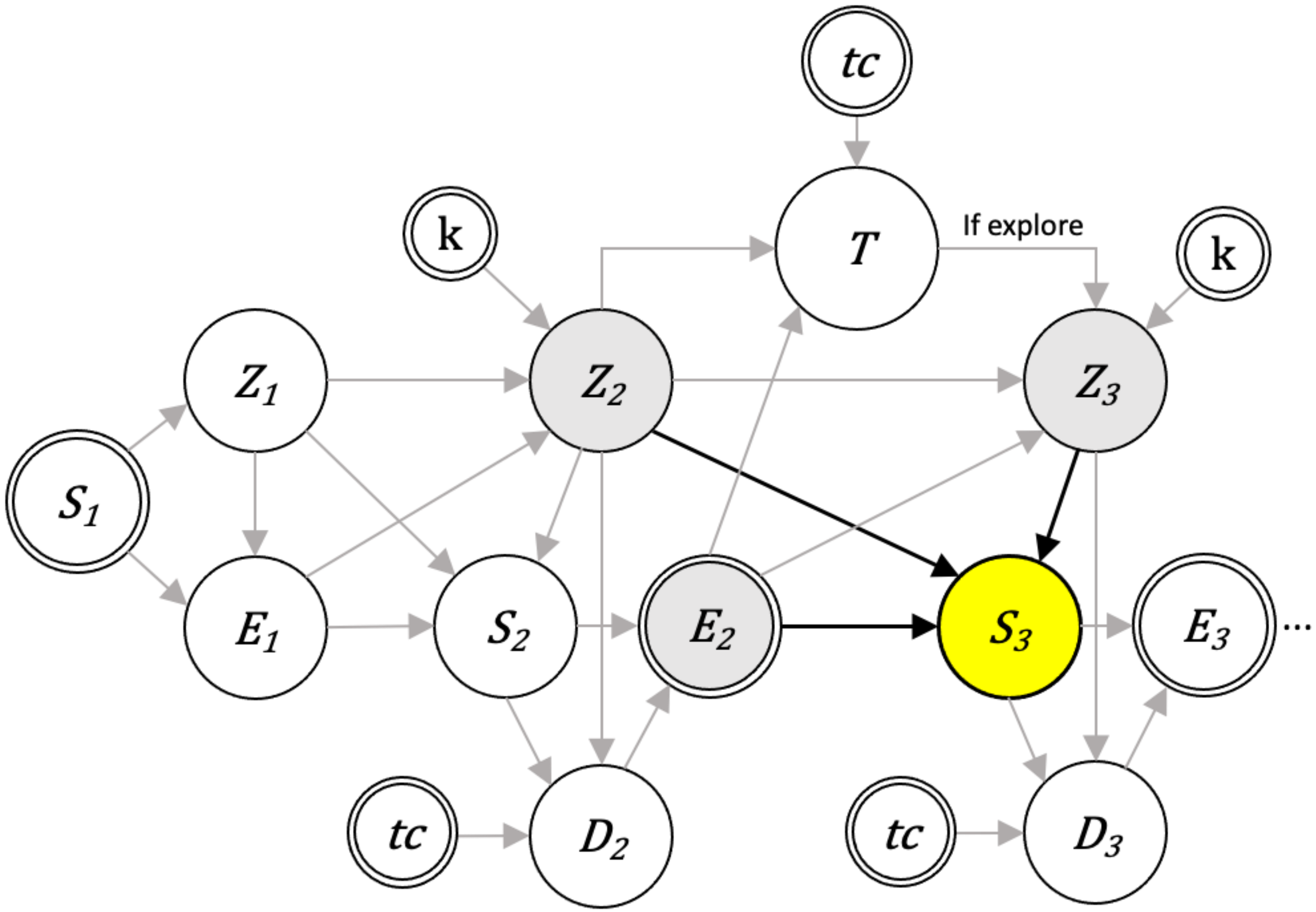


Z_1 = Tiong Bahru
S_1 = 00:00
E_1 = 08:00

Z_2 = Raffles Place
S_2 = 09:00
D_2 = 4 hours
E_2 = 13:00
T = 2 | tc = "01"

Z_3 = Maxwell

TX

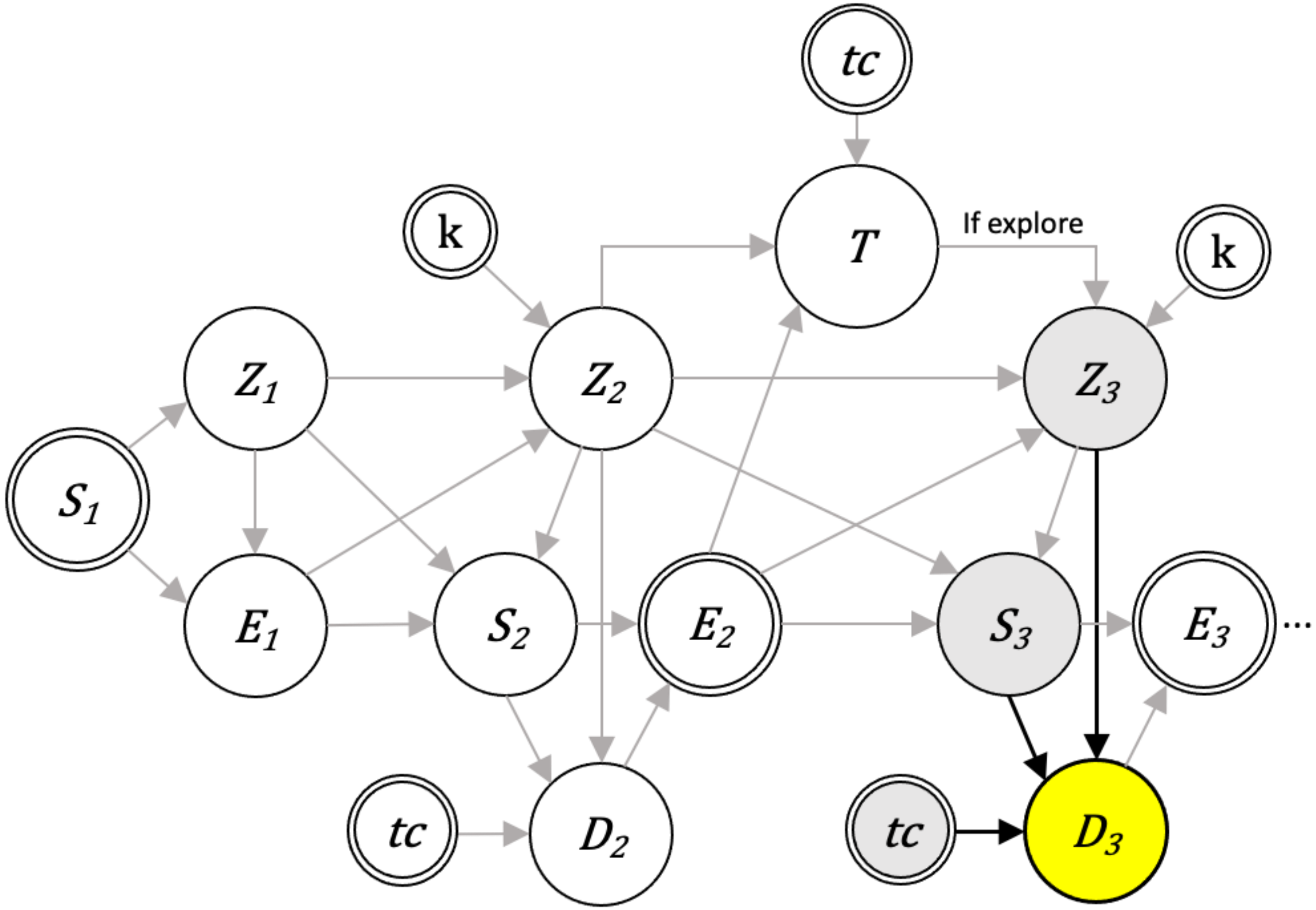


Z_1 = Tiong Bahru
S_1 = 00:00
E_1 = 08:00

Z_2 = Raffles Place
S_2 = 09:00
D_2 = 4 hours
E_2 = 13:00
T = 2 | tc = "01"

Z_3 = Maxwell
S_3 = 13:00

TX

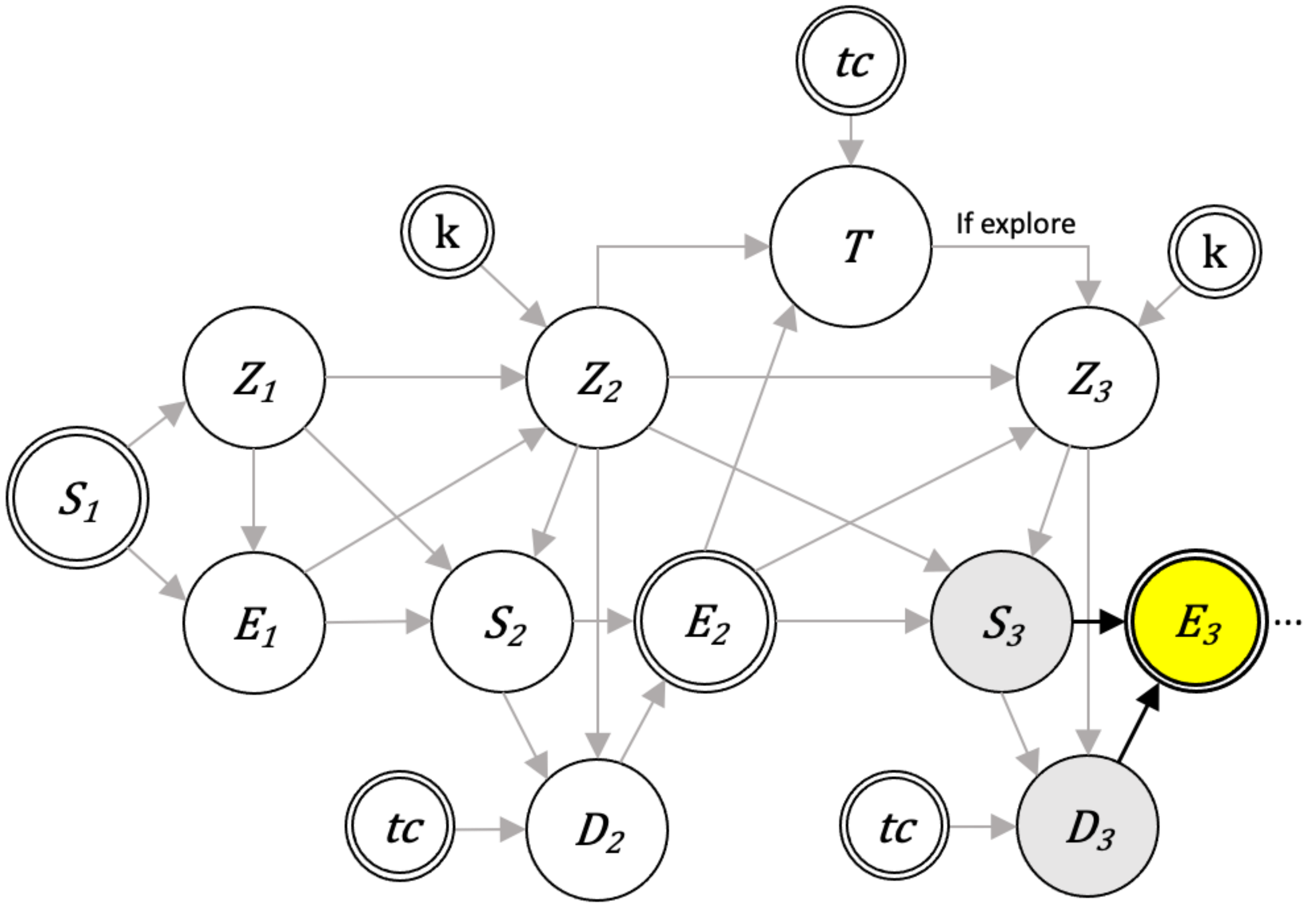


Z_1 = Tiong Bahru
S_1 = 00:00
E_1 = 08:00

Z_2 = Raffles Place
S_2 = 09:00
D_2 = 4 hours
E_2 = 13:00
T = 2 | tc = “01”

