Digitalization in the Service of Society: The Case of Big Vehicle Trajectory Data

Christian S. Jensen

www.cs.aau.dk/~csj

Setting: big data

Instrumentation and Digitization

- Instrumentation of reality (2022)
  - 6.6 billion smartphone users, 7.3 billion phone users
  - 10.6 billion mobile connections
  - World population: 7.9 billion
- Digitization of processes
  - E.g., e-commerce, public services, communications, social interactions, aspects of transportation
- We are at an unprecedented time in the history of humanity

https://www.bankmycell.com/blog/how-many-phones-are-in-the-world

Big Data

- Unprecedented data and computing infrastructure combine to offer potential for value creation.
  - “Data is the new oil” or “data is the new soil”

- To be competitive, society and businesses must be able to create value from data.
- Data-based decisions and data-driven processes
  - Decisions based on good data beat decisions based on feelings or opinions.

- Approximation: Every company is an IT company

- A finer granularity of services and entirely new services

Didi – Ridesharing and Beyond

"Didi Chuxing ("DiDi") is the world's leading mobile transportation and local services platform. The company offers a full range of app-based services for over 550 million users across Asia Pacific, Latin America and Russia, including taxi-hailing, private car-hailing, P2P rideshare, bus, bikes & e-bikes, designated driving, automobile solutions, delivery and logistics, and financial services. Tens of millions of car owners, drivers and delivery partners who find flexible work and income opportunities on DiDi platform provide over 10 billion passenger trips a year. Daily trip volume for DiDi's core mobility services exceeds 60 million during the first week of October 2020."

https://www.businesswire.com/news/home/20210128006172/en/2020-Our-Extraordinary-Year-in-DiDi-Numbers

Better Vehicular Transportation

- A safer, greener, and more cost-effective and predictable transportation infrastructure
  - Reduce accidents, adverse health effects, emissions, time spent in traffic jams, etc.

- Congestion cost US commuters $300+ billion in 2017

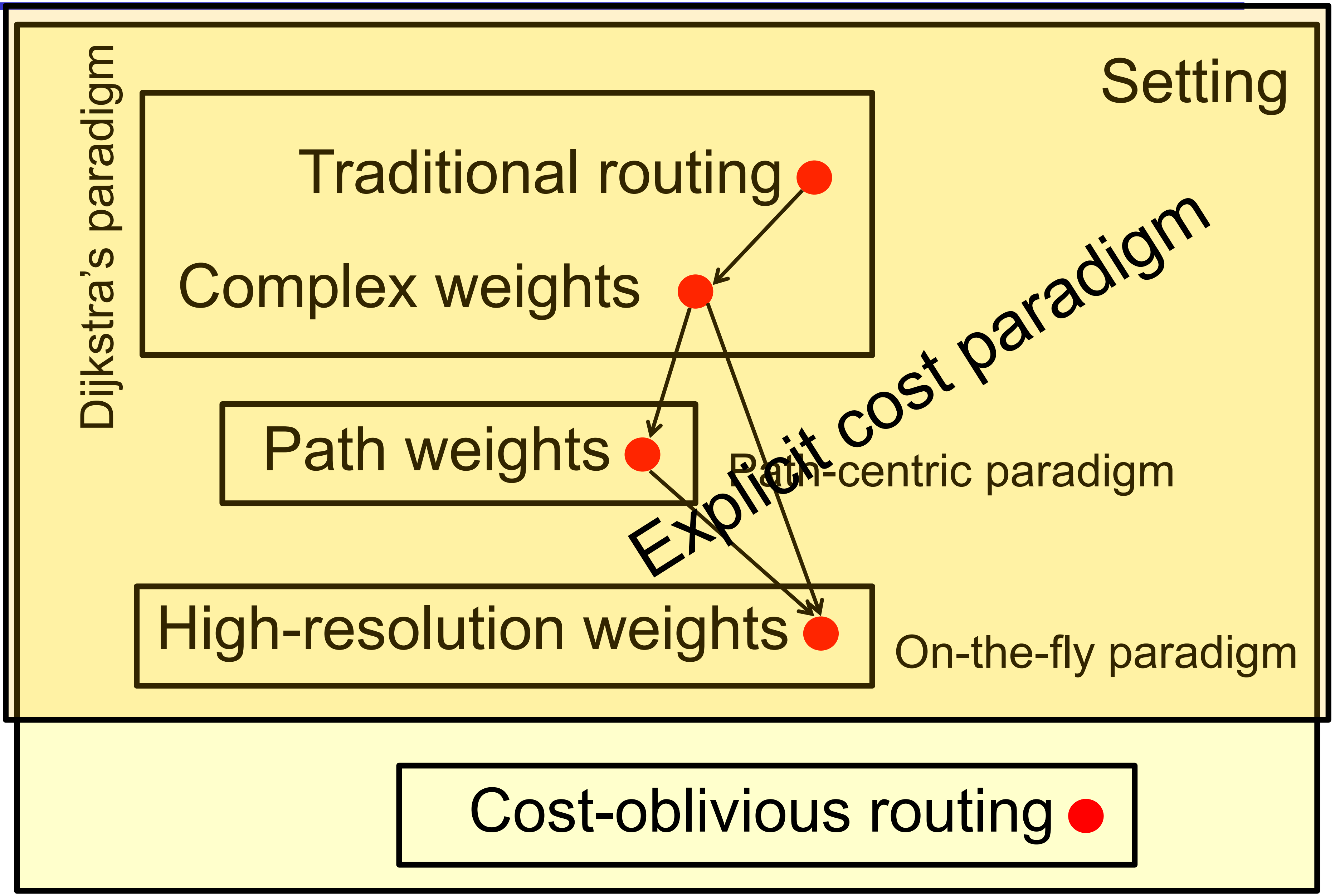
- The transportation sector is the second largest GHG emitting sector and causes substantial pollution.
  - By far, most of it comes from road transportation
- The reduction of GHG emissions from vehicular transportation is essential to combat climate change.

Process: vehicular transportation
Key data: vehicle trajectory data
Services: primarily routing

How Can We Exploit Big Trajectory Data?

- How can we obtain finer-granularity models?
  - Of transportation
  - Of individual transportation behavior
- How can these models fuel finer-granularity services?
  - Personalized services
  - More accurate services in general
- Can big trajectory data yield improved computational performance?
  - Replacing complex computations with look-ups
- Contending with, or benefitting from, skew?
  - The data is concentrated where and when it is needed the most.
- Contending with sparse data
  - Ensuring that coarse models are special cases of finer-granularity models

