Route analysis & Privacy

Adam Ludvig
258801

hunludvig@gmail.com

UNIVERSITY OF EASTERN FINLAND

Outline

- Far Out – Predicting long-term human mobility
  - Long-term prediction, GPS, Continuous and Cellular Pattern
- Personal continuous route pattern mining
  - Data mining, Route pattern, GPS, Privacy
- Unique in the Crowd – Privacy bounds of human mobility
  - Privacy, Anonym mobility dataset, 1.5M users, GSM

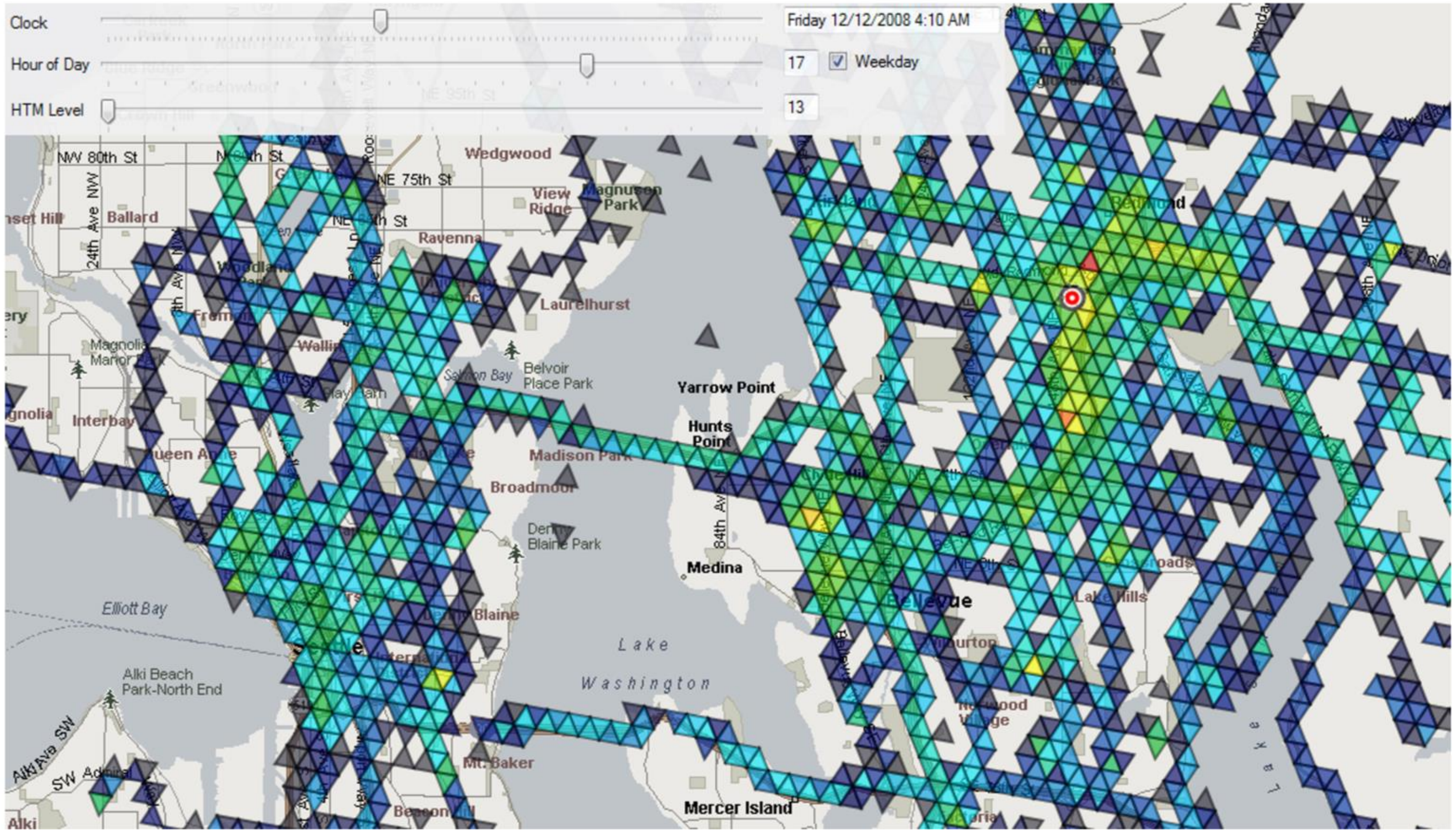
Far Out

- Where are you going to be 285 days from now at 2PM?
- A model of long-term human mobility
- Visualizing the patterns in a meaningful way
- “Need a haircut? In 4 days, you will be within 100 meters of a salon that will have a $5 special at that time.”

by Adam Sadilek & John Krumm @ Microsoft Research, 2012

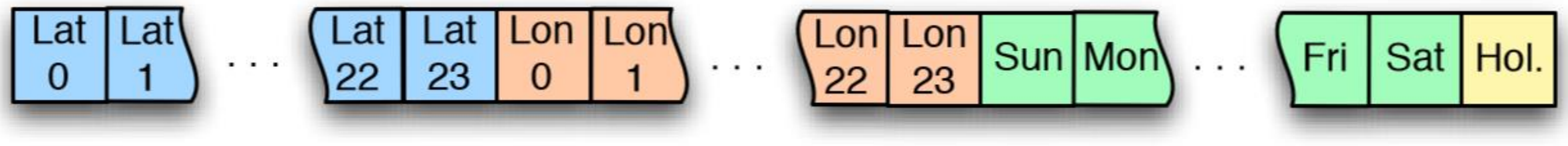
Data

- GPS
  - Seattle
- 307 people
- 396 vehicles
- 7-1247 days
  - Avg 46 days
  - Total 32000 days
- Triangular cells
  - Side 400m

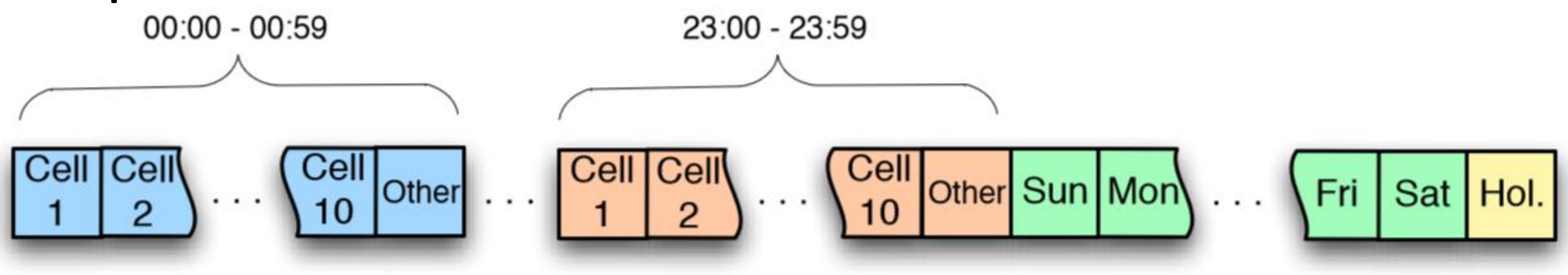


Model

- Fourier analysis to find periodicity
- PCA to extract strong patterns and eliminate insignificant features
- Continuous representation:



- Cellular representation:



Fourier analysis, find periodicity of data

- Discrete Fourier Transformation
- Find periodicity
- Complex representation
  - Latitude + i longitude
- O(NlogN) with FFT

[Wikipedia]

Principal Component Analysis

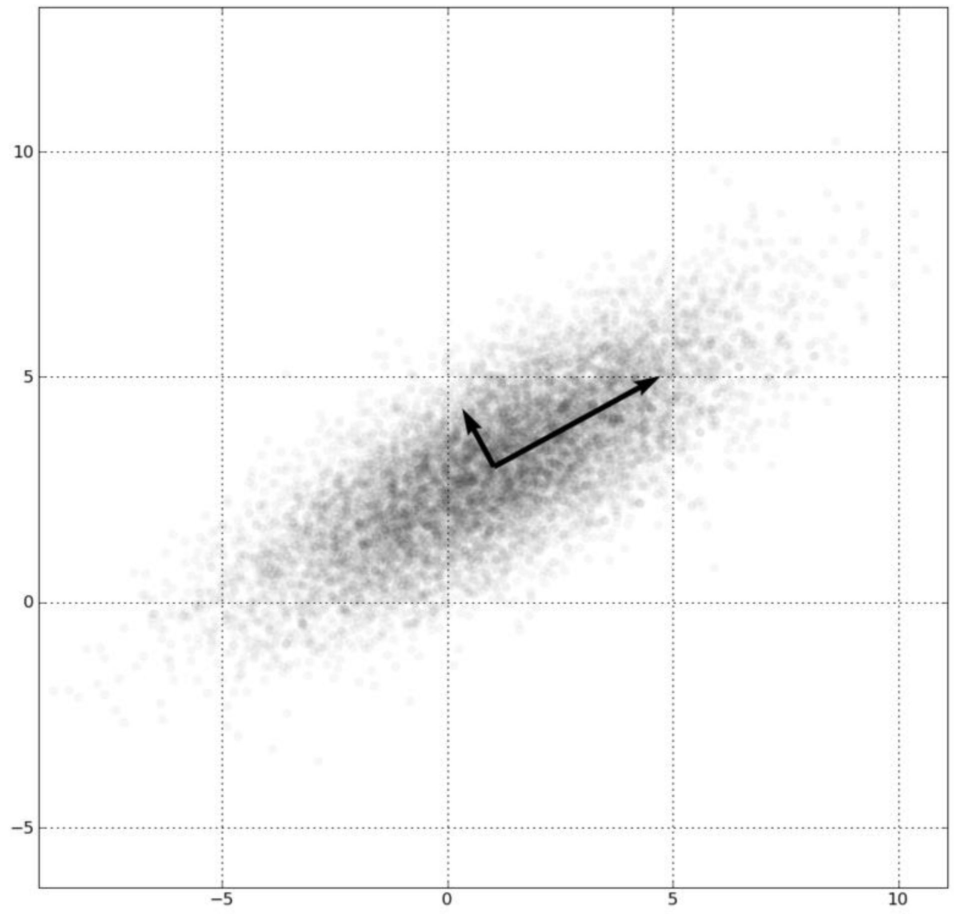
- Dimensionality reduction
- Find linearly uncorrelated, "principal" components
- Numerically stable algorithm by Singular Value Decomposition (SVD) $O(mn^2)$
- Decomposition of M $[m \times n]$ matrix

