ACM SIGMOD PODS 2022
Philadelphia, PA, USA

Aalborg Universitet

Roskilde Universitet

Monash University

Spatial Data Quality in the IoT Era

Management and Exploitation

Huan Li, Bo Tang, Hua Lu, Muhammad Aamir Cheema, Christian S. Jensen

3.
SID QUALITY MANAGEMENT

By Huan Li

Outlines

1. Location Refinement
2. Uncertainty Elimination
3. Outlier Removal
4. Fault Correction
5. Data Integration
6. Data Reduction

Definitions, Categories, and Representatives

1. Location Refinement (LR)

- ▷ Accompanies/follows localization $f : \mathbf{X} \mapsto \mathbf{Y}$

  IoT measurement *location*

- ▷ Adjust initial estimate
  - precision↑, accuracy↑, resolution↑

    $$\arg\max_{\hat{\mathbf{y}} \in \mathbf{Y}} \mathsf{P}(\mathbf{Y} \mid \mathbf{X}, F, C)$$

    optimal result *a family of positioning functions* *spatial constraints*

  - **Ensemble LR, Motion-based LR, Collaborative LR**

Ensemble LR

- ▷ $\mathbf{X}$: individual, multivariable, single time point
  - Different components measured by different sensors

- ▷ **Single-source methods**
  - Aggregate $\mathbf{y} = \{y_1, \ldots\}$ by a single process $\mathsf{f}(\mathbf{x})$
  - Weighted *k*NN and its variants [Fang et al., 2018]

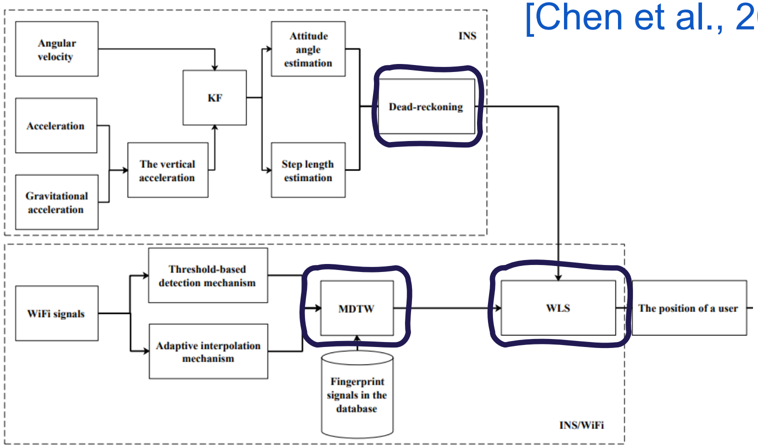
$$\hat{\mathbf{y}} = \sum_{j=1}^{k} \omega_j \cdot y_j$$

the likelihood $\mathsf{P}(y_j \mid \mathbf{x})$

Ensemble LR (Cont.)

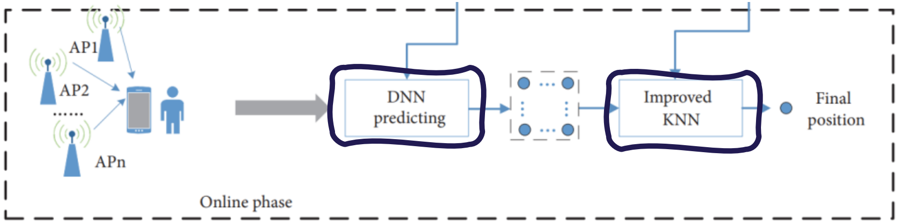
- **Multi-source methods**
  - Fuse multiple procedures $F = \{f_1, \dots\}$

[Chen et al., 2018]



weighted_least_squares (WiFi fingerprinting, Dead Reckoning)

[Dai et al., 2019]



improved_KNN (candidates <- DNN(·))

***NB**: multi-aspect information from a more complex deployment setting -> higher accuracy*

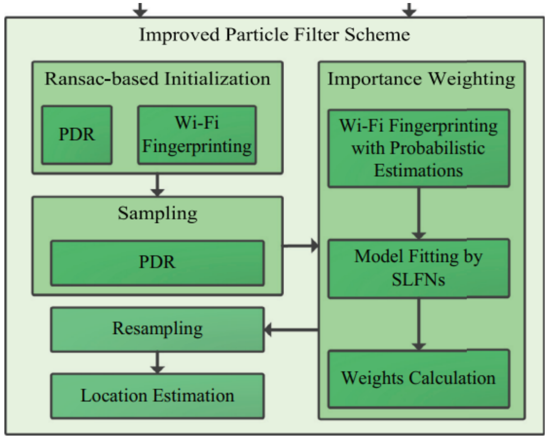
Motion-based LR

- $\mathbf{X}$: individual, sequential, single-variable or multivariable
  - Motion dynamics, spatiotemporal dependencies

Bayes Filters (Kalman filters, Particle filters, etc.) [Wu et al., 2016]	**Pro**: easy-to-implement **Con**: intricate dependencies
Probabilistic Graph Models (HMM, CRF, etc.) [Liu et al., 2012]	**Pro**: incorporate domain knowledge **Con**: non-discrete locations
Sequential Neural Networks (e.g., RNN [Hoang et al., 2019])	**Pro**: complex scenes **Con**: training data volume
Opportunity: decentralized computing setting?	

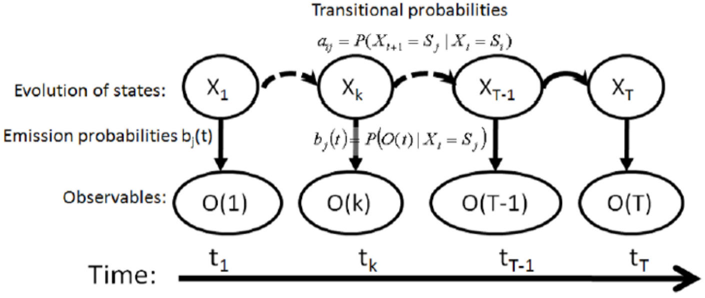
Motion-based LR (Cont.)