Routing with Massive Data



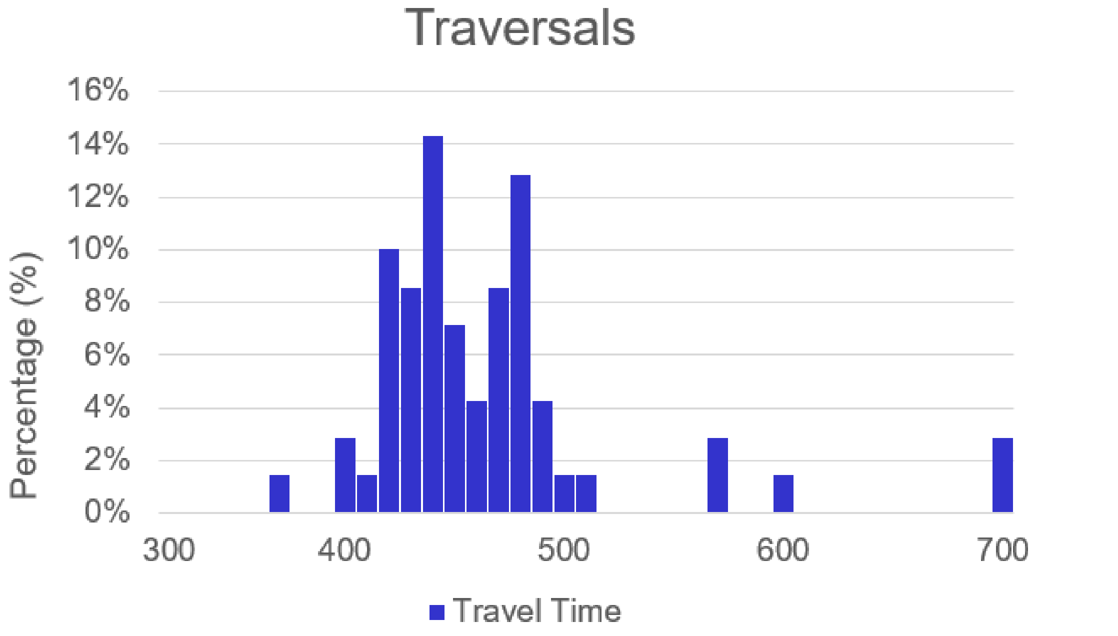
Routing with simple and complex weights

Routing with Simple Weights

- Model a road network as a graph
  - Map to (simple) directed graphs
  - Contend with turn restrictions, lanes
- Assign simple (real or integer) weights, or costs, to edges
  - Prototypical cost: distance
- Given an (s,d) pair, compute the lowest-cost path
  - Prototypical algorithm: Dijkstra’s algorithm
  - Hub labeling, contraction hierarchies, …

Complex Weights

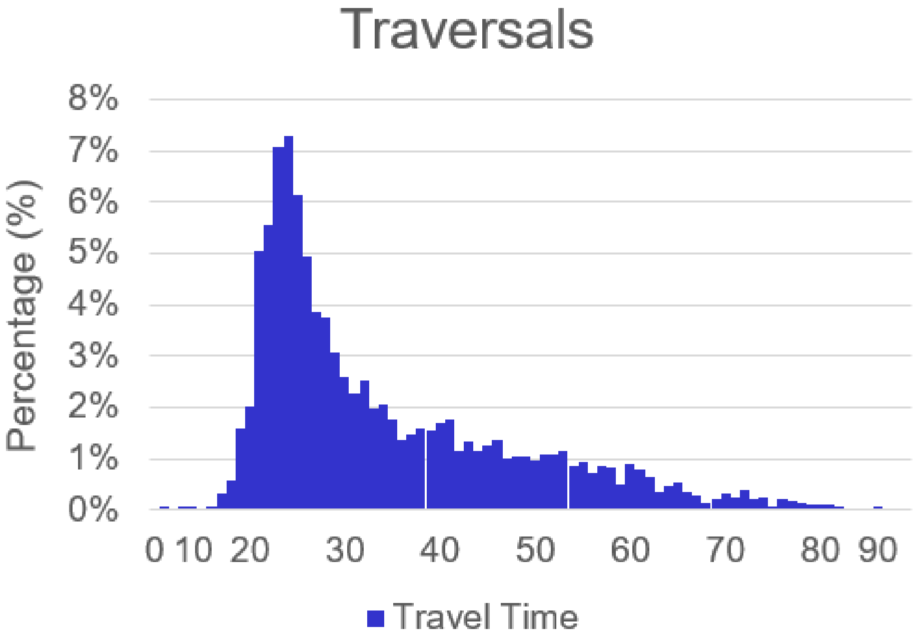
- Prototypical cost: travel time
  - Also GHG emissions, fuel consumption
- Time-dependency
  - Travel time is time varying, e.g., due to time-varying traffic.
- Uncertainty
  - At a single time, a single, deterministic time fails to accurately capture travel time; a distribution is necessary.

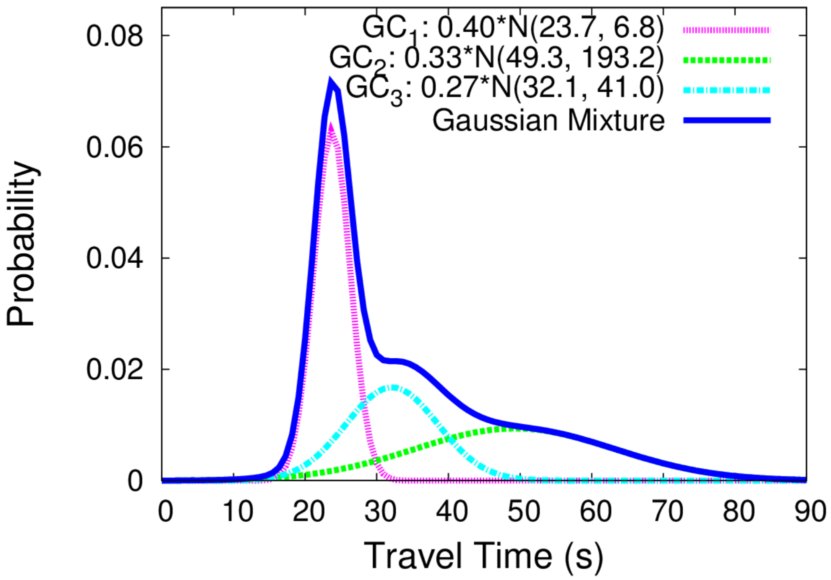


Distribution Construction: GMMs

- Each edge has a set of traversal records
- Partition time into intervals
  - E.g., 24 or 96 daily intervals; max: 7 x 96 = 672 weekly intervals
- For each cost, edge, and interval
  - Gaussian Mixture Model (GMM) based on the relevant records
  - A GMM is a weighted sum of *K* Gaussians.

$$\text{GMM}(x) = \sum_{k=1}^{K} m_k \cdot N(x|\mu_k, \delta_k^{\ 2})$$





Data Needs

- DK road network
  - 1.6 million edges
  - 672 intervals per edge (24 x 4 x 7)
  - 2.4 million cars

- Approx. 1 billion distributions

Route Cost

- Route $R_i = <r_1, r_2, \ldots, r_X>$
- $RC(R_i, t) = <RV_{TT}, RV_{GE}>$

- RV_{TT} is the *convolution* of the travel time distributions of the edges in R_i.
  - The travel time RV of edge r_1 depends on the trip start time t.
  - The travel time RV of edge r_k depends on the travel times of the previous k-1 edges, which may be uncertain.
- RV_{GE} is the *convolution* of the GHG emissions distributions of the edges in R_i.

- Routing: return the skyline routes
  - Use stochastic dominance among cost distributions

Path weights

Motivation for Path Weights

- With edge weights, we atomize trajectories.
- Good
  - Lots of information for each edge
- Bad
  - Destruction of information when trajectories are broken into unconnected atoms (turn times, traffic lights, specific driver behavior, etc.)

- Convolution assumes independence.
  - Used to compose distributions for paths
- Independence may not hold.
  - If one edge is congested, an adjacent edge may often also be congested.
  - An aggressive driver is fast on all edges.

- Solution: assign weights to paths (with enough data).

Dependency Example

Edge-Centric Paradigm

AB

Time	Prob
20	0.5
25	0.5

BC

Time	Prob
10	0.5
15	0.5

Path-Centric Paradigm

P = (AB, BC)

Time	Prob
20, 10	0.5
25, 15	0.5

Path Cost Computation in Edge-Centric Paradigm

Convolution

Time	Prob
20, 10	0.25
20, 15	0.25
25, 10	0.25
25, 15	0.25

↓

Time	Prob
30	0.25
35	0.50
40	0.25

Cost	Prob
30	0.5
40	0.5

Path Cost Computation in Path-Centric Paradigm

Learning Instead of Convolution

- Convolution assumes independence that often does not hold, yielding inaccurate results.
- The result has useful properties in terms of the input.
- Does not use data about the context of the convolution.

