Utilizing In-Store Sensors for Revisit Prediction

Sundong Kim and Jae-Gil Lee

Korea Advanced Institute of Science and Technology

https://github.com/kaist-dmlab/revisit

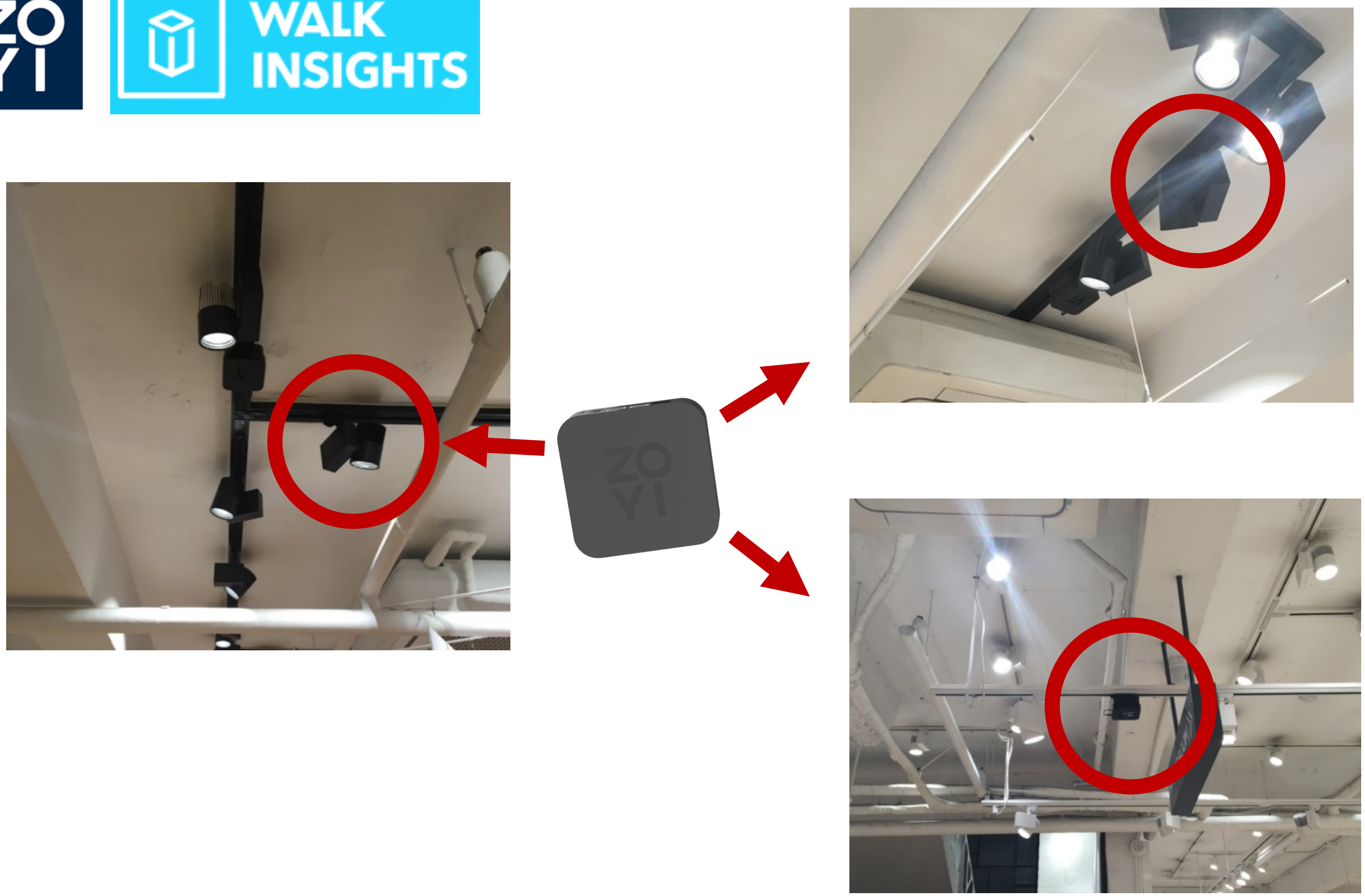
Presented in IEEE ICDM'2018

While You Are Shopping

Collecting Data with Wi-Fi APs

ZO YI

WALK INSIGHTS

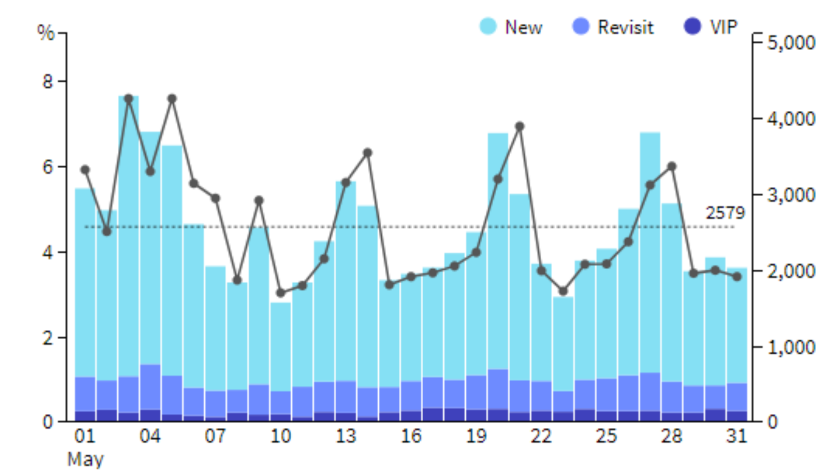


Retail Analytics

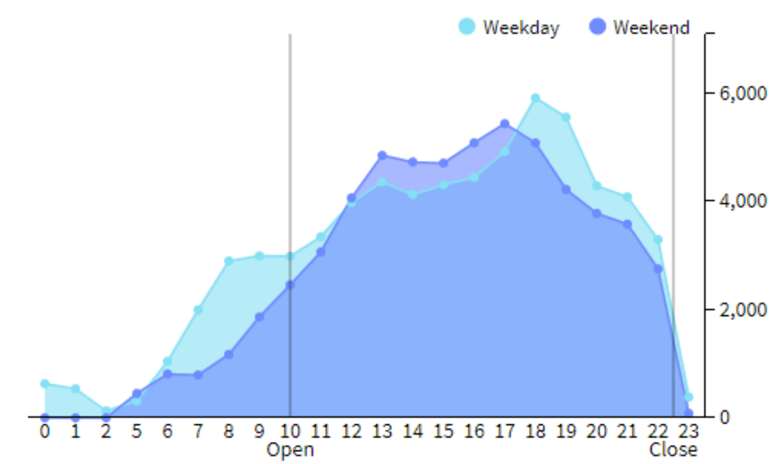
- Provide a dashboard, as well as consultancy services
- Data-driven monitoring examples:

DASHBOARD

Visitors



Traffic by Hour



Related Works & Our Study

Indoor Tracking
- Interior design, Museum
- Visitor locations → Measure interests → Display plan

Prediction (x)

Predictive Analytics
- Next location
- Customer life-time value
- Churn (On-line)