Z_3 = Maxwell
S_3 = 13:00
D_3 = 1 hours

TX



Z_1 = Tiong Bahru
S_1 = 00:00
E_1 = 08:00

Z_2 = Raffles Place
S_2 = 09:00
D_2 = 4 hours
E_2 = 13:00
T = 2 | tc = "01"

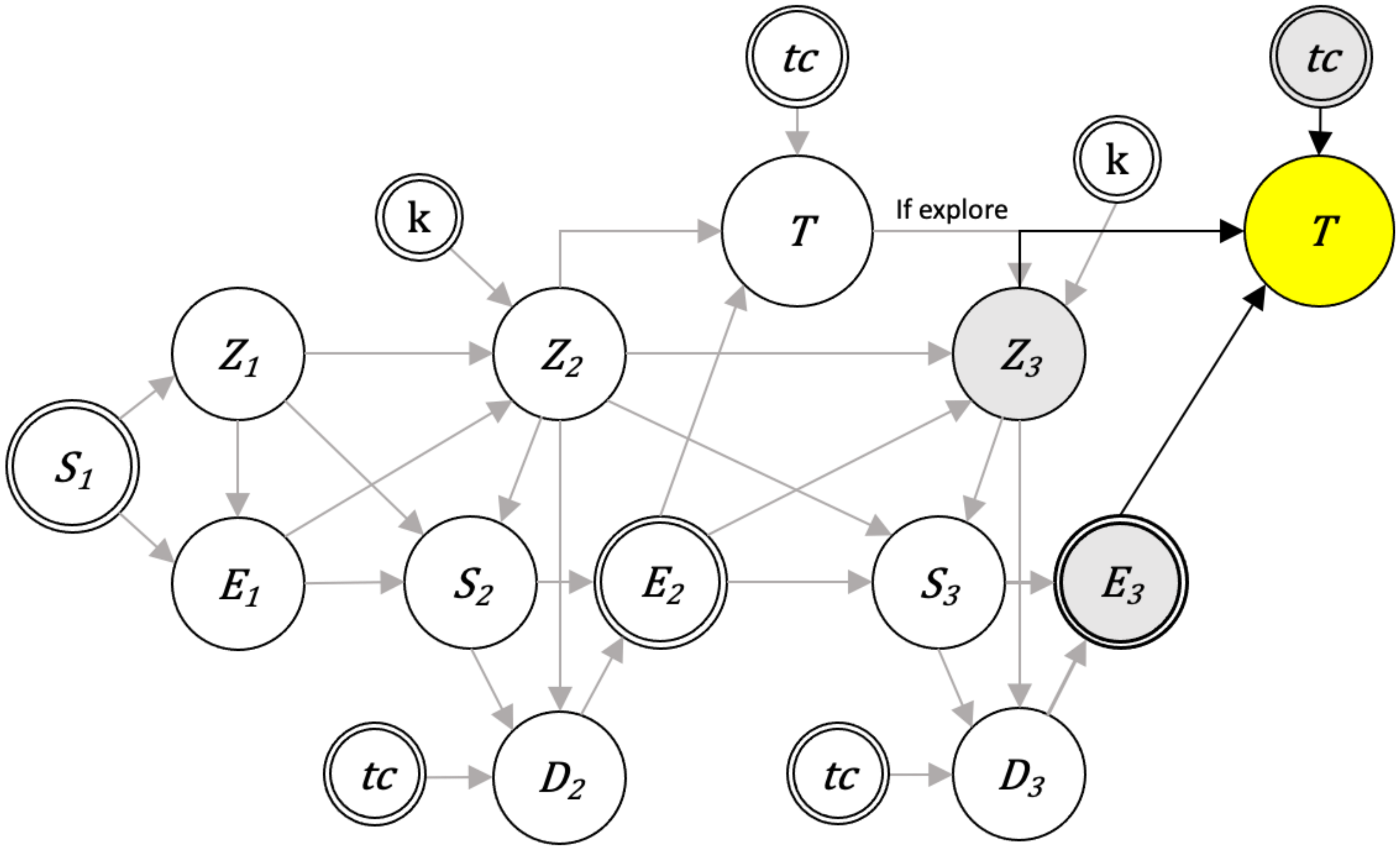
Z_3 = Maxwell
S_3 = 13:00
D_3 = 1 hours
E_3 = 14:00

Tour network: **012**

Options:

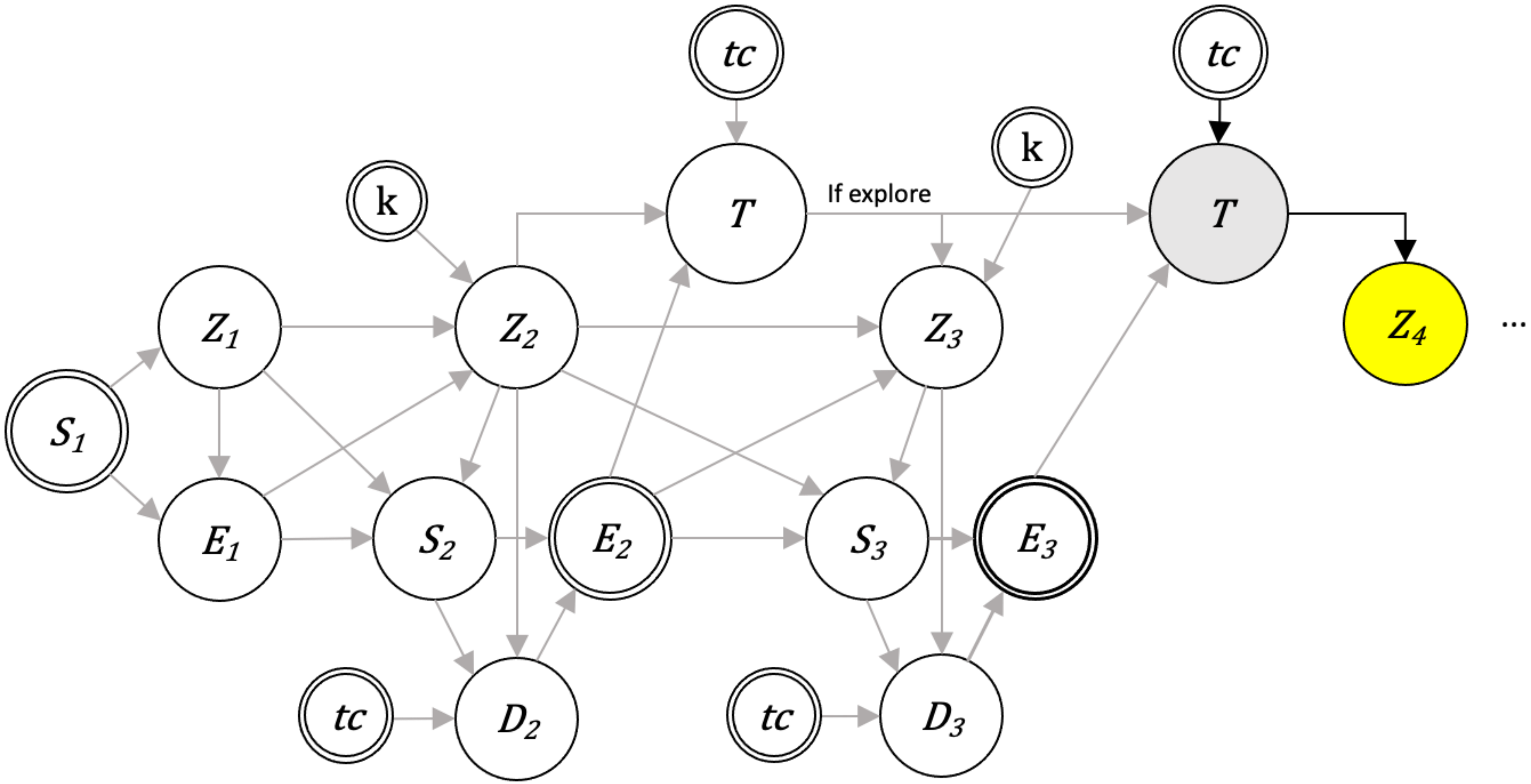
- **0** Tiong Bahru
- **1** Raffles Place
- **2** Maxwell
- **3** Any other

TX



Z_1 = Tiong Bahru	Z_2 = Raffles Place	Z_3 = Maxwell
S_1 = 00:00	S_2 = 09:00	S_3 = 13:00
E_1 = 08:00	D_2 = 4 hours	D_3 = 1 hours
	E_2 = 13:00	E_3 = 14:00
	T = 2 | tc = "01"	**T = 1 | tc = "012"**

TX



Z_1 = Tiong Bahru
S_1 = 00:00
E_1 = 08:00

Z_2 = Raffles Place
S_2 = 09:00
D_2 = 4 hours
E_2 = 13:00
T = 2 | tc = “01”

Z_3 = Maxwell
S_3 = 13:00
D_3 = 1 hours
E_3 = 14:00
T = 1 | tc = “012”

