Final Report for the International Collaboration Grant: OTKA-103244

Background

With the advent of new technologies both the way how we live our everyday life and how we do science are changing. Moore's law is pushing the exponentially accelerating pace of microelectronics development and on the one side better sensors and equipment with larger throughput are produced, and on the other side larger storage devices and faster processors are available to handle the information. In astronomy the breakthrough was the CCD sensor, which made it possible to draw the first ever 3D map of our Universe's large scale structure and to understand details of the complex structure and dynamics on extragalactic scale. In biology, the very same technology, the CCD chip is at the heart of the gene expression microarrays and next generation sequencing machines. The analogy does not stop here: the ever growing data sets bring a new challenge to the research community, transform significantly the methodology of the studies and require new skills from scientists. A few years ago most astronomers spent their time at telescopes and most biologists in their labs. Now with the advent of the Virtual Observatories and public genomic archives, significant part of the research work is done through computers, and instead of observer or lab skills, knowledge of programming and other expertise in advanced statistics and various fields of information technology became crucial. Organizing and analyzing scientific "big data" is a common issue of many fields making it possible to utilize and reuse technologies, methods and models in multidisciplinary collaborations.

Beyond the raw information-technological challenges (store and organize terabytes or soon petabytes of data) caused by this data deluge, the most urgent and the most intriguing question is how to handle complexity. When data is scarce researchers have the time and opportunity to closely look at and understand each piece of it and build models combining the original hypothesis, background knowledge and insights from the new data. For hundreds of millions of galaxies or thousands of billions nucleotide long human genome sequences this approach is not feasible. We need methods to winnow the wheat from the chaff, to select the essence of the information out of piles of raw data, to reduce the amount and dimensionality to the level where human intelligence and intuition can take over again with the analysis. While a couple of years ago one of science's biggest problems was the lack of data, now the greatest challenge and in the meantime the greatest opportunity lies in the question of how can we handle and analyze huge datasets.

Objectives

There are no readymade systems to handle this challenge. This is not just the question of buying regular computer hardware and hiring a few programmers. Like the creation of the first microscope or telescope was not just the question of glasswork: we need to develop a new scientific instrument, a Datascope. In this proposal we have joined the significantly larger effort,

led by Prof. Alexander Szalay at the Johns Hopkins University to develop such infrastructure and use it in research projects. In this sense the task was two folded: develop a scientific data management and analysis framework and do scientific investigations which leverage it. A key point in our efforts was also to show, that the methodology of scientific big data management is inherently multidisciplinary: the same or very similar computational tools can be used in seemingly independent disciplines. We consider it as a significant success of this project, that beyond using the framework for organizing and analyzing astronomical data, we could establish new collaborations and we could exploit our gathering science data management expertise to participate in genomics [2,3,16,28,34] and networks research [11,12,17,24,32,35,38,39,44,48,49].

Participants and tasks

Most of the benefits in terms of stipends, salaries and travel costs were provided for young researchers, from BSc students to postdocs, to involve them in the international collaboration and to support their work. Though the work was usually collaborative, involving people not just from the small group working on the project, but also colleagues from other institutes, we list various tasks under the name of the young participating researchers.

László Dobos, postdoctoral fellow took part in various aspects of the Datascope development in a strong collaboration with the JHU group. He has spent several months in Baltimore and he was the chief architect and developer of the SkyQuery and GrayWulf modules and responsible for system scalability [8,18,45,46]. SkyQuery is a federated cross-matching service built on top a GrayWulf distributed cluster, and provide a scalable, interactive SQL-like query engine for cross-identification of astronomical catalogs, that enables multi wavelength astronomy. The support of the project made possible to continue the work on these very technical but very important systems. The invitation of László Dobos to the prestigious Extremely Large Database (XLDB) Invitational Workshop at CERN [7] shows that during the project László Dobos became a leading expert of this field.

