Twitter-informed Crowd Flow Prediction

1st Workshop on Data-driven Intelligent Transportation (DIT 2018)

Gary Goh, Koh Jing Yu, Zhang Yue

17 November 2018

Singapore-MIT Alliance for Research and Technology

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN
Established in collaboration with MIT

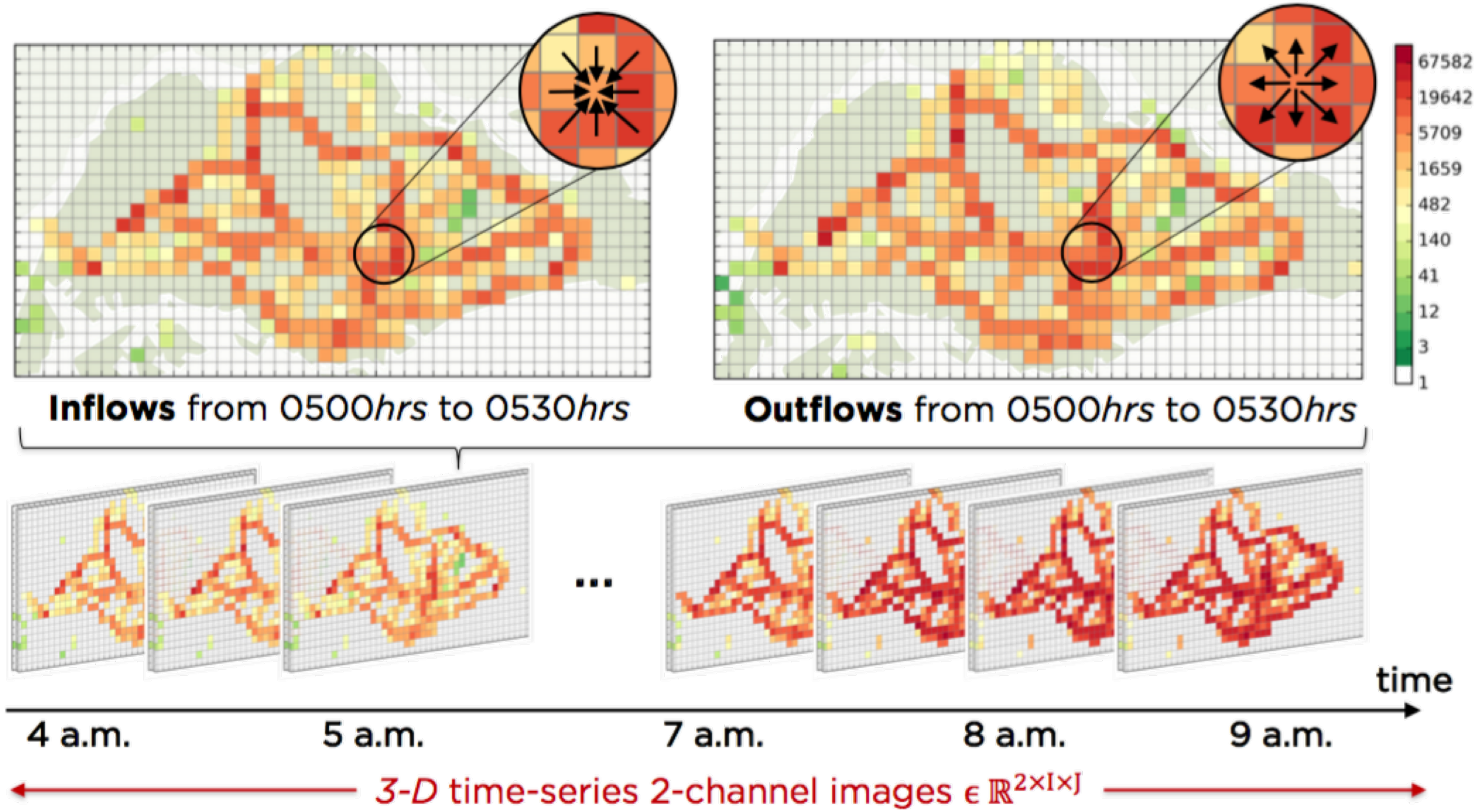
Research Statement

- **Traffic flow predictions:**
  - Most work rely on pattern recognition from historical traffic data
  - Generally effective; traffic is periodical and thus quite predictable
  - Areas poorly: non-recurring large-scale events
    e.g. traffic incidents, sports events, riots & demonstrations
- **Idea:** use real-time information from internet *(mainly twitter)* to capture these non-recurring events
- **Research questions:**
  1. Can tweets be useful to predict future traffic flows?
  2. How are tweets related to traffic flows?
- **Challenges:**
  - Noisy data
  - Spatio-temporal dependencies

