

DATA CONSOLIDATION AND ANALYSIS SYSTEM FOR BRAIN RESEARCH

**V.I. Volosnikov^{1,a}, V.V. Korkhov^{1,b}, A.O. Vorontsov¹, K.V. Gribkov¹,
A.B. Degtyarev¹, A.V. Bogdanov¹, N.M. Zalutskaya², N.G. Neznanov²,
N.I. Ananyeva²**

¹ *Saint Petersburg State University, 7/9 Universitetskaya nab., St. Petersburg, 199034, Russia*

² *V.M. Bekhterev Psychoneurological Research Institute, St. Petersburg, Russia*

E-mail: ^a Volosnikov.apmath@gmail.com, ^b v.korkhov@spbu.ru

Comprehensive studies in the field of brain pathology require strong information support for the consolidation of data from different sources. The heterogeneity of data sources and the resource-intensive nature of preprocessing make it difficult to conduct comprehensive interdisciplinary research. To solve this problem for brain studies, an information system with unified access to heterogeneous data is required. Effective implementation of such a system requires adapting preprocessing methods and creating a model for combining disparate data into a single information environment. We analyze the possibilities and methods of consolidation of clinical and biological data, build a model for the consolidation and interaction of heterogeneous data sources for brain research, implement the model as a cloud service, and provide a data interface in a format encapsulating a complex architecture from the user. We present the design and implementation of an information system; we show and discuss the results of the application of cluster analysis methods to differentiate various types of dementia with MRI data. Our results show that a study of the properties of cluster analysis data can significantly help neurophysiologists in the study of cognitive disorders such as Alzheimer's disease, especially with the possibilities provided by the proposed information system.

Keywords: brain, data analysis, data consolidation, cluster analysis, information system, neuroinformatics, Alzheimer's disease, cloud computing, service-oriented architecture

© 2018 Vladislav I. Volosnikov, Vladimir V. Korkhov, Andrey O. Vorontsov,
Kirill V. Gribkov, Alexander B. Degtyarev, Alexander V. Bogdanov,
Natalia M. Zalutskaya, Nikolay G. Neznanov, Natalia I. Ananyeva

1. The structure of the storage and processing system

A wide range of analyzes and measurements are used by specialists during the research of the human brain. In view of the extreme complexity of the development of diseases and disorders, heterogeneous indicators, such as, for example, the results of functional diagnostics, psychological tests, blood tests and DNA, should be considered as a whole without separation from each other.

While some data are presented in a relatively simple numerical form, a number of measurements have a complex structure, e.g. MRI and fMRI data containing information about the features of the functioning of the brain. Such data require huge computational power for processing and analysis, large amounts of memory for storage. At this stage of research in V.M. Bekhterev Psychoneurological Institute, medical examination results already occupy more than 20 TB and require approximately 12 hours for the preprocessing of new results on a fairly powerful computer.

Neuroinformatics tasks are focused on the creation, storage, processing, simulation and visualization of research results. So, all these stages affect the work with large amounts of data and require the development of special software for efficient operation. For these reasons, the implementation of the cloud approach is necessary for the optimization and expansion of research, which is shown in our previous article as part of a joint project of the V.M. Bekhterev Psychoneurological Institute and St. Petersburg State University [1].

Based on the foregoing, to ensure the effectiveness of research, a cloud system for analyzing and storing data based on computing resources of St. Petersburg State University and Bekhterev Institute is being developed and integrated into the practical work of medical researchers. The mentioned system consists of a number of separate services, as shown in a scheme (Figure 1). Through the use of containerization tools, such as Docker [2], the model of system is very flexible and changeable, which is extremely important for building a virtual data center [3]. In addition, containerization allows implementation of Continuous Integration approach, reducing development and deployment costs.