We “imported” the experiences and expertise and used them to build a smaller scale, - but still capable of holding dozens of terabytes -, version of DataScope to manage our local astronomical, genetics and social and financial networks data [17,27,32]. Also, the spherical indexing tools which were designed for the sky could be used with small modification to organize the spherical polygon data of countries [6,24]. During the period of the project L. Dobos was thesis adviser for several students, Bálint Ódor, Dezső Ribli, Bálint Ármin Pataki, who worked on project related topics and used our databases in their work [13,30,41,43].

In several cases raw observed data can be interpreted mathematically as multidimensional vectors. The dimensions are the observed quantities, like wavelengths in a spectrograph of gene expression values on micro-arrays. Each observation is a vector, or a point in the space expanded by these values. The dimension can be as high as several thousand, so some dimension reduction is needed to get interpretable relations. Together with the JHU group we have been working on

various dimension reduction methods and applied them to analyze scientific datasets. For the case of galaxy spectral energy distributions, defining the wavelength regions which are representative for galaxy parameters is an important question. Using our CUR Matrix Decomposition method we could identify regions, similar to the ones in the widely used Lick-indices and found the Dn(4000), H-beta and H-delta_A features to be most informative. The regions can be used to determine the stellar age and metallicity in early-type galaxies and the method can be applied to any set of spectra, so that we eliminate the need for a common, fixed-resolution index system [21,33].

Róbert Beck is a PhD student in physics. His thesis topic is also related to galaxy spectral energy distribution analysis and dimension reduction and machine learning algorithms: k-nearest neighbor finding, k-means clustering and support vector machines. Partly supported by this project, Róbert is spending a year at JHU to extend his studies and do joint research with our partner group. Together with the JHU researchers we analyzed the correlations between continuum properties and emission line equivalent widths of star-forming and active galaxies from the SDSS. Since upcoming large sky surveys will make broadband observations only, including strong emission lines into theoretical modeling of spectra will be essential to estimate physical properties of photometric galaxies. In a paper, submitted for publication [47] we have shown that emission line equivalent widths can be fairly well reconstructed from the stellar continuum using local multiple linear regression in the continuum principal component (PCA) space. We have also shown that, by combining PCA coefficients from the pure continuum and the emission lines, a plausible distinction can be made between weak active galactic nuclei and quiescent star-forming galaxies. The classification method is based on support vector machines, and allows a more refined separation of active and starburst galaxies than the empirical curve found by other researchers.

The regular task of our group is to generate the photometric redshift catalogue for each data release of the Sloan Digital Sky Survey. This time R. Beck led this effort. We refined the estimation method, compiled a new reference set and used our recently published composite spectral template catalogue for K-correction and spectral type estimation [51].

András Bodor was responsible for managing and developing our servers, to configure tools and databases and also to adjust the original system, which was mostly designed for astronomical data to be able to handle genomics and social network data. In collaboration with biomedical research colleagues from Semmelweis University, utilizing our databases and tools, he analyzed genomics data and helped identify colorectal cancer markers [3,16,34].

János Márk Szalai-Gindl, a PhD student in computer science, joined our group during the project interval. Initiated during his summer visit at JHU, he has developed a new method which uses parallel GPU code for Bayesian statistical analysis of galaxy luminosity function [50]. Beyond that, he is exploring ways to handle efficiently multidimensional point cloud (from gravitational N-body simulations, or abstract parameter spaces) datasets [25] and use column store and relational database management systems to load, process and search genomics data [31].

József Varga completed his PhD in astronomy during the span of the project. He used the database of the Sloan Digital Sky Survey (SDSS) and developed new statistical analysis methods. While investigating the correlation between galaxy orientations and large scale structure [26], we noticed that the galaxy position angle measurements, calculated by the surface brightness profile fitting code of the photometric pipeline of SDSS are strongly biased, especially in the case of almost face-on and highly inclined galaxies. To address this issue we developed a reliable algorithm which determines position angles by means of isophote fitting and created a catalogue [9,19,42].