Off-line Revisit (x)

Revisit Studies
- Marketing, tourism
- Questionnaire
- Qualitative Factors

Mobility (x)

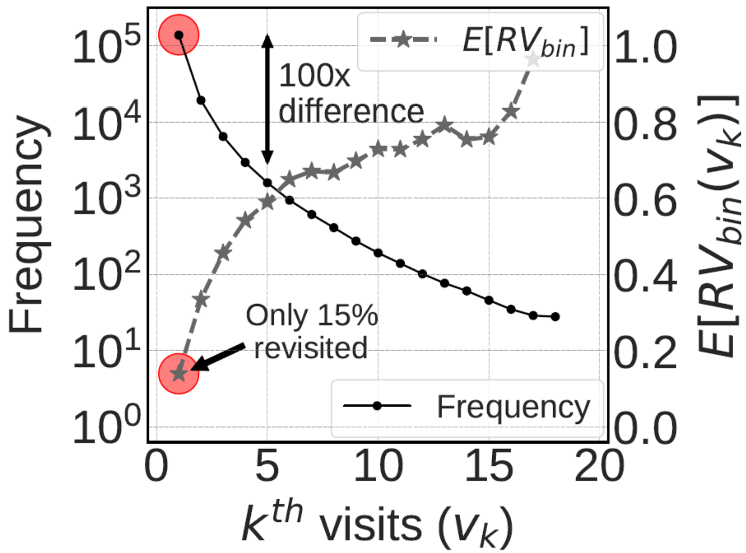
To Discover the Relation between

Customer Revisit and their Mobility

Revisit in Offline Market

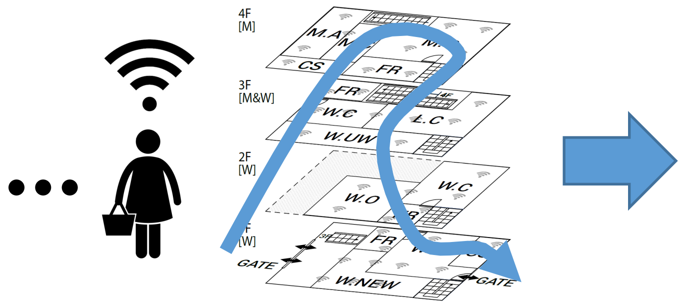
- **More than 85%** of retail purchases still happen **offline**. [Link]
- **Retaining customer** is very important. [Article]
  (5% more retention → 25-95% more profit)

↓ More than 70% of visits are first-time visits



← Rate of revisit

Research Questions



Revisit or Not

RQ1: How to predict customer revisits?

→ **Using a GBT model with carefully designed features.**

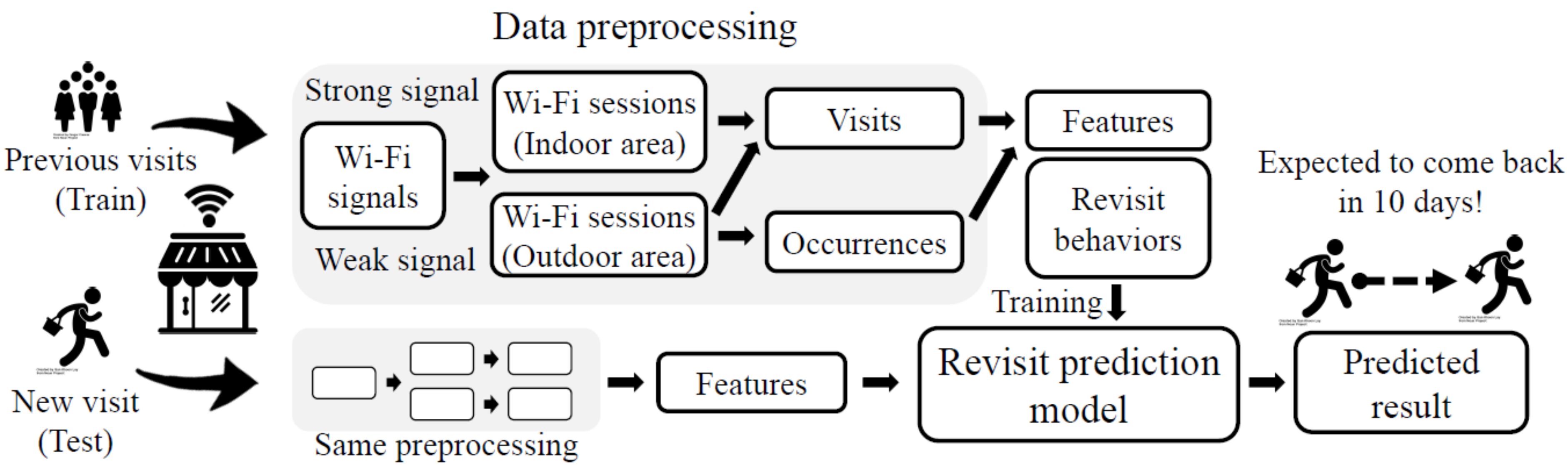
RQ2: How much effect of customer mobility has on prediction performance?