- Learning may exploit the context to improve accuracy.
- The result no longer has simple properties in terms of the input.

Data Needs

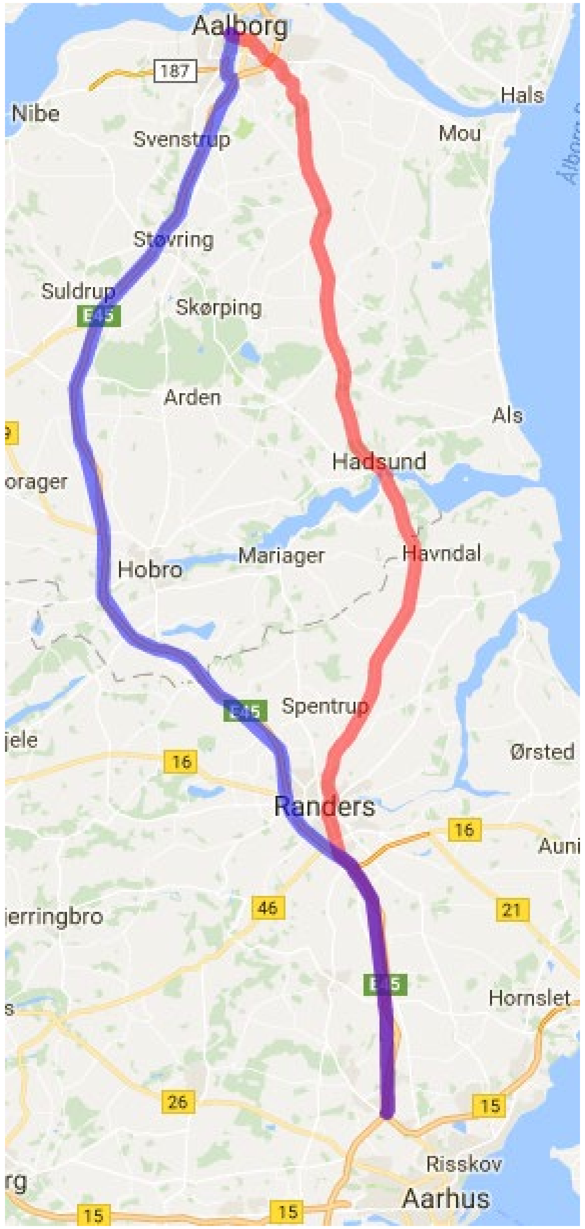
- DK road network
  - 1.6 million edges
  - 672 intervals per edge (24 x 4 x 7)
  - 2.4 million cars

- Edge weights: approx. 1 billion distributions
- Path weights
  - Assumption: one path from each edge to each edge on average (conservative)
  - Approx. 2.5 trillion paths
  - Approx. 1.7 quadrillion weights

