Telematica
Instituut

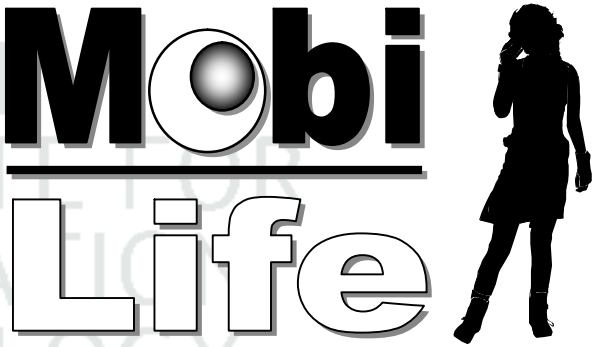
HELSINKI
INSTITUTE FOR
INFORMATION
TECHNOLOGY

Identifying Significant Locations

Petteri Nurmi[1], Johan Koolwaaij[2]

[1)] Helsinki Institute for Information Technology HIIT

[2)] Telematica Instituut (TELIN)



1. Introduction and motivation
2. Algorithms for location clustering
3. Experiments

What are significant locations?

1. Raw location data is usually meaningless to a user è
   - Data is clustered with the goal of finding clusters that are somehow **meaningful** to a user
   - Meaningful clusters are called *places* or *significant locations*
     - Examples: HOME, OFFICE, LIBRARY, ...
   - Usually a two phase process:
     1. Apply spatial clustering
     2. Prune out meaningless clusters using temporal information
2. Existing work differs regarding the source of data:
   - Continuous GPS streams
   - GSM cell identifiers
   - WiFi access point information

More on data sources

- Practical issues with different sources:
  - GPS coordinates
    - Continuous gathering infeasible
    - High communication and processing costs
  - WiFi
    - Locations of base stations must be known beforehand
  - GSM cell identifiers
    - Can be done on device, but poor location accuracy
- In our work, combine GPS and GSM data to achieve:
  - Better accuracy than with mere GSM identifiers
  - Resource costs higher than with GSM identifiers, but remain feasible to use in practice
  - Setting that is suitable for mobile devices

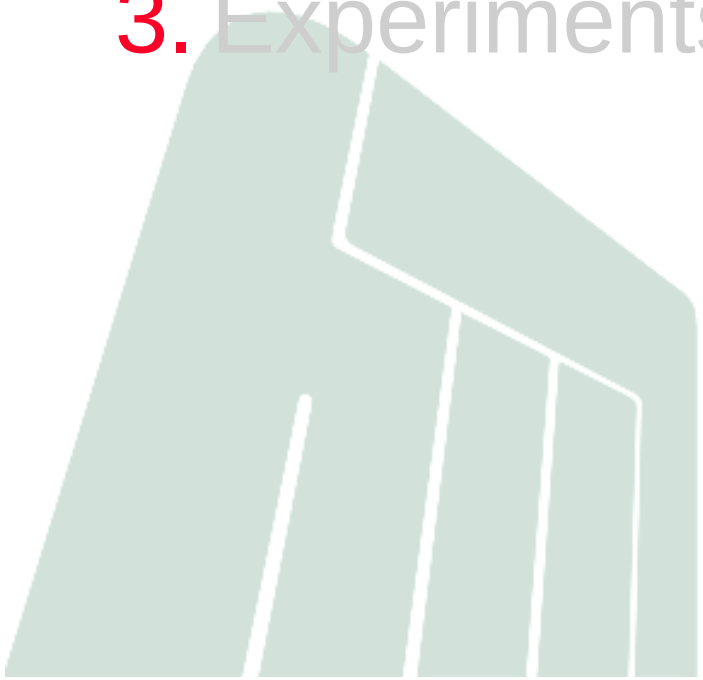
Why are some locations significant?

1. The motivation comes from (social) identity theory:
   - A person acts in several societal categories (roles)
   - To each role, the person associates behavioral expectations

   è When acting in a specific role, the observed behavior is influenced by the behavioral expectations

2. The link to significant locations:
   - Many locations, such as work and home are boundaries between different roles
   - Thus we can expect the location to influence the behavior of the user
   - However, not the only explanatory factor, but one, potentially useful source of information

1. Introduction and motivation
2. Algorithms for location clustering
3. Experiments



Goals for algorithms

1. Accuracy
   - Are our clusters correct? How much additional area they cover?
2. Time to reasonable results
   - How quickly can an application access clusters?
3. Scalability
   - Months of data, how to handle data overflow?
4. Adaptability
   - How well can the clusters be adapted over time? How well the clustering works in online or batch mode?
5. Meaningfulness of the clusters?
   - Do the users think the clusters that are captured are somehow meaningful?
6. Cluttering
   - How to avoid the effect that the diameter of the clusters grows over time?

