New Advances in Spatial Trajectory Analytics

Xiaofang Zhou

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

+ A Personal Journey

- 1994 – 1999 **CSIRO** Spatial Information Systems
  - SIRO DBMS used widely mainly to manage land and utility information
  - Worked with Dave Abel, Beng Chin Ooi, Kian-Lee Tan and Volker Gaede
  - Main focus: developing fast spatial join algorithms, spatial data sharing platforms and GIS applications for customers
- 1999 – now **University of Queensland**
  - Initially supported by Queensland State Govt on moving objects: green turtles!
  - Beijing taxi data made a big difference (~2008)
  - Worked with many people here
  - Main focus: trajectory analytics for the last 10 years

Trajectory Data

…data about moving objects

+ What is Spatial Trajectory Data

- Any data that record the locations of a moving object over time in a geographical space

- Simple form:

  $<ID, (p_1,t_1), (p_2,t_2) \ldots (p_n,t_n)>$

  ordered by time: $t_1 < t_2 < \ldots < t_n$

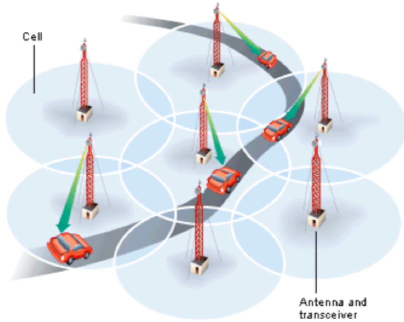
- General form:

  $<oID, tID, (p_1,t_1,a_1), (p_2,t_2,a_2) \ldots (p_n,t_n,a_n)>$

+ Where Trajectory Data Come From?

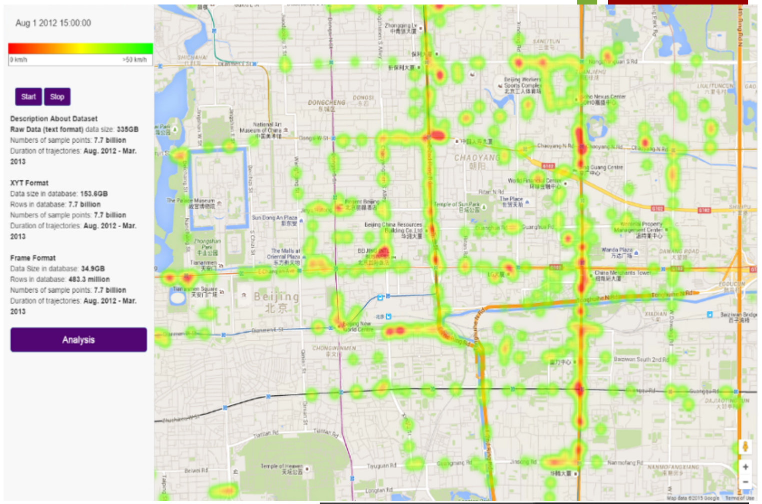
+ Massive Amount of GPS Data

+ Other Types of Trajectory Data

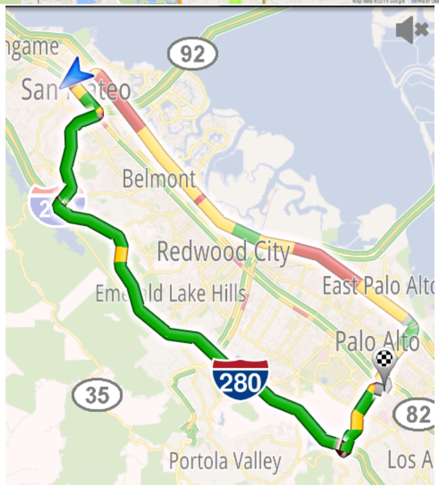


SENSORS

+ Trajectory Data is Useful

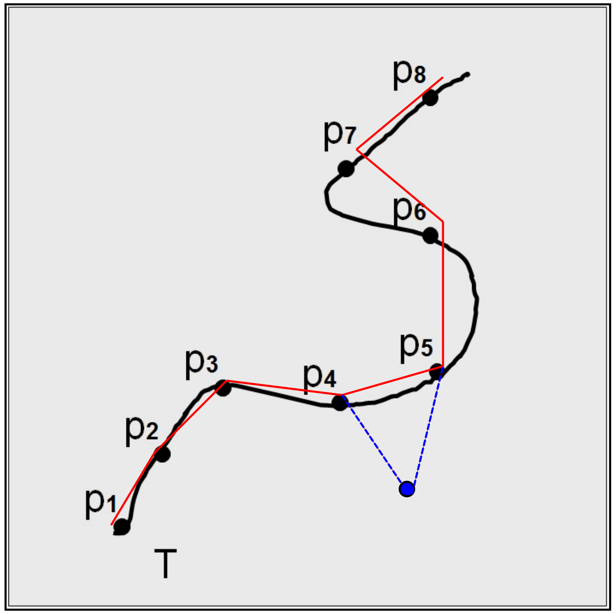


- Route planning
- POI recommendation
- LBS and advertisement
- Resource/object tracking and scheduling
- Intelligent transport systems
- Emergency responses
- Urban planning and smart cities…



+ Trajectory Data is Hard to Process

- Volume, velocity and variety…
- A trajectory is obtained from sampling the movement of an object
  - Some sampling strategies are used → not only data, but also models to generate data
  - Objects movement with constraints (e.g., by map) → not only data, but also environment data
  - There are many other factors which cannot be controlled → data quality issues
  - Data can be both redundant as well as sparse → compression, alignment and prediction
- It is non-trivial even to restore the original trace from a trajectory → harder to compare → much harder to use



+ Moving Objects/Trajectory Work

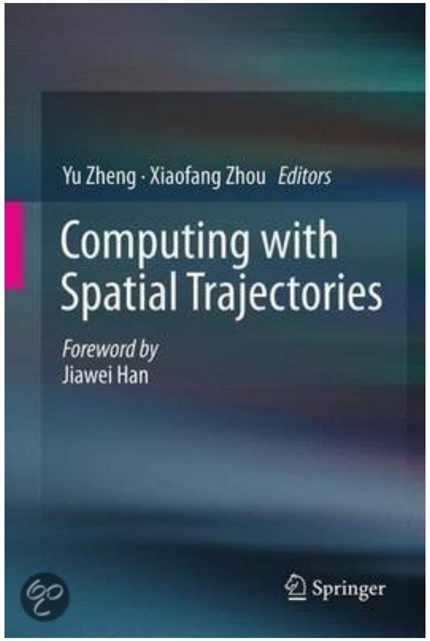
- Initially on foundations
  - Data representation, query languages and basic operations, indexing methods etc.
- Curiosity-driven
  - Imagine a special “novel” type of query, find a “novel” indexing method and then use “standard” methods to improve efficiency
- Not directly useful
  - Strong assumptions (not useful in practice)
  - Highly specialized indexes (cannot be implemented)
- Also active in other areas
  - Data mining, social networks, recommender systems…

+ Our Trajectory on Trajectories

Movement and path prediction [ICDE08, VLDBJ10], trajectory clustering [VLDB08], advanced spatial queries [SIGMOD09, SIGMOD10, VLDB17, ICDE19], most popular routes [ICDE11], probabilistic range query [EDBT11, ICDE12], materialized shortest paths [TODS12], spatial keyword search for trajectories [ICDE13,15,16, 19, TKDE19], trajectory calibration and repair [SIGMOD13, VLDBJ15, EDBT18], route and location recommendation [ICDE14, SIGKDD15, ICDE16, TOIS16, TIST18], trajectory summarization [ICDE15], routing algorithms [VLDB17, VLDBJ18, ICDE19], spatial crowdsourcing [2*TKDE19], in-memory trajectory databases [CIKM14, SIGMOD15], privacy-preserving trajectory search [ICDE15], data sparsity [MDM18], trajectory compression [TKDE19], ML for speed prediction [IJCAI18], tarjectory0based entity resolution [ICDE19], batch query processing [ADC 19, ICDE19]…