$$M = U\,S\,V$$

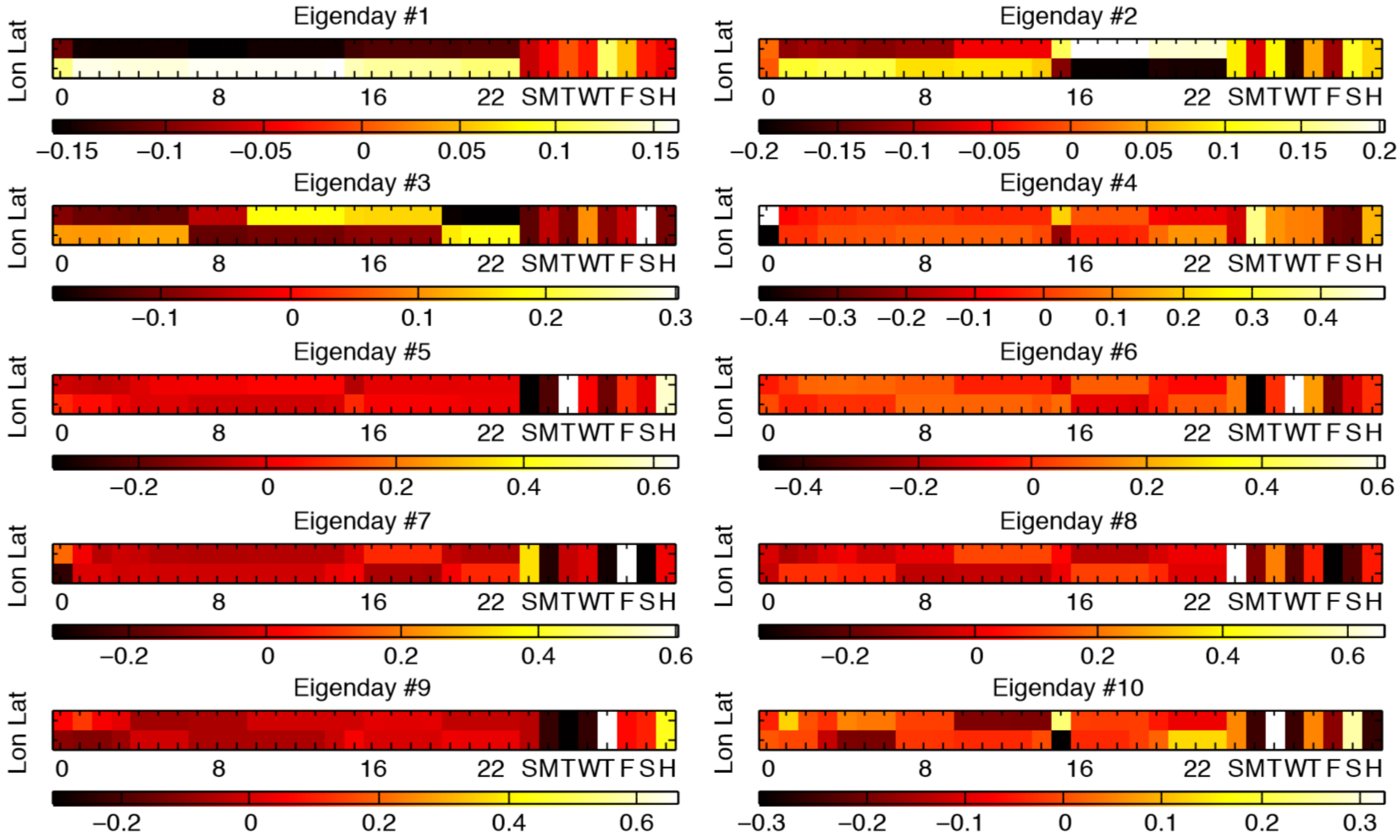
$U - [m \times m]$ complex unitary matrix

$S - [m \times n]$ rectangular diagonal matrix

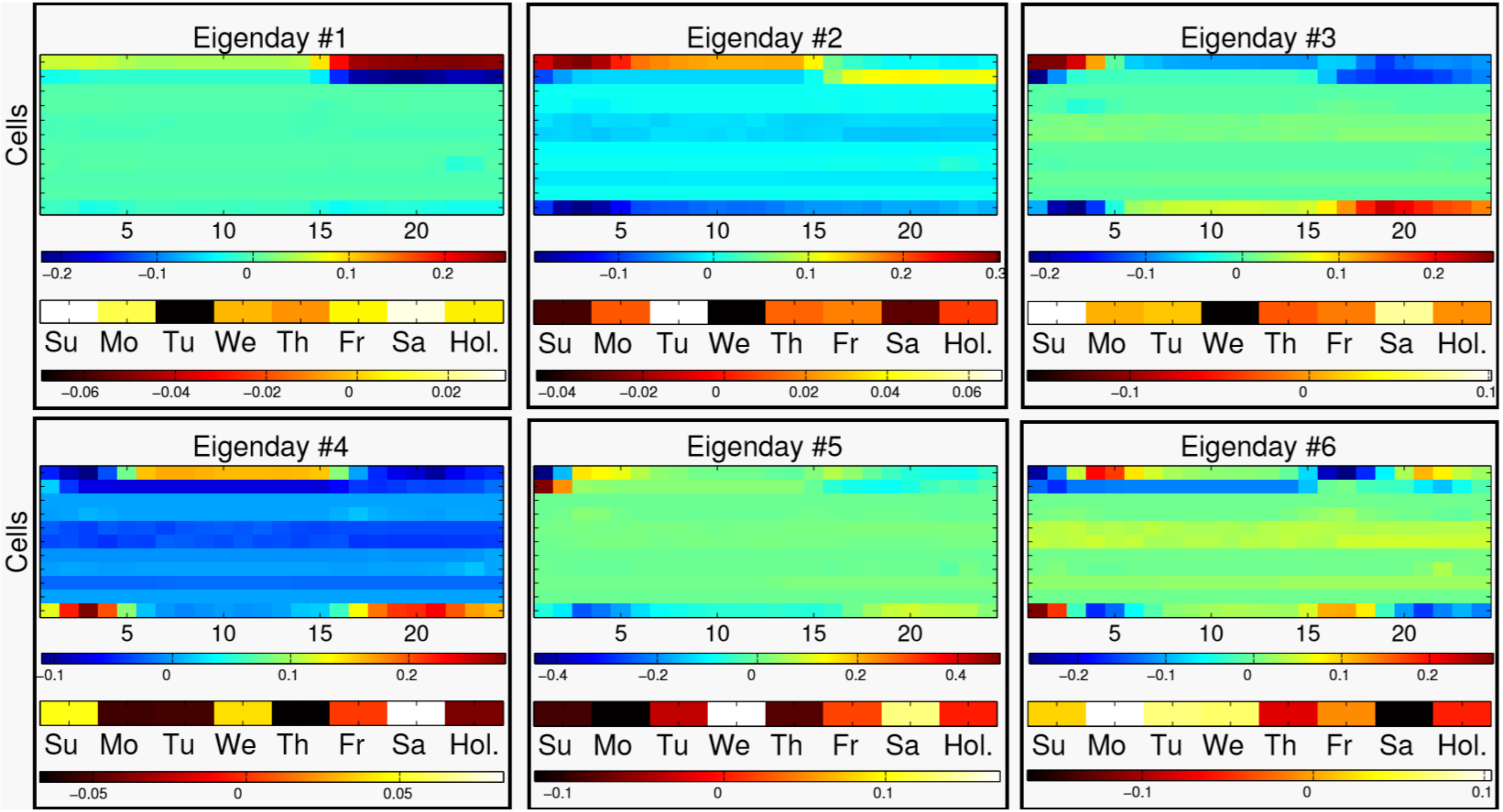
$V - [n \times n]$ complex unitary matrix



Ten most typical eigendays, continuous case



Six most typical eigendays, cellular case



Models of prediction

Extract ω observed feature vector from t time of prediction

E.g. Is t Monday? Is t holiday?

Models

- Mean Day Baseline Model
- Projected Eigenday Model
- Segregated Eigendays Model

Improve by

- Adapting to pattern drift

Mean Day Baseline Model

- Average Lat and Lon values for each hour and each day type
  $24 \times 7 \times 2 = 336$ hour type in this case
- Results the mean of all days matching ω

Projected Eigenday Model (PCA)

- Project ω onto features subspace of eigendays’ space
  - Projection provides w weights of eigenvectors
- Results the w weighted average of eigendays
- It is a least-squares fitting problem

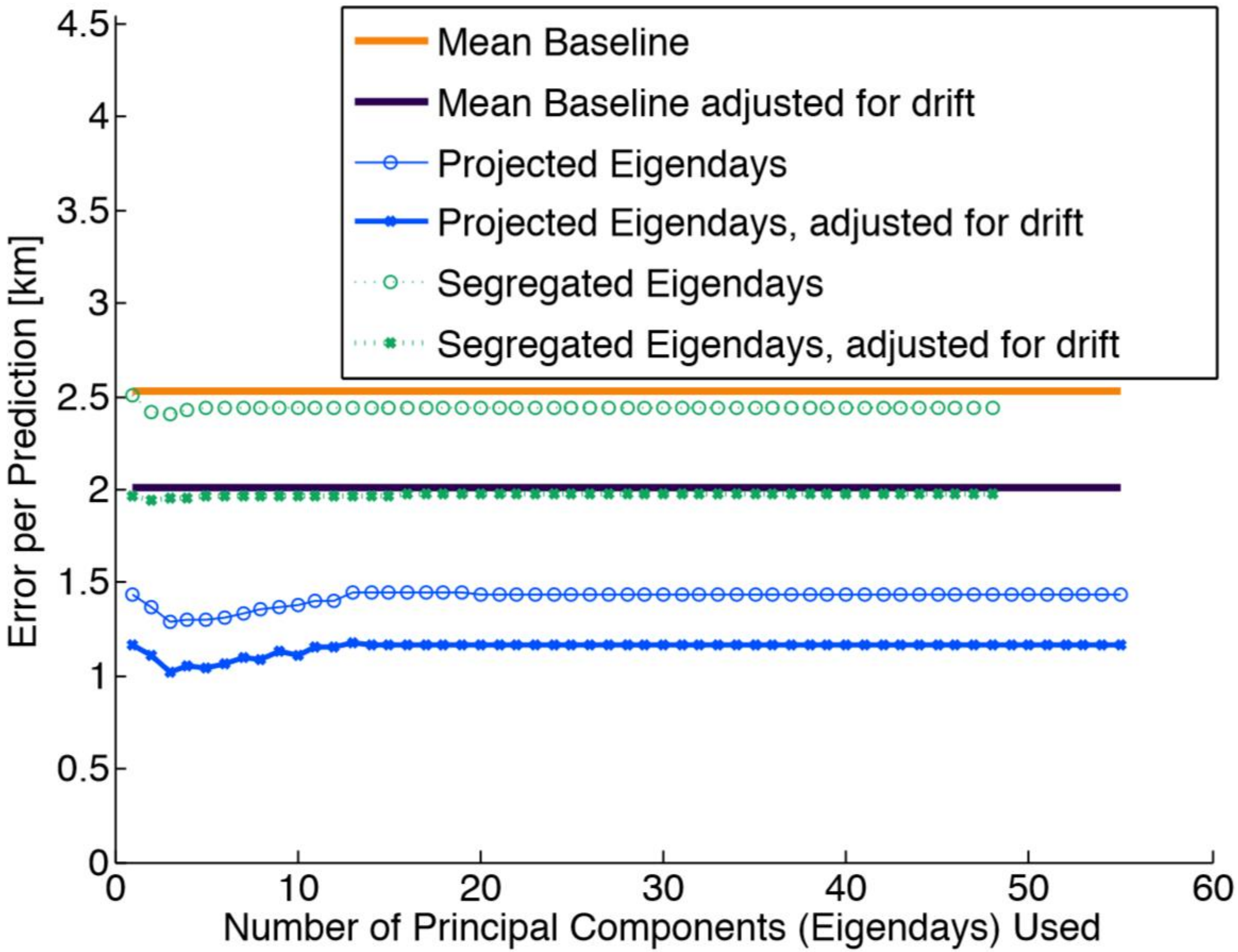
Segregated Eigendays Model (PCA)

