European Cooperation in the field of Scientific and Technical Research

- COST -

Brussels, 15 May 2014

COST 055/14

MEMORANDUM OF UNDERSTANDING

Subject :	Memorandum of Understanding for the implementation of a European Concerted Research Action designated as COST Action TD1403: Big Data Era in Sky and Earth Observation (BIG-SKY-EARTH)

Delegations will find attached the Memorandum of Understanding for COST Action TD1403 as approved by the COST Committee of Senior Officials (CSO) at its 190th meeting on 14 May 2014.

MEMORANDUM OF UNDERSTANDING
For the implementation of a European Concerted Research Action designated as

COST Action TD1403
BIG DATA ERA IN SKY AND EARTH OBSERVATION (BIG-SKY-EARTH)

The Parties to this Memorandum of Understanding, declaring their common intention to participate in the concerted Action referred to above and described in the technical Annex to the Memorandum, have reached the following understanding:

1. The Action will be carried out in accordance with the provisions of document COST 4114/13 "COST Action Management" and document COST 4112/13 "Rules for Participation in and Implementation of COST Activities" , or in any new document amending or replacing them, the contents of which the Parties are fully aware of.

2. The main objective of the Action is to frame the joint long-term agenda, set the stage, incubate new knowledge and defragment the existing knowledge in the fields of geo- and astro-informatics, and disseminate the results.

3. The economic dimension of the activities carried out under the Action has been estimated, on the basis of information available during the planning of the Action, at EUR 56 million in 2014 prices.

4. The Memorandum of Understanding will take effect on being accepted by at least five Parties.

5. The Memorandum of Understanding will remain in force for a period of 4 years, calculated from the date of the first meeting of the Management Committee, unless the duration of the Action is modified according to the provisions of section *2. Changes to a COST Action* in the document COST 4114/13.

GENERAL FEATURES

Initial Idea:

With the emergence of petabyte-scale astronomy and Earth observation, basic operations like data searching, analytics or visualization have become increasingly difficult and in many cases almost impossible. Simple database queries can now return results so large that they are incomprehensible — slow to handle, extremely hard to analyze and impossible to visualize with available tools. By leveraging similarities in data analytics, streams, collection, distribution and utilization between astronomy and Earth observation, solutions can be sought in close collaboration with computer scientists, exploiting new algorithms and emerging computer technologies, such as general-purpose computing on graphics processing units, large scale distributed filesystems and parallel processing frameworks. This Action will help identify common issues and cluster emergent solutions in the form of methodologies and tools from both research and industrial environments in astronomy, Earth observation and Big Data computer science. The Action results obtained by the multidisciplinary expert network, in a framework of common approaches towards simplified large scale data management and analysis, will promote techniques that are emerging as critical in the era of Big Data, such as the usage of data mining and statistics to discover new physics hidden in the data, unlike the traditional approach where the data are used to validate a priori defined models or theories. The Action will also help training a new generation of professionals capable to deal with the new technologies in an effective way. The Action's impact is ensured through raising the ability of participating parties in techniques developed and learned as a result of this Action.

Keywords: astronomy, Earth observations, remote sensing, Big Data, visualization, visual analytics, astroinformatics, geoinformatics

A. CHALLENGE

Big Data - defined as datasets that are too large to be processed with today's tools and methods - permeates our world. In 2012, about 2.5 exabytes were created each day, and that number is expected to double every 40 months. Astronomy and Earth sciences have also entered this new era. Projected data volumes for next generation surveys are so large that our ability to store and index the data, and efficiently extract knowledge from it, represents a fundamental limitation on our

ability to do science. Yet this is not an isolated problem: the challenges awaiting explorers in astronomy and Earth observation match fundamental challenges as identified from the Information Technology (IT) perspective.

With the current emergence of Terabyte(TB)- and very soon Petabyte(PB)-scale astronomical and Earth observation systems, the traditional approach has begun to fail spectacularly. Basic functions such as data searching, analytics or visualization are becoming increasingly difficult to handle. Simple database queries can result now in data subsets so large that they are incomprehensible, slow (or even impossible) to handle, and impossible to visualize with commodity visualization tools. Domains such as high energy physics or bioinformatics deal with observing artificial experiments with well modelled, stable instruments, and this has eased their transition into the era of Big Data. In contrast, astronomy and Earth observation handle highly heterogeneous data (different observatories producing multi-wavelength and multi-epoch data, airborne and space-borne sensors, application-specific heritage maps) and are concerned with a myriad of different, specific, individual “user defined” problems: in this respect, they can be considered as pathfinders for many other sciences.
In astronomy, many scientific databases, such as those derived from data obtained by the Sloan Digital Sky Survey, are already many tens of TB in size. In the next five years, the Panoramic Survey Telescope and Rapid Response System (Pan-STARRS) is expected to produce a scientific database of more than 100TB of data. The GAIA space mission, with the largest digital camera ever built for a space mission to create the three dimensional chart of our Galaxy, is expected to produce one PB of data through five years of exploration. The Large Synoptic Survey Telescope project will obtain 30TB of imaging data each night in the optical range and will operate for ten years, while in the radio regime the Low Frequency Array aims at producing PB of data per year, with each file (image or other data product) reaching TB sizes. Many other planned space and ground observatories will have a capacity to reach PBs during their operational lifetime. These new facilities will open the study of the transient sky, through high cadence observations of the sky, producing large and complex data. Pan-STARRS, the Catalina Real-Time Transient Survey and the Palomar Transient Factory are examples of current large scale surveys that explore transient phenomena and prepare the ground for even larger surveys. Efficient access to all these scattered datasets is a significant issue in astronomy, which prompted astronomers to work on the Virtual Observatory (VO, www.ivoa.net). VO provides standards describing all astronomical resources worldwide and supports standardized discovery and access to these collections, as well as powerful tools for scientific analysis and visualisation.