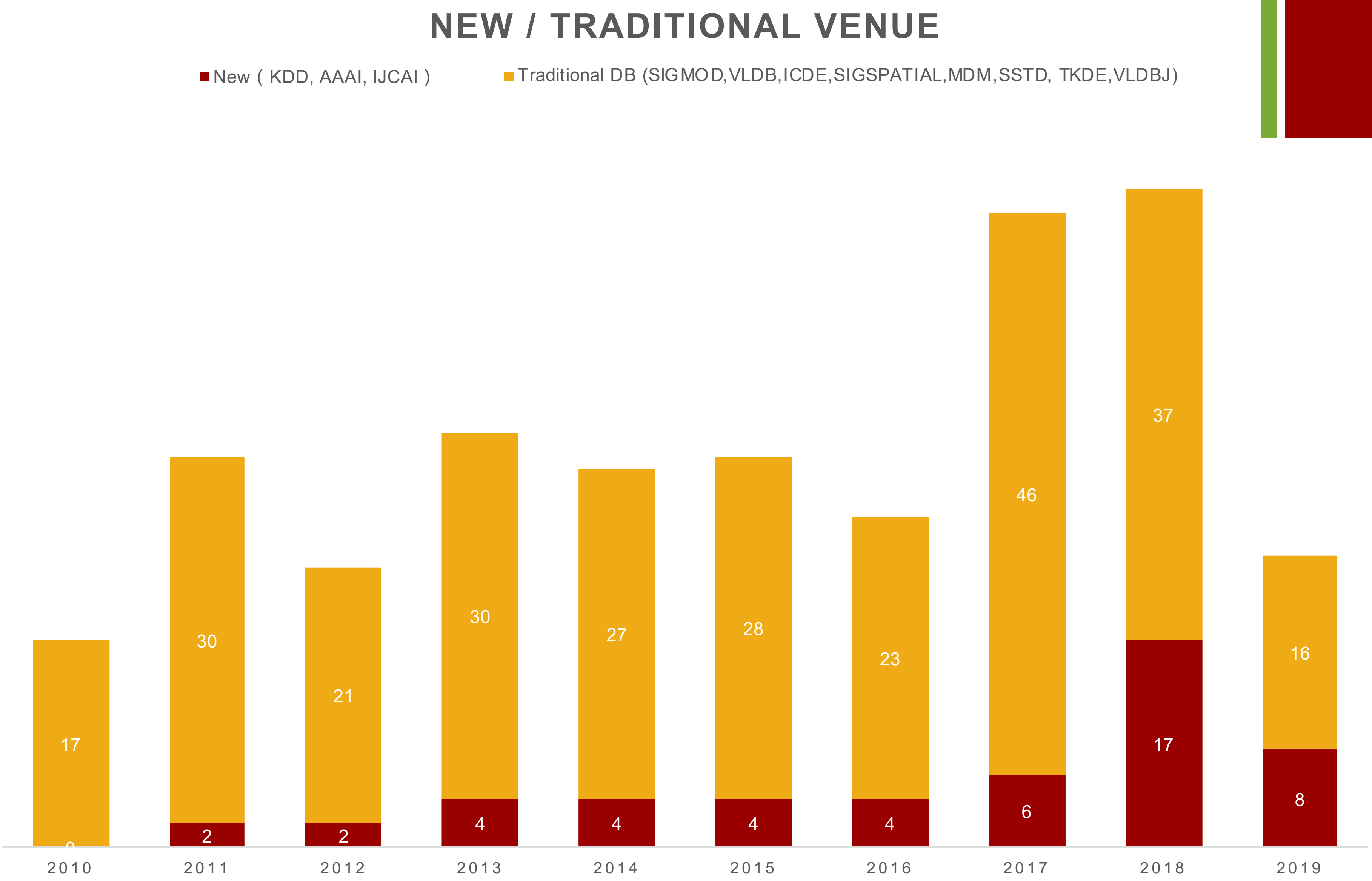
+ An Introduction Book

- ***Computing with Spatial Trajectories***
  - Yu Zheng and Xiaofang Zhou, 2011
- Part I Foundations
  - Trajectory Preprocessing *(W.-C. Lee, J.Krumm)*
  - Trajectory Indexing and Retrieval *(X. Zhou et al)*
- Part II Advanced Topics
  - Uncertainty in Spatial Trajectories *(G. Trajcevski)*
  - Privacy of Spatial Trajectories *(C.-Y. Chow, M. Mokbel)*
  - Trajectory Pattern Mining *(H. Young, K. L. Yiu, C. Jensen)*
  - Activity Recognition from Trajectory Data *(Y. Zhu, V. Zheng, Q. Yang)*
  - Trajectory Analysis for Driving *(J. Krumm)*
  - Location-Based Social Networks: Users *(Y. Zheng)*
  - Location-Based Social Networks: Locations *(Y. Zheng and X. Xie)*