Graph-based Algorithms

1. Heuristic graph clustering,
   - Builds a weighted graph where edge weights are distances between centroids of GPS measurements within individual GSM cells
   - Maps the distances into probabilities and uses (exponential) weight decay to merge GSM cells
2. Spectral clustering
   - Finds a minimum cut in the weighted graph using a spectral representation and a suitable objective function
   - In our case, form the spectral representation, calculate eigenvalues and use K-means to cluster top eigenvectors

Algorithms

1. Duration-based grid clustering
   - Accuracy of a location measurement is used to distribute the duration user spends in a single location (no transitions) to a grid
   - Clusters formed by merging grid points where the user spends ”enough” time

2. Frequent transitions
   - Builds a matrix of transitions between cells
   - Clusters two cells (A,B) when the transitions A à B and B à A both are significant

1. Introduction and motivation
2. Algorithms for location clustering
3. Experiments

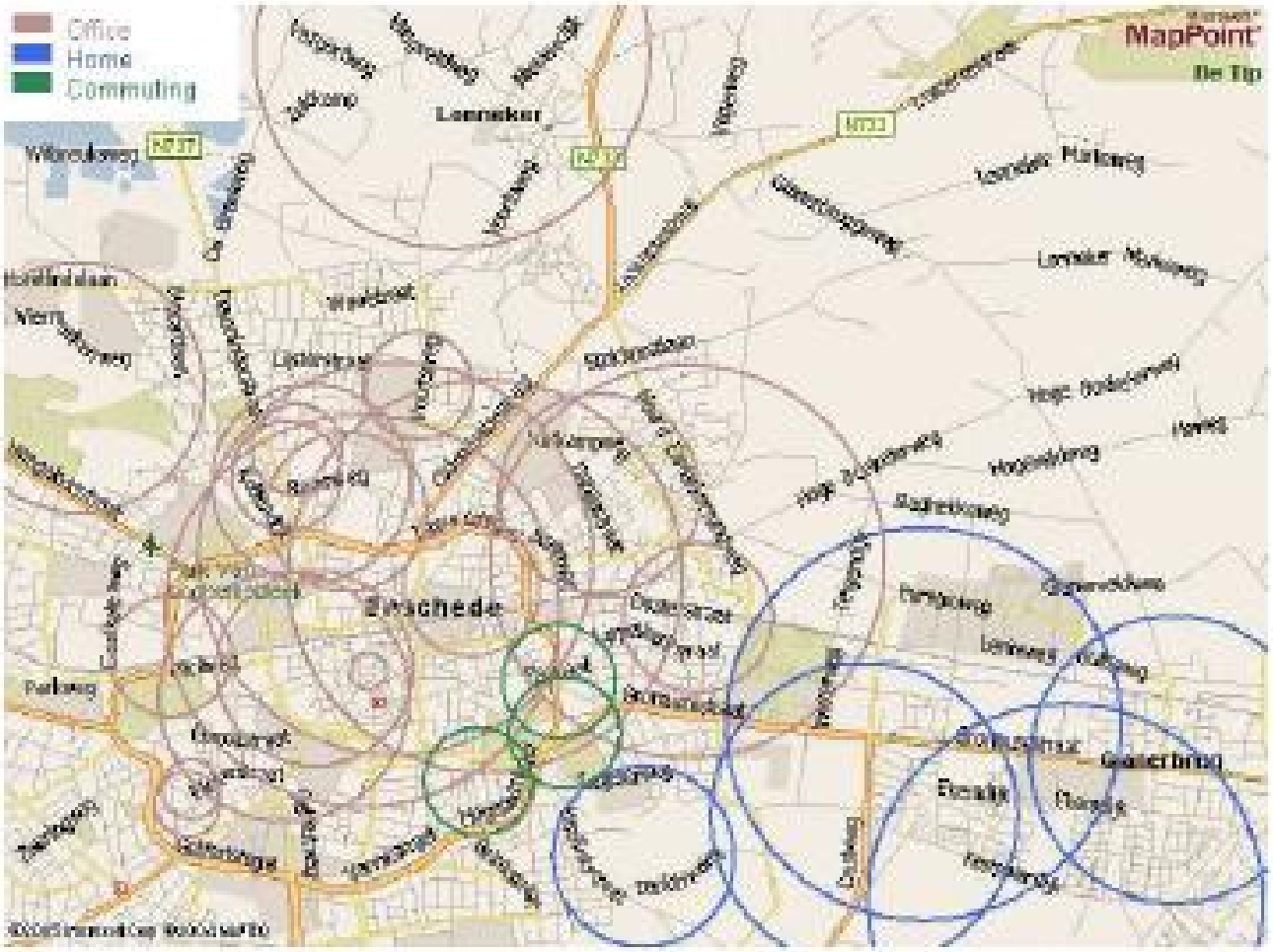
The test data

1. Data gathered by multiple users throughout Europe[2)] using the Context Watcher application http://www.lab.telin.nl/~koolwaaij/showcase/crf/cw.html (Google for *context watcher*, 1st hit)
2. In the picture you can see some example places where we have gathered data

