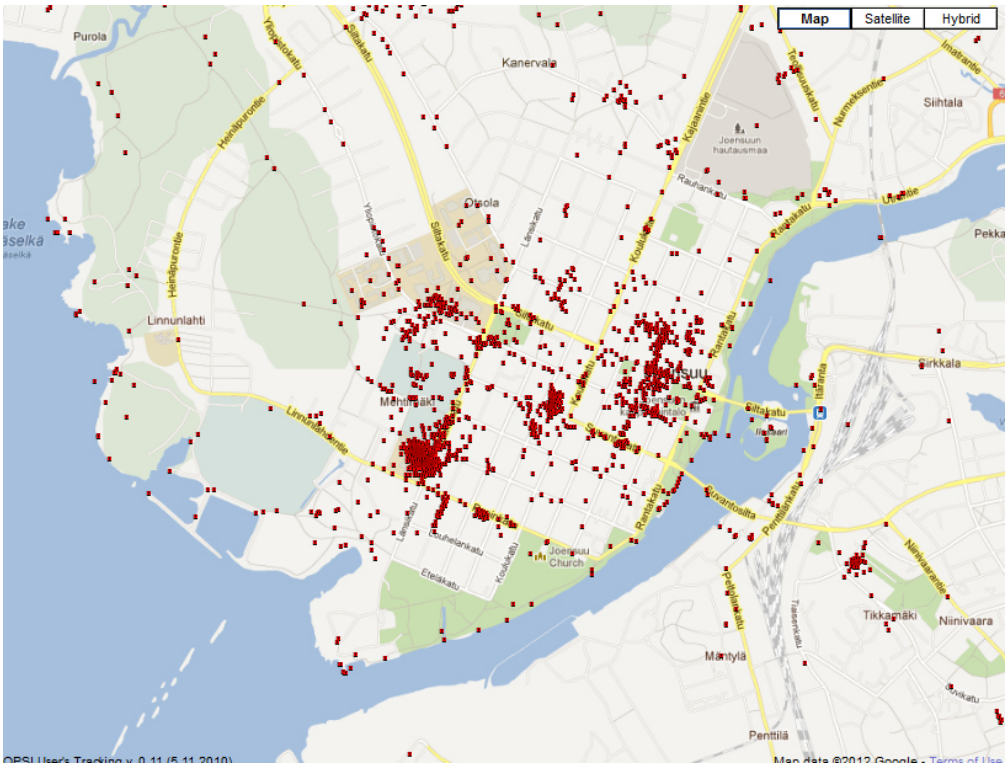
Pattern Recognition
Exercise set 2

1. **[2 points]** How do classification (supervised learning) and clustering (unsupervised learning) differ in terms of their objectives, data requirements, and implementation (in general)?

2. **[6 points] (k-means implementation)**

   - Use the attached GPS data to test your k-means implementation. Choose only the data points whose latitudes and longitudes are within ranges 62.59–62.61 and 29.7 – 29.8, respectively (Joensuu city centre). Then, divide longitude values by 2 to make coordinate axes (almost) equally scaled in terms of euclidean distance.
   - Implement k-means. Inputs: data points to be clustered and the number of clusters. Outputs: cluster centroids (means) and cluster indices that will tell to which cluster each clustered data point was assigned to.
   - Plot clustered data points and the obtained cluster centroids to the same figure (hint: `scatter; hold on`). Use different color for the cluster centroids. Try using different numbers of clusters.



3. **[3 points] (Principal component analysis (PCA) example)**
   Compute the mean vector μ (hint: `mean`) and the covariance matrix Σ (hint: `cov`) of the whole i-vector data (excluding labels). Obtain 2 principal directions of the data by computing eigenvectors and eigenvalues (hint: `eig`) of Σ and by choosing eigenvectors that correspond to the 2 largest eigenvalues. Assuming the principal directions are stored in 7×2 matrix P and *centered* i-vectors in 7×2000 matrix V, you can project the i-vector data into two-dimensional space by computing $P^T V$. I-vectors are centered by subtracting the data mean μ from all of the i-vectors. Finally, plot the projected 2-dimensional data points so that the points from different genders are plotted with different colors.

4. [**Bonus (4 points)**] **Agglomerative clustering (self-learning)**
   Find information about agglomerative clustering and then implement and test such clustering algorithm. If necessary, reduce the number of data points in the GPS data set in some way to make the algorithm run in a reasonable time. Plot the clustering results.

Bonus exercises can give you extra points, which makes it possible to get more than one-third of the course points from the exercises. For example, from this exercise set you could get 15/11 points i.e. 136% of the max. points.

Submit your answers to Moodle by November 23, Thu, 23:55.
Late submissions:
Before November 24, Fri, 7:00: -10% of points
Before November 25, Sat, 12:00: -30% of points
After November 25, Sat, 12:00: Not accepted without a good reason that has to be given well before the deadline.

The submission should be an archive (zip, tar, etc.) that contains following files:

- `answers.pdf`: Contains answers to questions in pdf-format. If you wish, you may include scanned (readable) handwritten answers in your answer file (to avoid math typesetting). In case you give some answers in program code comments, mention it in the `answers.pdf`. Include your full name.
- `main.m` / `main.py`: A script that outputs the answers for all programming tasks. Write your full name in the first line as a comment. If using Python, please use Python3 + NumPy + matplotlib.
- Possibly other code files that your main script calls.
- Include the required data set files (ivectors.txt) in the archive so that the main script is runnable right after unpacking the archive.

Use the following naming convention for the archive file:
`<first name>_<last name>_ex<exercise set number>`
For example:
`ville_vestman_ex2.zip`