Final Report for the International Collaboration Grant: OTKA NN 114560

Background

Large astronomical surveys systematically observe the sky at a wide range of wavelengths to produce photometric and spectroscopic catalogs of galactic and extragalactic object from the microwave to gamma. Multiwavelength astronomy depends on the combination of data of these catalogs but even though the same sky is observed by each instrument, cross-matching the catalogs based on celestial coordinates poses a series of problems. The varying angular resolution and instrument sensitivity, the different spectral energy distribution of the objects and the size of the data sets render cross-matching both an astrostatistical and a computational challenge. Catalog cross-matching is not a one-time problem, and although precomputed multi-instruments catalogs are often created and published according to the particular requirements of the project, most research projects have very specific requirements on how cross-matching is done. Also, cross-matched catalogs are usually created for a handful of instruments and cannot – due to the combinatorial explosion of the problem – be precomputed for just any arbitrary set of catalogs. The large size of the data sets poses a data management problem for which the solution of seldom in the toolset of an astronomer. For more than a decade, relational database management systems (RDBMS) have been widely used to address the data management side of survey catalogs, whereas no scalable solution was available to the cross-matching problem. Astronomers have learned how to access and process the large catalogs by accessing databases at remote data centers and write SQL queries to filter and aggregate the data, hence, it become an obvious requirement to support cross-matching in a way similar to how SQL databases work. The SkyQuery system, which was the main focus of this OTKA project, is a working solution that extends database technology with astronomical tools and solves the problem of on-demand cross-matching of billion-object-scale catalogs with a distributed, scalable approach. SkyQuery has been developed in close collaboration with research group of Prof. Alexander Szalay at the Johns Hopkins University within the NSF-funded project SciServer .

The most important application of multi-wavelength astronomy is to recover physical properties of objects that cannot be constrained from single observations. One such application is photometric redshift estimation, a field regaining new interest with the advent of large scale photometric surveys without spectroscopic follow-ups. Also, reliable photometric redshifts are necessary for weak lensing research. Our group has a long history in developing and evaluating both empirical and template-based photometric parameter estimation methods and although to original objective of the research proposal was to develop SkyQuery, as a byproduct, we published a series of papers on photo-z and developed software which will later be incorporated into SkyQuery.

Results in precision cosmology spurred a new interest in our research group towards cosmological simulations which resulted in two new approaches to numerical n-body simulations. We have been developing STePS, a simulation technique that uses a spherical compactification of the infinite Universe to circumvent the issues arising from periodic simulations. Most importantly, by simulating the model universe with a gradually changing resolution, our method can estimate the power spectrum with similar signal to noise ratio in a very broad interval of wavenumbers. To further pursue research in this field, our group has won a new OTKA NN grand starting fall 2018 in collaboration with the University of Hawaii.

Objectives and results

The original research proposal put the development of SkyQuery into the focus of the project. As primarily a software development project, the main output for the project was set to build a working system available publicly to the astronomer community, as well as provide the software in an open source way. Due to the way software are developed, and the requirement to work within the framework a larger project, SciServer, we proposed to develop and introduce new features gradually. The most important goals were to integrate the distributed database system, called Graywulf [2], which forms the basis of SkyQuery, into SciServer and develop SkyQuery to a stage where it is stable enough for public use. Another important objective was to compile a set of publicly available astronomical catalogs and convert them into the SkyQuery system to allow cross-matching. We also proposed to develop virtual observatory interfaces which would allow access to external data sources such as ESA and NASA archives and Vizier.

With close collaboration with our partners, we integrated SkyQuery with SciServer at the Johns Hopkins University which has the following benefits. A shared user database allows SkyQuery users to access all other services of SciServer, i.e. access the SkyServer CasJobs and MyDB to interact with the SDSS database and to store data in a personal database, SciServer Compute, a Jupyter infrastructure to process data, MyScratch to stage large amounts of temporary data and SciDrive, a drop-box-like system for large data transfers. The integration took about one and a half year and went parallel with converting about 50, including some very large astronomical databases into the SkyQuery format. Also, during this time, we designed and built the hardware for SkyQuery which was financed by Johns Hopkins and operates at their site. Three high-capacity, all-solid-state storage systems were built and configured which allowed installing the first stable public version of SkyQuery in 2016. A subsequent release still during 2016 extended the cross-match capabilities to restrict computations to an arbitrarily complex region of the sky. This was done by modifying and integrating the Spherical Geometry Library and HTM into the system. A third software release during 2017 addressed a series of stability and performance issues and the current public version is v1.3. Since last year, development was focusing on integrating the Footprint Service with SkyQuery to allow direct access to survey sky coverage information and Virtual Observatory integration by developing a VO standard plugin that allows access to external databases. These improvements to the system are in the final stage but not yet ready to be made public. A future release scheduled by end of January 2019 will likely contain footprint integration and VO integration will come at a later time.