→ **Accuracy improves by 5-12% compared to LBs.**

Outline

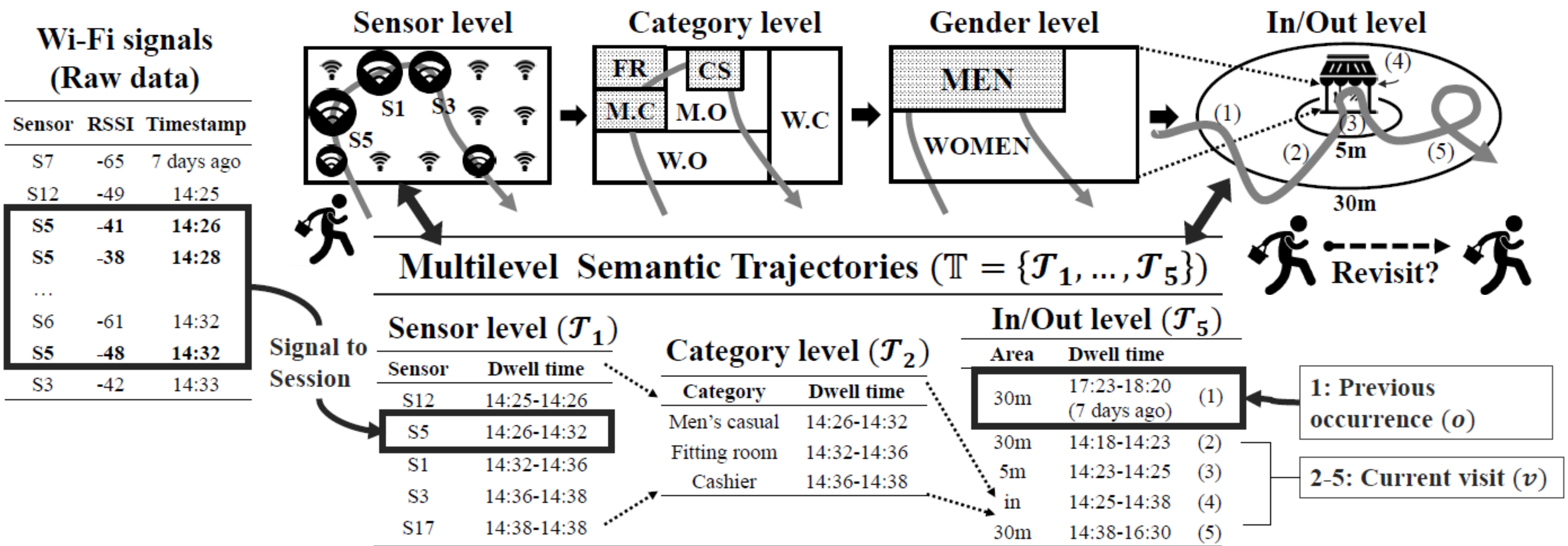
- Introduction
- **Prediction Framework <<**
- Features
- Performances
- Conclusion

