Highlights of Astronomy, Volume 16
XXVIIIth IAU General Assembly, August 2012
T. Montmerle, ed.

doi:10.1017/S1743921315001660

Summary: Special Session SpS15: Data Intensive Astronomy

A new paradigm in astronomical research has been emerging – "Data Intensive Astronomy" that utilizes large amounts of data combined with statistical data analyses.

The first research method in astronomy was observations by our eyes. It is well known that the invention of telescope impacted the human view on our Universe (although it was almost limited to the solar system), and lead to Keplerfs law that was later used by Newton to derive his mechanics. Newtonian mechanics then enabled astronomers to provide the theoretical explanation to the motion of the planets. Thus astronomers obtained the second paradigm, theoretical astronomy. Astronomers succeeded to apply various laws of physics to reconcile phenomena in the Universe; e.g., nuclear fusion was found to be the energy source of a star. Theoretical astronomy has been paired with observational astronomy to better understand the background physics in observed phenomena in the Universe. Although theoretical astronomy succeeded to provide good physical explanations qualitatively, it was not easy to have quantitative agreements with observations in the Universe. Since the invention of high-performance computers, however, astronomers succeeded to have the third research method, simulations, to get better agreements with observations. Simulation astronomy developed so rapidly along with the development of computer hardware (CPUs, GPUs, memories, storage systems, networks, and others) and simulation codes.

It has been well known that we need to conduct "statistical" analysis among and/or comparisons with various celestial objects to better understand astrophysical processes. However the limited sensitivity and amount of data in the past prohibited us to do so.

The rapid development of computer hardware depends strongly on semiconductor technologies, which, in turn, leads to large sensitive detectors that enabled astronomers to easily survey large sky areas. There are several challenging projects in the world to get large amount of data: ALMA with a few Petabytes of data product per year, LSST with expected product of 200 Petabytes for over ten years, Pan-STARRS that will produce several Terabytes per "night", together with the VISTA, ELT, TMT, SKA, and others. These projects cover a wide range of scientific themes: cosmology, the large-scale structure of the Universe, formation of galaxies, star formation, variable stars, transient phenomena such as the Gamma-ray bursts, small bodies in the solar system, extrasolar planets, life in the Universe, dark matter and dark energy, and others.

Thus a new era of astronomical research utilizing large amounts of data will soon come, and astronomers need to be well-prepared for this new era. Since the data production rate will be 100 to 1,000 times larger than the past, it will be crucial to have a combination of advanced machine learning technologies with immediate access to extant, distributed, multi-wavelength databases. Such an approach is necessary to make these assessments and to construct event notices that will be autonomously distributed to robotic observatories for near-real-time follow-up. Advanced data analyses combined with statistics and data mining will be essential to derive general "rules" and/or "knowledge" on various

https://doi.org/10.1017/S1743921315001660 Published online by Cambridge University Press

phenomena in the Universe, as the data volumes will make human inspection and analysis of the data impossible. The most important and exciting astronomical discoveries of the coming decade will rely on research and development in data science disciplines (including data management, access, integration, mining, visualization and analysis algorithms) that enable rapid information extraction, knowledge discovery, and scientific decision support for real-time astronomical research facility operations.

Significant scientific results are expected to be obtained from data-intensive astronomical research in the very near future and beyond under the fourth paradigm in astronomy – Data Intensive Astronomy.

— M. Ohishi, Past President of Comm. 5, October 2012

Summary of the Special Session

Special Session 15: Data Intensive Astronomy was held during the IAU General Assembly in Beijing in eight sessions from August 28 to August 31, 2012. The program began with three excellent review presentations that placed the age of data-intensive astronomy in context, focusing on the optical, radio, and X-ray spectral ranges. Subsequent talks made it abundantly clear that the data deluge is upon us now, from instruments and surveys such as SDSS, LAMOST, ALMA, VISTA, APERTIF, and LOFAR, and that the magnitude of the challenge will only get worse in the future, with, e.g., LSST, the SKA precursors, and SKA itself.