The source code of SkyQuery consists of several repositories openly available on GitHub. The two largest ones are the Graywulf base system which consists of 235,000 lines of code and the SkyQuery system which amounts to 50,000 lines of code. The total codebase, including the footprint libraries and SciServer integration plug-ins, mostly developed during the last three years, is roughly 500,000 lines[1]. URLs to the repositories are provided in a later section.

Within the scope of this project, based on the code primarily developed for SkyQuery, in a joint project with Konkoly Observatory and ESA, we processed the telescope pointing data of the Herschel Space Observatory to recover the exact sky coverage of the entire observation program of the satellite. Reconstruction of the footprints was far from trivial, partly due to the many observation modes of the

[11] Code contribution can be retreived from github for each related repository (see [24] for the list of submodules), for example for the Graywulf parallel database platform it is at https://github.com/sciserver/graywulf/graphs/contributors. As a comparison, the Illustris cosmological simulation team claims to work on a codebase of 200,000 lines.

satellite, partly the geometric complexity of the sky coverage. The results of the processing are available at our web service [31, 32] and was also incorporated into ESA's own Herschel archive [33]. We published the details of the project in [1] and [3]. The project comprised a section of a PhD thesis [3].

Significant work was done in the field of photometric redshift estimation. Using our own Local Linear Regression method, combined with template fitting for K-correction and a new deep spectroscopic training set, we compiled the "official" photo-z catalog for the Sloan Digital Sky Survey DR12 [7]. We also developed a new technique to generate realistic emission line galaxy model spectra to match photometric observations with the purpose of photo-z quality assessment [6, 20, 23, 26]. This sub-project has resulted in a PhD thesis [11] and a BSc thesis [20] so far. This is a major ongoing work that we will continue in the recently awarded OTKA project NN 129148.

With collaboration of our group and István Szapudi at the University of Hawaii, we developed a numerical method and implemented related software to run n-body simulations of structure evolution of the Universe in a spherically compactified setting [13, 16, 17, 21]. This type of simulation can yield the large-scale correlation function with equal signal-to noise ratio as on smaller scales while decreases the computational requirements significantly that arise in standard simulations from short-scale interactions on small scales. The software is implemented to run a cluster of GPUs but can also yield good results quickly on single GPU setups, consequently will be a useful tool to interpret the results of ongoing cosmological observations. This work will also be continued in OTKA NN 129148.

The project resulted in 11 peer-reviewed journal papers closely related to the main topics and several other papers in different fields by project participants with about 80 citations altogether. We presented at least 8 posters at various conferences and the core group is working on at least two journals papers at the moment. SkyQuery itself has at least 1200 registered users and about 4200 cross-match queries have been run.

Participants and Tasks

Most of the benefits in terms of stipends and salaries 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.

Evelin Bányai, research fellow and PhD student, was employed on the project for two years and worked on SkyQuery as an astronomer, software engineer and database developer. She contributed significantly to the new version of the Footprint Service (yet to be made public on-line) and worked extensively on curating, annotating and converting major astronomical databases to the SkyQuery format. As a result of the knowledge and skills she acquired during the project, she left to work at Konkoly Observatory as a system administrator lead.

Tamás Hajdu, research fellow and PhD student, was employed on the project as an astronomer and software test developer. He worked on client tools for SkyQuery and the Footprint Service and the Virtual Observatory interfaces to be integrated into SkyQuery.

Róbert Beck, PhD student, did a significant work on improving photometric redshift estimation for the Sloan Digital Sky Survey. He compiled a new training set of reliable high redshift galaxies and implemented

the Local Linear Regression method, an empirical photo-z technique capable of providing reliable error estimates on photo-z. While spending a year at Johns Hopkins University, he also developed a software package to run template-based photo-z inside a SQL database server. This tool will later be integrated into SkyQuery to allow spectral energy distribution fitting as a follow-up to catalog cross-matching. To improve and assess photo-z methods, Róbert developed a technique to generate realistic emission lines on top of synthetic galaxy spectrum stellar continua. Róbert defended his PhD thesis in 2017 and currently holds a post-doctoral position at the University of Hawaii.

János Márk Szalai-Gindl, PhD student in computer science, was working on the theoretical aspects of large multidimensional scientific databases and astronomical data modeling. He conducted research on the field of distributed point cloud databases that can be the basis of large data warehouses aimed to be the foundations of future statistical and machine learning application. In collaboration with the team at Johns Hopkins and other US universities, he developed and implemented a novel Bayesian approach to estimate the parameters of the galaxy luminosity function and implemented a Monte Carlo sampler to solve the problem using GPUs. He is expected to defend his thesis in 2019 based on the work he did within the scope of this project. As a result of the expertise he acquired working in our group, he recently obtained a position of assistant lecturer at the Faculty of Informatics, Eötvös University.