Bayes Filter [Wu et al., 2016]



Particle Filter, a sequential Monte Carlo process

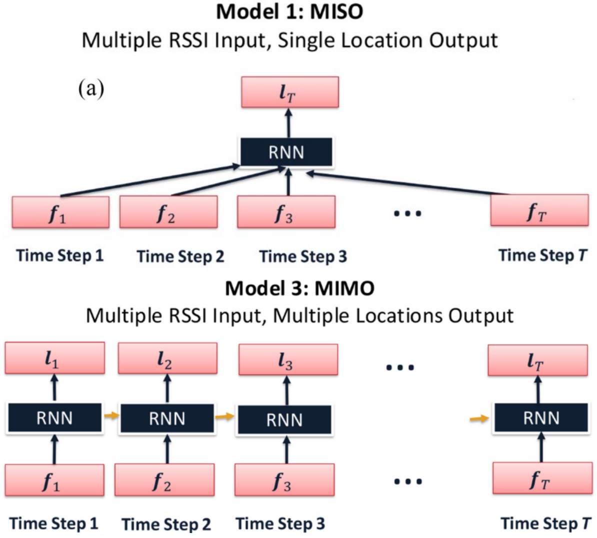
PGM [Liu et al., 2012]



HMM(S, 0, A, B, π)

S: grid-based locations

RNN [Hoang et al., 2019]



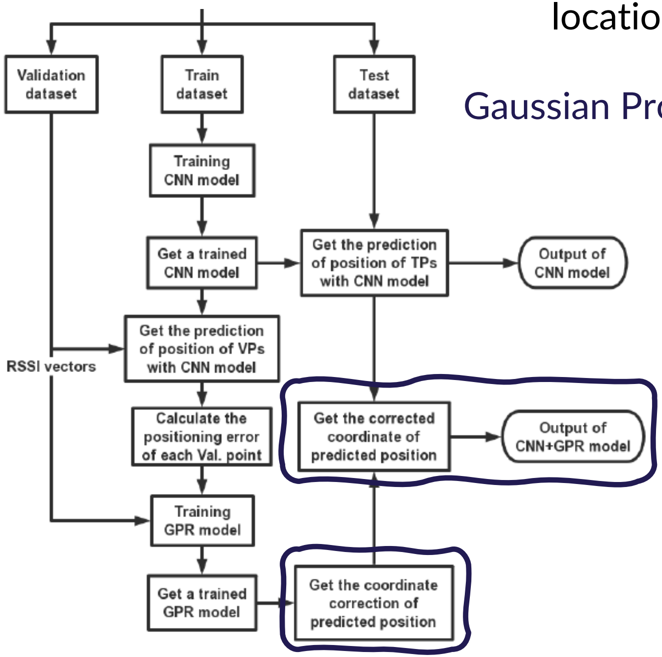
Collaborative LR

- $\mathbf{X}$: multiple objects, single time
  - Refine object locations collectively
- Joint Denoising [Zhang et al., 2019]
  - System noise, statistical hypothesis
- Iterative Optimization [Chen et al., 2017]
  - Random errors, evolutionary computation
- **Opportunity**: data and control coordination

Collaborative LR (Cont.)

Joint Denosing

[Zhang et al., 2019]



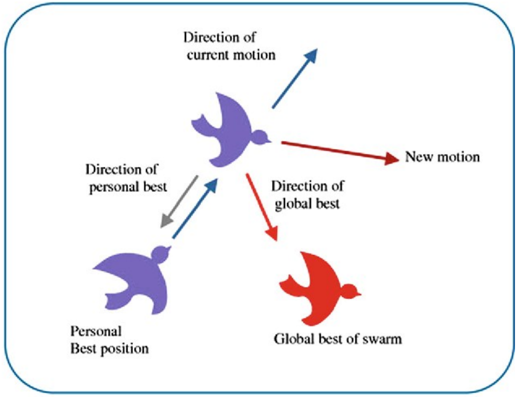
Gaussian errors for CNN location estimator

Gaussian Process Regression

Iterative Optimzation

[Chen et al., 2017]