- Separate library of eigendays for each day type (e.g. Monday-Holiday)
- Applied weights are proportional to the variance of eigenday on training data

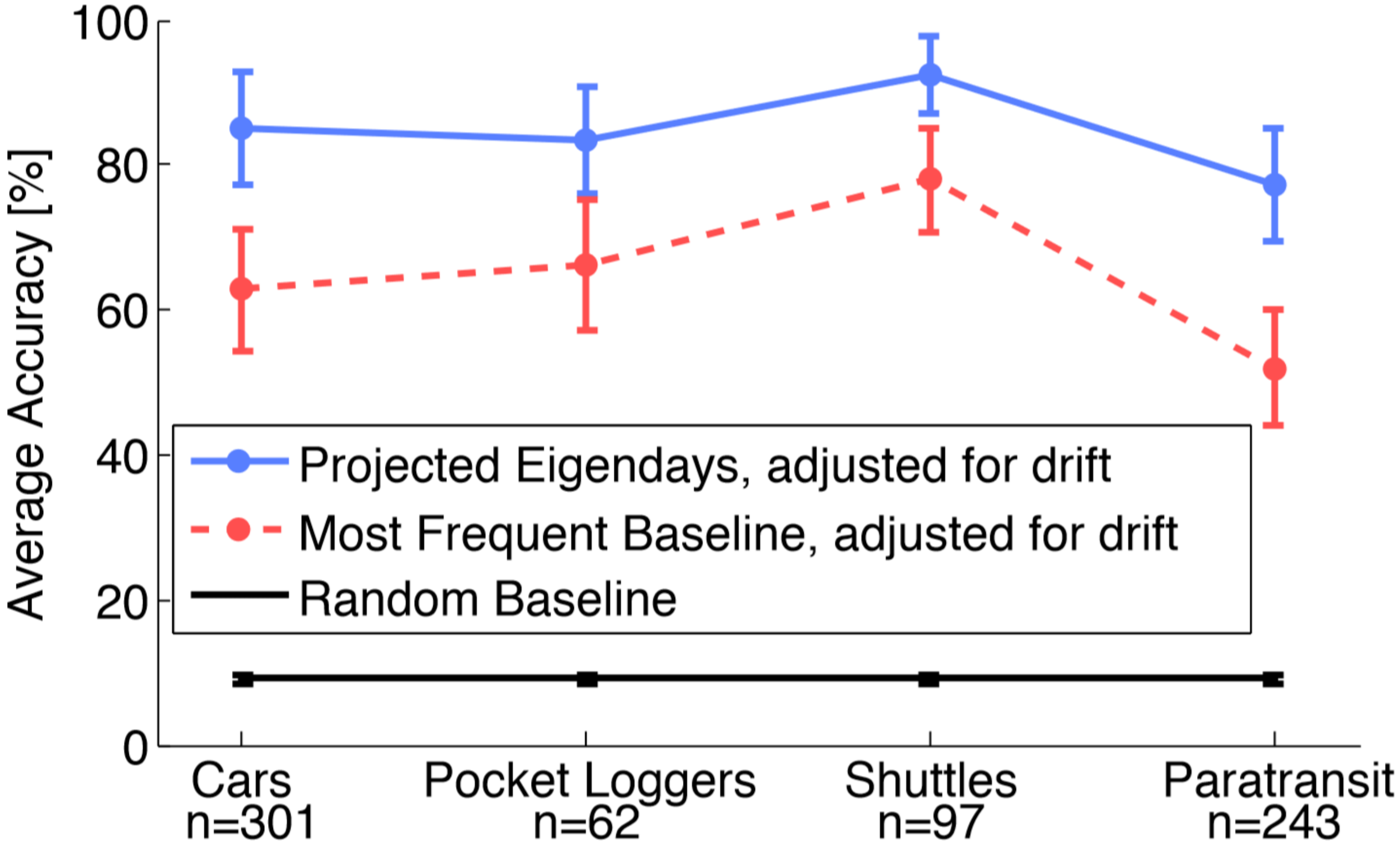
Adapting to pattern drift

- Linear decay to training data
- Applied to mean and variance calculation that are used to normalize data
- Reduces error by 27%

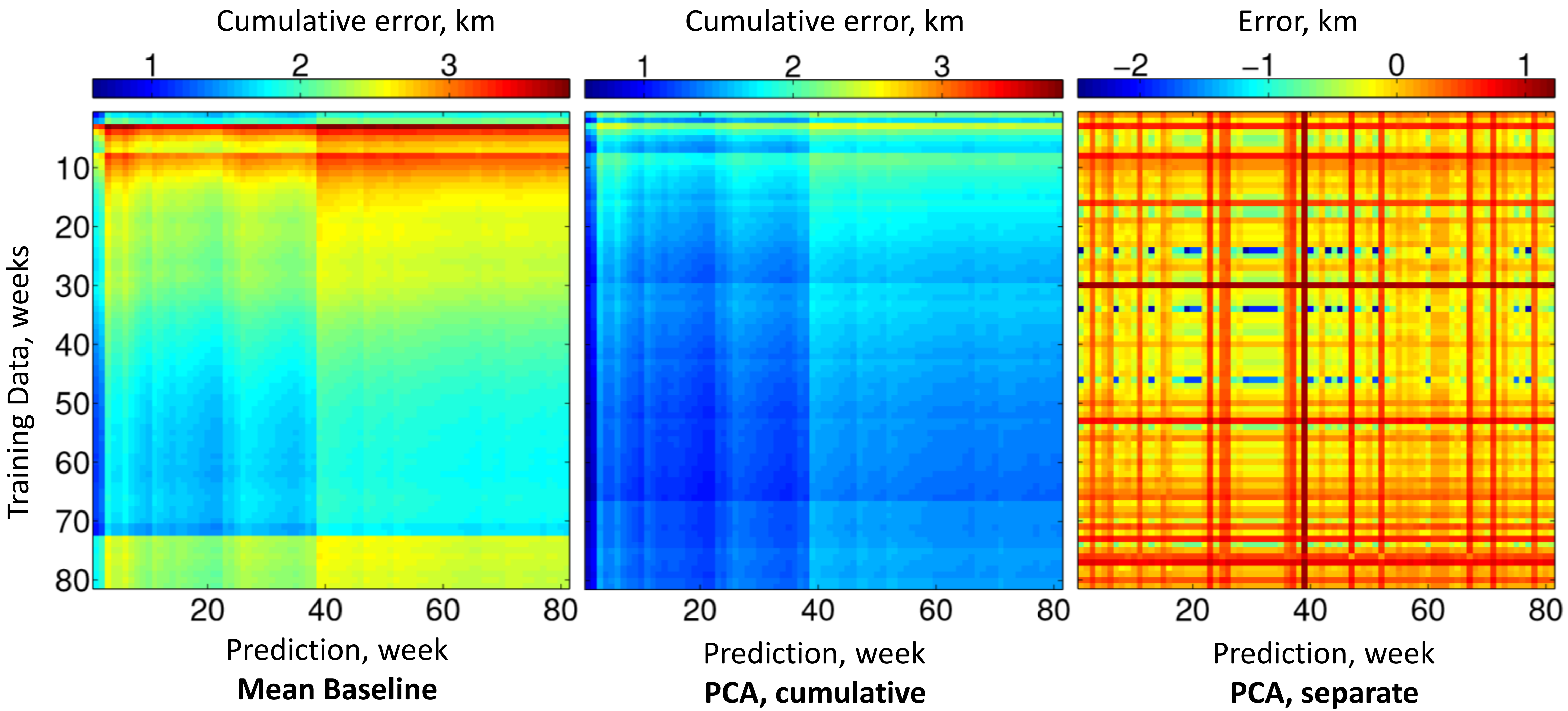
Number of eigendays



Experiment & Results



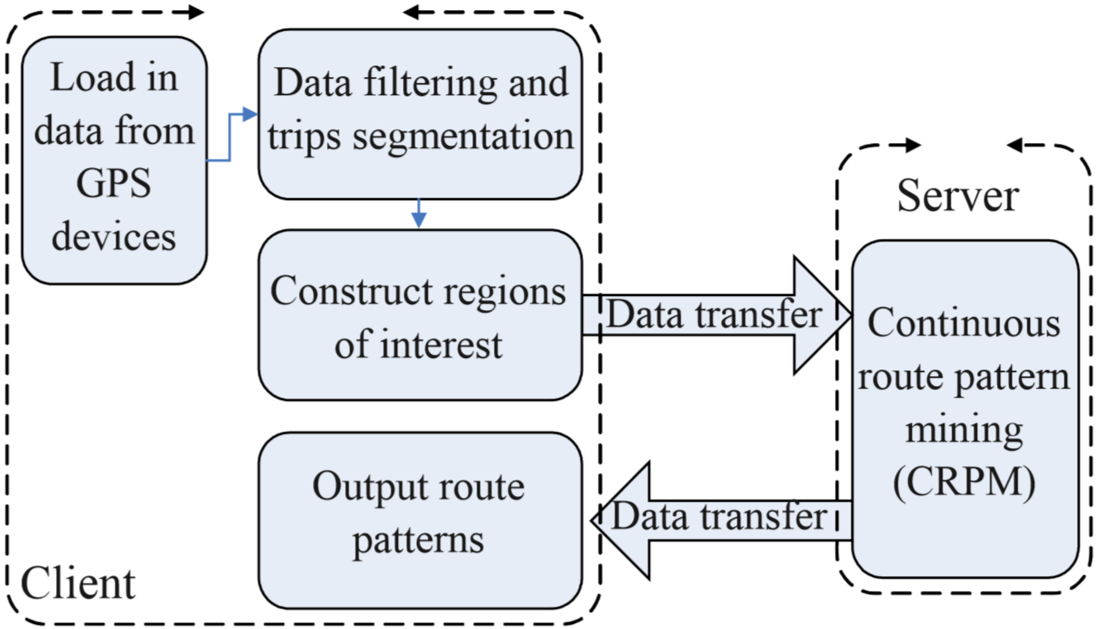
Prediction error



Personal continuous route pattern mining

- Record personal routes by GPS devices and smartphones
- Trajectory preprocessing on mobile device
- Spatially meaningless data sent to server to preserve privacy
- Route Pattern mining on the server

by Qian YE,
Ling CHEN,
Gen-cai CHEN
@ Zhejiang University
Hangzhou, 2008



Route preprocessing

- Adaptive sampling interval to reduce processed data amount
  - Decrease or increase interval based on estimated speed
- Trip filtering to remove measurement errors
  - 5 filters
- Spatio-Temporal Sequence into Regional-Temporal sequence