Gábor Rácz, PhD student, was doing an excellent work on novel cosmological simulations. He developed the toy model AveRA and related software to simulate a non-uniformly expanding universe. Although not proven rigorously in the framework of general relativity, results from this simulation could explain the acceleration effect known as dark energy. His work was very extensively reviewed throughout the world, see for example [34-39]. His recent work was focusing on the STePS cosmological simulation of the infinite Universe. Gábor published both theoretical and technical papers on his work and is expected to defend his PhD thesis in 2019.

Géza Csörnyei, BSc-MSc student, started working on the project as a BSc student continuing the work of Róbert Beck on photo-z of strong emission line galaxies. Géza further developed the recipe introduced by Róbert to generate realistic mock catalogs of emission line galaxies with the purpose of photo-z performance assessment. With his work, he won a 3rd prize at the student scientific conference (TDK) of the Faculty of Science and defended his BSc thesis.

István Csabai, senior participant, was responsible for managing the work, conduct and co-advise project related BSc, MSc and PhD thesis work and establish new interdisciplinary collaborations where our framework can be used. He also worked on disseminating the results (several public lectures at various universities). István was very active in supervising students Dezső Ribli (PhD) and Bálint Pataki (MSc) in the field of deep neural networks. Their recent work on estimating cosmological properties from simulated weak lensing maps [24] was published in Nature Astronomy and revealed that the information content of the maps is way beyond what peak counting can recover and suggested a new way of performing the statistics.

László Dobos, PI, was mostly working on the scientific, software development and database engineering tasks of SkyQuery. During the span of the project, László spent several months at the partner institution to work with the partner team at Johns Hopkins in tight collaboration. He also worked on managing the project and advise and co-advise project-related BSc, MSc and Phd thesis work [5, 11, 20].

Publications closely related to the project

[1] Verebélyi, Erika; Dobos, László; Kiss, Csaba: Footprint Database and web services for the Herschel space observatory, 2015 in Proceedings of the IAU General Assembly, Meeting #29, id.2236977

[2] Laszlo Dobos, Tamas Budavari, Evelin Banyai, Istvan Csabai and Alexander Szalay: SKYQUERY: A PARALLEL DATABASE PLATFORM FOR ON-DEMAND CROSS-MATCHING, 2016 in Proceedings of the ESA Conference on Big Data from Space (BIDS)

[3] Erika Verebelyi, Laszlo Dobos, Eva Verdugo, David Teyssier, Katrina Exter, Ivan Valtchanov and Csaba Kiss: A FOOTPRINT DATABASE OF THE HERSCHEL SPACE OBSERVATORY, 2016 in Proceedings of the ESA Conference on Big Data from Space (BIDS)

[4] Dobos, László; Varga-Verebélyi, Erika; Verdugo, Eva; Teyssier, David; Exter, Katrina; Valtchanov, Ivan; Budavári, Tamás; Kiss, Csaba: The Footprint Database and Web Services of the Herschel Space Observatory, 2016 in Experimental Astronomy, Volume 42, Issue 2, pp.139-164

[5] Varga-Verebélyi Erika: Hideg galaktikus molekulafelhők vizsgálata, 2016 PhD thesis, ELTE, 7. fejezet

[6] Beck, Róbert; Dobos, László; Yip, Ching-Wa; Szalay, Alexander S.; Csabai, István: Quantifying correlations between galaxy emission lines and stellar continua, 2016 in Monthly Notices of the Royal Astronomical Society, Volume 457, Issue 1, p.362-374

[7] Beck, Róbert; Dobos, László; Budavári, Tamás; Szalay, Alexander S.; Csabai, István: Photometric redshifts for the SDSS Data Release 12, 2016 in Monthly Notices of the Royal Astronomical Society, Volume 460, Issue 2, pp. 1371-1381

[8] Bagoly, Zsolt; Szécsi, Dorottya; Balázs, Lajos G.; Csabai, István; Horváth, István; Dobos, László; Lichtenberger, János; Tóth, L. Viktor: Searching for electromagnetic counterpart of LIGO gravitational waves in the Fermi GBM data with ADWO, 2016 in Astronomy & Astrophysics, Volume 593, id.L10, 4 pp.

[9] Plachy, E.; Molnar, L.; Szabo, R.; Kolenberg, K.; Banyai, E: Target selection of classical pulsating variables for space-based photometry, 2016 in Communications from the Konkoly Observatory, Vol. 105, p. 19-22

[10] Beck, R.; Dobos, L.; Budavári, T.; Szalay, A. S.; Csabai, I.: Photo-z-SQL: Integrated, flexible photometric redshift computation in a database, 2017 in Astronomy and Computing, Volume 19, p. 34-44.

