Document version 1.0 (July 25, 2017)

Prague, Czech Republic, June 30 & July 1, 2017

Subject | Minutes of the WG 1&2 Meeting of COST Action TD1403 “Big data Era on Sky and Earth Observation (BIG-SKY-EARTH)”

Prague, Czech Republic: EWASS conference venue and the Faculty of Information Technology at the Czech Technical University in Prague

Meeting Agenda

Friday, June 30, 2017 (shared with the EWASS conference)
Chaired by Petr Skoda

09:00-09:20	LSST data products Darko Jevremović - Astronomical Observatory, Belgrade, Serbia
09:20-09:40	Big Data in Space- Big data in our Computers Edwin A. Valentijn - Kapteyn Astronomical Institute, Groningen, Netherland
09:40-10:00	Large Scale Data Management of Astronomical Surveys with AstroSpark Karine Zeitouni
10:00-10:20	Challenges of Big Data processing and machine learning in meteor science Dejan Vinkovic - Hipersfera Ltd., Zagreb, Croatia
10:20-10:30	Fully non-linear statistical analysis of Large scale structure data for wide and deep surveys Guilhem Lavaux - IAP / CNRS, Paris, France
	Break (informal discussions)
14:00-14:20	The art of getting science from astronomical data deluge Giuseppe Longo - University Federico Ii , Napoli, Italy
14:20-14:40	Space and cyberspace: hidden patterns in astrophysical datasets Aleksandra Solarz - National Centre for Nuclear Research, Warsaw, Poland
14:40-15:00	The Big Picture from the Bottom Up Ashish Mahabal - California Institute Of Technology, Pasadena, United States
15:00-15:20	Domain adaptation and active learning for SNe photometric classification Emille Ishida - U. Blaise Pascal, France, Clermont Ferrand, France
15:20-15:30	Exploring large spectroscopic surveys using t-SNE reduction of spectral information Gregor Traven - University of Ljubljana, Faculty of Mathematics and Physics, Ljubljana, Slovenia
	break
16:00-16:20	A data-driven probabilistic approach for emission-line galaxy classification Rafael De Souza - Eötvös Loránd University, Budapest, Hungary
16:20-16:40	SUNDIAL: combining astronomy and computer science to understand the formation and evolution of galaxies Johan H. Knapen - Instituto De Astrofisica De Canarias, La Laguna, Spain
16:40-16:50	Photo-Z redshift reconstruction using a constructive multilayer perceptron Engelbert Mephu Nguifo - Limos - Cnrs - University Clermont Auvergne, Aubière, France
16:50-17:00	Deep Learning in Large Astronomical Spectra Archives Ondřej Podsztavek - FIT CVUT, Prague, Czech Republic
17:00-17:10	Light Curves Classifier - Package for obtaining and classifying light curves Martin Vo - Masaryk University, Brno, Czech Republic
17:10-17:20	Search for UV Ceti type stars in astronomical surveys using machine learning methods with Python Jan Okleštěk - Masaryk University, Brno, Czech Republic
17:20-17:30	Conclusions of the first day Petr Skoda

Saturday, July 1, 2017
Chaired by Dejan Vinković

09:00-09:30	Introduction of discussion participants
09:30-10:00	Introduction to BigSkyEarth and its activities
10:00-10:15	Dejan Vinkovic: Airship technology for astronomy and remote sensing
10:15-10:20	Areg Mickaelian: World Data System
10:20-10:30	Christian Muller: Attempt to have a threaded balloon in the stratosphere
	Break
11:00-12:00	Discussion on book and repository activities
12:00-13:00	Summary of the Astroinformatics symposium and discussion on how to harvest this and previous BigSkyEarth (co-)events for book and repository materials
	Break
14:00-16:00	Work/discussion in groups
	Break
16:30-18:00	Summary of work in groups; Joint discussion about the next steps

Discussion about the project initiatives:

- After the participants quickly introduced themselves, Dejan Vinkovic presented a general overview of the Action. This included an overview of currently opened suggestions/ideas for joint project proposals:
  - Boris Antic, ITN proposal “Engaging Mobile Data to Increase Disaster Resilience in Europe”. Boris will restart discussion about the proposal, but it can also lead to an R&D project.
  - Unknown lead, ITN proposal (?): hyperspectral imaging in astronomy and remote sensing: Still missing the lead partner and more concrete focus, but it has attracted a wide interest. This project needs a new round of discussion.
  - Johan Knapen proposed we work together on a RISE proposal. This could be a natural follow-up on BigSkyEarth, but focused on well-defined R&D problems!
    Proposals should include at least three partners, which can be universities, research institutions, or non-academic organisations. Small and medium-sized enterprises (SMEs) are encouraged to participate. Partner organisations should be from three different countries. Proposals should highlight networking opportunities, sharing of knowledge and the skills development of staff members. The grant supports the secondment of staff members for one month to one year. They must be engaged in or linked to research and innovation activities for at least six months prior to the secondment. Funding for a RISE project can last up to four years.
  - Petr Baumann (from Rasdaman) suggested at the workshop in Sporon a cross-domain (Earth Observation + Sky Surveys + Planetary Observations) project on fast, flexible, scalable engine for data-intensive computations on spatio-temporal datacubes. The project would implement a standardisation on innovative data access interfaces across domains that need massive array processing. H2020 calls would be targeted.
    Possible work packages are something like: “Knowledge and innovation exchange on data infrastructure for…”, “Knowledge and innovation exchange on data- mining/machine- learning for…”, “Knowledge exchange on modeling of…”, “Training, sharing and dissemination of knowledge”, “Management, Communication”
  - Jouni Peltoniemi: he is interested in leading a project on Big Data hyperspectral remote sensing in forestry (and possibly agriculture) using drones and other aerial remote sensing platforms. Dejan said he is interested to join through his SME Hipersfera and possibly bring some other SMEs into the project (from a hyperspectral camera manufacturer to wood processing companies. Uros Kostic is also interested.
  - There is an open call by ESA for contributions to their Stratospheric High Altitude Pseudo-Satellites (HAPS) programme for Earth Observation, Telecommunications and Navigation. They organize a conference in October to discuss possible science cases and HAPS designs. There has been discussion within BigSkyEarth on possible science cases for a stratospheric research platform for astronomy, aeronomy and remote sensing. Hipersfera (Dejan Vinkovic) will participate at the conference with a suggestion for an airship design to reach stratospheric heights and play the role of HAPS. Everyone is invited to join in to this effort.

- Victor Debattista explains that his current professorship includes work on science outreach. He suggests we could also put together a project in that area, too, combining attractive visualizations in astronomy (e.g. GAIA data on Milky Way structure) and Earth Observations

- Giuseppe Longo suggests the “FET-Open research and innovation actions” H2020 call as a good target for spatio-temporal databases as a technology. The call has no limits in project topics, the ideas have to be innovative and with envision impact in the next 5 years, all costs (people, infrastructure) are funded, and the deadline is in September.
- Areg Mickaelian is promoting a possible collaboration with the World Data System - an interdisciplinary body of the International Council for Science. Their goal is to create trustworthy data services for global science through quality assurance, long-term stewardship, standardisations and compliances.
- Christian Muller described efforts from the 1970’s to deploy tethered stratospheric balloons. Dejan described plans by Hiperfsera company to deploy tethered mid&lower-troposphere airships and have tethered balloons up to a few 100s of meteors above the ground to continuously collected meteo-data along the entire vertical air column.

Discussion about the book and repository:

- Dejan presented results of a small survey that registered participants could fill out prior to the meeting. The question was: “Are you interested in contributing to one or both of the following book activities: OPTION 1: A research book on BigSkyEarth topics (knowledge discovery in Big Data in Earth and Sky observations) covering areas of: theory, programming, algorithms, learning, etc. OPTION 2: An online repository of training material combining short theoretical introduction and practical code snippets (and maybe a book based on it later on), OPTION 3: None (I am interested only in research networking and possible joint research projects)”. Out of 35 responses, the distribution is as follows:
  OPTION 1: 26, OPTION 2: 21, OPTION 1&2: 14