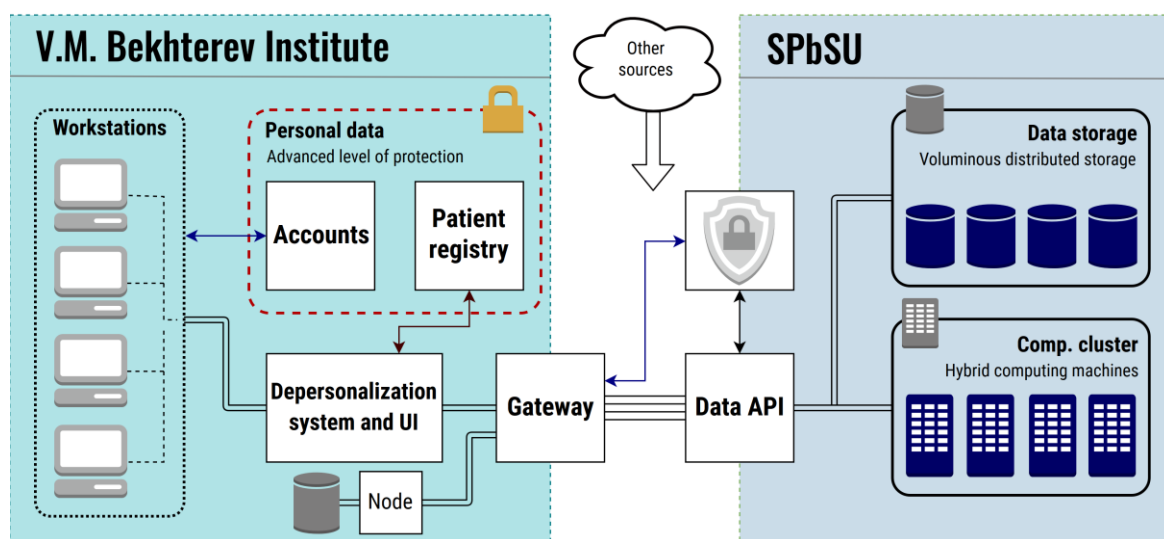


Figure 1. A schema of data storage and processing system

The use of service-oriented architecture (SOA) poses to solve a number of problems arising during the creation of the system. One of the requirements is compliance with the law in the field of personal data – the results of the research contain information about patients, which strictly should not go beyond the Bekhterev Institute. Employee data used for authorization in the system should also be stored only at the Institute.

Another important advantage of using SOA is the ease of scaling and use of new resources. Due to this, the model of the system may be integrated into existing collaborations on the study of the human brain.

2. Data preprocessing and consolidation

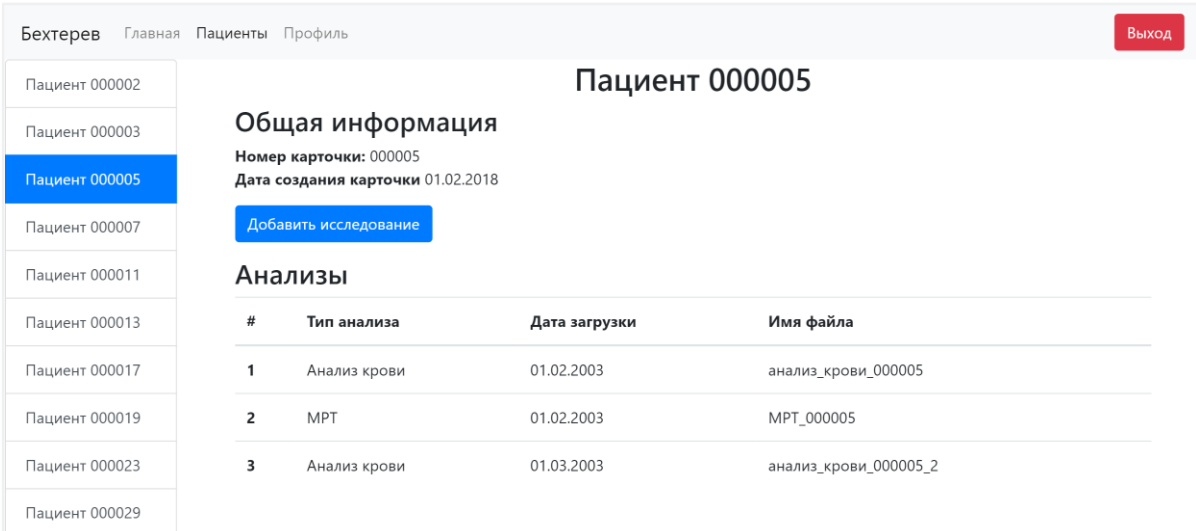
As mentioned above, the data under study is extremely heterogeneous. Due to the different structure and nature of the values being processed, each type of data requires individual tools and methods for preprocessing. It should be noted that the preprocessing of certain types of data is very costly in terms of time and computational resources – processing of raw MRI results is carrying out with the FreeSurfer package [4] and requires about 12 hours to process one record in the presence of large amounts of RAM. In other cases, there are data difficult to formalize and interpret. Some types, such as EEG, are presented in specific formats requiring a special approach [5].

The transfer of the data preparation process to the cloud system leads to a significant acceleration due to the use of computing devices with the optimal configuration for each specific case. Furthermore, distributed storage and processing systems provide an opportunity to consolidate and analyze data in a complex, which is a requirement for successful research in this subject area.