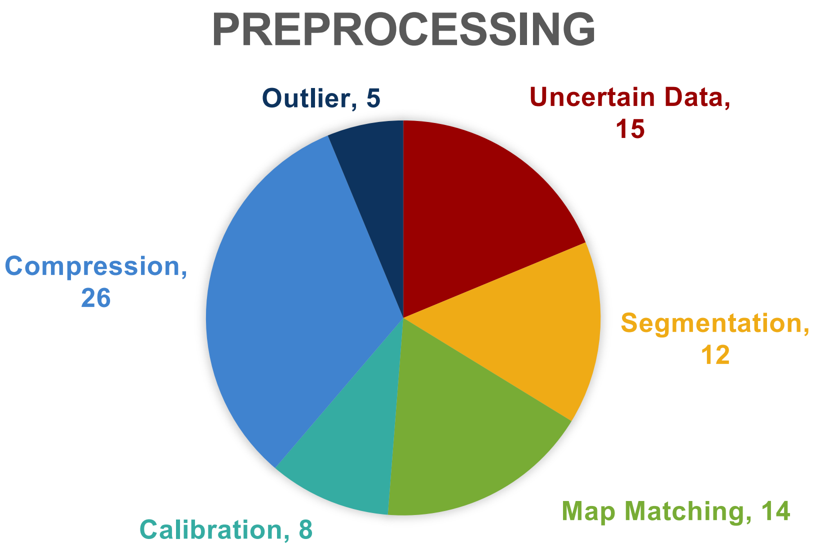


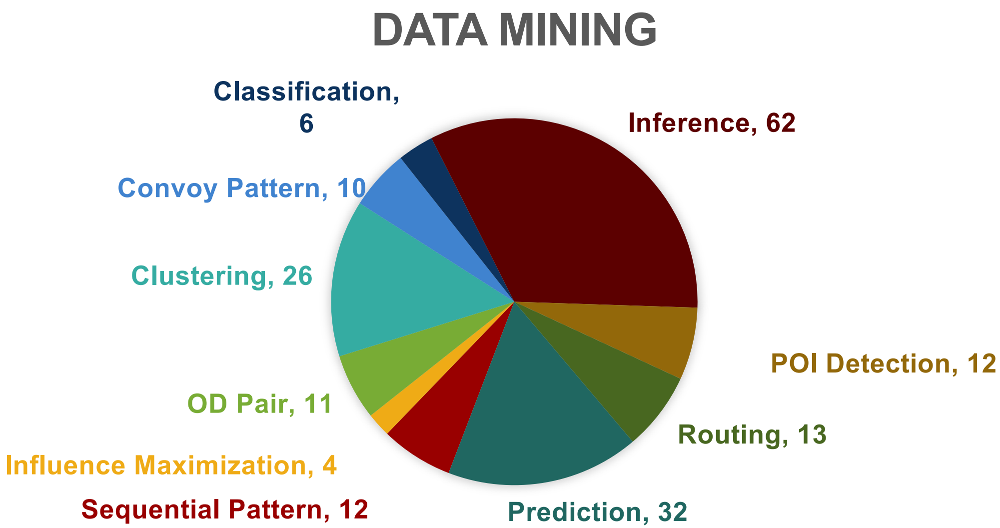
+ Popular Words

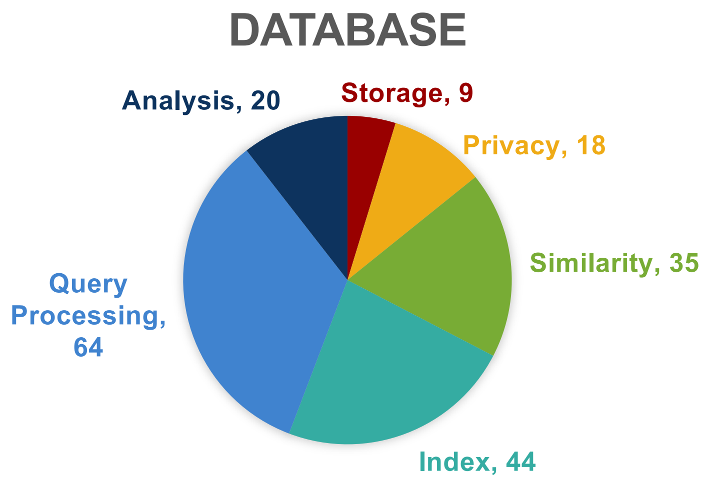
+ Paper Counts



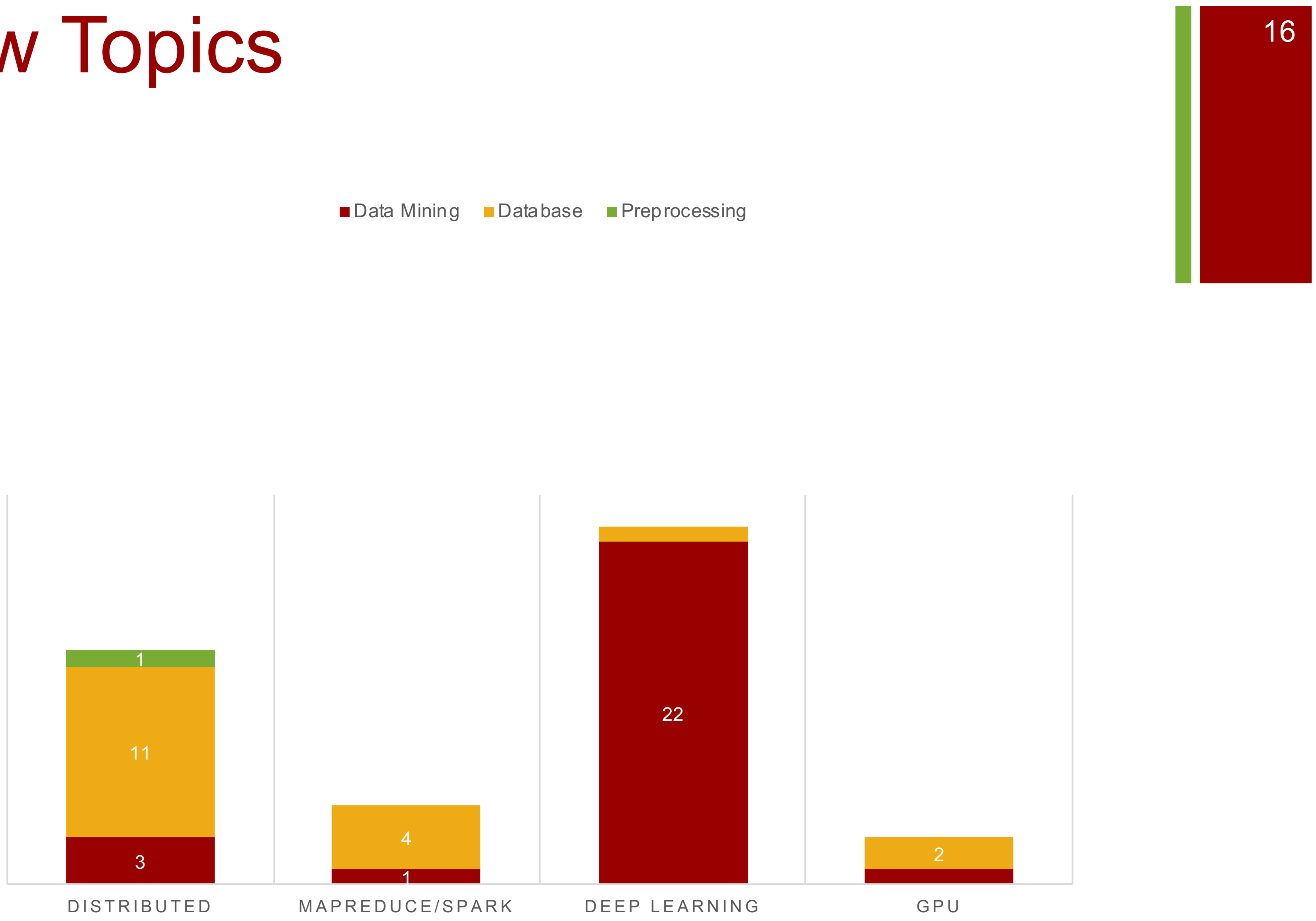
+ Traditional Topics







+ New Topics



Data Mining
Database
Preprocessing
16
1
11
3
4
1
22
2
DISTRIBUTED
MAPREDUCE/SPARK
DEEP LEARNING
GPU

+ Trajectory Data in a Company (2014)

- A car navigation service provider
- Total trajectory data: 32 TB in size, 10.9 billion matched trajectories

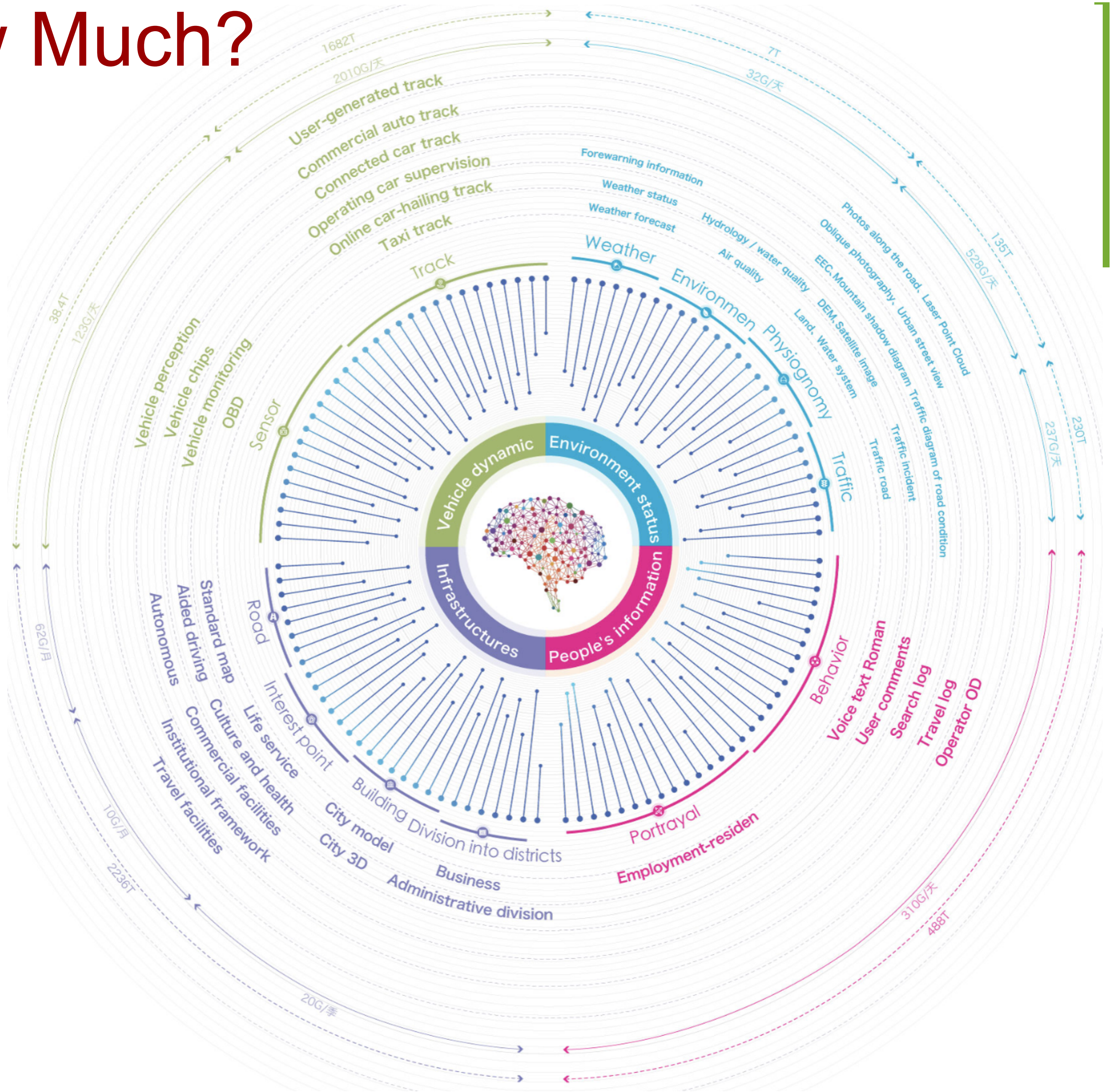
	Current	Daily
Company X (in-car navigation provider)	17.6TB	15M trajectories
Company Y (map app provider)	14.5TB	5M trajectories
Company Z (social network)	0.68TB	18M trajectories

- Every day, ~40M new trajectories, ~4 billion points
- Sampling rates: 50% ~2s, 99% < 10s

NavInfo DataHIVE (minedata.cn, 2018)

Vehicle	Infrastructure	Environment	People
Trajectories:	Standard maps	Weather	Voice and text
- taxis	High res maps	Events	User comments
- uber-like	Services POIs	Air quality	Search log
- monitored	Culture POIs	Water quality	Travel log
- commercial	Commercial POIs	Land & water info	Operators' OD
- user generated	Health POIs	DEM & EEC	Workplace info
Sensor/OBD data	Travel POIs	Satellite image	
Perception data	City models	Street views	
	City 3D Models	Roadside pictures	
	Business districts	Laser point cloud	
	Admin boundaries	Road condition	
	Organization maps	Traffic condition	
		Traffic incidents	

How Much?


1682T
2010G/天
User-generated track
Commercial auto track
Connected car track
Operating car supervision
Online car-hailing track
Taxi track
Track
7T
32G/天
Forewarning information
Weather status
Weather forecast
Hydrology / water quality
Air quality
Weather
Environmen
Physiognomy
Traffic
Photos along the road, Laser Point Cloud
Oblique photography, Urban street view
EEC,Mountain shadow diagram Traffic diagram of road condition
DEM,Satellite image
Land, Water system
Traffic incident
Traffic road
135T
528G/天
237G/天
230T
38.4T
123G/天
Vehicle perception
Vehicle chips
Vehicle monitoring
OBD
Sensor
Vehicle dynamic
Environment status
Infrastructures
People's information
Road
Standard map
Aided driving
Autonomous
62G/月
Interest point
Life service
Culture and health
Commercial facilities
Institutional framework
Travel facilities
Building
City model
City 3D
Division into districts
Business
Administrative division
10G/月
2236T
20G/季
Behavior
Voice text Roman
User comments
Search log
Travel log
Operator OD
Portrayal
Employment-residen
310G/天
488T

+ A Lot of Data!

