ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

SeTraStream:

Semantic-Aware Trajectory Construction Over Streaming Movement Data

Zhixian Yan* Nikos Giatrakos† Vangelis Katsikaros†

Nikos Pelekis† Yannis Theodoridis†

*Distributed Information Systems Lab
Swiss Federal Institute of Technology
(EPFL), Lausanne, Switzerland

† Information Management Lab
University of Piraeus,
Piraeus, Greece

12th International Symposium on Spatial and Temporal Databases

Minneapolis, MN, USA, 26 August 2011

Outline

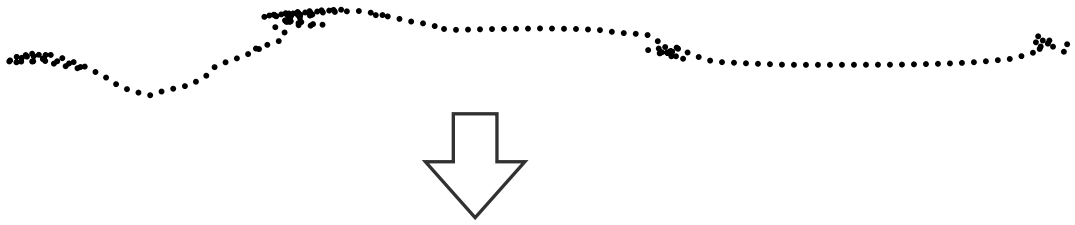
- Introduction
  - semantic trajectories…
  - …over streaming movement data?
- Related Work
- SeTraStream Framework
  - Big Picture
  - Details of each module
    - Data Cleaning
    - Data Compression
    - Segmentation – Episode Identification
- Experimental Evaluation
- Conclusions

Outline

- Introduction
  - semantic trajectories…
  - …over streaming movement data?
- Related Work
- SeTraStream Framework
  - Big Picture
  - Details of each module
    - Data Cleaning
    - Data Compression
    - Segmentation – Episode Identification
- Experimental Evaluation
- Conclusions

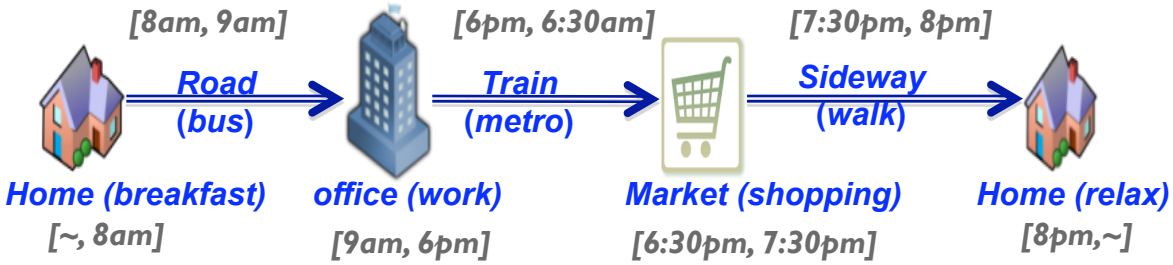
What is semantic trajectory?

raw mobility data
sequence (x,y,t) points
e.g., GPS feeds



meaningful mobility tuples

<place, $time_{in}$, $time_{out}$, tags>



- Semantic Trajectory: $T=\{e_{first},\ldots,e_{last}\}$
- Episode: $e_i=(t_{from}, t_{to}, place, tag)$

Why semantic trajectories?

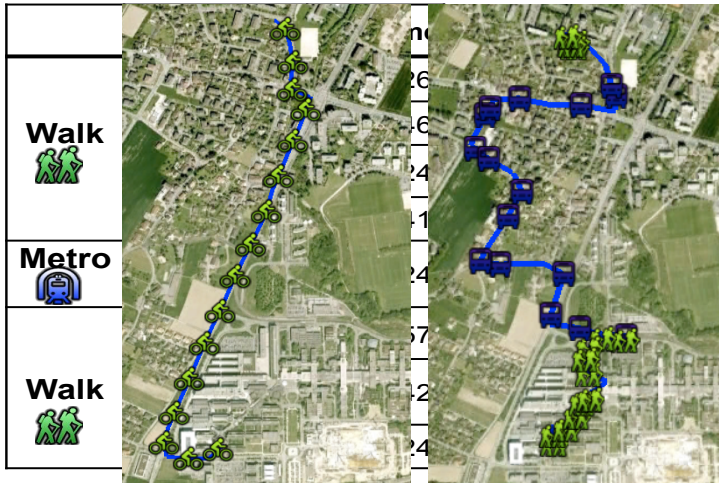
- Detection of homogenous fractions of movement,
  - Trajectory is recreated **as a sequence of episodes (stops/moves)**
  - E.g., home, shopping, move with bus, in train …
- Semantic data abstraction & compression (efficiency/effectiveness)
- Better mobility understanding & LBS

Home-office trajectory examples

Raw GPS Points

Trajectory Notion of Segments

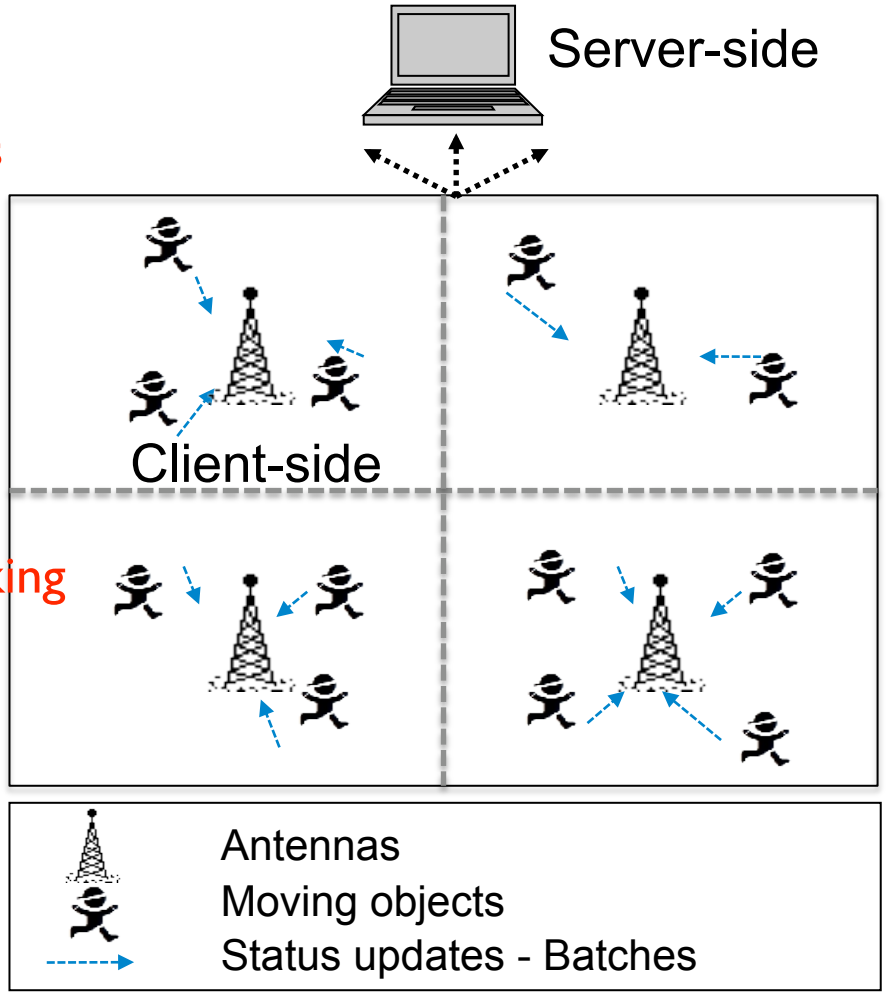
Semantic-Aware Trajectory



(a) HomeOffice via Bike (b) HomeOffice via Bus

Why on streaming mobility data?

