UNIVERSITY OF
EASTERN FINLAND

21.2.2019

Similarity of mobile users based on sparse location history

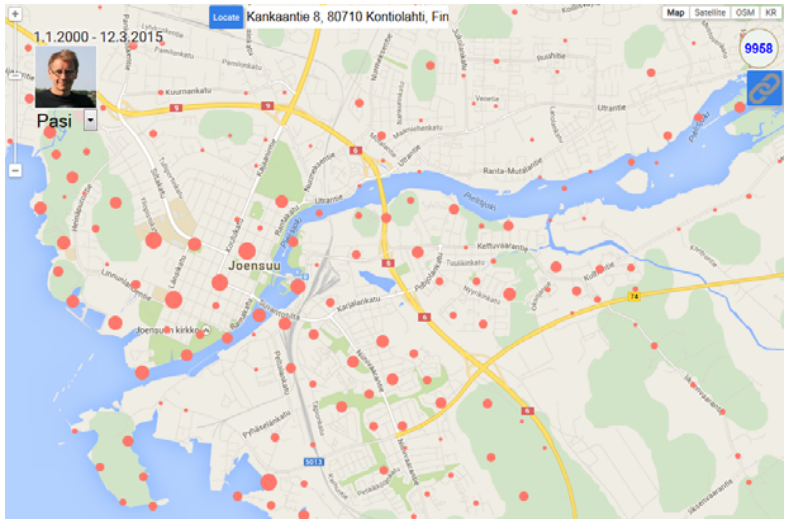
Pasi Fränti
Radu Mariescu-Istodor
Karol Waga

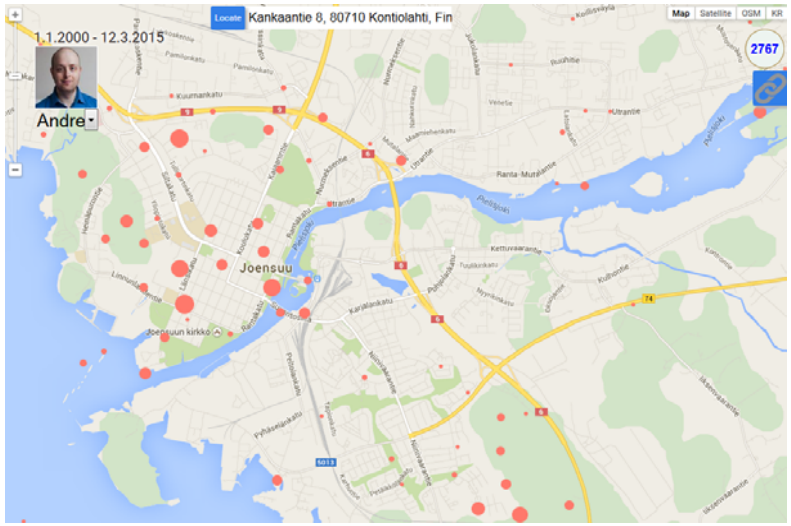
P. Fränti, R. Mariescu-Istodor and K. Waga
"Similarity of mobile users based on sparse location history"
Int. Conf. Artificial Intelligence and Soft Computing (*ICAISC*)
Zakopane, Poland, 593-603, June 2018

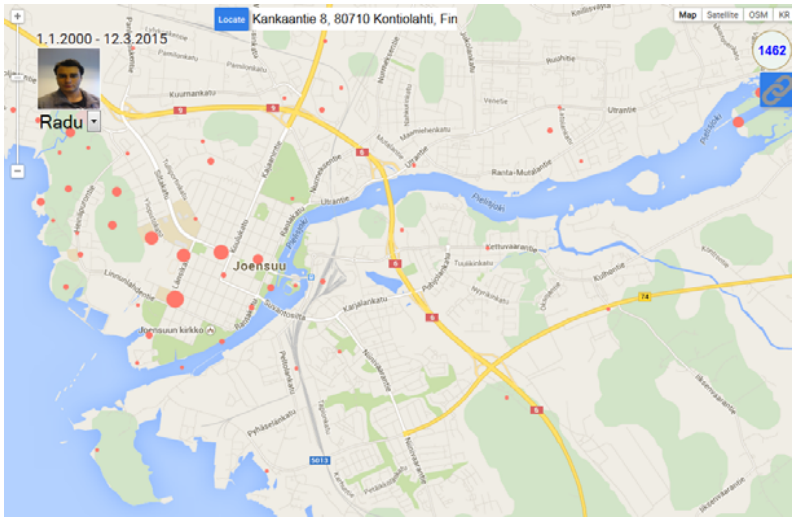
Introduction

Motivation

- Cold start users in Recommender systems
- Activity in the same area
- Opinions of local experts more valuable [Bao'2012]
- Identify the same user