Large astronomical surveys, like the SDSS make possible studies which cannot be done with few targeted observations. More than half of the sources identified by recent radio sky surveys have not been detected by optical surveys, hence the nature of these objects is not known. We have developed an image double-stacking technique and applied it to detect the optical emission from unresolved, isolated radio sources of the Very Large Array (VLA) Faint Images of the Radio Sky at Twenty-cm (FIRST) survey that have no identified optical counterparts even in the in the deep SDSS Stripe 82 co-added data set. Double-stacking means, that the already co-added Stripe 82 images, centered on the positions of the "invisible" radio sources were added together yielding an even deeper composite image which revealed an "average" source and its properties could be analyzed [4,20,42].

Gyöngyi Kerekes is a PhD student in astronomy. Getting spectra at good signal-to-noise ratios takes orders of magnitudes more time than photometric observations. Utilizing our machine learning technique previously developed for photometric redshift estimation of galaxies [14] she developed a nonparametric method for estimating the chemical composition of galactic stars using only photometric information instead of spectroscopy. We investigated the efficiency of our method using spectroscopically determined stellar metallicities from the SDSS database. The technique is generic in the sense that it is not restricted to certain stellar types or stellar parameter ranges and makes it possible to obtain metallicities and error estimates for a much larger sample than spectroscopic surveys would allow. Our method performs well, especially for brighter stars and higher metallicities and, in contrast to many other techniques, we were able to reliably estimate the error of the predicted metallicities [10].

Áron Süli joined our group for a short period as a post-doctoral fellow to develop GPU-accelerated astronomical software. Our group was experimenting to use this technology for luminosity function estimation [50] and to assemble galaxy spectral energy distributions with stellar population modeling [30]. Together with Á. Süli we have developed *gSOLARIS*, a GPU accelerated N-body code for astronomical simulations [22]. The support from the project also made possible for Á. Süli to complete his work on another gravitational simulation problem, and investigate how to control chaos in the vicinity of the Earth–Moon L5 Lagrangian point and keep a spacecraft in orbit [37].

Gábor Rácz, a graduating MSc physics student also worked on gravitational N-body problems, but on a different scale. The JHU group has created the so called *Indra Simulation suite* is a set of 512

cosmological N-body simulations in a 1Gpc/h-sided box producing over 100 TB of data, We have been involved to create the SQL database, that store the snapshots of the simulations and which is indexed with Peano-Hilbert curve to support efficient searches. Possible queries involve sampling all of the particle data for a particular snapshot, such as computation of particle topologies like filaments, voids, and clusters. With Gábor Rácz we investigated how the simulations can be improved, to handle the inhomogeneities. State of art simulation codes, like the GADGET2 N-body code used for Indra, assume that the matter distribution in the Universe is homogeneous and isotropic. Gábor has developed a new code that takes into account the back-reaction of clustering matter, and showed that this back-reaction alone can yield an accelerated expansion without dark energy [15,40].

István Csabai, PI, was responsible for managing the work, conduct and co-advise project related BSc, MSc and PhD thesis works [5,12,14,15,28,35,38,39,40,42], establish new interdisciplinary collaborations where our framework can be used, and disseminate the results (several public lectures at Eötvös University, Johns Hopkins University, Corvinus University, Semmelweis University, Hungarian Academy of Sciences, summer schools, etc.) [36].

Since “scientific databases” can be considered as a new emerging discipline by itself we have started a new course at the Eötvös Loránd University with the title “Design and implementation of scientific databases” which was very popular among physics and computer science students, and another one “Astronomical databases” specifically for astronomers.

In summary, the support from OTKA made it possible to continue our long term collaboration with the partner group at the Johns Hopkins University, start new interdisciplinary collaborations and research topics, continue to build the scientific data management and analysis framework, publish results at/in refereed international conferences/journals and at last but not at least the grant made it possible for many young researchers to join this exciting endeavor.