		Total	Per Period
Vehicle Dynamics	Track (GPS and others)	1682 T	2010 G/day
	Sensor (OBD, cameras etc)	39 T	123 G/day
Environment Status	Weather and air/water quality	7 T	32 G/day
	Physiognomy	135 T	528 G/day
	Traffic	230 T	237 G/day
Infrastructures	Road	2236 T	62 G/mth
	POI		10 G/mth
	Building and admin boundary		20 G/quarter
People Information	Profile and behavior	488 T	310 G/day

+ Some New Trends

- Trajectory analytics now becomes a new frontier for business intelligence
- It is imperative for many businesses to derive values form their trajectory data
- Strong interest from a wide range of industries
- Trajectory data is often used together with other types of data
- Many things we have done so far need to be revisited in the new context

+ New Challenges

- An enterprise-wide spatial information system
- Prefer a general-purposes trajectory management systems
  - For monitoring and managing trajectory data
  - For supporting current and future analytics and mining applications
  - Taking advantages of fast and scalable computing platforms
- Data Integration and Data quality management
- Scalable algorithms
  - For billions of trajectories and millions of concurrent queries

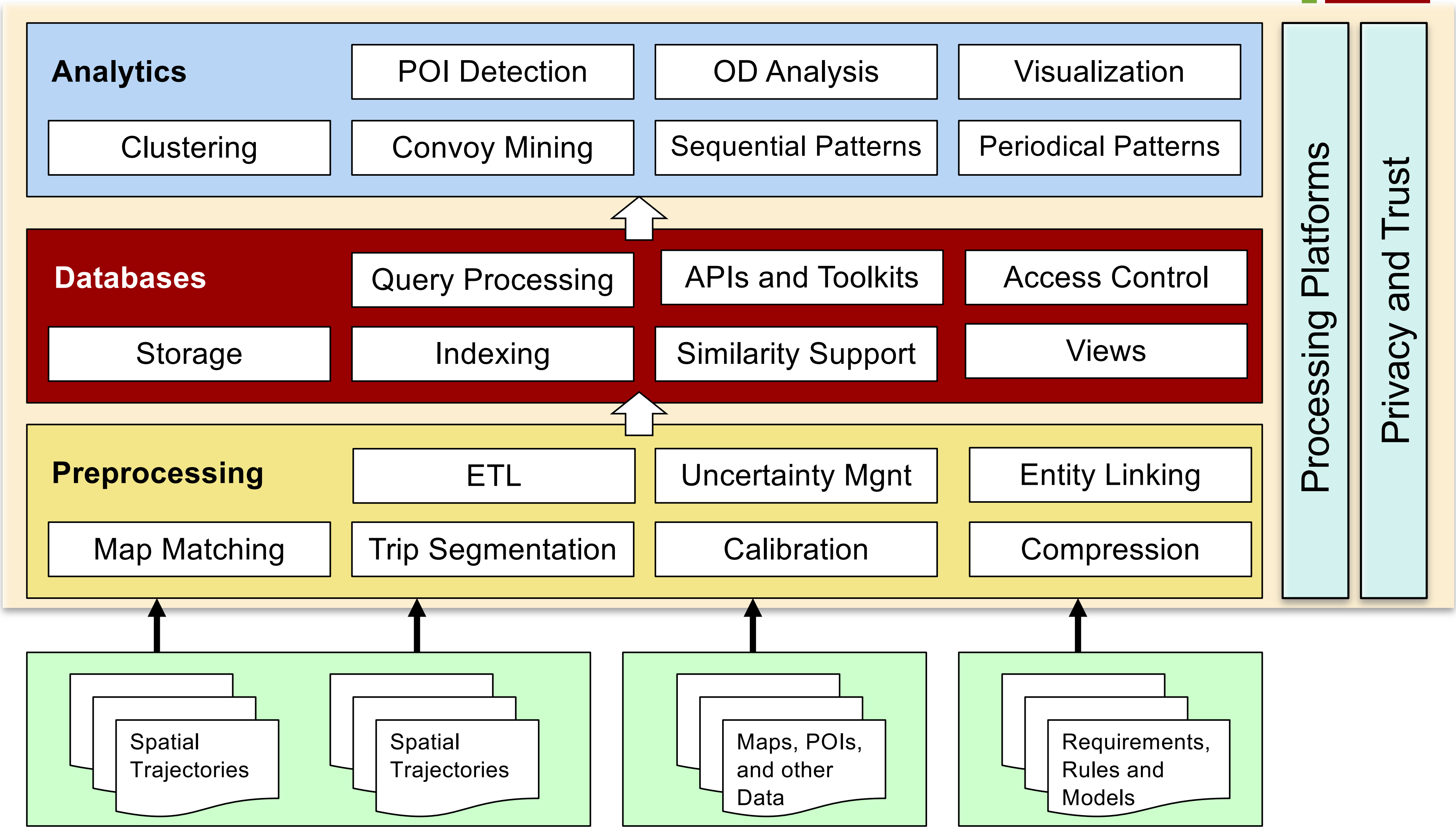
A Trajectory DBMS?

…for monitoring, managing and analyzing

+ Why a Common Platform?

- Universal
  - GPS, telecom tokens, social apps…
- Shared enterprise data
  - For monitoring, predication, business insights…
- Separation of conceptual, logical and physical design
  - Especially different computing platforms to consider today
- Other benefits we took for granted
  - Optimization for data storage and query processing, scheduling, concurrency control…