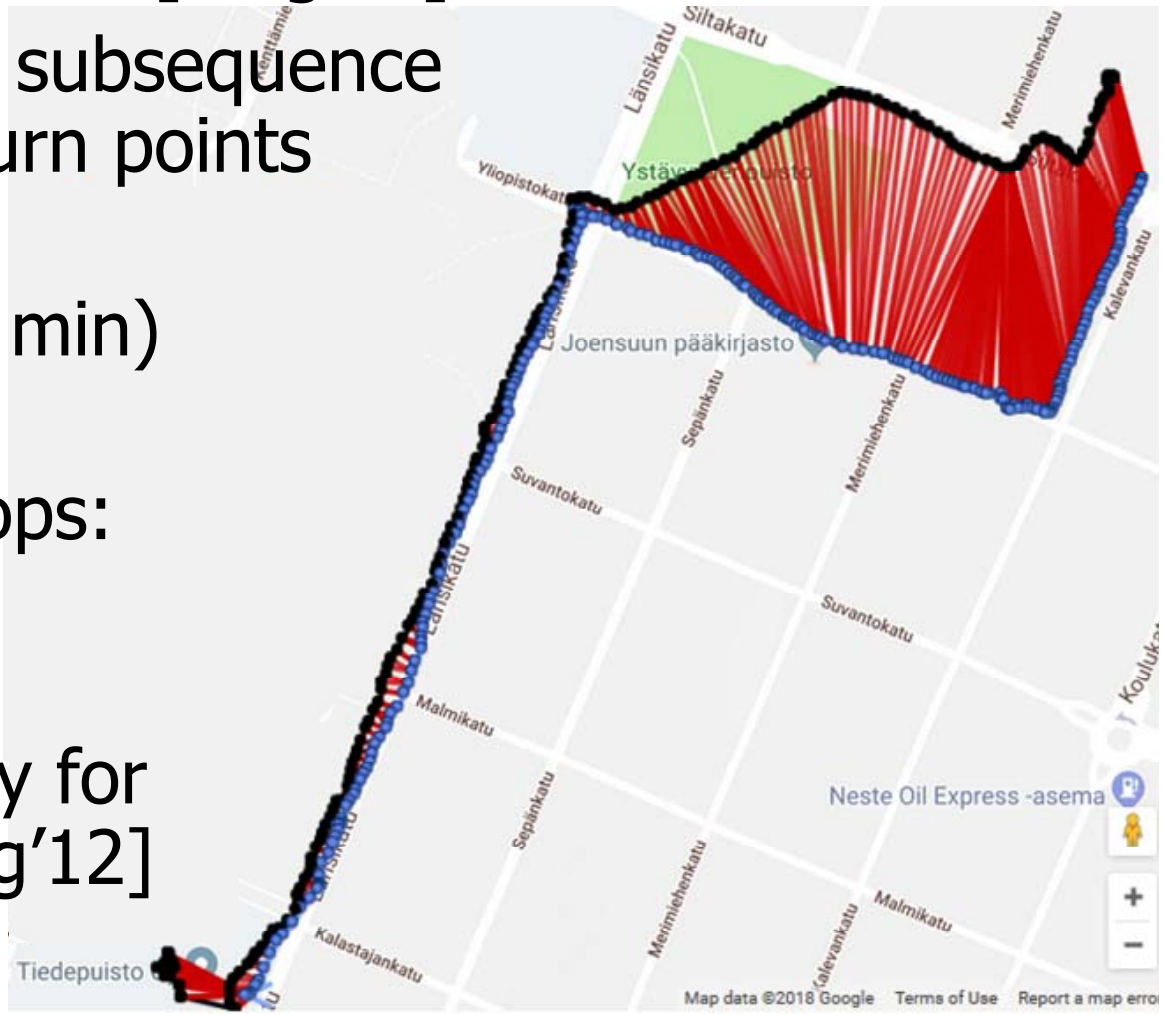






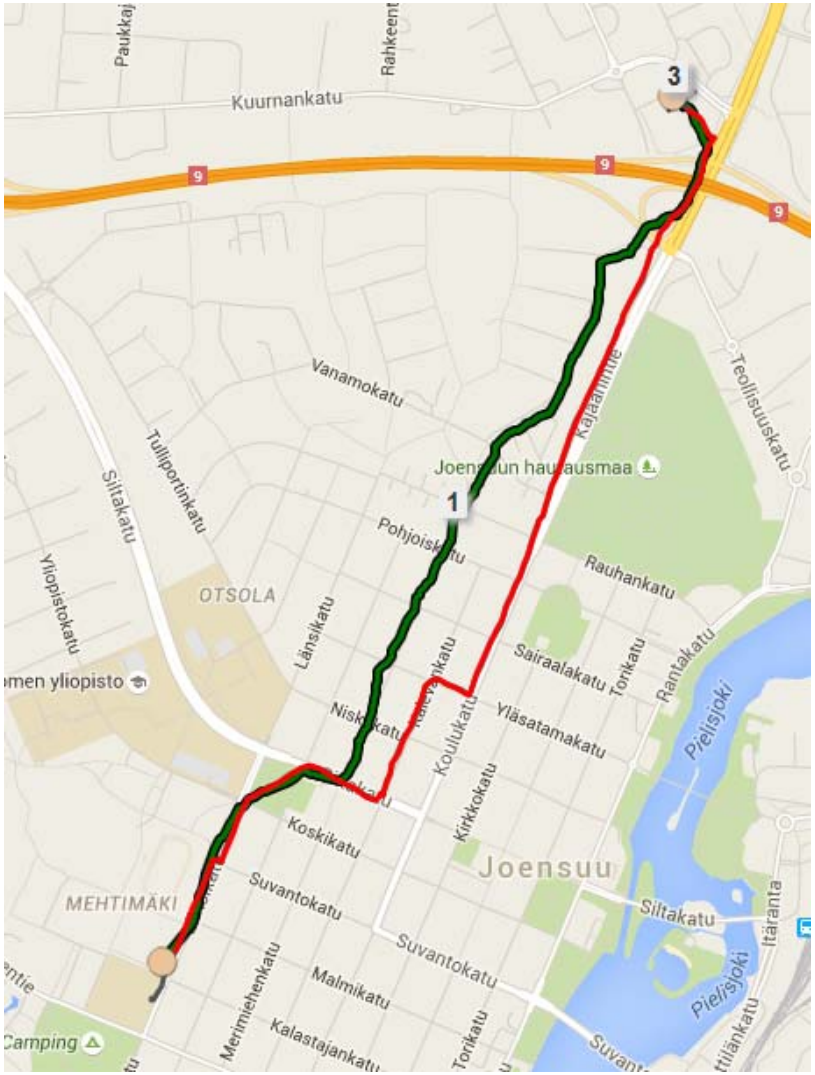
Existing methods

- Complete trajectories [Ying'10]
- Longest common subsequence with speed and turn points [Liu'12]
- Stop points (>30 min) [Zheng'10]
- Most common stops: home and work [Biagioni'13]
- User given priority for parts more [Wang'12]

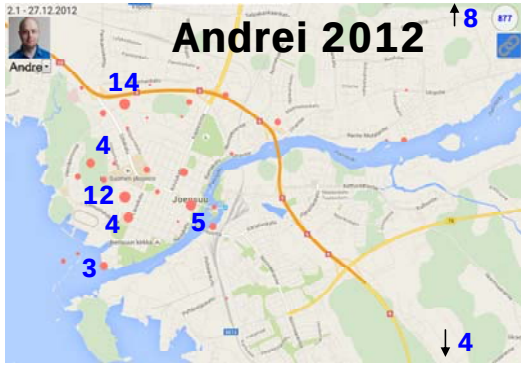


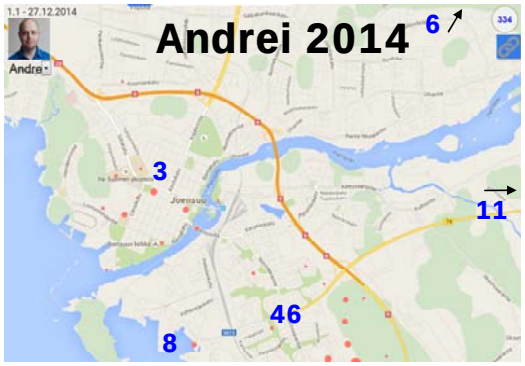
Problems?

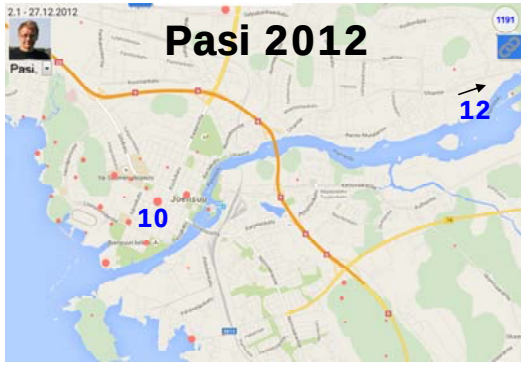
- Full GPS trajectories not always stored
- Areas of activity are often more important than the exact tracks
- Explicit activities such as visits, check-ins and photo taking are widely recorded.