Trip filtering

- Duplication filter

  Drop latter measurement point within λ_{dup} distance of previous

- Speed filter

  Drop latter point if the speed calculated by previous point is unreasonable

- Acceleration filter

  Compare speed to the previous segment and drop point if over threshold

- Total-distance filter

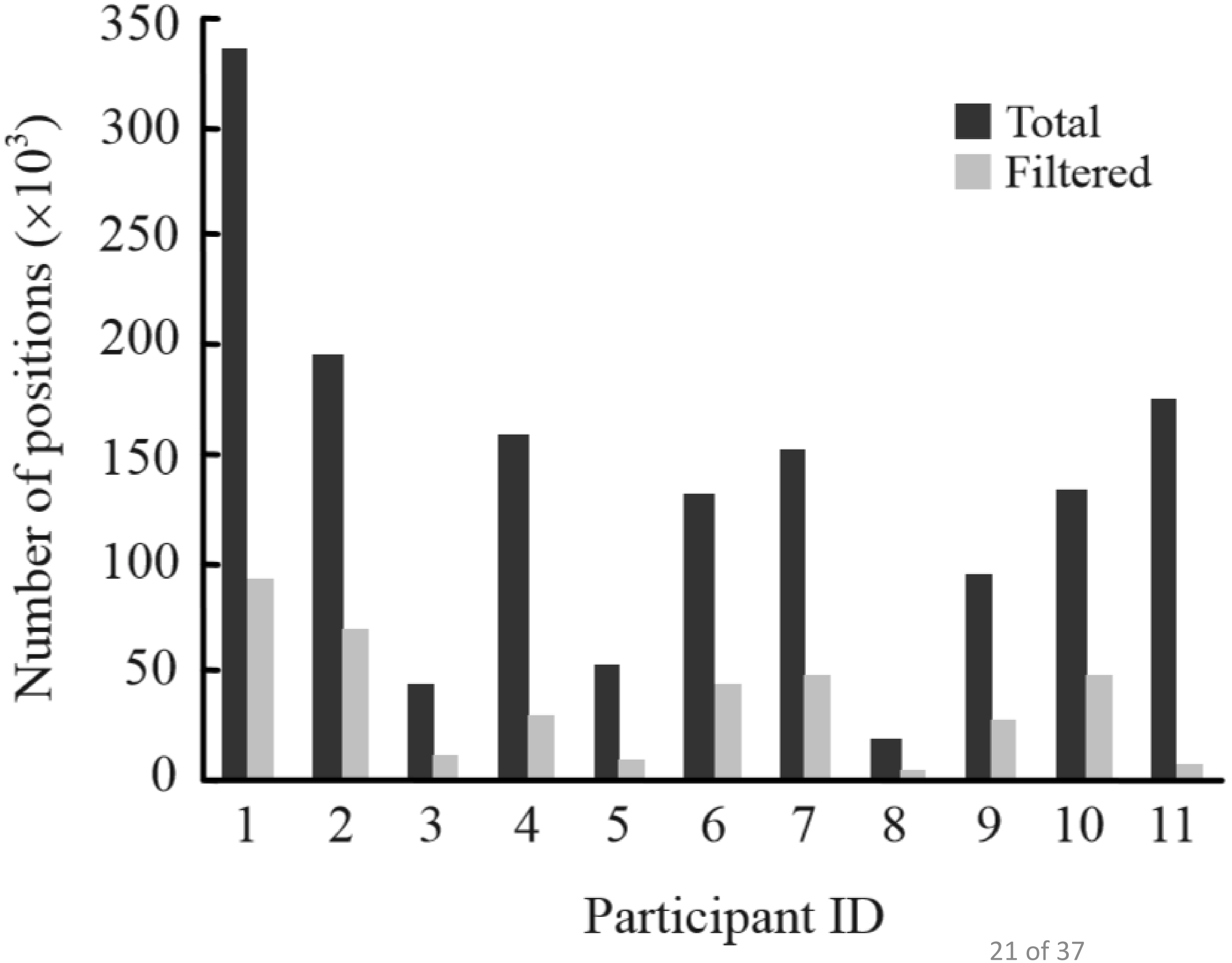
  A trip is dropped if all the data points are within a λ_{tdis} distance of its centroid

- Angle filter

  Drop the middle point if three consecutive points in a short time form a sharp angle,
  Smooth the trip

Result of filtering

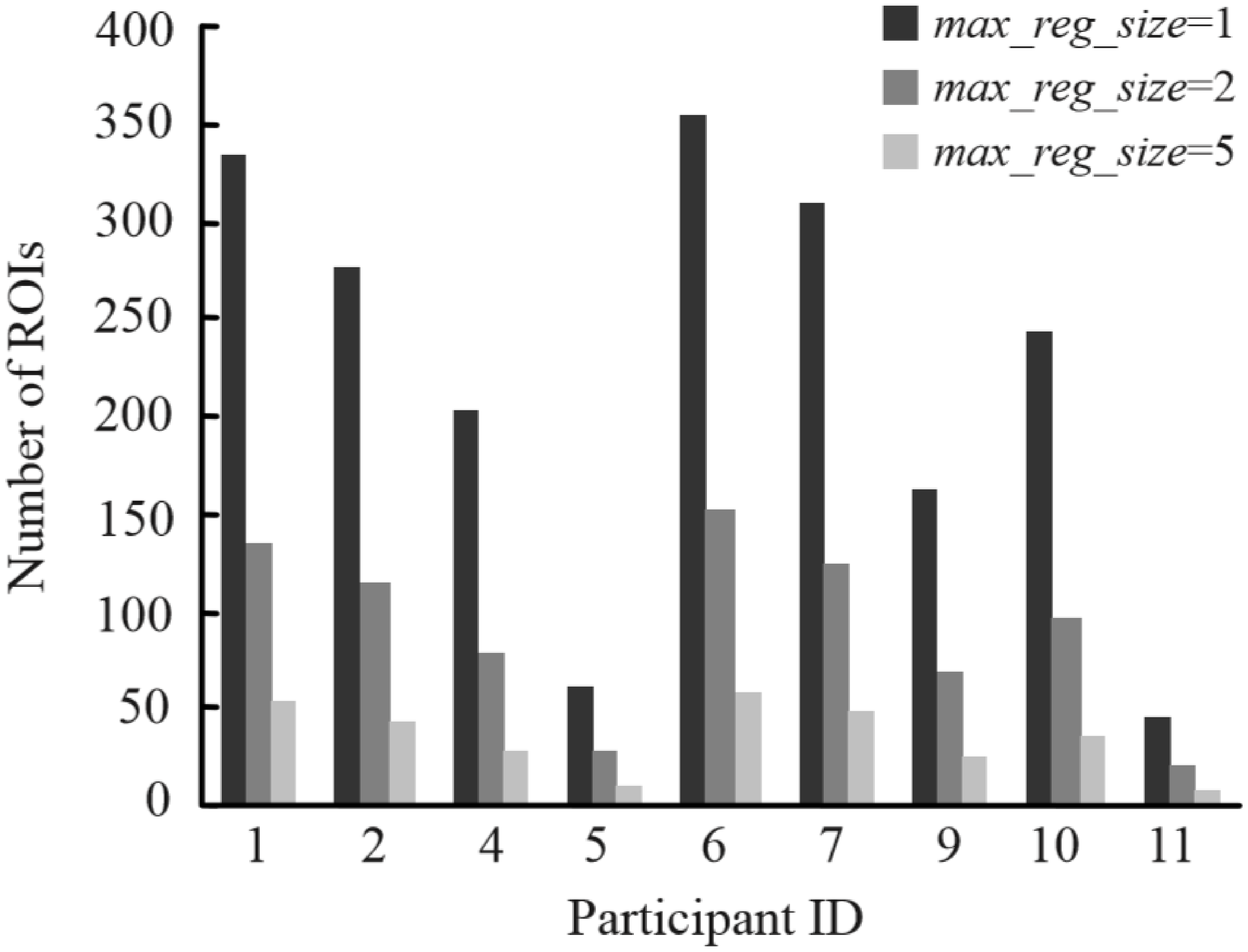
- Not tested separately
- Fixed order
- Not independent



Regional-Temporal Sequence

- Spatio-Temporal Sequence (STS) $\{\ldots,(x_i, y_i, t_i),\ldots\}$
- Create square grid
- Assign trajectory points to cells (CTS)
- Compute cell density of cells
- Merge successive cells of similar density into Regions (RTS)
- Maximal region size 2..5 cells
- $\{\ldots, (R_i\ , T_{in}^i, T_{out}^i), \ldots\}$ is transferred