Our Framework



Multi-level Trajectories

- Multi-level descriptions of the customer visit



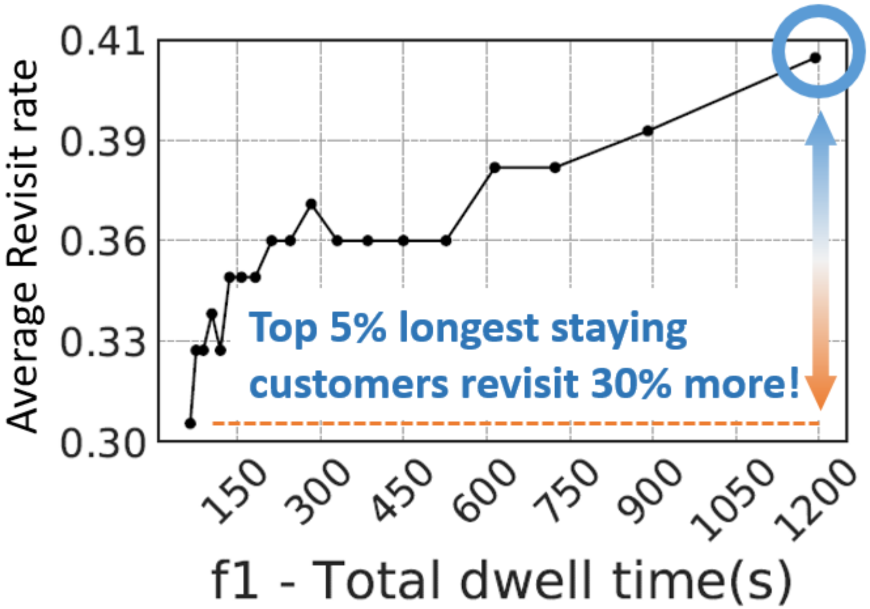
Outline

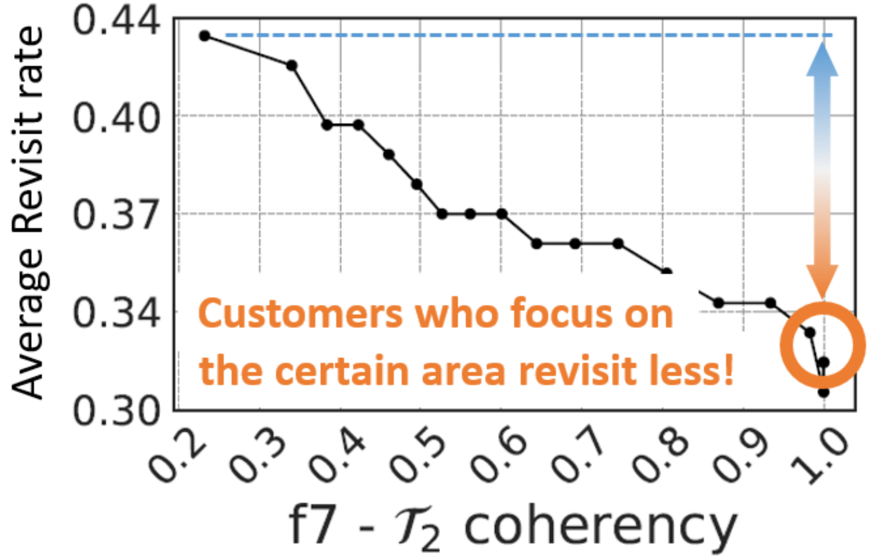
- Introduction
- Prediction Framework
- **Features <<**
- Performances
- Conclusion

Feature Engineering

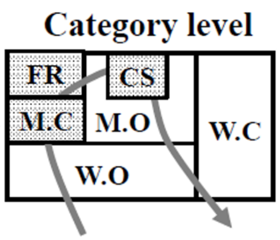
- Considered feature groups:
  - Overall statistics
  - Travel distance/speed/acceleration
  - Area preference
  - Entrance and exit pattern
  - Heuristics
  - Statistics of each area
  - Time of visit
  - Upcoming events
  - Store accessibility
  - Group movement

Feature Value & Revisit Rate (1)

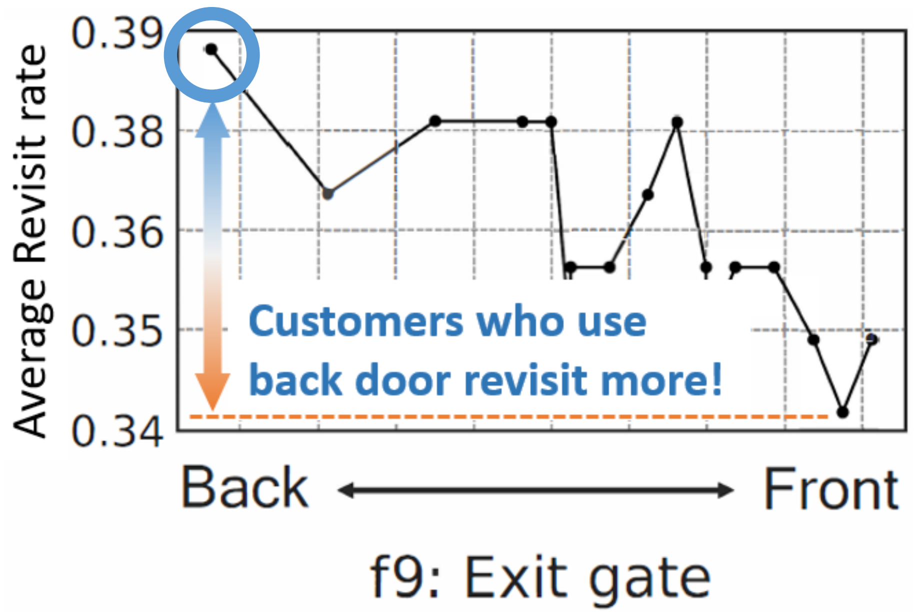


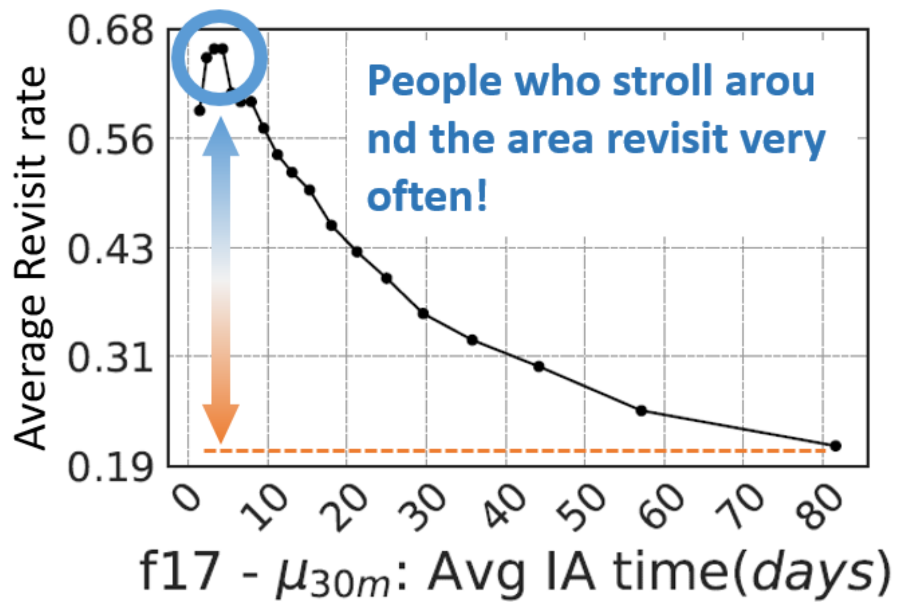


- $\mathcal{T}_2$ level:



Feature Value & Revisit Rate (2)