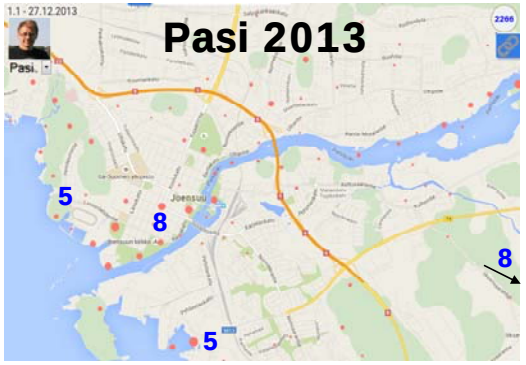


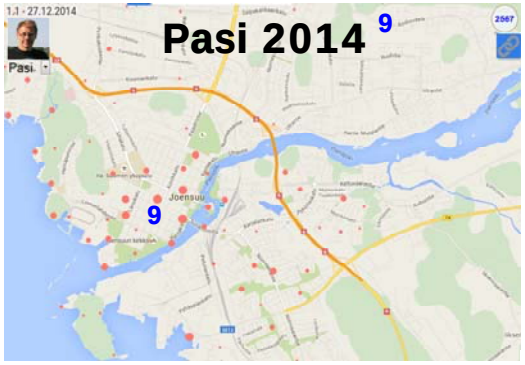
Test case: APR trio

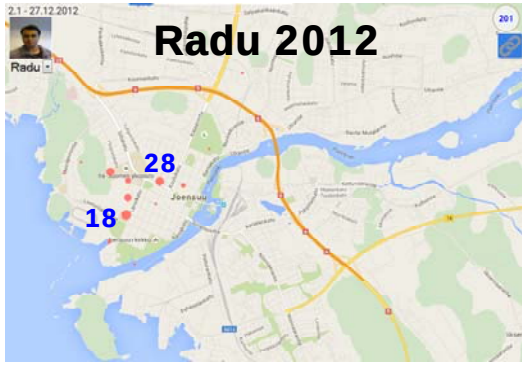


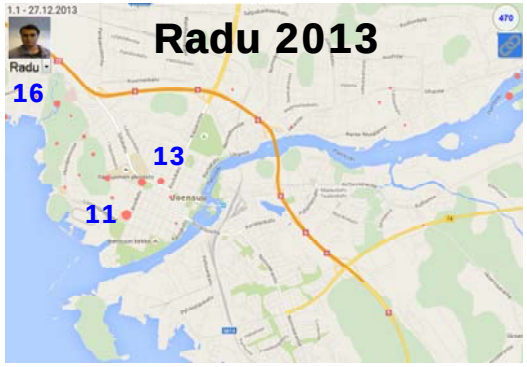


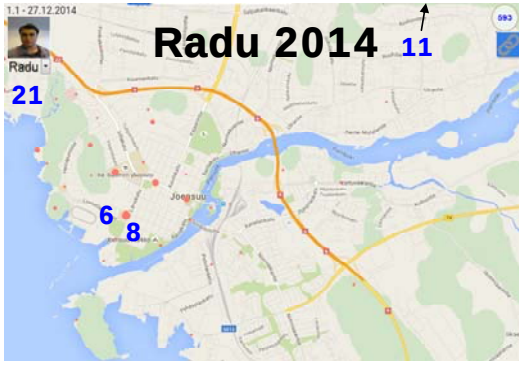




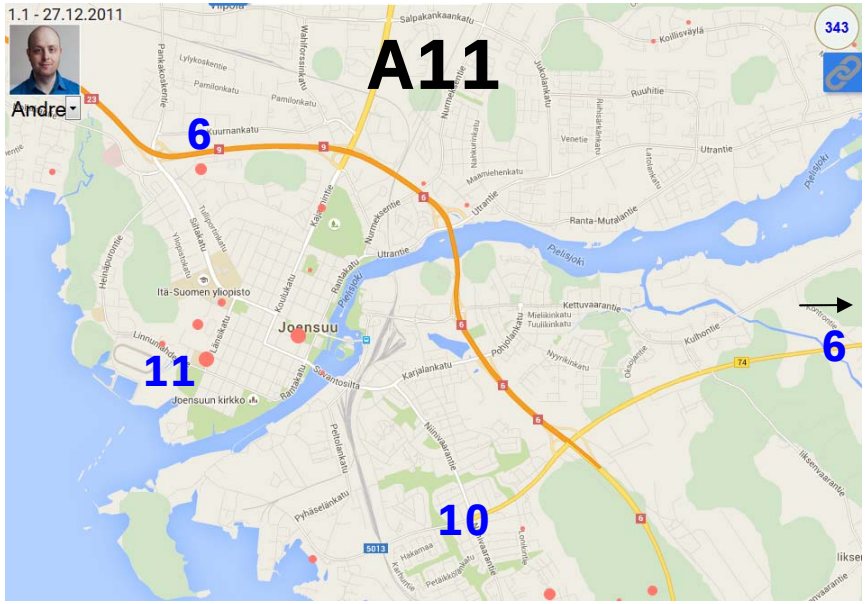


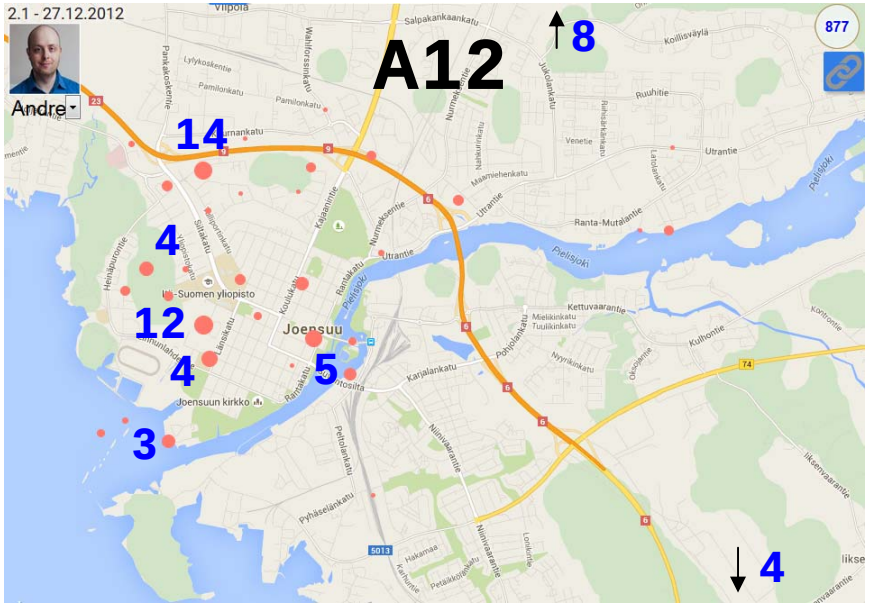


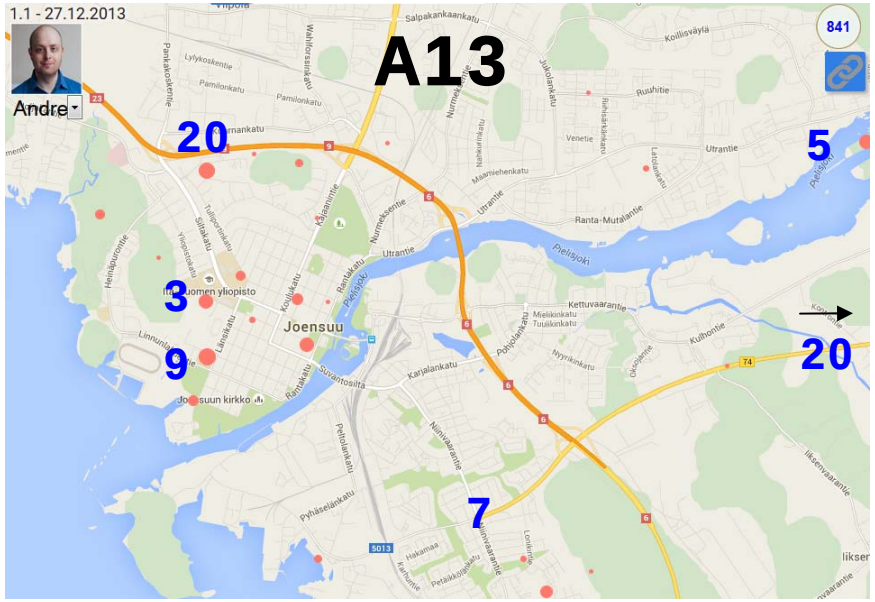


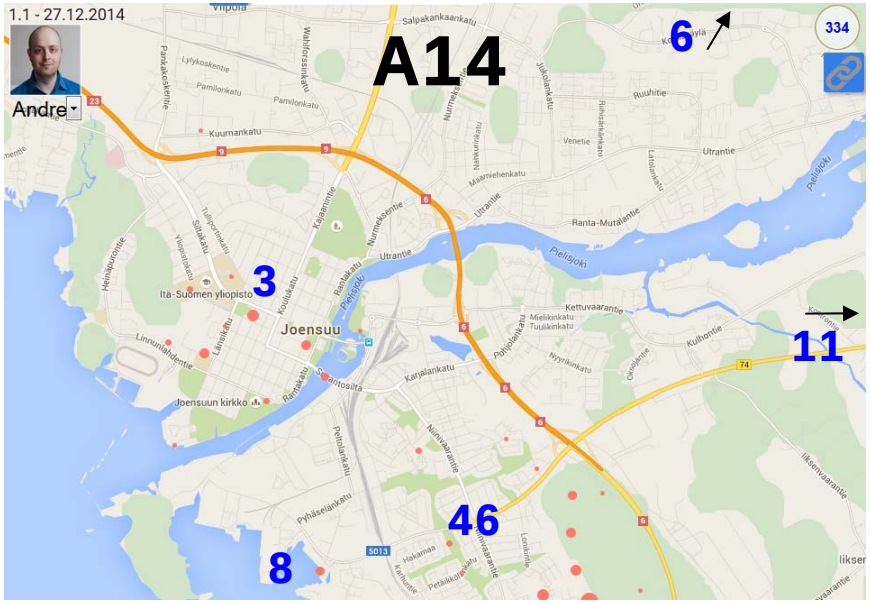


Example of User A

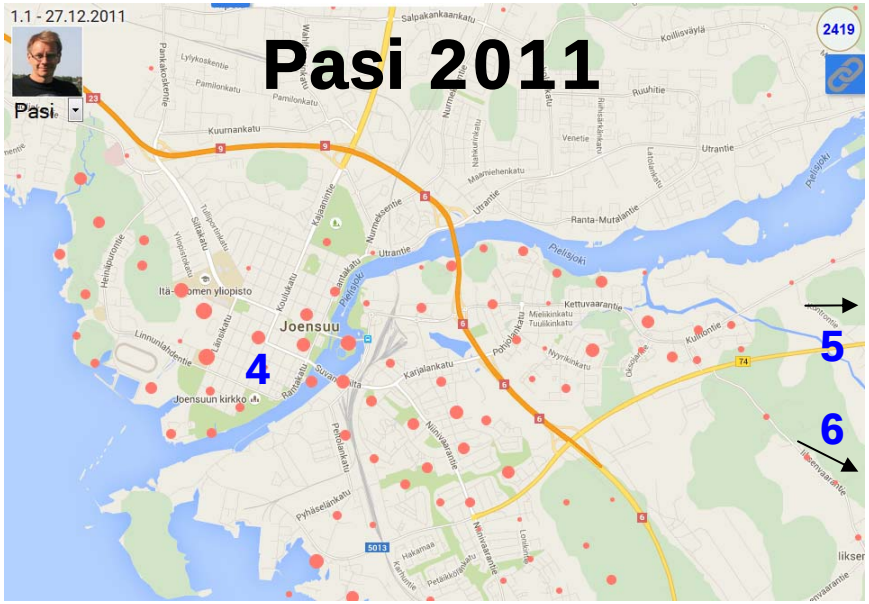


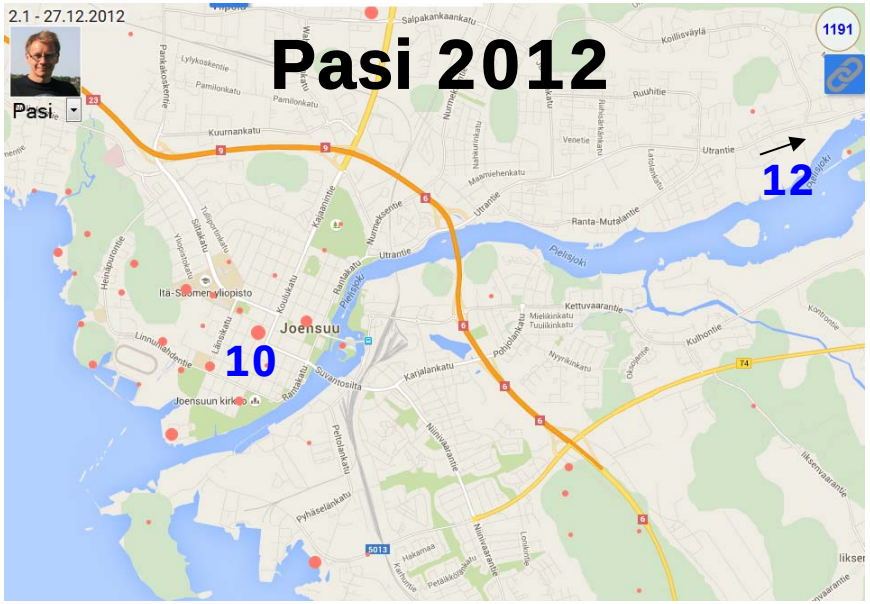


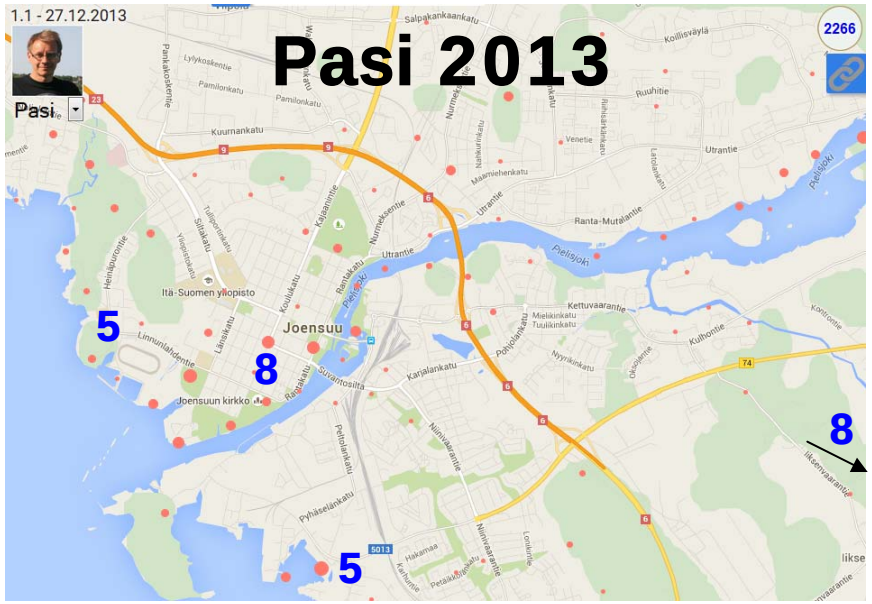