Publications

Including the key publications listed in the specific sections of the report we list here all results that were completely or partially supported by this grant.

1. Dobos László: Galaxispopulációk fizikai paramétereinek meghatározása és Virtuális Obszervatóriumok, PhD disszertáció, ELTE, 2012
2. S. Spisak et al.: Plazmaminták szabad dns frakciójának teljes genom szintű újraszekvenálása és elemzése, VII. Magyar Sejtanalitikai Konferencia, 2012
3. S. Spisak, N. Solymosi, P. Ittzes, A. Bodor, D. Kondor, G. Vattay, B. Bartak, F. Sipos, O. Galamb, Z. Tulassay, Z. Szallasi, S. Ramussen, T. Sicheritz-Ponten, S. Brunak, B. Molnar, I. Csabai: Metagenome analysis of human plasma samples from inflammatory bowel disease, colorectal adenoma and colorectal cancer patients using next generation sequencing, UEGW Week 2012, poster, 2012

4. Varga J, Csabai I, Dobos L: Revealing a strongly reddened, faint active galactic nucleus population by stacking deep co-added images, MON NOT R ASTRON SOC 426: (2) 833-850, 2012
5. Beck Róbert (Csabai István témavezető): Dinamikus gráfautomaták - diszkretizált gravitáció szimuláció, Msc diplomamunka, ELTE, 2013
6. Budavári, Tamás; Dobos, László Fekete, György; Gray, Jim; Szalay, Alex: Spherical: Geometry operations and searches on spherical surfaces, Astrophysics Source Code Library, record ascl:1309.004, 2013
7. Dobos L, Budavari T, Szalay AS, Csabai I: Sky Query: A distributed query engine for astronomy, Extremely Large Database (XLDB) Invitational Workshop, CERN, 2013
8. Dobos L, Csabai I, Szalay AS, Budavári T, Li N: Graywulf: A platform for federated scientific databases and services, In: s n (szerk.) (szerk.) 25th International Conference on Scientific and Statistical Database Management, SSDBM 2013. New York: ACM Press, 2013. (ACM International Conference Proceeding Series), 2013
9. J. Varga, I. Csabai, L. Dobos: Correct measurements of galaxy orientation angles and its implications to angular correlation studies, Ripples in the Cosmos Conference at Durham University 22-26 July 2013, 2013
10. Kerekes G, Csabai I, Dobos L, Trencséni M: Photo-Met: A non-parametric method for estimating stellar metallicity from photometric observations, ASTRON NACHRICH 334: (9) 1012-1015, 2013
11. Kondor D, Matray P, Csabai I, Vattay G: Measuring the dimension of partially embedded networks, PHYSICA A 392: (18) 4160-4171, 2013
12. Mátray Péter (Csabai István témavezető): Az Internet térbeli szerkezetének elemzése és a Hálózati Mérések Virtuális Obszervatóriuma, PhD disszertáció, ELTE, 2013
13. Ódor Bálint (Dobos László témavezető):: Aktív galaxisok spektroszkópiai vizsgálata, Bsc szakdolgozat, ELTE Fizika Intézet, 2013
14. Purger Norbert (Csabai István témavezető): Fotometrikus vöröseltolódás-becslési módszerek továbbfejlesztése, PhD disszertáció, ELTE, 2013
15. Rácz Gábor (Csabai István témavezető): Kozmológiai struktúraképződés, Bsc Szakdolgozat, ELTE, 2013
16. S. Spisak, N. Solymosi, P. Ittzes, A. Bodor, D. Kondor, G. Vattay, B. Bartak, F. Sipos, O. Galamb, Z. Tulassay, Z. Szallasi, S. Ramussen, T. Sicheritz-Ponten, S. Brunak, B. Molnar, I. Csabai: Complete genes may pass from food to human blood, PLoS ONE 8(7): e69805, 2013
17. T. Sebok, Zs. Kallus, S. Laki, P. Matray, J. Steger, L. Dobos, I. Csabai, G. Vattay: The Network Measurement Virtual Observatory: An Integrated Database Environment for Internet Measurements and Data Analysis, 25th International Conference on Scientific and Statistical Database Management (SSDBM 2013), July 29-31, 2013, Baltimore, Maryland, USA, 2013
18. Tamas Budavari, Laszlo Dobos, Alexander S. Szalay: SkyQuery: Federating Astronomy Archives, Computing in Science and Engineering, vol. 15, no. 3, pp. 12-20, 2013