- Dejan then presented a brief overview of the conclusions reached at the last meeting regarding the initiative of working together on a book and repository:

  About the book preparations:
  - Decide if the book is going to a) include teaching material and exercises to qualify as a curse book or b) focus on purely research topics
  - Decide if the book is going to be a part of some book series (e.g. a special issue of Big Data Research by Elsevier)
  - MC members who volunteered to be Editors: Petr Skoda, Peter Butka, Areg Mickaelian. One more is needed to cover remote sensing. The Editors now need to guide the preparations for the book writing.
  - There will be an open call for contributions, but the editors need to converge toward the procedure and book concept. Meeting in Prague is an opportunity to discuss this.

  About the training repository and possible book based on it:
  - Python would be the main language used in the repository, including additions like numba, cython, datashader, xarray dask, pandas, etc.
  - Some contributions could be connected into a sequence suitable for university courses
  - The repository can be published on www.zenodo.org to get a DOI number for each submitted contribution, with additional links to code examples on Github.
  - Possible chapters include:
    - Preparation/pre-processing/handling of Big Data (what to be aware of, columns-vs-rows-vs-unstructured, binning and re-binning, missing values, etc.)
    - De-nosing and feature extraction
    - Feature analysis, simplification and synthesis
    - Simulations of stochastic processes
    - Matching data of different structures
    - Machine learning
    - Visualization
    - Data processing architectures and database structures

    - GPGPU implementation
    - applied statistics and modelling in astronomy
  - The repository can be split into thematic chapters and we would need an editor for each chapter
  - volunteers so far for editing: Marco Qurtulli, Victor Debattista, Dejan Vinkovic, Peter Butka

- Rafael De Souza described his experience on collaborative work with repositories and that was done in bursts of activities, with a senior researcher overviewing what to do and what not to do. The materials evolved into publishable papers. People meet in person to give ideas on topics to be covered and then they worked for a week. This meeting in person was important. He suggests that we need a few people in the beginning who will start the movement and then have open calls after something specific is built by about 5 people. A book can be a final goal for some of the material developed within the repository. He emphasized that lots of useful material young researchers put on their blogs because they have a desire to share their knowledge and experience with the community. Some prolific blog authors could be invited to contribute to this repository effort, which would give them a chance to formalize their so far informal blog material.
- Giuseppe Longo suggests that the book refers to the repository, such that all the teaching materials can be prepared in the repository while the book keeps a more scientific content. Since preparing the book will take about 2 years, it is hard to make it useful if it talks about the technological aspects that change faster than that. The repository can be updated appropriately to the changes in technological trends, which does not affect the book if it covers a general introduction into the applications of methods, while the specific implementations of those methods are put into the repository.
- Ashish Mahabal seconds the Giuseppe's suggestion and suggests the repository built on the book articles. This way the repository can show how methods are connected to different fields. This way the book covers a scientific topic that stays relevant for more than 2 years, while the repository has practical applications connected to the scientific method.
- Emille Ishida emphasises the need for having explanations on how different disciplines call the same method (e.g. regression). It is about semantics, but this is very important for people who want to understand utilization of methods in different fields.
- Emille also suggests that the repository needs to cover the same examples in at least python and R, or maybe some other languages, too. The examples should not be complicated.
- Ashish thinks that pseudocodes could be presented in the book, such that they are then translated to whatever programming language.
- Petr Skoda gives an example – the book can have “classification” as a chapter and then the repository covers this with using astropy, remote sensing libraries, and from the computer science perspective
- Emille thinks young people need to be involved in the repository development and their work needs to be recognized. But how to motivate people to contribute and how to make people to use it? She thinks about 10 students are needed who will be very active.
- Adam Fathalrahman asks what is the advantage of using zenodo over GitHub?
- Giuseppe says zenodo provides recognition through the ability to be cited (thanks to the DOI number), but he points out that this becomes useful only if the repository becomes viral (i.e. widely used)
- Ashish warns that we cannot force people to publish on zenodo, but we have to convince them it is in their interest to use it – to have searchable contributions in literature databases and then become cited
- Petr emphasises that often participation in books is counted/scored by funding agencies or institutions for promotions
- Petr discussed with Springer a possibility to publish a book with them. They like the idea of interdisciplinarity and a book of about 10 to 15 chapters. General suggestions are:
  - 2-3 editors of the whole book/volume (they communicate with the main editor in Springer)
  - So far we have volunteers for editors from astronomy: we need a geophysics editor, too
  - Those editors have to deliver the book on time
  - Chapter have authors (from one to many)
  - A general advice is to overshoot the number of chapters because about 1/3 will drop out (not be finished on time). This means at least 15 chapters initially for us.
  - Books are freely available to people at institutions that have subscription with Springer
  - An attractive overall topic is Astro-Geo-Informatics (geophysics, astronomy, remote sensing)
- Aleksandra Nina volunteers for the geo-editor and suggests that she can help write (with some co-authors) a chapter on ionospheric studies (relevant in astronomy, planetary science, remote sensing, aeronomy, geophyiscs)