+ Trajectory Processing Framework

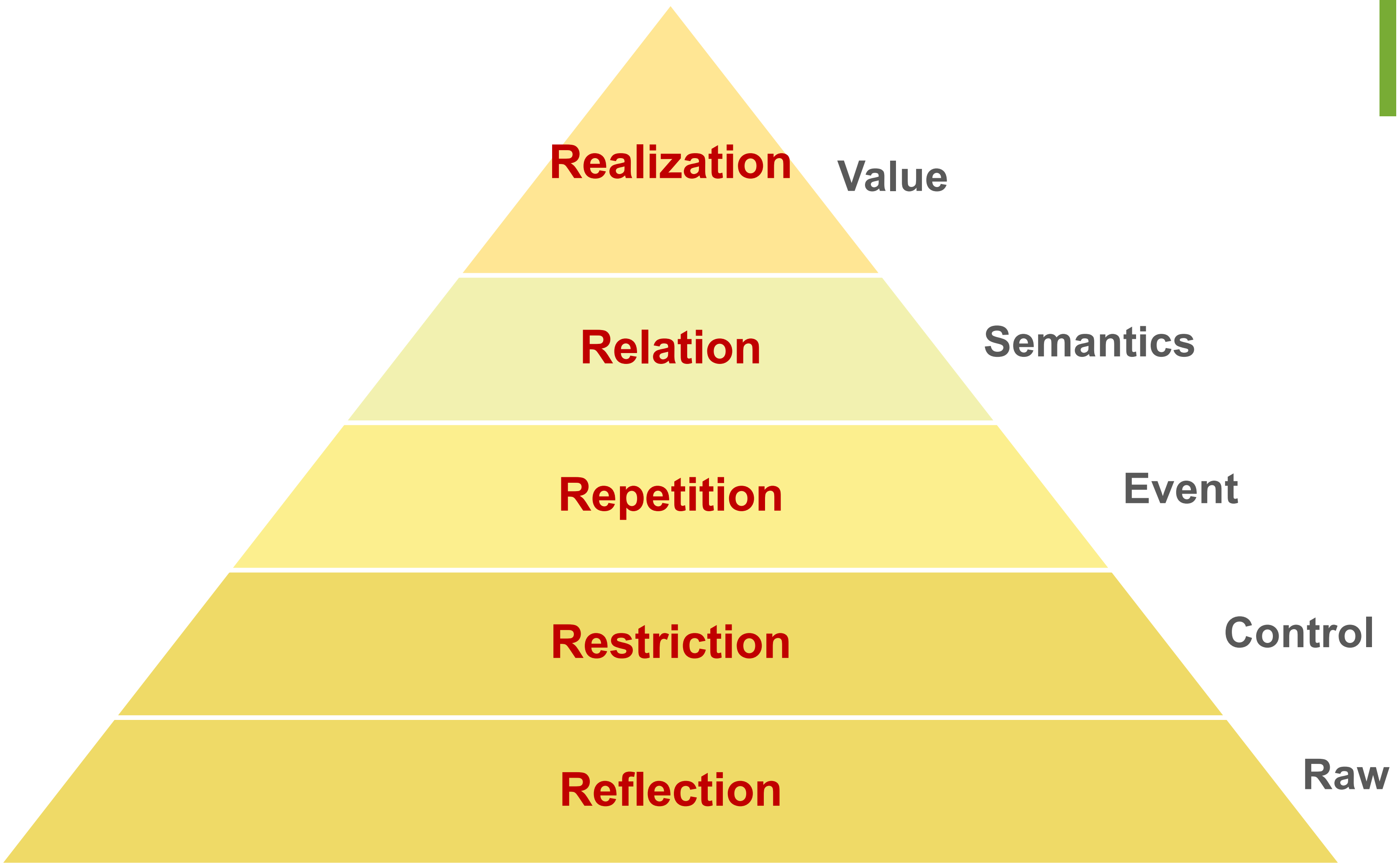


+ The Large-Scale Space Problem

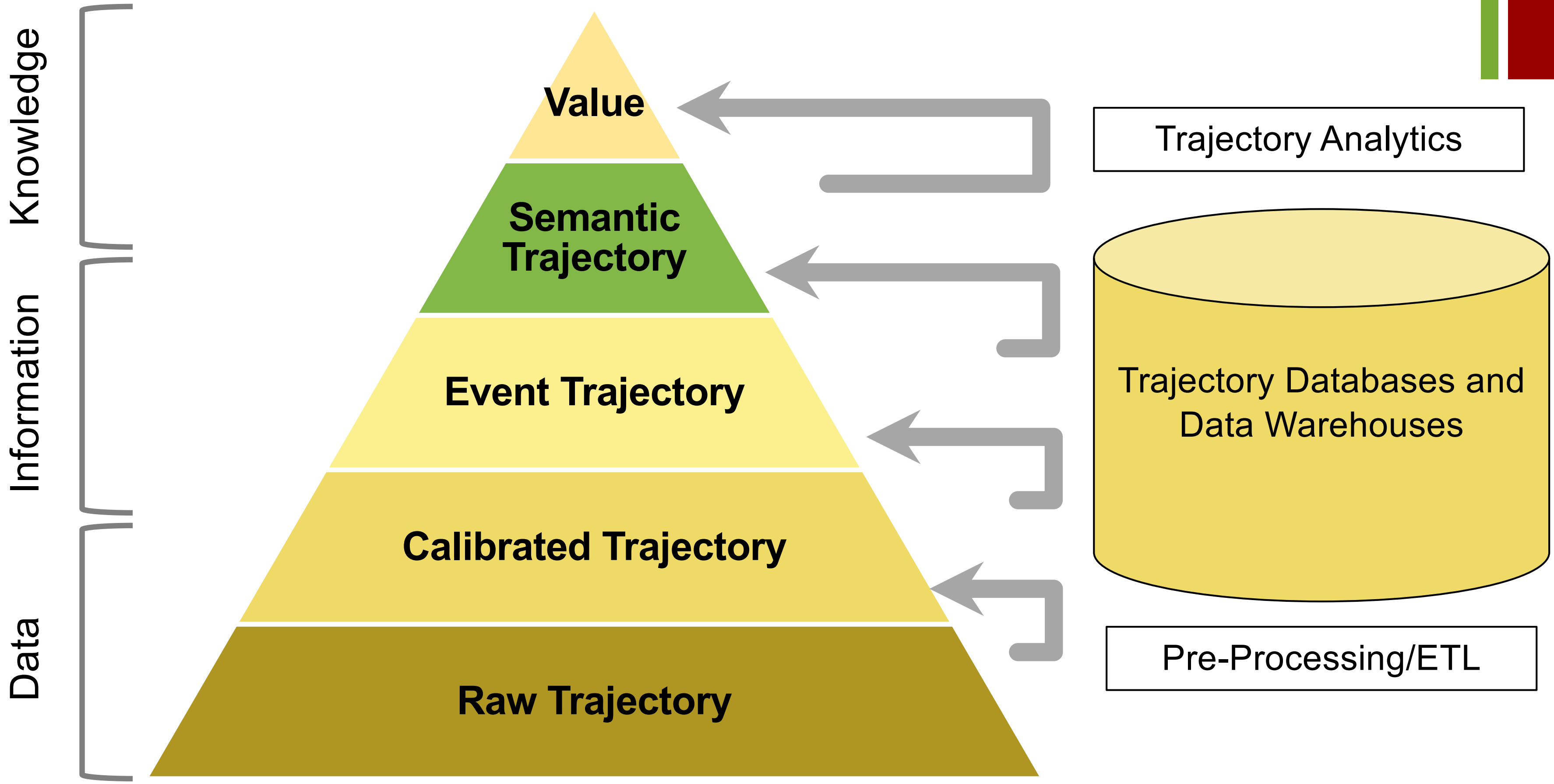
- A space whose structure is at a much larger scale than the sensory horizon of the agent
  - Therefore, a knowledge model is needed to understand the space
- It consists of multiple <u>interacting</u> representations, each with its own ontology, given the agent
  - More expressive power for incomplete knowledge
  - More robustness in sensorimotor uncertainty and computational limitations

Benjamin Kuipers, "The Spatial Semantic Hierarchy", Artificial Intelligence, 2000

+ The 5R Approach



+ A Spatiotemporal Pyramid



+ SparkDB

- A time-centric storage and processing system for trajectories
- Designed for in-memory computers
- A more ambitious system is under development, following the proposed processing framework
- Now supported by a couple of users

H. Wang, K. Zheng, X. Zhou and S. Sadiq, "SharkDB: An In-memory Column-oriented Trajectory Storage", **CIKM** 2014
Haozhou Wang, Kai Zheng, Xiaofang Zhou, Shazia Sadiq, "SharkDB: An In-Memory Storage System for Massive Trajectory Data", **SIGMOD** 2015 (demo)

Data Quality

…fitness for use

+ Data Quality in General

- Data quality is about “fitness for use”
- Four many criteria
  - Accuracy
  - Completeness
  - Timeliness
  - Consistency
- Many other aspects
  - Entity linking
  - Data provenance

+ Trajectory Data Quality Issues

- Inaccuracy
  - Measurement errors and sampling issues
  - Rule-based data calibration and uncertainty management
- Redundancy
  - Low value density vs high redundancy
  - Data reduction and compression
- Data sparsity (i.e., incompleteness)
  - No matter how much data you have, you don't have enough
- Lack of structure
  - Trip information, entity information
- Lack of semantics
  - Transportation mode, activity, contextual information…

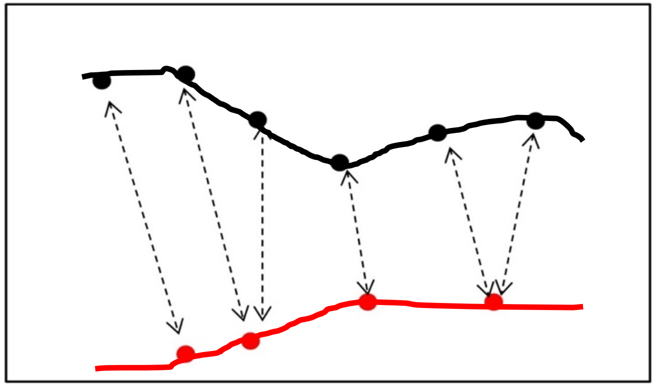
+ Dealing With Low Sampling Data

- Where an object goes between two sampling points which are 10 minutes apart?
  - Interpolation based on the map
  - Interpolation based on other moving objects
  - Results: locations and paths ranked by probabilities
  - Probabilistic query processing is not always desirable but sometimes unavoidable
- And now?
  - Telecoms tokens
  - Social networks check-ins…

Kai Zheng, Goce Trajcevski, Xiaofang Zhou, Peter Scheuermann, "Probabilistic Range Queries for Uncertain Trajectories on Road Networks", **EDBT** 2011
Kai Zheng, Yu Zheng, Xing Xie, Xing Zhou, "Reducing Uncertainty of Low-Sampling-Rate Trajectories", **ICDE** 2012

+ Trajectory Calibration

- Popular trajectory distance measures
  - Euclidean distance, LCSS, DTW, EDR

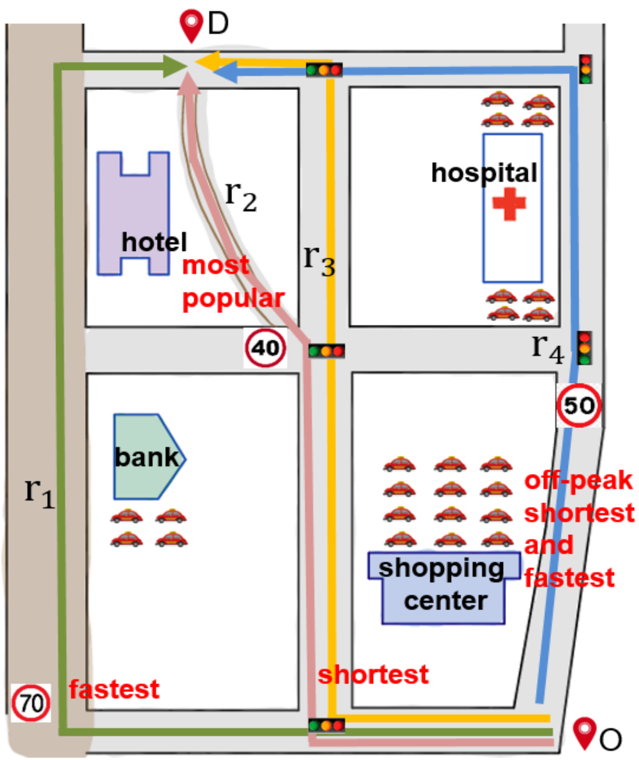


- How distance measures work?
  - Sample points alignment
  - Aggregating differences of aligned pairs
- Experiments
  - Ground Truth: 11,000 high-sampling-rate real trajectories
  - Derived Trajectory Datasets: re-sampling, shifting, jumping
- Need to calibrate – rewrite using points in a common reference set