Within the field of remote sensing and Earth observation, Big Data has emerged as an important challenge that has been fully recognized by the players in the field – from space agencies to space industry to the science community. On the one hand, there is a range of technical issues such as data rates, storage capacity and processing throughput. On the other hand, the remote sensing community has recognized the need for new concepts and approaches to handle, manage, and understand Big Data. For example, the Global Earth Observation System of Systems is projected to be producing 1PB per day; the ESA Sentinel-1 two-satellite constellation will by itself generate up to 4.4TByte per day of compressed raw data; ESA Sentinel-2 requires a ground segment with a throughput of 20Gbytes per second. Large archives are being set up while existing data lies under-exploited: 96% of the acquired images having been never seen by a human. Thus, Big Data is a challenge that must be met by innovative data representation within databases that allow rapid content-oriented access to datasets as well as by multi-dimensional visualization of internal relationships. These relationships and their dependencies have to be explored interactively within new domains. Typical examples are data mining in feature and cluster spaces or the use of metadata descriptors for semantic annotation and search. There is a need for innovative and robust tools that allow users to browse large data repositories. This will require the development of new techniques in (visual) data mining, efficient databases, knowledge discovery, and (visual) data analytics.

Measures are urgently needed to make sure that the scientific community continues with delivering its societal and knowledge benefits even in the data domain that we have already reached. Astronomy and remote sensing complement each other in that sense, as they are on the quest for new Big Data interpretation capabilities: both disciplines have peculiar data, typical data processing and analysis chains, and specific models to be fed with data. However, both disciplines lack the capabilities for easily accessible semantics-oriented browsing (usage of higher level descriptive expressions) in large data archives. Therefore, joint efforts to design and develop innovative Big Data tools should help users in many different fields and set new standards for many communities. This has identified several broad challenges to this line of reasoning that need multidisciplinary approach through international networking of experts and professionals. These challenges are then channelled into Action Objectives.

Challenge A: Digital curation and data access

The growth of data availability and program-based queries has driven up the usage of the data archives. This trend will accelerate as new data come in, attracting more users who query large

volumes of data. The consequences are already visible in the slower response times to queries. Preventing escalation of this problem is not possible by simply adding infrastructure as usage increases, due to conditions of limited project budgets and the current practice of analysing downloaded data on users' local machines. A solution has to be sought in advances in database management systems (DBMS), as well as in transforming "the design and operation of archives as places that not only make data accessible to users, but also support in situ processing of these data with the end users' software". From the users' point of view, Big Data brings challenges on algorithms and knowledge-extraction methods, resulting in the fact that those methods require to be upgraded or redesigned. The problem escalates when a user wants to compare data from different repositories. Large observational facilities often host specialised data centres, which collect data produced by their specific mission or instrument. Most of the interesting research is based on and motivated by the possibility to correlate information from one mission with information from other surveys. Transferring large volumes of data between facilities whenever analysis of both sets is requested is becoming inoperable, thus the software has to be moved to the data. This is not a critical issue in other disciplines (such as high energy physics) that do not work with multiple comparisons of such a large volume of diverse data, but in astronomy and Earth observation, numerous users use different combinations of the same datasets with different algorithms and for different purposes. This requires a study of the interoperability standards between data archives.

Challenge B: New frontiers in visualization

Scientific visualization is a very active research field because it provides powerful means of making sense of data. Human cognition critically depends on visual stimuli, hence visualization maps digital data attributes to visual properties such as position, size, shape, movement, and colour. This helps users discern and interpret patterns within data. But identifying relevant data is difficult in the Big Data domain. The necessity to visualize large data flows in a precise and interactive way is often a research imperative. Moreover, astronomical and remote sensing datasets have special characteristics that make visualization difficult: various data formats, often customized to the needs of a specific research subfield; low signal-to-noise ratio and high dynamic range; multidimensional parameter space not suitable for typical 2D/3D visualization tools on the market; massive datasets where typical visualization methods fail. The new visualization tools should be developed jointly by scientists and visualization experts, but this requires funding and appropriate academic valorisation within astronomical and Earth observation communities. Commercialization of remote sensing data gives some impetus for such visualization explorations, but this is not so in astronomy.
Surprisingly, even though astronomy is a very much visual science, the use of advanced computer

based scientific visualization in astronomy is still a young field. Hence, some tools developed for Earth data visualization might be adapted to the astronomical data (geo-coordinates replaced by sky-coordinates). Google Sky for astronomical images based on Google Maps backend tools is a popular example. Overall, the era of Big Data is a challenge for both disciplines and it requires a close collaboration with experts in computer graphics to develop visualization tools adequate for Big Data exploration.