3. Specialist working environment

Through the use of service-oriented architecture, we have the opportunity to freely select technology stacks for different loosely coupled parts of the system. Along with the use of distributed databases and data analysis tools in the Python ecosystem for analyzing and storing heterogeneous data, we use the MEAN (Mongo, Angular, Express, Node) stack to implement the user interface and work environment. The validity of this decision was demonstrated in a previous paper [1].

The primary task of the interface is to provide an intuitive understanding of the process of working with the system, adapting it to the requests and tasks of a specialist in the medical field. Based on the organization of the work of a specialist, the main object in this subsystem is the patient page, which contains the functionality of adding the results of various analyzes and studies, monitoring the fill level and correctness of information (Figure 2). According to the experience of using such systems, automatic entry of the functional diagnostic' results into the database and control of compliance with a particular patient are a necessary conditions for preserving the integrity and relevance of the data. Therefore, the development of a working environment that meets the requirements set by specialists is one of the priorities.



The screenshot displays a web application interface for patient management. At the top, there is a navigation bar with the name 'Бехтерев' and links for 'Главная', 'Пациенты', and 'Профиль'. A red 'Выход' button is located in the top right corner. The main content area is titled 'Пациент 000005' and is divided into two sections: 'Общая информация' and 'Анализы'. The 'Общая информация' section includes fields for 'Номер карточки: 000005' and 'Дата создания карточки 01.02.2018', along with a blue 'Добавить исследование' button. The 'Анализы' section contains a table with three columns: '#', 'Тип анализа', 'Дата загрузки', and 'Имя файла'. The table lists three analyses: 1. 'Анализ крови' (01.02.2003, анализ_крови_000005), 2. 'МРТ' (01.02.2003, МРТ_000005), and 3. 'Анализ крови' (01.03.2003, анализ_крови_000005_2). A sidebar on the left lists other patients, with 'Пациент 000005' highlighted in blue.

#	Тип анализа	Дата загрузки	Имя файла
1	Анализ крови	01.02.2003	анализ_крови_000005
2	МРТ	01.02.2003	МРТ_000005
3	Анализ крови	01.03.2003	анализ_крови_000005_2

Figure 2. A user interface example

4. Application of cluster analysis methods

An important task of the developed system is a comprehensive analysis of data and search for correlations in them. To provide this functionality, the architecture provides the possibility of using a wide range of analysis methods – using of single Data API provides transparent access to consolidated data, which makes it possible to avoid various difficulties with different approaches to analysis.

An example of implemented methods is a cluster analysis toolkit, which potential in assisting a specialist in making a diagnosis was shown in previous papers on this topic [1, 6]. It should be noted that structural and functional brain changes occur long before the obvious manifestations of cognitive impairment. In this connection, the methods of automatic neuroimaging and analysis are very useful in medical practice.

The implemented subsystem provides an opportunity to carry out cluster analysis on various brain lobes and their combinations. Convenient visual presentation, together with automatic statistical analysis, simplify the search for optimal parameters for the partitions of the required significance. Due to lack of information about clusters count and shapes, the main class of used methods are density-based algorithms. The existing articles on this topic also note the high efficiency of other approaches – random SVM and deep learning methods [7, 8]. Due to the flexibility of the architecture, these methods also will be implemented.

At this stage of development, the purpose of cluster analysis was to clearly separate the control group from patients with pronounced signs of cognitive impairment, such as Alzheimer’s disease and other dementia types. As can be seen in the graphs obtained as a result of the data processing (Figure 3), the results are consistent with expectations with a sufficient level of statistical significance. Obvious neurodegenerative changes could be separated from conditionally healthy volunteers even without taking into account already known regions of interest (ROI) and patterns. There is reason to believe that working together with experts in the field of neurophysiology on the application of a number of well-known rules and patterns can lead to a significant improvement in results and the introduction of tools into medical practice.

