Welcome to

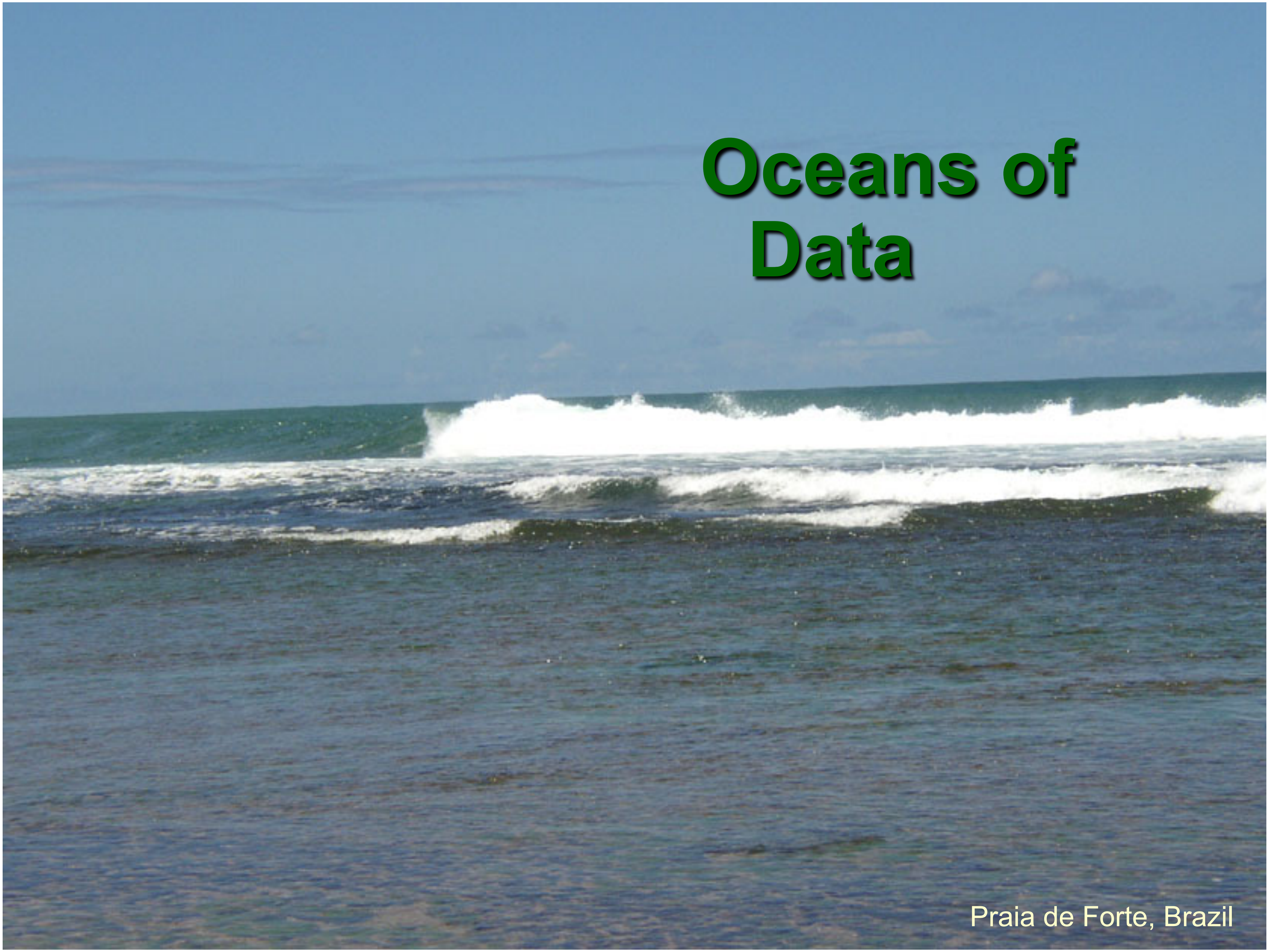
DS504/CS586: Big Data Analytics

Data Pre-processing and Cleaning

Prof. Yanhua Li

Time: 6:00pm – 8:50pm R
Location: AK 232
Fall 2016

Oceans of Data



Praia de Forte, Brazil

Data Quality Dimensions

- Accuracy
  - Errors in data

  Example:"Jhn" vs. "John"

- Currency
  - Lack of updated data

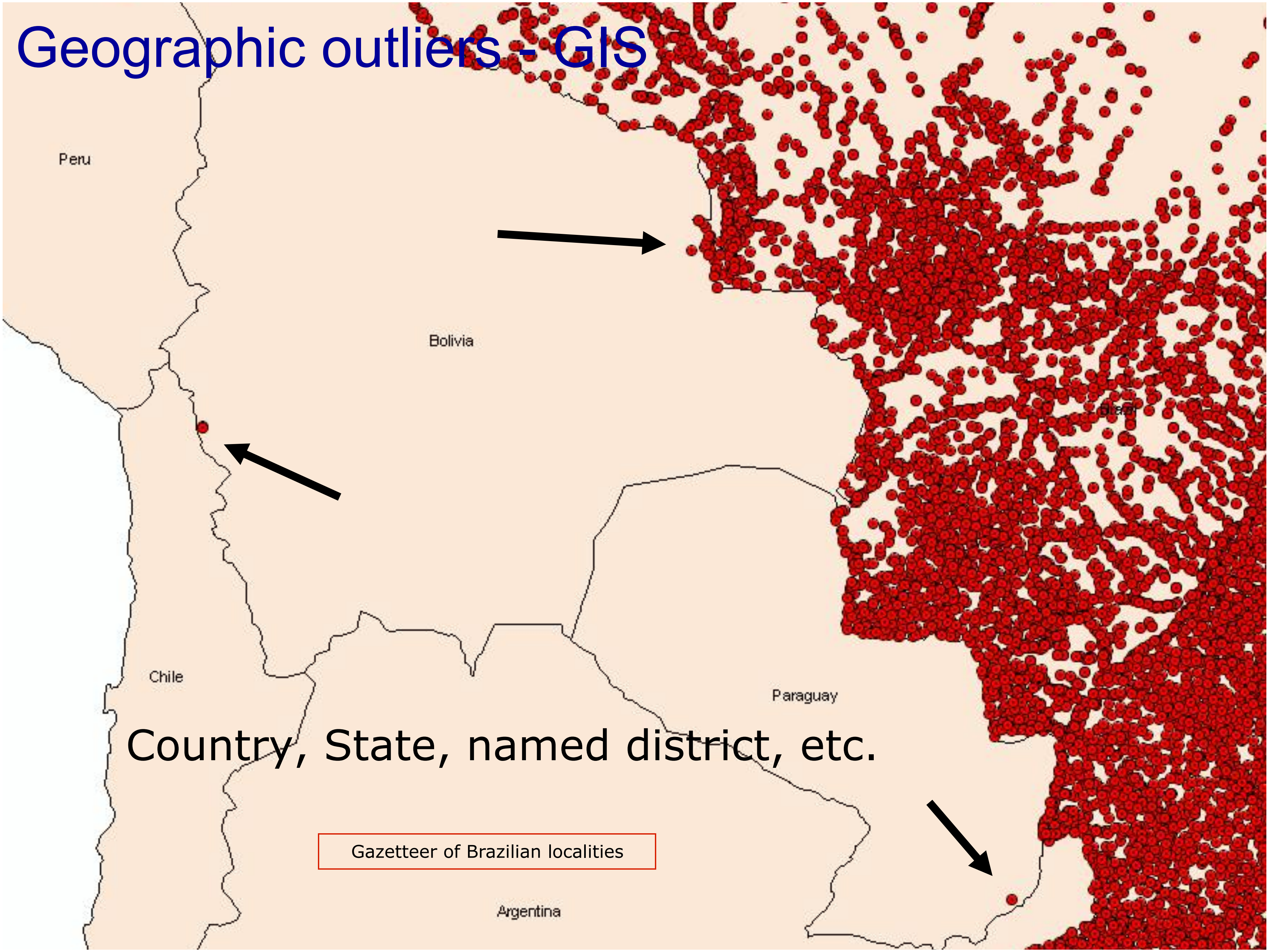
  Example: Residence (Permanent) Address: out-dated vs. up-to-dated