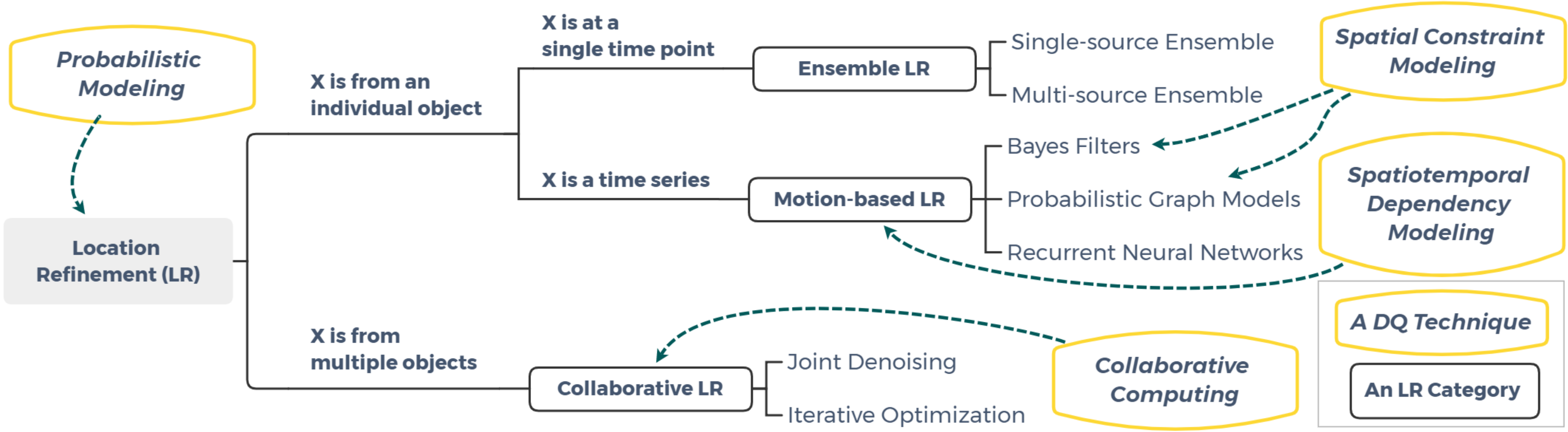
1



Trilateral estimates are particles

Particle Swarm Optimization (PSO)

[1] image originally from [Ahmadi et al. 2010]



- Most LR techniques rely on probabilistic modeling
- Markovian widely utilized in motion-based LR;
  - Spatial constraints -> Particle Filters and PGMs
- Motion-based LR compared to Ensemble/Collaborative LR
  - Higher accuracy but more ground truth to parameterize models

2. Uncertainty Elimination (UE)

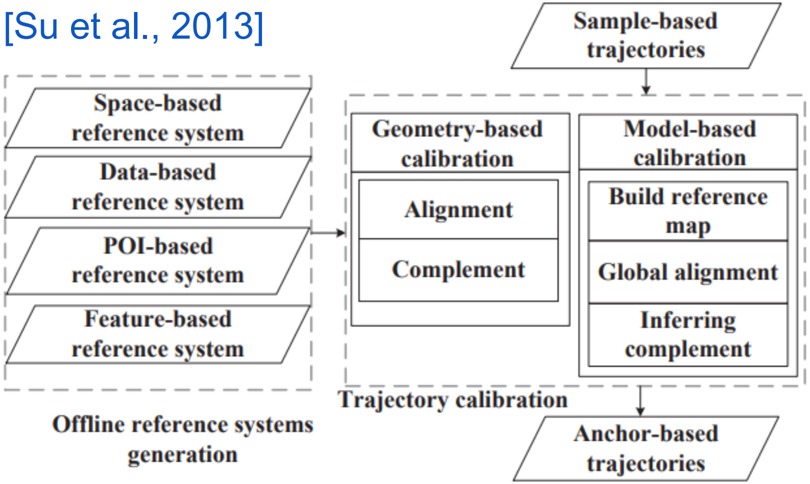
- Reduces uncertain/imprecise measurements, imputes values at unsampled points
  - precision↑, completeness↑, resolution↑, time sparsity↓
- UE for **trajectories** and spatiotemporal IoT data (**STID**)

Trajectory UE

- **Smoothing-based**
  - Temporal autocorrelation, volatility
  - Moving averages, exponential smoothing, and random walks
- **Calibration-based**
  - Reference points/ranges from maps [Su et al., 2013] or extracted from collective trajectory data [Li et al., 2020]
- **Inference-based**
  - Structural regularities, restore underlying path
  - Explicit (topology) and implicit (observations) [Wu et al., 2016]

Trajectory UE (Cont.)

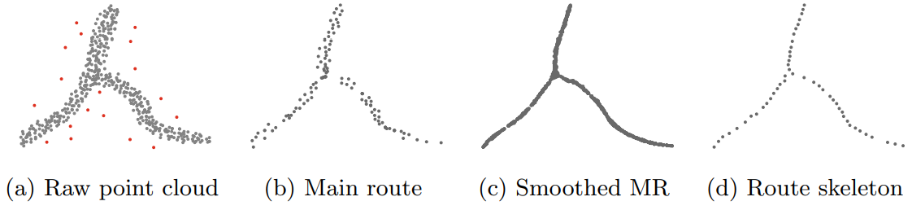
Calibration-based
[Su et al., 2013]



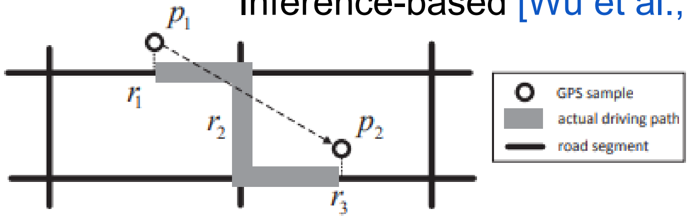
Anchors from maps and trajectories

Calibration-based
[Li et al., 2020]

Skeleton points from trajectories only



Inference-based [Wu et al., 2016]