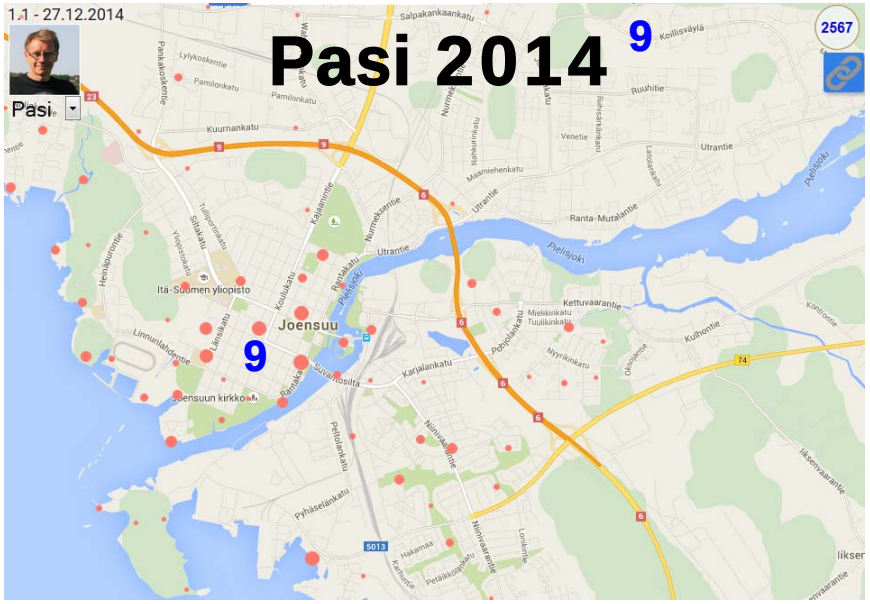


Example of User P

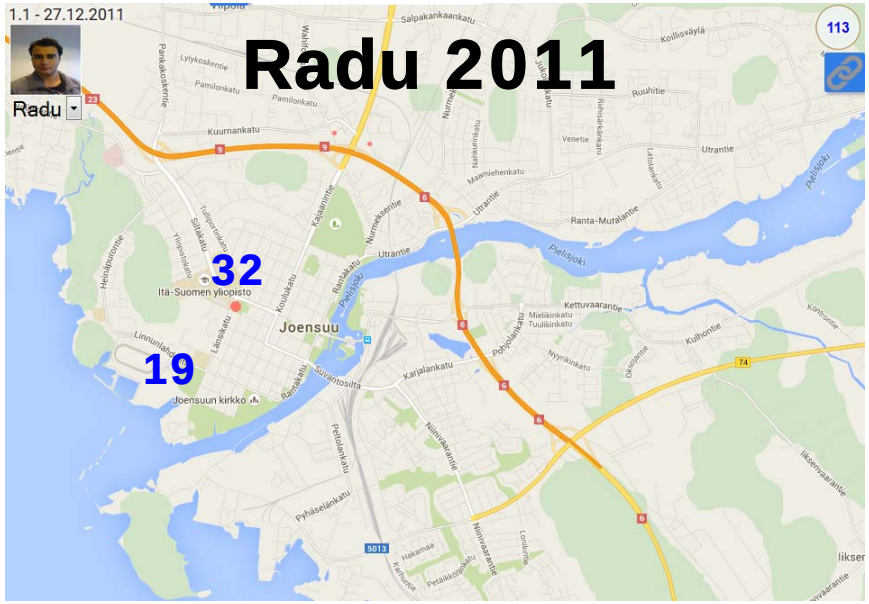


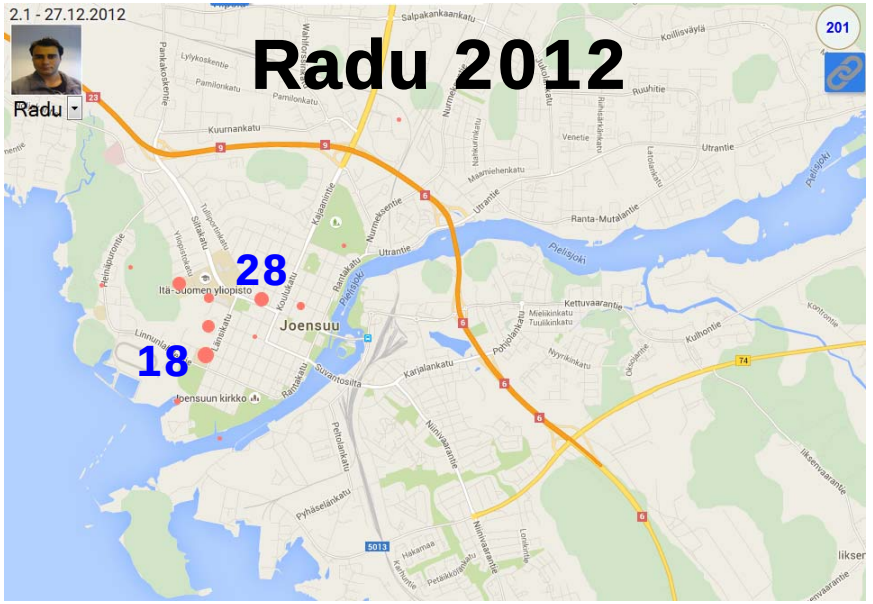


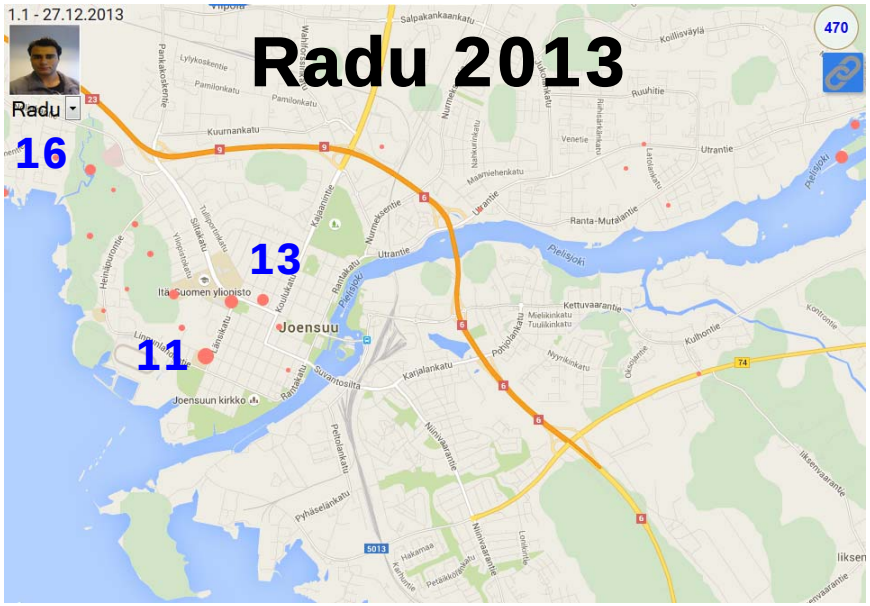


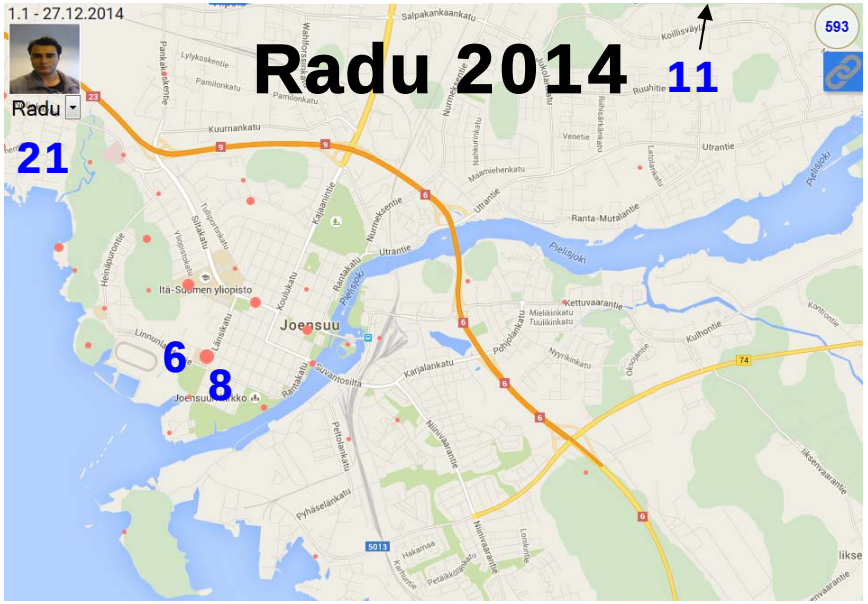


Example of User R









Two approaches

1. Histogram comparison
   - Construct histogram from some common data
   - Map points to the bins
   - Histogram matching of the two counts
2. Clustering comparison
   - Cluster each data separately
   - Map centroids between the data
   - Calculate similarity based on the mapping