Z_4 = Raffles Place

Digital Twin Travellers

Finding traveller archetypes from mobile phone data

Finding traveller archetypes from mobile phone data

1. Train a generative model per type of traveller

Traveller type X_1

Traveller type X_2

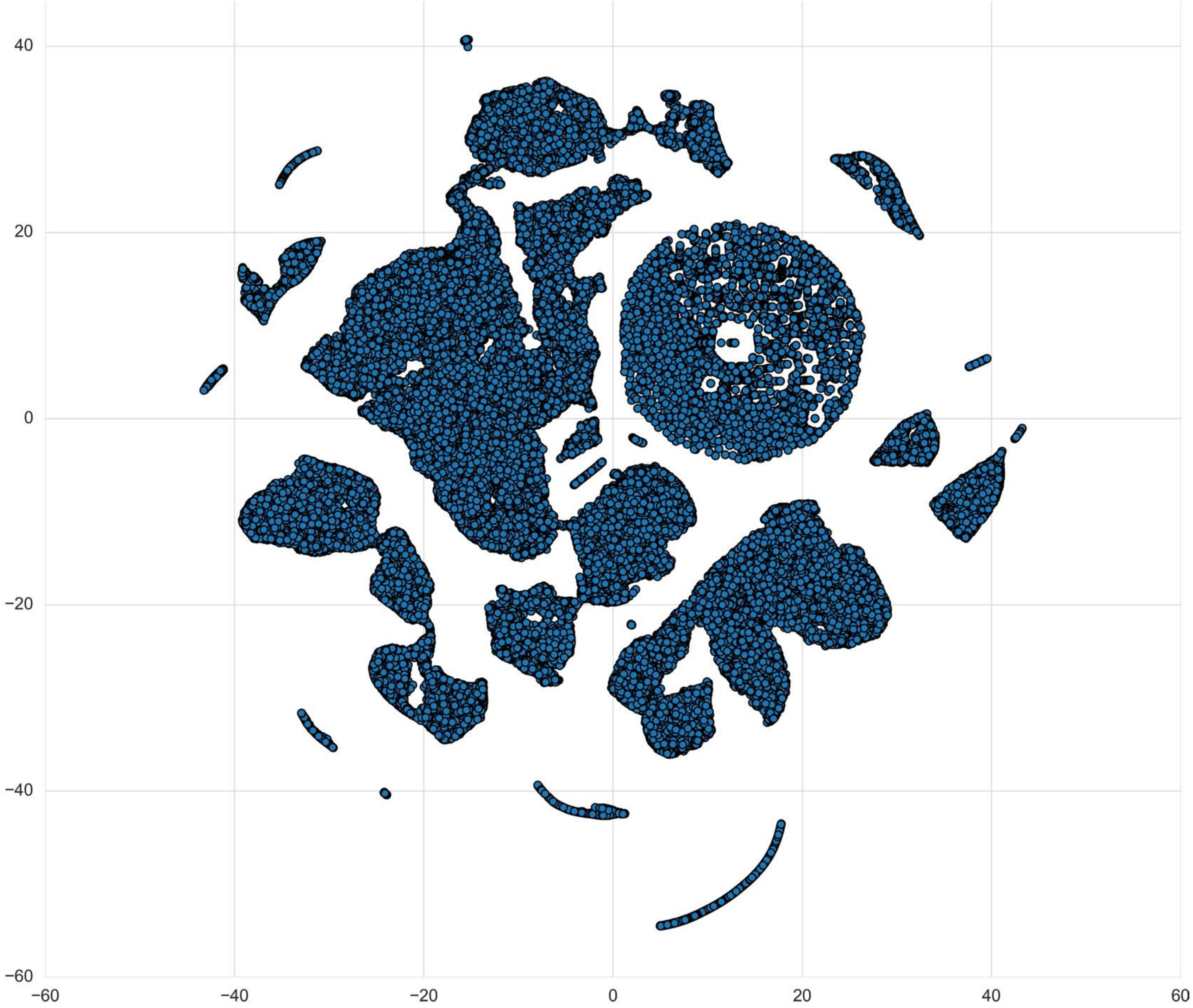
Traveller type X_n

- ❌ socio-demographics
- ❌ activity/trip information
- ✅ one-day mobile data locations and timings

