Spatiotemporal Information Pyramid and Databases

周晓方

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

dke
DATA & KNOWLEDGE ENGINEERING

Soochow Advanced Data Analytics Lab
苏州大学先进数据分析研究中心

Outline

- What is spatial trajectory data?
- Why is it important?
- Processing spatial trajectories
- Conclusions

Outline

- What is spatial trajectory data?
- Why is it important?
- Processing spatial trajectories
- Conclusions

What is 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:

  $<objId, trajID, trajProperties,$

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

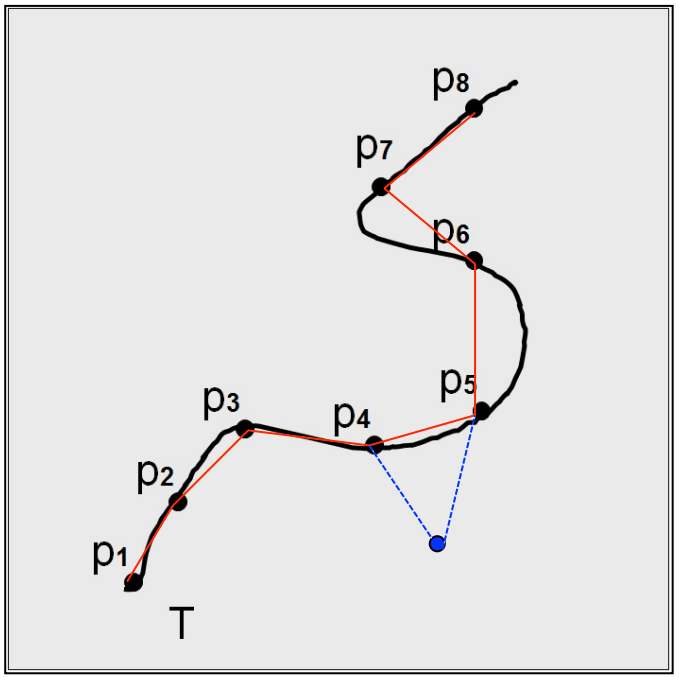
Many Dimensions of Trajectory Data

- Basic dimensions
  - Spatial dimension: locations $p_1, p_2 \ldots p_n$
  - Temporal dimension: time-stamps $t_1, t_2 \ldots t_n$
  - Attribute dimension: other data of interest $a_1, a_2 \ldots a_n$

- Other dimensions
  - Entity dimension: what type of objects?
  - Environment dimension: road networks, water systems, sensor networks
  - Semantic dimension: what activities at a location or time?

More about Trajectory Data

- A trajectory is obtained from sampling the movement of an object
  - Some sampling strategy is used
  - There are many other factors which cannot be controlled
  - Data can be noisy, and there can be both redundant data as well as missing data

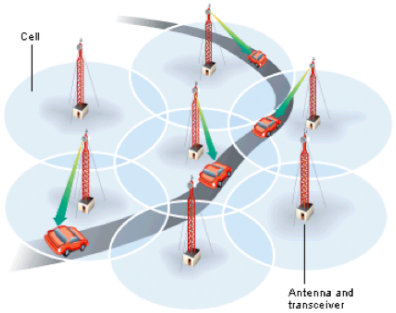


- It is non-trivial to restore the original trace from a trajectory

Where Trajectory Data Come From?

A Lot More GPS Trajectory Data

Beyond GPS Recordings



SENSORS

How Much Trajectory Data?

- A back-of-the-envelope calculation:
  - A simple point data (`x,y,t`): 24 bytes
  - A car can generate 85KB a day (10 hours a day, 10 seconds interval)
  - Beijing has 60,000 taxis, that is 5GB a day, or 1.72 TB a year

- Actual trajectory data could be much larger
  - A multiplier of X: There are much more information than just a point data: taxi ID, trip ID, job status, direction, velocity, acceleration, fuel consumption, other sensor data (M2M data)
  - Even larger once processed: original data, plus map-matched data, other derived data, other forms of representation (e.g., OpenLR)

Public Transport Data

- Brisbane: ~2M population, buses, ferries and trains in southeast Queensland
  - ~1000 routes, every days with 200K passengers, 320K journeys, 430K trips
  - Each transaction contains:

    operator, date, direction (inbound or outbound), ticket number, card ID, boarding time, alighting time, number of passengers per ticket, boarding stop, alighting stop, journey ID, trip ID
  - Now with streaming bus running data
  - We have 6 mths data, ~100M records (actual operations use ~6 years data)
  - New projects to link smartcards to online accounts and bank accounts

M2M Data

- An integral part of IoT, Industry 4.0, connected cars
- Unlimited monitoring and control applications in industry automation, logistics, smart grid, smart cities, health, defense

Program Group Number (pgn) 65262 – **Engine Temperature 1 – ET1**

Transmission rate:	**1 sec**
Data Length:	**8 bytes**
Data Page:	**0**
PDU Format:	**254**
PDU Specific:	**238**
Default Priority:	**6**
Parameter Group Number:	**65262 (00FEEE hex)**

Bit Start Position	Bytes Length	Suspect Parameter Number SPN Description	SPN
1	**1 byte**	**Engine Coolant Temperature**	**110**
2	**1 byte**	**Fuel Temperature**	**174**
3-4	**2 bytes**	**Engine Oil Temperature 1**	**175**
5-6	**2 bytes**	**Turbo Oil Temperature**	**176**
7	**1 byte**	**Engine Intercooler Temperature**	**52**
8	**1 byte**	**Engine Intercooler Thermostat Opening**	**1134**

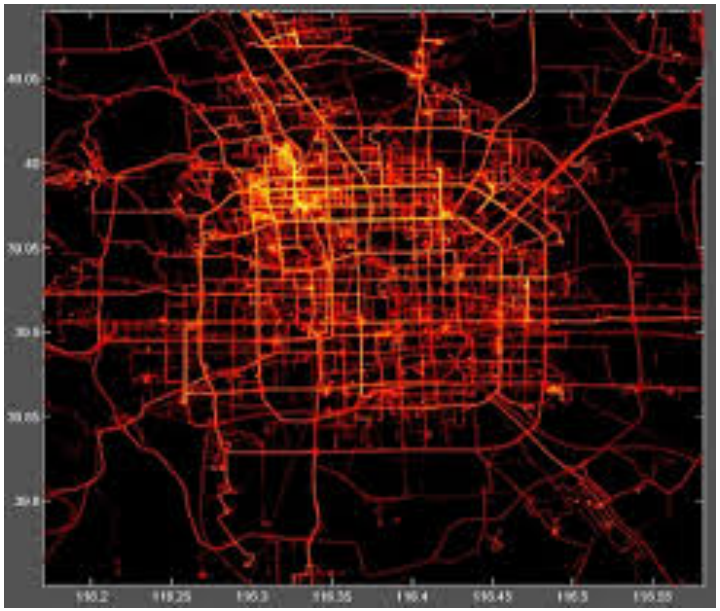
In-vehicle M2M data (telematics)

- 90 million in 2010 ➔1.4 billion in 2020, with 300 million vehicle platforms plus 1+ billion application-specific “aftermarket” devices (according to *Machina Research*)
- To improve efficiency, safety, maintenance, logistics, customer loyalty, with applications such as Asset tracking, road freight, fleet management, route optimization, security, driver/emergency/roadside assistance, vehicle recovery, traffic information, navigation, pricing and payment, advertising

Are Spatial Trajectory Data Useful?