- Consistency
  - Discrepancies into the data

  Example: ZIP Code and City consistent

- Completeness
  - Lack of data
  - Partial knowledge of the records in a table

Geographic outliers - GIS



Country, State, named district, etc.

What do we mean by ‘Data Quality’ ?

An essential or distinguishing characteristic necessary for data to be fit for use.

SDTS 02/92

The general intent of describing the quality of a particular dataset or record is to describe the fitness of that dataset or record for a particular use that one may have in mind for the data.

Chrisman, 1991

Loss of data quality

Loss of data quality can occur at many stages:

- At the time of collection
- During digitisation
- During documentation
- During storage and archiving
- During analysis and manipulation
- At time of presentation
- And through the use to which they are put

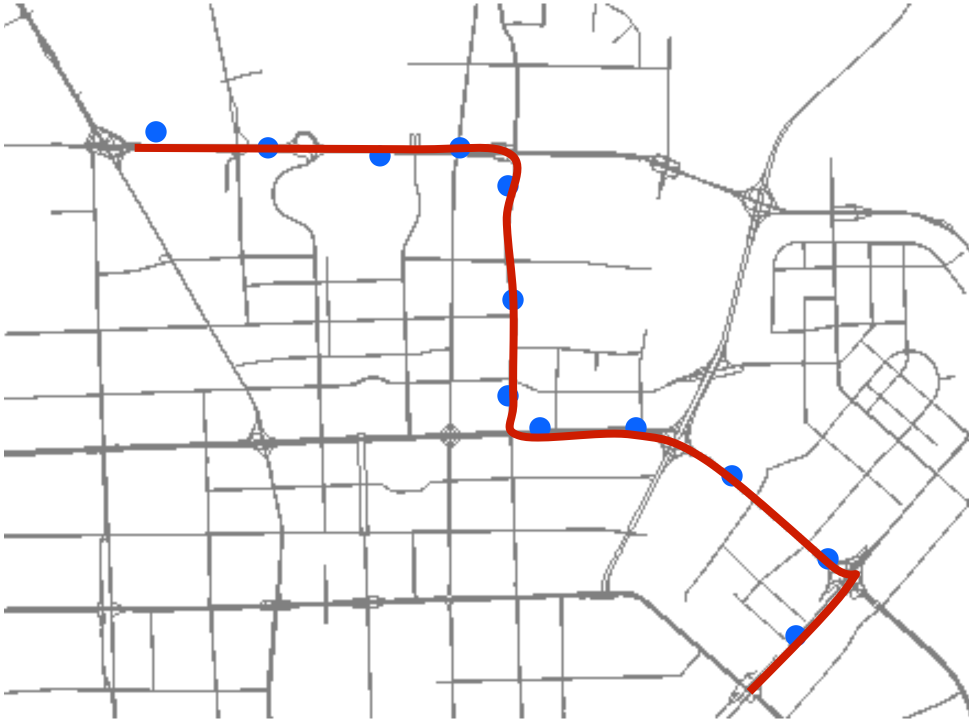
Data Cleaning

- Data cleaning tasks
  - **Accuracy:** Smooth out noisy data
  - **Currency:** Update the records
  - **Consistency:** Correct inconsistent data
  - **Completeness:** Fill in missing values