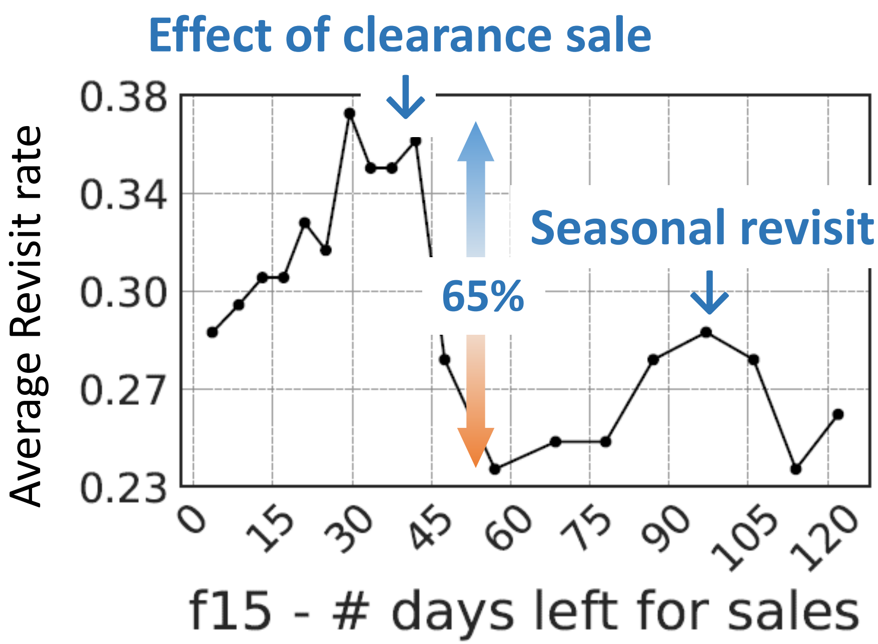


- AVG IA time: Average Inter-arrival time

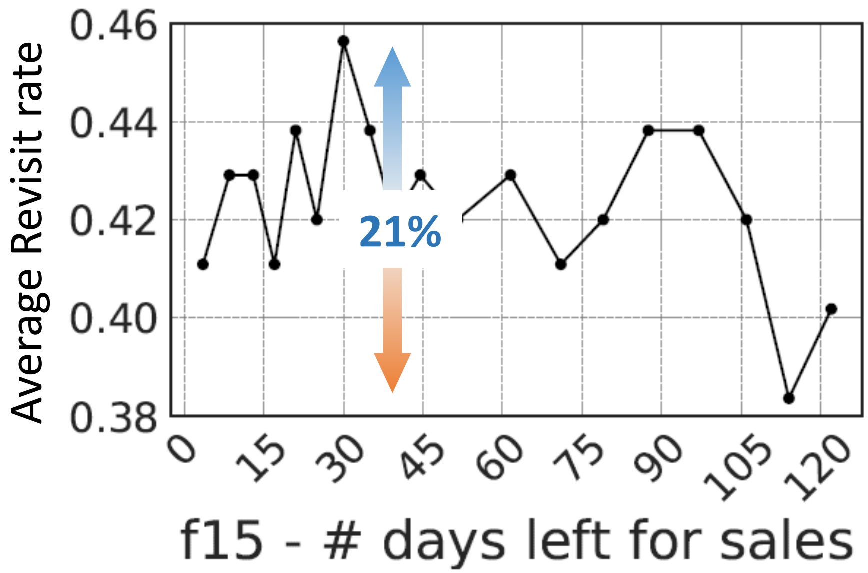
“Sale” For First-Time Visitors

<u>Number of days left for sales</u>:

- Feature with non-linear relationship



(a) First-time visitors: Prone to special events.



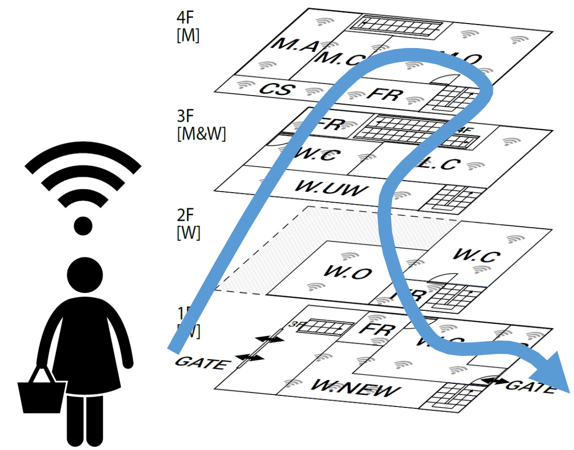
(b) All visitors: Indifferent to events.

Outline

- Introduction
- Prediction Framework
- Features
- **Performances <<**
- Conclusion

Mobility Data from In-Store Sensors

- 7 Flagship stores
- 220-990 days collected
- 110K-2M visits/stores
- Avg. # areas = 6.56

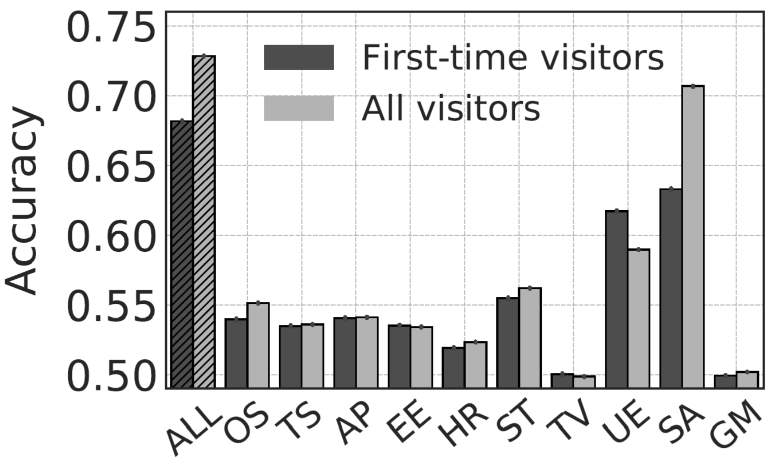


Shop ID	A_GN	A_MD	E_GN	E_SC	L_GA	L_MD	O_MD
Location	Seoul, Korea						
Length (days)	222	220	300	373	990	747	698
# sensors	16	27	40	22	14	11	27
Data size	15GB	77GB	148GB	99GB	164GB	242GB	567GB
# visits > 60s	0.11M	0.33M	0.18M	0.27M	1.06M	1.72M	2.01M
Revisit rate	11.73%	31.99%	21.18%	36.55%	21.22%	32.98%	48.73%