Significant interest in

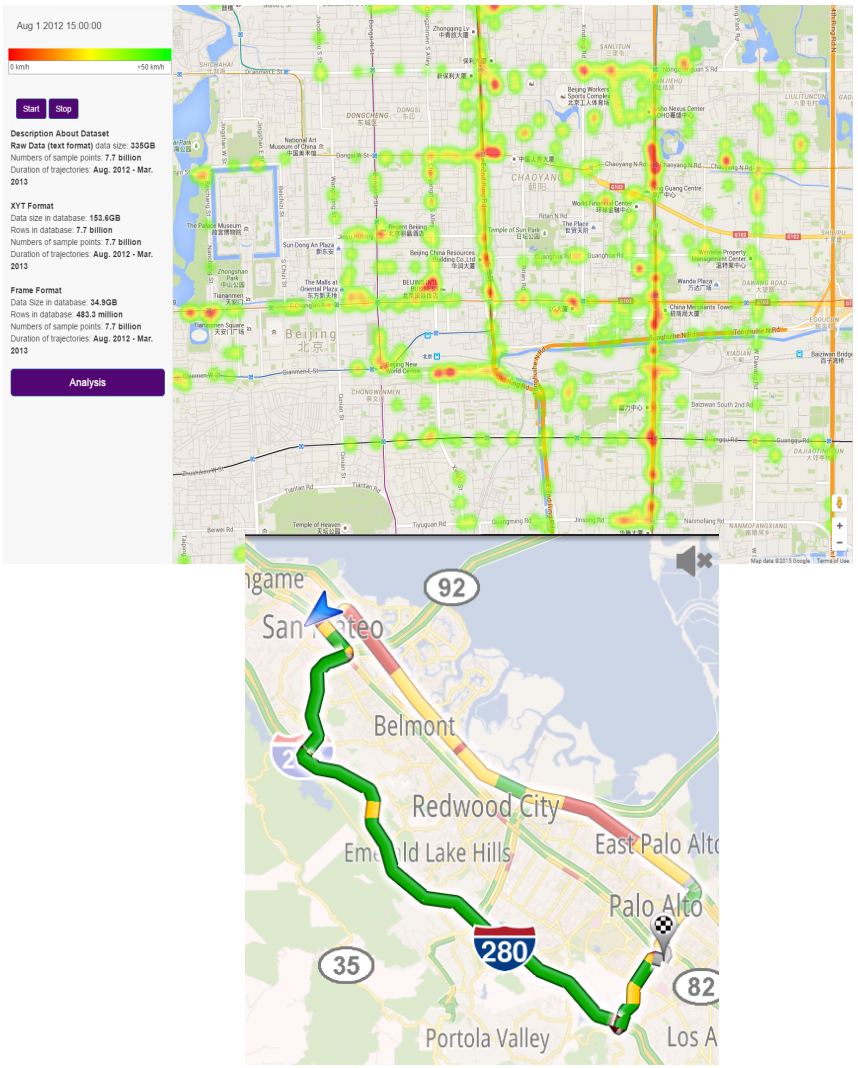
- Route planning and recommendation
- LBS and mobile advertisement
- Resource tracking and scheduling
- Transport & Logistics
- Urban planning and smart cities
- Road safety
- Emergency responses
- Environment monitoring…



Trajectory analytics now becomes a new frontier for business intelligence, especially when done in real-time and combined with other types of data

Example 1: Floating Car Data (FCD)

- Essential source for traffic information for ITS today
  - Add more useful data to traffic cameras and roadside induction loops
- Can monitor traffic in real-time, and identify traffic congestion, travel time
- Both real-time and historical data are useful
  - Many applications

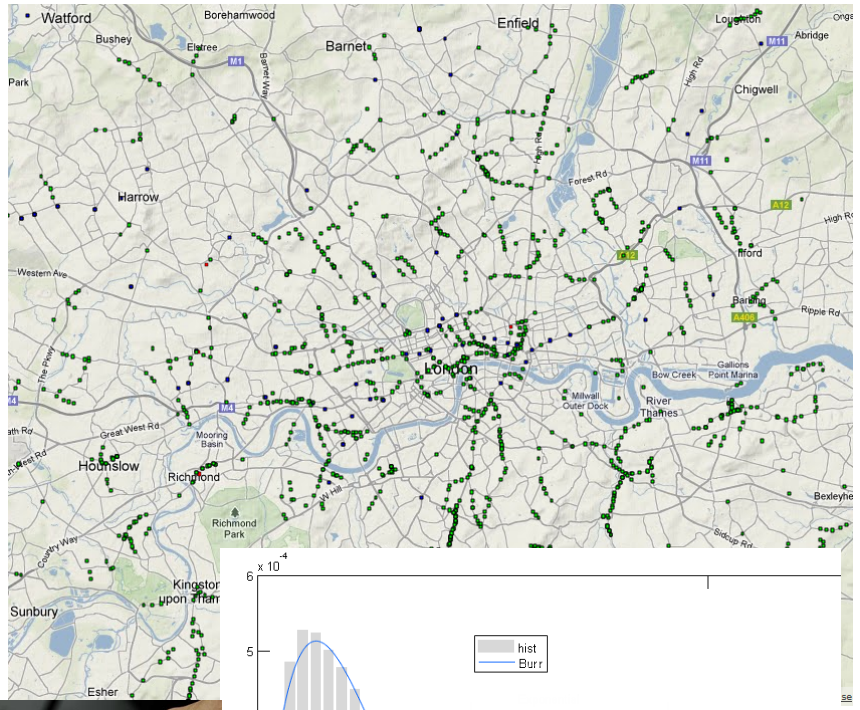


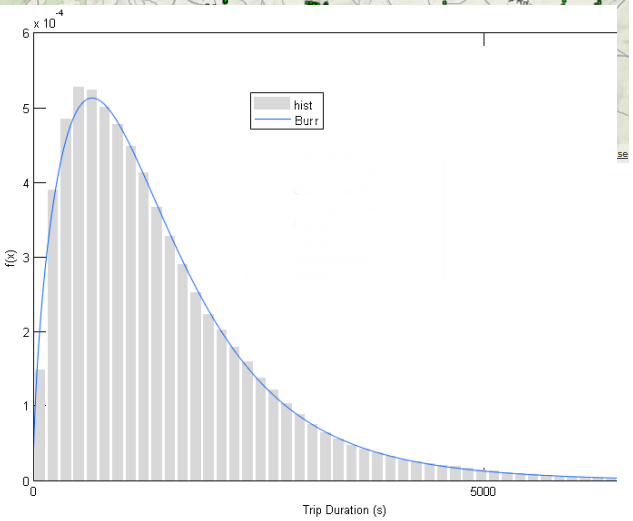
Example 2: Public Transport Smart Data

- Vehicle locations
- Ticketing analysis
- Trip analysis & profitability
- Performance monitoring

- Fraud analysis
- Stop utilization
- Fleet maintenance & utilization

- On-time running
- Service performance
- OD analysis
- Service planning & improvement