[11] Beck, Róbert: Empirical and spectral template-based approaches in the analysis of galaxy data, 2017 PhD thesis, ELTE

[12] János M Szalai-Gindl, László Dobos, István Csabai: Tiling Strategies for Distributed Point Cloud Databases, 2017 in Proceedings of the 29th International Conference on Scientific and Statistical Database Management pp. 32

[13] Rácz, Gábor; Dobos, László; Beck, Róbert; Szapudi, István; Csabai, István: Concordance cosmology without dark energy, 2017 in Monthly Notices of the Royal Astronomical Society: Letters, Volume 469, Issue 1, p.L1-L5

[14] Beck, R., et al.: On the realistic validation of photometric redshifts, 2017 in Monthly Notices of the Royal Astronomical Society Volume 468 Issue 4 pp.4323-4339.

[15] István Csabai, László Dobos, Attila Kiss and János M. Szalai-Gindl: Some Mathematical Properties of the Performance Measures Applied for Point Cloud Databases, 2018 in Annales Universitatis Scientiarum Budapestinensis de Rolando Eötvös Nominatae, Sectio Computatorica, Volume 47, pp. 197–209

[16] G Rácz, I Szapudi, I Csabai, L Dobos: Compactified Cosmological Simulations of the Infinite Universe, 2018 in Monthly Notices of the Royal Astronomical Society, Volume 477, Issue 2, p.1949-1957

[17] R Beck, I Csabai, G Rácz, I Szapudi: The integrated Sachs-Wolfe effect in the AvERA cosmology, 2018 in Monthly Notices of the Royal Astronomical Society, Volume 479, Issue 3, p.3582-3591

[18] G Dálya, G Galgóczi, L Dobos, Z Frei, I S Heng, R Macas, C Messenger, P Raffai, R S. de Souza: GLADE: A Galaxy Catalogue for Multi-Messenger Searches in the Advanced Gravitational-Wave Detector Era, 2018 in Monthly Notices of the Royal Astronomical Society, Volume 279, Issue 2, pp. 2374-2381

[19] János M. Szalai-Gindl, Tamás Budavári, Thomas J. Loredo, Brandon C. Kelly, István Csabai, László Dobos: Hierarchical Bayesian Method for Estimating Luminosity Function, 2018 accepted for publication in Astronomy and Computing

[20] G Csörnyei: Fotometriai vöröseltolódás-becslések pontosítása, 2018, BSc thesis at Eötvös University

[21] G Rácz, L Dobos, I Szapudi, I Csabai: Multi-GPU simulations of the infinite Universe with STEreographically Projected cosmological Simulations, 2018 talk presented at the Wigner GPU Day.

[22] L Dobos, T Budavári, E Bányai, T Hajdu, Alexander S. Szalay: SkyQuery: a web service for fast cross-matching of the largest astronomical catalogs, 2018 poster presented at Division B of IAU General Assembly, Vienna.

[23] G Csörnyei, L Dobos: Characterizing the effect of emission lines on photometric redshift estimation, 2018 poster presented at Division J of IAU General Assembly, Vienna.

[24] Ribli, Dezső, Bálint Ármin Pataki, and István Csabai: An improved cosmological parameter inference scheme motivated by deep learning, 2018 in Nature Astronomy doi:10.1038/s41550-018-0596-8

[25] G Rácz, I Szapudi, L Dobos, I Csabai, A Szalay: StePS: A Multi-GPU Cosmological N-body Code for Compactified Simulations, 2018 in preparation

[26] L Dobos, G Csörnyei: The effect of emission lines on the performance of photometric redshift estimation algorithms, 2018 in preparation

URLs of software source code repositories and web pages accompanying papers

[27] https://github.com/sciserver/skyquery-all

[28] https://github.com/sciserver/footprint-all

[29] https://github.com/eltevo/avera

[30] https://github.com/eltevo/StePS

[31] https://github.com/eltevo/HerschelDB

[32] http://herschel.vo.elte.hu

[32] http://www.cosmos.esa.int/web/herschel/science-archive,

[33] https://github.com/beckrob/Photo-z-SQL

References to our work in the media

[34] https://www.csillagaszat.hu/hirek/sotet-energia-nelkul-is-megertheto-az-univerzum-gyorsulo-tagulasa/

[35] http://www.sciencemag.org/news/2017/04/dark-energy-illusion

[35] https://phys.org/news/2017-03-expansion-universe-dark-energy.html

[36] https://www.sciencedaily.com/releases/2017/03/170330115254.htm

[37] https://www.dailymail.co.uk/sciencetech/article-4368950/Simulation-suggests-68-universe-not-exist.html

[38] https://www.iflscience.com/space/dark-energy-may-not-actually-exist/

[39] https://www.elespanol.com/ciencia/salud/20170410/207479687_0.html