H. Su, K. Zheng, H. Wang and X. Zhou, Calibrating Trajectory Data for Similarity-based Analysis, **SIGMOD** 2013

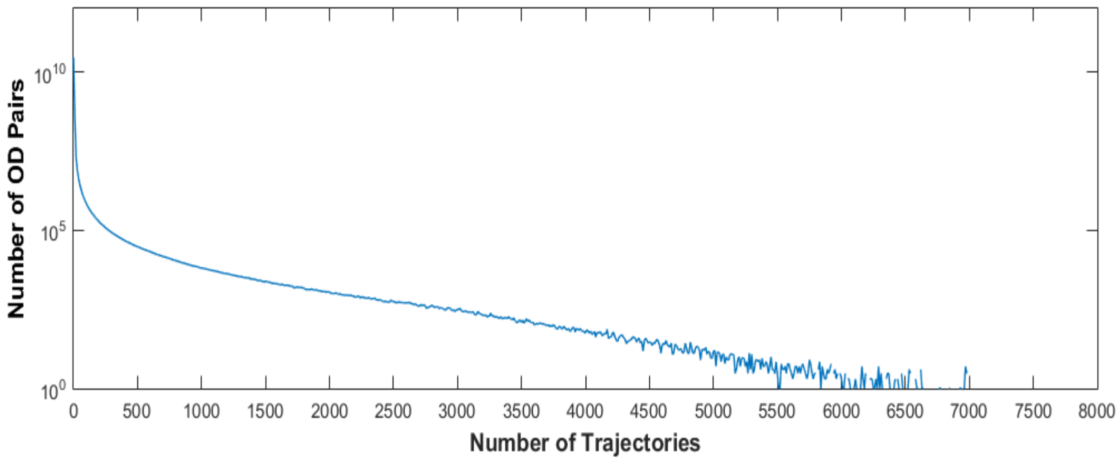
+ Trajectory Clustering and Labeling

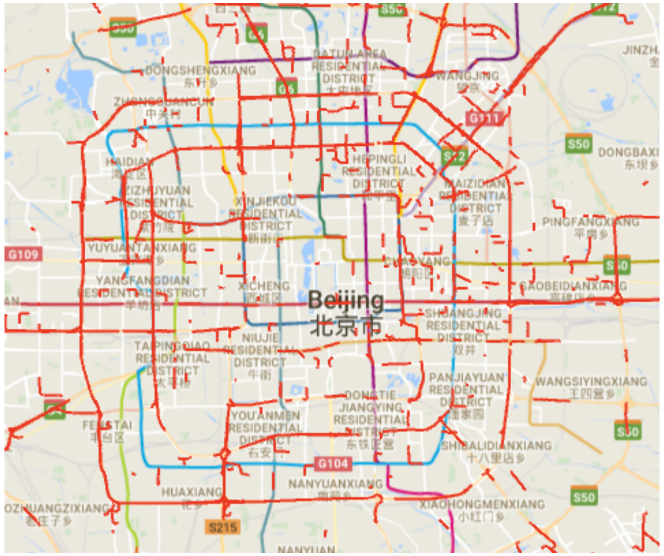
- Applications
  - Moving behaviors analysis
  - Personalized routing
- Clustering
  - OD-specific trajectories
- Labeling
  - Features: fastest, shortest, most popular, time-related



+ Trajectory Augmentation

- Data augmentation approach
  - Factorization-based [1] : tensor decomposition with extra data sources (geospatial, temporal, and historical correlation)
  - Concatenation-based [2] : sub-trajectories
  - Correctne3ss check [3]: similar distribution





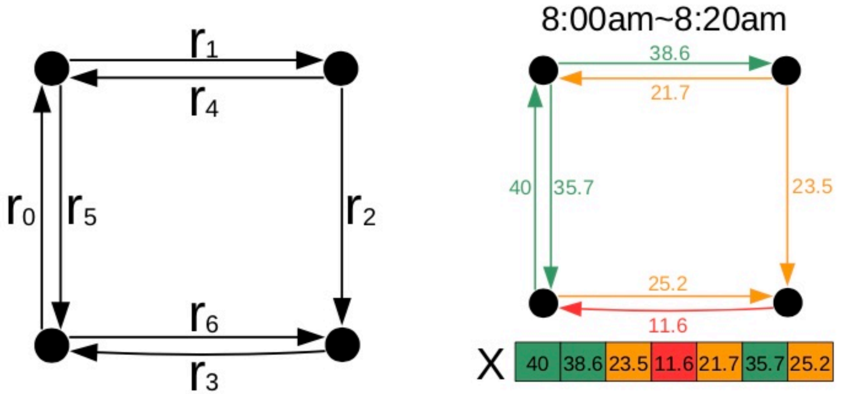
[1]. Yilun Wang, Yu Zheng, Yexiang Xue. "Travel time estimation of a path using sparse trajectories" ***SIGKDD***, ***2014***.
[2]. Dai Jian, Bin Yang, Chenjuan Guo, Zhiming Ding. "Personalized route recommendation using big trajectory data." ***ICDE, 2015***
[3] D. He, B. Ruan, B. Zheng, X. Zhou, Origin-Destination Trajectory Diversity Analysis: Efficient Top-k Diversified Search, **MDM 2018**

+ Deep Learning for Predication

- Given:
  - A road map (as a directed graph)
  - A sequence of speed vectors, each vector is the speed at each road segment during a time interval

$$X_t = [x_t^{r_0}, x_t^{r_1}, ..., x_t^{r_{|E|-1}}],$$

Problem: Given the historical observations $\{X_i | i = 1, ..., t\}$, this paper aims to predict $Y_t = \{X_j | j = t+1, ..., t+z\}$, where z is the number of time intervals to be predicted.

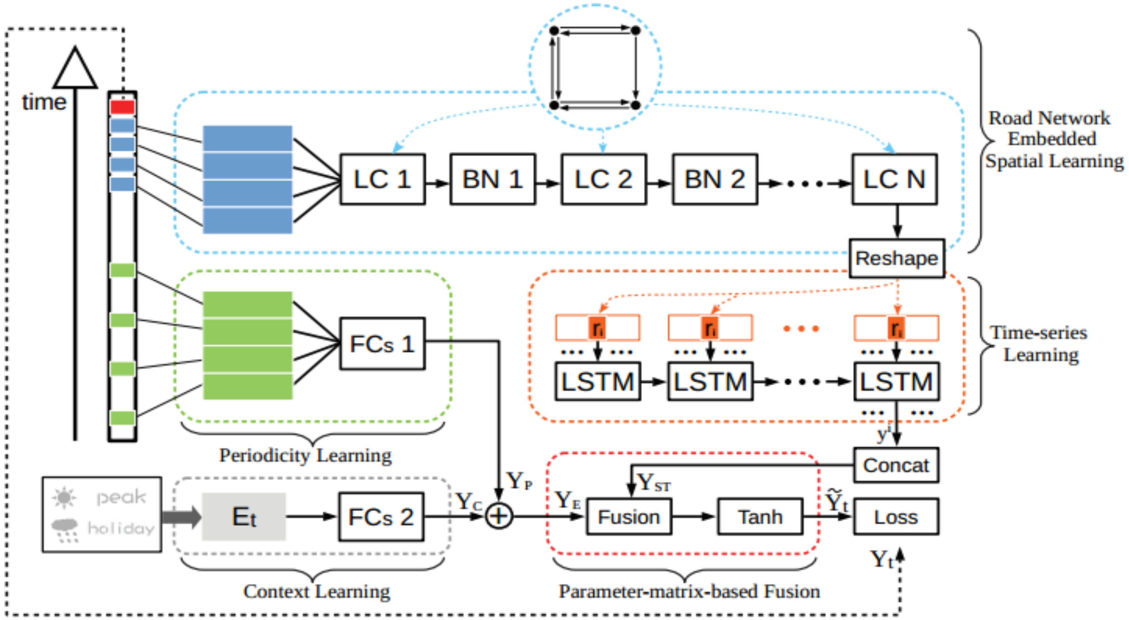


(a) A road network

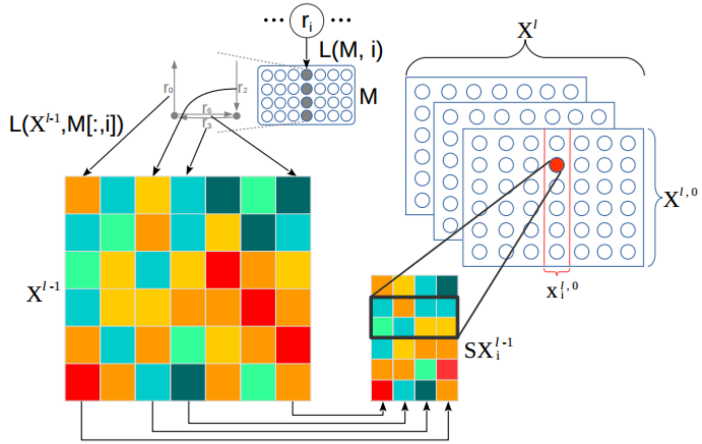
(b) A speed vector at 8:00am - 8:20am

+ LC-RNN Model

- ARIMA based (conventional), RNN based (consider time only), CNN based (spatial information but previously ony at grid level)
- Look-up Convolution (LC): learn the latent features of surrounding area
- LSTM components: learn the time-series pattern that is aware of surrounding area dynamics