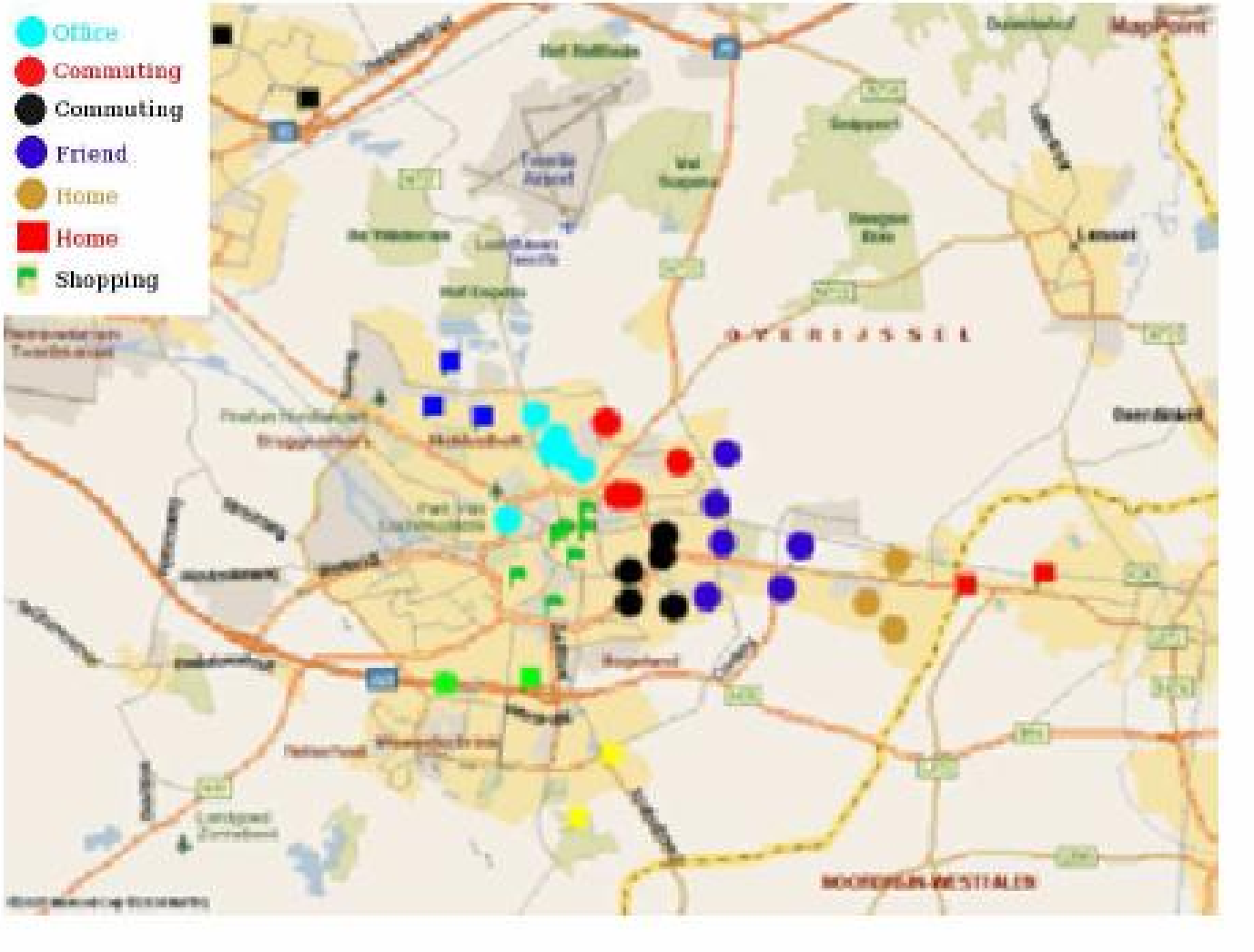
è We have a rich and heterogeneous collection of location data as our test set