19. Varga J, Csabai I, Dobos L: Refined position angle measurements for galaxies of the SDSS Stripe 82 co-added dataset, ASTRON NACHRICH 334: (9) 1016-1019, 2013
20. Varga J, Csabai I, Dobos L: Deep co-add stack (DCS) sample (Varga+, 2012), VizieR Online Data Catalog 742: 60833, 2013
21. Yip, Ching-Wa; Mahoney, M. W.; Szalay, A. S.; Csabai, I.; Budavari, T.; Wyse, R. F.; Dobos, L.: Objective Identification of Informative Wavelength Regions in Galaxy Spectra, American Astronomical Society, AAS Meeting #221, #303.04, 2013
22. A. Suli, L. Dobos, E. Forgacsne-Dajka, I. Csabai: gSOLARIS: a GPU accelerated N-body code for astronomical simulations, https://github.com/suliaron/solaris.cuda, 2014
23. Csabai I, Dobos L, Rácz G, Rudd B, Beck R: Struktúraképződés sztochasztikusan táguló térben, Statisztikus Fizikai Nap, Magyar Tudományos Akadémia, 2014
24. Dániel Kondor, László Dobos, István Csabai, András Bodor,Tamás Budavári, Alexander S. Szalay: Efficient classification of billions of points into complex geographic regions using hierarchical triangular mesh, 26th International Conference on Scientific and Statistical Database Management, 2014
25. Dobos L., Csabai I., Szalai-Gindl J.M., Budavari T., Szalay A.S.: Point Cloud Databases, 26th International Conference on Scientific and Statistical Database Management, 2014, 2014
26. J. Varga, I. Csabai, L. Dobos: Intrinsic alignment between galaxies and the large scale structure, Alpine Cosmology Workshop Ausztria, Gschnitztal 2014, 2014
27. L. Dobos, B, Pinczel, A. Kiss, G. Racz, T. Eiler: A comparative evaluation of NoSQL database systems, Annales Univ. Sci. Budapest, Sect. Comp. 42. 173-198 2014, 2014
28. Pipek Orsolya Anna (Csabai István témavezető): A genom, mint komplex rendszer, Msc diplomamunka, ELTE Fizika Intézet, 2014
29. R. Beck, L. Dobos and I. Csabai: Quantifying correlations between galaxy emission lines and stellar continua using a PCA-based technique, Statistical Challenges in 21st Century Cosmology Proceedings IAU Symposium No. 306, 2014 A. F. Heavens, J.-L. Starck & A. Krone-Martins, eds. International Astronomical Union doi:10.1017/S1743921314010990201 4 , 2015
30. Ribli Dezső (Dobos László témavezető): Galaxisspektrumok modellezése GPU-n, Bsc szakdolgozat, ELTE Fizika Intézet, 2014, 2014
31. Szalai-Gindl J.M, Dobos L, Csabai I: LoaderToolkit: a parallelized SQL Server loader for big science datasets, https://github.com/szalaigj/LoaderToolkit, 2014
32. Tamás Sebők, Zsófia Kallus, Sándor Laki, Péter Mátray, József Stéger, János Szüle, László Dobos, István Csabai, Gábor Vattay: Network Measurement Virtual Observatory: An Integrated Database Environment for Internet Research and Experimentation, Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering Volume 137, pp 65-74, 2014, 2014
33. Yip C-W, Mahoney MW, Szalay AS, Csabai I, Budavári T, Wyse RFG, Dobos L: Objective identification of informative wavelength regions in galaxy spectra, ASTRON J 147: (5) , 2014