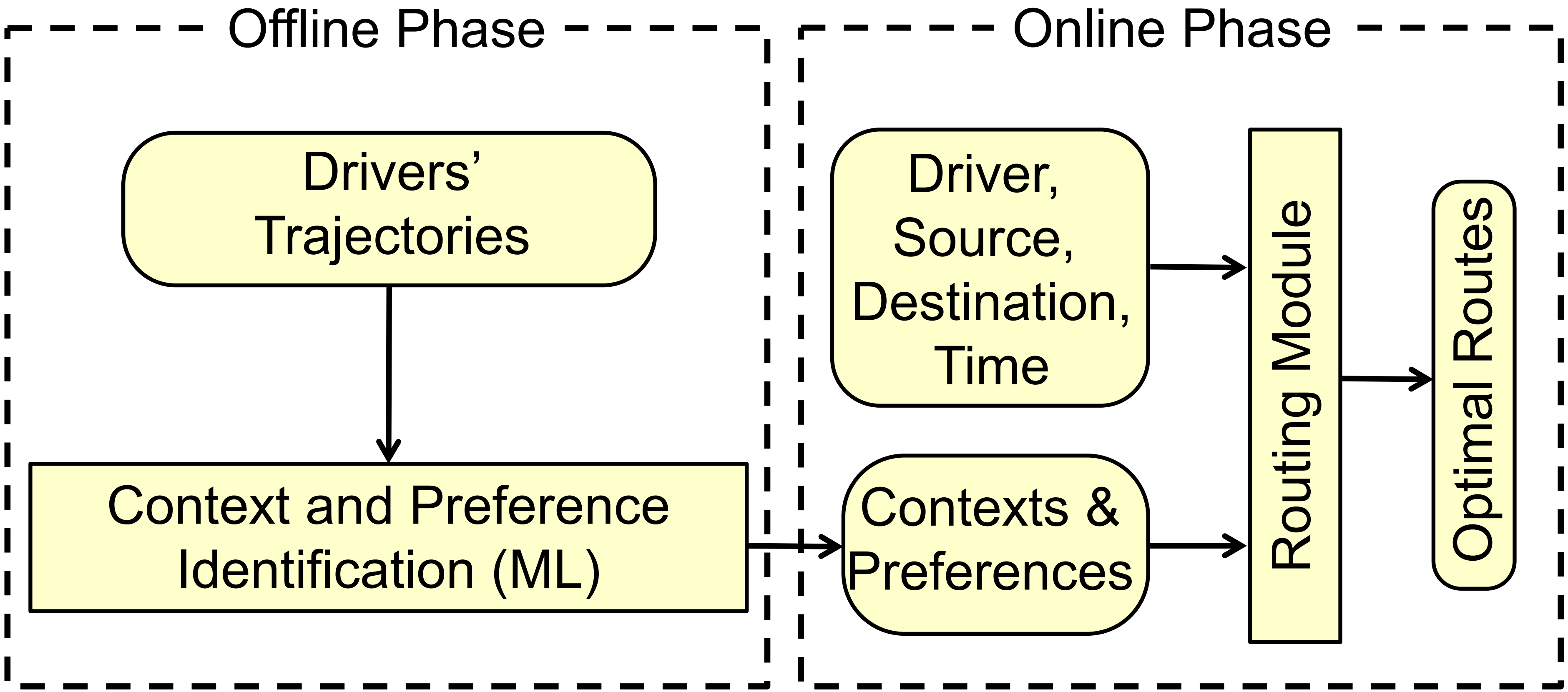
High-resolution weights

Personalized Routing

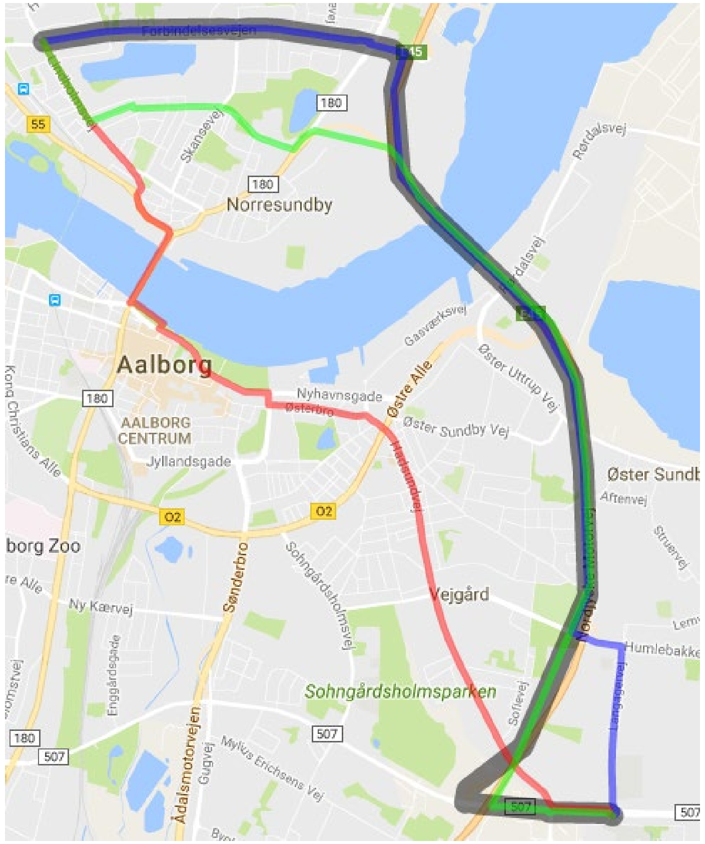
- Different drivers may take different routes due to different preferences.
- The same driver may take different routes in different contexts.
  - Morning: try to save time to avoid being late.
  - Weekend afternoon: try to save fuel.
- Supported by data
- Challenges
  - Identify contexts for drivers and identify driving preference in each context.
  - Contend with time-dependent uncertain travel costs while considering individual drivers' driving behaviors.

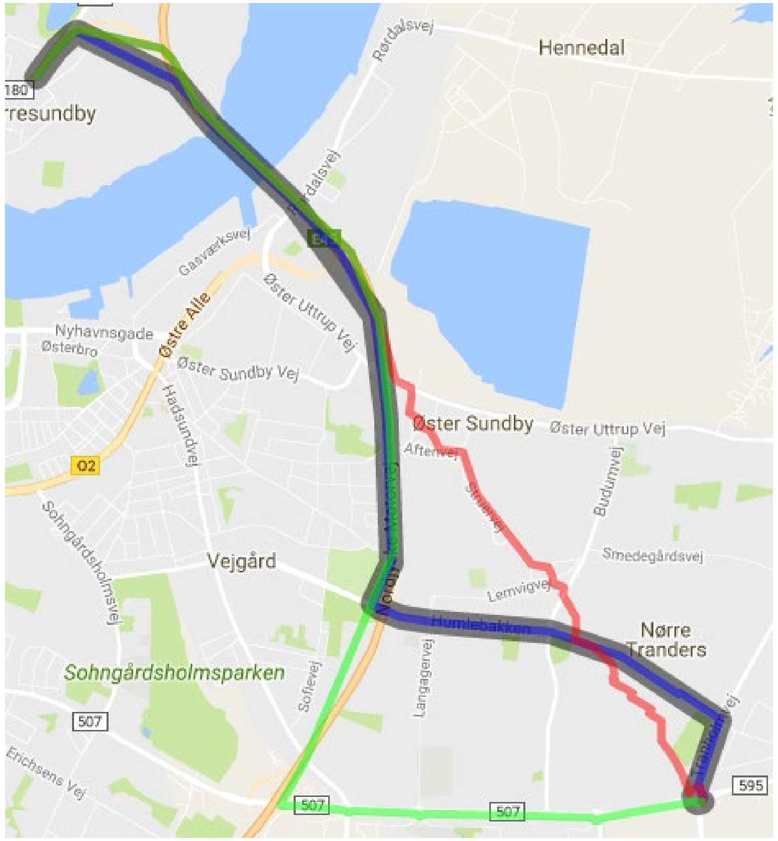


Framework


Offline Phase
Online Phase
Drivers' Trajectories
Context and Preference Identification (ML)
Driver, Source, Destination, Time
Contexts & Preferences
Routing Module
Optimal Routes

Example Results





- Dark, bold routes: actual routes used by drivers.
- Red routes: shortest routes.
- Green routes: fastest routes.
- Blue routes: predicted routes using the identified contexts and driving preferences.

Data Needs

- DK road network
  - 1.6 million edges
  - 672 intervals per edge (24 x 4 x 7)
  - 2.4 million cars
- Approx. 1 billion distributions for one-size-fits-all routing with edge weights

- Each car (driver) has 2 contexts.
- Result
  - Edge weights: 5 million x 1 billion = 5 quadrillion distributions (2.4 million x 2 x 1 billion)
  - Path weights: 5 million x 1.7 quadrillion = approx. 8 sextillion distributions (10^21)
- For each routing query, compute weights using a different set of trajectories.
- Weights cannot be precomputed.
- A new on-the-fly paradigm is needed.

Cost-oblivious routing

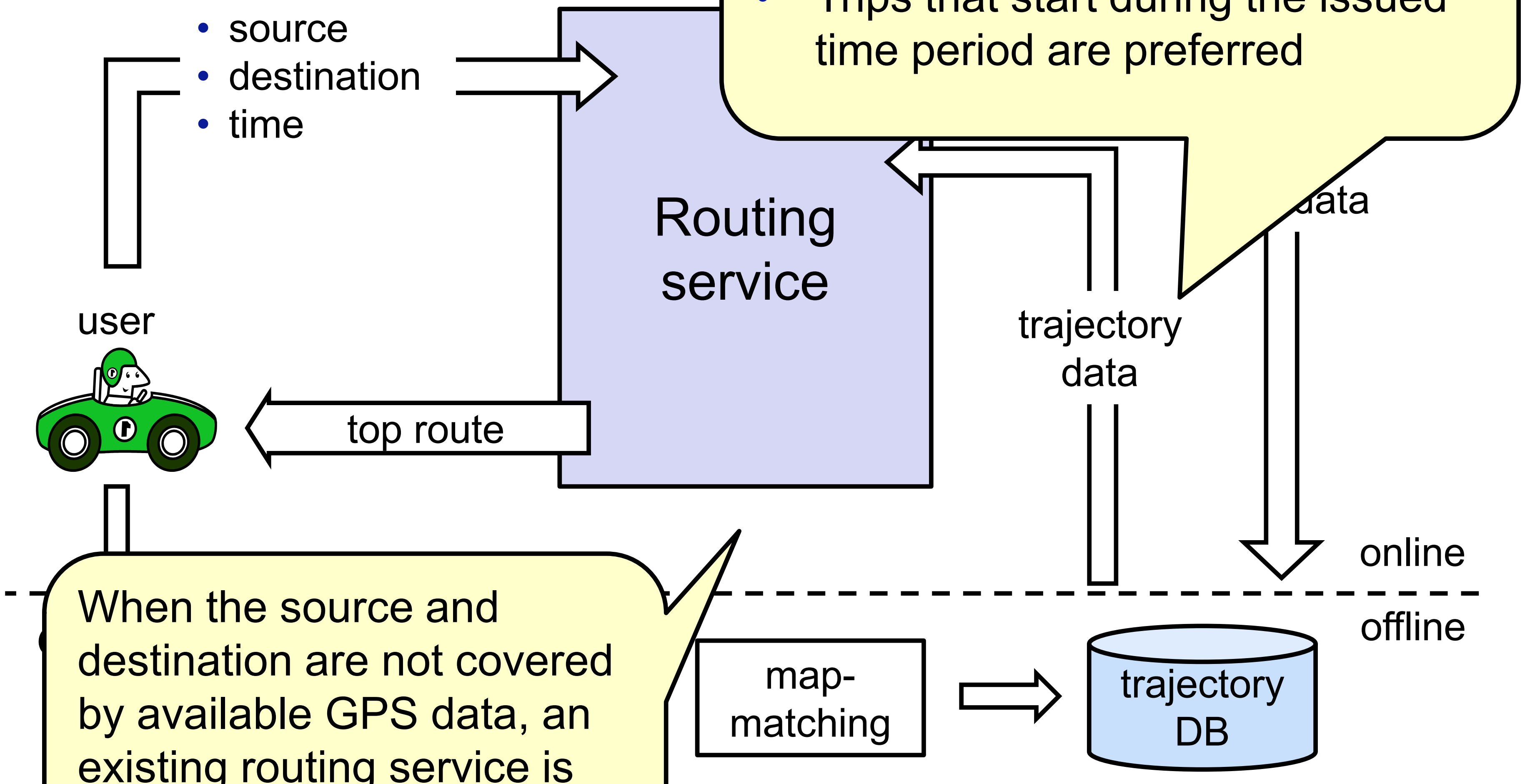
Overview

- Idea: Use the knowledge of local drivers for routing.

