Abstract for the general public

The sequence of a human genome encodes an informative history of the development of a single cell into a profound organism. A human body consists of tens of trillions of cells, which, despite the small size, store many substructures. The genome consists of four basic building units - nucleotides - stored in 23 pairs of chromosomes. Chromosomes are built cumulatively of tens of thousands of genes that define the instructions of the organism's development. The numbers listed capture the scale of information encapsulated in each cell. A natural question arises - what can be learned if we understand the language programming the organism development? Genomics, a branch of science derived from genetics, deals with the answer to this question. Contemporary genomics is a very interdisciplinary science that tries to understand the functions and evolution of genomic structures. One of the main tools used in genomics is sequencing, the result of which is the information about genome composition. For example, we can get a gene expression matrix that counts the frequency of gene products. However, the massive sequencing data makes gene-by-gene analysis practically impossible. This leaves a vast field for mathematicians and computer scientists to create tools and software for processing the data and extracting essential information that biologists can easily interpret in a significantly shorter time.

During the last decade, the *single-cell sequencing* technologies have developed rapidly. Previously, all analyses were carried out on groups of cells (*bulk sequencing*), which is a technologically easier, cheaper, but much less accurate approach. The resolution of single-cell sequencing allows the examination of each cell separately, which was a breakthrough in genomics. The single-cell sequencing was awarded the *Method of the Year 2013* by *Nature Methods*. Due to further technological developments, single cells can now be sequenced at multiple molecular levels, e.g., chromatin, RNA, or proteins. One of the most state-of-the-art approaches is multimodal sequencing - simultaneously sequencing the genome at different molecular levels in the same cell. As a result, we have much more profound information about each cell, as none of the molecular levels can encode the complete genome information on its own (the central dogma of molecular biology).

The proposed project aims to develop a new *topic modeling* algorithm able to compress data from multimodal single-cell sequencing into a form that is drastically smaller in size and easily interpreted. Topic modeling stems from the analysis of large text corpora. The underlying idea is to find homogenous groups of words that occur in similar contexts. To date, one of the most successful topic modeling approaches is *generative Bayesian modeling*, which uses probabilistic language to describe the relationships between words and text documents. It turns out that topic modeling (when interpreting text documents as cells, and words as, i.e., genes or proteins) copes very well with genomic data processing. However, to the best of our knowledge, no one has yet attempted to model multimodal topics for single-cell sequencing data. As a result, currently used topic models are less accurate. Using the complete multimodal information to create topics will unlock much more detailed analyzes that are impossible with current methods. We will apply modern techniques from the field of Bayesian probability theory to carry out individual research tasks. We will propose different model architectures based on existing well-functioning models and adapt them to single-cell data. We will improve the models iteratively. The resulting models will be programmed and made available in a public repository as an open-source software package.