34. Alexandra Kalmár, Bálint Péterfia, Péter Hollósi, Barnabás Wichmann, András Bodor, Árpád V. Patai, Andrea Schöller, Tibor Krenács, Zsolt Tulassay, Béla Molnár: Bisulfite-Based DNA Methylation Analysis from Recent and Archived Formalin-Fixed, Paraffin Embedded Colorectal Tissue Samples, Pathology & Oncology Research, 1219-4956 1-8, 2015
35. Bokányi Eszter (Csabai István konzulens): A Twitter fizikája: szociálishálózat-kutatás a komplex rendszerek kutatási módszereivel, MSc diplomamunka, ELTE, 2015
36. Csabai István: Adat-intenzív megközelítés a modern természettudományokban1, Magyar Tudomány, 2015
37. Judit Slíz, Áron Süli, Tamás Kovács: Control of chaos in the vicinity of the Earth-Moon L5 Lagrangian point to keep a spacecraft in orbit, Astronomische Nachrichten Volume 336, Issue 1, pages 23–31, February 2015
38. Kondor Dániel (Csabai István konzulens): Empirical analysis of complex social and financial networks, PhD disszertáció, ELTE, 2015
39. Laki Sándor (Csabai István témavezető): Analysis of complex communication networks: from measurement control to optimization, PhD diszzertáció, ELTE, 2015
40. Rácz Gábor (Csabai István témavezető): Az Univerzum nagyskálás szerkezetének vizsgálata gravitációs N-test szimulációkkal, MSc diplomamunka, ELTE, 2015
41. Pataki Bálint Ármin (Dobos László témavezető): Galaxishalmazok azonosítása, BSc szakdolgozat, ELTE 2015
42. Varga József (Csabai István témavezető): Modern képfeldolgozó eljárások alkalmazása csillagászati égboltfelmérésekben, PhD disszertáció, ELTE, 2015
43. Zsidi Gabriella (Dobos László konzulens): A napaktivitás vizsgálata, BSc szakdolgozat, ELTE, 2015
44. Dobos L, Szüle J, Sebők T: TwitterToolkit, a software library to build a relational database from twitter social network data, source code at http://github.com/twtoolkit
45. Dobos L: Graywulf: system for distributed database cluster management, source code at http://github.com/idies/graywulf
46. Dobos L: SkyQuery: astronomical cross-match engine, source code at http://github.com/idies/skyquery

Submitted and ready to submit publications:

47. R. Beck, L. Dobos, I. Csabai, CW. Yip, AS Szalay: Quantifying correlations between galaxy emission lines and stellar continua, submitted to Mon. Not. R. Astron. Soc, 2015
48. S. Kisfaludi-Bak, T. Sebok, Z. Kiraly, I. Csabai: Guaranteed Milgram routing using near-optimal address lengths, submitted to IEEE Transactions on Communications, 2015
49. E. Bokanyi, D. Kondor, L. Dobos, T. Sebok, J. Steger, I. Csabai, G. Vattay: Unsupervised Twitter Analysis Reveals Dominant Language and Demographic Patterns in the United States, to be submitted to Proceedings of the National Academy of Sciences, 2015

50. J. Szalai-Gindl, T. Budavari, TJ. Loredo, BC. Kelly, I. Csabai, L. Dobos: Hierarchical Bayesian Method for Estimating Luminosity Function, to be submitted to Journal of Computational and Graphical Statistics, 2015
51. R. Beck, L. Dobos, T. Budavári, AS. Szalay, I. Csabai: Photometric redshifts for SDSS Data Release 12, In preparation, 2015. The catalogue and a short description is already online at http://www.sdss.org/dr12/algorithms/photo-z/ and http://skyserver.sdss.org/dr12/en/help/browser/browser.aspx#&&history=description+Photoz+U