- Utilize trajectories that capture local driver behavior.
  - Local drivers often follow paths that are neither shortest nor fastest.
  - Exploits possibly hard-to-formalize insight into local conditions
- Recommend routes based on this behavior.

- Study 1 [Ceikute and Jensen MDM 2013, 2015]
  - Use only “directly related” trajectories.
- Study 2 [Guo et al. ICDE 2018]
  - “Extend” study 1 to contend with sparse data.

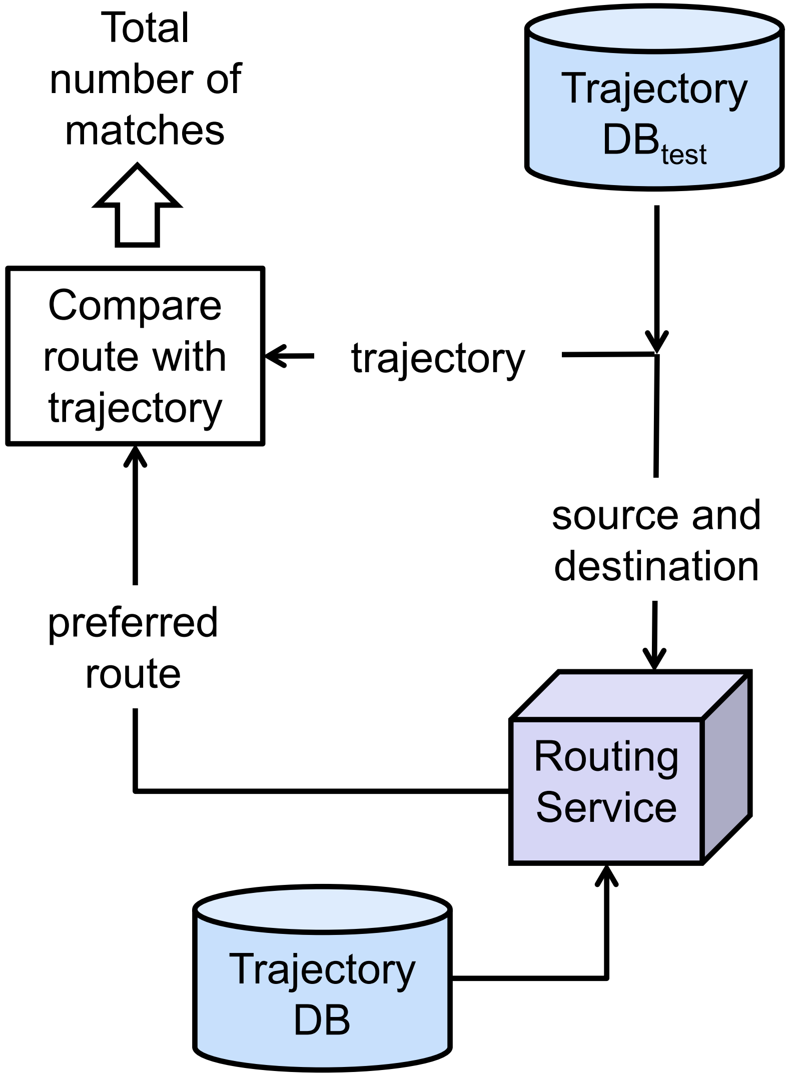
Study 1: Framework



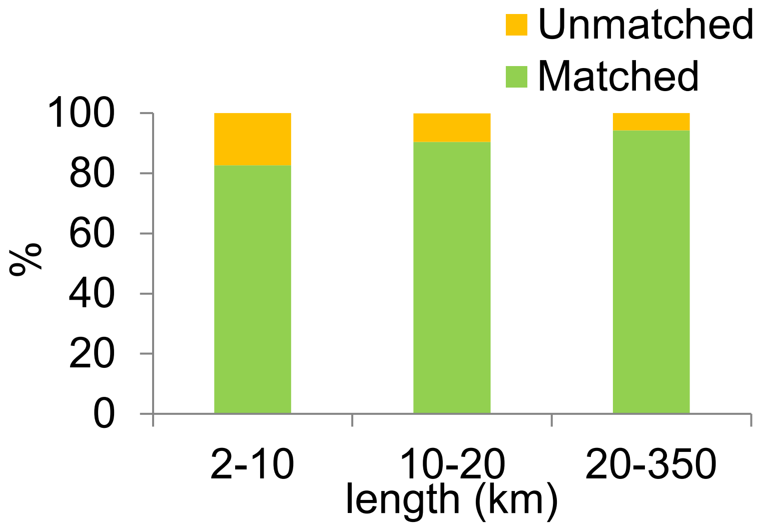
Trips used:
- Start or end at, or go through, source and destination locations
- Trips that start during the issued time period are preferred

When the source and destination are not covered by available GPS data, an existing routing service is used.

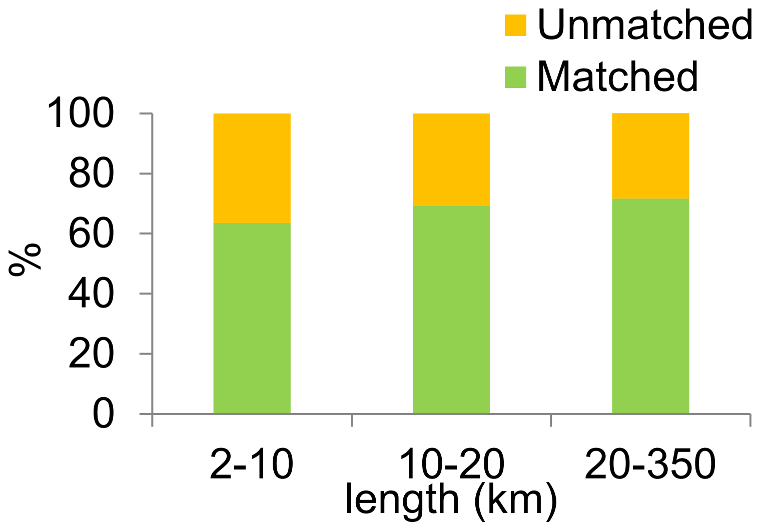
Routing Quality: Match



Routing Quality Evaluation



Our Proposal



Google Directions API (top-1)