2. Feature Engineering

1. Activity durations mean
2. Activity durations std
3. Bias morning / night
4. First departure
5. Last arrival

T-SNE representation

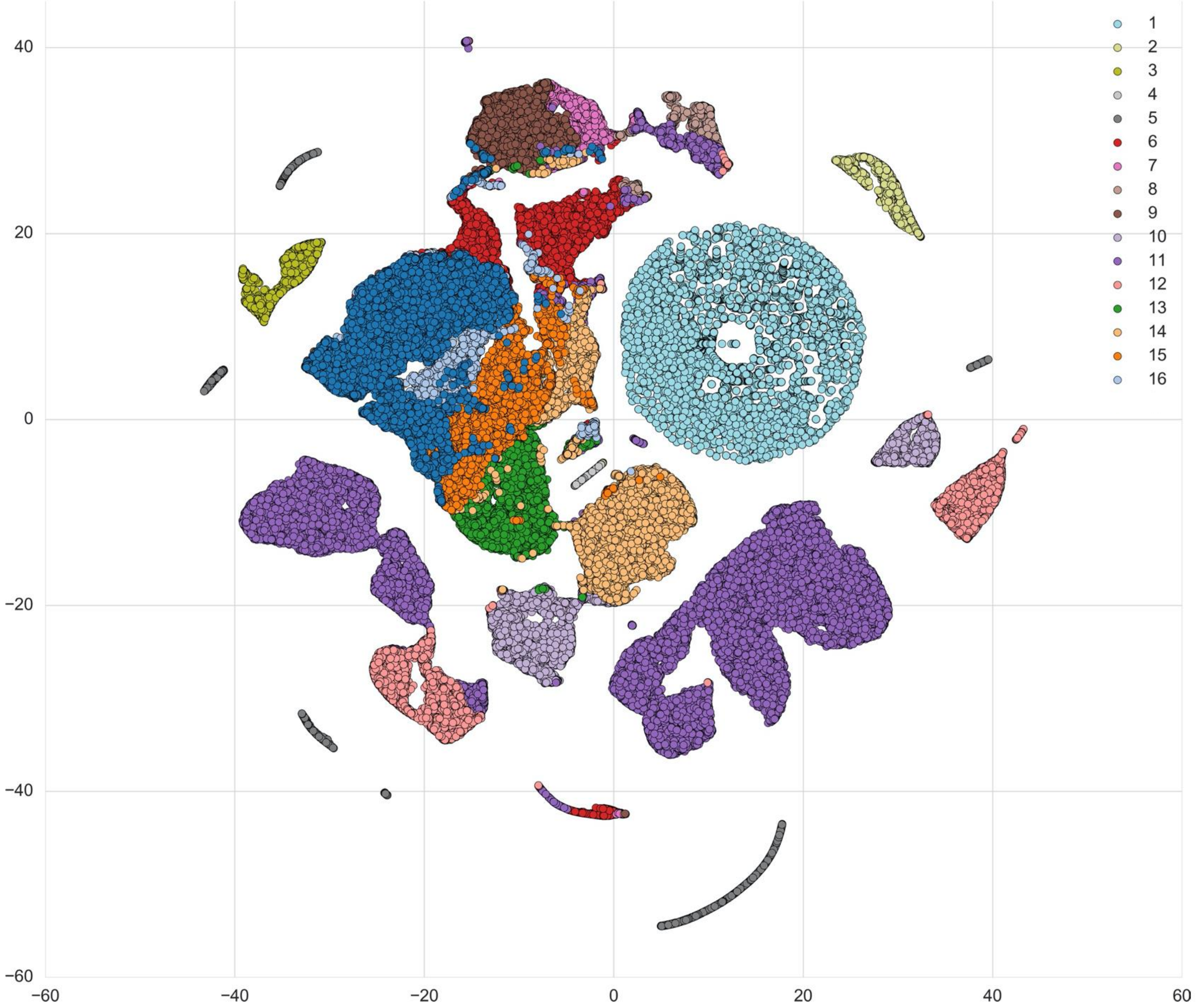


DBSCAN results

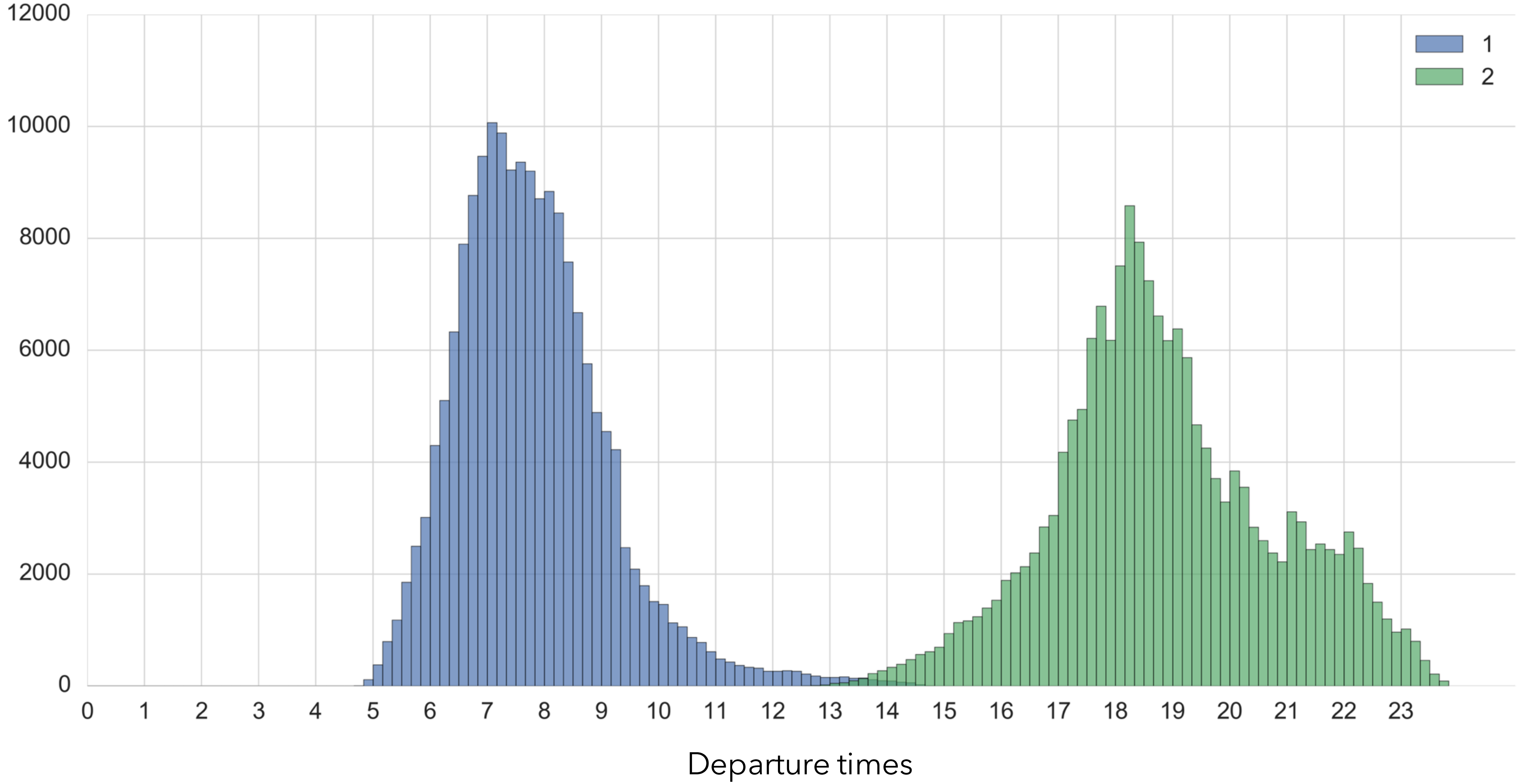
16 types of travellers

in Singapore

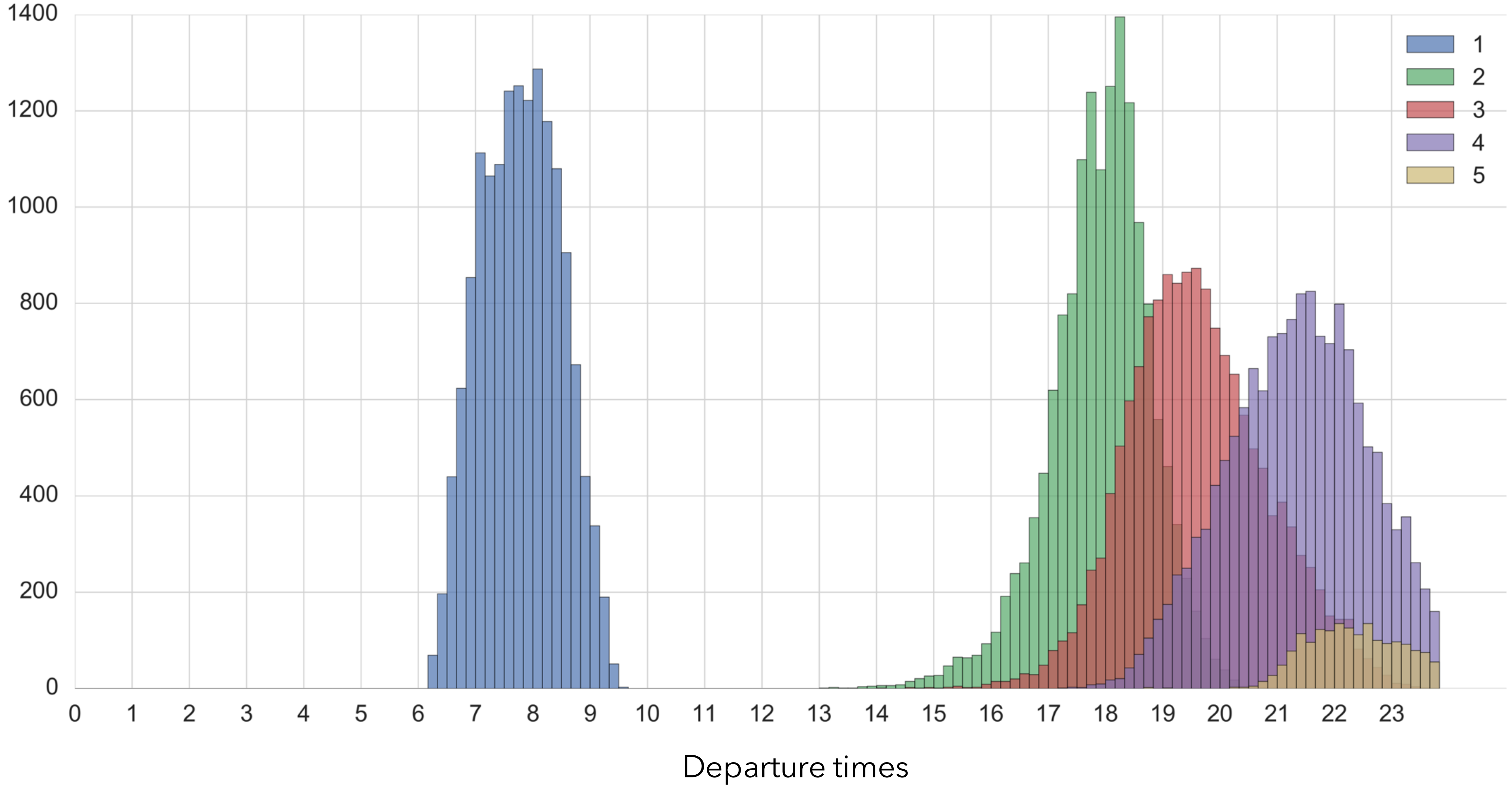
in the given day



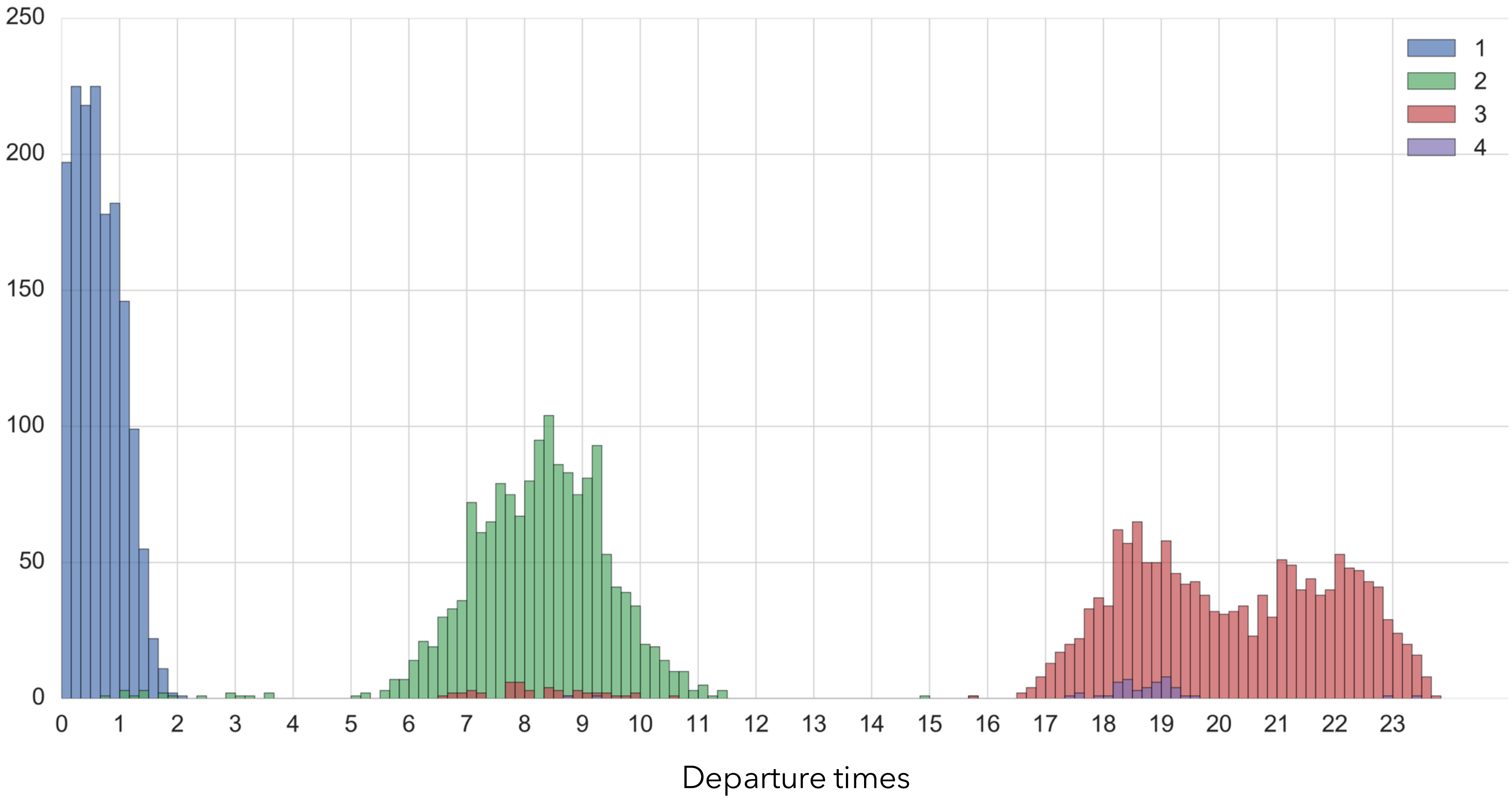
Cluster 11: The Commuter



Cluster 13: The Commuter with multiple evening activities



Cluster 7: The Commuter with a late-night activity



Agenda

1. Mobile network signalling data
2. Motivation
3. Intuitive example
4. Digital Twin Travellers
5. Experiment and results
6. Application
7. Conclusion

Experimental setting

Generate a synthetic mobility population with …

BM	Baseline Markov
XR	Explore and Return
XR-C	Explore and Return with clusters
TX	Tour Explicit
TX-C	Tour Explicit with clusters