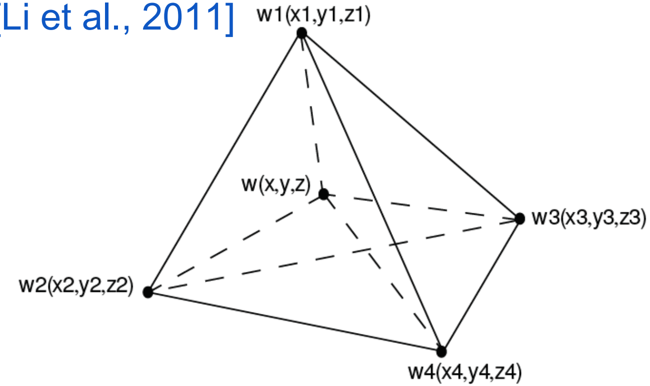
Posterior over historical trajectories

STID UE

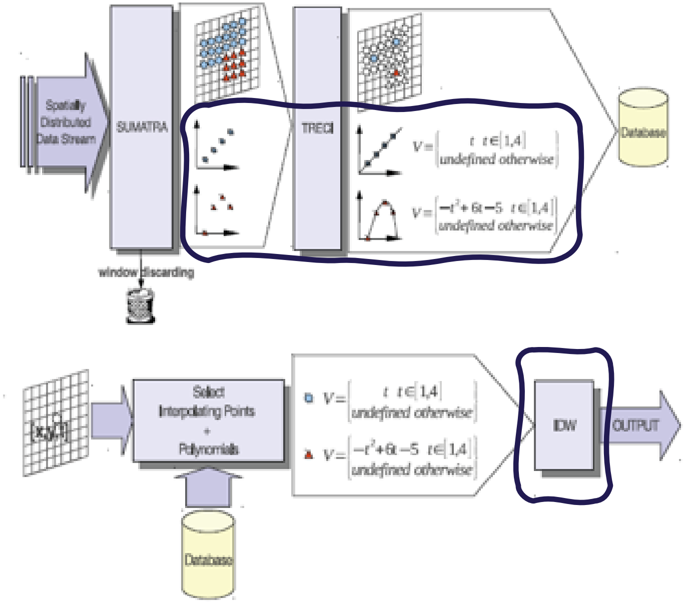
- **Spatiotemporal interpolation**
  - Unsampled location-time points
  - *Spatial-interpolation-primitive* (shape function, inverse distance weighting, Kriging, etc.)
    - **Tobler's first law**: near things more related than distant things
  - *Time-interpolation-primitive* (neighbor-based, regression-based, matrix factorization, LSTM/GRU, etc.)
  - *Space and time simultaneously* [Li et al., 2011] [Appice et al., 2013]
- **Data fusion**: calibration models [Okafor et al., 2020]
  - Additional relevant and reliable data sources?

STID UE (Cont.)

[Li et al., 2011]



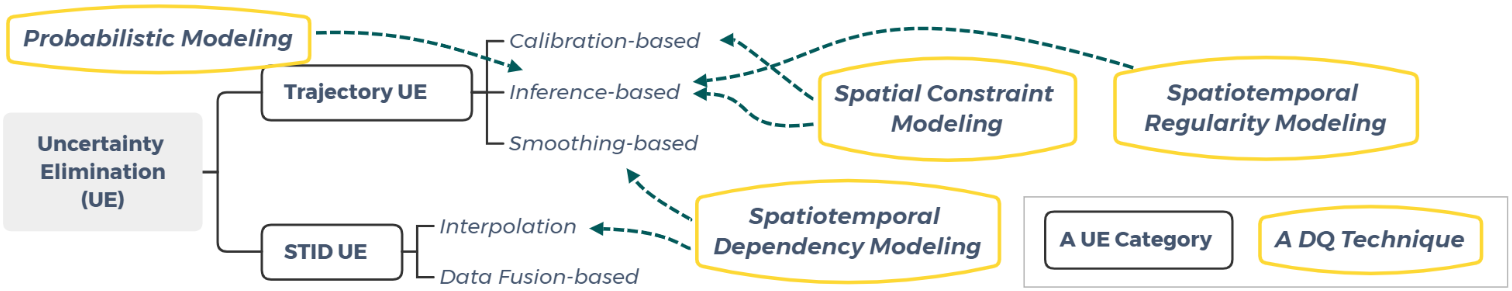
Shape function: time as imaginary third dimension z.



[Appice et al., 2013]

Offline: trend cluster -> key sensors as a polynomial time interpolator

Online: nearby key sensors, temporally interpolated values -> IDW based interpolation



- Calibration/inference-based UE utilize spatial constraints and collective trajectories

- Smoothing-based UE -> varying smoothly and Markovian
    - Stream computing, fog/edge computing

- Interpolation -> spatiotemporal dependencies
    - Varying smoothly, spatially autocorrelated/anisotropic

3. Outlier Removal (OR)

- Deletes items that do not conform to their context
  - precision↑, accuracy↑, consistency↑
- OR for **trajectory points** and **STID**
  - Anomaly trajectories? A business layer task

Trajectory Point OR

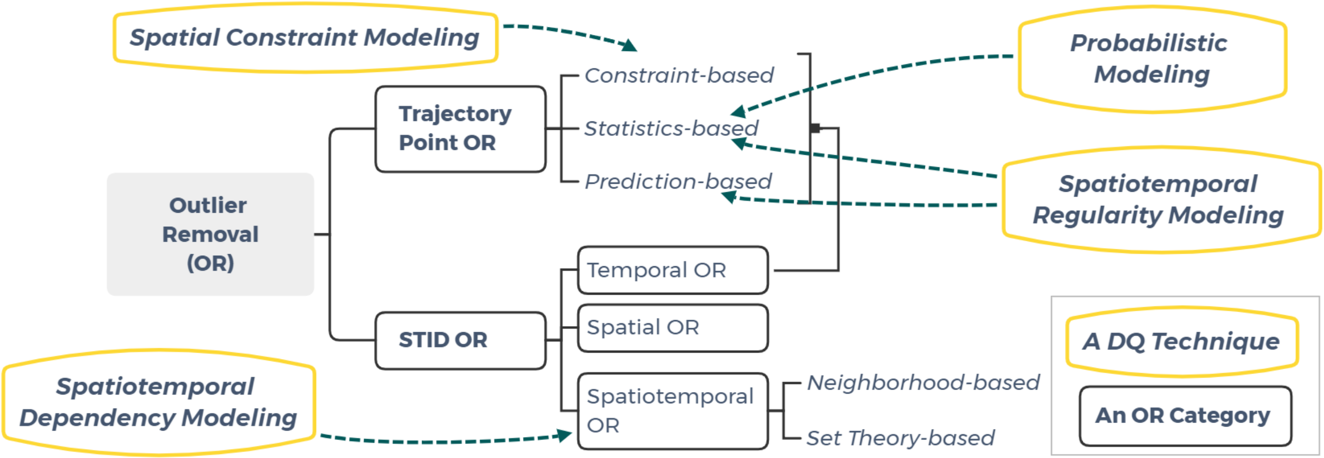
▷ Location points corresponding to unexpected *abnormal* mobility behavior