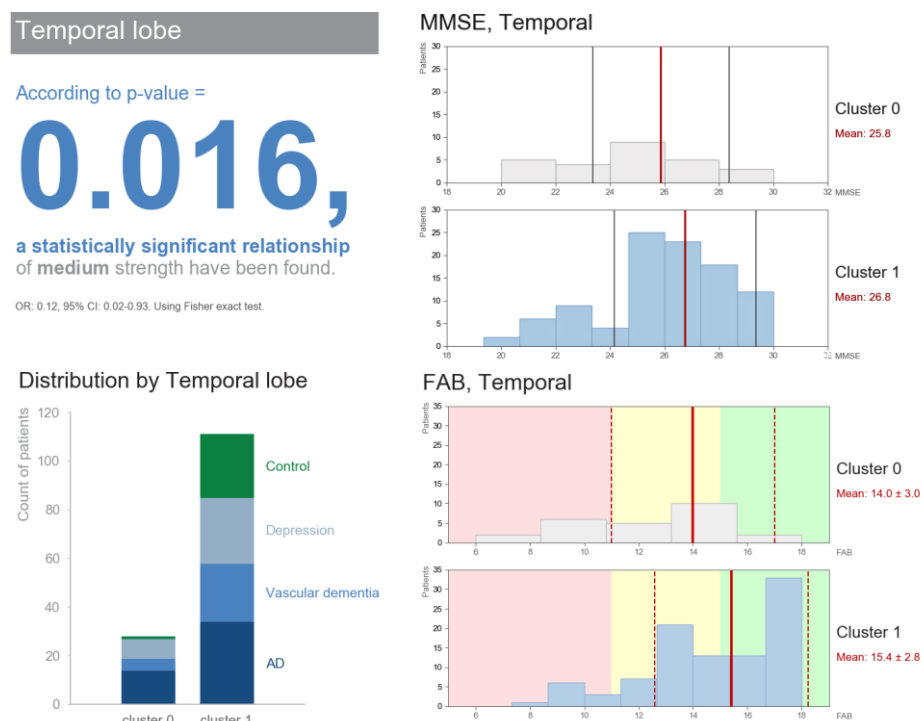


Figure 3. A result of temporal lobe analysis

5. Conclusions and future work

The developed system is necessary for successful research in the field of the human brain. In view of the presence of huge arrays of heterogeneous information that requires a complex analysis, after a certain point, work using standard tools becomes impossible – tasks require the use of fundamentally different approaches applicable to processing Big Data. These changes cease to be quantitative and acquire a qualitative character. The consequence of the above is the transition to distributed cloud computing with the prospect of integration into existing scientific collaborations.

In addition, the use of machine analysis methods is also an unavoidable necessity for conducting research of such high complexity. This statement is confirmed by the successes achieved in conducting cluster analysis – one of the tools that can simplify the work of a specialist and minimize his mistakes.

The constructed platform opens up broad opportunities for the further development of the project. We plan to expand the scope of the used data, implement a number of highly efficient methods of analysis, develop a decision support system, optimize and expand the functionality of the specialist's work environment.

6. Acknowledgement

The work on data consolidation and analysis system was supported by the grant of Saint Petersburg State University no. 26520170 and the Russian Foundation for Basic Research (RFBR), grant #16-07-00886.

References

- [1] V. Korkhov, V. Volosnikov, A. Vorontsov, K. Gribkov, N. Zalutskaya, A. Degtyarev, A. Bogdanov. Data storage, processing and analysis system to support brain research // *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2018, vol. 10963, pp. 78–90, ISBN: 978-331962403-7.
- [2] V. Korkhov, I. Gankevich, A. Degtyarev, A. Bogdanov, V. Gaiduchok, N. Ahmed, A. Cubahiro. Experience in building virtual private supercomputer // *Proceedings of International Conference on Computer Science and Information Technologies (CSIT)*, 2015, pp. 220–223, ISBN: 978-5-8080-0797-0.
- [3] A. Bogdanov, A. Degtyarev, V. Korkhov. Desktop supercomputer: what can it do? // *Physics of Particles and Nuclei Letters*, 2017, vol. 14 (7), pp. 985–992, DOI: 10.1134/S1547477117070032.
- [4] FreeSurfer // <http://surfer.nmr.mgh.harvard.edu/> (accessed: 04 Nov 2018).
- [5] WinEEG // <http://www.mitsar-medical.com/eeg-software/qeeg-software/> (accessed: 04 Nov 2018).
- [6] Volosnikov V.I. Primenenie metodov klasternogo analiza dlya diagnostirovaniya bolezni Al`czgejmery [The application of cluster analysis methods for diagnosing Alzheimer's disease] // *Control Processes and Stability*, 2018, vol. 5 (21), pp. 267–271, ISSN: 2313-7304 (in Russian).
- [7] Xia-an Bi, Qing Shu, Qi Sun, Qian Xu. Random support vector machine cluster analysis of resting-state fMRI in Alzheimer's disease // *PLoS One*, 2018, vol. 13(3), DOI: [10.1371/journal.pone.0194479](https://doi.org/10.1371/journal.pone.0194479).
- [8] Suk HI, Shen D. Deep Learning-Based Feature Representation for AD/MCI Classification // *Med Image Comput Comput Assist Interv.*, 2013, vol. 16 (pt. 2), pp. 583–590.