Selected approaches

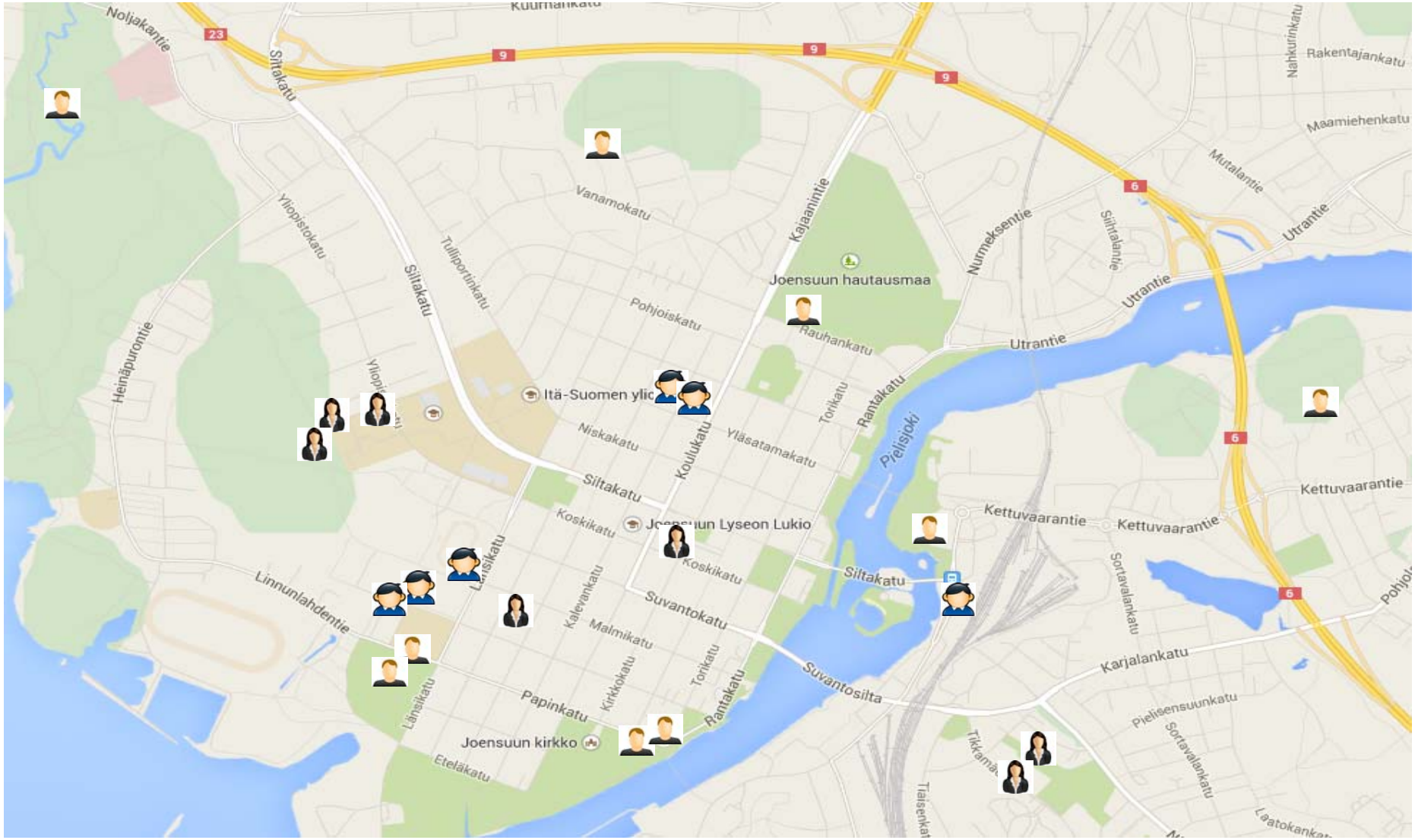
1. Histogram comparison
   - Construct histogram from some common data
   - Map points to the bins
   - Histogram matching of the two counts
2. Clustering comparison
   - Cluster each data separately
   - Map centroids between the data
   - Calculate similarity based on the mapping

Approach 1:

Histograms

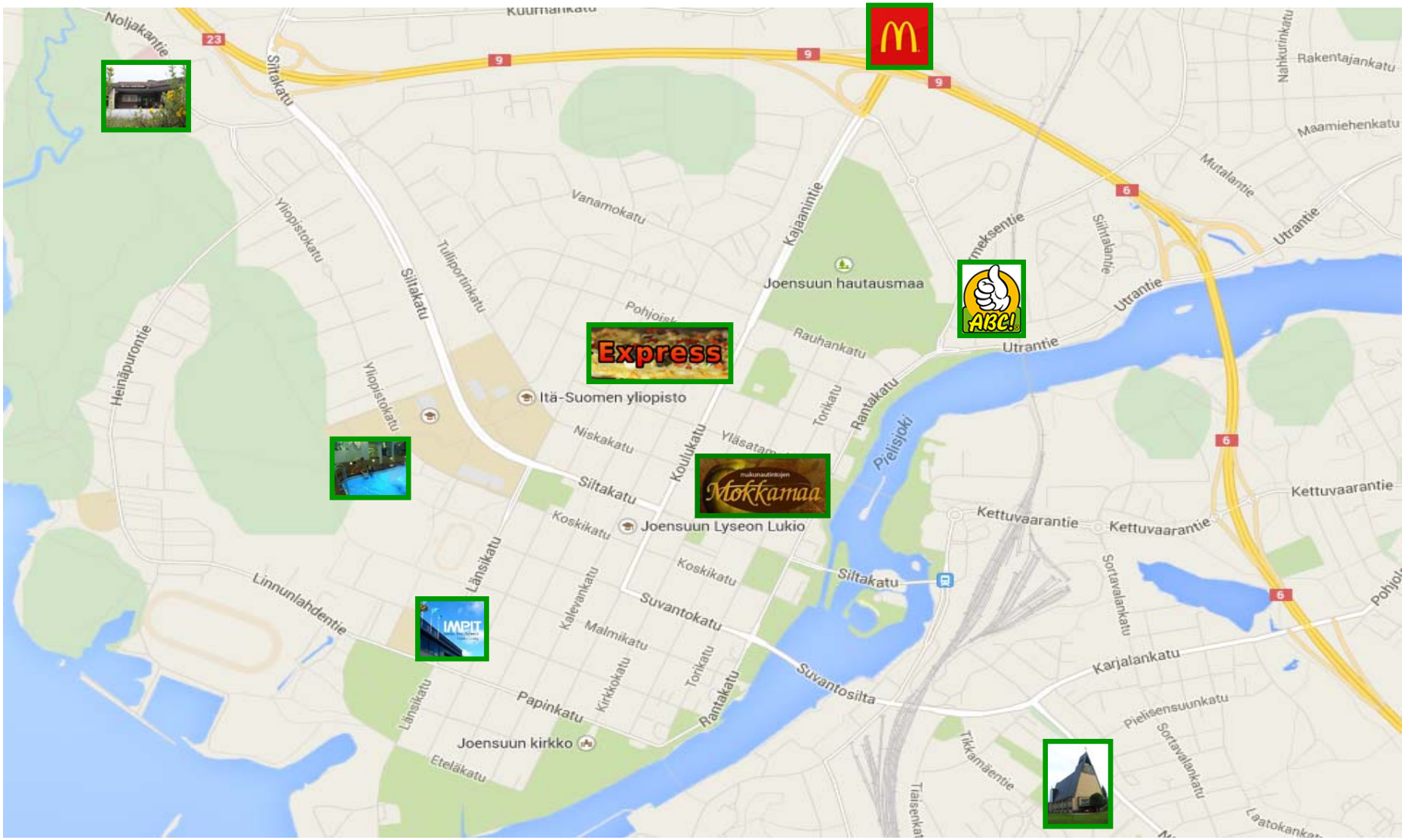
Location history

Activity points of users



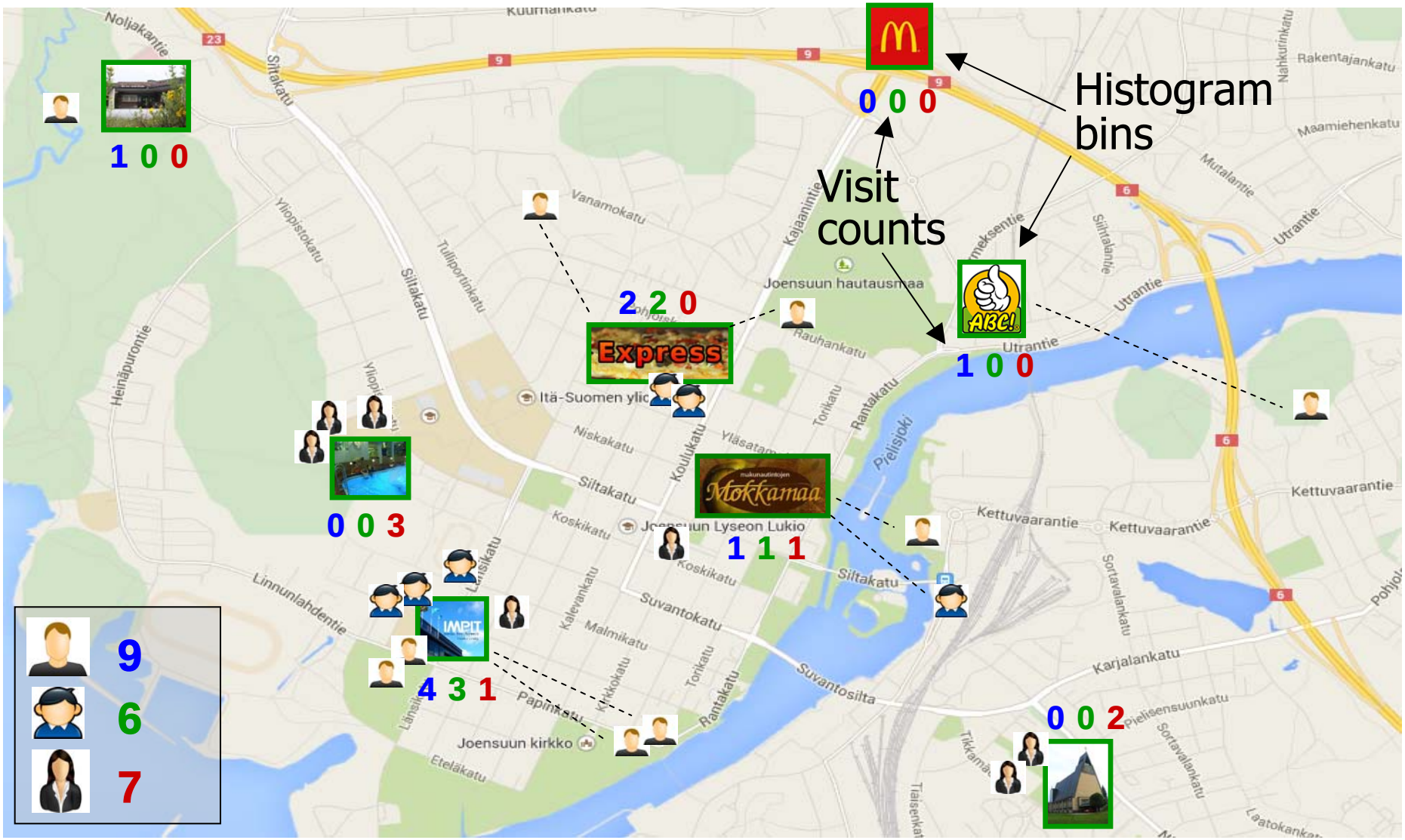
Histogram matching

Create bins from popular locations



Location similarity

Count visit statistics



Histogram of the toy example

	Frequencies				Relative frequencies		
	Pekka	Bartek	Julia		Pekka	Bartek	Julia
	4	3	1	8	0.44	0.50	0.14
	2	2	0	4	0.22	0.33	0.00
	0	0	3	3	0.00	0.00	0.43
	1	1	1	3	0.11	0.17	0.14
	0	0	2	2	0.00	0.00	0.28
	1	0	0	1	0.11	0.00	0.00
	1	0	0	1	0.11	0.00	0.00
	0	0	0	0	0.00	0.00	0.00
	9	**6**	**7**				

Distance measures

$$L_1 = \sum_i |p_i - q_i|$$

$$L_2 = \sum_i (p_i - q_i)^2$$

$$L_\infty = max|p_i - q_i| \forall i$$

$$d_{ChiSq} = \sum_i \frac{(p_i - q_i)^2}{p_i + q_i}$$

$$d_{KLD} = \sum_i \left(p_i \cdot \log \frac{p_i}{q_i} + q_i \cdot \log \frac{q_i}{p_i} \right)$$

$$S_{BC} = \sum_i \sqrt{p_i \cdot q_i}$$

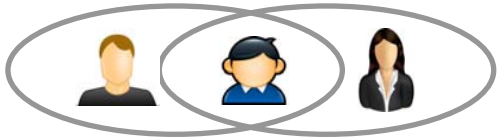
Bhattacharyya distance

$$D = -\ln \sum \sqrt{p_i \cdot q_i}$$

	Frequencies				Relative frequencies			Pekka vs Bartek	Bartek vs Julia
	Pekka	Bartek	Julia		Pekka	Bartek	Julia		
	4	3	1	8	0.44	0.50	0.14	**0.47**	**0.26**
	2	2	0	4	0.22	0.33	0.00	**0.27**	0.00
	0	0	3	3	0.00	0.00	0.43	0.00	0.00
	1	1	1	3	0.11	0.17	0.14	**0.14**	**0.15**
	0	0	2	2	0.00	0.00	0.28	0.00	0.00
	1	0	0	1	0.11	0.00	0.00	0.00	0.00
	1	0	0	1	0.11	0.00	0.00	0.00	0.00
	0	0	0	0	0.00	0.00	0.00	0.00	0.00
	9	**6**	**7**					Σ = 0.88 -ln = 0.13	0.41 0.89

L_1 distance

$$L_1 = \sum_i |p_i - q_i|$$



	4	3	1	8	0.44	0.50	0.14	0.06	**0.36**
	2	2	0	4	0.22	0.33	0.00	**0.11**	**0.33**
	0	0	3	3	0.00	0.00	0.43	0.00	**0.43**
	1	1	1	3	0.11	0.17	0.14	0.06	0.03
	0	0	2	2	0.00	0.00	0.28	0.00	**0.28**
	1	0	0	1	0.11	0.00	0.00	**0.11**	0.00
	1	0	0	1	0.11	0.00	0.00	**0.11**	0.00
	0	0	0	0	0.00	0.00	0.00	0.00	0.00
	9	**6**	**7**					Σ = 0.42	1.43

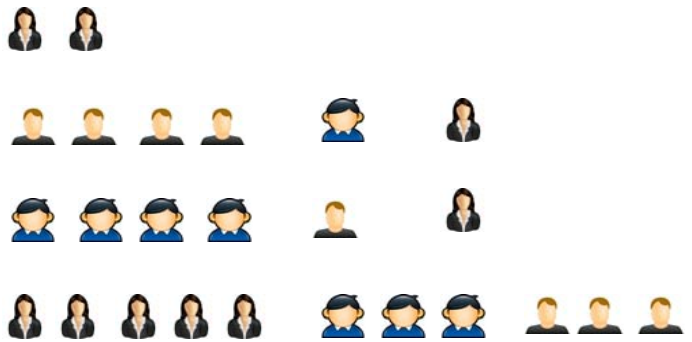
L_2 distance

$$L_2 = \sum_i (p_i - q_i)^2$$

	4	3	1	8	0.44	0.50	0.14	0.36%	**13 %**
	2	2	0	4	0.22	0.33	0.00	1.21%	**11 %**
	0	0	3	3	0.00	0.00	0.43	0.00%	**18 %**
	1	1	1	3	0.11	0.17	0.14	0.36%	0%
	0	0	2	2	0.00	0.00	0.28	0.00%	**8 %**
	1	0	0	1	0.11	0.00	0.00	1.21%	0%
	1	0	0	1	0.11	0.00	0.00	1.21%	0%
	0	0	0	0	0.00	0.00	0.00	0.00%	0%
	9	**6**	**7**					Σ = 0.04	0.50

How to create histogram bins

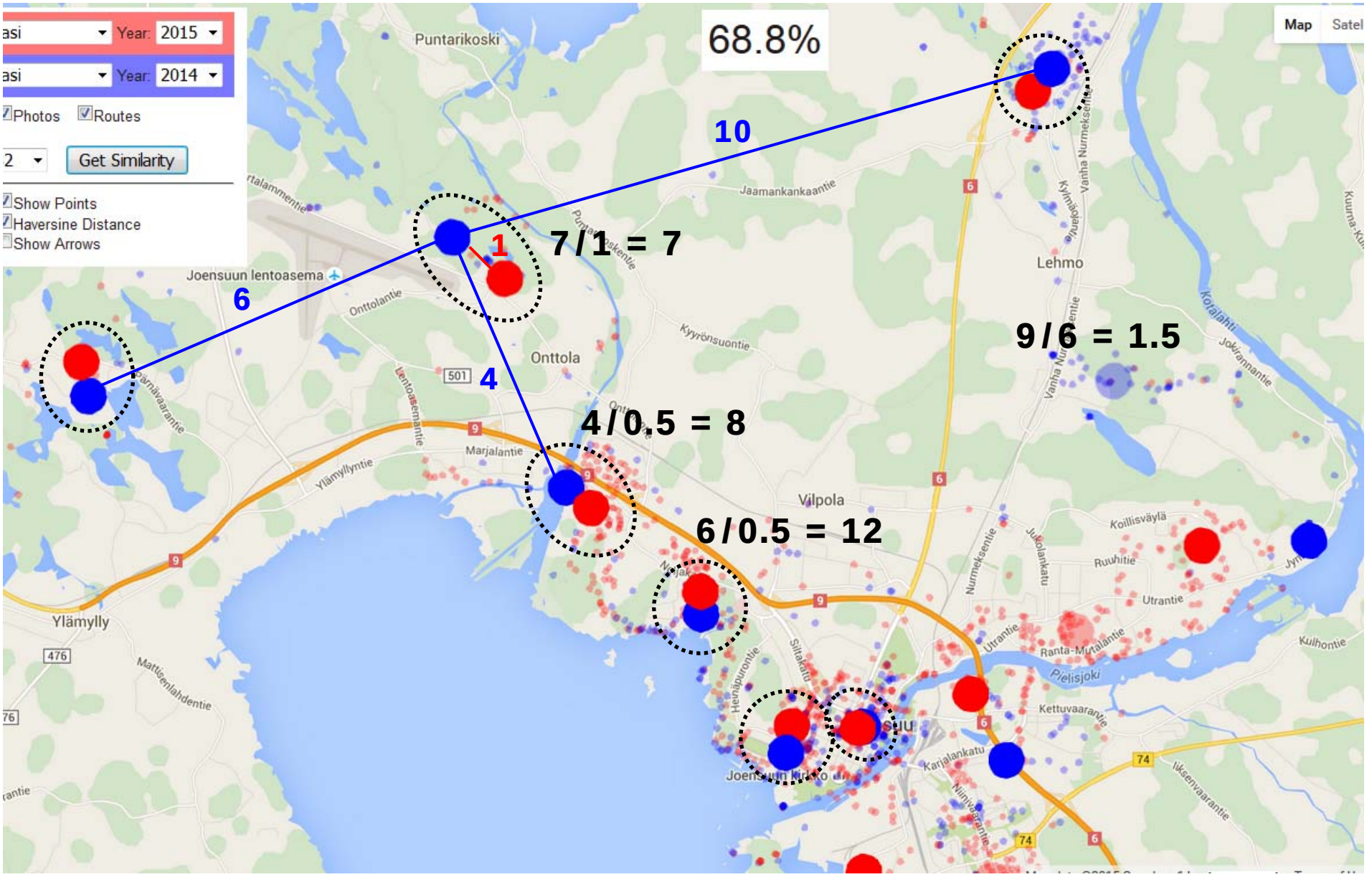
- Uniform global grid (e.g. 25×25 m)
- **Some known places (Mopsi services)**
- All user data from Mopsi
- The two data as such:
  - Cluster the joint activity data
  - Each cluster represents one bin
  - Count blues and reds in the cluster using any histogram matching technique