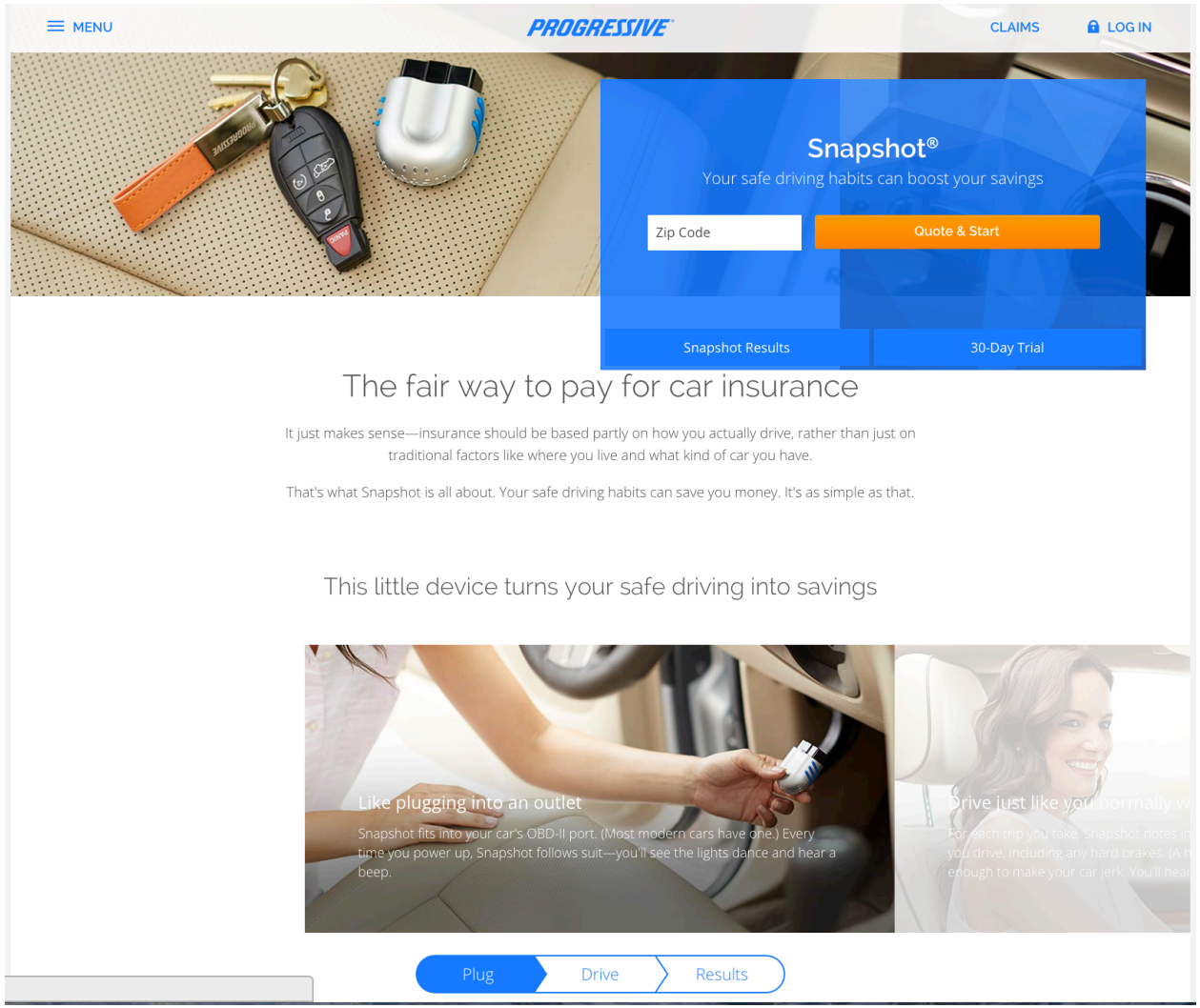




Example 3: Car Insurance

An M2M device plugs into a car's OBD-II diagnostic connector and tracks driving habits over a 30-day period, including the number of miles driven, how hard the driver brakes, and how often she drives between midnight and 4 a.m.

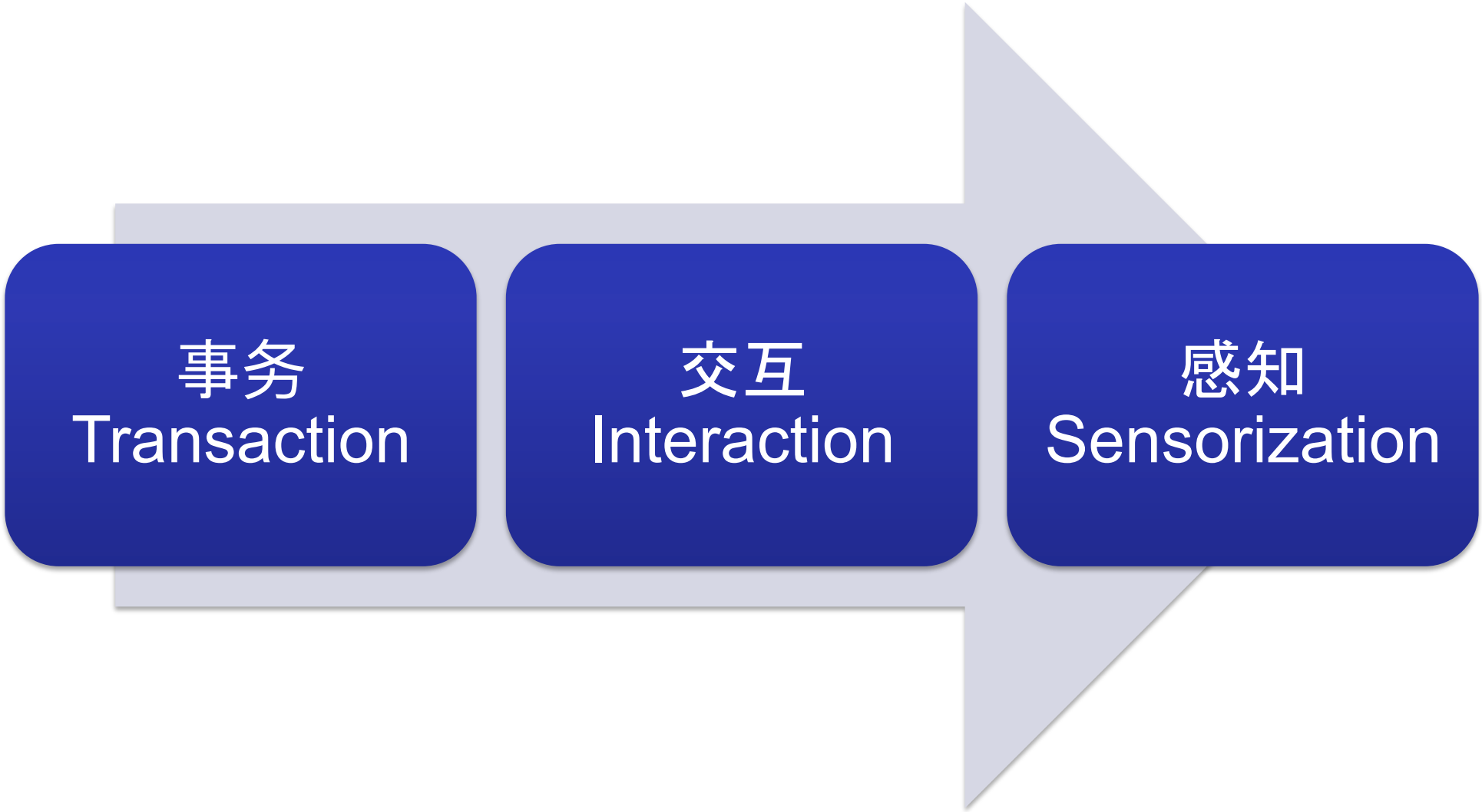
Cautious drivers can save up 30% on their insurance premiums.