Constraint-based neighborhood information	Speed thresholding [Zheng, 2015]	**Pros**: easy-to-implement **Cons**: dynamic and noisy trajectories?
Statistics-based statistical profiling of a trajectory or a trajectory set	Z-test using a combination of distance, speed and acceleration [Patil et al., 2018]	**Pros**: controllable and explainable **Cons**: availability of historical trajectories?
Prediction-based compare with predictions	Iterative minimum repair with an ARX model [Zhang et al., 2017]	**Pros**: data repairing **Cons**: achieve accurate predictions?

STID OR

- W.r.t spatial/temporal/spatiotemporal neighbors
  - Temporal outliers [Gupta et al., 2013] [Blázquez-García et al., 2021]
  - Spatial outliers (fundamental step) -> spatiotemporal outliers [Aggarwal, 2015]
- Spatiotemporal OR

Neighborhood-based	Spatiotemporal DBSCAN [Birant et al., 2007]	Decoupling of spatial and temporal aspects
Set theory-based	Rough/kernel set [Albanese et al., 2012]	Holistic, simple data attributes



- Probabilistic modeling, spatiotemporal dependencies and regularity, spatial constraints
  - Some follow unsupervised learning paradigm

- Constraint/prediction-based approaches can be implemented a stream computing fashion

4. Fault Correction (FC)

- Repairs wrong and conflicting data values
  - accuracy↑, consistency↑, completeness↑

- **Symbolic trajectories** and **STID**
  - Each location in a symbolic trajectory is an ID of the sensor that detected that object at that time, e.g., RFID tracking sequences

Symbolic Trajectory FC

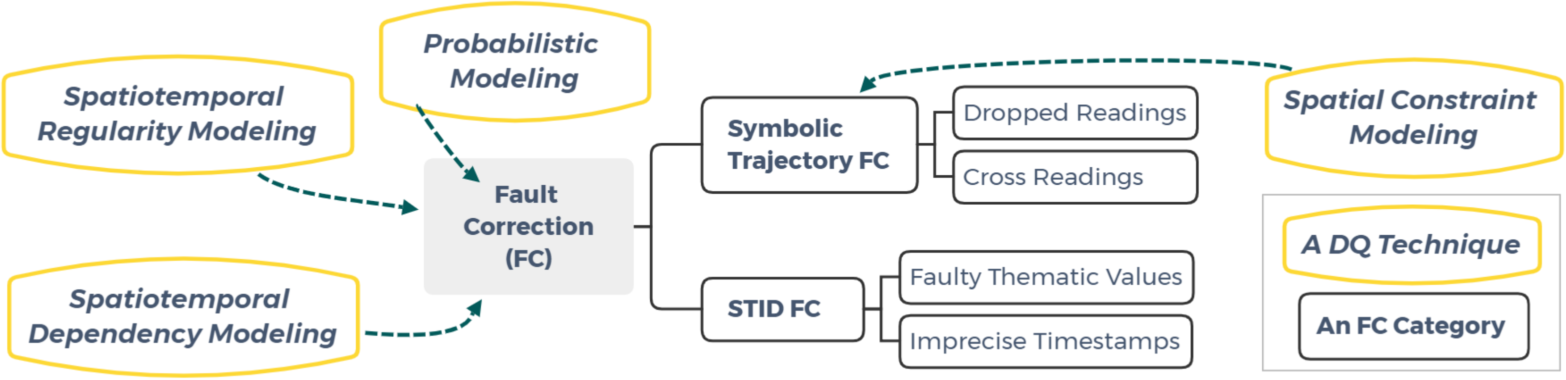
- False Negatives (dropped readings) – a sensor fails to detect a tag (object)
- False Positives (duplicated readings) – a sensor fails to detect tag movement

	Probabilistic modeling	Regularities of sensor-tag interactions	Spatiotemporal dependencies	Spatial constraints
[Jeffery et al., 2006]	Tag as random sample	Per-tag and multi-tag cleaning	Smoothing filter	
[Chen et al., 2010]	Bayesian inference	Likelihood that a reader reports an object	MCMC-based sampler	Resource descriptors
[Fazzinga et al., 2016]	Probabilistic trajectories	Conditioned trajectory graph	Conditioning over time	Unreachability, traveling time, latency
[Baba et al., 2016]	Multi-variate HMM	Emission probabilities	Transition probabilities	Deployment, hidden state semantics

STID FC

- Correct **faulty thematic values**
  - Neighboring/correlated homogeneous sensors [Pumpichet et al., 2012]
  - Cross-validation of heterogeneous sensory information [Kuemper et al., 2018]

- Correct **imprecise timestamps**
  - **Staleness**: spatiotemporal dependencies [Milani et al., 2019]
  - **Inconsistency:** temporal constraint violations [Song et al., 2016]
  - **Disorder**: K-slack that buffers the arriving data for K time units for reordering, distributed setting [Mutschler et al., 2013], heterogeneous network setting [Liu et al., 2009]



- Spatiotemporal regularities and dependencies from existing data
- Correcting incoming symbolic trajectories by data-driven models
- K-slack for disorder resolution in a stream and/or distributed computing mode

5. Data Integration (DI)

- Unified data representation
- Comparing, combining, and fusing data collections from multiple sources
  - accuracy↑, completeness↑, data volume↑, resolution↑, interpretability↑