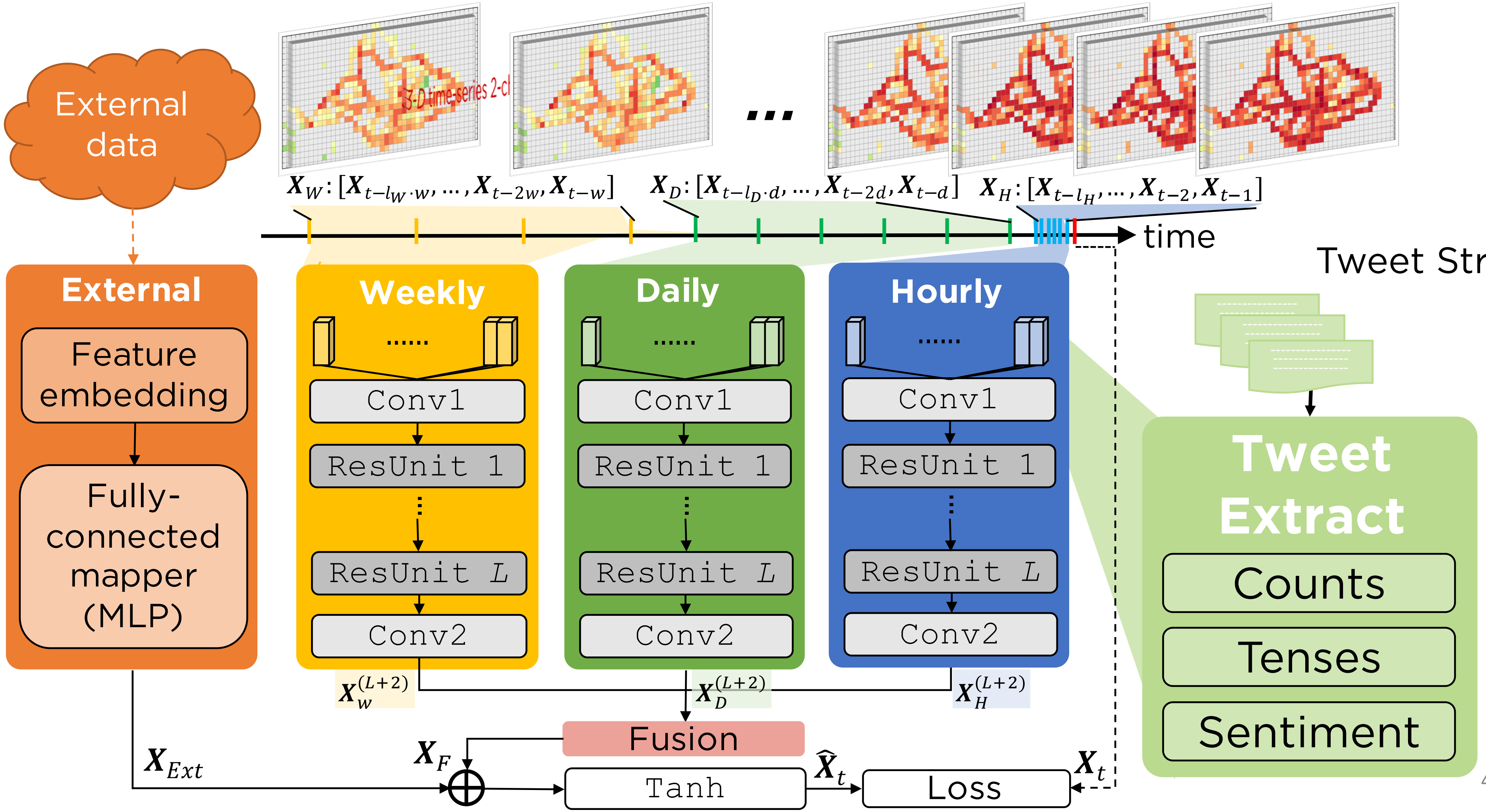
Crowd Flow Prediction

- **Aim:** Predict future traffic flows (or traffic conditions)