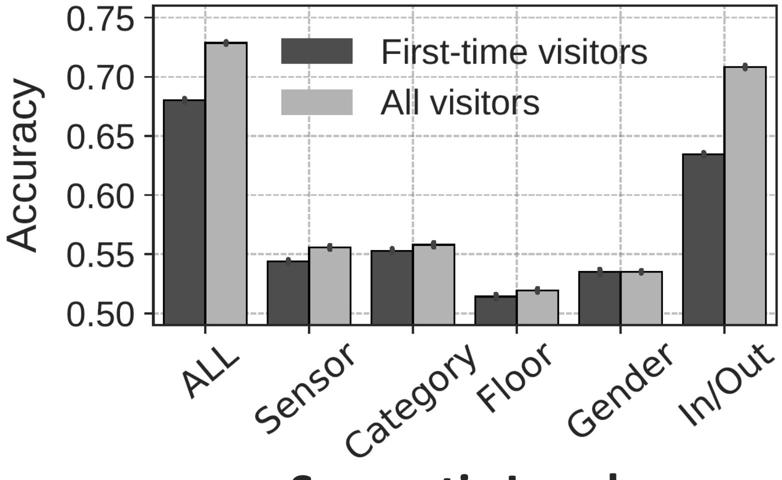
Results: Prediction Accuracy

Store	Accuracy (First)	Accuracy (All)
A_GN	0.6336	0.6689
A_MD	0.6930	0.7412
E_GN	0.6663	0.7050
E_SC	0.6818	0.7288
L_GA	0.7173	0.7789
L_MD	0.6799	0.7991
O_MD	0.6645	0.7599

Accuracy of 7 stores using a XGBoost Classifier.



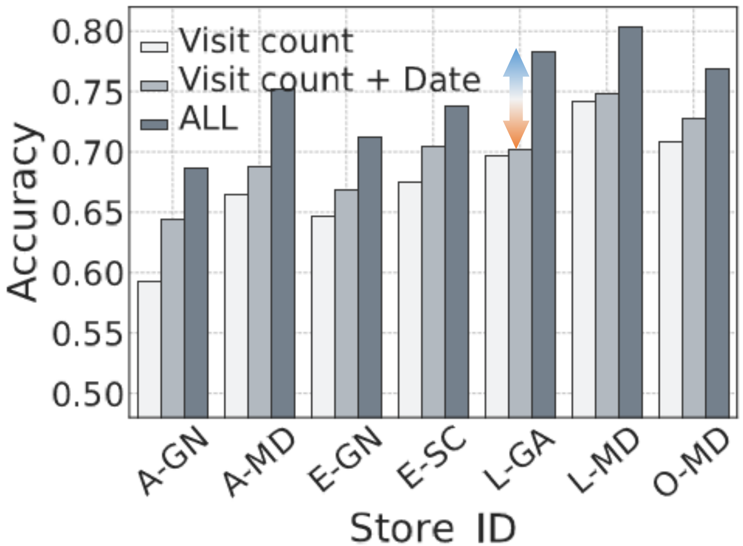
Feature Group.



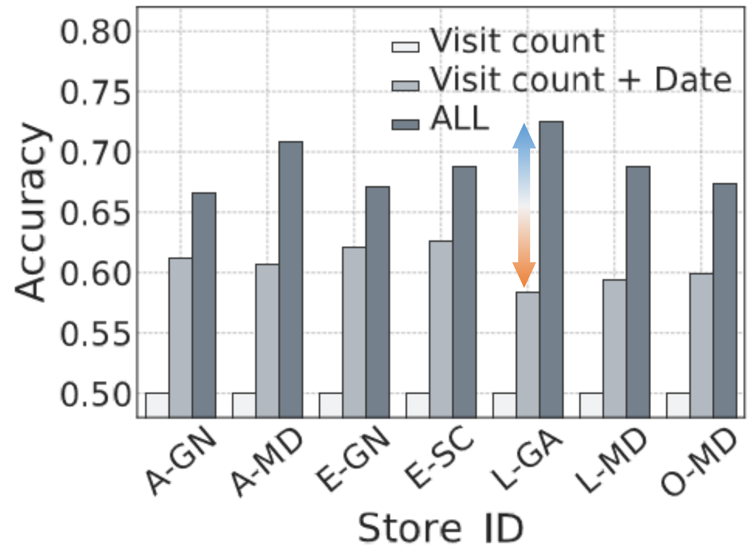
Semantic Level.

Effectiveness of the Feature Set

- Baseline 1 (LB): By only knowing the **number of visits**
- Baseline 2: By knowing the **number of visits** & **date of the visit**
- By utilizing features derived from **Wi-Fi signals**, we achieved **significant** performance improvement on **revisit prediction**.