skip

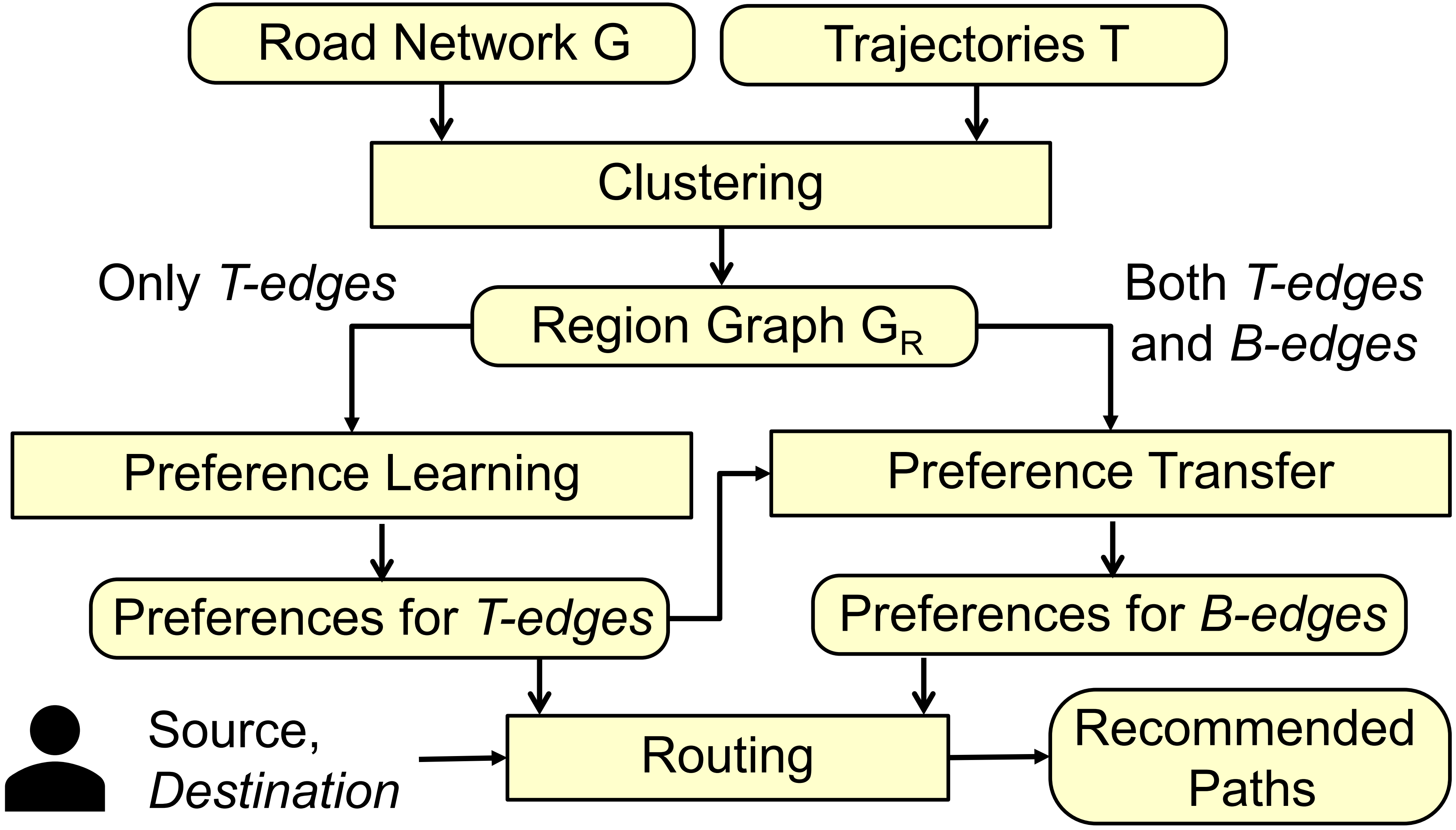
Study 2: Learning to Route

- Even given massive data, the data is skewed and will not cover all (s,d) pairs at all times.
- How can we better utilize the data we have?

- Re-use trajectories for near-by (s,d)’s.
- Re-use trajectories for similar (s,d)’s

