Area: Data Analytics / Cognitive Science

Supervisor: Niko Troje, Professor (troje@queensu.ca)

Department of Psychology, Queen's University

Skill required: Knowledge of machine learning and data analytics tools and technologies. Knowledge of Psychology would be an asset.

Description: Mine a big data base of containing "semantic scaling" data. Over many years, visitors of the lab website have rated a set of 100 point-light walkers (simple depiction of individual walkers) according to attributes which they could come up with themselves. The data basically explore the interindividual variability in the way people walk and try to assign semantic contents to it. Examples are the "axes" depicted in the BMLwalker demo of the biomotion lab at http://www.biomotionlab.ca/Demos/BMLwalker.html (sex, mood, nervousness, weight).

If you go one click further (http://www.biomotionlab.ca/Demos/BMLrating.html) you get to the tool that we used to collect the rating data. The user can come up with ANY attribute, say, "voting behaviour" which might range from "liberal" to "conservative". Or "neighbourhood" which might range from "Harlem" to "East Side Central Park". Whatever you want. Once you have chosen your attribute, you are displayed with individual walkers -- each one depicting the walking style of a particular person in our data base. You have to rate each walker according to a scale defined by the attribute you chose. At the end we regress these rating onto the walkers and derive a linear function in walker space. The sliders in BMLwalker allow you to visualize walkers along these linear discriminant functions.

We collected close to 100,000 such data sets and they are sitting in a mySQL database on our server. Many of them are useless but many, many are super interesting. The data are just sitting there awaiting smart analysis. Questions are:

1. How do attributes cluster? What are principle attributes that characterize an individual? How are they called? The walkers are represented in 10 dimensional space spanned by the principal components derived from a representation that is based in the frequency domain (Fourier components). Each individual walker is a point in 10-dimensional space, and each of the semantic axes (voting behaviour, neighbourhood, etc.) is also a vector in 10-D space. The 100,000 dimensions that we captured are probably not randomly distributed in walker space. There will be hot spots -- areas that people chose more often than others. What do these clusters represent? Which labels did people use for these attributes?

2. What can we learn about the walkers themselves? Which attributes often go together and which ones don't. Starting with a space spanned by all attributes that people used, can we identify a lower dimensional subspace which contains typical walkers? Can we recover the Big Five, the personality traits that classic personality researchers have long claimed to be the most important ones?

3. How does the rating reflect the information about the raters themselves. We don't know too much about them, but we have their age and we have their nationality. Do people in East Asia have

different concepts of, say, confidence, attractiveness, or trustworthyness than people in Europe have?