- **Semantic** DI and **non-semantic** DI

Semantic DI

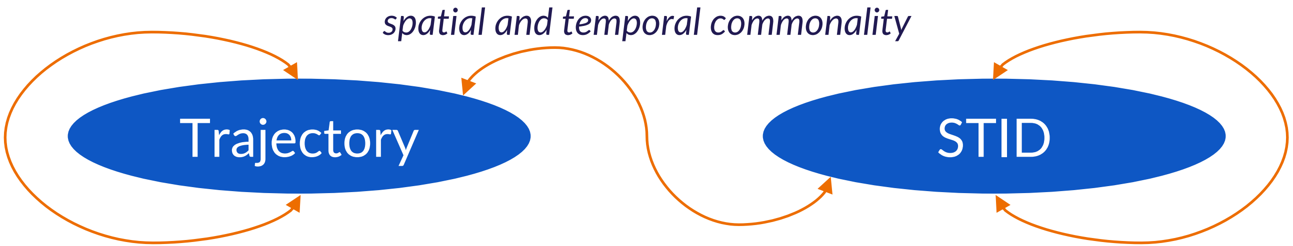
- Enriches interpretability of SID
- Semantic DI for **trajectories**: concepts or events -> raw traces

	[Wu et al., 2015]	[Nogueira et al., 2018]	[Liao et al., 2007]	[Yan et al., 2013]
Semantic aspect	Social media posts	Semantic web concepts	Transportation location/mode	Stop/move and POI categories
Method	Relevant word extraction using kernel density estimation (KDE)	Self-defined functions to map spatial features to tags/ontology instances	Hierarchical CRF to map GPS data to transportation concepts	Speed-based stop identification. HMM to infer POI category
Significance	Dynamic semantics enrichment	Reasoning/analysis of trajectories	Unsupervised (EM) method	Third-party semantic sources

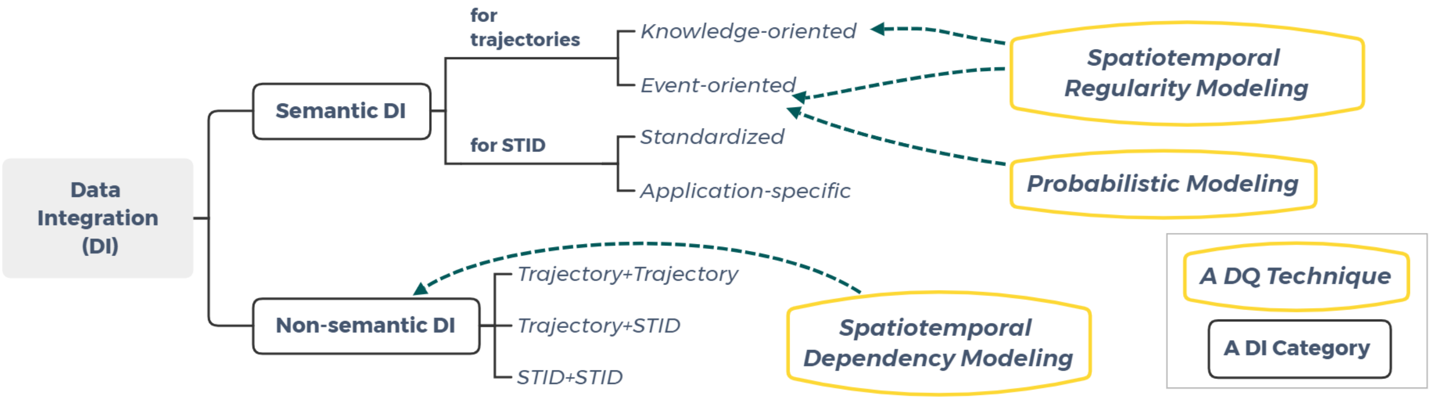
- Semantic DI for **STID**: geo-semantic meta information
  - Reasoning system [Maarala et al., 2016], Web of Things ontology [Wu et al., 2017]
  - **Opportunity**: dynamically evolving semantics

Non-semantic DI

- Multi-angle spatiotemporal observations
  - Consistency and reliability



unified trajectories	**enriched trajectories**	**measurement fusion**
formats [Peixota et al., 2018] resolutions [Zhang et al., 2019-2] ID systems [Jin et al., 2020]	comprehensive urban analytics [Zhang et al., 2014]	multisource remote sensing data [Zhu et al., 2018]



- Semantic DI for trajectories: spatiotemporal data regularity
  - Geo-semantics, spatial constraints, personal preferences
- Semantic DI for STID
  - Edge computing and stream computing
- Non-semantic DI utilizes spatial and temporal commodities

6. Data Reduction (DR)

- ▷ Converts a data collection into a corrected and simplified form
- ▷ Eliminate meaningless items, Reconstruction/Summary
  - data volume↓, latency↓, redundancy↓
- ▷ **Trajectory compression**
- ▷ **STID reduction**

Trajectory Compression

▷ **Lossy solution**: Compression ratio (size, number) vs compression loss (error, cost)