LC-RNN model

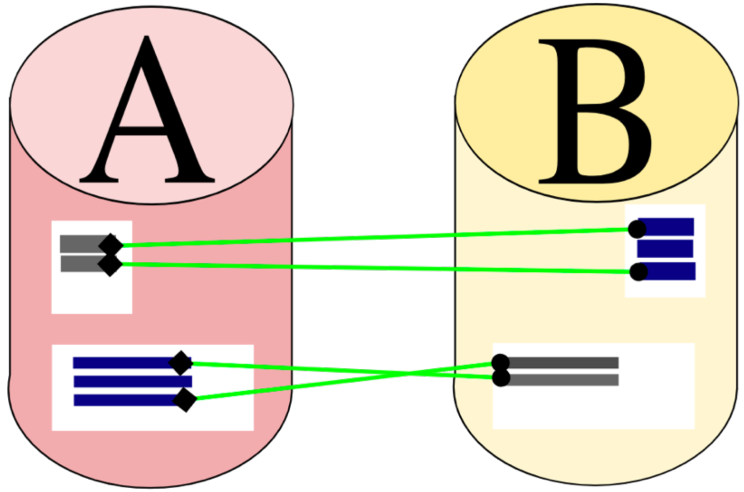


Look-up Convolution

Z. Lv, J. Xu, K. Zheng, P. Zhao, H. Yin and X. Zhou, "LC-RNN: A Deep Learning Model for Traffic Speed Prediction", **IJCAI** 2018.

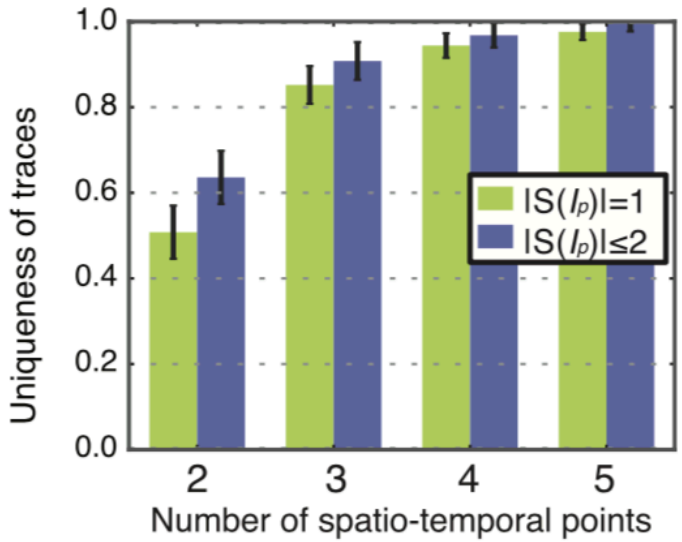
+ Spatiotemporal Entity Resolution

- Linking entities based on their trajectory data
- Understanding the extent to which spatiotemporal data are distinctive is crucial to:
  - Entity resolution and data integration
  - Location privacy protection
- Data sources
  - Check-ins
  - Card transactions
  - Phone tokens/call records
  - Vehicle trajectories
  - Many social networks…



+ Uniqueness of Individual Mobility

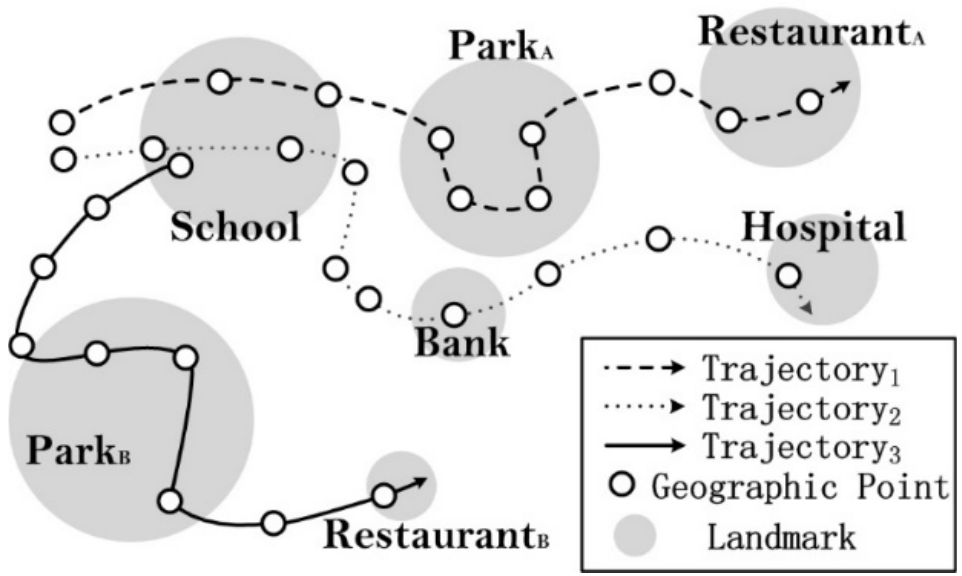
- "4 randomly sampled spatiotemporal points can uniquely identify 95% of individuals."[1]
  - Dataset
    - 1.5 M mobile phone users over 15 mths
    - Only when/where to make/receive calls
  - As for another real-world taxi dataset
    - 12,000 taxis over one month
    - <15% of taxis were successfully identified



[1] Montjoye Y A D et al. Unique in the Crowd: The privacy bounds of human mobility[J]. Scientific Reports, 2013, 3(6):1376.

+ Everyone Has Mobility Signature?

- Spatial signature?
  - Commonality: you visit frequently, such as your office building
  - Unicity: you can be distinguished from others, like personal home address



+ Signature Representations

- Sequential signature
  - q-gram and generalized Jaccard coefficient
- Temporal signature
  - Temporal histogram and Earth Mover's Distance (EMD)
- Spatial signature
  - TF-IDF weighted vector and cosine similarity
  - $f(o) = (< p_1, w(p_1) >, \ldots, < p_d, w(p_d) >)$
    - p: a spatial point
    - $w(p)$: TF-IDF weight of p
      - TF: measures the frequency of p in $T(o)$ - commonality
      - IDF: measures how much distinctiveness p provides – unic
- Spatiotemporal signature
  - TF-IDF weighted vector and cosine similarity
  - Each dimension is a spatiotemporal pair (p, T)

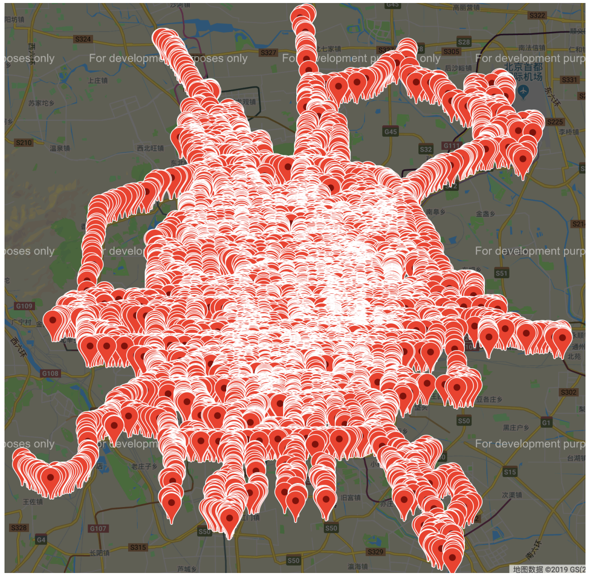
+ Signature Reduction

- Baselines
  - Principal component analysis (PCA) [1]
  - Locality sensitive hashing (LSH) [2-3]
- CUT – simple but very effective
  - Signature exhibits a power-law distribution – CUT long tail
  - Preserve top-m points with largest weights – minor information loss
  - Signature's spatial shrinking

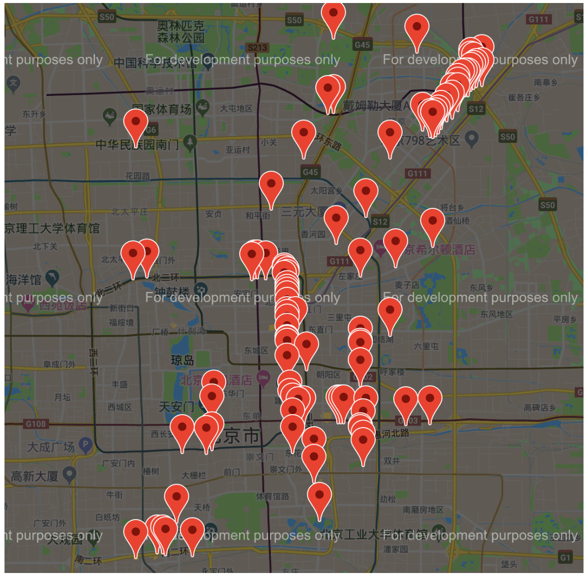
[1] K. P. F.R.S., "Liii. on lines and planes of closest fit to systems of points in space", *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 1901
[2] P. Indyk, "Approximate nearest neighbors: Towards removing the curse of dimensionality", STOC '8
[3] A. Gionis, P. Indyk, and R. Motwani, "Similarity search in high dimensions via hashing", VLDB 99