Outline

- What is spatial trajectory data?
- Why is it important?
- Processing spatial trajectories
- Conclusions

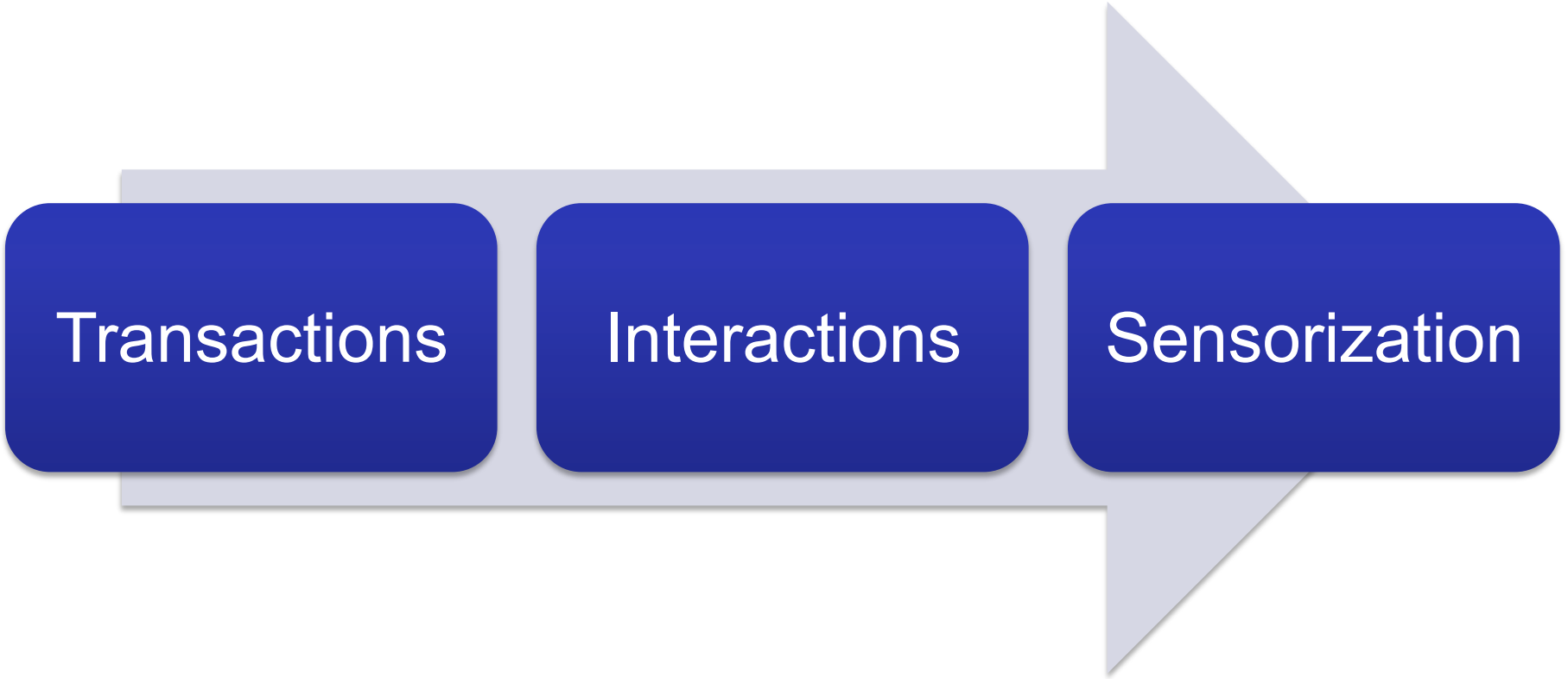
Change of Data



Data Characteristics

	Transaction	Interaction	Sensorization
Applications	Banks, airlines, enterprises… Recording, reporting and BI	Web, e-Commence, social media… Aggregation, association and predication	Sensors, GPS, M2M… Monitoring, control and process optimization
Characteristics			
Volume	Small (GB-TB)	Large (TB-PB)	Large (TB-PB)
Velocity	Periodical change	Streaming	Streaming
Structure	Well	Semi	Poor
Data quality	Good	Fair	Bad
Redundancy	No	Medium	High
Value density	High	Low	Very low

Data Processing Technologies



RDBMS **Data Warehousing** **Data Mining**	**HDFS/Hadoop** **NoSQL** **Column-stores**	**We need to work here!**

A General View about Trajectory Data

- Spatial trajectory data is a representative type of “big data”
  - Volume, velocity, variety and value
- It also exemplifies the future generation of data processing problems
  - Transaction ➔ Interaction ➔ Senorization
- It is a typical type of low-value data
  - Not about storing “normal” things, but to monitor “interesting” things
  - Compression is not good enough; What data to throw away?
- It contains not only data, but also models to generate data
  - Location sampling strategies, data cleaning algorithms, dependency on the base map data, feature extraction algorithms…
- It has a good foundation to start

Spatial Semantic Hierarchy (SSH)