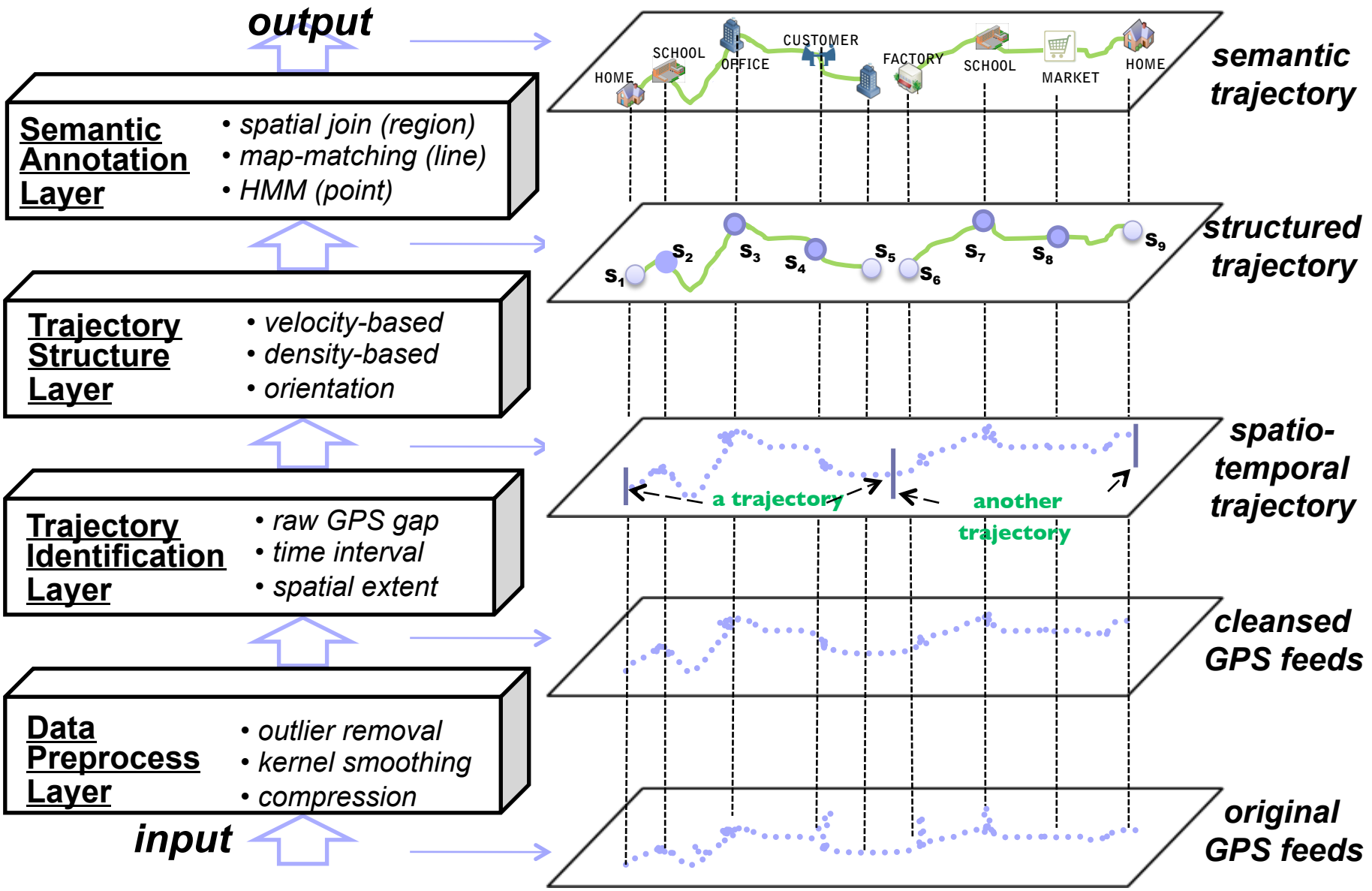
- Offline vs. Real-time
  - Offline: past trajectories
  - mobility streams: ongoing trajectories
  - efficient computation
- Real-life scenarios
  - Traffic Control Scenarios: real time placement & rearrangement of traffic wardens
  - Modern Navigation & Social Networking Services e.g. www.waze.com
  - …
- Distributed setting
  - local site vs. coordinator
  - client vs. server side



Outline

- Introduction
  - semantic trajectories…
  - …over streaming movement data?
- Related Work
- SeTraStream Framework
  - Big Picture
  - Details of each module
    - Data Cleaning
    - Data Compression
    - Segmentation – Episode Identification
- Experimental Evaluation
- Conclusions

Offline Construction of Semantic Trajectories (ESWC '10, EDBT '11)



Related Work & Motivation

- Semantic Trajectories (DKE '08, ESWC '10, EDBT '11)
  - High-level trajectory concepts like episodes (e.g., stops/moves), trajectory ontologies
  - Offline training & tuning parameters (particularly on raw movement features like velocity/direction/density)
  - Tuning parameters, not efficient in real-time settings

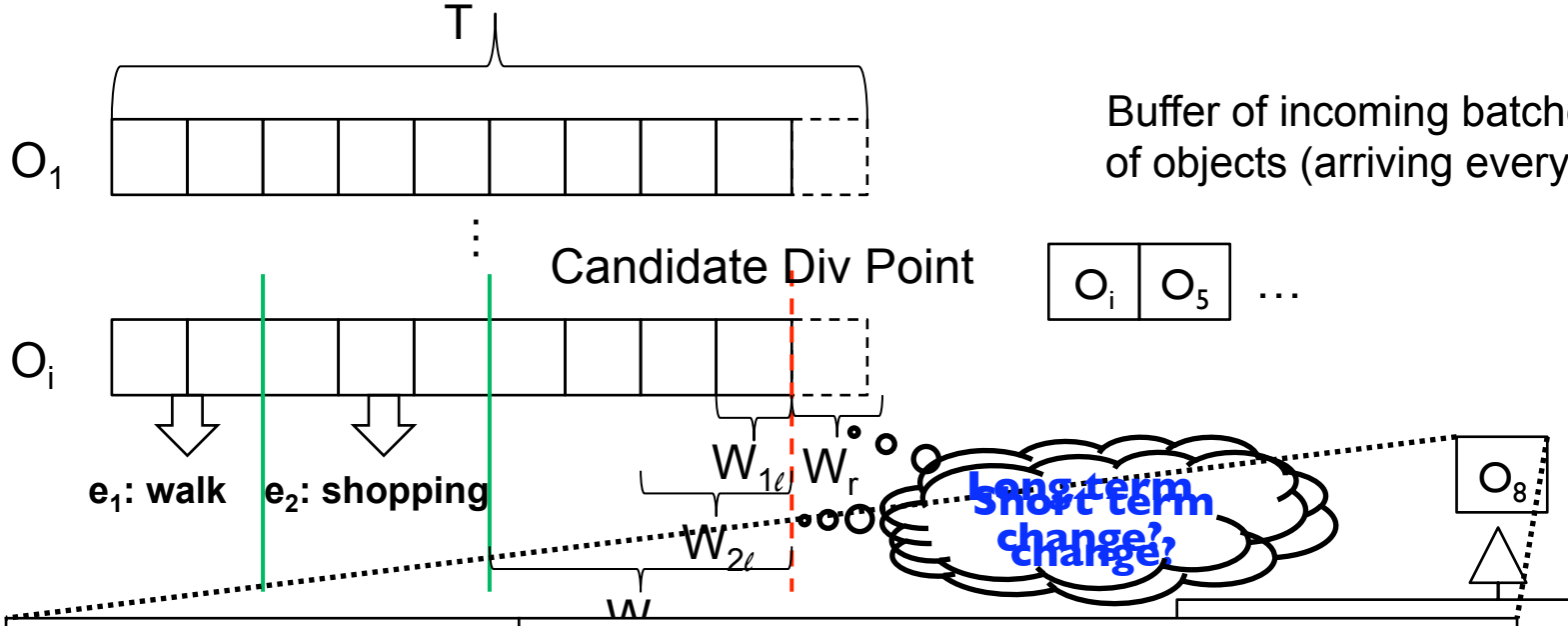
- Streaming data processing
  - Online mobility data compression (e.g., Honle @GIS '10)
  - Time series online segmentation (e.g., Keogh @ICDM '01)
  - Tilted time window specification (Giannotti '02)