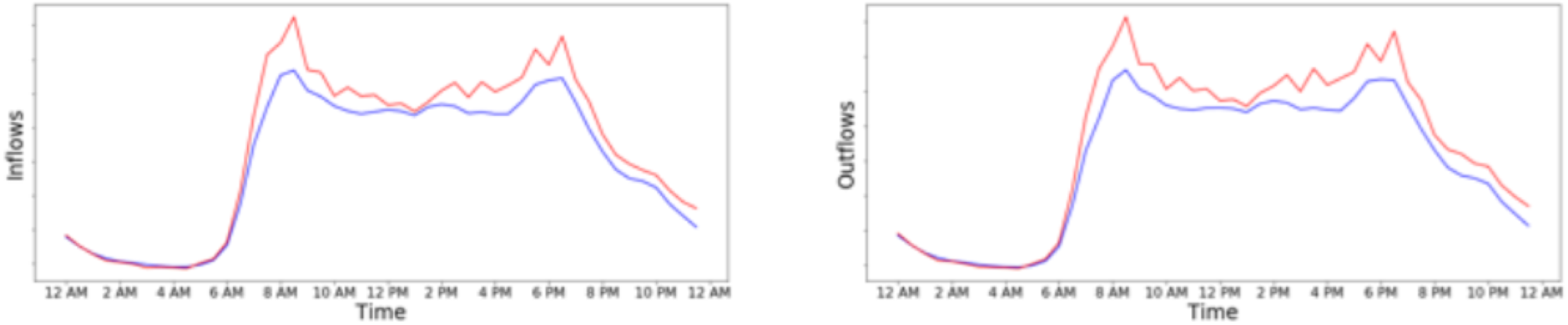
  *** aggregated traffic flows = crowd flows**



Spatio-temporal Residual Network (ST-ResNet)

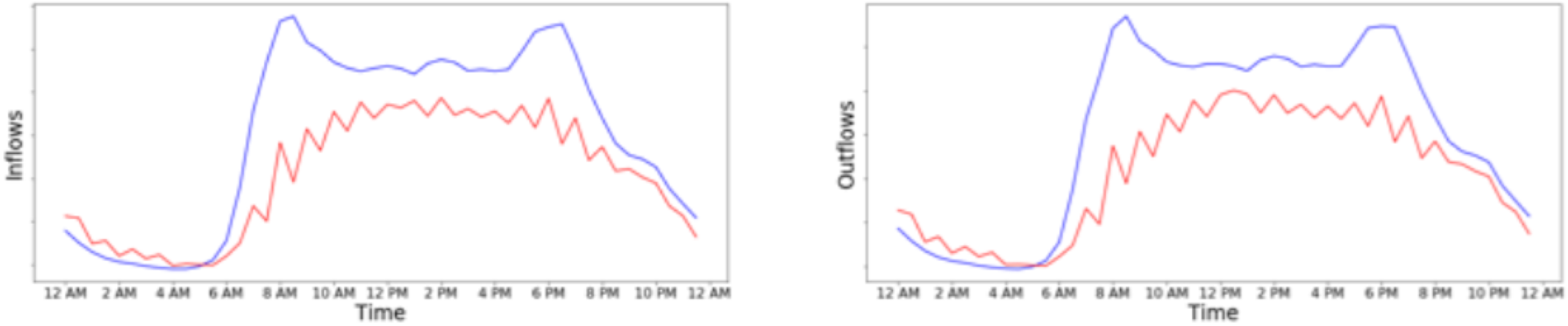


Time	Tweet text
17:38	Now at Jalan Besar stadium alr!!
17:41	Now going to jalan besar for ball picker
19:12	Finally the jam cleared. Now en route to Jalan Besar !!
19:34	Kickoff at the Jalan Besar Stadium - Albirex Niigata (S) 0-0 Home United! #SLeague
19:50	A bit late but thrilled to be watching #ALB v #HUFC @Jalan Besar Stadium



(a) VLDs' crowd flows (bottom) in nearby regions during a present large-scale event on 1 March 2013 around 7.30pm and a sample of the relevant tweets (top). Crowd flows are higher than expected.