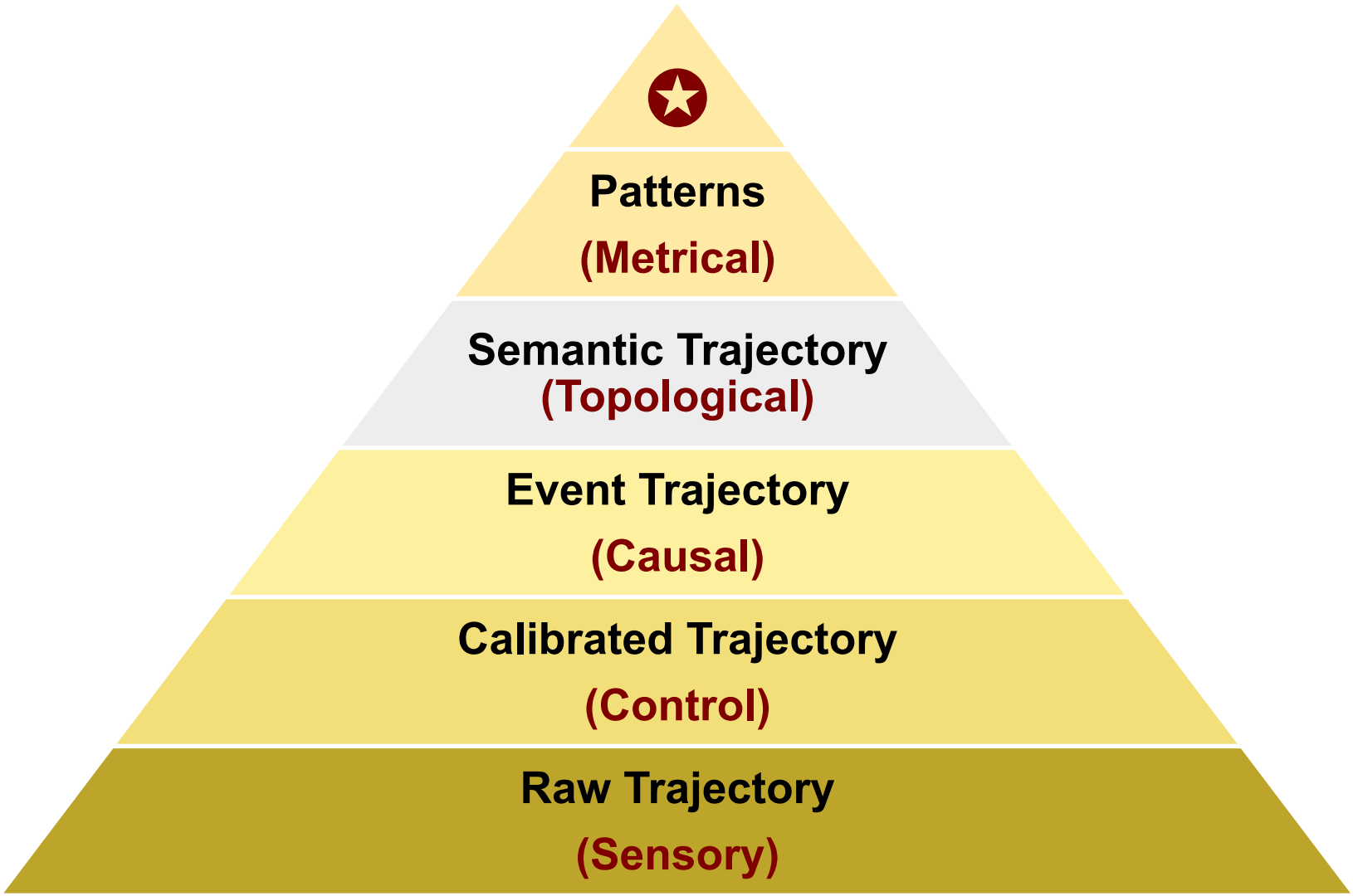
- Large-scale space
  - A space whose structure is at a much larger scale than the sensory horizon of the agent
  - Therefore, a knowledge model is needed to understand the space

- It consists of multiple interacting representations, each with its own ontology, given the agent
  - More expressive power for incomplete knowledge
  - More robustness in sensorimotor uncertainty and computational limitations

Knowledge Representation Levels

- Sensory
  - Continuous data from the agent's sensory system
- Control
  - Descriptions in terms of control laws behind the agent and its environment (e.g., Hill-climbing and trajectory-following control laws)
- Causal
  - Abstracts in a discrete model (e.g., views, actions, turning angles)
- Topological
  - Introduction of place, paths and regions, and their connectivity, order and containment relations
- Metrical
  - Representation of a global geometric map of the environment in a single frame or reference

SSH For Trajectories



A New Challenge for Database People

- DBMS has never been good at managing and processing trajectory data
  - Variable length, spatial and temporal dimensions, high redundancy, poor quality
- Processing trajectory is hard
  - Incompatible data, ad hoc similarity measures
  - Data fusion is essential
- Real-time is harder
  - Data volume, complex similarity measures, ineffective approximation, user feedback…
- Redundancy
  - Most information can be derived from the raw data, but with assumptions, models and algorithms
  - Derived data are stored for performance and usability reasons
  - Links many different levels of representations, and query along those links
- Reduction
  - Compression is important, but often not enough
  - What data to throw away is a new research problem

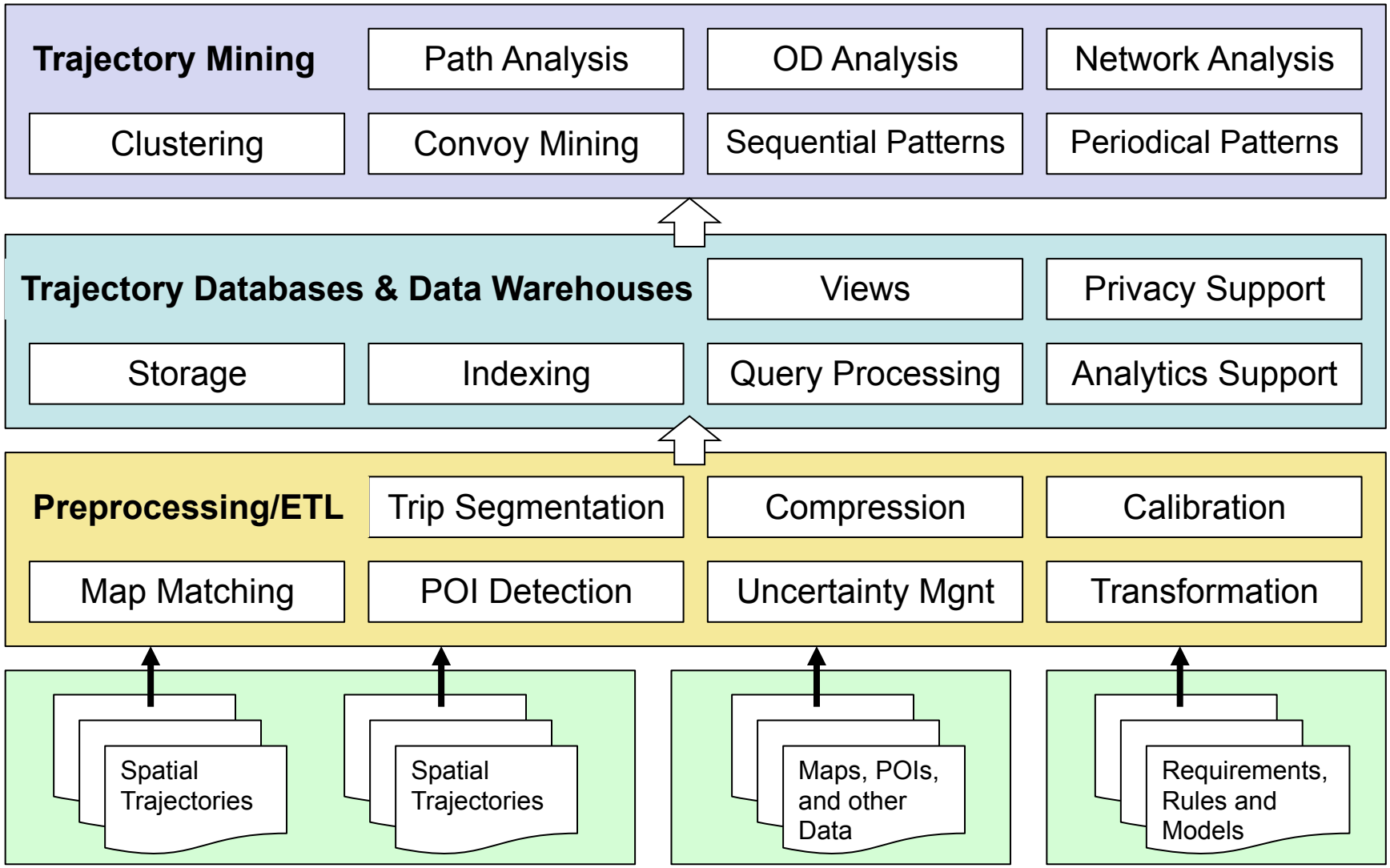
Outline

- What is spatial trajectory data?
- Why is it important?
- Processing spatial trajectories
- Conclusions