skip

L2R Framework



Road Network G

Trajectories T

Clustering

Only *T-edges*

Region Graph G_R

Both *T-edges* and *B-edges*

Preference Learning

Preference Transfer

Preferences for *T-edges*

Preferences for *B-edges*

Source, *Destination*

Routing

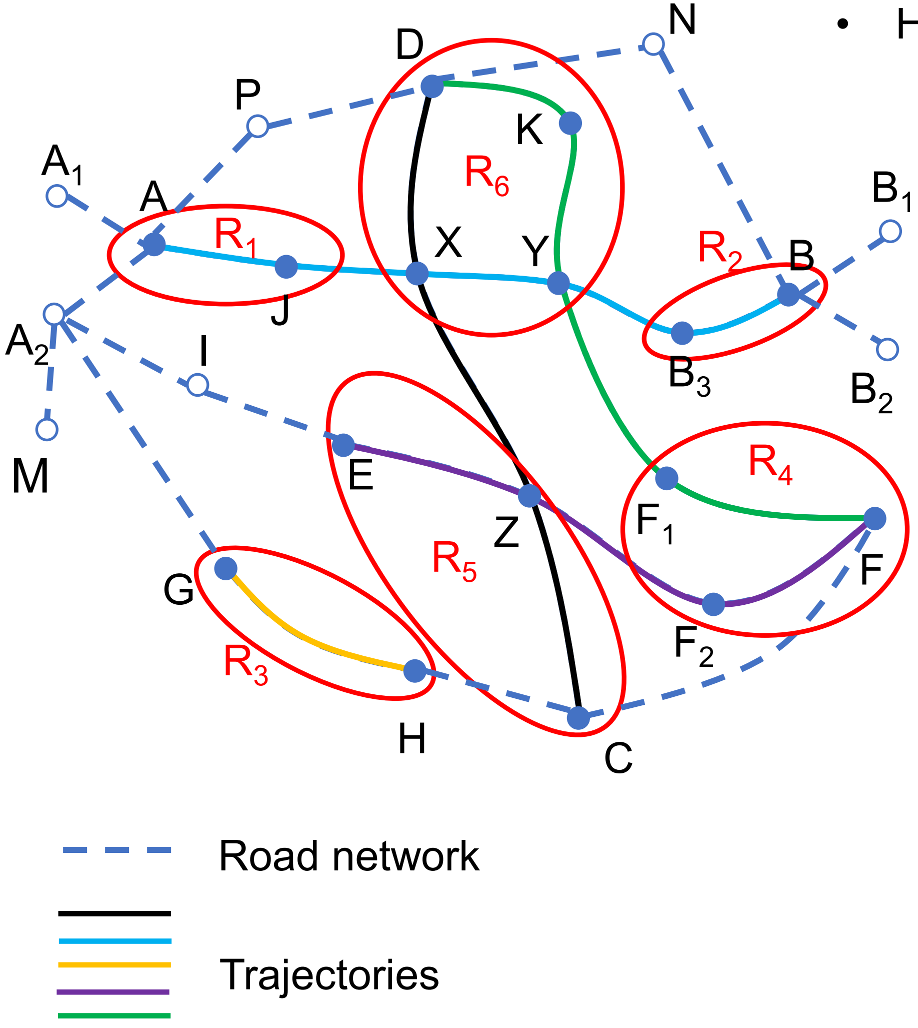
Recommended Paths

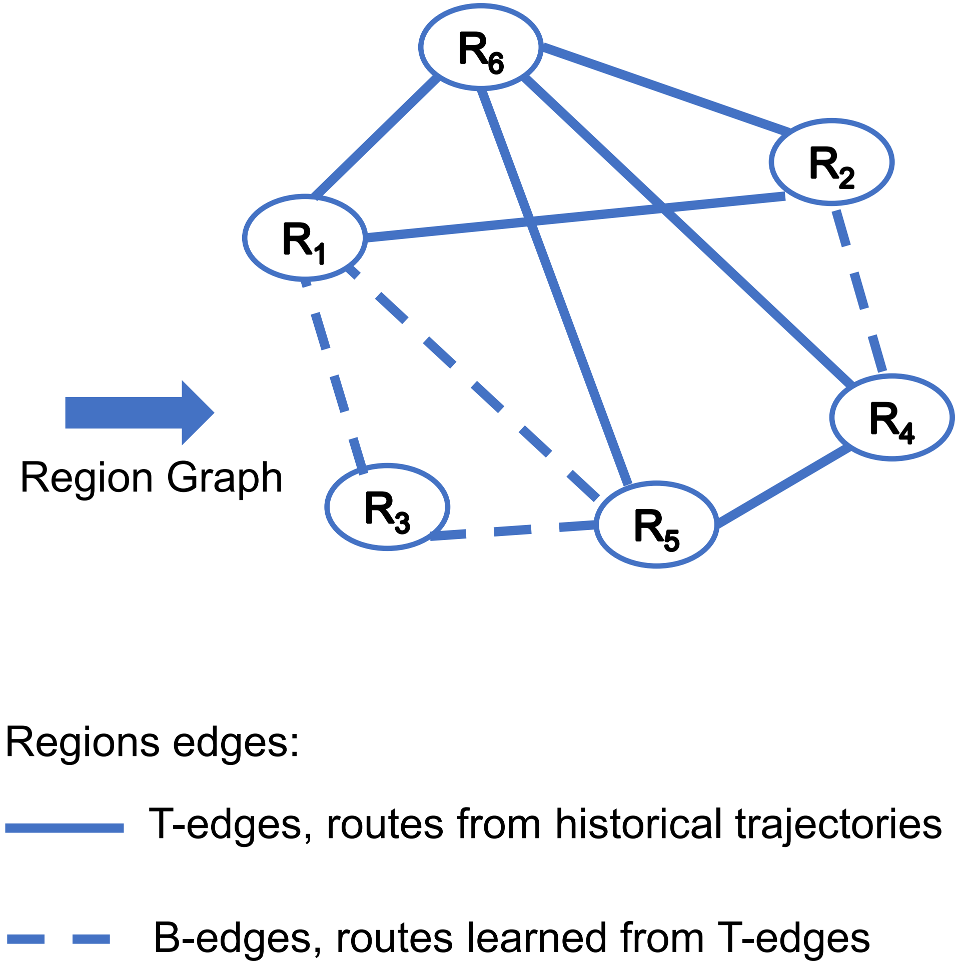
skip

Region Graph

How to utilize the trajectories for routing?

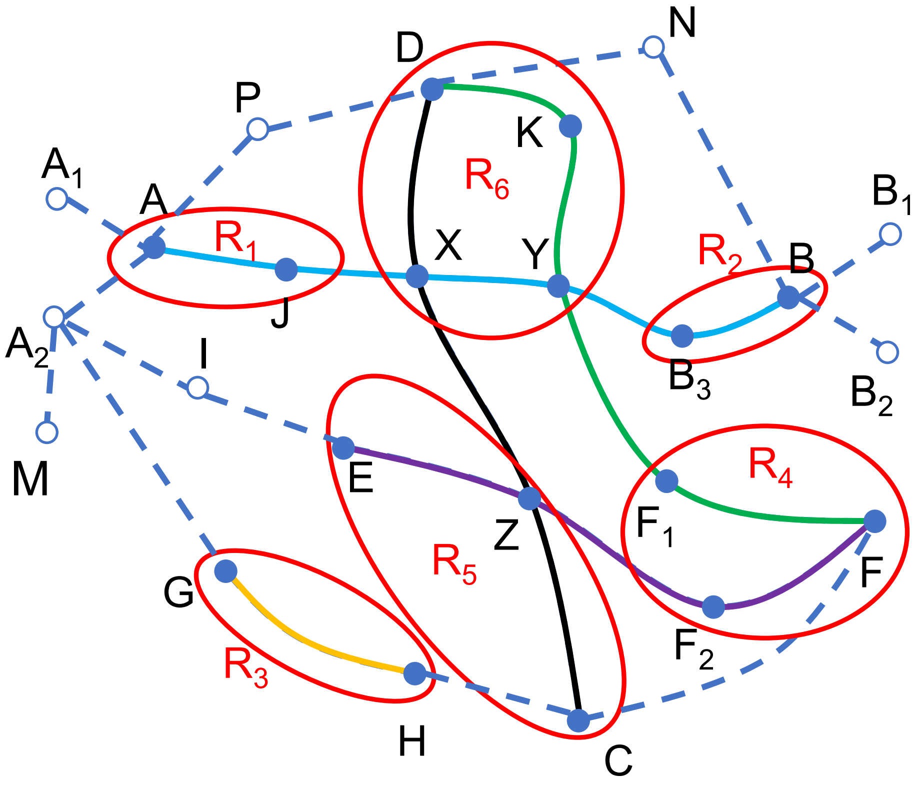
- $A_1 – B_1$
- H – F

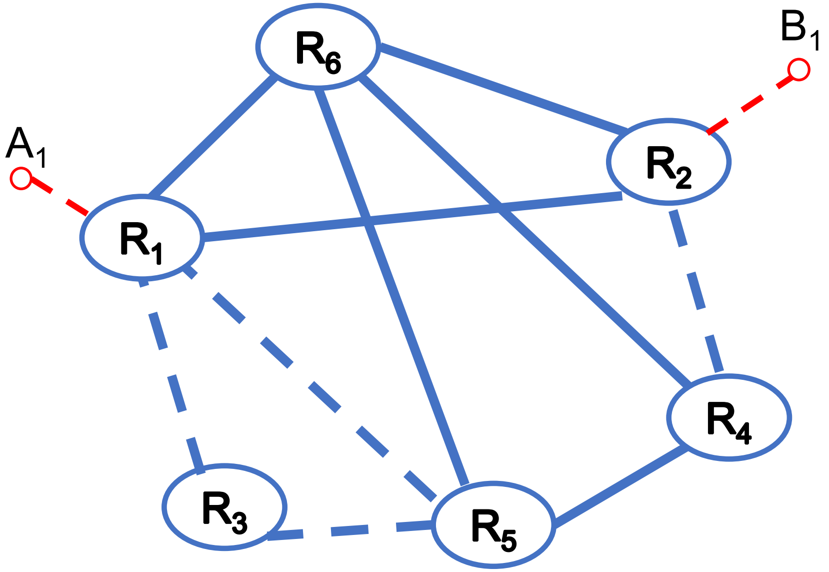




Region Graph
R_6
R_2
R_1
R_4
R_3
R_5
Regions edges:
T-edges, routes from historical trajectories
B-edges, routes learned from T-edges