Map matching

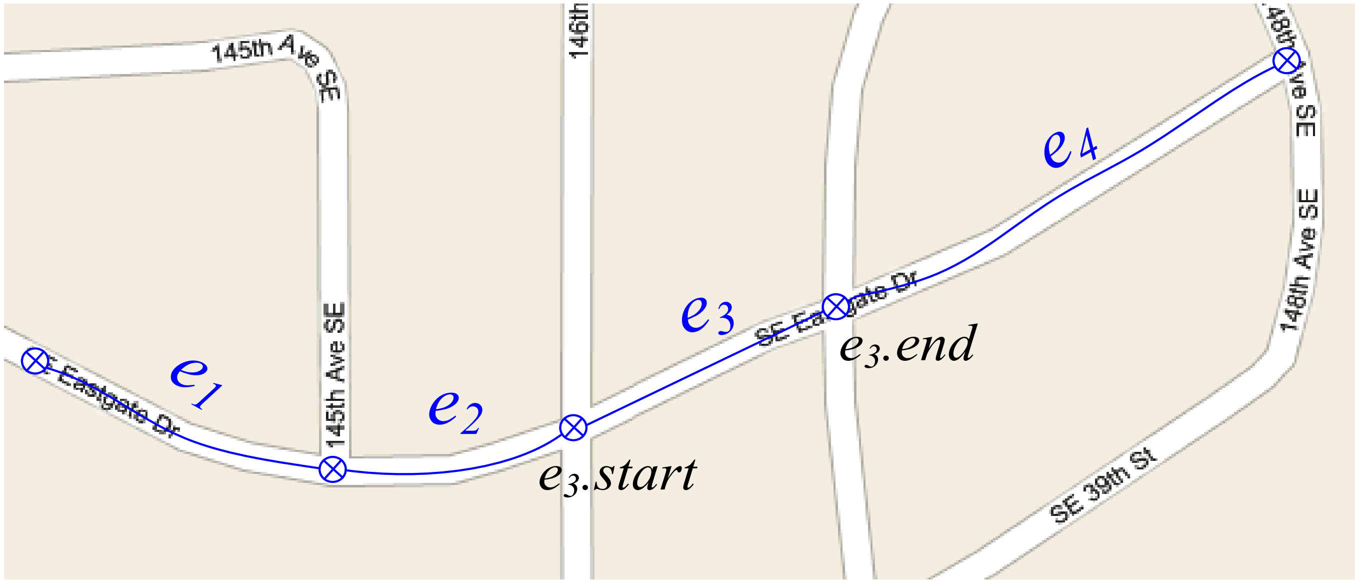
Map-matching

- Problem: (Sampled data)
  - GPS trajectory = a sequence of GPS locations with time stamps
  - Map a GPS trajectory onto a road network
  - a sequence of GPS points → a sequence of road segments



Spatial Data

- Road network: G=(V, E)
  - V is a set of nodes
  - E is a set of road segments
  - $e \in E$, consists of two terminal nodes and a sequence of intermediate points describing the segment with a polyline
  - Properties: $e.len$, $e.dir$, $e.lanes$

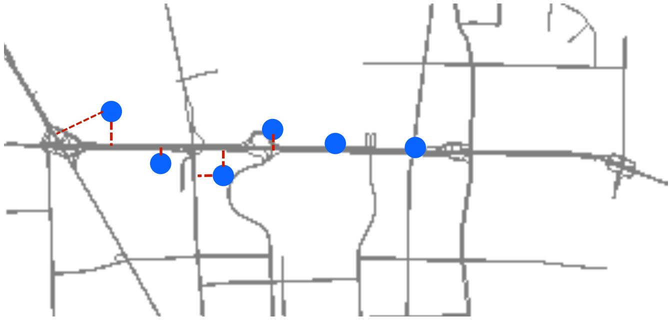


Map-Matching

- Why it is important
  - A fundamental step in many transportation applications
    - Navigation and driving
    - Traffic analysis
    - Taxi dispatching and recommendations
  - Examples:
    - Find the vehicles passing Institute Road
    - Calculate the average travel time from WPI campus to MIT campus
    - When will the Bus 3 arrive at stop Highland St & North Ashland St

Map-Matching

- Simple solution for high-sampling-rate data
  - Weighted distance



7th Ave to Jones Blvd.

1/4 mi.

Remaining:
1:30

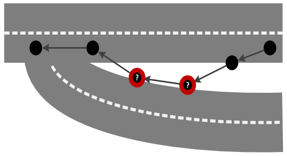
15.5 mi

ETA 1:45 PM

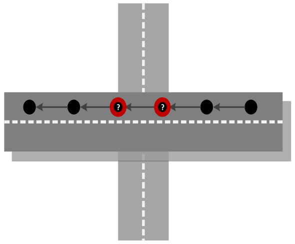
Yin Lou, Chengyang Zhang, **Yu Zheng**, et al. Map-Matching for Low-Sampling-Rate GPS Trajectories. In ACM SIGSPATIAL GIS 2009

Map-Matching (for low sampling rate)