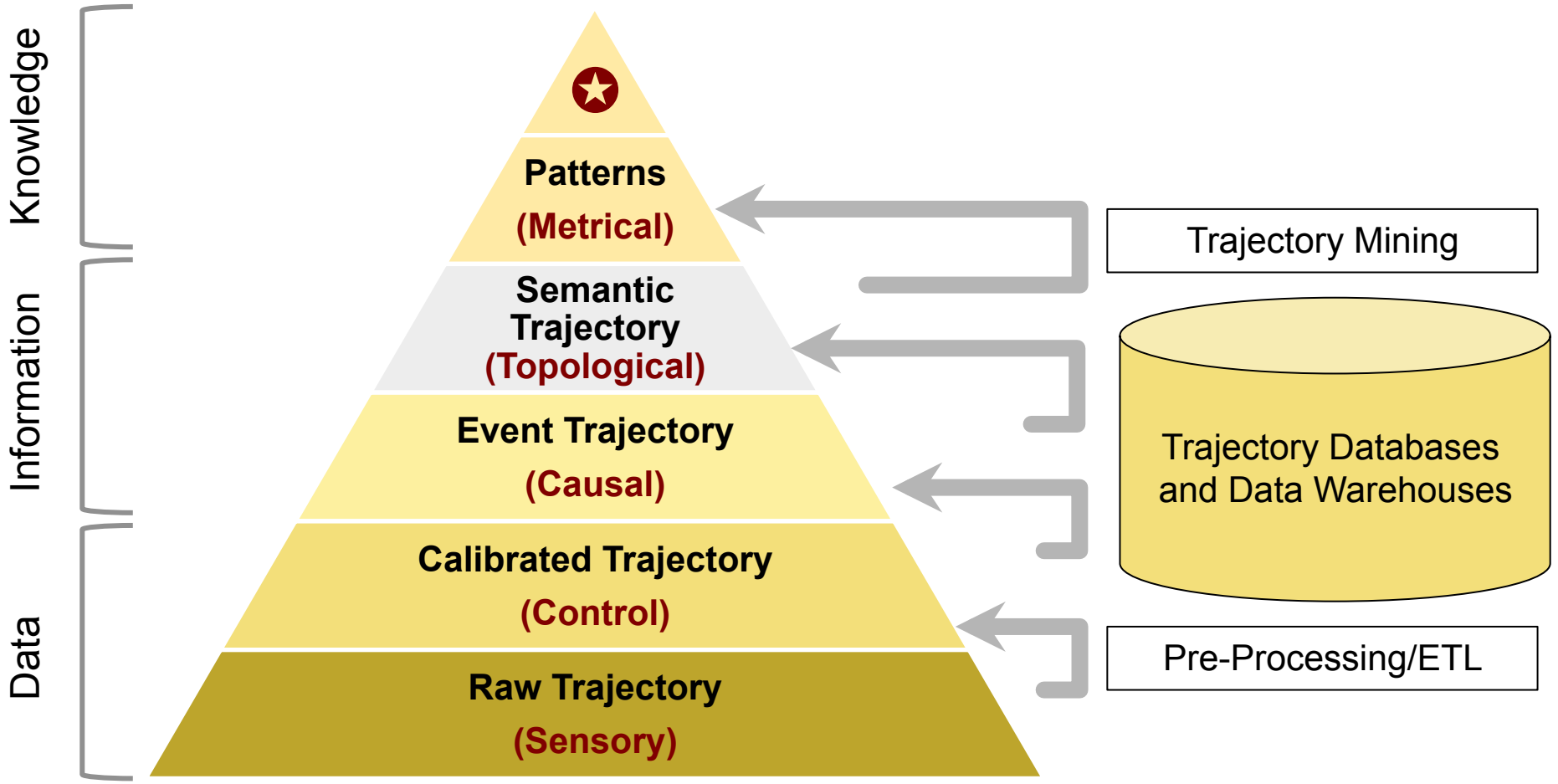
Is Trajectory Processing Hard?

- Extremely high level of redundancy
  - A lot of derivable data
- Poor data quality
  - Outliers, different sampling rates (from every second to a few days) and missing data (start and end of a trip, gap), and lack of semantics
- Poorly structured
  - Variable length, with both local and global operations
  - Can be unstructured data, especially for M2M data
- Easy to monitor whereabouts, but hard to monitor events, trends and performance in a system
- Large volume, rapid streaming and low value
  - Lack of semantic information
  - Values can only be unlocked when combined with other data

Trajectory Processing Framework



A Spatiotemporal Pyramid



Research Output Statistics

Topic	#Papers
Trajectory data preprocessing (simplification, compression, segmentation, map-matching) and **data quality management** (cleaning, uncertainty, calibration)	35
Trajectory databases (indexing, search, query processing, including distance measures)	31
Trajectory data mining and predication (co-traveller pattern, clustering, frequent pattern, popular routes, significant locations)	53
Semantic trajectory	4
Trajectory privacy protection	9
SUM	132

Since 2005, full papers published in major conferences including SIGMOD, VLDB, ICDE, EDBT, CIKM, SIGKDD, ICDM, ACM GIS

Four Main Research Challenges

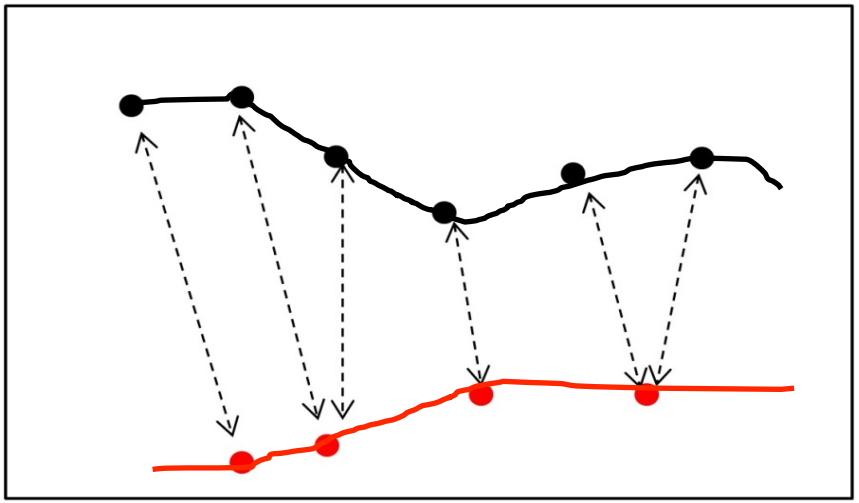
- Storage
  - How to store trajectory data (and what to store)?
  - Need to consider applications, compression, and parallel processing
- Similarity
  - How to measure trajectory similarity?
  - Have major impacts on data storage, indexing and processing
- Scalability
  - Can we process large trajectory datasets, for millions of users, in real-time?
  - Use of modern computing infrastructures
- Semantics
  - Semantics for stops/trips and identification of events/trends
  - Add “what” to “where” and “when”

Measuring Trajectory Similarities

- Spatial trajectory is the object movement history in a space
  - Continuous in nature, but discrete once captured and stored
- Many location-update strategies
  - By time, by distance, by deviation…
  - A trade-off between accuracy and other types of overhead
  - Variations may not always be under control
- Movement can be in a free space (e.g., flying), or a constrained space (e.g., cars in road networks)
- Trajectory similarity measure is a fundamental operations
- Have we done it right?