Routing from A_1 to B_1

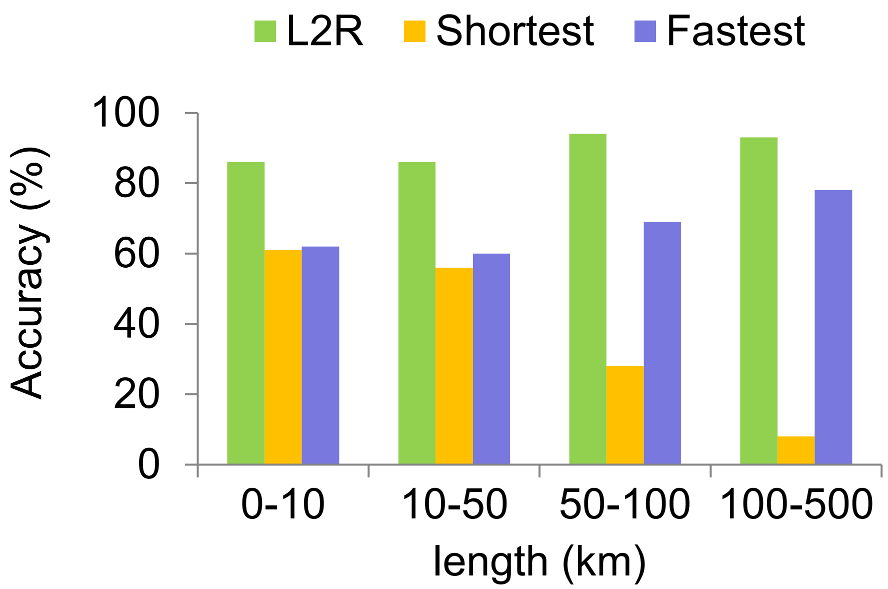




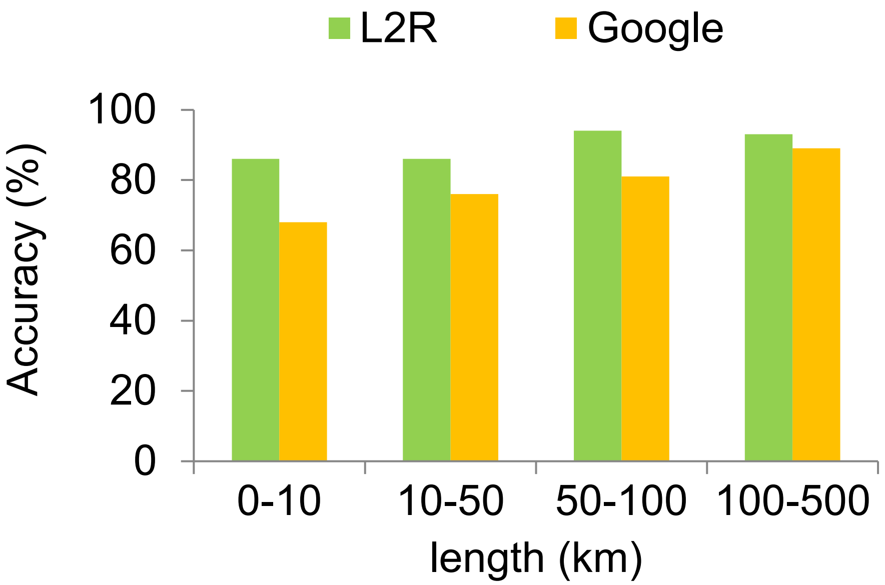
Routing $A_1 - B_1$:

1. A_1 is close to R_1: $A_1 - R_1$
2. B_1 is close to R_2: $R_2 - B_1$
3. Region graph routing: $R_1 - R_2$
4. Region route $R_1 - R_2$ is recovered as $A - J - X - Y - B_3 - B$
5. Return $A_1 - A - J - X - Y - B_3 - B - B_1$

Comparison with Cost-Based Routing



By Distance



By Distance

skip

Closing

Quantification of Vehicular Emissions

- "If you can't measure it, you can't improve it."

- Road administrators would like to quantify the emissions from their roads
  - Based on number of vehicles, vehicle types, distances driven, speeds
- Road administrators would like to be able to study the effects of changes
  - Construction projects, road pricing-type schemes, future distributions of vehicle types, incentives for bicycling, etc.

- Trajectory data is an important part of the picture.

- We are working with Danish company Rambøll to enable quantification of vehicular emissions.

skip

Summary

- Illustrated trends in trajectory-based data analytics, using vehicle routing as an example.
  - Data-driven focus
  - Integration of ML
- Trajectory data fuels many other mobility services, and I was unable to cover many exciting advances...
  - Analytics systems based on Spark: e.g., UITraMan (offline), Dragoon (offline, online)
  - Co-movement, clustering
  - Data cleaning, e.g., IHCS
  - Graph Convolutional Networks for Road Networks

Readings

- C. Guo, C. S. Jensen, B. Yang: Towards Total Traffic Awareness. SIGMOD Record 43(3):18-23 (2014)
- J. Hu, B. Yang, C. S. Jensen, Y. Ma: Enabling time-dependent uncertain eco-weights for road networks. GeoInformatica 21(1): 57-88 (2017)
- B. Yang, C. Guo, C. S. Jensen, M. Kaul, S. Shang: Stochastic skyline route planning under time-varying uncertainty. ICDE 2014:136-147
- J. Dai, B. Yang, C. Guo, C. S. Jensen, J. Hu: Path Cost Distribution Estimation Using Trajectory Data. PVLDB 10(3): 85-96 (2016)
- B. Yang, J. Dai, C. Guo, C. S. Jensen, J. Hu: PACE: a PAth-CEntric paradigm for stochastic path finding, VLDBJ 27(2): 153-178 (2018)
- B. Yang, C. Guo, Y. Ma, C. S. Jensen: Toward personalized, context-aware routing. VLDB J. 24(2):297-318 (2015)
- B. Yang, C. Guo, C. S. Jensen: Travel Cost Inference from Sparse, Spatio-Temporally Correlated Time Series Using Markov Models. PVLDB 6(9): 769-780 (2013)
- B. Yang, M. Kaul, C. S. Jensen: Using Incomplete Information for Complete Weight Annotation of Road Networks. IEEE TKDE 26(5):1267-1279 (2014)
- V. Ceikute, C. S. Jensen: Routing Service Quality - Local Driver Behavior Versus Routing Services. MDM 2013: 97-106
- V. Ceikute, C. S. Jensen: Vehicle Routing with User-Generated Trajectory Data. MDM 2015: 14-23
- C. Guo, B. Yang, O. Andersen, C. S. Jensen, K. Torp: EcoMark 2.0: empowering eco-routing with vehicular environmental models and actual vehicle fuel consumption data. GeoInformatica 19(3):567-599 (2015)
- C. Guo, B. Yang, J. Hu, C. S. Jensen: Learning to Route with Sparse Trajectory Sets. ICDE 2018: 1073-1084
- C. Guo, B. Yang, J. Hu, C. S. Jensen, L. Chen: Context-aware, preference-based vehicle routing. VLDBJ 29(5): 1149-1170 (2020)
- T. Li, R. Huang, L. Chen, C. S. Jensen, T, B. Pedersen: Compression of Uncertain Trajectories in Road Networks. PVLDB 13(7): 1050-1063 (2020)
- T. Li, L. Chen, C. S. Jensen, T. B. Pedersen: TRACE: Real-time Compression of Streaming Trajectories in Road Networks. PVLDB 14(7): 1175-1187 (2021)
- S. Aa. Pedersen, B. Yang, C. S. Jensen: A Hybrid Learning Approach to Stochastic Routing. ICDE 2020: 1910-1913
- S. Aa. Pedersen, B. Yang, C. S. Jensen: Fast stochastic routing under time-varying uncertainty. VLDBJ 29(4): 819-839 (2020)
- S. Aa. Pedersen, B. Yang, C. S. Jensen: Anytime Stochastic Routing with Hybrid Learning. PVLDB 13(9): 1555-1567 *(2020)*
- T. S. Jepsen, C. S. Jensen, T. D. Nielsen: UniTE - The Best of Both Worlds: Unifying Function-Fitting and Aggregation-Based Approaches to Travel Time and Travel Speed Estimation. CoRR abs/2104.13321 (2021)
- J. Qi, R. Zhang, C. S. Jensen, K. Ramamohanarao, J. He: Continuous Spatial Query Processing: A Survey of Safe Region Based Techniques. ACM CSurv 51(3): 64:1-64:39 (2018)