Semantic Trajectories + Online Algorithms

Outline

- Introduction
  - semantic trajectories…
  - …over streaming movement data?
- Related Work
- SeTraStream Framework
  - Big Picture
  - Details of each module
    - Data Cleaning
    - Data Compression
    - Segmentation – Episode Identification
- Experimental Evaluation
- Conclusions

SeTraStream - Server Side



Location Stream Instances	Complementary Feature Instances		
<x,y,t>	Position in Lane	Distance to Headway Vehicle	Steering Wheel Activity
123.34, 121.21, 18:35:43	0.1m	1m	π/36
…	…	…	…
120.34, 125.21, 18:36:59	0.05m	3m	π/16

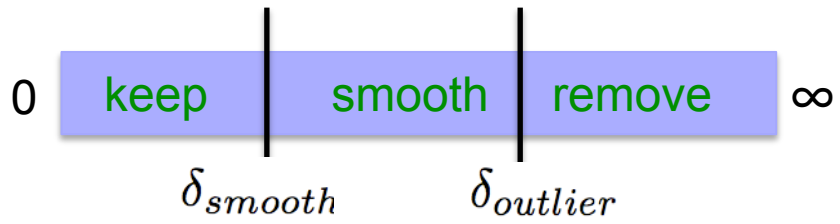
…isy data
…ss batch
…ment Feature
…ors

Online Cleaning (1)

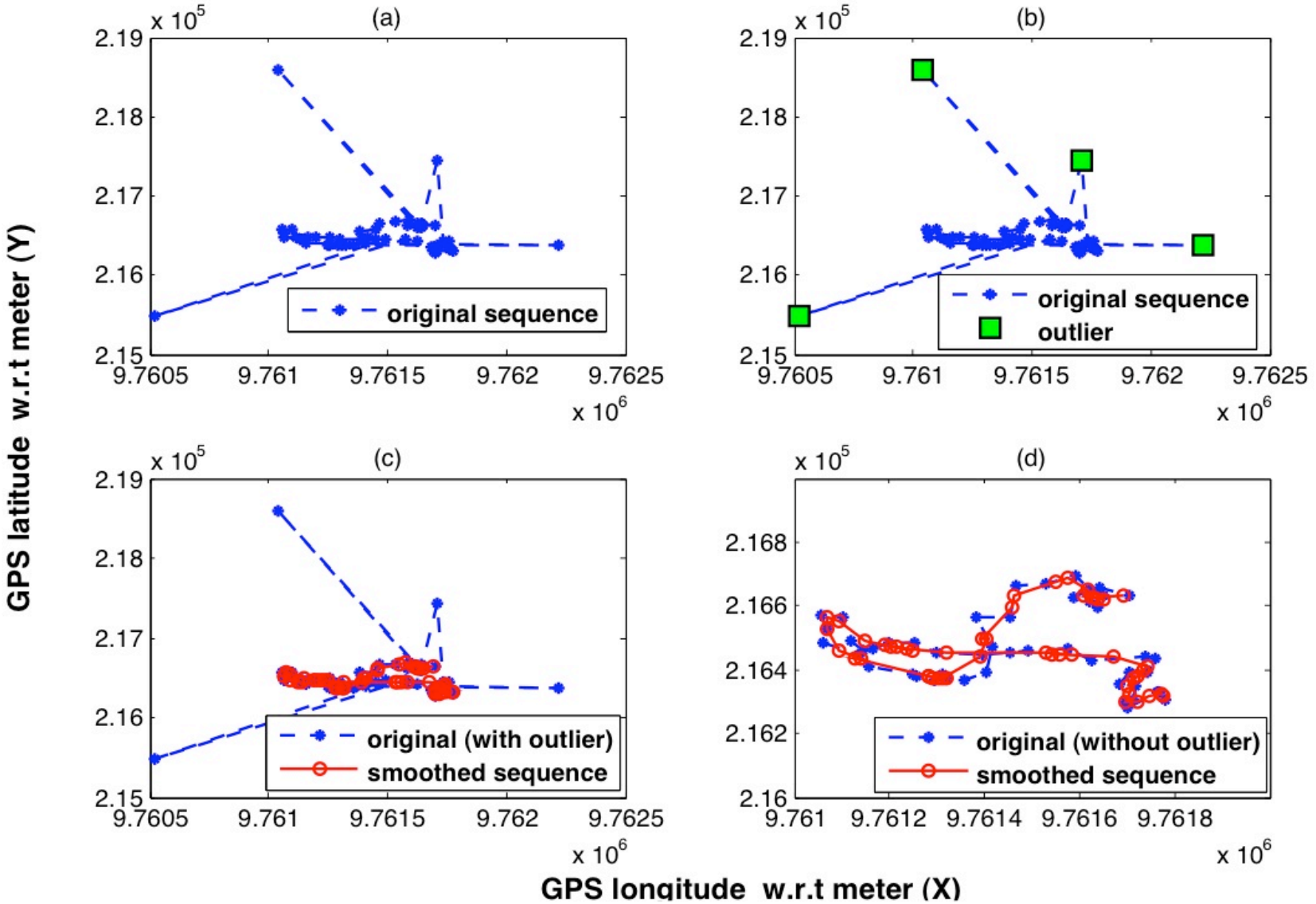
- Two types of GPS errors
  - systematic errors (outlier) - removing
  - random errors (e.g. ±15 meter) – smoothing
- ONE LOOP
  - build Kernal smooth $(\widehat{x},\widehat{y}) = \frac{\sum_i k(t_i)(x_{t_i},y_{t_i})}{\sum_i k(t_i)}$ $k(t_i) = e^{-\frac{(t_i-t)^2}{2\beta^2}}$
  - calculate residual $res = \sqrt{(\widehat{x}-x)^2+(\widehat{y}-y)^2}$
  - calculate the outlier bound & the smooth bound
  - filter outlier or smooth error

$$\delta_{outlier} = v_{limit} \times \left(t_{Q_p^{ls}} - t_{Q_{p-1}^{ls}}\right)$$