Trajectory Distance Measures

- Trajectory distance measures
  - ✓ Euclidean distance
  - ✓ LCSS (longest common sequence)
  - ✓ DTW (Dynamic time warping)
  - ✓ EDR (Edit distance on Real sequences)
- How trajectory distance measures work?
  - ✓ Align sample points on both trajectories
  - ✓ Accumulate sum of aligned sample pairs



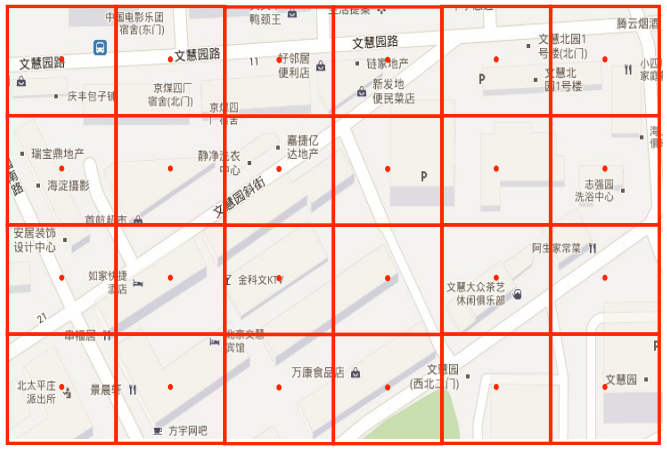
You Just Have to Rewrite…

Sampling rate	ED	DTW	LCSS	EDR
10	0.35	0.21	0.41	0.55
20	0.21	0.09	0.27	0.37
30	0	0	0	0
60	0.24	0.15	0.33	0.23
100	0.25	0.21	0.45	0.28

Choose a Reference System

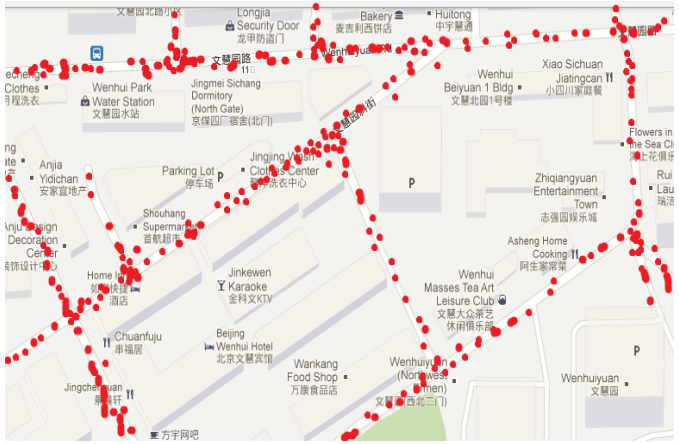
Goal:

Find a stable reference system for all trajectories



Grid Centroids

POIs



Achieved Points

Turning Points

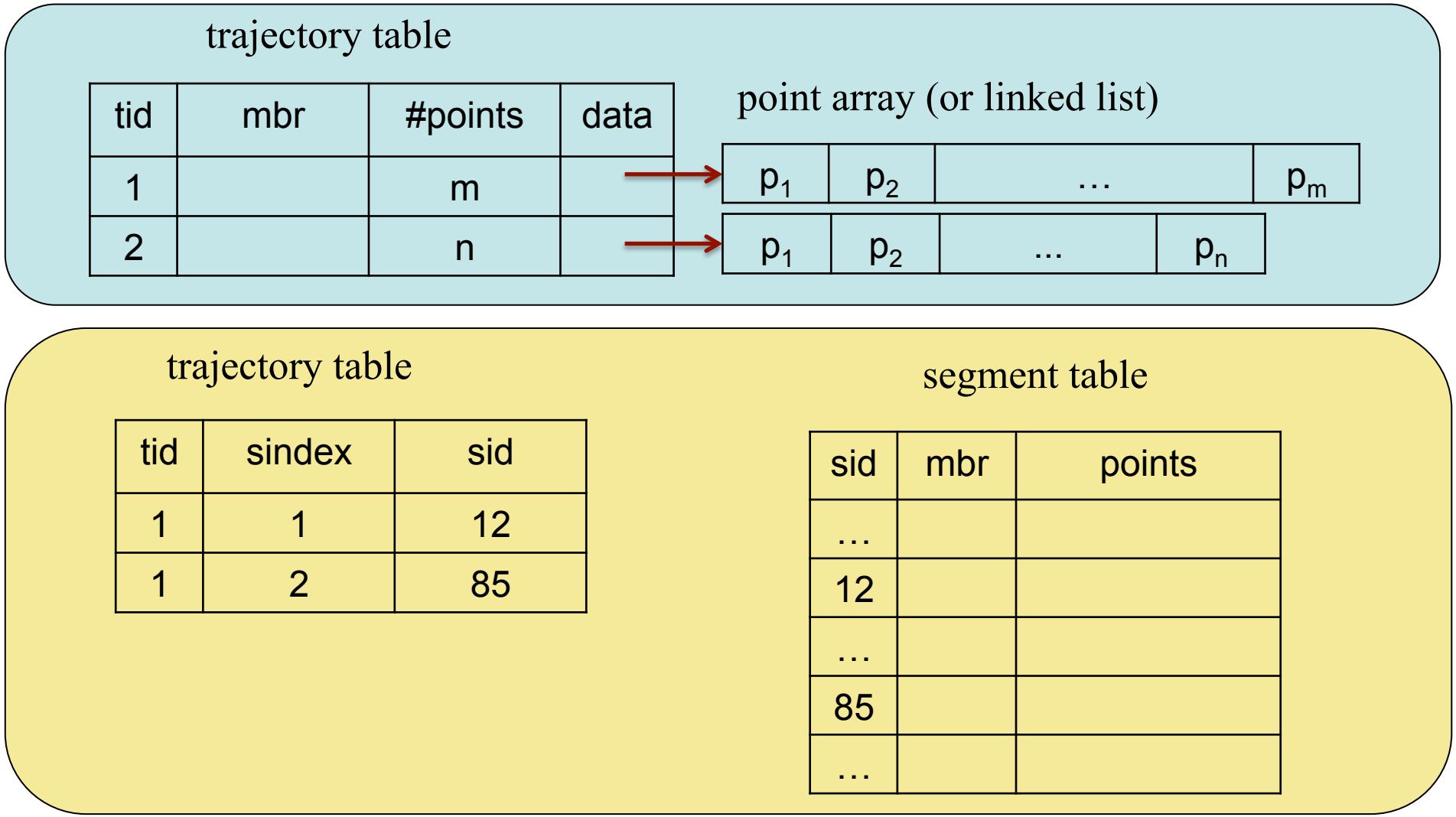
Data Calibration

Goal: Trajectory rewriting using the anchor points in the reference system

- Alignment:
  - Map each sample point of trajectory to an anchor points or drop it
  - Reduce redundant points
- Complement:
  - Insert missing anchor points to aligned trajectory
  - Increase trajectory size