Challenge C: Adaptation to new high performance computing (HPC) technologies

The importance of parallel computing is on the rise. Astronomers and remote sensing scientists have relied to Moore's Law for a long time and benefited greatly from it. Their codes required minimal maintenance as CPUs were doubling their speed every two years on average. This worked well for a design of complex software for astronomical and remote sensing facilities, which often take more than a decade to become operational, where future computational technology developments have to be anticipated far in advance. About eight years ago, the increase in CPU speed due to miniaturization reached its physical limits caused by energy consumption and heat dissipation issues. Instead, CPUs became “multicore”, with the increasing number of cores soon in the three-figure domain for off-the-shelf products (as for the Intel Phi architecture). On the other side, currently available GPUs on the market already have hundreds, and soon thousands, of cores, where most of the silicon is devoted to numerical operations. This property, combined with low cost, pushed GPUs toward gradually becoming the mainstream accelerators in the heterogeneous supercomputing environment. They are particularly well suited to many of the scientific and engineering computational workloads. When used in cluster environments, GPUs also have the potential of significantly reducing the overall space, power, and cooling demands. GPUs are today treated basically as numerical coprocessors, with a technological trend of becoming fused with CPUs on the same wafer in the future. Not surprisingly, even though utilization of GPUs requires new programming techniques, their application to astronomical and remote sensing problems has already started and it is continuously growing. The interest is growing also for astronomy in cloud and remote sensing in cloud, although not as extensively as for GPUs. Harvesting all these new HPC technologies requires invention of a new processing and algorithm coding paradigm. Not all problems are suitable for porting to these new technologies, but orientation toward implementation of massively parallel processors (locally or in cloud) in knowledge extraction from large astronomical and planetary (including Earth) databases seems in fact inevitable. This trend is slowed down by the fact that this is a fast evolving technology which induces a lack of stability to develop durable solutions. Nonetheless, opportunities are especially abundant in the case of

complex image-access queries requiring sweeps through the databases to perform pattern matching. Such algorithms demand orders of magnitude higher computational capabilities to sustain real-time performance. While it is clear that not all data mining algorithms will benefit from massive parallelisation, it is already obvious that some of the most common will greatly benefit from them (such as Genetic algorithms, nearest neighbour, Decision trees, etc).

Challenge D: New generation of scientists in the age of interdisciplinarity

All these trends in modern astronomy and Earth observation make those fields increasingly dependent on computer science. Actually, so dependent that often they have to be considered as a new interdisciplinary fields of research - astroinformatics and geoinformatics. Astroinformatics can be formally defined as the formalization of data-intensive astronomy and astrophysics for research and education, with data mining as one of the core areas. Geoinformatics can be viewed as a scientific discipline designed for handling of geospatial information that encompasses the acquisition, storage and retrieval, modelling, management and analysis, and presentation and visualisation of geo-processes. New research fields imply new type of professionals who are experts in these fields. Hence, we expect professional networking activities within the fields, specialized publications and literature, and university degree programs on this topic. Geoinformatics has already grown to such a level of recognition and maturity, but astroinformatics is still in its infancy - the information is scattered across numerous journals and conference proceedings, and university programs in astroinformatics are scarce, with Europe lagging far behind the global trends. These new disciplines challenge the current paradigm of what constitutes astronomical or remote sensing education.

For example, natural scientists would need skills not only to manage and maintain software, but also to develop software that is environment-agnostic and scalable to large data sets. Obviously, a significant effort has to be invested into education of young natural scientists with IT/computer science (CS) professionals. This problem is summarized in the literature as following: "It turns out that all “good examples” were developed either by professional astronomers with very strong IT/CS background or by IT/CS professionals working closely with astronomers for years and understanding astronomy. One cannot simply hire an industrial software engineer to develop astronomical software and/or an archive and/or a database. A possible non-trivial solution is to change the teaching paradigm for students in astronomy. Basic courses in algorithms, programming, software development and maintenance have to be made mandatory in the education of modern astronomers and physicists; advanced courses should be recommended to some of them". But how to make such a significant "cultural change" within natural sciences, when many raise valid

concerns about computer science work defocusing young scientists from their "actual research"? A compromise might be found in setting up on-line courses for interested individuals, who might also see these interdisciplinary fields as a good set of skills sought after by private companies and public administrations. The bottom-line is that whatever the problems have been to date within astroinformatics and geoinformatics, in the Big Data era these problems will expand and seriously jeopardize the big facilities' efforts to utilize PB datasets.