- Emille suggests we pick e.g. 15 the most important methods/issues/things for people who work in those areas as chapters
- Areg Mickaelian asks what is the general focus of the book. Ognyan Kounchev suggests the title “Big Data in Astro-Geo-Informatics”, with subtitle “Knowledge Discovery in Earth and Sky Observations”. Petr suggests the title “Big Data Knowledge Discovery in Earth and Sky Observations”.
- Areg suggests we have an open call for chapter contributions. The call should ask for the chapter title, a small summary and a list of proposed authors. This call can be opened for one month, but before that we should ask invited speakers to give suggestions for the first topics and then in the second wave of solicitation the call is sent to the entire BigSkyEarth community and beyond
- Gottfried Schwarz asks if climate research is included into our consideration. If it is then he has some authors to suggest. The problem is that sometimes events like the Big Data from Space conference do not include topics like climate research, atmospheric pollution, gravity research, etc.
- Dejan points out that all these contributions are welcomed if we demonstrate how some method is used in their situations.
- Petr emphasises that we need a clear background and purpose for the book. He thinks that interdisciplinarity poses a problem for cross-field communication and method usage/exchange. Therefore, our book can benefit the community.
- Peter Butka asks if we want to invite people from industry to give talks at our next meeting (in Bulgaria) and help us with preparing the book chapters with the industry point of view.
- Nima Sedaghat says the book also needs examples from computer science. He can contribute in writing about machine learning in computers science (application on image recognition), since its application is not the same in computer science, astronomy and geoscience.
- Emille says she can also help with machine learning contributions. She also warns that chapters should be names after applications, not methods.
- Christian Muller says he can help with topics in operational meteorology, neutral atmosphere, and climate topics
- Gottfried says he can help with discussion on difference between correlation and causation
- Kathrine Zeitouni says she can help with a topic on optimization of databases for faster access in geo-astro Big Data applications
- Areg can help write an introduction on Big Data in astronomy and geoscience + examples of codes that enable access to these data
- Gottfried asks what about simulated data – how to exploit it and combine it with observations?
- Areg warns that we need to show the benefit of astro- and geo-community working together.
- Gottfried suggests we also cover the issue of error propagation
- Ognyan emphasizes the importance of covering the topic of ionosphere research (as mentioned before by Aleksandra), because both communities (geo and astro) work together to understand it better, while it also has very practical daily implications in telecommunications and GPS signal propagation (error)
- Atanas Hristov points out that we could have a chapter from the computer science view on programming methods for efficient ways to utilize databases. E.g. streaming processing to keep up with the processing time.
- Kathrine suggests (see also Emille’s suggestion before) that we need a text on differences in data produces between astronomy and geoscience (there are differences in, e.g., error propagation, validation, calibration modalities, etc.)
- Petr and Victor Debattista suggest we also have a chapter on visualization – form general theory to repository examples of Big Data visualization.
- Dejan suggests that the repository goes beyond just this book as it can have its own additional storylines useful as teaching materials
- Engelbert Mephu Nguifo suggests we also ask for contributions to the question on what are the problems in astro-geo that still need to be solved with the machine learning techniques. This is a sort of “New Challenges” chapter.