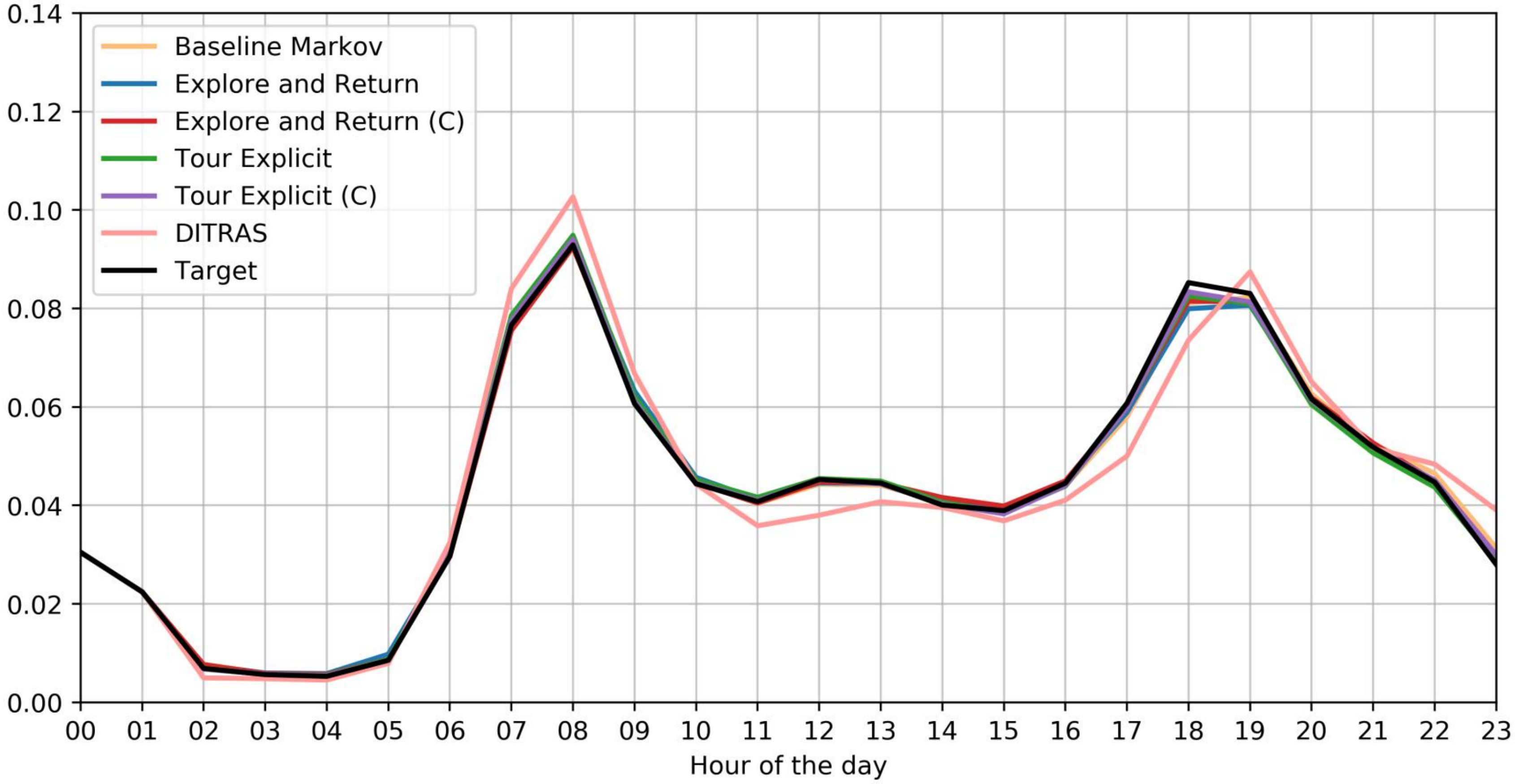
DITRAS External benchmark

Pappalardo, L., & Simini, F. (2018). Data-driven generation of spatio-temporal routines in human mobility. *Data Mining and Knowledge Discovery*, *32*(3), 787-829.

… and compare with …

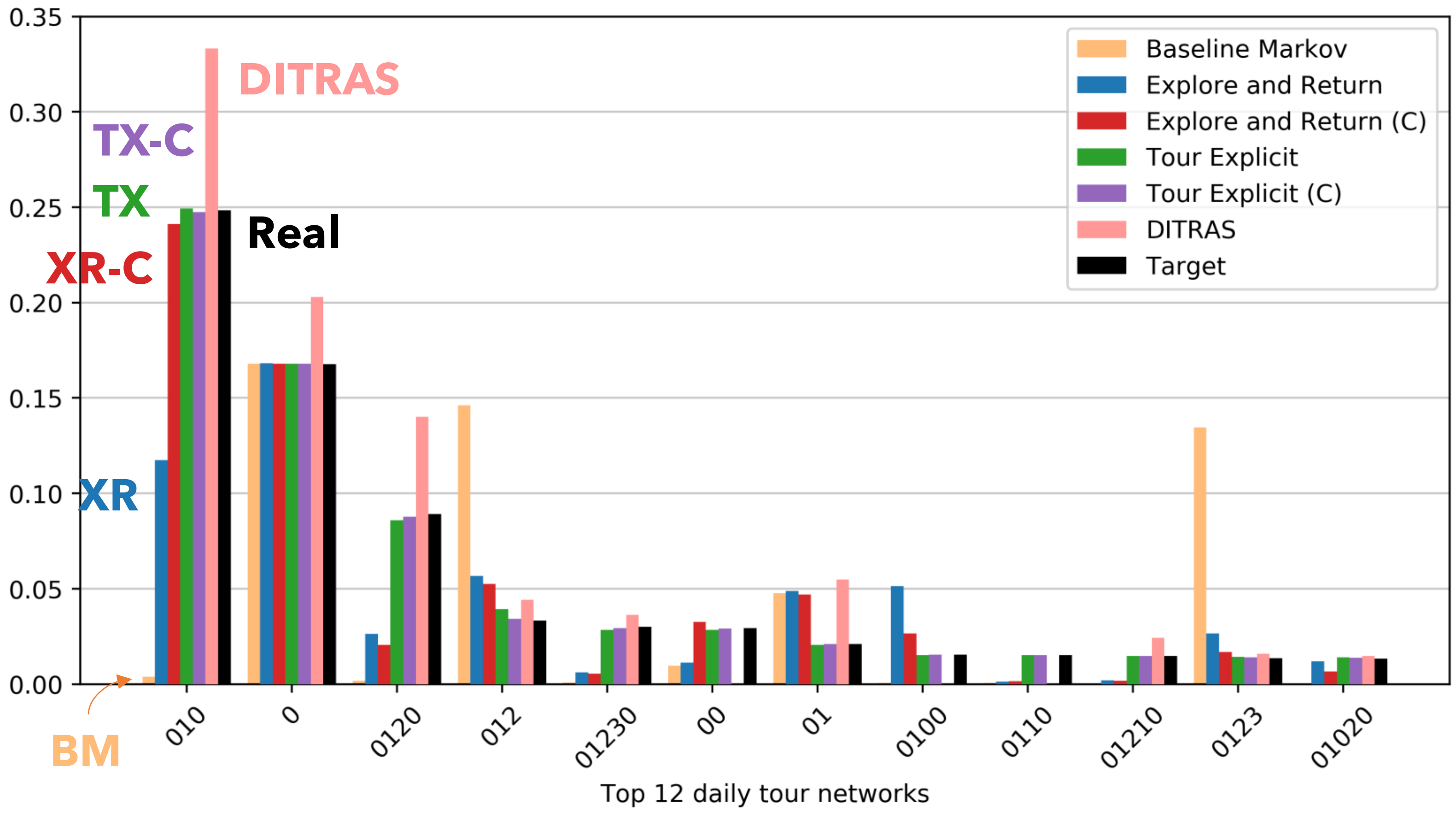
Real Mobile users' staypoint data

Results: Activity start time



Model	Error (%)
BM	1.20
XR	1.11
XR-C	0.92
TX	0.81
TX-C	**0.59**
DITRAS	4.98

Results: Tour networks



Model	Error (%)
BM	65.32
XR	35.87
XR-C	21.98
TX	1.36
TX-C	**0.32**
DITRAS	13.59

Results: All

One-day Mobility Population Accuracy Score

10 distributions validated

1. Activity start time
2. Activity duration
3. Number of trips
4. Tour network
5. Dynamic spatial error
6. Total distance travelled
7. Radius of gyration
8. Mobility entropy
9. Activity space
10. Semantic similarity *(**new**)*

Model	OMPAS (%)
BM	79.19
XR	89.79
XR-C	93.32
TX	97.65
TX-C	97.97
DITRAS	83.05

Conclusion 1:

XR and TX architectures mitigate the first-order Markov constraint

Results: All

10 distributions validated

1. Activity start time
2. Activity duration
3. Number of trips
4. Tour network
5. Dynamic spatial error
6. Total distance travelled
7. Radius of gyration
8. Mobility entropy
9. Activity space
10. Semantic similarity (***new***)

One-day **M**obility **P**opulation **A**ccuracy **S**core

Model	OMPAS (%)
BM	79.19
XR	89.79
XR-C	93.32
TX	97.65
TX-C	97.97
DITRAS	83.05

Conclusion 2:

Clustering improves the accuracy score

Results: All

One-day **M**obility **P**opulation **A**ccuracy **S**core

10 distributions validated

1. Activity start time
2. Activity duration
3. Number of trips
4. Tour network
5. Dynamic spatial error
6. Total distance travelled
7. Radius of gyration
8. Mobility entropy
9. Activity space
10. Semantic similarity *(**new**)*

complexity

Model	OMPAS (%)
BM	79.19
XR	89.79
XR-C	93.32
TX	97.65
TX-C	97.97
DITRAS	83.05

Conclusion 3:

If the model designs are reasonable, larger complexity, higher accuracy score

Results: All

One-day Mobility Population Accuracy Score

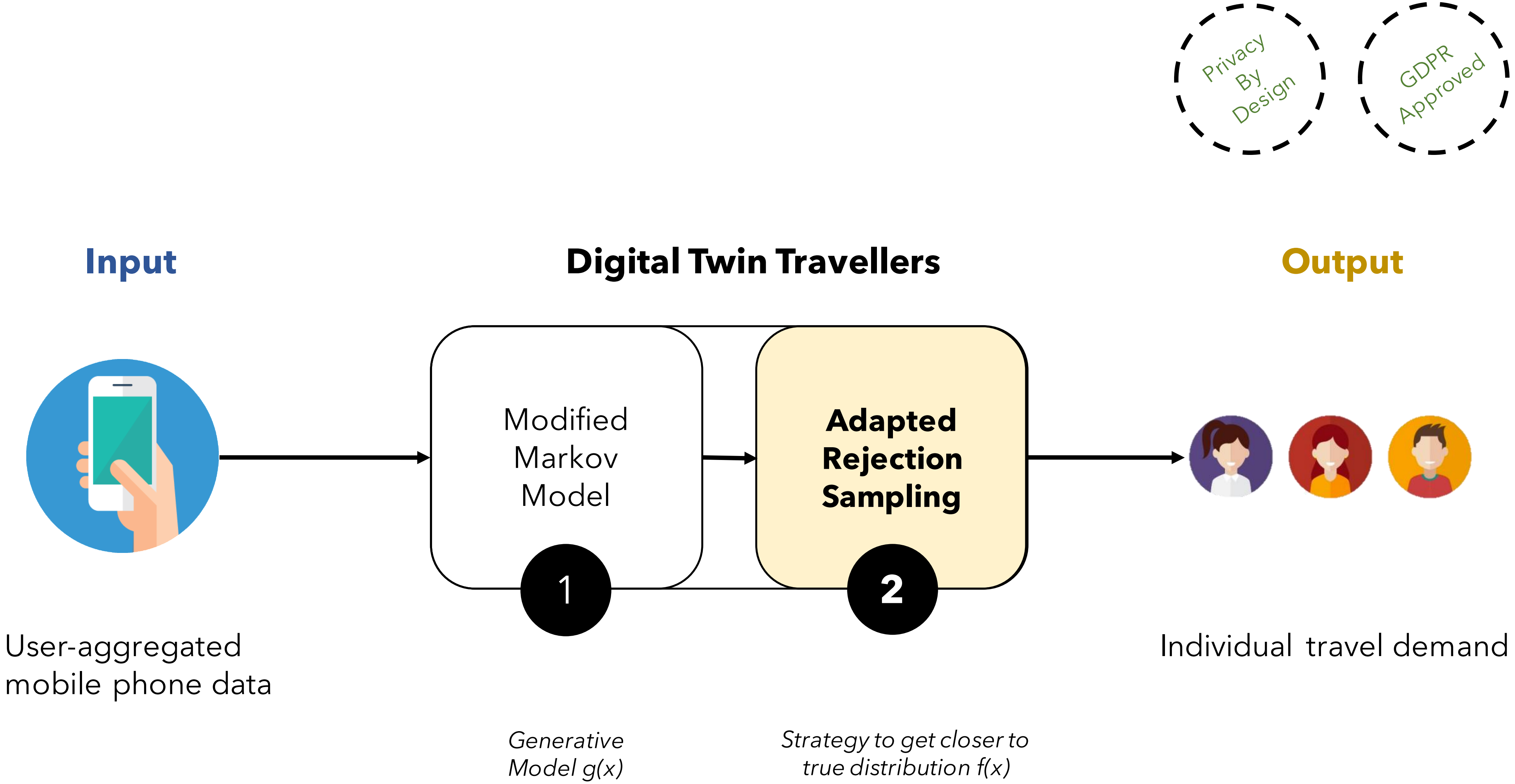
10 distributions validated

1. Activity start time
2. Activity duration
3. Number of trips
4. Tour network
5. Dynamic spatial error
6. Total distance travelled
7. Radius of gyration
8. Mobility entropy
9. Activity space
10. Semantic similarity *(**new**)*

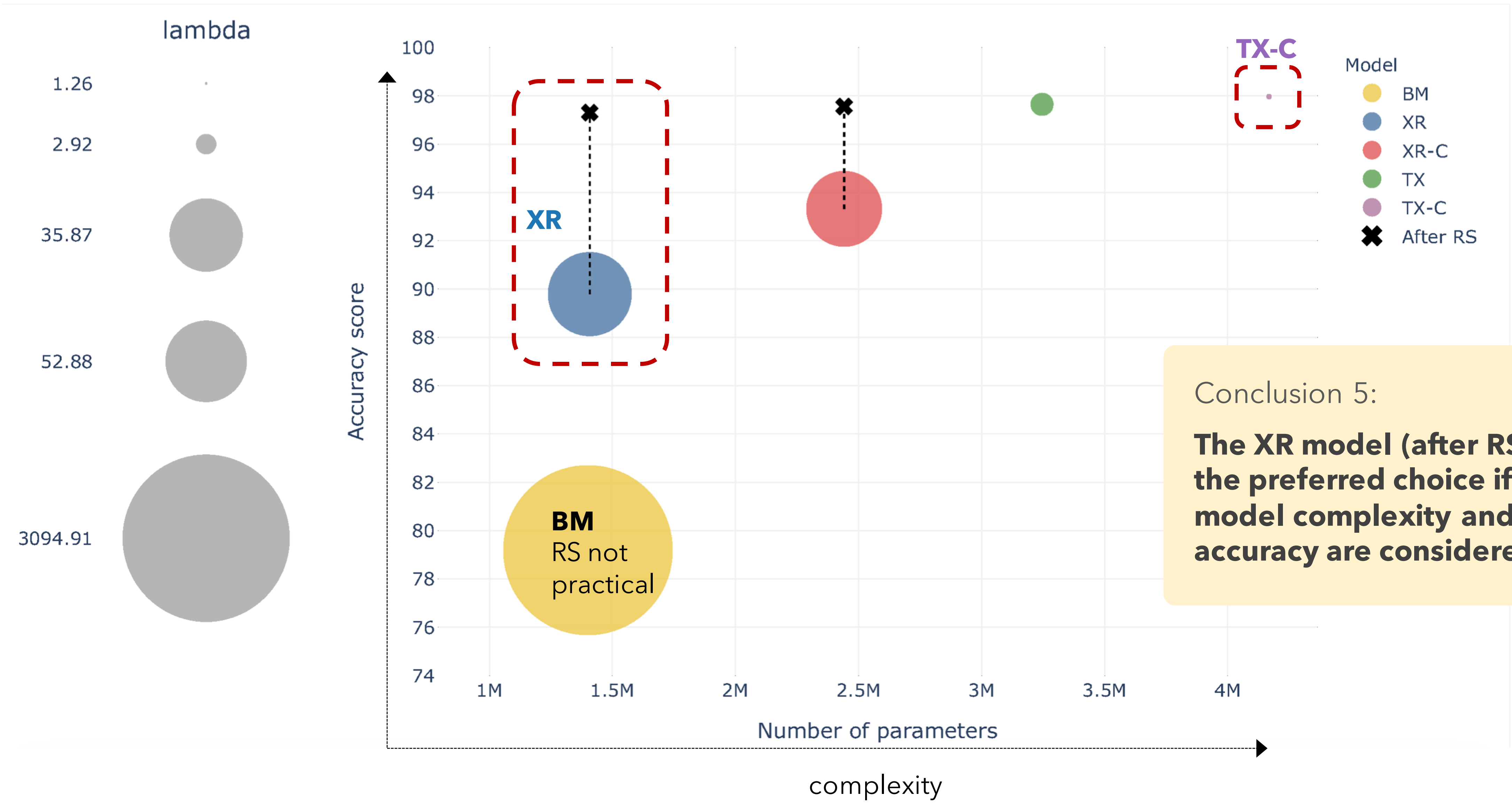
Model	OMPAS (%)
BM	79.19
XR	89.79
XR-C	93.32
TX	97.65
TX-C	**97.97**
DITRAS	83.05

Conclusion 4:

TX-C model outperforms all others


Privacy By Design
GDPR Approved
Input
Digital Twin Travellers
Output
Modified Markov Model
1
Adapted Rejection Sampling
2
User-aggregated mobile phone data
Individual travel demand
Generative Model g(x)
Strategy to get closer to true distribution f(x)

Results: Model complexity vs accuracy



Conclusion 5:

The XR model (after RS) is the preferred choice if both model complexity and accuracy are considered

Agenda

1. Mobile network signalling data
2. Motivation
3. Intuitive example
4. Digital Twin Travellers
5. Experiment and results
6. Application
7. Conclusion

Application

1. Mode choice

Do the Digital Twin Travellers **choose** the same **modes of transport** than the real population?

2. Agent-based simulation

Do the Digital Twin Travellers **congest the city** in a similar way than the real population?

Mode choice

Model: Random Forest Classifier

Training data: One week of trips extracted from mobile phone data (DataSpark)

Modes:
Car
Subway
Bus
Walk
Road

Feature set:
1. Start time
2. Distance
3. Travel time
4. Origin Car share
5. Origin Subway share
6. Origin Bus share
7. Origin Walk share
8. Origin Road share
9. Destination Car share
10. Destination Subway share
11. Destination Bus share
12. Destination Walk share
13. Destination Road share

Results

Origin: **All** Destination: **All** Time: **All**

Main Mode	Digital Twin Travellers	Itineraries extracted from mobile data
Car	**62.03**	**62.15**
Subway	**18.60**	**17.84**
Walk	**10.07**	**10.74**
Road	**4.76**	**4.79**
Bus	**4.53**	**4.45**

Mode choice

Model: Random Forest Classifier

Training data: One week of trips extracted from mobile phone data (DataSpark)

Modes:
Car
Subway
Bus
Walk
Road

Feature set:

1. Start time
2. Distance
3. Travel time
4. Origin Car share
5. Origin Subway share
6. Origin Bus share
7. Origin Walk share
8. Origin Road share
9. Destination Car share
10. Destination Subway share
11. Destination Bus share
12. Destination Walk share
13. Destination Road share

Results

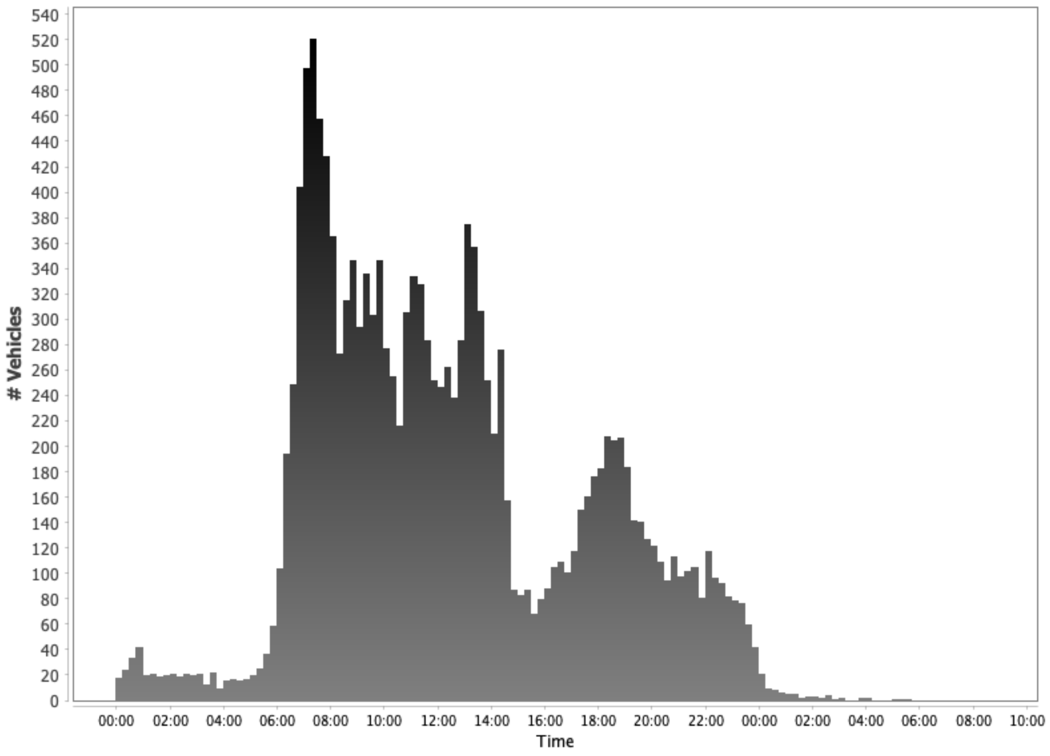
Origin: **All** Destination: **Raffles P.** Time: **8 am**