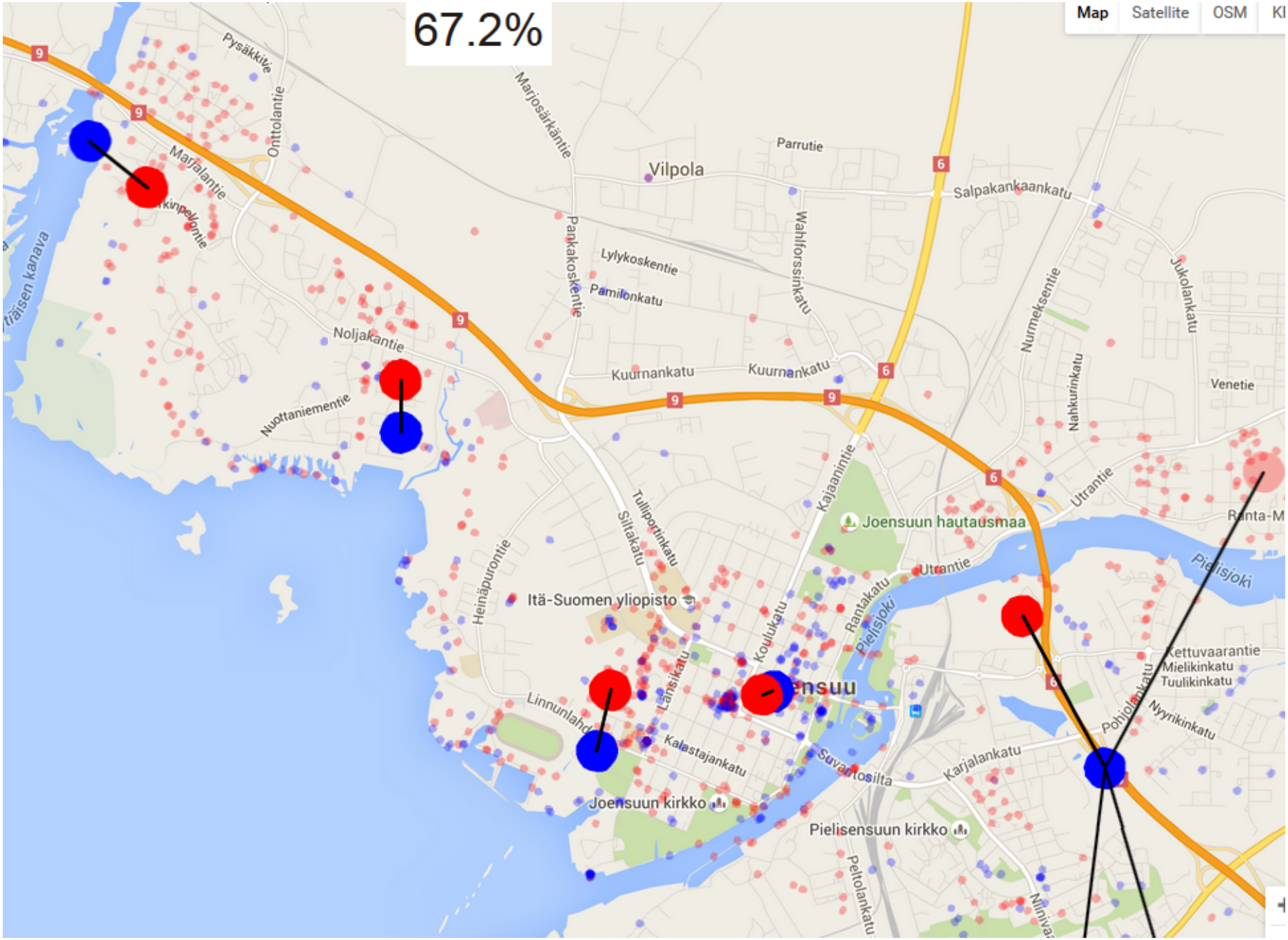


Approach 2:

Clustering

Clustering of two distributions

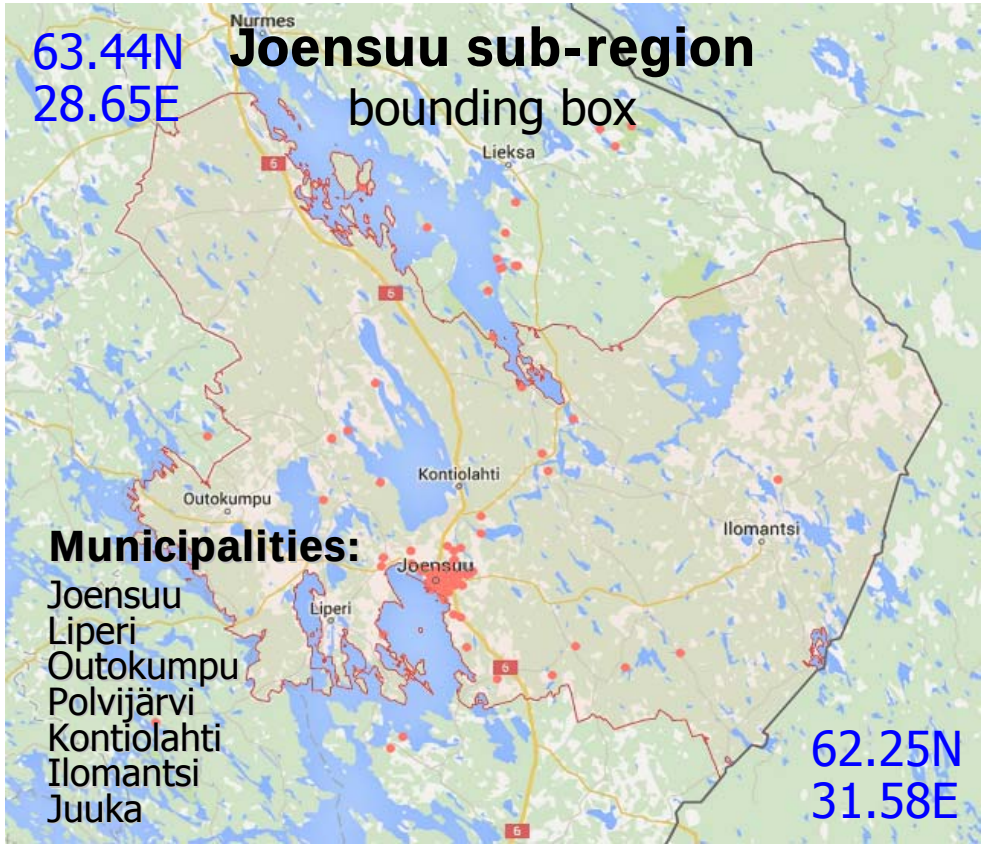



67.2%
Map
Satellite
OSM
Vilpola
Parrutie
Salpakankaankatu
Marjosärkäntie
Pysäkkitie
Onttolantie
Marjalantie
Pankakoskentie
Lylykoskentie
Pamilonkatu
Wahlforssinkatu
Noljakantie
Kuurnankatu
Nuottaniementie
Nurmeksentie
Nahkurinkatu
Jukolankatu
Venetie
Kajaanintie
Utrantie
Joensuun hautausmaa
Tullipuistinkatu
Siltakatu
Heinäpurontie
Itä-Suomen yliopisto
Koulukatu
Rantakatu
Pielisjoki
Linnunlahdenkatu
Kalastajankatu
Suvantosilta
Joensuun kirkko
Pielisensuun kirkko
Karjalankatu
Kettuvaarantie
Mielikinkatu
Tuulikinkatu
Pohjolankatu
Nyyrikinkatu
Peltolankatu