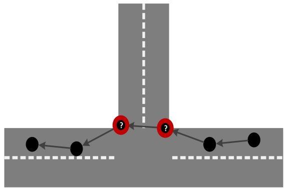
- Why difficult



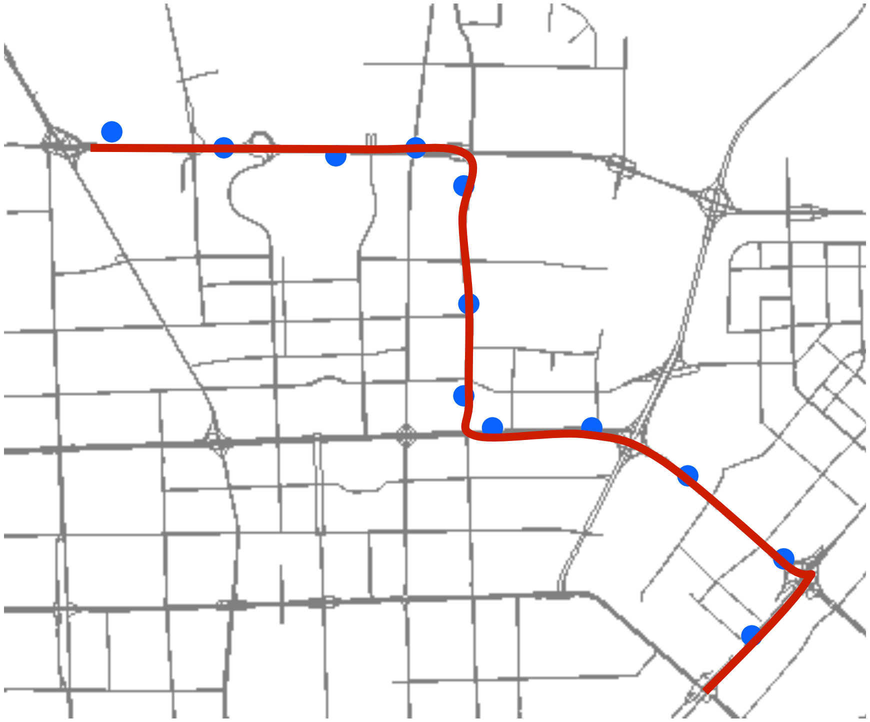
(a) Parallel roads

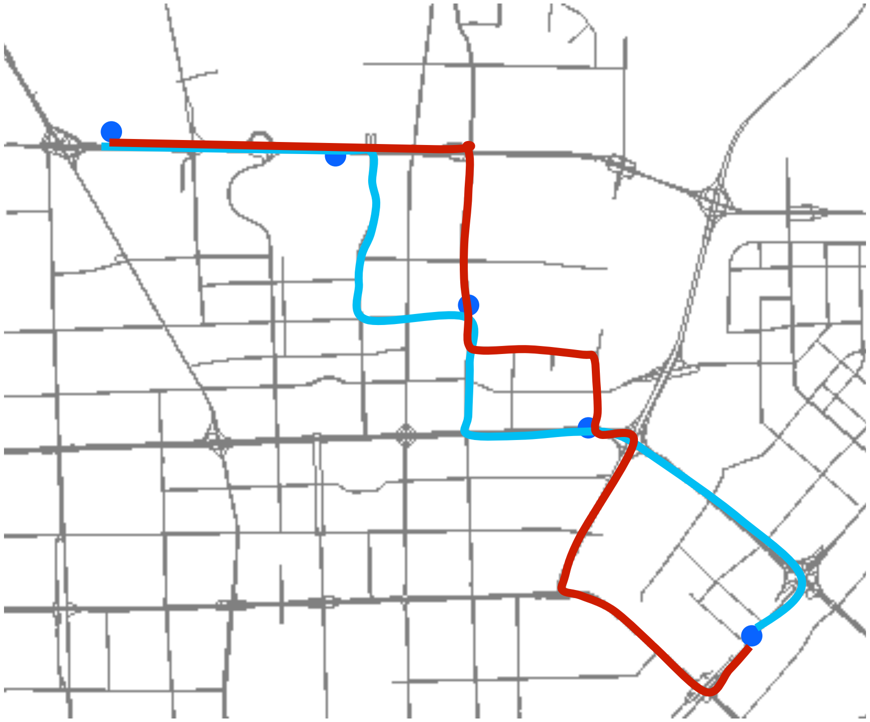


b) Overpass



c) Spur





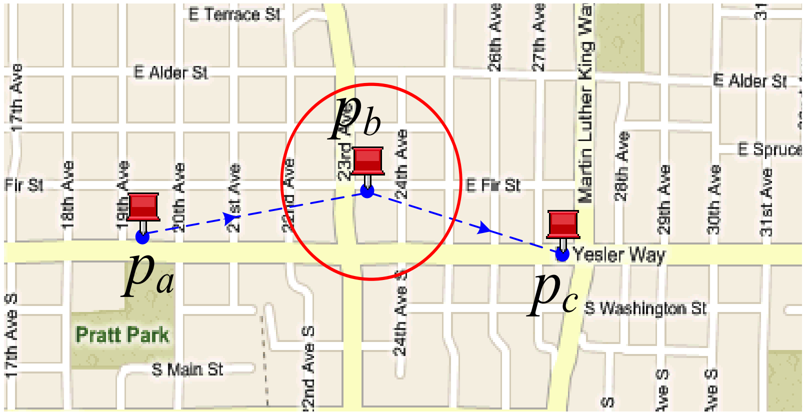
Map-Matching

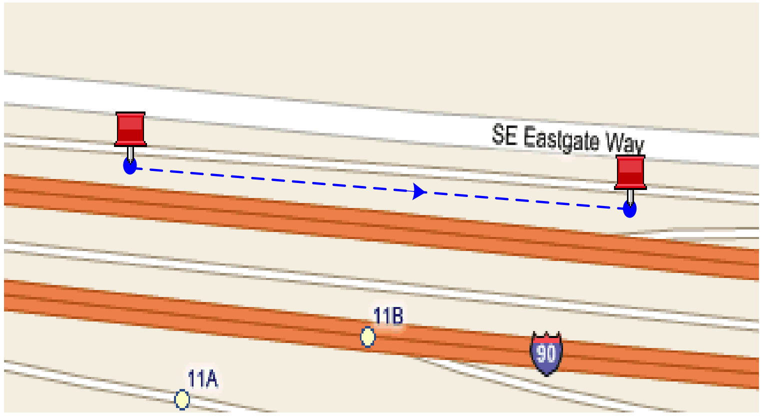
- According to the additional information used
  - Geometric
  - Topological
  - Probabilistic
  - Advanced techniques
- According to the range of sampling points
  - Local/incremental
  - Global

Yu Zheng. Trajectory Data Mining: An Overview. ACM Transaction on Intelligent Systems and Technology, 6, 3, 2015.

Map-matching

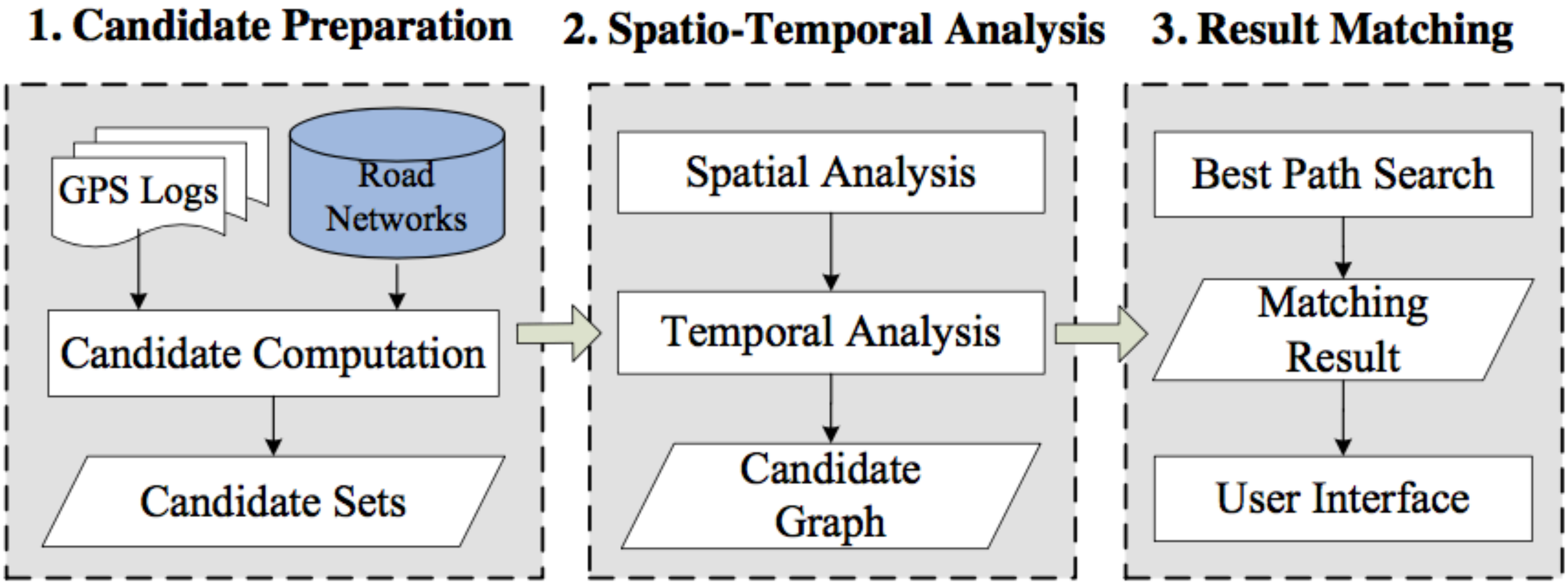
- Insights
  - Consider both local and global information
  - Incorporating both spatial and temporal features