Regions



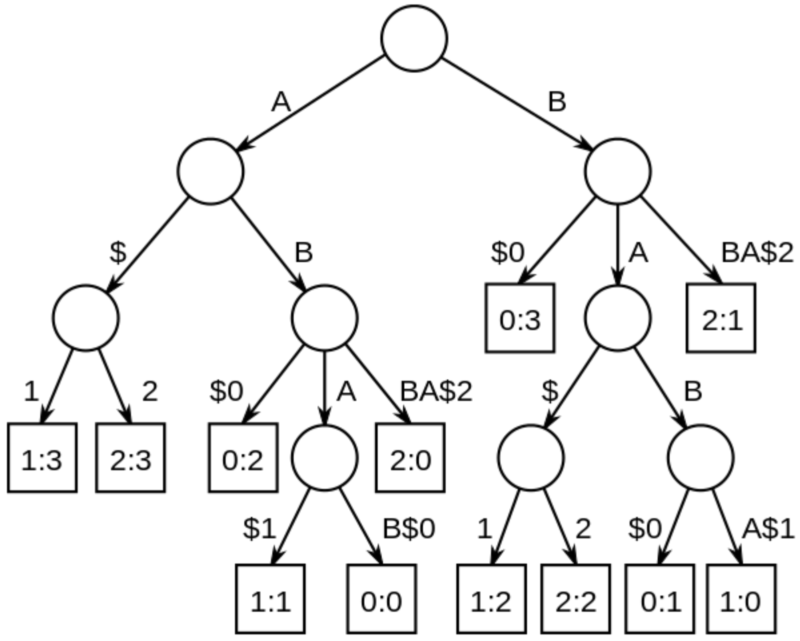
max_reg_size=1
max_reg_size=2
max_reg_size=5
Number of ROIs
400
350
300
250
200
150
100
50
0
1
2
4
5
6
7
9
10
11
Participant ID

Ways of pattern mining on RTS

- Find frequent patterns in set of RTSs
- λ_{time} threshold parameter of gap between regions
  - λ_{time}=0, only continuous subsequences are handled, same problem as longest common substring
  - $\lambda_{time}=\infty$, any gap is accepted between regions, same as PrefixSpan algorithm
- Extended PrefixSpan algorithm to take λ_{time} into consideration

Substring

- $S = \{\ldots, S_i, \ldots\}$ set of K character strings, $|S_i| = n_i, \sum n_i = N$
- λ_{min_sup} parameter of support (5)
- Find all common substrings that it is supported by at least λ_{min_sup} strings from S
- DP solution $O(\prod n_i)$
- Generalized suffix tree solution $O(K \times N)$



"ABAB", "BABA" and "ABBA"

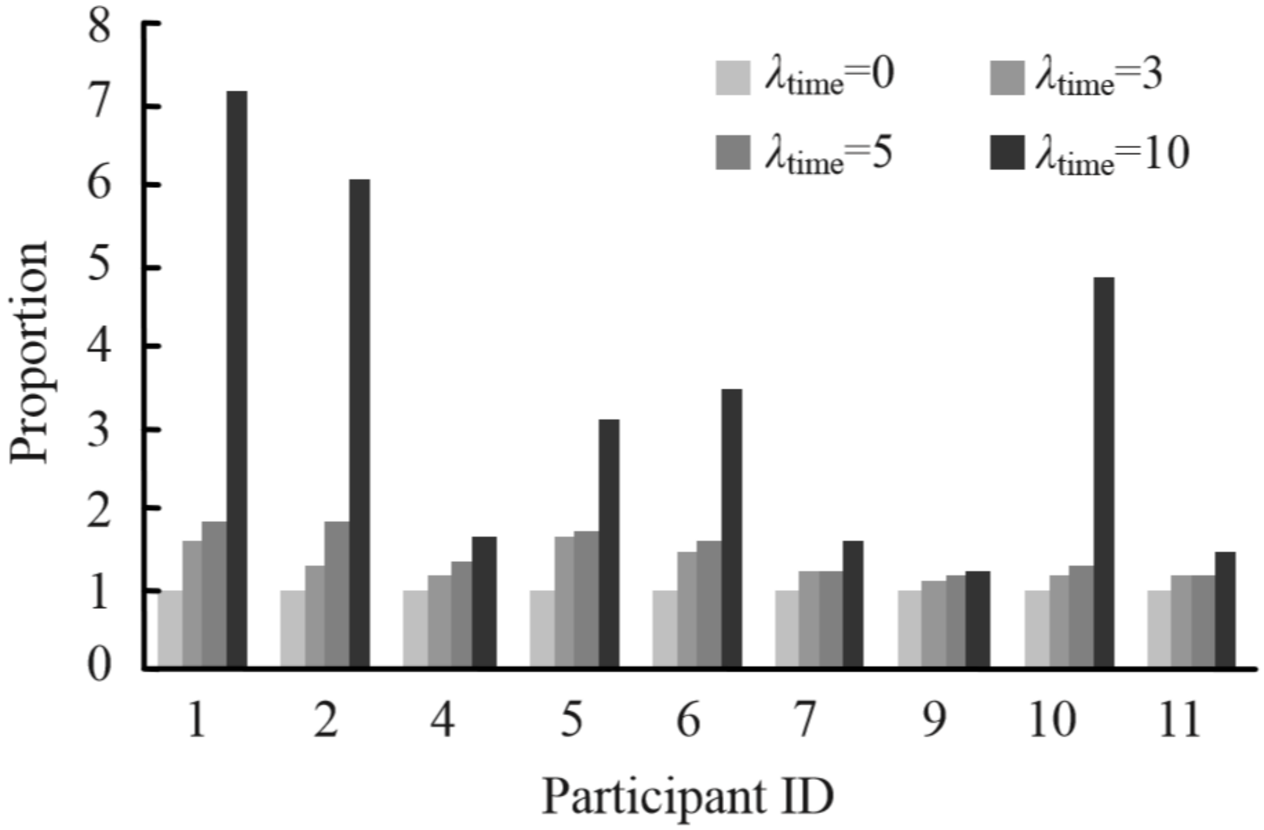
PrefixSpan

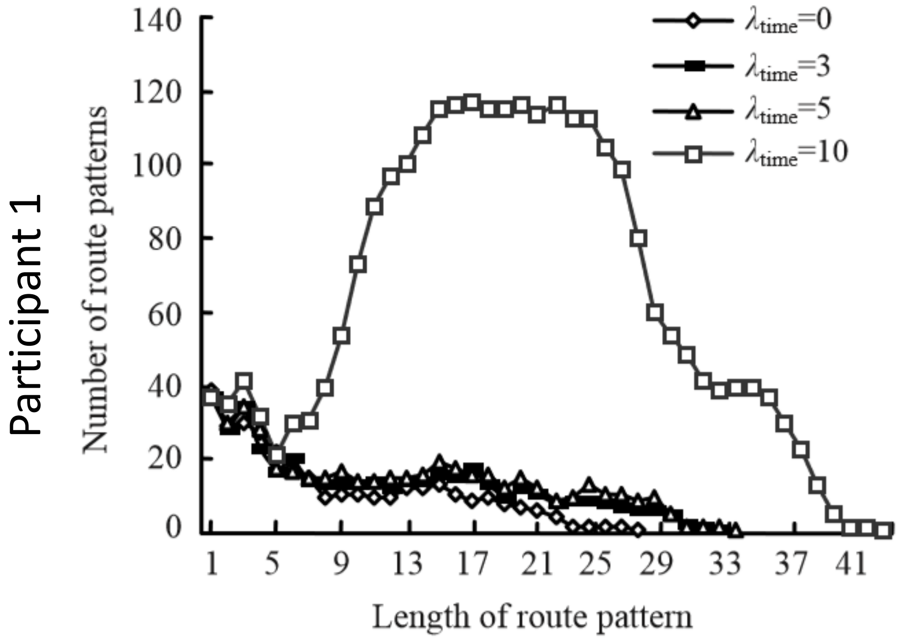
- Given two sequences $\alpha = (a_1, a_2, \dots, a_n)$ and $\beta = (b_1, b_2, \dots, b_m)$
- α is subsequence of β, α ⊆ β,
  if $\exists\ 1 \le j_1 < j_2 < \cdots < j_n \le m$ such that $\forall i\ a_i \subseteq b_{j_i}$
- Find all subsequences of S set of sequences supported by at least λ_{min_sup} sequences
- Pseudo-polynomial time complexity

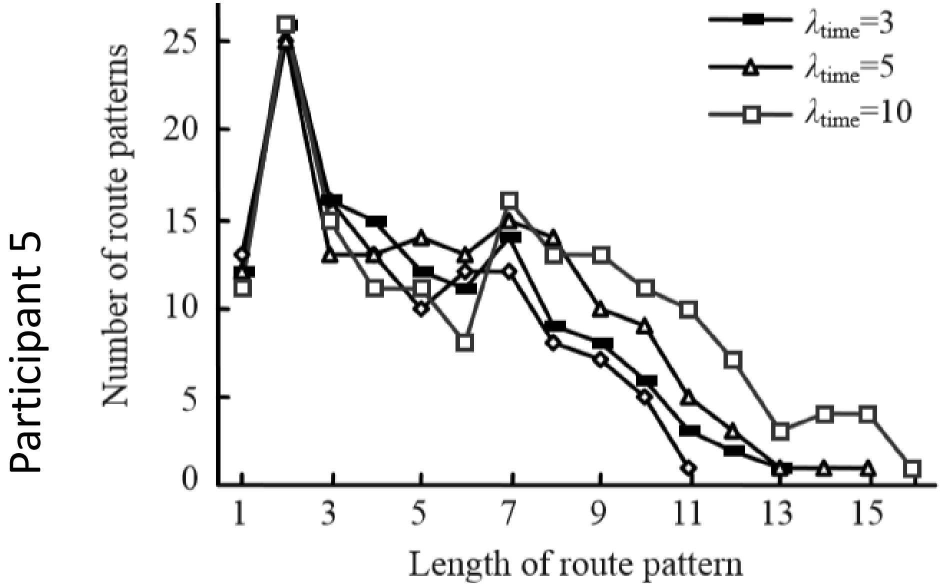
Example: β =<a(abc)(ac)d(cf)>

α_1=<aa(ac)d(c)> ⊆ β	α_4=<df(cf)> ⊈ β
α_2=<(ac)(ac)d(cf)> ⊆ β	α_5=<(cf)d> ⊈ β
α_3=<ac> ⊆ β	α_6=<(abc)dcf> ⊈ β