	Offline (all points are accessible)	**Online (only buffered points are accessible)**
Free-space (raw) Trajectory	• ε-simplification (Hausdorff) with least space-location points (min-# problem), Douglas–Peucker (DP) [Cao et al., 2006] • Min-distance-preserving-error with a fixed #, binary search strategy [Long et al., 2014] • Min-max (DTW) of using sub-trajectories as references, greedy and optimal algorithms [Zhao et al., 2018]	• ε-bounded and time-limited, wireless communication cost reduction with dead reckoning [Lange et al., 2011] • Min-SED, priority queue [Muckell et al., 2011] • Min-geometric-error, convex hull bounding [Liu et al., 2015] • Min-SED, cone intersection [Lin et al., 2019] • Min-error as MDP [Wang et al., 2021]
Network-constrained (map-matched) trajectory	• Min-# against road segment discontinuity, adapted model + DP + SED metric [Popa et al., 2015] • Encoding spatial paths/time sequences [Han et al., 2017] • TED (encoding timestamps, relative spatial path, and distances) [Yang et al., 2017] • Retrieval, compressed substring index [Koide et al., 2018] • Probabilistic trajectories, referential +TED [Li et al., 2020-2]	• Heading change based compression [Chen et al., 2019] • Transmission cost at edge, referential representation online fashion [Li et al., 2021]

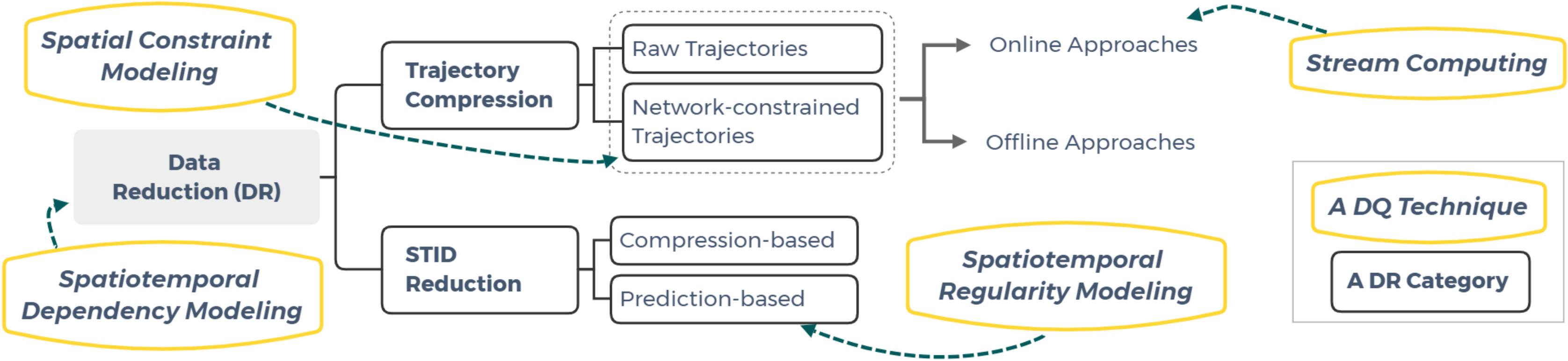
STID Reduction

▷ **Compression-based**: batch processing

Lossless accuracy-oriented	• Golomb-Rice codes [Tate et al., 2015] • Gaussian approximation + lossless margins [Abuadbba et al., 2017]
Lossy higher compression ratio	• Lightweight temporal compression algorithm [Li et al., 2018] • SVD (singular value decomposition) [de Souza et al., 2015] • Compressive sampling + Gaussian Mixture Model [Tripathi et al., 2018]

▷ **Prediction-based**: drop data if predicted error is acceptable

- Reduce communications among IoT nodes: Regression [Carvalho et al., 2011], KF [Yin et al., 2015], CNN+LSTM [Zhang et al., 2018]
- Robustness/timeliness?



- Spatiotemporal data dependencies widely utilized for DR
- Prediction-based DR challenged by robustness and timeliness of prediction models
- Edge computing
  - Reduce data at resource-scarce IoT edge devices

References

- [Fang et al., 2018] Optimal weighted K-nearest neighbour algorithm for wireless sensor network fingerprint localisation in noisy environment. *IET Communications*.
- [Chen et al., 2018] An INS/WiFi indoor localization system based on the weighted least squares. *Sensors*.
- [Dai et al., 2019] Combination of DNN and improved KNN for indoor location fingerprinting. *Wirel. Commun. Mob. Comput.*.
- [Wu et al., 2016] Improved particle filter based on WLAN RSSI fingerprinting and smart sensors for indoor localization. *Computer Communications*.
- [Liu et al., 2012] A hybrid smartphone indoor positioning solution for mobile LBS. *Sensors*.
- [Hoang et al., 2019] Recurrent neural networks for accurate RSSI indoor localization. *IEEE IoTJ*.

References

- [Zhang et al., 2019] Wireless indoor localization using convolutional neural network and Gaussian process regression. *Sensors*.
- [Chen et al., 2017] Improved Wi-Fi indoor positioning based on particle swarm optimization. *IEEE Sensors Journal*.
- [Su et al., 2013] Calibrating trajectory data for similarity-based analysis. *SIGMOD*.
- [Li et al., 2020] A data-driven approach for GPS trajectory data cleaning. *DASFAA*.
- [Wu et al., 2016] Probabilistic robust route recovery with spatio-temporal dynamics. *KDD*.
- [Li et al., 2011] Spatiotemporal interpolation methods for air pollution exposure. *SARA*.
- [Appice et al., 2013] Using trend clusters for spatiotemporal interpolation of missing data in a sensor network. *J. Spat. Inf. Sci.*.
- [Zheng, 2015] Trajectory data mining: An overview. *TIST*.

References

- ▷ [Patil et al., 2018] Geosclean: Secure cleaning of GPS trajectory data using anomaly detection. *MIPR*.
- ▷ [Zhang et al., 2017] Time series data cleaning: From anomaly detection to anomaly repairing. *PVLDB*.
- ▷ [Gupta et al., 2013] Outlier detection for temporal data: A survey. *TKDE*.
- ▷ [Blázquez-García et al., 2021] A review on outlier/anomaly detection in time series data. *CSUR*.
- ▷ [Aggarwal, 2015] Outlier analysis. *Data Mining* Book.
- ▷ [Birant et al., 2007] ST-DBSCAN: An algorithm for clustering spatial-temporal data. *DKE*.
- ▷ [Albanese et al., 2012] Rough sets, kernel set, and spatiotemporal outlier detection. *TKDE*.