Yin Lou, Chengyang Zhang, **Yu Zheng**, et al. Map-Matching for Low-Sampling-Rate GPS Trajectories. In ACM SIGSPATIAL GIS 2009

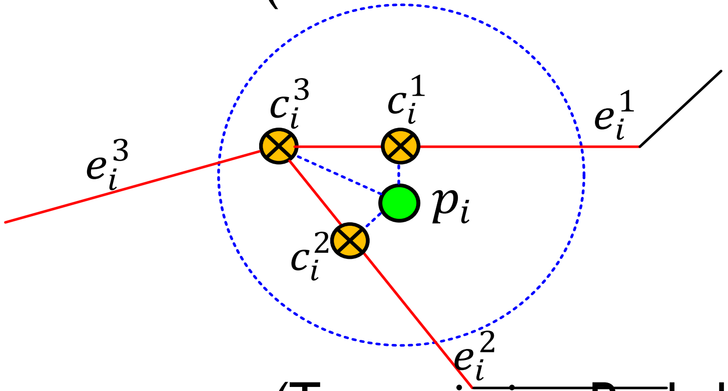
Map-matching framework



Yin Lou, Chengyang Zhang, **Yu Zheng**, et al. Map-Matching for Low-Sampling-Rate GPS Trajectories. In ACM SIGSPATIAL GIS 2009

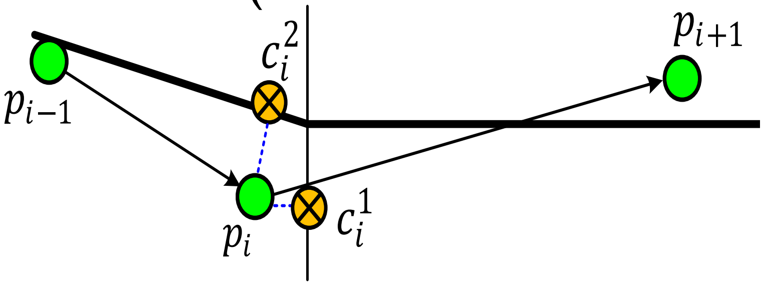
Map-matching

- Solution (incorporating spatial information)
  - (Observation Probability) Model local possibility



$$N(c_i^j) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{(x_i^j - \mu)^2}{2\sigma^2}}$$

  - (Transmission Probability) Considering context (global)



$$V(c_{i-1}^t \to c_i^s) = \frac{d_{i-1\to i}}{w_{(i-1,t)\to(i,s)}}$$

  - Spatial analysis function

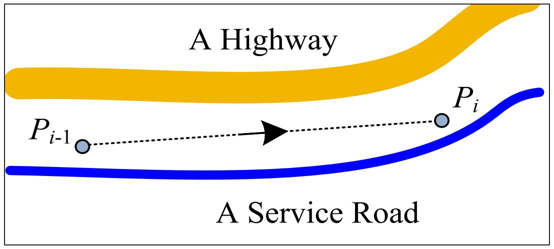
$$F_s(c_{i-1}^t \to c_i^s) = V(c_{i-1}^t \to c_i^s) * N(c_i^s)$$

Yin Lou, Chengyang Zhang, **Yu Zheng**, et al. Map-Matching for Low-Sampling-Rate GPS Trajectories. In ACM SIGSPATIAL GIS 2009

Map-matching

- Solution (Cosine Similarity)
  - Temporal analysis function (Considering temporal information)
  - Shortest path is used.

$$F_t(c_{i-1}^t \rightarrow c_i^s) = \frac{\sum_{u=1}^{k}(e'_u.v \times \bar{v}_{(i-1,t)\rightarrow(i,s)})}{\sqrt{\sum_{u=1}^{k}(e'_u.v)^2} \times \sqrt{\sum_{u=1}^{k} \bar{v}^2_{(i-1,t)\rightarrow(i,s)}}}$$

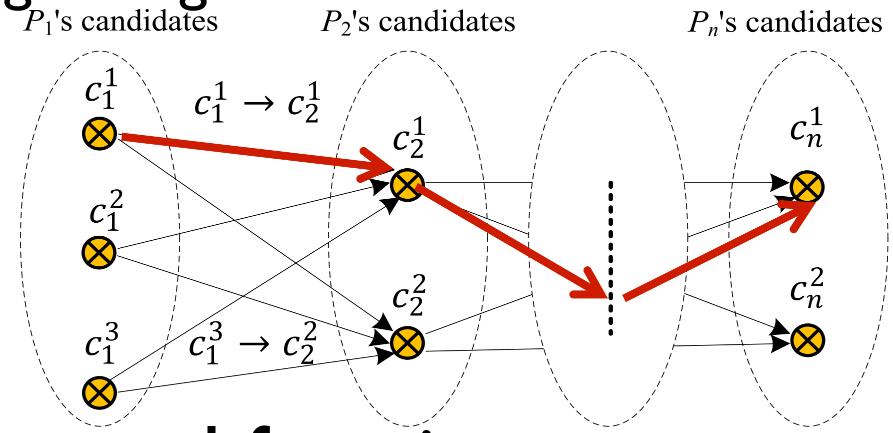


$$\bar{v}_{(i-1,t)\rightarrow(i,s)} = \frac{\sum_{u=1}^{k} l_u}{\Delta t_{i-1\rightarrow i}}$$

Yin Lou, Chengyang Zhang, **Yu Zheng**, et al. Map-Matching for Low-Sampling-Rate GPS Trajectories. In ACM SIGSPATIAL GIS 2009

Map-matching

- Aggregating
  - Spatial and temporal information
  - Local and global information
- Dynamic programing