Main Mode	Digital Twin Travellers	Itineraries extracted from mobile data
Car	**62.38**	**62.93**
Subway	**18.18**	**17.56**
Walk	**10.58**	**10.48**
Road	**4.66**	**4.64**
Bus	**4.18**	**4.37**

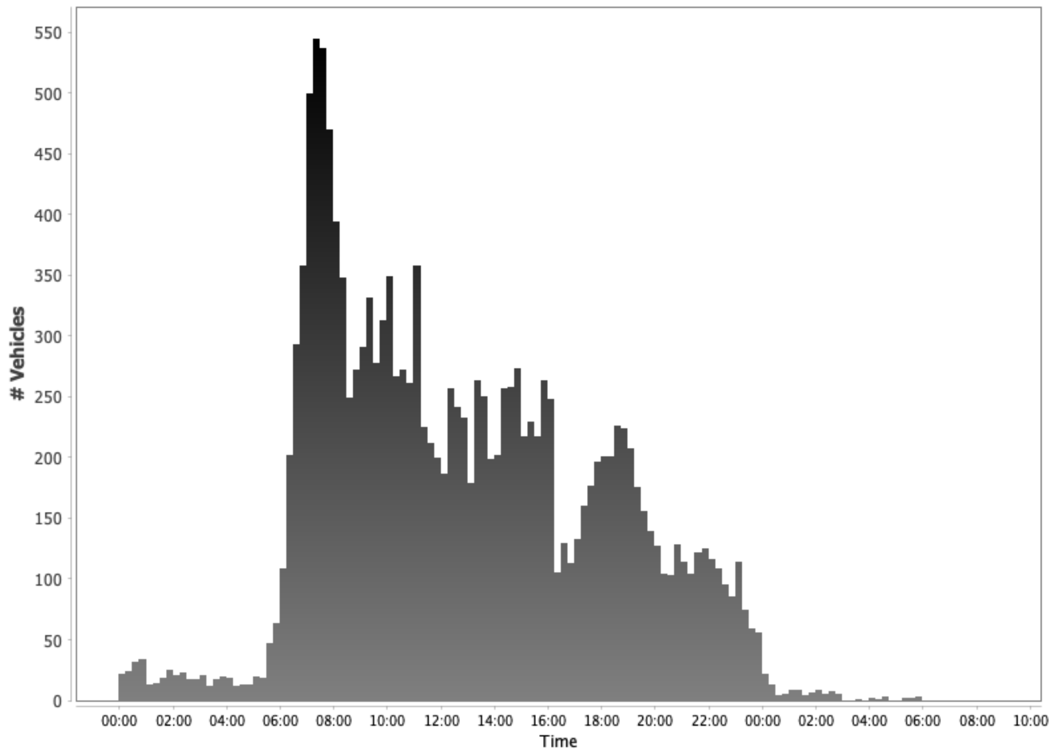
Digital Twin Travellers vs Real mobile users data
in an agent-based simulation

Road counts
Geylang Rd (between KPE and Lor 22 Geylang)