General Objective: Since the identified challenges are similar in astronomy and Earth observations, with computer science as the common denominator, the Action aims at boosting the communication within and between disciplines by identifying and clustering relevant common solutions developed within research and industrial environments. These solutions can be aided by methodologies and tools for large distributed data management and processing, developed by computer scientists in academia or industry. For example, metadata is extensively exploited in multimedia Digital Asset Management to provide effective access to deep repositories of audio-visual content. This approach can contribute a valuable know-how to natural scientists working with similar type of data structures in large databases. Visual Analytics is another example of a growing field in computer science, with interesting implications for astronomy and Earth observation that inherently depend on visual datasets. Therefore, objectives are set in a logical framework where a diverse network of experts first identifies the issues to be addressed, and then tackles the problems via the joint utilization of existing resources, putting emphasis on building bridges between disciplines needed for success, and in disseminating the acquired knowledge, know-how and results to a wider circle of stakeholders. Following this framework the specific Action objectives are:

Objective 1: Framing the Joint Long-term Agenda

Addressing the challenges described above requires a systematic approach, where experts from different disciplines build a framework needed for joint work on comparison and performance assessment of resources used for Big Data environments in each discipline. Hence, the Objective 1 is focused on tacit knowledge of the Action participants to identify, compare and assess the common narrative, methods, techniques and tools used in astro-, geo- and computer sciences. This includes, but is not limited to, discussing performance parameters of data archives, critical evaluation of optimal tools used by users, debating the optimal incorporation of emerging technologies and how to boost interdisciplinary education of young scientists. This Objective sets the stage for activities within other Objectives.

Objective 2: Incubation of New Knowledge

As the outcomes of Objective 1 start to materialize, they will be immediately followed up by activities that aim at incubating new and improved knowledge needed to advance the common narrative, methods, techniques and tools discussed in Objective 1 and develop a new ground for long term collaborations between disciplines on the topic of Big Data science. This step exploits existing tacit and codified knowledge within the Action and, when combined with the accumulated technology knowledge, it leads to the development of solutions to the challenges described above. For example, this can be an implementation of a better DBMS for some data archive, or the standardization of some data communication task across disciplines, or a joint visualization tool or joint education materials. The specifics will depend on the resources readily available to the Action participants at the time and the priority list devised in Objective 1.

Objective 3: Defragmentation of Existing Knowledge

One of the key components of bridging separate fields of science and creating a new interdisciplinary field is to simplify access to the knowledge critical to this field. The knowledge needed for the challenges described above is dispersed between and within the communities of astronomers, Earth observers and computer scientists. This knowledge emerges spontaneously at places where some practical problems have to be solved. Hence, the aim of this Objective is to look at a bigger picture of bridging these disciplines and use international collaboration to defragment and systematize the Big Data knowledge they create. Experts from each discipline will provide their perspective on the importance of certain knowledge, which then can be put into the appropriate context. For example, visualization and image recognitions have a long history in computer science, including the gaming industry, but astronomers and Earth observers have a highly specialized knowledge how to identify objects in their imaging databases based on the specifics of science behind it (e.g. spectral, spatial and morphological properties of stars and galaxies or Earth surface categories).

Objective 4: Dissemination

A common component to all efforts in the Action is to spread tacit knowledge, especially toward Early Stage Researchers (ESRs), and distribute codified knowledge. This objective aims at reaching a larger audience and spreading the acquired knowledge. Objectives 1-3 will create a significant amount of material that will be useful to the larger community of experts in academia and business. However, a considerable effort will be invested into adapting this material to different levels of expertise. For example, some target groups, such as ESRs or experts from different fields, need not

only help with numerical tools, but also some help with the underlying science or a statistical method.

B. ADDED VALUE OF NETWORKING

The challenges described in Part A are inherently global and transdisciplinary. The nature of public Big Data initiatives is that facilities are built and/or maintained by larger collaborations, often spanning a number of countries. However, the challenges that we have identified are of global nature, affecting all such initiatives and have deeper social roots within the research community. Individual projects and initiatives cope with these problems in different ways and typically under multiple constraints - limited budgets with inflexible budget structure, limited access to Big Data experts, lack of understanding of "cultural" changes that Big Data imposes on the infrastructure, project management issues, "on the fly" training of ESRs, etc.

The net result is a diverse set of tools used for data storage, handling, and utilization. The knowledge and hands-on experience is scattered among various groups, who are becoming increasingly aware that we are facing deeper systematic problems that can be resolved on a long term only through a world-wide networking between natural scientists and computer experts.

Hence, it is becoming increasingly important to spread good practices between such different projects. For example, successful initiatives have a lot to say about budgeting Big Data projects and management structure, since their success is largely based on their awareness of the importance of software development and communication between scientists and engineers involved in the project. Or, when it comes to societal changes within the research communities, the Earth observation researchers, who have been promoting geoinformatics for some time now, might have a lot of very useful advice for the emerging community of astroinformatics. Some national initiatives already exist in this respect, where an astro-geo-informatics community spontaneously emerged. However, while being important to leverage funds and to initiate partnerships at the local level, national coordination is not sufficient in this case, when the expertise and resources are spread among several countries around the same field of interest, thus making the international scale the relevant yard for project-wise solution comparisons.