Time	Tweet text
10:58	Wah f*ck the road near paya lebar mrt flood like shit its fucking knee deep in water !!!
10:59	wow rain till gt flood at paya
10:59	Oh my gosh... Paya Lebar is flooded. Literally.. Like shin-level.
11:09	Water level falls below 90%. High Flood Risk.11:09:13 #SgFlood,,,#SgFlood



(c) VLDs' Crowd flows (bottom) in nearby regions during a negatively-potrayed event on 28 April 2013 around 11am and the relevant tweets (top). Crowd flows are lower than expected.

Singapore Experiments: Data sources

- **Traffic Flow**
  1. Vehicular Loop Detectors
  2. Mobile Phone Signals

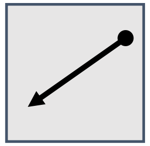
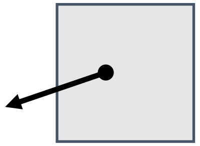
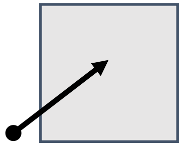
Dataset	VLD	MPS
Data type	Vehicular counts	Origin-destination pairs
Time span	Set 1: 1/3/2013 – 30/6/2013 Set 2: 1/9/2014 – 31/12/2014 Set 3: 1/12/2015 – 31/3/2016	1/8/2017 – 30/11/2017
Time interval	30 minutes	1 hour
# of time intervals (T)	5,856	1,464
Grid map size (I, J)	(89, 49)	(90, 54)

- **External Data:**
  1. Weather information
  2. Public holidays & weekends

- **Tweets Data:**
  - 369 search keywords of critical locations' names across the city

Input Data Description

- **Inflow/outflow matrix:** aggregate flows spatially and temporally

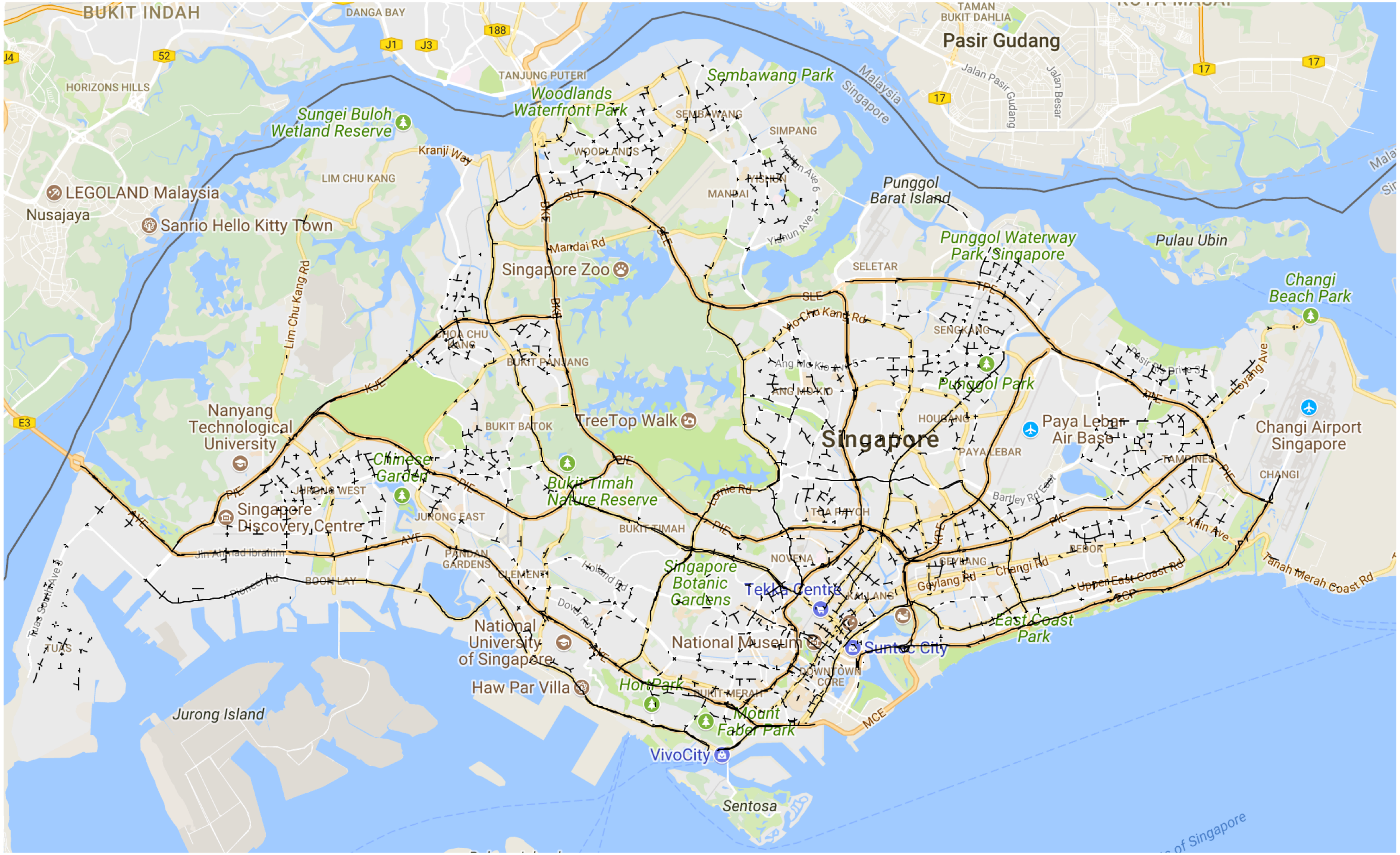
 $ignore$  $x_t^{out,i,j} = \sum_{k^{out} \in (i,j)} f_{k^{out}}^t$  $x_t^{in,i,j} = \sum_{k^{in} \in (i,j)} f_{k^{in}}^t$