- Spatio-temporal function

$$F(c_{i-1}^t \to c_i^s) = F_s(c_{i-1}^t \to c_i^s) * F_t(c_{i-1}^t \to c_i^s), 2 \le i \le n$$

Yin Lou, Chengyang Zhang, **Yu Zheng**, et al. Map-Matching for Low-Sampling-Rate GPS Trajectories. In ACM SIGSPATIAL GIS 2009

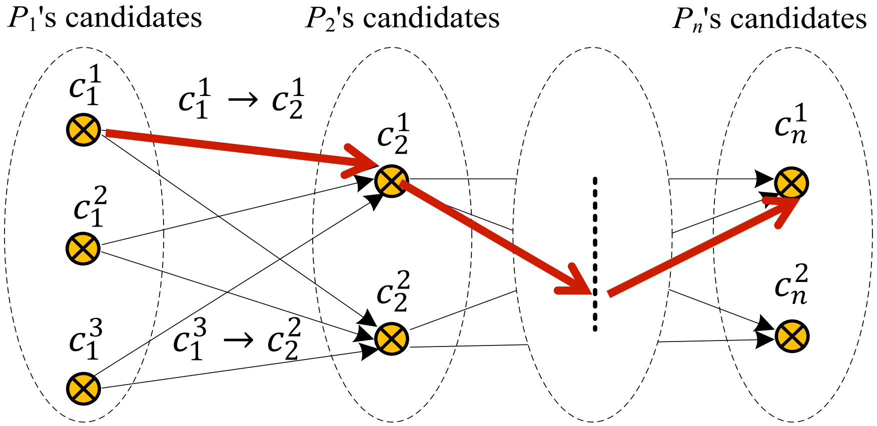
Map-matching

- Path Selection

$$F(P_c) = N(c_1^{s_1}) + \sum_{i=2}^{n} F(c_{i-1}^{s_{i-1}} \to c_i^{s_i})$$

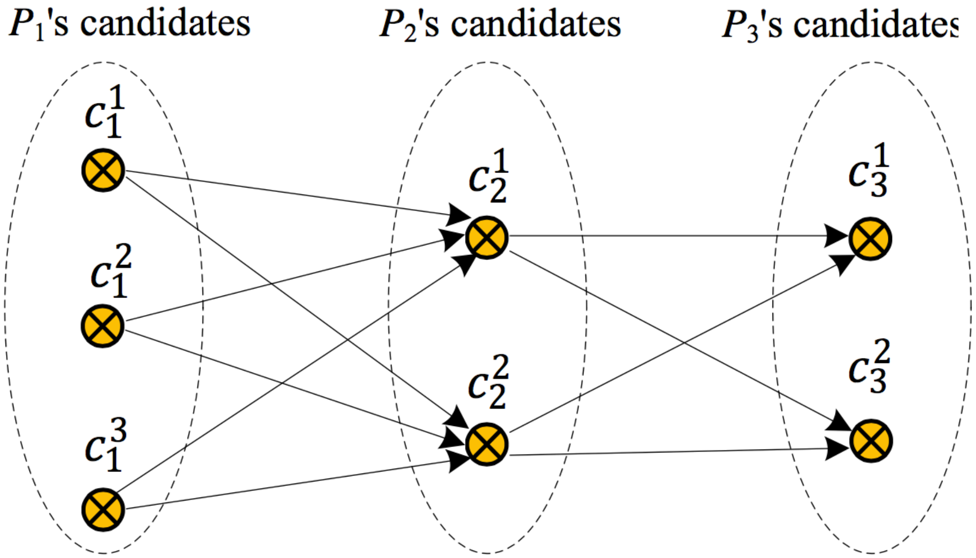
$$P = argmax_{P_c} F(P_c)$$

- Dynamic programing



Yin Lou, Chengyang Zhang, **Yu Zheng**, et al. Map-Matching for Low-Sampling-Rate GPS Trajectories. In ACM SIGSPATIAL GIS 2009

Map-matching Example



$f[\,]$:

c_1^1	c_1^2	c_1^3	c_2^1	c_2^2	c_3^1	c_3^2
0.8	0.2	0.5	▨	▨	▨	▨

N:

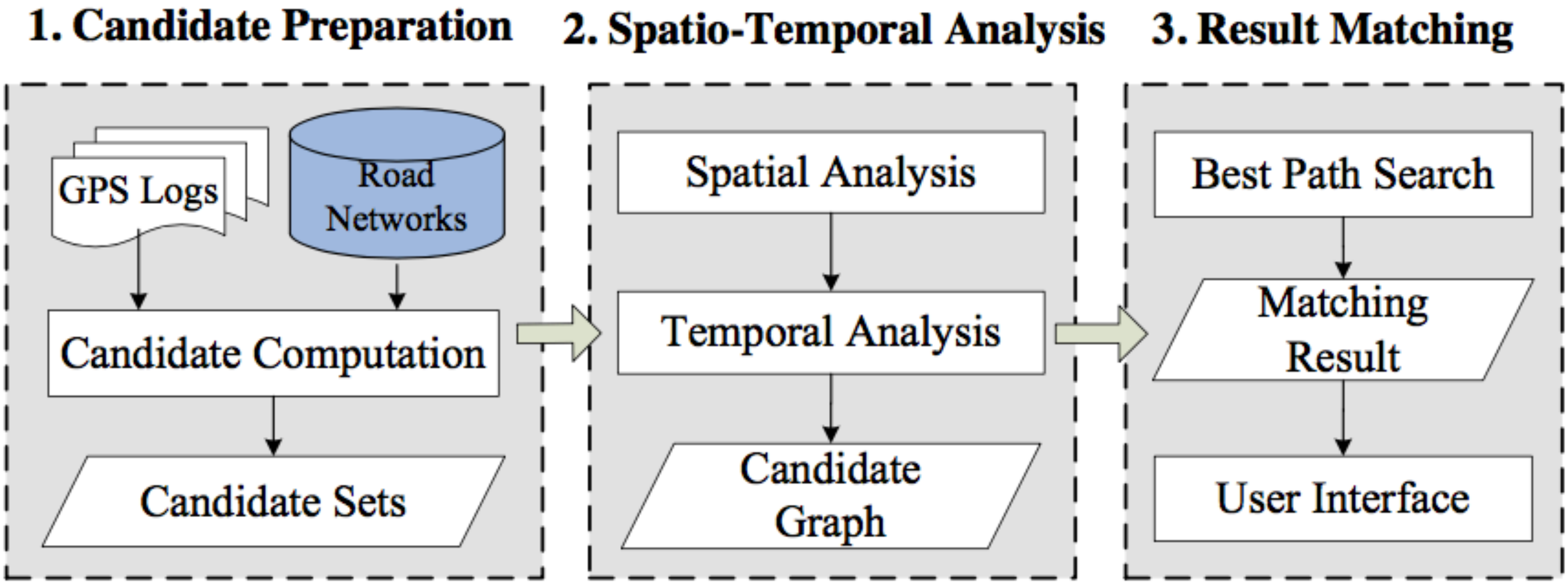
c_1^1	c_1^2	c_1^3	c_2^1	c_2^2	c_3^1	c_3^2
0.8	0.2	0.5	0.6	0.6	0.4	0.3