The transdisciplinary aspect of the Action brings benefits to astro- and geo-science communities as they share common IT/CS specific issues like handling the spherical geometry for visualization,

data organization and/or localization, or the heterogeneity of data sources. As mentioned in part A, Earth science has an earlier start of networking with IT/CS scientists than astrophysics and this specific expertise will be an added value of the network. The IT/CS community also has its interests in participating in such a networking effort. For example, the biggest datasets are property of private companies, which makes them difficult to obtain for academic research needed for creative exploration of new ways to exploit Big Data for the public good. The private sector is also a provider of Big Data related issues, but often linked to well-defined problems to be solved, which is not the optimal way of exploring the opportunities opened by Big Data science. On the other hand, astro- and geo-sciences will provide not only datasets open for academic research, but also specific challenges for the IT/CS community: efficient data access for versatile queries coming from research needs, inclusion of measurement errors linked with data properties into the data-mining algorithms, integration of heterogeneous datasets leading to novel data structures, exploration of new algorithms and software architectures for carrying out data analytics, prediction and visualization tasks, etc. It will be possible to test current CS paradigms and if necessary work on development of new ones. Diversity of data sources and data content may require development of so called 'Domain Specific Languages' for a particular sets of problems. The need for such a task will be determined by the interaction of computer scientists with natural scientists. International networking also enables optimization of scientific (hence economic and technological) return from the large public investments made so far in space and ground base astronomical and Earth observation facilities. The interoperability of the data and the possibility to merge information, thus increasing their scientific value and return to the society, adds value and extends the duration of legacy data.

This Action therefore aims at setting the ground for a long-term networking that should lead to spreading good practices and establishing collaborations and joint research proposals that would not be initiated without this Action.

Thus, this networking is expected to last much longer than just the official time span of the Action. The interest in such an approach is visible already from the fact that 32 research groups from 16 countries already expressed their interest at the time of preparing this COST Action, which gives a good head start needed to achieve desired networking effects. Many of those groups are involved, directly or indirectly, in projects aimed at generating and exploiting large datasets. This Action enables them to build a layer of communication, collaboration and knowledge sharing beyond their individual projects, which is typically not covered by their project budgets, especially not in a way

that would bring together a diverse set of experts from different fields to collaborate on resolving challenges on the level of entire research community.

The Acion will actively promote and implement gender balance, in particular through the involvement of female participants in all Action activities. This problem is directly related to a low interest of female students in computer science majors, resulting in a decline of female computer workers within the last two decades. One reason for this could be the negative stereotypes perpetuated by the media, which means that increased efforts to promote positive images of women role models in computer science are much needed. It is interesting to note that computer science shares many similarities with mathematics, but women are highly represented among mathematical workers. Therefore, throughout this Action it will be ensured that women will appropriately be involved and contribute to achieve the Action objectives.

As described above, international transdisciplinary networking is necessary for addressing and documenting various issues in Big Data science from both the natural science and engineering side. The Action will identify and address such issues, among which we can list some examples based on the challenges described in part A:

- What are the hard limits of the existing Big Data back-end tools in astrophysical and geophysical Earth observation sciences with respect to (1) real-world usability/programmability by scientific personnel (2) large scale scalability? How can these tools be improved to ease data access and manipulation?
- What criteria can be used to decide when an archive should adopt technologies such as GPUs or cloud computing? What kinds of technologies are needed to manage distribution of data time, computation-intensive data-access jobs, and end-user processing jobs?
- Approaches exist for large distributed data management and processing that focus either on (1) optimizing data analysis inside scientific databases hosted on dedicated servers, and (2) distributed data analysis in elastic clouds. How to flexibly choose between the two in real-world scenarios? When and how can a transition between the two provide faster results?
- How can metadata layers improve or advance current data mining in astronomy and Earth Observation?
- How can this metadata layers be built with little or no human intervention?
- How has the role of visualization and front-end tools changed in the era of Big Data astronomy and Earth observation? Can approaches be imported from computer science with respect to these issues?

- How can we facilitate the transfer of new technologies in ICT from the computer sciences domain to the scientific domains?
- Rare events are crucial in both astronomy and Earth observation, but what kind of tools is needed for finding rare events or outliers in large multidimensional data cubes?
- How to deal with missing or incomplete information?
- How to efficiently explore temporal data in real time - transient, periodic and irregular events - in the Big Data era when detectors trigger millions of real time events per night for follow up observations?
- How can we form a new generation of scientists working at the crossroads between domain expertise and computer sciences?

To achieve the stated Objectives, the Action will implement goal-oriented Short-Term Scientific Missions (STSMs) allowing researchers from participanting institutions to work on joint tasks and support joint student supervisions. The Action will make best use of COST networking tools to organize meetings, workshops and conferences, organize hands-on training schools targeting mostly ESRs and young experts, and initiate joint activities with other networks. The additional value from the Action comes in a form of outputs such as scientific publications - summarizing the efforts that the Action takes, various documents useful to stakeholders within the expert community, books of abstracts from the events, education and training materials, and a website that will aggregate all the information produced by the Action. Also, the networking effects will be strengthened by organizing the Action's meetings and workshops jointly with (or as satellite events at) some major conferences or events of other networks - e.g. Astronomical Data Analysis Software and Systems conference series, astroinformatics and geoinformatics conferences, IEEE International Geoscience and Remote Sensing Symposium, Image Information Mining conference, Big Data computer science conferences.