2) Also small amounts of data outside Europe: USA, Russia, Canada, Ukraine

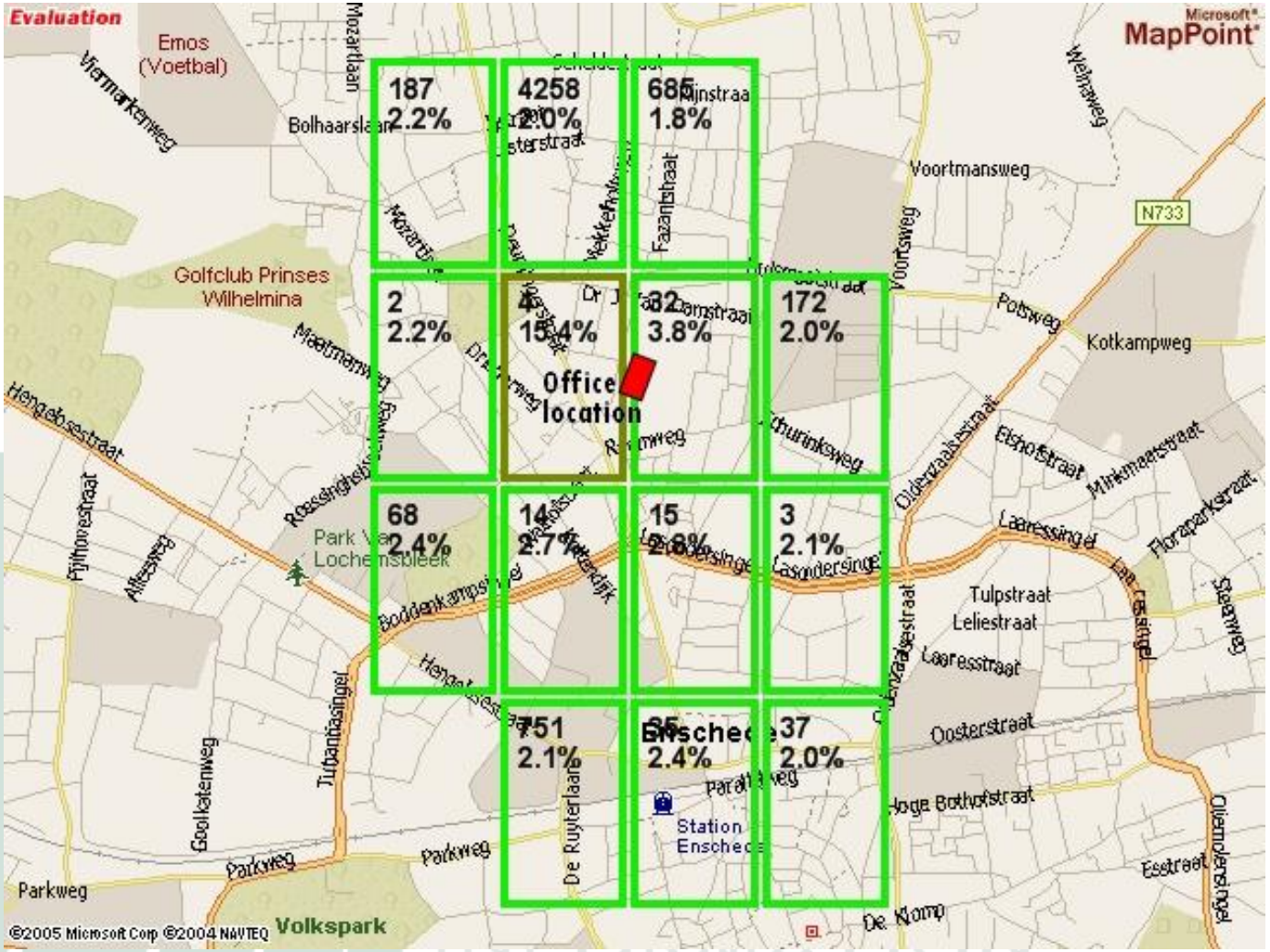
Results: Heuristic graph clustering



Results: Spectral clustering



Example results: Grid clustering



Results: Transition-based clustering

Comparison of the algorithms

	Heuristic Graph clustering	Spectral clustering	Grid clustering	Transition-based clustering
Accuracy	Ok	Good	Ok	Good
Time to results	Good	Ok	Good	Ok
Scalability	Ok	Bad	Good	Ok
Adaptability	Good	Bad	Good	Ok
Cluttering	Bad	Good	Ok	Ok

Summary

- Why are some locations meaningful?
  - Identity theory: a person acts in different roles, to which (s)he associates behavioral expectations
  - Thus, if the location is a boundary between two roles, we can expect to observe different kind of behavior
- We introduced a novel domain for (mobile) location data:
  - GSM cell identifiers + GPS coordinates
  - Presented four algorithms for the problem
  - Compared the algorithms using real data
- In addition, we discussed what are desirable properties for a location clustering algorithm