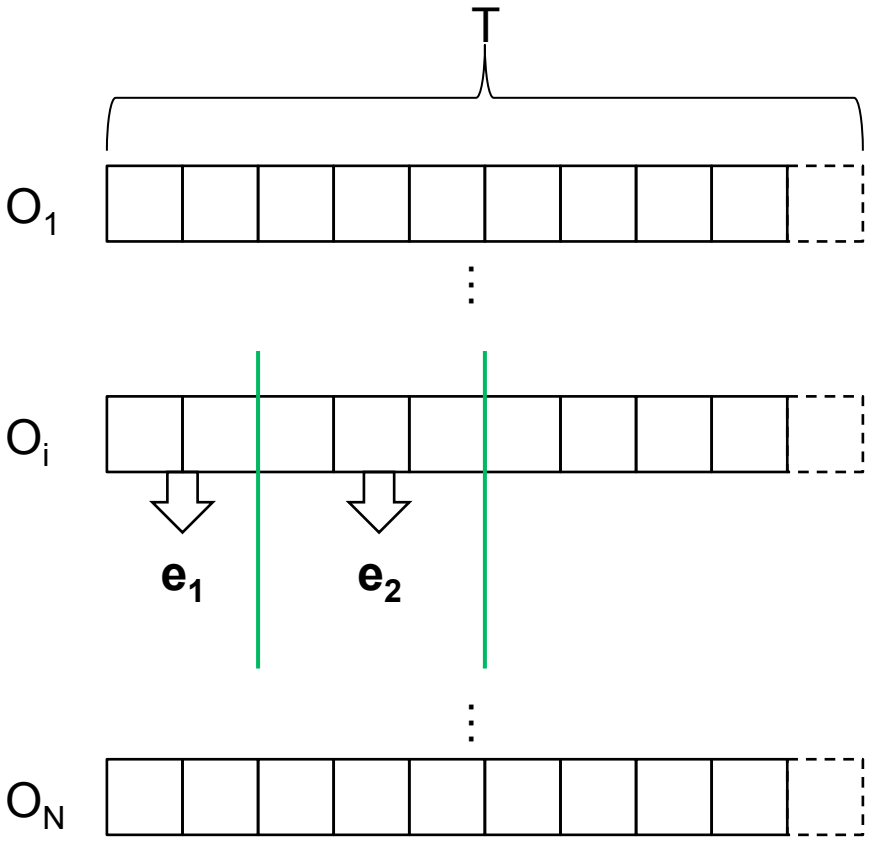
$$\delta_{smooth} = v_{Q_{p-1}^{ls}} \times \left(t_{Q_p^{ls}} - t_{Q_{p-1}^{ls}}\right) \times 120\%$$



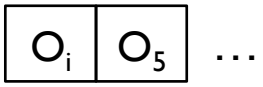
Online Cleaning (2)

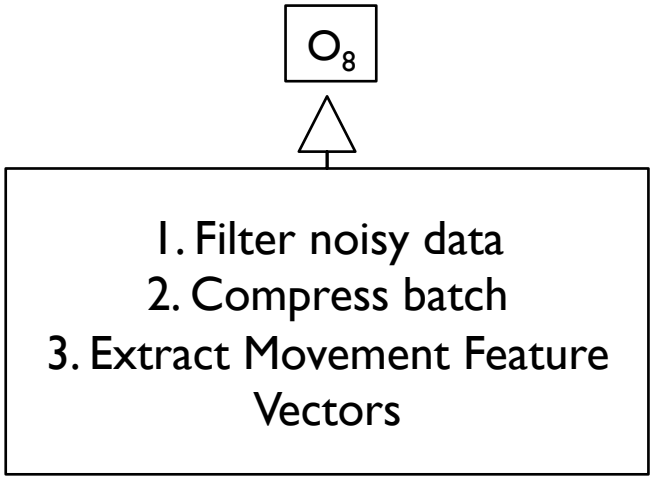


SeTraStream - Compression



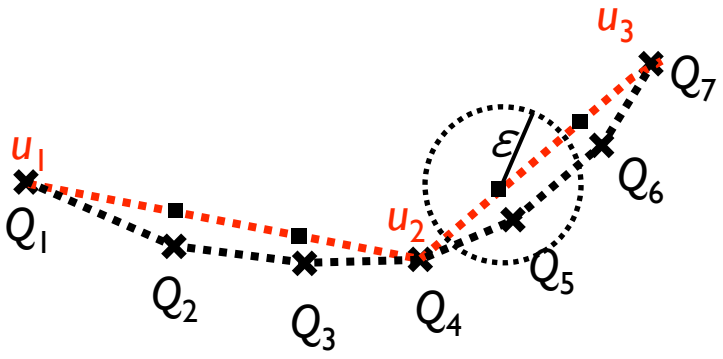
Buffer of incoming batches of objects (arriving every τ)





Online Compression (1)

- Why Compression?
  - Data continuously growing
  - Remove "redundant" data points
  - Reduce transmission cost (local?)
  - Fast computation, application performance

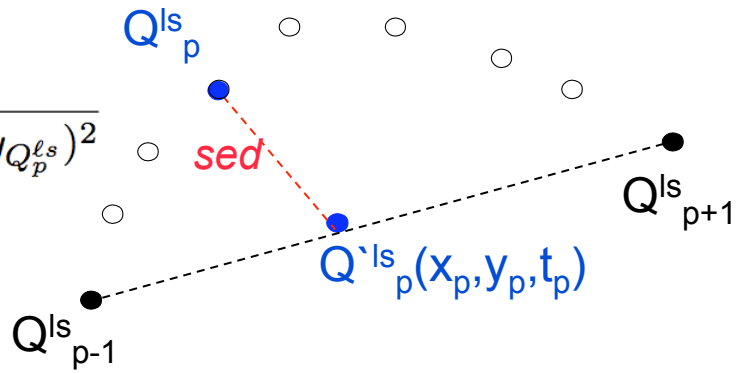


SED (Synchronized Euclidean Distance)

$$sed(Q_p^{ls}, Q_{p-1}^{ls}, Q_{p+1}^{ls}) = \sqrt{(x_{Q_p'^{ls}} - x_{Q_p^{ls}})^2 + (y_{Q_p'^{ls}} - y_{Q_p^{ls}})^2}$$

$$x_{Q_p'^{ls}} = x_{Q_{p-1}^{ls}} + v^x_{Q_{p-1}^{ls} Q_{p+1}^{ls}} \cdot (t_{Q_p^{ls}} - t_{Q_{p-1}^{ls}})$$

$$y_{Q_p'^{ls}} = y_{Q_{p-1}^{ls}} + v^y_{Q_{p-1}^{ls} Q_{p+1}^{ls}} \cdot (t_{Q_p^{ls}} - t_{Q_{p-1}^{ls}})$$



Online Compression (2)

- **SED (Synchronized Euclidean Distance)**
  - Relative Spatio-Temporal Significance
- **SCC (Synchronized Correlation Coefficient)**
  - Relative Significance of the Complementary Features

$$scc(Q_p'^{cf}, Q_p^{cf}) = \frac{E(Q_p'^{cf} Q_p^{cf}) - E(Q_p'^{cf}) E(Q_p^{cf})}{\sqrt{(E((Q_p'^{cf})^2) - E^2(Q_p'^{cf}))(E((Q_p^{cf})^2) - E^2(Q_p^{cf}))}}$$