(V, F_t):

	$\rightarrow c_2^1$	$\rightarrow c_2^2$
c_1^1	(0.5,0.5)	(0.8,0.5)
c_1^2	(0.3,0.4)	(0.1,0.9)
c_1^3	(0.4,0.6)	(0.9,0.9)

	$\rightarrow c_3^1$	$\rightarrow c_3^2$
c_2^1	(0.2,0.6)	(0.1,0.7)
c_2^2	(0.3,0.3)	(0.5,0.9)

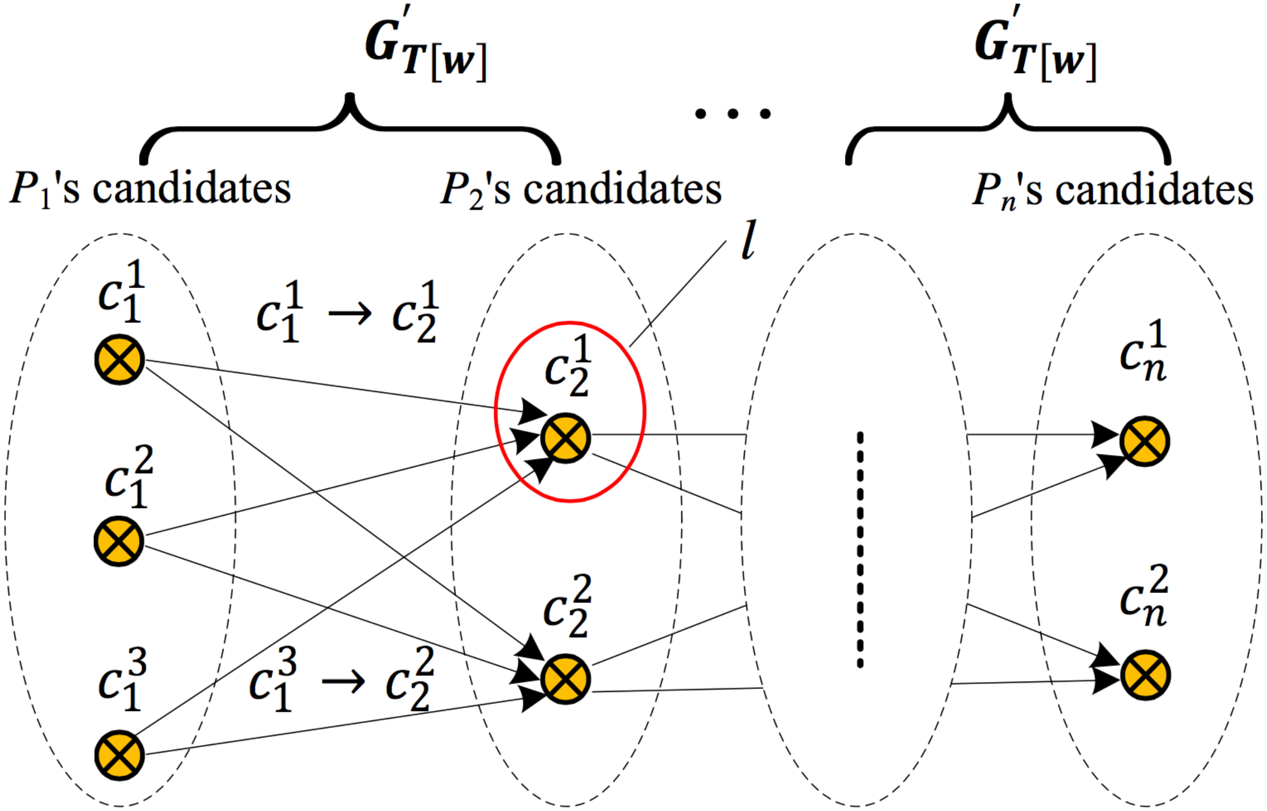
Map-matching framework



Yin Lou, Chengyang Zhang, **Yu Zheng**, et al. Map-Matching for Low-Sampling-Rate GPS Trajectories. In ACM SIGSPATIAL GIS 2009