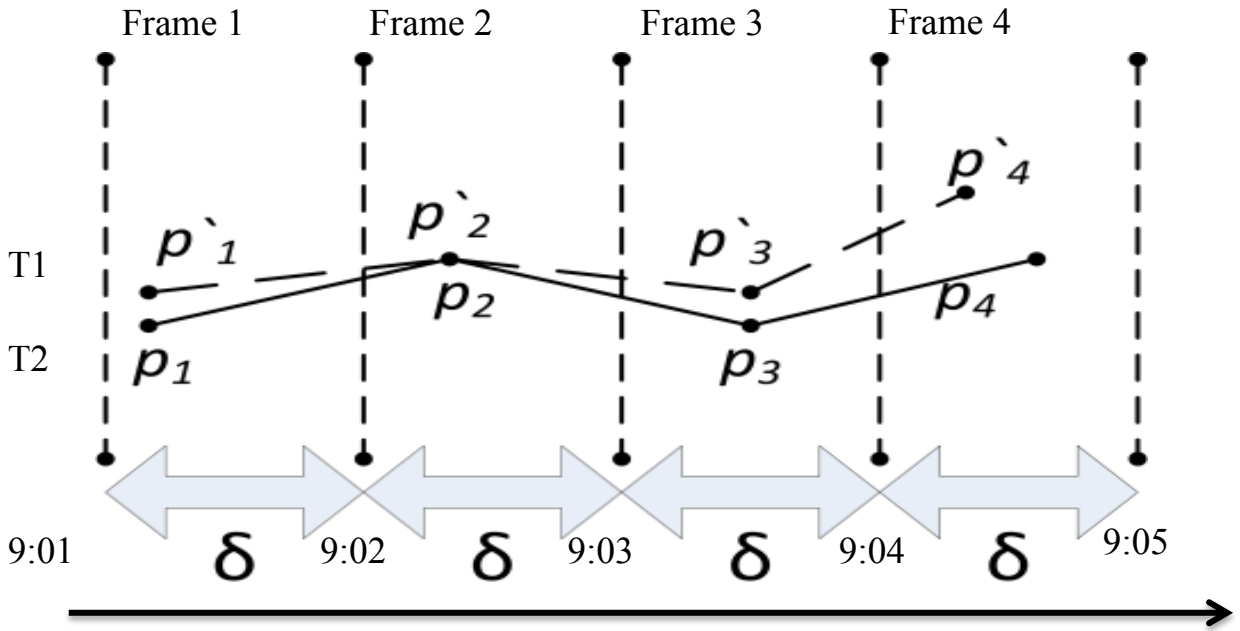
	Alignment	**Complement**
Geometry-based calibration	Nearest neighbour mapping	Linear insertion
Model-based calibration	Global alignment	Probability-based insertion

Row-based Structures



Frame Table - Concepts

- A frame contains all points at a particular time
- A fixed time-interval is chosen by the system for a given dataset
- All points are aligned to their nearest frame (possibly with linear interpolation and skipping)
- Subsequent frames can be delta-encoded, and possibly with some techniques used in video encoding (eg, IP-frame encoding)



SharkDB

- An in-memory trajectory database systems
  - Also implemented on SAP HANA
- Advantages
  - Column-based structure for high performance analytics
  - Can benefit many operations, such as
    - I-frame based localization, generalization & approximation
    - I-frame based indexing
  - Naturally designed for parallel processing
- Evaluation
  - Compression ratios, accuracy, and performance for typical operations
    - For a range of DB operations and advanced trajectory operations

Conclusions

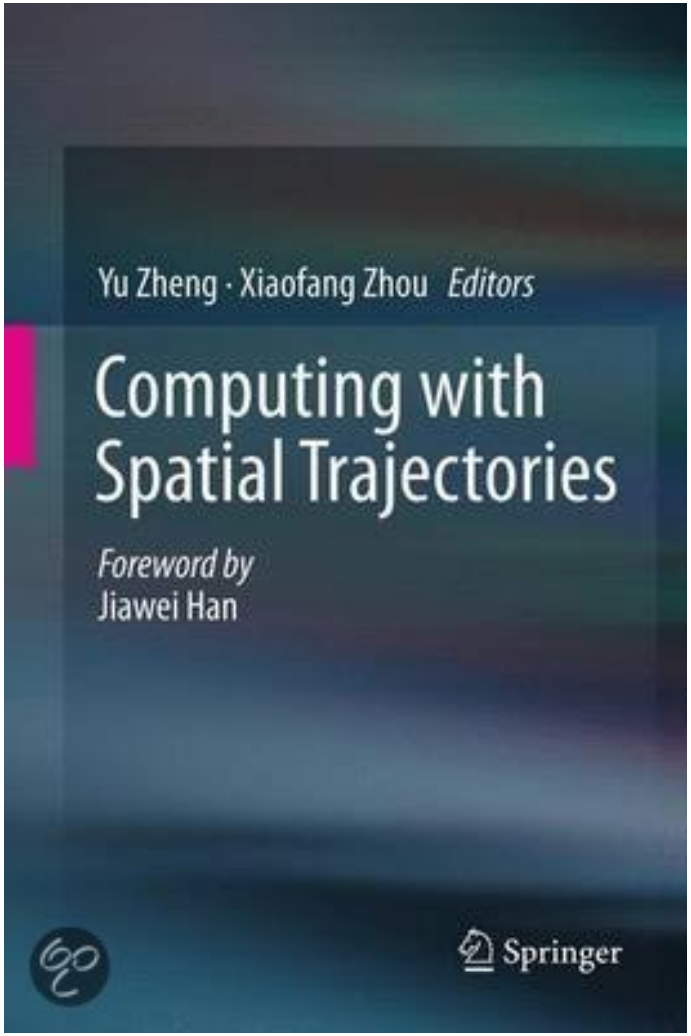
- Trajectory data is an important type of data
  - For its abundance, usefulness and complexity
- A system approach is essential
  - To manage, among others, data redundancy, relationship and reduction
  - To be a new data integration base for many applications
  - The spatiotemporal pyramid is a useful knowledge model
- More research effort needed
  - To address problems in storage, similarity, semantics and scalability
  - And also to address problems in privacy, crowd sourcing, data exploration and visualization

Some of Our Publications

Prediction of movement [ICDE 08] and paths [VLDBJ 10], trajectory simplification with error bound [VLDB 08], path nearest neighbor query [SIGMOD 09], searching trajectory by locations [SIGMOD 10], most popular routes [ICDE 11], probabilistic range query [EDBT 11, ICDE12], materialized shortest paths [TODS 12], spatial keyword search for trajectories [ICDE 13,15], trajectory calibration [SIGMOD 13], route and location recommendation [ICDE 14, SIGKDD 15], trajectory exploration and summarization [ICDE 15], in-memory spatial databases [CIKM14, SIGMOD 15], privacy-preserving search [ICDE 15]

An Introduction Book

- ***Computing with Spatial Trajectories***
  - Yu Zheng and Xiaofang Zhou, 2011
- Part I Foundations
  - Trajectory Preprocessing
  - Trajectory Indexing and Retrieval
- Part II Advanced Topics
  - Uncertainty in Spatial Trajectories
  - Privacy of Spatial Trajectories
  - Trajectory Pattern Mining
  - Activity Recognition from Trajectory Data
  - Trajectory Analysis for Driving
  - Location-Based Social Networks: Users
  - Location-Based Social Networks: Locations.



THANKS

zxf@suda.edu.cn