  - Flows are normalized to be in the range $[-1, 1]$, i.e. $\widetilde{\boldsymbol{X}}_t = 2 \cdot \frac{\boldsymbol{X}_t - \min(\boldsymbol{X}_t)}{\max(\boldsymbol{X}_t) - \min(\boldsymbol{X}_t)} - 1$

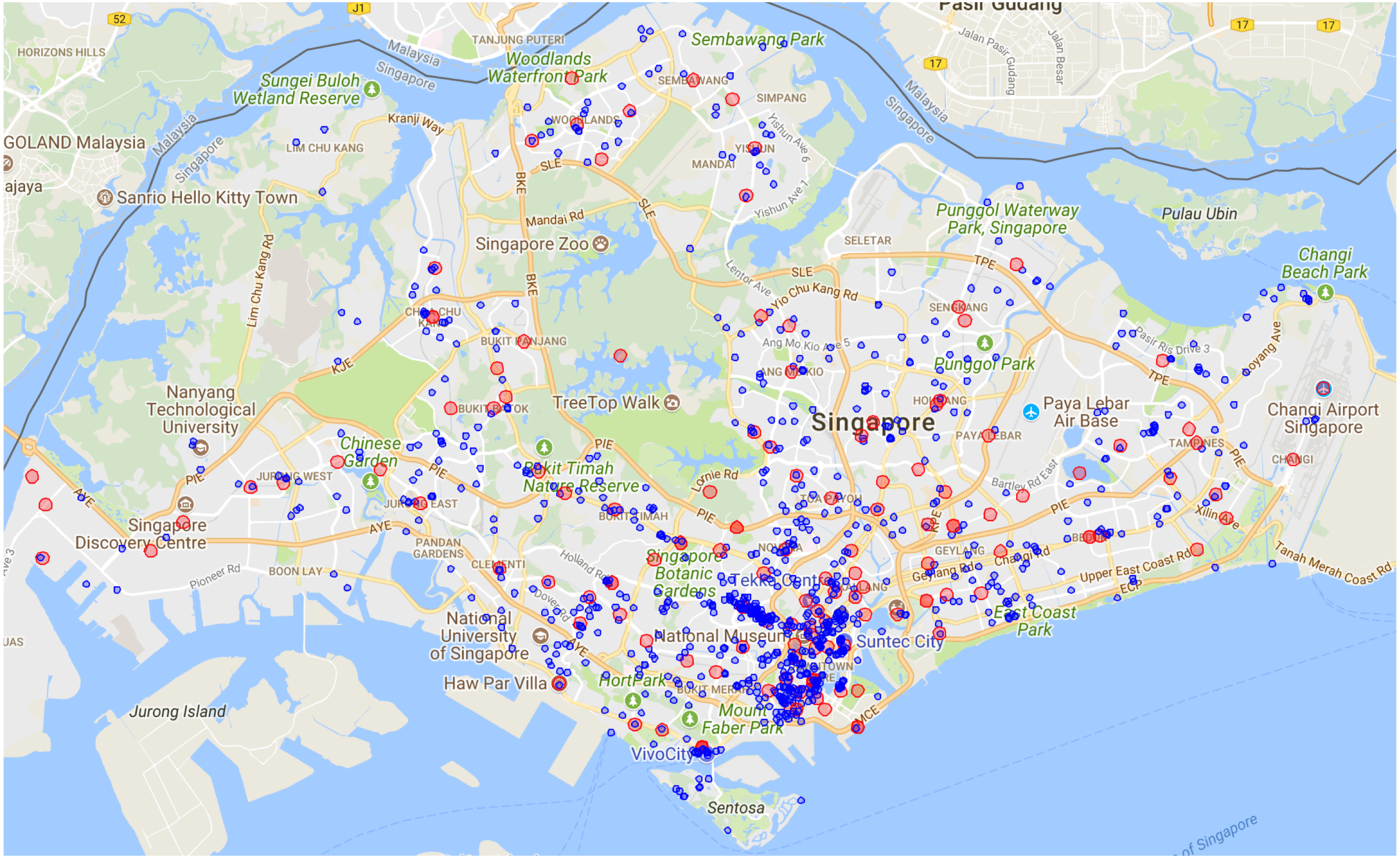
- **External data:**
  - **Holiday data:** list of dates e.g. `[2016-01-01, 2016-02-08, …]`
  - **Metrological data:**
    - List of corresponding temperature (℃) e.g. `[27, 28, 29, 30, 31, 31, 29, …]`
    - List of corresponding wind speed (km/h) e.g. `[4, 0, 0, 2, 3, 6, 12, 14, 20, …]`
    - List of corresponding weather (one-hot vectors) e.g.
```
[[ 1 0 0 0 0 0 0 0],
 [ 0 0 0 1 0 0 0 0],
 [ 0 1 0 0 0 0 0 0], …]
```

where index location of 1 represents the following mapping:
0 = sunny, 1 = cloudy, 2 = overcast, 3 = rain, 4 = light rain, 5 = heavy rain, 6 = fog, 7 = haze

Vehicular Loop Detectors



Selected Places (search keywords)



SG Grid Map (500m x 500m)

Hyperparameters

- **Recent limit** (l_R): 4
- **Period interval** (p): 48 ; **Period limit** (l_P): 1
- **Season interval** (s): 48 X 7 = 336 **Season limit** (l_S): 1
- **Number of filters in** `Conv1` **&** `ResUnit`: 64
- **Number of Residual units** (L): 2
- **Optimization:** Backpropagation with Adam
  - **Learning rate** (α): 0.0002
  - **Development test**: last 10%; **Development training # of epochs**: 500 with early stopping
  - **Cont. training (on full training set) # of epochs**: 100

Experiment Setup

- **Evaluation metric**: Root Mean Square Error (RMSE)

$$RMSE = \sqrt{\frac{1}{z}\sum_i^z (x_i - \hat{x}_i)^2}$$

where $\hat{x}$ is the predicted value and x is the ground truth;
z is the # of all predicted values $= 48\times 4\times 7\times 2\times I\times J$

- **Simple baselines**:
  - **Historical average**: use average flow of the same time slot and day of the week in the train dataset
  - **Persistence**: use previous time slot flow values as current prediction values