Special attention will be given to the Action's efforts to educate a new generation of experts. The natural sciences are faced with a structural problem of computer science essentially becoming a "new math" for future generations of natural scientists. But this fact is not easy to transfer into student training, since a quality education in computer science would reduce the time students spend on learning natural science materials. Another major problem in teaching brand-new and rapidly developing technologies is that good and up-to-speed teachers are very hard to come by. Much more so than in more established areas of engineering or physical sciences, which can perfectly well be taught by traditional university lecturers and professors, while the best expert

teachers in new IT fields cannot be hired widely and should ideally teach students across local and national boundaries. The only viable resolution to these problems at this point is to offer a new approach where interested individuals (not limited to students) can access specific knowledge through online materials and/or at specialized training schools. International networking is a crucial component of this effort, since this knowledge combining natural science and IT/CS is highly specialized and globally fragmented over various institutions and research facilities. Hence, this Action treats educational materials as one of the most valuable outputs from its activities. Transdisciplinarity is an important innovative feature of these materials. It will enhance the scope of mutual exchange of information and know-how among the scientists in the areas of Astronomy and Earth observation and novelties in the IT/CS area. It will also bring a wider range of topics and experts involved in activities such as training schools and workshops.

C. MILESTONES AND DELIVERABLES: CONTENTS AND TIME FRAMES STRATEGY

Objective 1 (A.4) - Type: Comparison and/or performance assessment of theory/model/ scenario/projection/simulation/narrative/methodology/technology/technique

1. Science and Technology Event or Meeting, Action Workshop.
2. Scientific Publication (including Science and Technology study and excluding handbooks, guidelines and best practices. Excluding Joint Peer-Reviewed Publication), open access.
3. Action Science and Technology Meeting, Working Group.
4. Science and Technology Output, Education and/or Training Material.
5. Science and Technology Coordination, Short-Term Scientific Missions (STSM).

Objective 2 (A.5) - Type: Development of knowledge needing international coordination: new or improved theory/model/scenario/projection/simulation/narrative/methodology/technology/ technique

1. Science and Technology Event or Meeting, Action Workshop.
2. Scientific Publication (including Science and Technology study and excluding handbooks, guidelines and best practices. Excluding Joint Peer-Reviewed Publication), open access.
3. Action Science and Technology Meeting, Working Group.
4. Science and Technology Output, Education and/or Training Material.
5. Science and Technology Coordination, Short-Term Scientific Missions (STSM).

Objective 3 (B.13) - Type: Bridging separate fields of science/disciplines to achieve breakthroughs that require an interdisciplinary approach

1. Science and Technology Event or Meeting, Action Workshop.
2. Science and Technology Coordination, Short-Term Scientific Missions (STSM).
3. Scientific Publication (including Science and Technology study and excluding handbooks, guidelines and best practices. Excluding Joint Peer-Reviewed Publication), open access.
4. Action Science and Technology Meeting, Working Group.
5. Science and Technology Coordination, Joint Student Supervision (at Master's or Doctoral Level).

Objective 4 (A.10) - Type: Dissemination of research results to stakeholders (excluding specific input in view of knowledge application, as per objective 7)

1. Action Science and Technology Meeting, Working Group.
2. Science and Technology Event or Meeting, Training School.
3. Internal and External Communication, Production of dissemination material for distribution.
4. Internal and External Communication, Website.
5. Science and Technology Event or Meeting, Action Conference.

The Action is guided by the following four objectives:

- Objective 1 (Framing the Joint Long-term Agenda) is aligned with category A.4
There are many activities currently underway within natural sciences and IT/CS related to Big Data. This Objective aims at reviewing the existing solutions and making decisions on what to focus on within the transdisciplinarity of this Action. For example, this can be an implementation of column store DBMS, or exploration of possibilities that metadata brings to knowledge extraction, or evaluations of possible standards in data formats or higher level languages - e.g. already quite popular Python extensions such as SciPy and NumPy for scientific computing and data handling.

- Objective 2 (Incubation of New Knowledge) is aligned with category A.5
Given that many of the Action networking members already work on various state-of-the-art Big Data projects, this Objective will provide an environment for a synergy between them. Through this Objective the Action will foster transdisciplinary collaborations that will lead to new knowledge and solutions. For example, pattern recognition is a widely explored topic in CS and, at the same time, one of the key problems in astronomy (e.g. classification of diffuse objects) and remote

sensing (e.g. identification of Earth surface features). Or the Action might consider jointly exploring possible utilization of GPUs, since peculiarities in their programming and operation (intensive memory access and massive vector operations) require an educated decision on where and for what should GPUs be used. Usage of GPUs for pattern recognition and interactive visualization is one of the possibilities.

- Objective 3 (Defragmentation of Existing Knowledge) is aligned with category B.13
The Action will bring together experts from different fields and the Objective aims at them working jointly to build a bridge between their disciplines. For example, this can be a work on similarities between visualization tools in astroinformatics and geoinformatics. One of the tools needed by the former is projection of large point clouds with attached vectorized data (spectra, temporal data, etc.) to the sky, while the latter performs a similar data operation to the Earth surface (e.g. various types of GIS data). Even though a simple transposition of these tools between disciplines is not possible, the underlying methods, algorithms and backend IT/CS solutions might be very similar. Since these two science fields should communicate between each more often, this objective fosters their joint work on merging their knowledge.