+ Signature's Spatial Shrinking

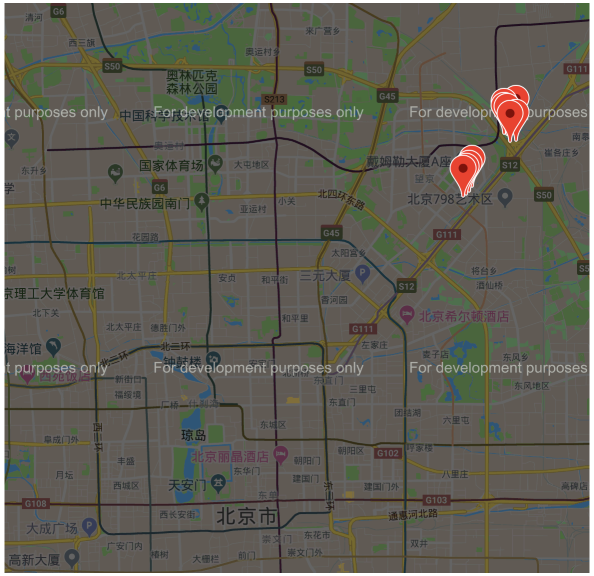
- After CUT, the ratio of spatial overlapping between objects is reduced from almost 100% to 1% when dimensionality is reduced to $m = 10$



Original



$m = 100$



$m = 10$

+ Efficient Moving Object Linking

- Formalize the linking problem as a kNN search on the collection of signatures

- Baselines:
  - Cosine similarity search algorithms
    - e.g. AllPairs, APT, MMJoin, L2AP[1] …
  - Efficient kNN search methods in Euclidean space
    - Spatial indexing (e.g. R-tree)
    - Approximate k-NN search (e.g. LSH) [2]

[1] D. C. Anastasiu and G. Karypis, "L2AP: Fast cosine similarity search with prefix l-2 norm bounds," ***ICDE 2014.***
[2] A. Gionis, P. Indyk, and R. Motwani, "Similarity search in high dimensions via hashing," ***VLDB 1999***

+ Weighted R-Tree (WR-tree)

- Transform the high-dimensional kNN search to 2D space
  - Combine weight and spatial information
    - $MBR(o)$: the minimum bounding rectangle of a weighted signature stored in the node
  - Two pruning strategies
    - Pruning by spatial overlapping – 2D R-tree
    - Pruning by signature similarity

+ Experiments

- A real-world taxi dataset
  - 12,000 taxis in total
  - 160,000 unique points in total after trajectory calibration

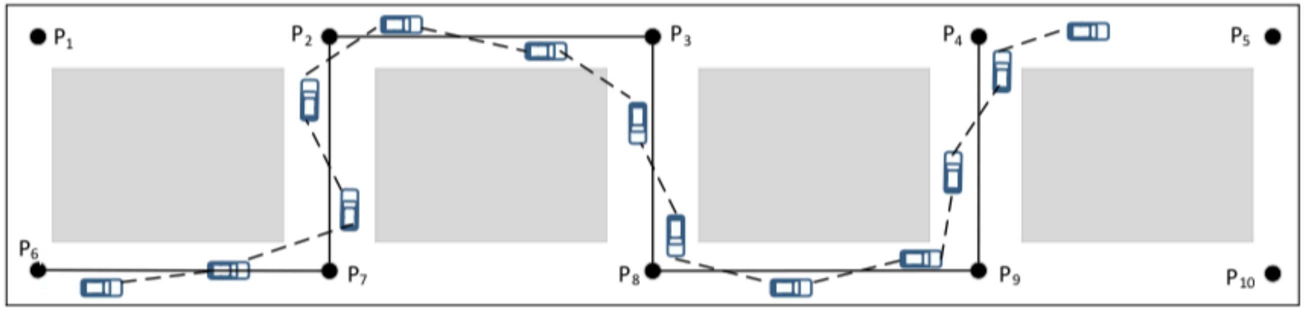


Fig. 1. An example of vehicle trace calibration.

- Evaluation metric
  - Acc@k – Effectiveness
  - Time cost – Efficiency

+ Signature Effectiveness Study

- Spatial signature is the most effective: 85.5% Acc@1
- Sequential and temporal features are not important for the task of moving object linking

Methods	Sequential (q)					Temporal (Δt)							Spatial	Spatiotemporal (# of grids)		
Parameters	1	2	3	4	5	1h	2h	3h	4h	6h	8h	12h	N/A	100^2	200^2	300^2
$Acc@1$	0.681	0.679	0.649	0.627	0.604	0.127	0.123	0.104	0.087	0.042	0.018	0.004	0.855	0.535	0.567	0.583
$Acc@2$	0.721	0.718	0.695	0.681	0.664	0.169	0.167	0.145	0.124	0.074	0.033	0.007	0.904	0.587	0.613	0.630
$Acc@3$	0.745	0.741	0.724	0.708	0.698	0.195	0.186	0.172	0.150	0.092	0.046	0.009	0.928	0.612	0.64	0.651
$Acc@4$	0.760	0.758	0.741	0.726	0.724	0.216	0.205	0.198	0.174	0.113	0.057	0.011	0.940	0.632	0.659	0.681
$Acc@5$	0.768	0.768	0.755	0.741	0.741	0.233	0.220	0.216	0.192	0.131	0.071	0.013	0.948	0.647	0.673	0.693

Spatial signature is the most effective empirically. We only consider spatial signature from here.

+ Reduction Effectiveness Study

- CUT outperforms PCA and LSH
  - The superiority of CUT is most obvious when m is small
- CUT can reduce dimensionality dramatically with a slight accuracy decrease (< 5%)

Methods	PCA					LSH					CUT					Original
m	10	50	100	500	1000	10	50	100	500	1000	10	50	100	500	1000	160,000
$Acc@1$	0.007	0.050	0.113	0.542	0.697	0.046	0.476	0.638	0.795	0.824	0.806	0.827	0.831	0.836	0.838	0.855
$Acc@2$	0.012	0.088	0.187	0.686	0.801	0.079	0.542	0.705	0.847	0.870	0.866	0.877	0.880	0.885	0.886	0.904
$Acc@3$	0.018	0.123	0.243	0.765	0.846	0.097	0.577	0.731	0.872	0.893	0.893	0.903	0.907	0.913	0.916	0.928
$Acc@4$	0.023	0.150	0.289	0.809	0.875	0.118	0.597	0.748	0.891	0.912	0.906	0.919	0.920	0.928	0.929	0.940
$Acc@5$	0.031	0.176	0.333	0.835	0.892	0.130	0.617	0.760	0.900	0.924	0.917	0.929	0.930	0.937	0.939	0.948

We will use reduced signatures obtained by CUT algorithm with $m = 10$ in the following.

+ Search Efficiency Study

- 2D R-tree and WR-tree are more efficient than others
  - The importance of pruning by spatial overlapping
- WR-tree is better than 2D R-tree
  - The significance of pruning by signature similarity

TIME COST (S) OF DIFFERENT LINKING ALGORITHMS ($m = 10$, $k = 1$).

	Linear	L2AP	LSH	2D R-tree	WR-tree
$\|D\| = 3000$	2.269	3.090	1.769	0.651	0.140
$\|D\| = 6000$	8.182	14.557	6.652	2.801	0.633
$\|D\| = 9000$	19.733	36.541	15.642	5.122	0.908
$\|D\| = 12000$	27.183	70.440	38.131	18.876	1.403

Fengmei Jin, Wen Hua, Jiajie Xu, Xiaofang Zhou, "Moving Object Linking Based on Historical Trace", **ICDE** 2019.

+ More To Be Done…

- What are those selected points?
- More efficiency improvement, and for join queries too
- How to safe guide the process?
  - Minimum amount of data? Drifting?
- Heterogeneous data sources
  - Mobile phone token data
  - Social media data
  - Both data and ground truth are difficulty to get…
- How to protect privacy with trajectory data?

Algorithms Revisited

…old problems, new challenges

+ New Context

- More data, more queries, more applications, more computing platforms, and more tools
- Example 1: batch shortest path query processing
- Example 2: correctness-aware kNN query processing

Mengxuan Zhang, Lei Li, Wen Hua and Xiaofang Zhou, "Batch Processing of Shortest Path Queries in Road Networks", **ADC** 2019.
Dan He, Sibo Wang, Xiaofang Zhou and Reynold Cheng, "An Efficient Framework for Correctness-Aware kNN Queries on Road Networks", **ICDE** 2019.

+ Conclusions

- We have discussed:
  - More data, more queries, more applications, more tools
  - The need for a general-purpose and open platform
  - Data quality again is a key issue
  - Many things now need to be revisited
- Some of our current research problems
  - Large-scale space problems
  - Dynamic road networks and contained-based routing
  - Massive concurrent queries and updates
  - Trajectories as a focal point for data integration
  - Time for a trajectory DBMS?
- Now it’s the most exciting time to work on trajectories!