T. Tyson made a prescient observation in his presentation on LSST, in which he described LSST as an integrated survey system. That is to say, the future of large surveys is not about a telescope, detector, or data processing system, but about the complete integration of these subsystems. The data flows and complexities are such that all components must be designed from the outset as a balanced whole (with, we might add, the software and data management components developed in concert with the hardware, not as an under-funded afterthought). We are, he said, entering the age of data abundance and the era of the data-driven astronomer.

The data volumes and the complexities and interdependencies in the data being produced now and in the coming decade necessitate innovation in our software tools and data management infrastructure:

- Algorithmic research and cross-fertilization with other fields such as computer science, statistics, knowledge discovery, and data mining, with clever implementations that defeat the standard scaling constraints.
- Automated and semi-automated classification.
- Common workflow development environments and shared workflows.
- Common data representations; where FITS may no longer be sufficient in terms of data volume or structural complexity, consider the CASA measurement set or the HDF5 formats.
- Access to atomic and molecular databases for spectral line analysis.
- Access to numerical simulations and comparison of simulations to observations, through "virtual observations" that take into account telescope, detector, and possible atmospheric effects.
- Parallel computing, including parallel implementations of large (peta-scale) databases, and adaptations of algorithms to exploit these architectures.

https://doi.org/10.1017/S1743921315001660 Published online by Cambridge University Press

- Server-side and cloud-based computation; moving the algorithm to the data rather than the data to the algorithm.
- GPUs and other innovative computer architectures.
- Efficient visualization of high-dimension data (with server-side and client-side implementations).
- Versatile, global data discovery, access, and interoperability through the Virtual Observatory.

A closing session concerning public outreach and education emphasized how astronomical discoveries and the data underlying them are having a huge societal impact. Millions of students, educators, and members of the general public explore the universe with tools such as World Wide Telescope and Google Sky. Thousands of people contribute to real-world research problems through citizen science and crowd-sourcing initiatives such as the Zooniverse. Such efforts help to promote logical thinking and create an informed society, and an informed society is one that will be more supportive of scientific research in general.

Special Session 15 drew a diverse audience. Members of the SOC recognized a number of attendees, but more important was the number that we did not recognize. The community recognizes the challenges we face and wishes to help in finding solutions.

Special thanks are due to M. Ohishi for his extraordinary efforts as chair of the SOC, to C. Cui (China) for special support with computer networking, and to all of the contributors to the program.

Special Session 15 was sponsored by Commission 5: Astronomical Data and Documentation, which as part of the reorganization of the IAU will move to the new Division B: Facilities, Technologies, and Data Science (D. Silva, president). We believe this is a good move for Comm. 5, and will provide excellent opportunities for dialog amongst facility and instrument developers, data processing experts, data managers, and the virtual observatory.

We must embrace our data-intensive present–and even more-data-intensive future–with energy and enthusiasm, as this is the way new astronomical discoveries will emerge, from planetary science and stellar astrophysics to cosmology and all fields in between.

— R. Hanisch, President of Comm. 5, October 2012

Scientific Organizing Committee

Masatoshi Ohishi (Japan, Chair)
Kirk Borne (United States)
Janet Drew (United Kingdom)
Robert Hanisch (United States)
Melaine Johnston-Hollitt (New Zealand)
Nick Kaiser (United States)
Ajit Kembhavi (India)
Oleg Malkov (Russian Federation)
Bob Mann (United Kingdom)
Raffaella Morganti (Netherlands)
Paolo Padovani (Germany)
Hu Zhan (China Nanjing)

https://doi.org/10.1017/S1743921315001660 Published online by Cambridge University Press

Robert J. Hanisch
Space Telescope Science Institute, 3700 San Martin Drive,
Baltimore, MD 21218, USA
email: hanisch@stsci.edu

Masatoshi Ohishi
National Astronomical Observatory of Japan, 2-21-1,
Osawa, Mitaka, Tokyo, 181-8588, Japan
email: masatoshi.ohishi@nao.ac.jp

https://doi.org/10.1017/S1743921315001660 Published online by Cambridge University Press