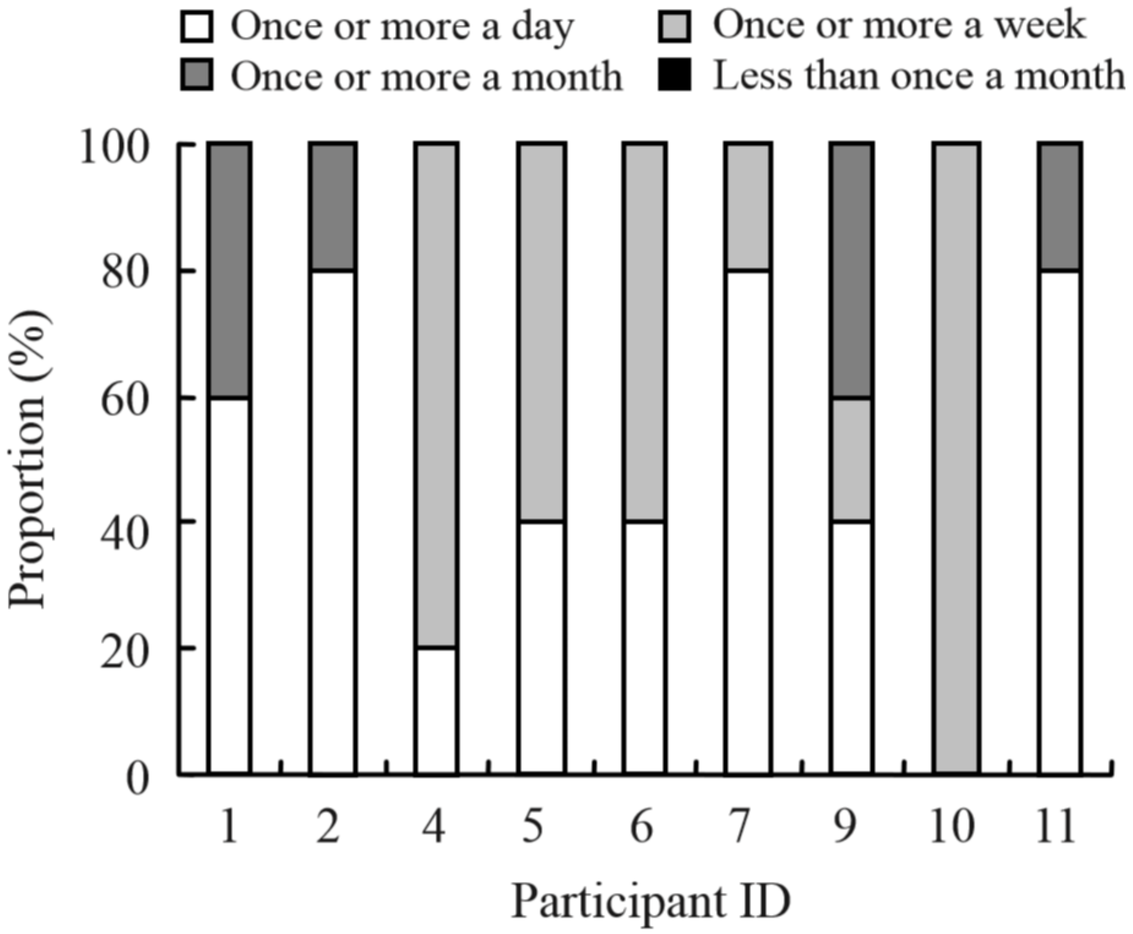
Found route patterns



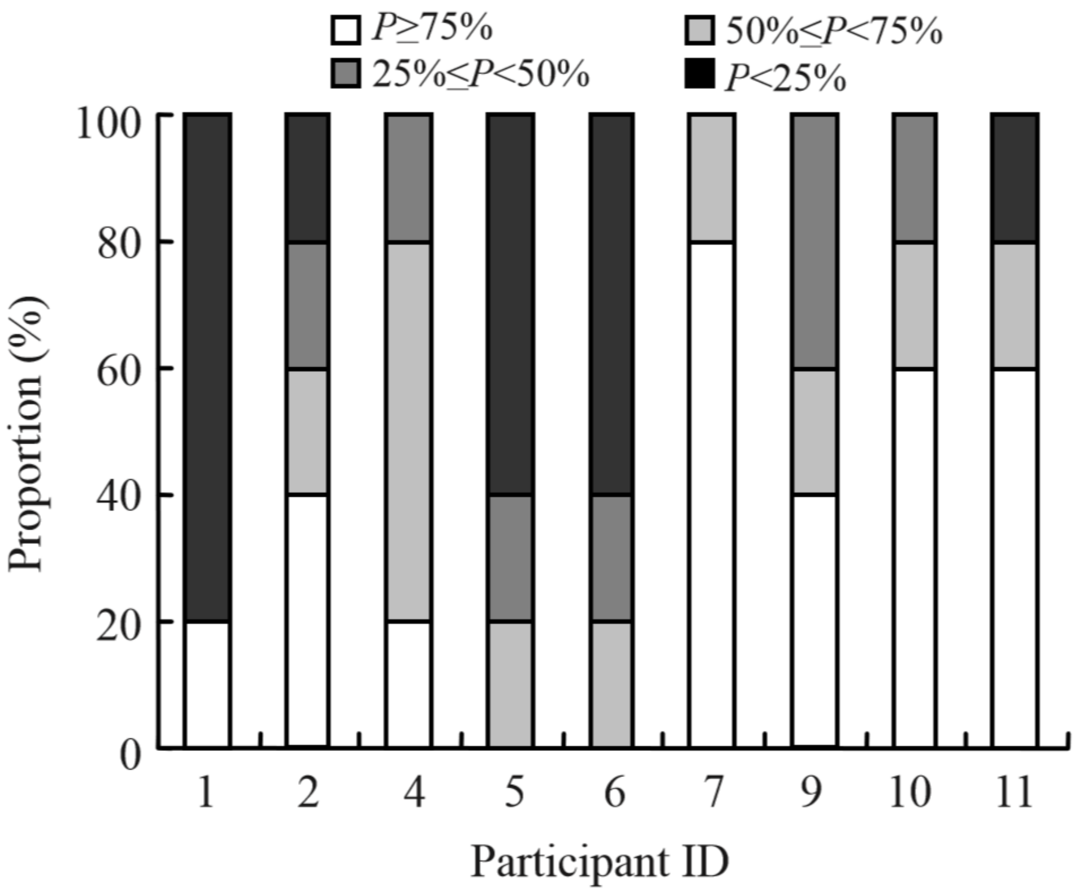




Personal pattern frequency estimation



What's your traveling frequency on the route?



What's the proportion of the route pattern to your regular trip, containing the route, in length?

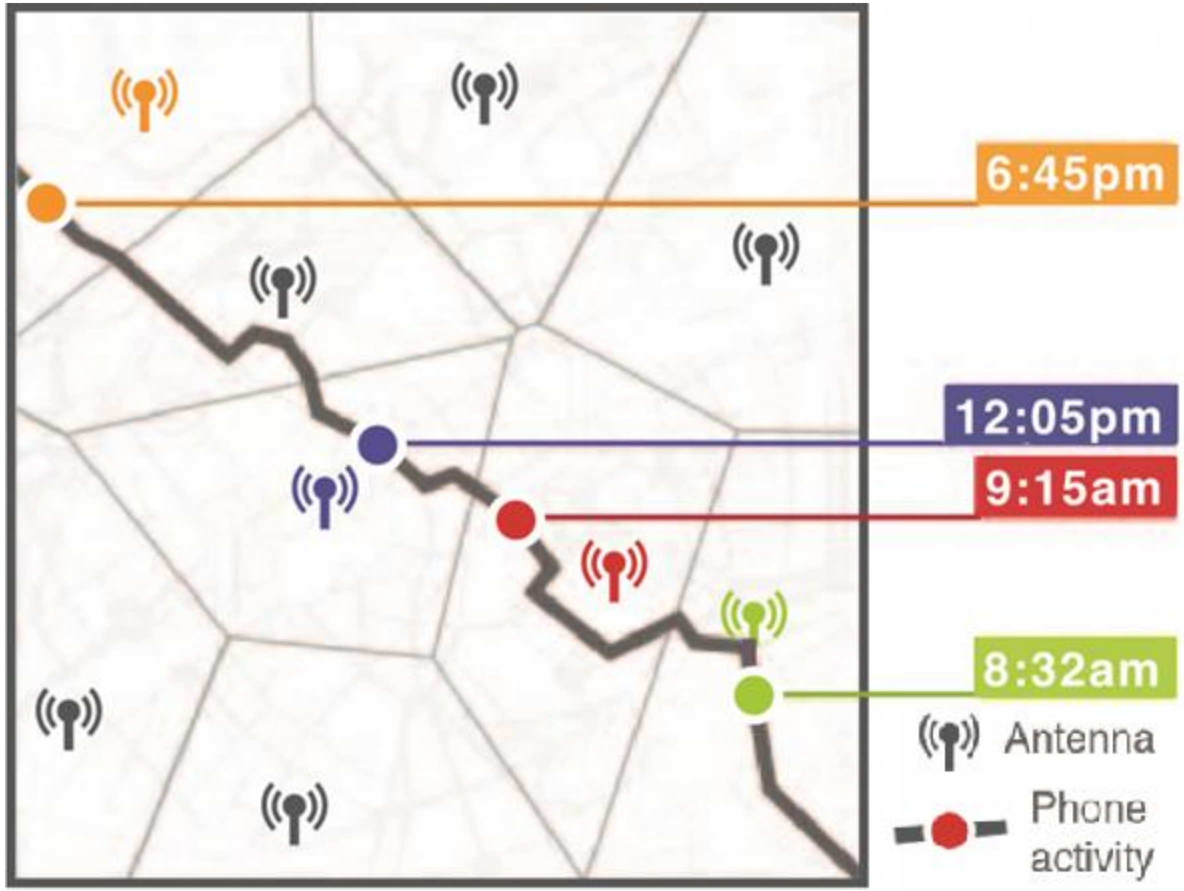
Unique in the crowd

- Four spatio-temporal point is enough to uniquely identify 95% of people
- Electronic Frontier Foundation published about inferring potentially sensitive information out of mobility trace
- 33% of App Store applications access geo-location
- Medical DB combined with voters list to extract health record of governor of Massachusetts

by Yves-Alexandre de Montjoye, César A. Hidalgo, Michel Verleysen, Vincent D. Blondel @ MIT, Harvard, … , 2013