- Objective 4 (Dissemination) is aligned with category A.10
Astroinformatics and geoinformatics are faced with a problem of inadequate CS training of students studying natural sciences, while on the other side CS students do not have a specialized knowledge in astronomy and geophysics. This led to various channels of informal learning - self-taught with the help from colleagues involved in the same project or from specialized training schools or online courses. The Action will aggregate a significant amount of specialized knowledge that will be disseminated (over the Internet and at workshops/conferences/training schools) to the researchers and professionals interested in these interdisciplinary fields. This also includes project leaders and managers of projects handling Big Data, who have a difficult task of making educated budget projections and deciding on the balance of different types of personnel. The Action will disseminate examples of good practice to this type of stakeholders, too. The Action will also make a point in trying to stimulate the participation of women in this type of researches. A special attention will be also given to ways how to adapt visualizations to be enjoyable and educational for the general public.

The Action's work will be organized through Working Groups (WG). Even though some WGs will be naturally more focused on a subset of these goals, all WGs will be guided by these Objectives

because of a strong interaction between them. For example, Objective 1 can be applied to the evaluation of existing GPU usage in Big Data environments, while Objective 2 may lead to work on a new joint code exploiting GPUs . For these tasks Objective 3 has to aggregate the existing knowledge (from academic and business sectors), while Objective 4 will disseminate these results targeting researchers, specialists in the private sector, students interested in astroinformatics and geoinformatics and managers of Big Data initiatives and facilities.

The Action will operate through activities that combine four levels of networking work. Among those involving a small group of people, the Action supports Short-Term Scientific Missions (STSM) that consist of short-term visits to some facilities or expert groups participating in the Action. The visits will have a well defined goal - work on joint tasks delegated by the Action's Management Committee, fulfilment of some part of the Objectives that requires on-site visits, or initiation of joint projects requiring a visit. This work can be also combined with the activity of joint student supervision. At a Working Group level, the Action will involve all members of the Action who will work together toward the stated objectives either over the Internet or during the Action's science and technology meetings. The Action will enable interaction among a broader community of experts during the Action workshops and training schools, both organized once a year. The workshops will have a well defined topic narrowed down to practical issues of Big Data science (e.g. on the topic of digital curation and data access). Training schools will target mostly ESRs and young experts, who will be given hands-on training in topics related to science using large datasets. The highest level of networking are conferences, every second year, where a larger community is expected to participate and present their work and results, as well as to share best practices.

Specifically, the Objectives will be achieved by the following milestones:
- Short-Term Scientific Missions (contributes to: A.4, A.5, B.13),
- Joint Student Supervision (at Master's or Doctoral Level) (B.13),
- Action Science and Technology Meeting (WG Meetings and annual meetings) (A.4, A.5, B.13, A.10),
- Action Conference (A.10),
- Action Workshop (A.4, A.5, B.13),
- Training School (A.10),
- Website (A.4, A.5, B.13, A.10),
- Conference Attendance for Action Dissemination Purposes (B.13, A.10),
- Participation in Activities of Other Networks (B.13, A.10).

As described before, WGs will coordinate their activities. This will result in a number of deliverables, where all working groups will contribute, either with input materials or with work on formulation of the output material: Scientific Publications, Documents to be Used as Input to Stakeholders, Book of Abstracts or Proceedings of COST Action Conferences and Workshops, Education and Training Materials, Tutorials and Handbooks (possible textbooks), Virtual Network (the largest fraction of communication will happen over the Internet), and Production of dissemination material for distribution.

The timeline to achieve these goals is the following:

Year 1 Milestones:
- 1st MC ("Kick-Off") meeting and 2 WG meetings
- Core Group (CG) and Management Committee (MC) meetings (every 6 months, preferably combined with some conference of other networks)
- At least 6 STSMs
- Joint Student Supervision (combined with STSMs)
- Action workshop and training school (one workshop and one school: winter or summer)
- website operating (by M6)

Year 1 Deliverables:
- reports from STSM visits and meetings (1st MC - "kick-off", WGs, CG, MC), annual report, joint scientific publications, drafting the initial version of a document to stakeholders (position paper and recommendations), proceedings from the workshop, education and training materials for the training school, dissemination material (leaflets, posters, etc.), virtual network tools up and running (email lists, online collaboration tools, etc.)

Year 2 Milestones:
- 2 WG meetings, CG and MC meetings (combined with conferences of other networks and with the Action conference)
- at least 8 STSMs
- Joint Student Supervision (combined with STSMs)
- Action workshop and training school (one workshop and one school: winter or summer)
- Action conference combined with annual meeting

- mid-term evaluation

Year 2 Deliverables:
- reports from STSM visits and meetings (WGs, CG, MC, annual), annual report, joint scientific publications, release of the first version of a document to stakeholders (position paper and recommendations), proceedings from the workshop and conference, education and training materials for the training school, dissemination material (leaflets, posters, etc.)