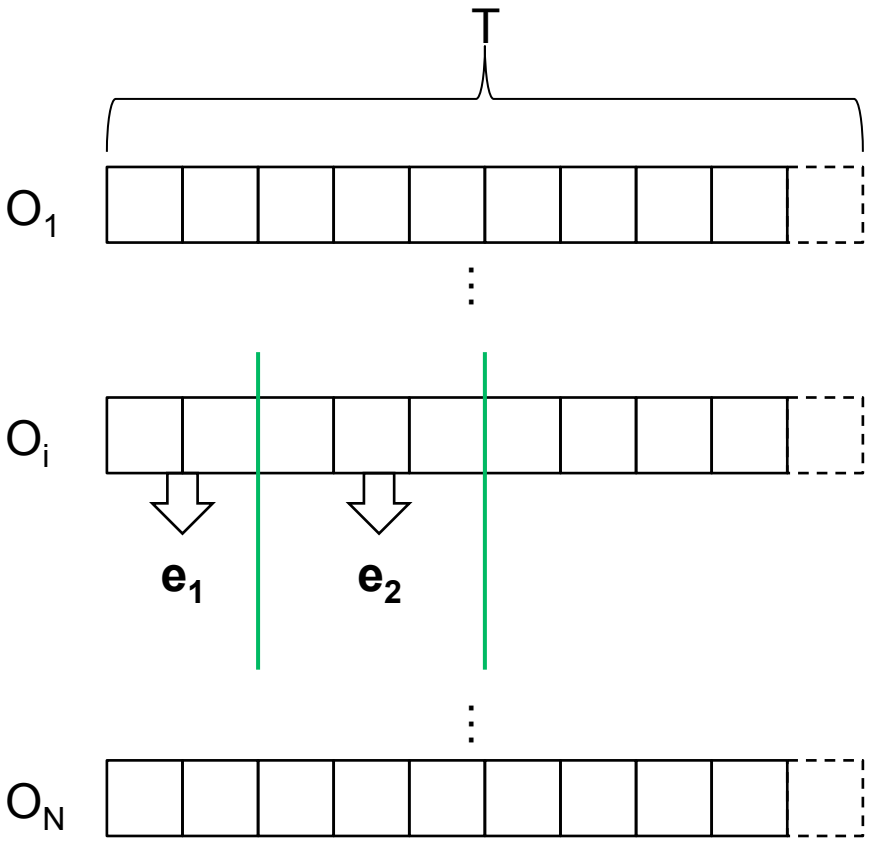
Normalization:

$$Sig^{SP}(Q_p^{ls}) = \frac{sed(Q_p^{ls}, Q_{p-1}^{ls}, Q_{p+1}^{ls})}{max_{sed}} \qquad Sig^{C}(Q_p^{cf}) = \frac{1 - scc(Q_p'^{cf}, Q_p^{cf})}{2max\{(1 - scc)\}}$$

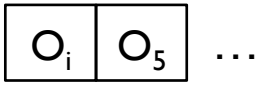
Simple combination:

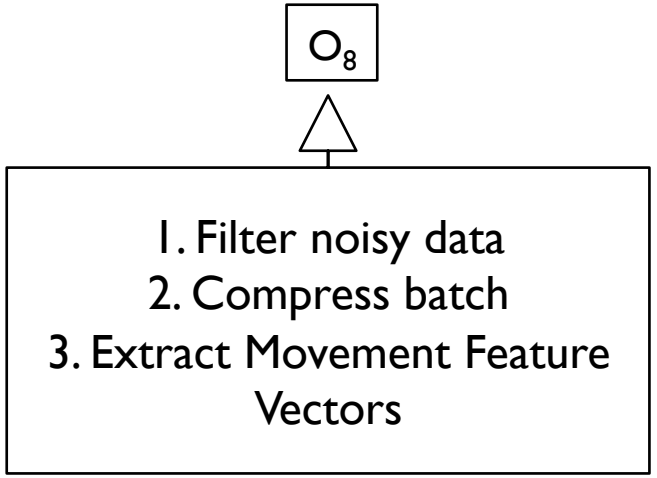
$$Sig(Q_p) = \frac{1}{2}(Sig^{SP}(Q_p^{ls}) + Sig^{C}(Q_p^{cf}))$$

SeTraStream - Feature Extraction



Buffer of incoming batches of objects (arriving every τ)





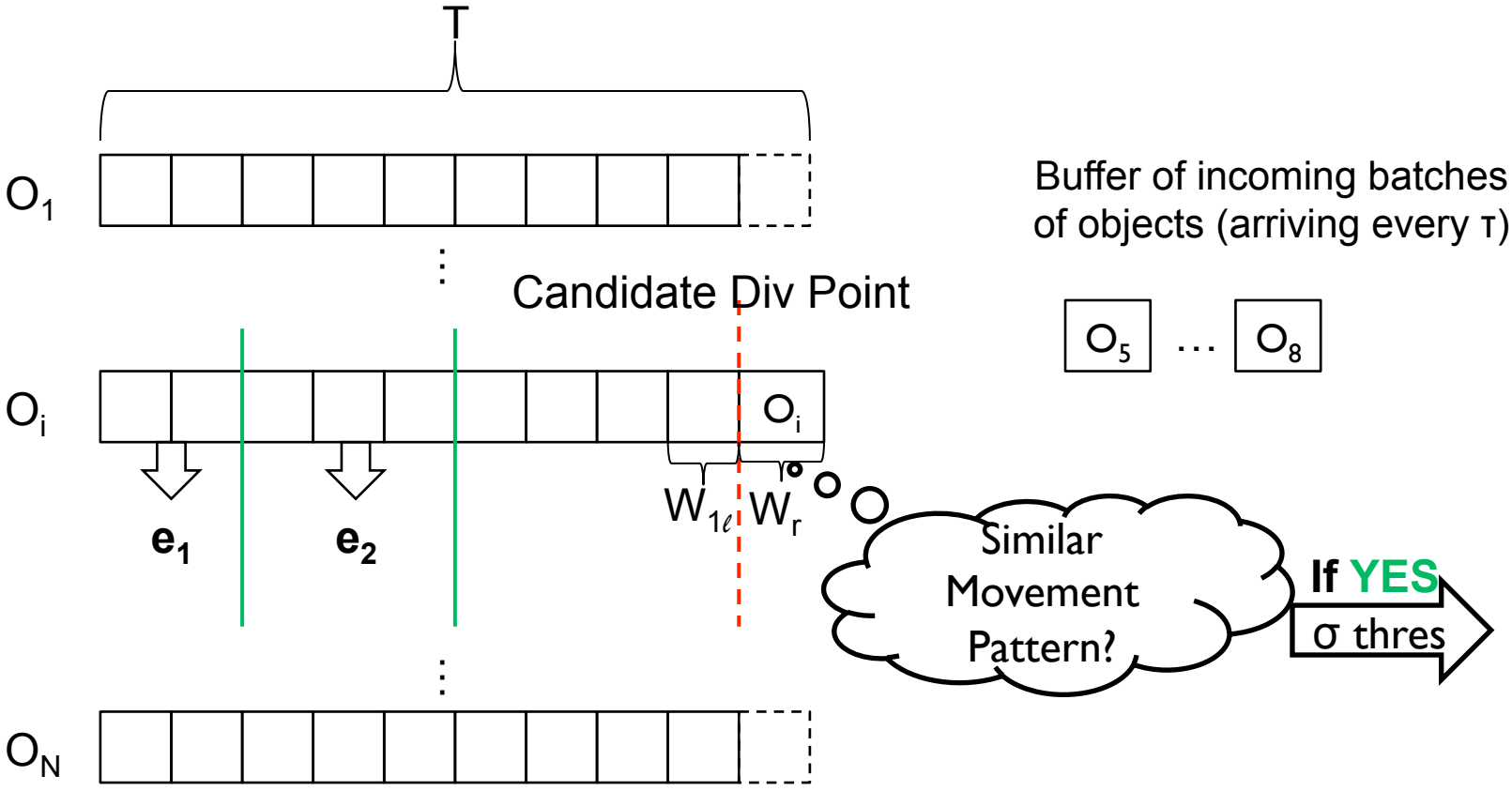
Movement Feature Vectors (MFVs)

