Astroinformatics: Data-Oriented Astronomy

Kirk Borne[1]

Computational and Data Sciences, George Mason University, Fairfax, VA 22030, USA

Abstract. We describe Astroinformatics, the new data science paradigm for astronomy research and education, with a focus on preparing the next generation of scientists with skills in data-oriented science: discover and access large distributed data repositories, conduct meaningful scientific inquiries into the data, mine and analyze the data, and make data-driven scientific discoveries.

New modes of discovery are enabled by the growth of data and computational resources in the sciences. This cyberinfrastructure includes databases, virtual observatories (distributed data), high-performance computing (clusters and petascale machines), distributed computing (Grid, Cloud), intelligent search and discovery tools, and innovative visualization environments [1]. Data volumes from astronomical sky surveys have grown from gigabytes into terabytes during the past decade, and will grow from terabytes into tens (or hundreds) of petabytes in the next decade. This plethora of new data both enables and challenges effective astronomical research, requiring new approaches. Thus far, astronomy has tended to address these challenges in an informal and ad hoc manner, with the necessary special expertise being assigned to e-Science [1] or survey science. However, we see an even wider scope and therefore promote a broader vision of this data-driven revolution in astronomical research: the creation of a major new discipline, which we call *Astroinformatics* [2]. By virtue of its new stature, Astroinformatics needs to be integrated into Astronomy as a formal sub-discipline within agency funding plans, university research programs, graduate training, and undergraduate education [3]. *Astroinformatics is an essential methodology for data-oriented astronomical research.* The future of astronomy depends on it.

Within the scientific domain, data science is becoming a recognized academic discipline. Iwata states that *"without the productivity of new disciplines based on data, we cannot solve important problems of the world"* [4]. The report of the 2007 NSF workshop on data repositories states: *"Data-driven science is becoming a new scientific paradigm – ranking with theory, experimentation, and computational science"* [5]. Consequently, astronomy and other scientific disciplines are developing sub-disciplines that are information-rich and data-intensive to such an extent that these are now becoming (or have already become) recognized stand-alone research disciplines and full-fledged academic programs on their own merits. The latter include bioinformatics and geoinformatics, but will soon include astroinformatics, health informatics, and data science.

In sky survey astronomy, the LSST (Large Synoptic Survey Telescope) will produce one 56Kx56K (3-Gigapixel) image of the sky every 20 seconds, generat-

ing nearly 30 terabytes of data daily, growing into a 100-petabyte data collection within 10 years [6]. To cope with such enormous data volumes, an informatics approach is required. What is informatics? Informatics has recently been defined as *"the use of digital data, information, and related services for research and knowledge generation"* [7]. This description complements the usual definition: *informatics is the discipline of organizing, accessing, integrating, and mining data from multiple sources for discovery and decision support* [8]. Therefore, we believe that the discipline of Astroinformatics will include a set of naturally-related specialties including data modeling, data organization, data description, transformation and normalization methods for data integration and information visualization and knowledge extraction, indexing techniques, information retrieval and data mining methods, content-based and context-based information representations, consensus semantic annotation tags, classification taxonomies, astronomical concept ontologies, and astrostatistics [2]. These enable scientific knowledge discovery across heterogeneous massive data collections. But they do more – they also enable collaborative research and data re-use (both in the research environment and in learning settings). Astroinformatics provides a natural context for the integration of research and education – the excitement and experience of research and discovery are enabled and infused within the classroom through easy re-use of data. Astroinformatics enables many other science use cases: repurposing of archival data for new projects, semantic integration of data within different contexts (*e.g.*, follow-up observations with robotic telescopes, collaborative research environments, search engines), literature-data linkages, personalization and recommendation services in astronomical data archives, intelligent retrieval of data, autonomous classification of objects, quantitative scoring of astronomical classifications of new objects, discovery of "interesting" objects and new classes of objects, information retrieval metrics on archive queries (precision and recall metrics), decision support for new observations and instrument-steering, query-by-example functionality in astronomical databases, and development of an astronomical genome. The latter will lead to the specification of key "genes" that define each class of astronomical object [3].

References

1. Eastman, T., Borne, K., Green, J, Grayzeck, E., McGuire, R., & Sawyer, D.: eScience and Archiving for Space Science. Data Science Journal, 4, 67-76 (2005)
2. Borne, K. D. Astroinformatics: The New eScience Paradigm for Astronomy Research and Education. Microsoft eScience Workshop at RENCI (2007)
3. Borne, K.: Astroinformatics: A 21st Century Approach to Astronomy. http://mason.gmu.edu/∼kborne/Borne_astroinformatics_CDH_FFP_APP.pdf (2009)
4. Iwata, S.: Scientific "Agenda" of Data Science. Data Science Journal, 7, 54 (2008)
5. NSF/JISC Repositories Workshop, http://www.sis.pitt.edu/∼repwkshop/ (2007)
6. Becla, J., et al.: arxiv.org/abs/cs/0604112 (2006)
7. Baker, D. N.: Informatics and the 2007-2008 Electronic Geophysical Year. EOS, 89, 485 (2008)
8. Downloaded from http://www.google.com/search?q=define%3A+informatics