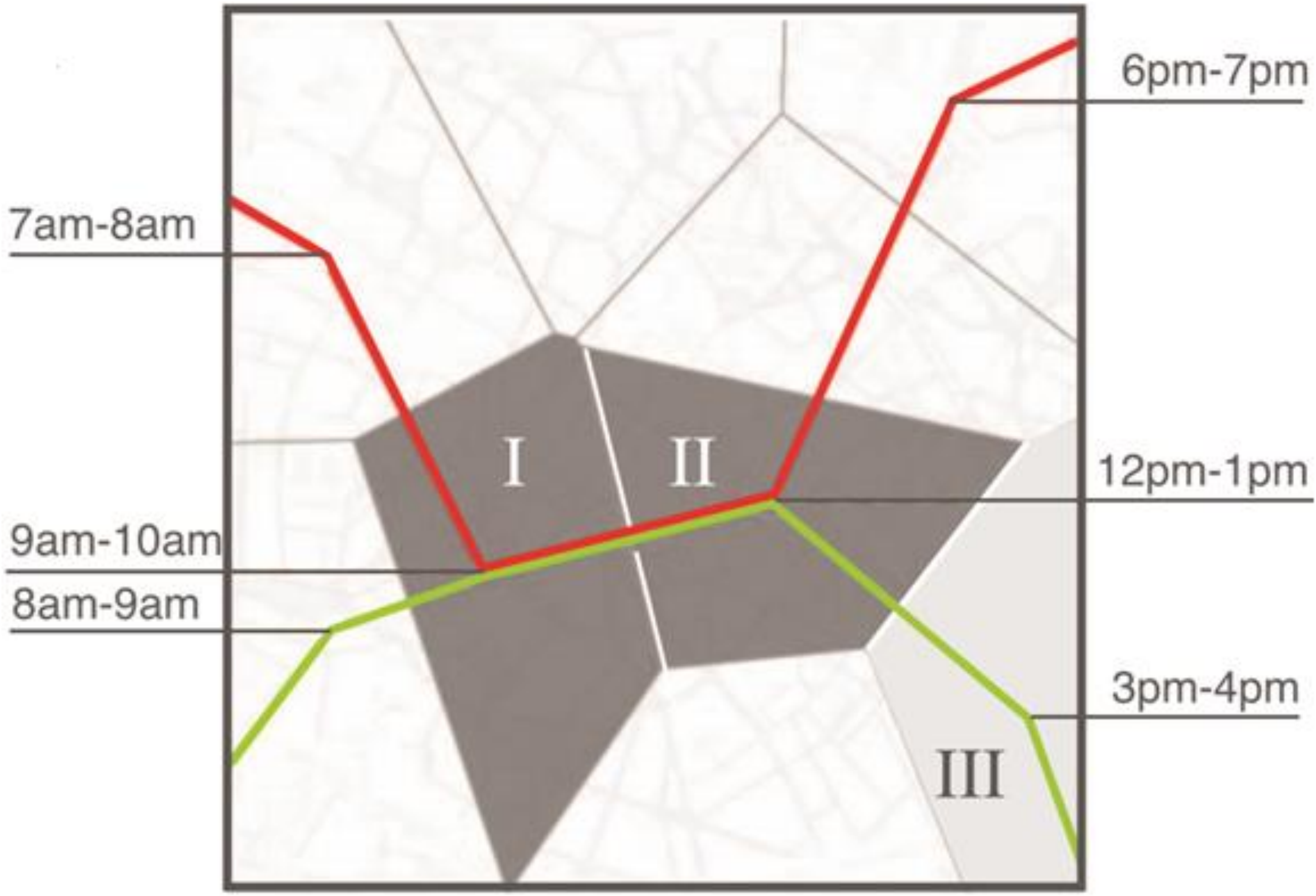
Anonymized dataset

- No personal info (name, address, phone number, email)
- Rough spatial records based on GSM cells and hourly temporal resolution
- 1.5M individuals, all subscribers of a nameless European operator
- 1.5 years of data
- Not continuous

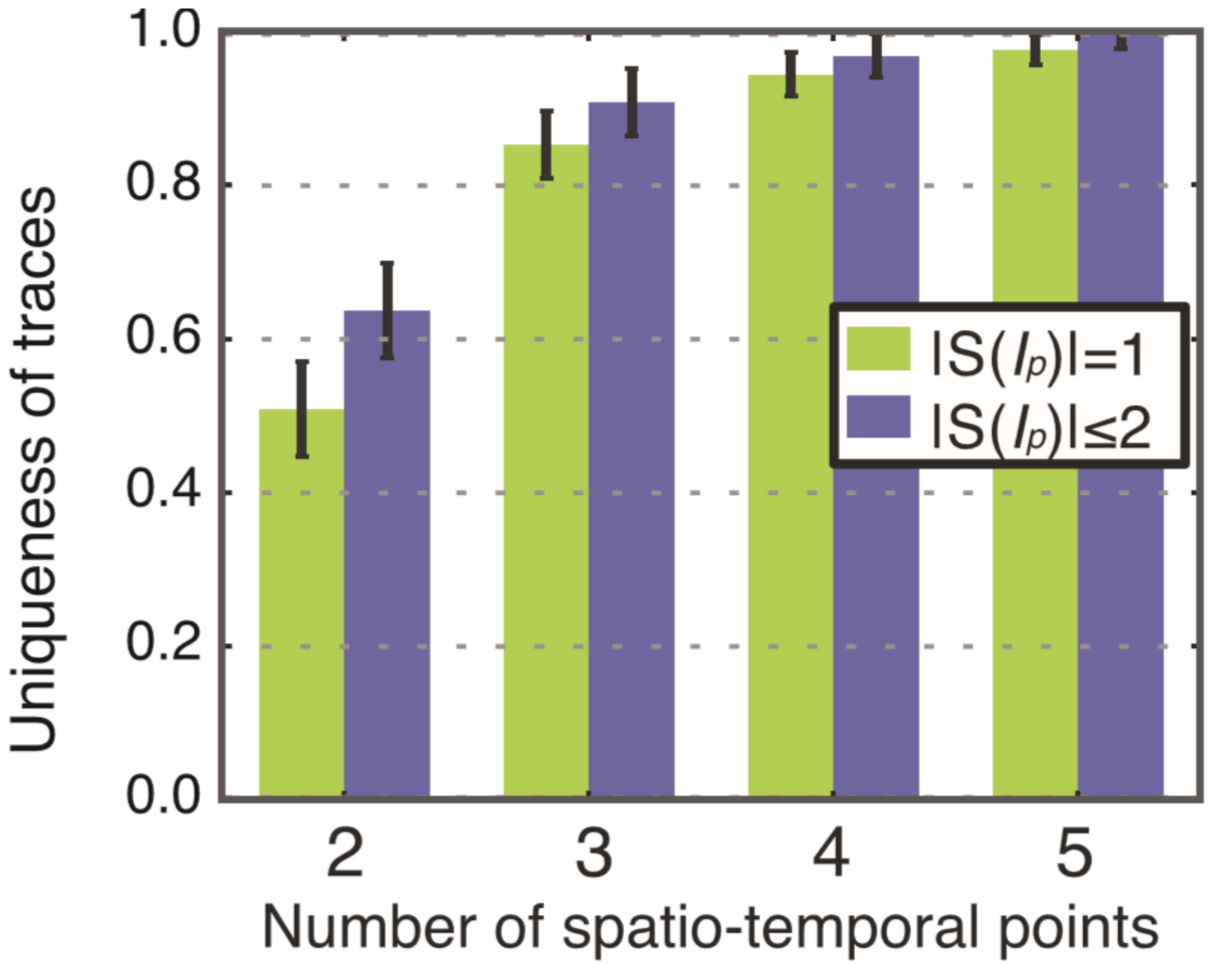


Uniqueness, ε

- D a simply anonymized dataset
- I_p a p size set of spatio-temporal points
- $S(I_p)$ subset of D matching I_p
- ε uniqueness, the probability of $|S(Ip)| = 1$ by choosing the points of I_p uniformly distributed among the range of spatio-temporal points of D