<x,y,t>	Position in Lane	Distance to Headway Vehicle	Steering Wheel Activity
123.34, 121.21, 18:35:43	0.1m	1m	π/36
…	…	…	…
120.34, 125.21, 18:36:59	0.05m	3m	π/16

speed	direction	acceleration
35 m/s	76°	40 m/s^2
…	…	…
60 m/s	85°	55 m/s^2

MFVs in Batch make up a Matrix		
35	…	60
76	…	85
40	…	55
0.1	…	0.05
1	…	3
π/36	…	π/16

SeTraStream - Segmentation



Which types of similarity measurement?

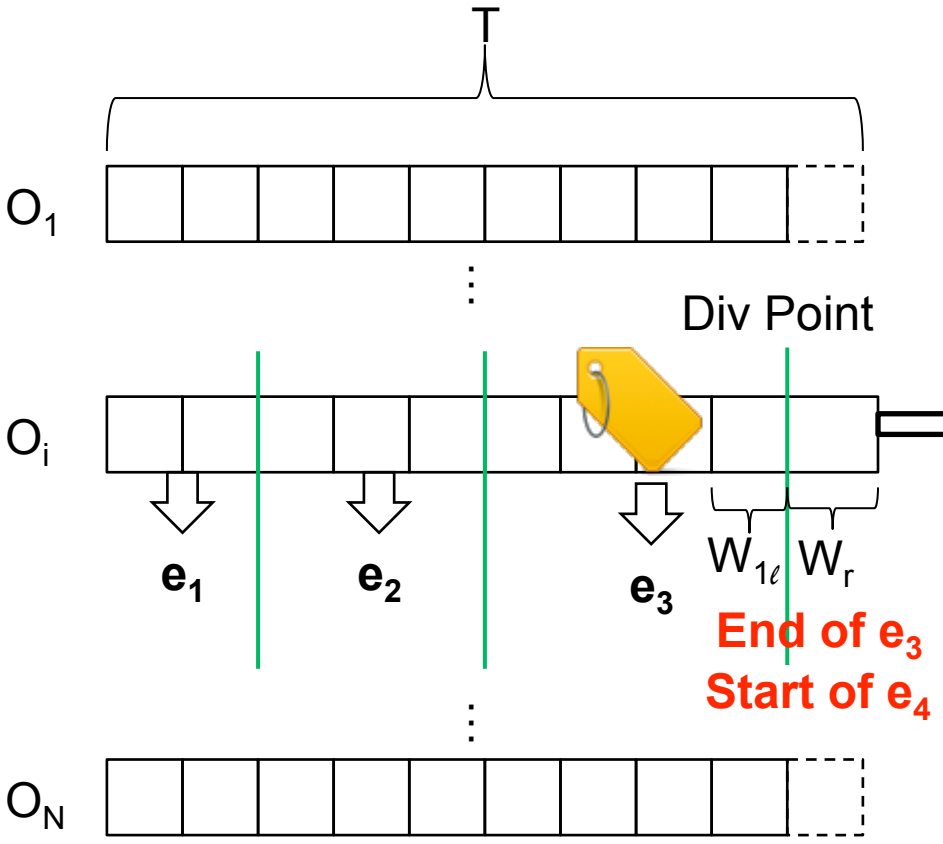
Movement Similarity

- Existing trajectory computing:
  - Offline, thresholds on movement features like velocity/direction/density
- Online solution:
  - Similarity on movement patterns (not individual attributes)
  - Threshold on movement pattern alteration

$$RV(W_\ell, W_r) = \frac{Tr(W_\ell W_\ell' W_r W_r')}{\sqrt{Tr([W_\ell W_\ell']^2) Tr([W_r W_r']^2)}}$$

- RV-coefficient:
  - A multivariate correlation coefficient, focusing on "trend" similarity; NOT on absolute differences
  - Measures the relative resemblance of two sequences of vectors
  - Dimension independent since $W_l W_l'$, $W_r W_r'$ possess d * d dimension – *d the number of features*

Short-term Movement Change



Buffer of incoming batches of objects (arriving every τ)