(a) On all visitors



(b) On first-time visitors

Data Collection Period

- To find the **right amount of data** to study revisit
- To maintain sensors for **securing enough profit**

→ **Find the minimum sufficient** amount of data T to predict revisit without accuracy loss

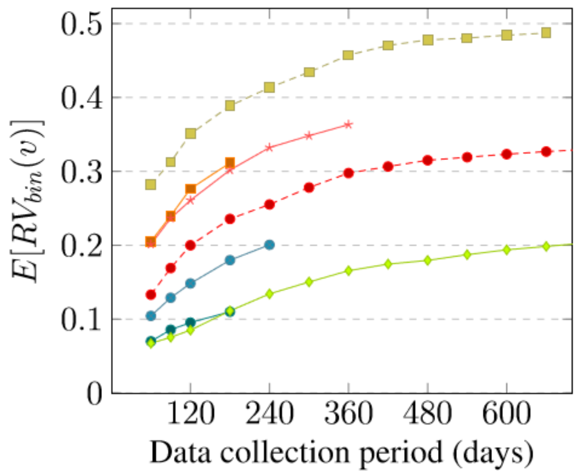
Data Length

Short ⟷ **Long**

Low cost & Small effort

Less evidence

Changes in revisit rate



More evidence

Closer to the steady state

Capture most of revisits

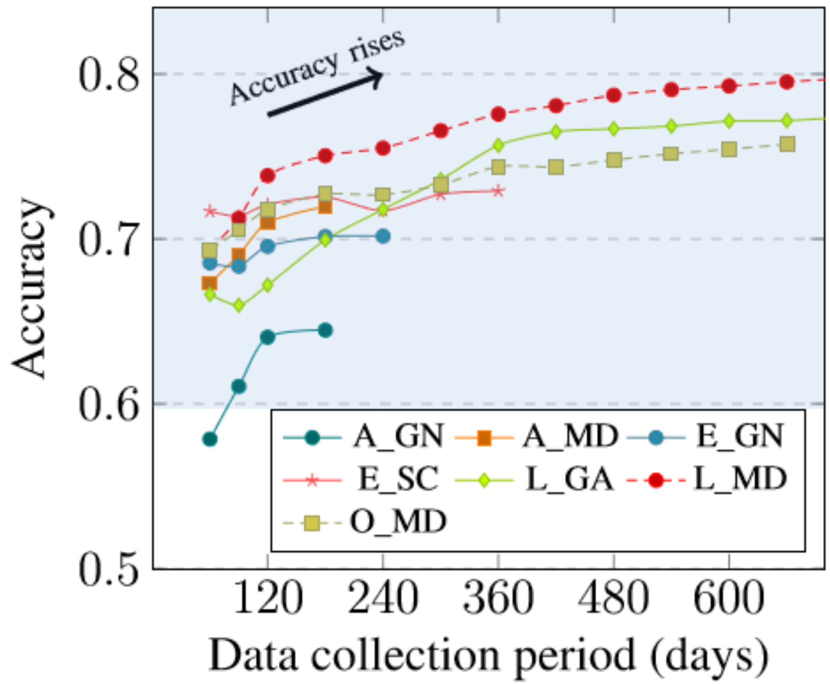
Hard to persuade clients

Impact on Prediction Accuracy

1) On all visitors

∴ # Regular customers ↑

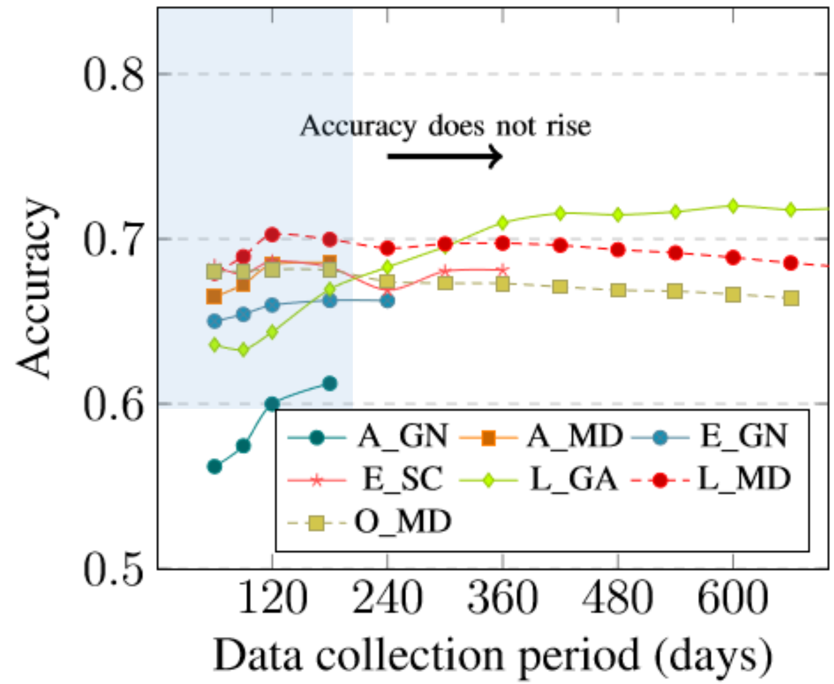
∴ Accuracy **gradually increases.**



2) On first-time visitors

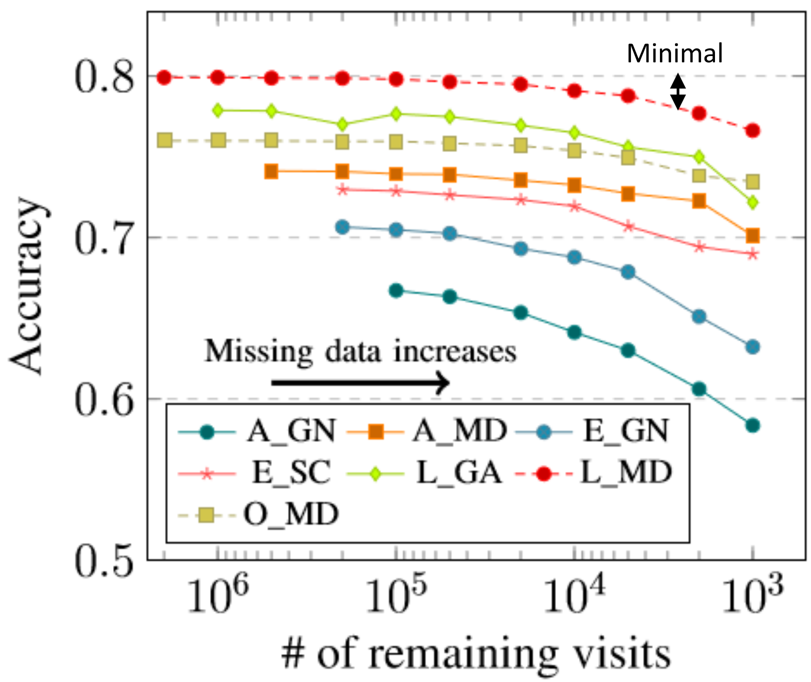
∴ Change of floor plans.