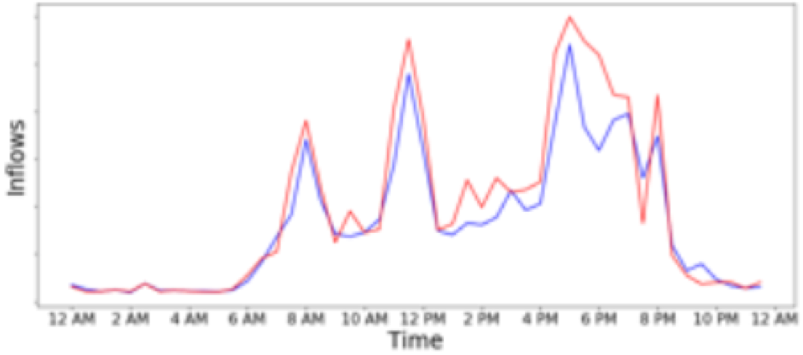
Experiment Results

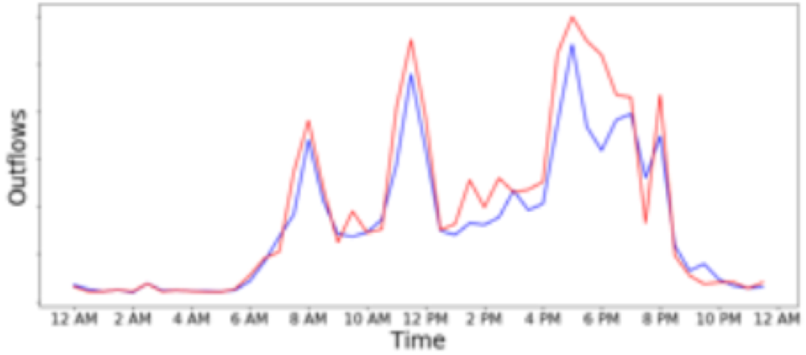
Model	Dataset				Average
	VLD1	VLD2	VLD3	MPS	
ST-ResNet (*Main baseline*)	**3.1278**	**3.4302**	**3.4586**	**2.2520**	**3.0672**
Historical average	5.2428	5.6838	5.0124	2.3585	4.5744
Persistence model	4.3864	4.2329	4.9451	4.9921	4.6391
ST-ResNet + Tweet Counts	**3.1073**	3.2965	3.2345	2.2369	**2.9688**
ST-ResNet + Tweet Tenses	3.1113	3.4238	3.2664	<u>2.2581</u>†	3.0149
ST-ResNet + Tweet Counts + Tenses	<u>3.1459</u>	3.3231	**3.2294**	2.2271	2.9814
ST-ResNet + Tweet Sentiment	<u>3.1300</u>	<u>3.4255</u>†	<u>3.4609</u>†	2.2441	3.0651
ST-ResNet + Tweet Counts + Sentiment	<u>3.1984</u>	3.2578	3.2498	<u>2.3321</u>†	3.0095
ST-ResNet + Tweet Counts + Tenses + Sentiment	<u>3.1578</u>	**3.2455**	3.3409	**2.2072**	2.9879

[a] *Underlined – did not beat baseline; Bold – best score for dataset; † – statistical insignificant.*

- With tweet counts, error reduced by 3.28% on average
- Feature misrepresentation by tenses and sentiment

Time	Tweet text
18:23	walao mrt breakdown at serangoon omg zzzz
18:29	F*cking train stalling at serangoon
18:41	Train broke down at ne line. Now waiting for bus at serangoon . But bus stop overcrowded.
18:50	Human traffic at serangoon is insane
18:57	You gotta be f*cking kidding me. The entire serangoon mrt breakdown





(a) Negative sentiment event on 19 June 2013 at 6pm; relevant tweets (top) and VLDs' crowd flows (bottom). Crowd flows are higher than expected. Note: Blue lines represent daily mean flows, while the red lines represent specific flows on the day.

Conclusion

- Summary:
  - Use tweets as additional source of information for crowd flow prediction by extending upon an existing prediction model
  - 3.28% error reduction on average
- Potential Future work:
  - Work on better encoding of tweets information
  - Look into multilingual processing

Thank you ☺