Users mobile data

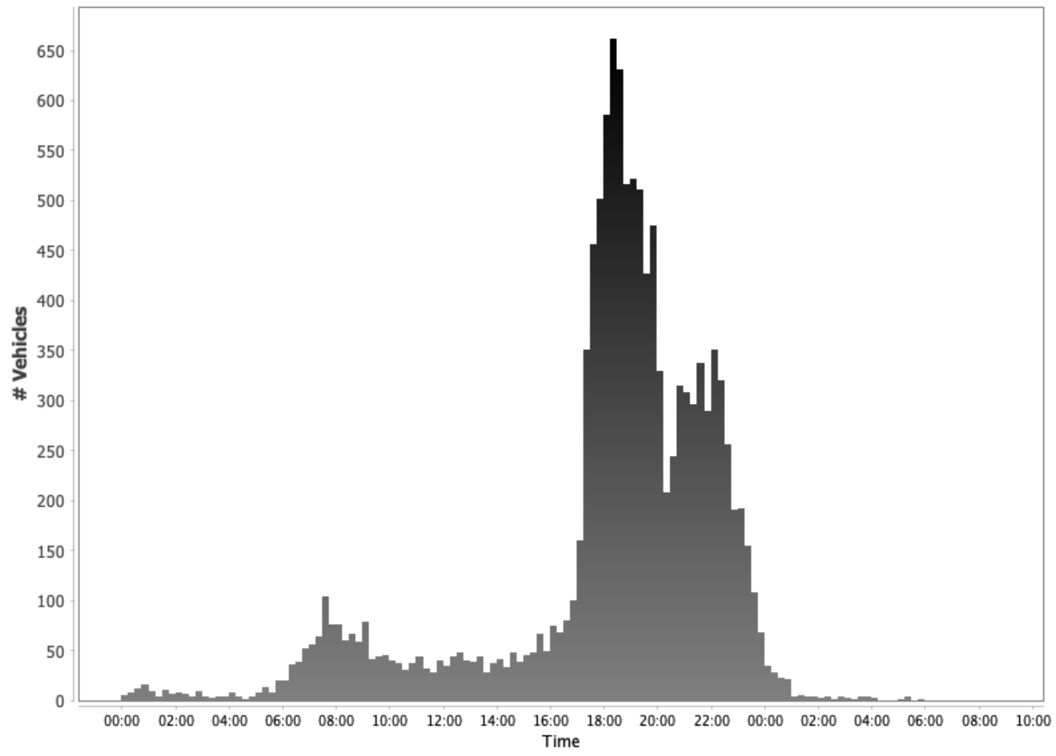


Digital Twin Travellers

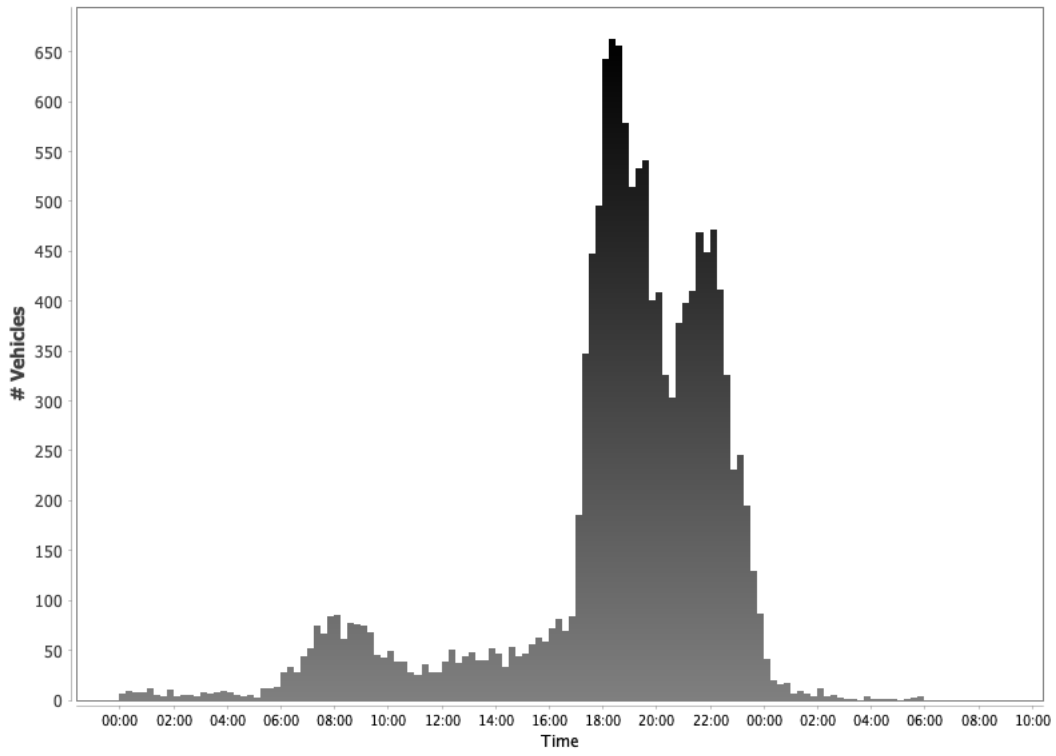


Road counts
MCE Tunnel

Users mobile data

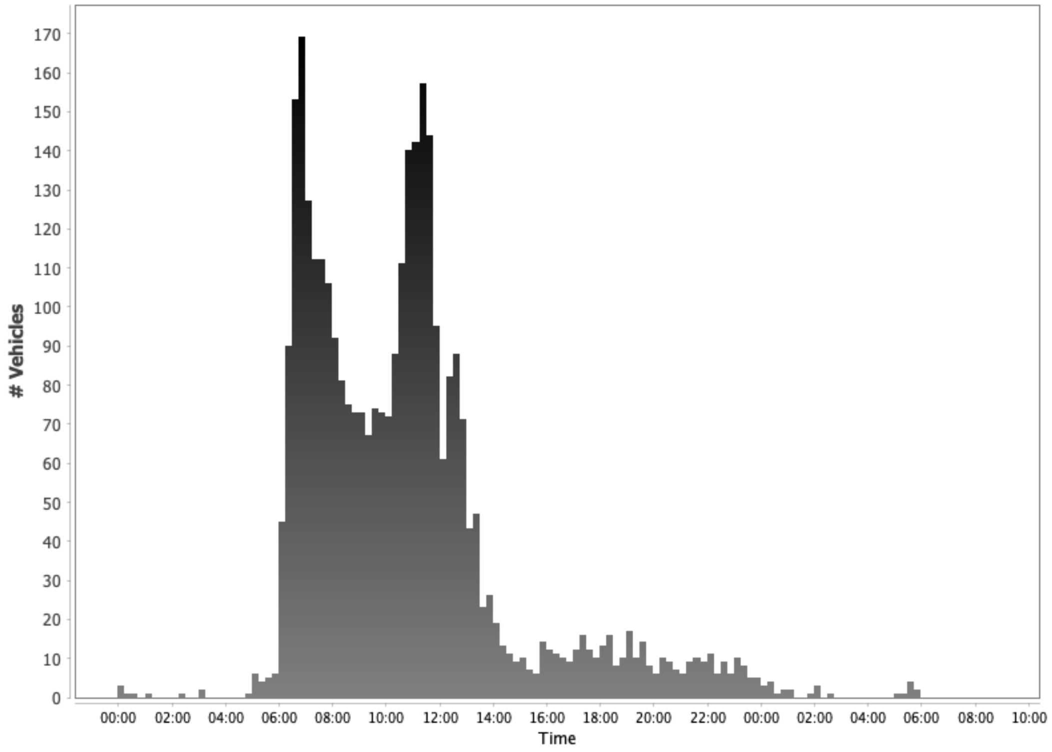


Digital Twin Travellers

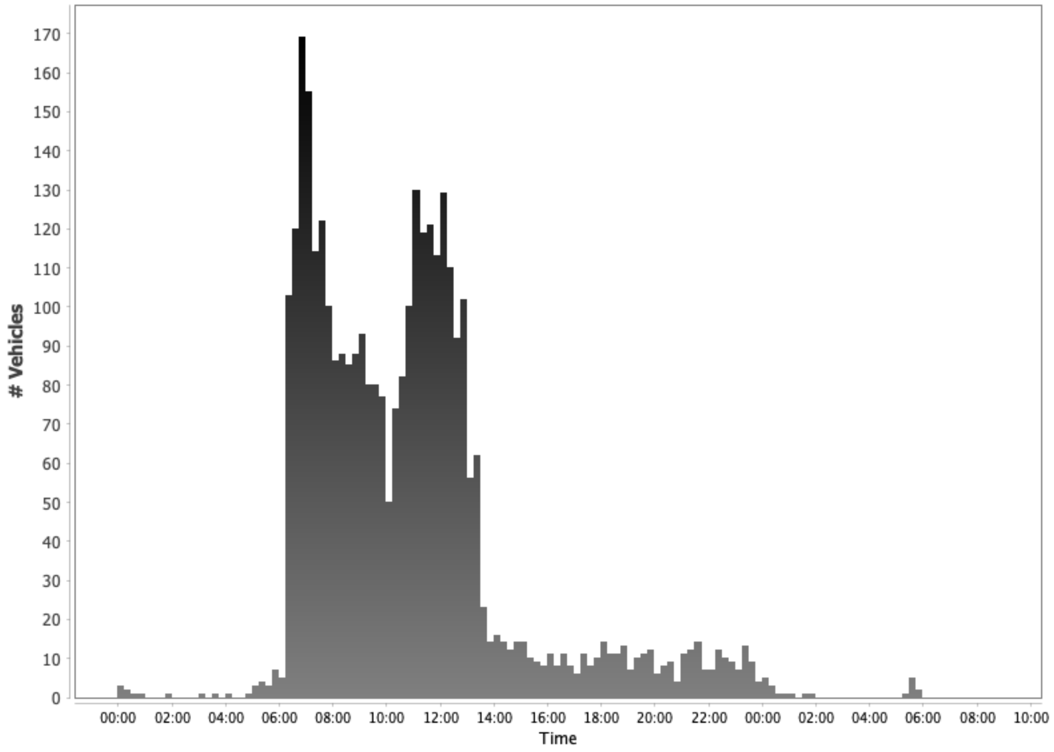


Road counts
BKE (Between Bukit Panjang Rd and Diary Farm Rd)

Users mobile data

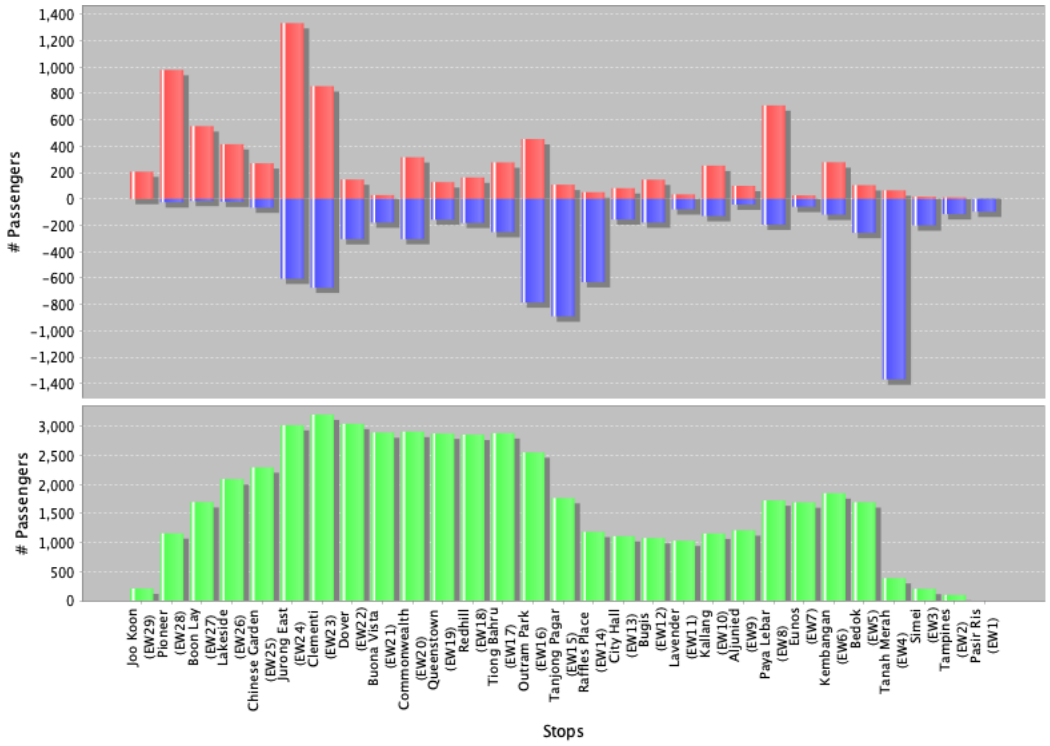


Digital Twin Travellers

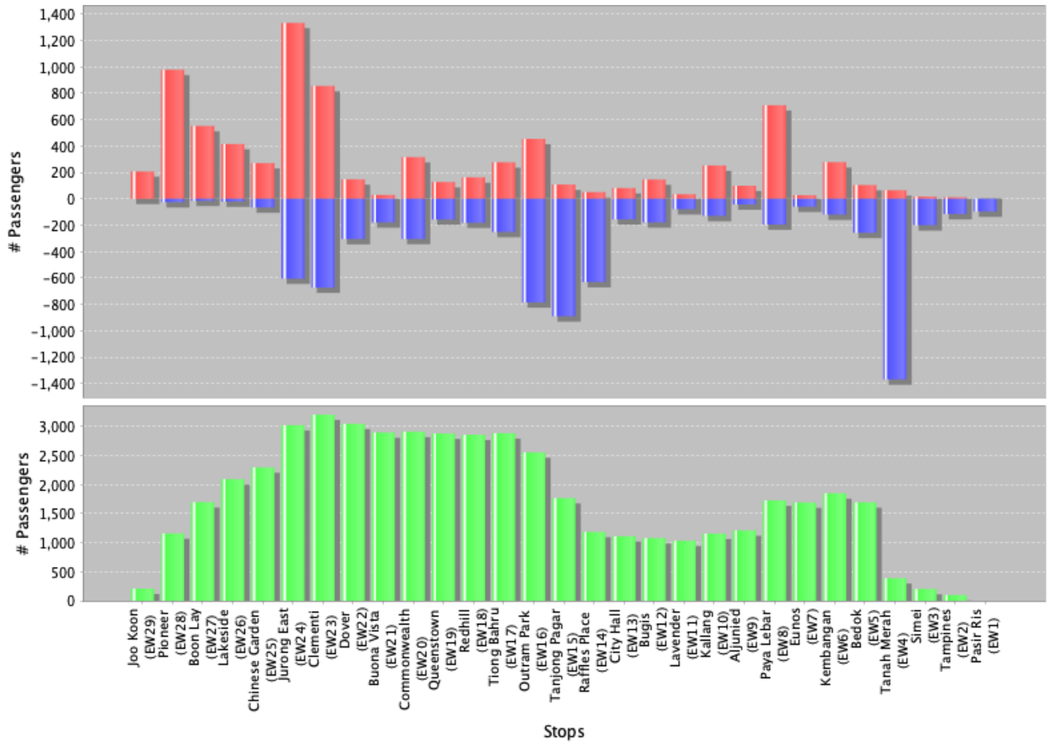


East-West Line occupancy
Joo Koon – Pasir Ris
Morning peak

Users mobile data



Digital Twin Travellers



Agenda

1. Mobile network signalling data
2. Motivation
3. Intuitive example
4. Digital Twin Travellers
5. Experiment and results
6. Application
7. Conclusion

Conclusion

Contribution

Digital Twin Travellers: framework to synthesise realistic and individual travel demand from aggregates of mobile phone data

Integrates mobile phone data into **disaggregated travel demand models** and **simulations**

Privacy-by-Design approach and inline with data privacy regulations enables the framework to be used in **practice**

Contribution

Baseline Markov (**BM**), Explore and Return (**XR**), Tour Explicit (**TX**) generative models for individual travel demand

Method to find **types of travellers** based uniquely on their daily itinerary extracted from mobile phone data

Two new metrics to measure the similarity between two populations consisting of one-day itineraries: **OMPAS** and **SS**

Contribution

Digital Twin Travellers compared with users' itineraries extracted from mobile phone data in a **mode choice** model and an **agent based simulation**

Application

A transport modeller requires only **7** different types of **histograms** from the TSP

Future research

Digital Twin Travellers for future/alternative scenarios

Digital Twin Travellers with socio-demographic and trip purpose information

Publications

1	Literature review	Journal	Anda, C., A. Erath and P. J. Fourie (2017) Transport modelling in the age of big data, *International Journal of Urban Sciences*, 21 (sup1) 19–42.
2	Dynamic Bayesian Networks for activity-based models	Conference	Anda, C. and S. A. Ordoñez Medina(2017) Extending the hidden Markov model for activity scheduling, paper presented at the *6th Symposium of the European Association for Research in Transportation (hEART 2017)*, Haifa, September 2017.
3	Clustering types of travellers from mobile phone data	Conference	Anda, C. and S. A. Ordoñez Medina (2018a) Archetypes of urban travellers: Clustering of mobile phone users in Singapore, paper presented at the *Mobile Tartu Conference 2018*, Tartu, June 2018.
4	BM model	Conference	Anda, C. and S. A. Ordoñez Medina (2018b) A time-space model of disaggregated urban mobility from aggregated mobile phone data, paper presented at the *15th International Conference on Travel Behavior Research (IATBR 2018)*, Santa Barbara, July 2018.
5	XR and TX models	Journal	Anda, C., S. A. Ordoñez Medina and K. W. Axhausen (2021) Synthesising digital twin travellers: Individual travel demand from aggregated mobile phone data, *Transportation Re- search Part C: Emerging Technologies*, 128, 103118.
6	Agent-based simulation from mobile phone data	Journal	Anda, C., S. A. Ordoñez Medina and P. Fourie (2018) Multi-agent urban transport simulations using od matrices from mobile phone data, *Procedia Computer Science*, 130, 803–809.

Github: github.com/candac/DigitalTwinTravellers

Thanks