Uniqueness, ε

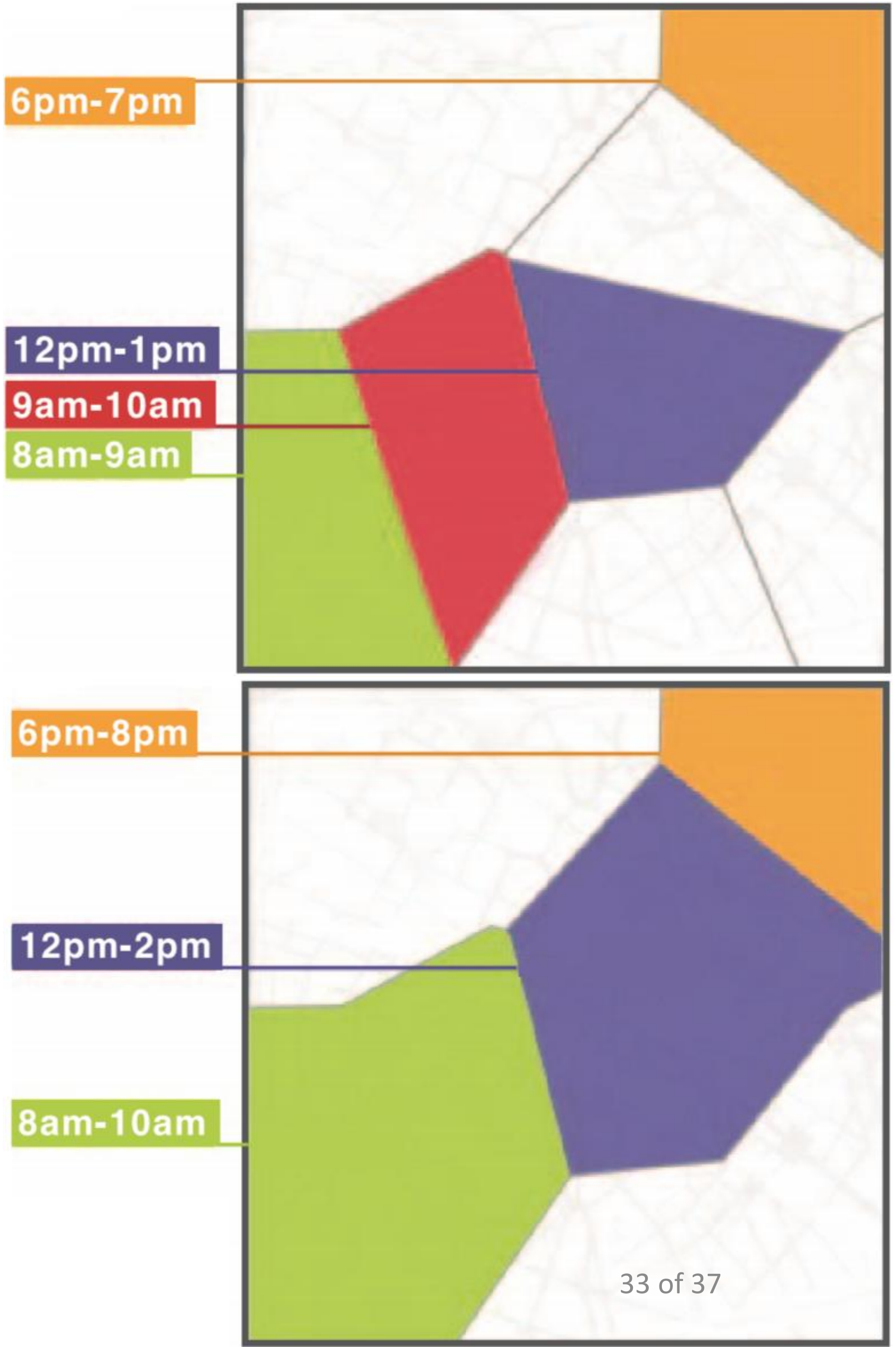


Scaling properties

- Decrease spatial and temporal resolution

  Merge cells and increase the observation time window

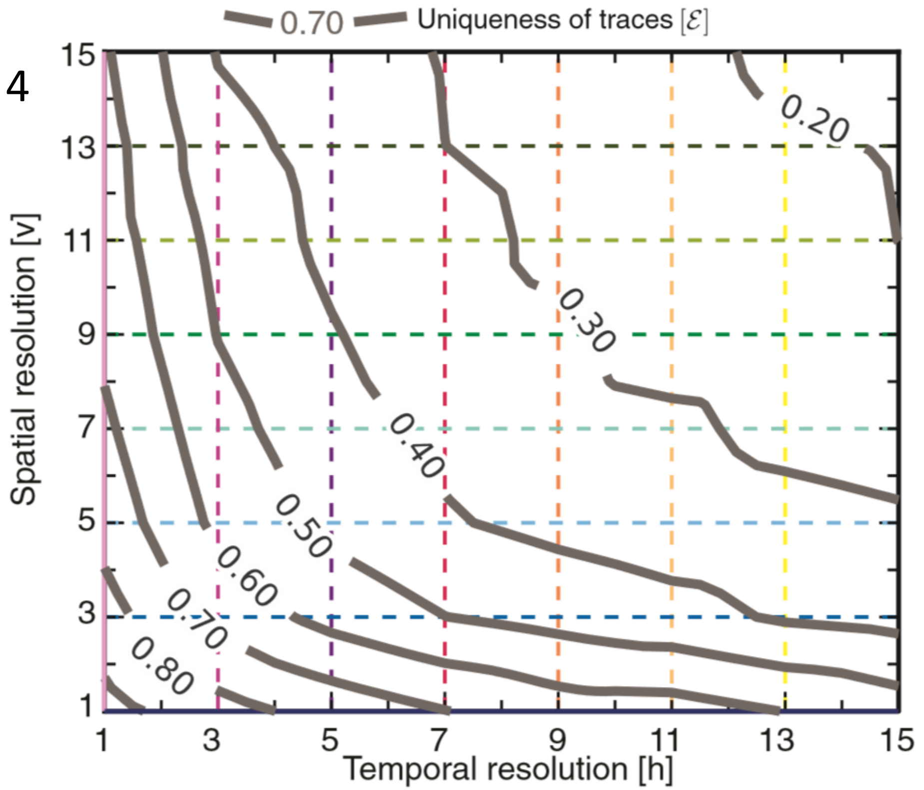
- h – proportion of time window to original 1 hour
- v – number of merged cells



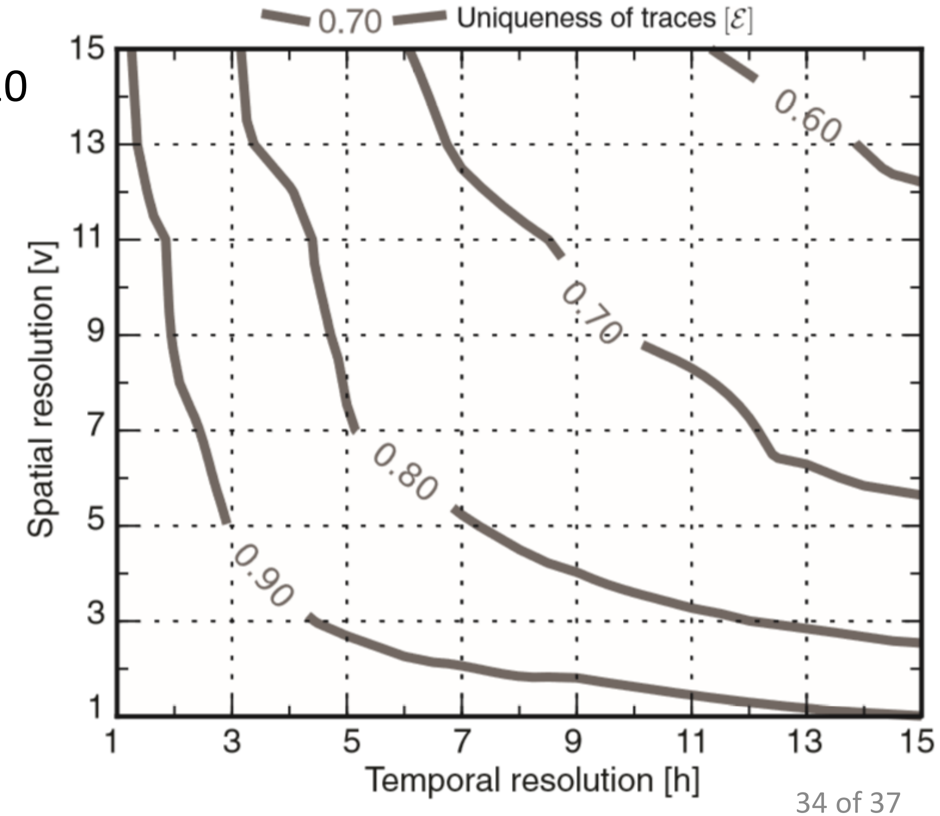
Uniqueness of traces

- Easier to attack if dataset is coarse on one dimension but fine on the other than mid-grained on both dimensions

p = 4

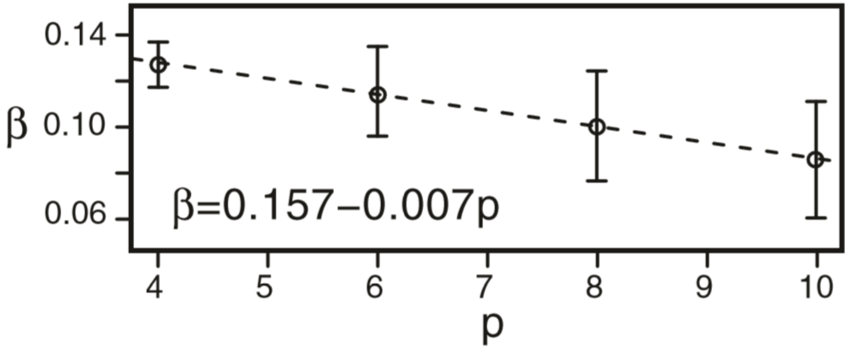


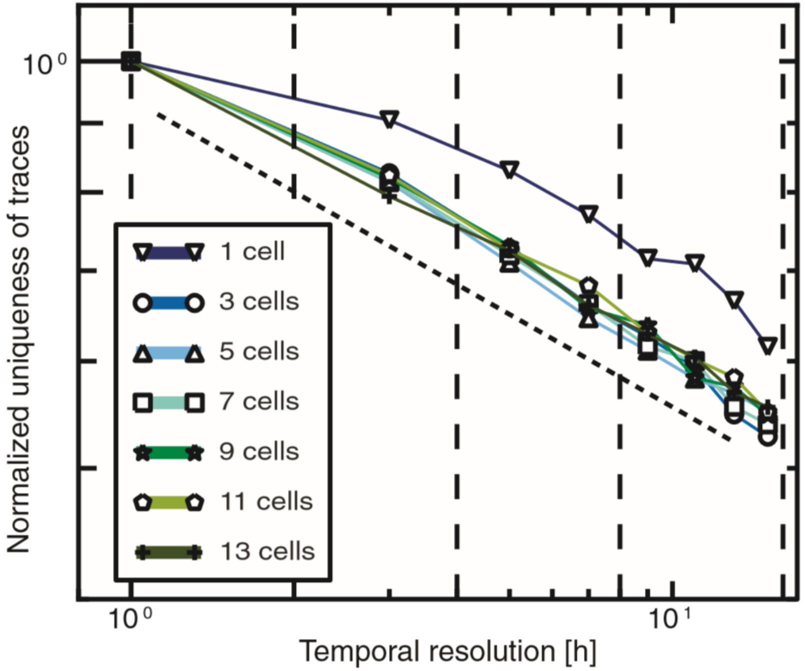
p = 10

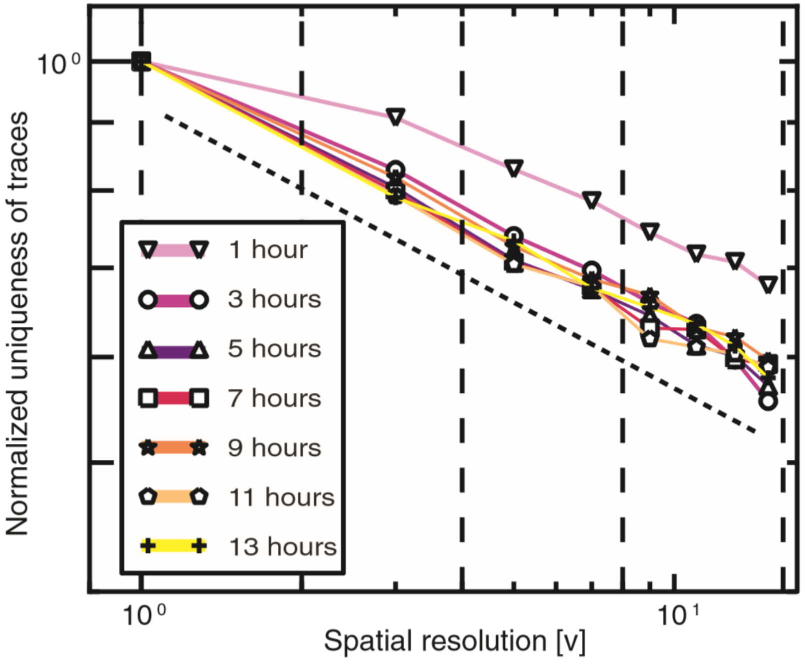


Uniqueness as a function of resolution

- Power function fits data
- $\varepsilon = \alpha - (vh)^{\beta}$
- β is linear function of number of points
  - If the resolution halves the uniqueness decreases by constant factor $2^{-\beta}$
  - Privacy is increasingly hard to gain by lowering the resolution







Lessons

- Privacy is increasingly hard to achieve
- Re-identification is possible even in sparse, large scale, coarse dataset
- Knowing the bounds of individual privacy is important for future policies and information technologies

Questions

UNIVERSITY OF
EASTERN FINLAND

Thank you for the attention!

Adam Ludvig
258801

hunludvig@gmail.com