∴ **Increases** depends on the store.

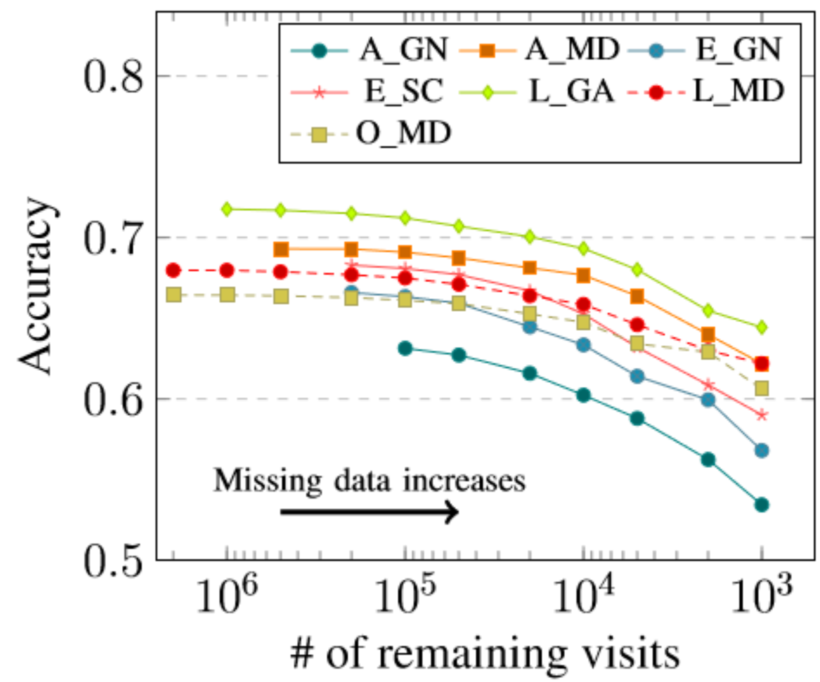


Robustness on Missing Customers

- Over 95% of the performance is maintained with a very small fraction of the dataset (e.g., **0.5%** for L_MD)



(a) On all visitors



(b) On first-time visitors

Real Behavior vs. Collected Data

Significantly different!

Actual ratio	Wi-Fi status	Probability of each state	Observed as	Observed ratio (data)
Group ($p_y = 56\%$)		$p_y p_x{}^2$	**Group**	Group ($p_{yo} = 15.6\%$)
		$p_y p_x(1-p_x)$	Individual	
		$p_y(1-p_x)p_x$	Individual	
		$p_y(1-p_x)^2$	None	
Individual $(1-p_y)$		$(1-p_y)p_x$	Individual	Individual $(1-p_{yo})$
		$(1-p_y)(1-p_x)$	None	

The value to be observed in the data

p_x = Wi-Fi turn on rate (39.2%)

p_y: Ratio of customers with companion (Observed at the spot)

p_{yo}: Ratio of customers with companion (To be observed in the data)

Conclusions

- **Goal:** To discover the relation between **customer revisit** and their **mobility**
- **Data:**
  - Customer mobility data captured in seven stores
- **Findings:**
  - Prediction models using handcrafted features
  - Predictive powers of each feature groups
  - Performance improvement by utilizing indoor trajectories
  - Predictive powers by collecting longer period
  - Robustness on missing data

Thank you!

https://github.com/kaist-dmlab/revisit

Scan me for details ☺

(Paper, Slides, Datasets, Tutorial)

Detailed Feature List

- Systemically generated features:

Data sources	Data sources (low-level)	Feature groups	Twenty representative features (Among 866 features of store A)	Semantic level of features					
				Sensor	Category	Floor	Gender	In/Out	None
Moving pattern of the visit (Sec. 4.2)	From the entire trajectory	Overall statistics **[OS]** (4.2.1)	f1 = Total dwell time					✓	
			f2 = Trajectory length	✓	✓	✓	✓		
			f3 = Skewness of dwell time of each area	✓	✓		✓		
		Travel Distance/ Speed/Acceleration **[TS]** (4.2.2)	f4 = Total distance traveled in the store		✓				
			f5 = Speed based on transition time	✓	✓	✓	✓		
			f6 = First-k HWT coefficients of acceleration	✓	✓	✓	✓		
		Area preference **[AP]** (4.2.3)	f7 = Coherency of dwell time for each level		✓	✓	✓		
			f8 = Top-k-area dwell time	✓	✓	✓	✓		
		Entrance and Exit pattern **[EE]** (4.2.4)	f9 = Exit gate	✓					
			f10 = Daily visit count of the customer					✓	
		Heuristics **[HR]**	f11 = Wears clothes but does not buy		✓				
	From the subsequence	Statistics of each area **[ST]** (4.2.6)	f12 = Number of time sensed in the area	✓	✓	✓	✓		
			f13 = Stdev of dwell time for the area	✓	✓	✓	✓		
Temporal information of the visit (4.3)	From the time of visit and events calender	Time of visit **[TV]**	f14 = Day of the week						✓
		Upcoming events **[UE]** (4.3.2)	f15 = Remaining day until the next sale						✓
			f16 = Number of holidays for next 30 days						✓
Occurrences before the visit	From access intervals	Store accessibility **[SA]** (4.4)	f17 = Number of days since the last access					✓	
			f18 = Average interarrival time					✓	
Simultaneous visits	From the entrance and exit time	Group movement **[GM]** (4.5)	f19 = Presence of companions					✓	
			f20 = Number of companions					✓	