References

- [Jeffery et al., 2016] Adaptive cleaning for RFID data streams. *PVLDB*.
- [Chen et al., 2016] Leveraging spatio-temporal redundancy for RFID data cleansing. *SIGMOD*.
- [Fazzinga et al., 2016] Exploiting integrity constraints for cleaning trajectories of RFID-monitored objects. *TODS*.
- [Baba et al., 2016] Learning-based cleansing for indoor RFID data. *SIGMOD*.
- [Pumpichet et al., 2012] Belief-based cleaning in trajectory sensor streams. *ICC*.
- [Kuemper et al., 2018] Valid.IoT: A framework for sensor data quality analysis and interpolation. *MMSys*.
- [Milani et al., 2019] CurrentClean: Spatio-temporal cleaning of stale data. *ICDE*.
- [Song et al., 2016] Cleaning timestamps with temporal constraints. *PVLDB*.

References

- [Mutschler et al., 2013] Distributed low-latency out-of-order event processing for high data rate sensor streams. *IPDPS*.
- [Liu et al., 2009] Sequence pattern query processing over out-of-order event streams. *ICDE*.
- [Wu et al., 2015] Semantic annotation of mobility data using social media. *WWW*.
- [Nogueira et al., 2018] FrameSTEP: A framework for annotating semantic trajectories based on episodes. *Expert Syst. Appl.*.
- [Liao et al., 2007] Learning and inferring transportation routines. *Artif. Intell.*.
- [Yan et al., 2013] Semantic trajectories: Mobility data computation and annotation. *TIST*.
- [Maarala et al., 2016] Semantic reasoning for context-aware Internet of Things applications. *IEEE IoTJ*.

References

- [Wu et al., 2017] Towards a semantic web of things: A hybrid semantic annotation, extraction, and reasoning framework for cyber-physical system. *Sensors*.
- [Peixoto et al., 2018] A system for spatial-temporal trajectory data integration and representation. *DASFAA*.
- [Zhang et al., 2019-2] National-scale traffic model calibration in real time with multi-source incomplete data. *ACM Trans. Cyber-Phys. Syst.*.
- [Jin et al., 2020] Trajectory-based spatiotemporal entity linking. *TKDE*.
- [Zhang et al., 2014] Exploring human mobility with multi-source data at extremely large metropolitan scales. MobiCom.
- [Zhu et al., 2018] Spatiotemporal fusion of multisource remote sensing data: Literature survey, taxonomy, principles, applications, and future directions. Remote Sensing.
- [Cao et al., 2006] Spatio-temporal data reduction with deterministic error bounds. *VLDBJ*.

References

- [Long et al., 2014] Trajectory simplification: On minimizing the direction-based error. *PVLDB*.
- [Zhao et al., 2018] REST: A reference-based framework for spatio-temporal trajectory compression. *KDD*.
- [Lange et al., 2011] Efficient real-time trajectory tracking. *VLDBJ*.
- [Muckell et al., 2011] SQUISH: An online approach for GPS trajectory compression. *Com.Geo*.
- [Liu et al., 2015] Bounded quadrant *system*: Error-bounded trajectory compression on the go. *ICDE*.
- [Lin et al., 2019] One-pass trajectory simplification using the synchronous Euclidean distance. *VLDBJ*.
- [Wang et al., 2021] Trajectory simplification with reinforcement learning. *ICDE*.

References

- [Popo et al., 2015] Spatio-temporal compression of trajectories in road networks. *GeoInformatica*.
- [Han et al., 2017] COMPRESS: A comprehensive framework of trajectory compression in road networks. *TODS*.
- [Yang et al., 2017] A novel representation and compression for queries on trajectories in road networks. *TKDE*.
- [Koide et al., 2018] CiNCT: Compression and retrieval for massive vehicular trajectories via relative movement labeling. *ICDE*.
- [Li et al., 2020-2] Compression of uncertain trajectories in road networks. *PVLDB*.
- [Chen et al., 2019] TrajCompressor: An online map-matching-based trajectory compression framework leveraging vehicle heading direction and change. *TIST*.

References

- [Li et al., 2021] TRACE: Real-time compression of streaming trajectories in road networks. *PVLDB*.
- [Tate et al., 2015] Preprocessing and Golomb-Rice encoding for lossless compression of phasor angle data. *IEEE Trans Smart Grid*.
- [Abuadbba et al., 2017] Gaussian approximation-based lossless compression of smart meter readings. *IEEE Trans Smart Grid*.
- [Li et al., 2018] A multi-dimensional extension of the lightweight temporal compression method. *IEEE Big Data*.
- [de Souza et al., 2015] Data compression in smart distribution systems via singular value decomposition. *IEEE Trans Smart Grid*.
- [Tripathi el al. 2018] An efficient data characterization and reduction scheme for smart metering infrastructure. *IEEE TII*.

References

- [Carvalho et al., 2011] Improving prediction accuracy for WSN data reduction by applying multivariate spatio-temporal correlation. *Sensors*.
- [Yin et al., 2015] An efficient data compression model based on spatial clustering and principal component analysis in wireless sensor networks. *Sensors*.
- [Zhang et al., 2018] An efficient neural-network-based microseismic monitoring platform for hydraulic fracture on an edge computing architecture. *Sensors*.
- [Ahmadi et al., 2010] Flocking based approach for data clustering. *Nat. Comput.*.
- [Okafor et al., 2020] Improving data quality of low-cost IoT sensors in environmental monitoring networks using data fusion and machine learning approach. *ICT Express*.