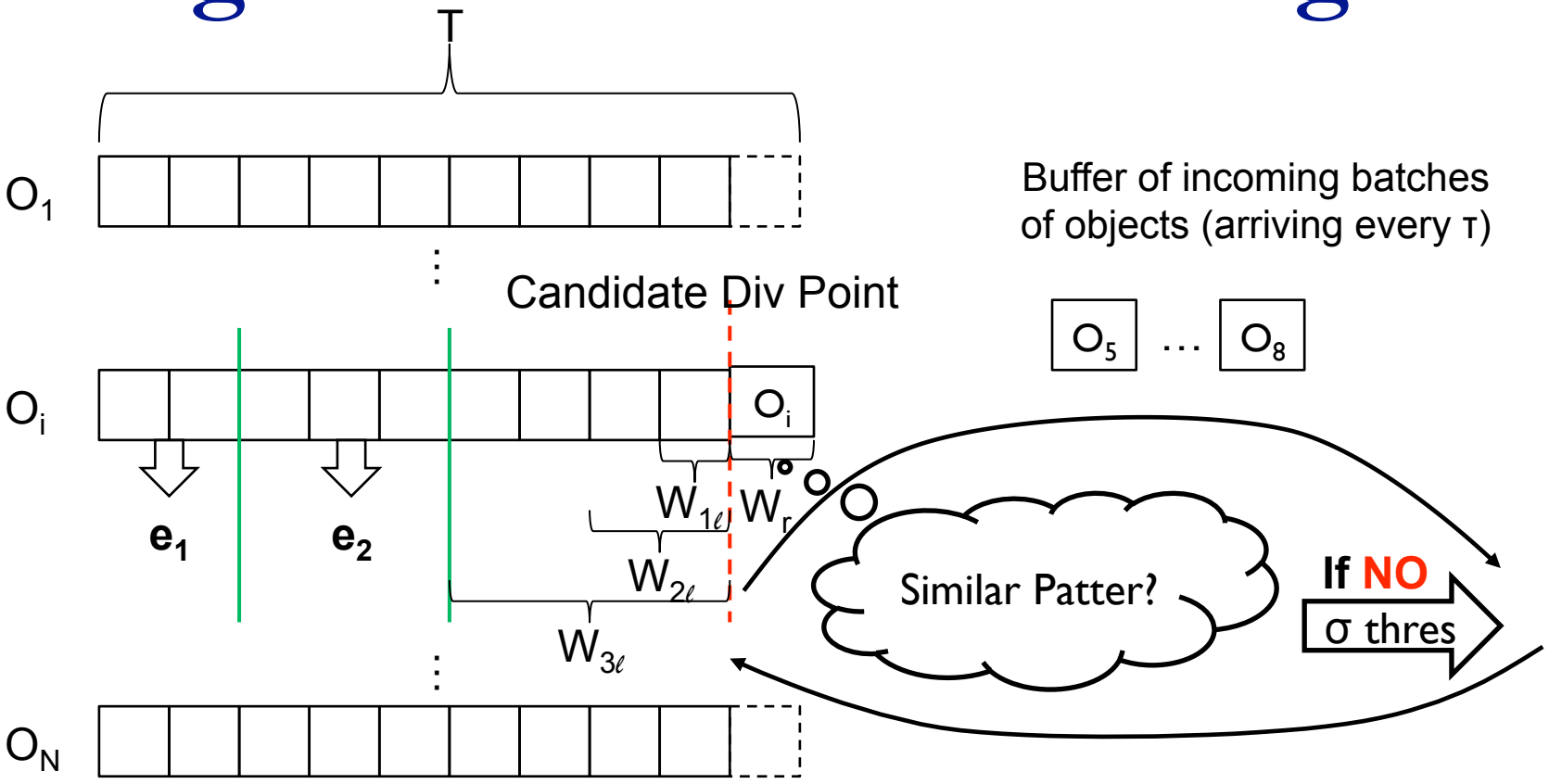
O_5 … O_8

As soon as we find an episode, we tag it

Tagging Episodes:

Training offline, tagging online

Long-term Movement Change



Similarity (W1, W2)
e.g. RV-coefficient (W1, W2)

$$RV(W_\ell, W_r) = \frac{Tr(W_\ell W_\ell' W_r W_r')}{\sqrt{Tr([W_\ell W_\ell']^2) Tr([W_r W_r']^2)}}$$

Outline

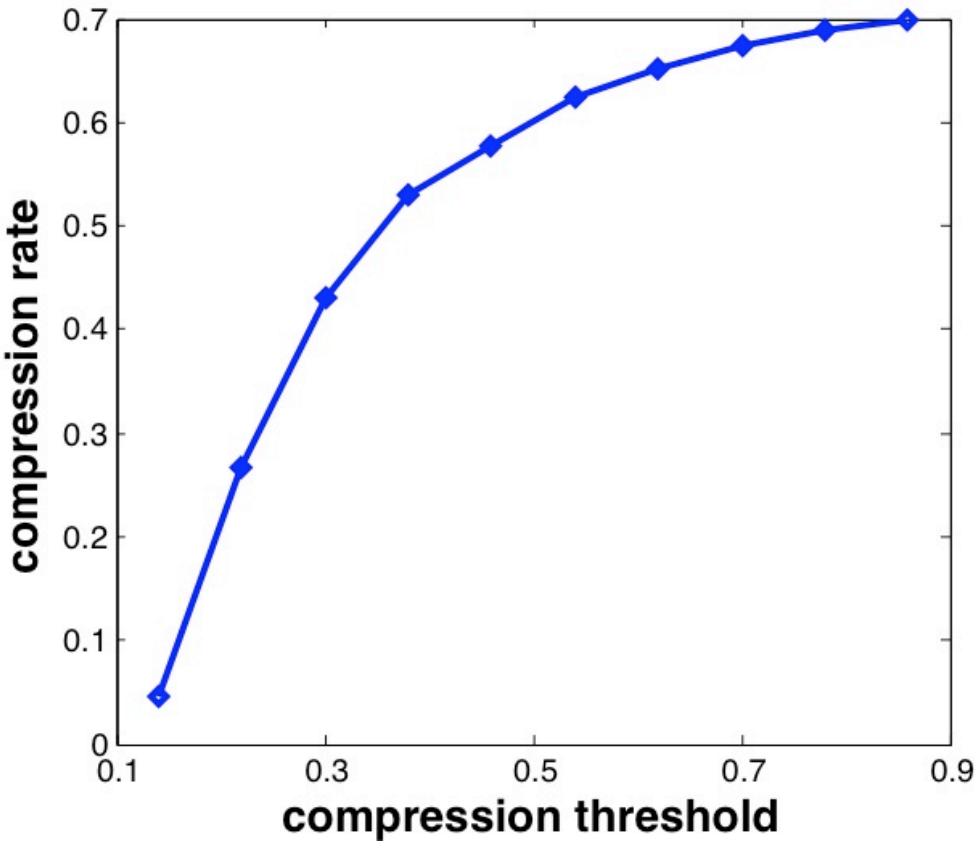
- Introduction
  - semantic trajectories…
  - …over streaming movement data?
- Related Work
- SeTraStream Framework
  - Big Picture
  - Details of each module
    - Data Cleaning
    - Data Compression
    - Segmentation – Episode Identification
- Experimental Evaluation
- Conclusions

Experiment - Dataset

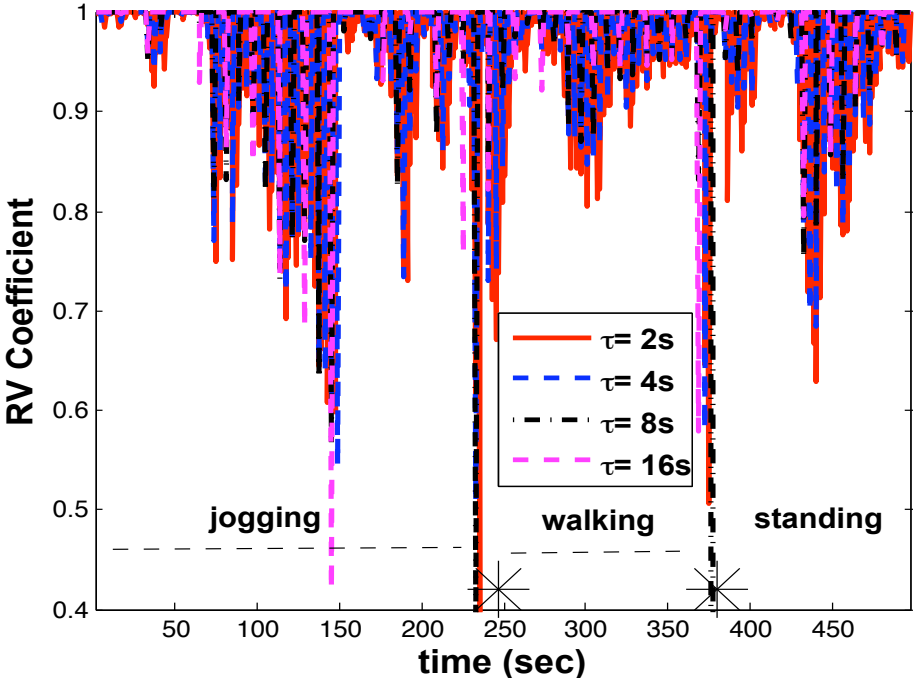
- GPS data from Nokia Research Center @ Lausanne
- User tags: home_cook, office_work, stand, jog, walk, bus ….