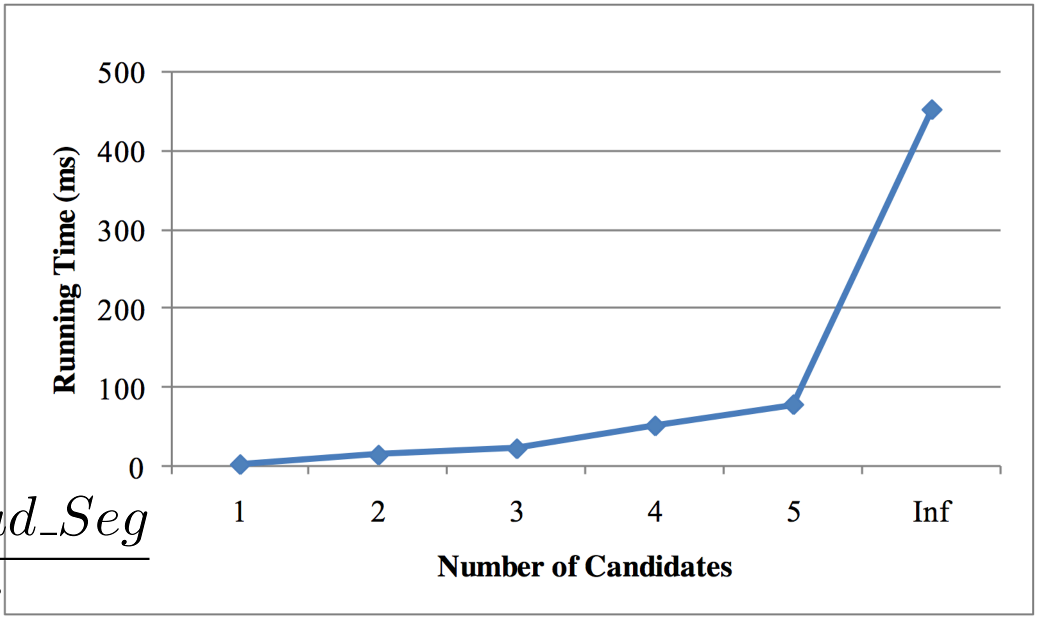
Localized ST-Matching Strategy

- Path Selection

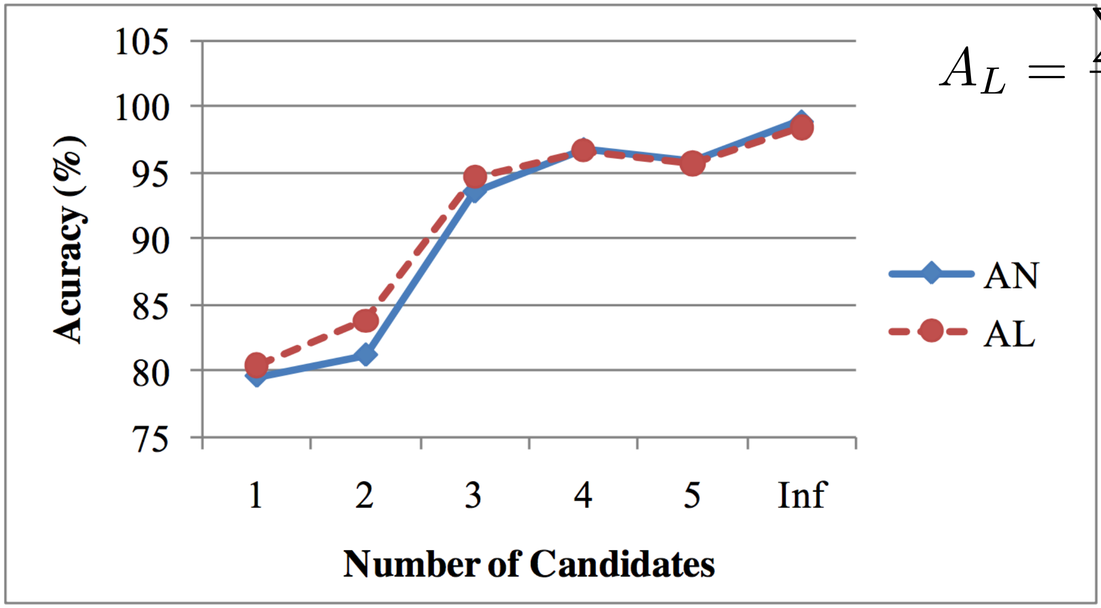


Yin Lou, Chengyang Zhang, **Yu Zheng**, et al. Map-Matching for Low-Sampling-Rate GPS Trajectories. In ACM SIGSPATIAL GIS 2009

Evaluations



$$A_N = \frac{\#Correctly\_Matched\_Road\_Seg}{\#all\_road\_segments}$$



$$A_L = \frac{\sum Length\_Matched\_Road\_Seg}{Length\_of\_the\_trajectory}$$

Yin Lou, Chengyang Zhang, **Yu Zheng**, et al. Map-Matching for Low-Sampling-Rate GPS Trajectories. In ACM SIGSPATIAL GIS 2009

Course Project

Project 1 directions

What is your project goal?

- What new story you want to tell?
- New contents to sample?
- New sampling methods via API?
- New statistics of YouTube, view count distribution, dynamics, or # uploaders/active users?
- Analysis on other websites, Twitter, Facebook, Foursquare, Yelp, with API interfaces

Broad impacts? (Keep in mind)

- How YouTube is evolving?
  - More business or personal videos? How to distinguish the two
  - How special events, e.g., NBA game, breaking news, affect the uploading, viewing behaviors
- Online Marketing, advertising?

A Few Words on Course Project I

Project I: Collecting and Measuring Online Data

- Team work; each team 3-4 students.
- Starting date: Week 3 (9/8 R)
- Proposal Due: Week 4 (9/15 R) 2 pages roughly
- Due date/time: Before Class on Week 8 (10/13 R)
- Presentation date/time: Class on Week 8 (10/13 R)
  - Selected teams only
- Requiring Programming in C/C++, Java, Python, and, etc

- Choose one online site/service with APIs to download data, or use existing datasets.
- Examples:
- (1) estimate site statistics, or
- (2) applying machine learning methods to predict future trends, or
- (3) perform time-series analysis to capture dynamic patterns,
- or something else, as long as your work can potentially bring research value to the community.

A Few Words on Course Project I

- Group meeting with Prof Li by appointment)
  - **Week 3 (9/8 R),** Starting date
  - **Week 4 (9/15 R),** Proposal Due: 2 pages roughly (upload it to discussion board)
  - **Week 5 (9/22 R)**, Methodology due (upload it to discussion board)
  - **Week 6 (9/29 R)**, Results due (upload it to discussion board)
  - **Week 7 (10/6 R)**, Conclusion due (upload it class discussion board)
  - **Week 8 (10/13 R)**, Final Report due at 11:59pm EST & Self and Cross-evaluation due at 11:59pm EST
  - **Week 8 (10/13 R)**, In-class Presentation (10 min) (Selected teams only)

Course Project II

- Projects will be in groups!
  - 3-5 students per group, depending on enrollment

- Topics on your choice (related to big data analytics)
  - Application-driven
  - Fundamental data analytics research
  - Data sources on course website http://wpi.edu/~yli15/courses/DS504Fall16/Resources.html

Talk to me once you have an idea.

Next Class: Data Management

- Do assigned readings before class
  - Be prepared, read and review required readings *on your own in advance!*
  - *Do literature survey: find and read related papers if any*
  - *Bring your questions to the class and look for answers during the class.*
- Submit reviews/critiques
  - In myWPI before class
  - Bring 2 hardcopies to the class
  - Hand in one copy, and keep one copy with you.

Review Writing:
http://users.wpi.edu/~yli15/courses/DS504Fall16/Critiques.html

- Attend in-class discussions
  - Please ask and answer questions in (and out of) class!
  - Let's try to make the class interactive and fun!