Year 3 Milestones:
- 2 WG meetings, CG and MC meetings (combined with some conference of other networks and with the annual meeting)
- at least 8 STSMs
- Joint Student Supervision (combined with STSMs)
- Action workshop and training school (one workshop and one school: winter or summer)
- Annual meeting

Year 3 Deliverables:
- reports from STSM visits and meetings (WGs, CG, MC, annual)
- annual report
- joint scientific publications
- drafting the second version of a document to stakeholders (position paper and recommendations)
- proceedings from the workshop
- education and training materials for the training school
- dissemination material (leaflets, posters, etc.)

Year 4 Milestones:
- 2 WG meetings, CG and MC meetings (combined with some conference of other networks and with the Action conference)
- at least 8 STSMs
- Joint Student Supervision (combined with STSMs)
- Action workshop and training school (one workshop and one school: winter or summer)
- Action conference combined with the final annual meeting
- final evaluation

Year 4 Deliverables:

- reports from STSM visits and meetings (WGs, CG, MC)
- annual report combined with the final evaluation report
- joint scientific publications
- release of the final version of a document to stakeholders (position paper and recommendations)
- proceedings from the workshop and conference
- education and training materials for the training school
- dissemination material (leaflets, posters, etc.)

D. ACTION STRUCTURE AND PARTICIPATION – WORKING GROUPS, MANAGEMENT, INTERNAL PROCEDURES

Working groups: The Action Objectives are setting the framework for work within the Action, while the actual work is performed within the Action's four Working Groups (WGs). The WGs follow all four Objectives and they express the nature of challenges described in part A:

- WG1: Optimisation of database tools in astro- and geophysics contexts:
this group is focused more on the back-end tools providing support for knowledge extraction from large datasets (database management systems, hardware configurations, heterogeneous environments, location of processing of large users' data-requests, etc).

- WG2: Data mining and machine learning in the petabyte era as frontiers in astronomy and Earth observation: this group is inclined more toward front-end solutions visible to the users, where often tools used to date fail or are too slow on petabyte datasets.

- WG3: Education of a new generation of experts in knowledge extraction from massive datasets: this group has the task of identifying critical gaps in the Big Data users' skills, organizing materials needed for training and education of users, and organizing training sessions.

- WG4: Visualization of high dimensional data: this group explores various aspects of visualization of large datasets under scientific and outreach requirements, promotes the role of visualization in data-mining (visual analytics) and assesses various visualization tools.

The management structure is the following:

Management Committee (MC): in charge of management and supervision of the Action. Delegates to the MC are expected to play a key role in the work of the Action, participating in meetings, linking (in both directions) the Action with work underway in other important areas, and liaising with scientists and research groups in their own countries. The MC will meet on a regular basis, at least once per year. The meetings will target some important milestones, where important reports are available, so as to perform a joint peer review and to quickly identify and resolve any issues that require attention by the MC. The MC will promote an active role in implementing COST policies, in particular encouraging active involvement of young and senior scientists, with a priority for PhD and young postdoc students, including joint student supervision.

MC Chair: The role of the MC Chair is to provide the reference point of the Action, to chair the MC meetings and to prepare all scientific reports such as the annual progress and the final report. The MC Chair will be elected during the first meeting of MC. The MC chair will coordinate assessments of milestones as a part of the materials needed for MC meetings.

MC Vice-Chair: The Vice-Chair is elected by the MC and should come from a different research field than the Chair (e.g. one is from astronomy another from Earth observation or computer science). The Vice-Chair will take care of practical issues and will represent the MC in the relations with the external world, thus allowing the Chair to concentrate on scientific and MC coordination issues. The MC Vice-Chair will be responsible for reviewing financial and progress reports and for the fulfilment of web-based communication and project dissemination tasks, maintenance and logistics for all events and monitoring and evaluation of Action activities.

WG Leaders and co-Leaders: Each WG will have two leaders from different communities involved in this Action, appointed by the MC. They will coordinate the WG networking and capacity building activities, stimulate STSMs and seek synergy with other WGs. They will provide the input to the annual reports and help preparing working materials for annual meetings/workshops/conferences.

The Training, Dissemination and Liaison Manager (TDL) will support the MC Chair and the WG Leaders for the management and organization of the MC meetings and WG events. He/she will be coordinating organization of training schools and dissemination efforts. The TDL will be in charge of creating cooperation with other projects or industrial activities.

The Technical Manager (TM) will support the MC Chair and Vice-Chair in the preparation and running of the meetings, by keeping updated the Key Performance Indicators for Action monitoring and evaluation of the activities and will manage the website through a content management system, so that the site can be kept constantly up-to-date.

The Core Group (CG) comprises the MC Chair, the MC Vice-Chair, the WG Leaders, the TDL and the TM. The CG serves as a coordinating body and provides directions at strategic, tactical and operational levels. The monitoring of the Action's performance is performed by the Core Group, who will continuously check the status of tasks within the network, report to the Management Committee potential risks of missing deadlines because of some slow task execution, and take care of some logistical details required for reaching the milestones. This approach will prevent confusions about deadlines and responsibilities.
The CG will meet every six months (the meeting will be joined to the MC meeting and to the Action Conference), and should address most of the issues on the fly by e-mail and/or phone/video conference.