Experimental results

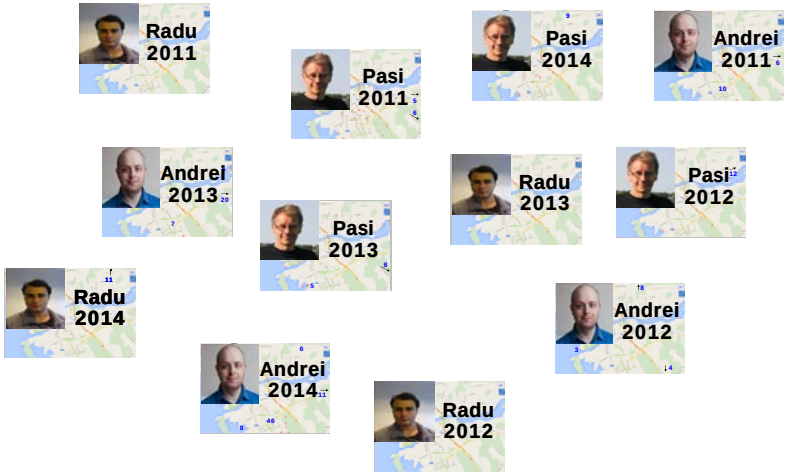
Collected data

- Histogram from **293** places (Mopsi services)
- User activities until 31.12.2014
  - Photos taken
  - Tracking started or ended



Summary of APR trio data

2011-2014



Distinct users: 3
Total users: 12
Sample sizes: 37 - 1263

	2011	2012	2013	2014
Andrei	206	757	432	329
Pasi	1263	545	636	751
Radu	37	292	324	259

Ten most popular histogram bins

Frequencies shown

Home

	Andrei				Pasi				Radu				
	2011	2012	2013	2014	2011	2012	2013	2014	2011	2012	2013	2014	Total
1	20	0	29	**150**	47	7	6	8	1	2	2	3	**275**
2	13	11	87	36	64	17	16	7	0	1	12	2	**266**
3	0	0	0	0	**51**	**54**	**54**	**69**	0	1	0	0	**229**
4	**12**	**107**	**87**	0	0	0	1	2	0	0	2	0	**211**
5	1	0	0	1	34	9	20	11	0	0	52	**54**	**182**
6	6	29	10	3	6	2	1	3	7	54	35	16	**172**
7	7	92	6	0	18	4	7	15	0	13	4	2	**168**
8	**22**	**6**	**5**	**4**	**36**	**9**	**6**	**18**	1	**12**	**11**	**21**	**151**
9	0	3	4	0	0	0	0	1	**12**	**82**	**41**	7	**150**
10	0	0	0	0	73	6	48	13	0	0	0	0	**140**

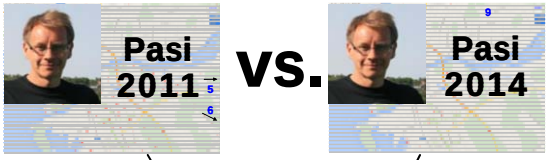
Work

Home

Test protocol

Expected result

True positive:



Distance

SAME

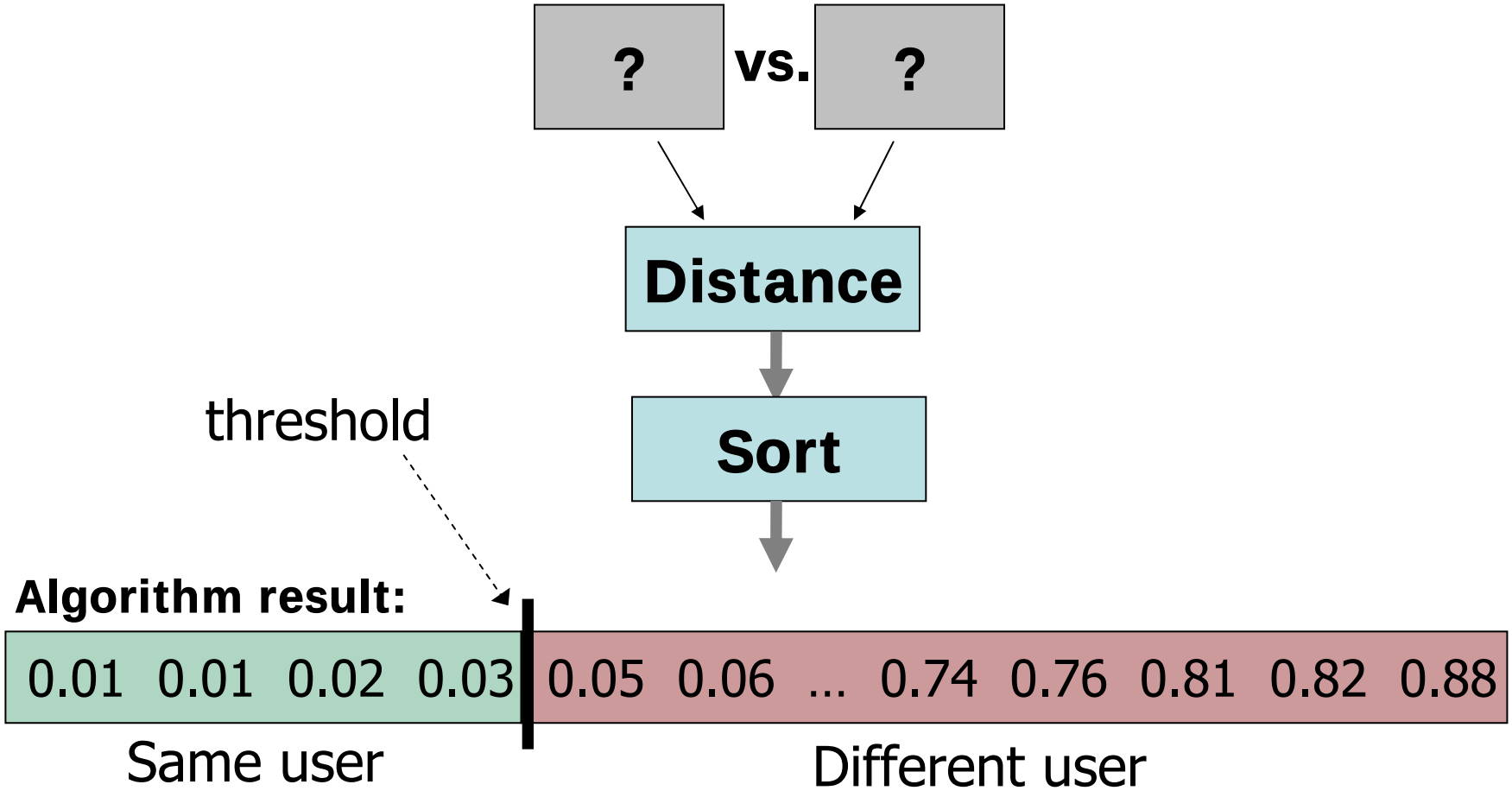
True negative:

Distance

DIFFERENT

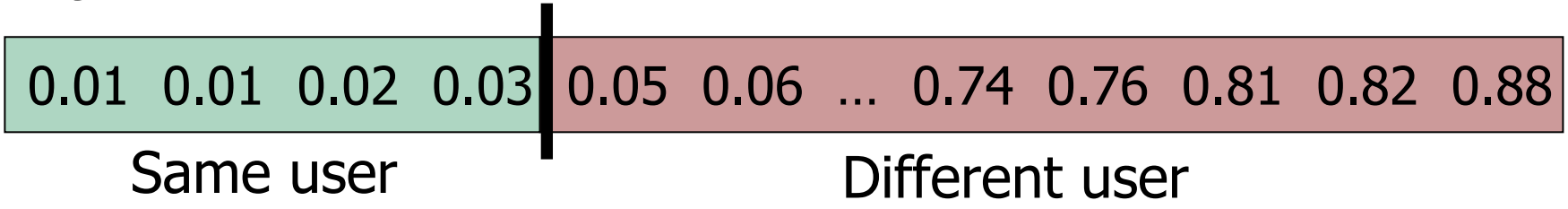
- 12*12 = 144 comparisons in total
- 3*4*4 = 48 true positives (33%)
- 3*8*8 = 96 true negatives (67%)

Comparison of all pairs



Classification error

Algorithm result:



Ground truth:

same same **not** same | **same same** … not not not not not

Errors made: **3**
Total comparisons: **144**
Classification error: **2%**

Classification results

Average distance

A priori Top-33%

Best possible

	Threshold (Crisp)			Error (Crisp)			Error (fuzzy)		
	Mean	Apr.	Oracle	Mean	Apr.	Oracle	Mean	Apr.	Oracle
L_1	0.31	0.27	0.28	8%	8%	7%	10%	10%	10%
Chi2	1.22	1.24	1.18	8%	7%	7%	17%	11%	10%
BHA	0.46	0.46	0.48	10%	10%	8%	15%	14%	11%
KLD	0.82	0.89	0.88	11%	11%	10%	36%	21%	15%
L_2	0.84	0.80	0.88	35%	47%	15%	35%	49%	14%
L_∞	0.79	0.72	0.87	43%	43%	18%	38%	47%	21%

Fuzzy worse than crisp

What if top-10 places removed?

Method:	All data:	Excluding Top-10:	Observation:
L_1	8%	24%	Loses its ability
Chi^2	8%	13%	A11 becomes similar with P11, P13, P14
BHA	10%	13%	A11 becomes similar with P11, P13, P14. P11-R14 no longer matches.
KLD	11%	13%	A11 and R14 become similar, no other effects.
L_2	35%	40%	Works slightly worse.
L_∞	42%	43%	Works slightly worse.

Conclusions

Main results:

- People's identity can be recognized with high accuracy
- All except L_2 and L_∞
- Fuzzy works worse than crisp

Future research:

- Clustering-based approach
- K-NN classifier

Thank you

Time for questions!