All dataset	user-id	*from-date*	*to-date*	*#days-with-gps*	*#GPS*
185 smartphone users	1	2009-02-17	2010-04-27	191	50,274
23,188 daily trajectories	2	2009-02-25	2010-05-16	330	200,418
7,306,044 GPS records	3	2009-09-14	2010-05-16	166	62,272
from date: 2009-02-01	4	2009-11-19	2010-05-16	161	66,304
to date: 2010-08-16	5	2009-12-18	2010-05-16	140	69,467
	6	2010-01-25	2010-05-16	89	45,137

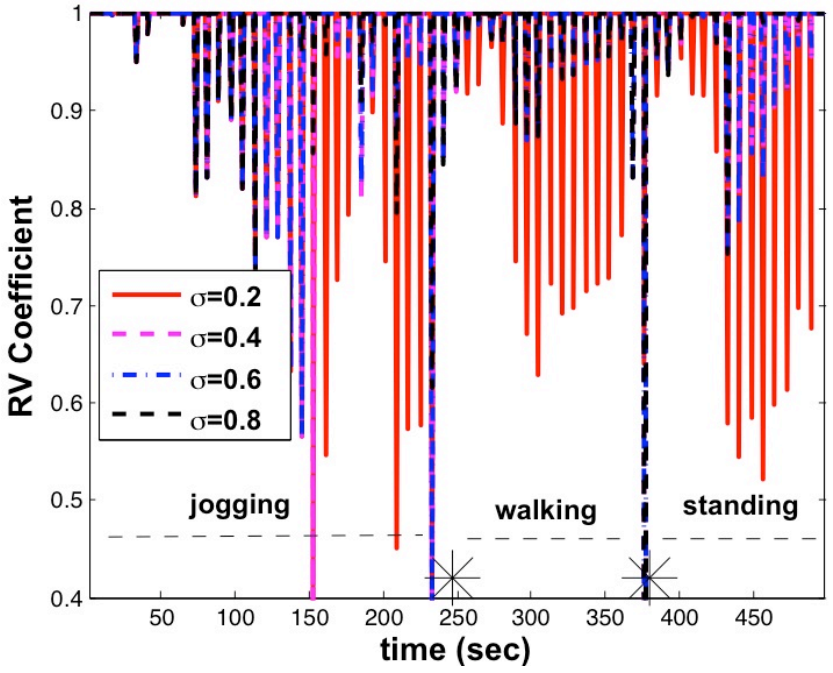
Experiment - Compression



Experiment - Segmentation



Different batch sizes



Different RV threshold

Experiment - Latency

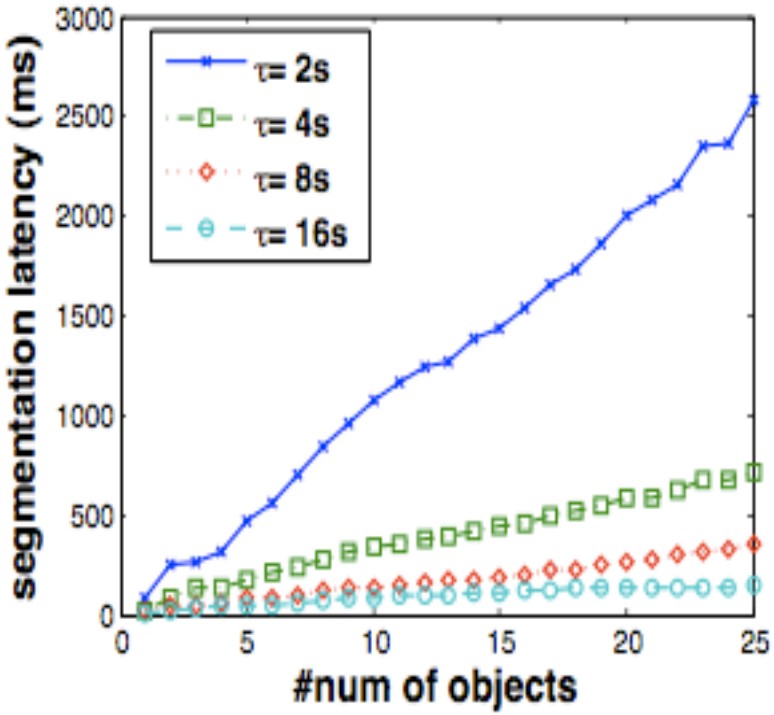


Fig. 7: Segmentation latency with different τ sizes (σ=0.6)

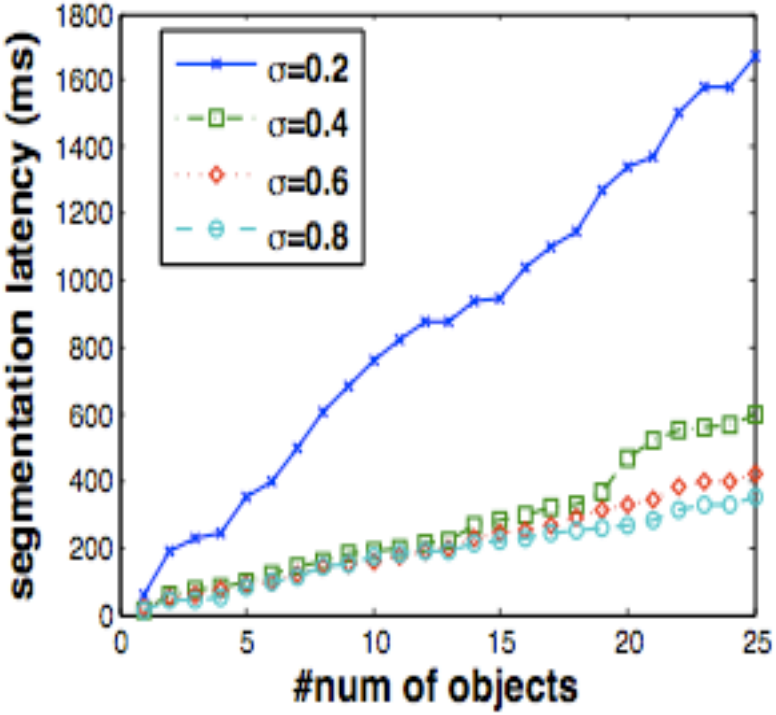


Fig. 8: Segmentation latency with different σ thresholds ($\tau = 8s$)

Outline

- Introduction
  - semantic trajectories…
  - …over streaming movement data?
- SeTraStream Framework
  - Big Picture
  - Details of each module
    - Data Cleaning
    - Data Compression
    - Segmentation – Episode Identification
- Experimental Evaluation
- Related Work
- Conclusions

Conclusion and Future Work

- We developed SeTraStream
  - Online Semantic Trajectory Construction
  - Complete Framework
    - Data Cleaning, Load Shedding, Trajectory Segmentation, Tagging
  - To our knowledge, the first work tackles with semantic trajectories in the context of streaming movement data
- Future Work
  - Explore new similarity measurement (rather than RV-coefficients)
    - …and still allow W_1 expansion so as to seek for long term motion pattern changes (e.g. Sketch Summaries ?)
  - Further experimentation with larger datasets
  - Extensions to distributed settings: Local vs. global computation
    - Can any part of the computation be conducted locally?
    - Most likely only cleaning & load shedding can be done locally ☹

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Thank You!

Zhixian Yan* Nikos Giatrakos† Vangelis Katsikaros†
Nikos Pelekis† Yannis Theodoridis†

*Distributed Information Systems Lab
Swiss Federal Institute of Technology
(EPFL), Lausanne, Switzerland

† Information Management Lab
University of Piraeus,
Piraeus, Greece

12th International Symposium on Spatial and Temporal Databases

Minneapolis, MN, USA 26 August 2011