ASTROINFORMATICS IN CANADA

Nicholas M. Ball[1], David Schade[1]

LRP 2010 White Paper

ABSTRACT

We detail the significance of the emerging field of *astroinformatics* to the next decade of Canadian astronomy. Following a brief introduction defining the subject, we explain its importance, summarize work in Canada in the past decade, explain the current state of the field, give examples of astroinformatics enabling improved science, and conclude with recommendations. Our recommendations can be summarized, in order of importance, as: (1) The existing infrastructure at CADC should be maintained and utilized as the basis for enabling astroinformatics methodologies to be widely adopted within the Canadian community. Given implementations of the methodology should remain science-driven. (2) The established services should not attempt to duplicate the in-house efforts of large international collaborations such as LSST. (3) The full exploitation of the potential of astroinformatics for new discoveries should be furthered by the hiring of specialists in this area, especially at the postdoctoral level. (4) The skills in astroinformatics demonstrated by such specialists should play a strong role, alongside their science, in the assessment of such specialists in their astronomical careers.

Subject headings: methods: data analysis — astronomical databases: miscellaneous — catalogs

1. WHAT IS ASTROINFORMATICS?

In the past two decades, astronomy has gone from being starved for data to being flooded by it. This onslaught has now reached the stage where the exploitation of these data has become a named subdiscipline in its own right, *astroinformatics.* The naming follows in analogy from the already established fields of bio- and geoinformatics, which contain their own journals and funding. Following Borne et al. (2009), we define astroinformatics as a subject that **includes a set of naturally-related specialties including data organization, data description, astronomical classification taxonomies, astronomical concept ontologies, data mining, visualization, and statistics**. The place of the subject is encapsulated in a broad schema representing the process of observational astronomy, shown in Figure 1.

Although astroinformatics is stated in terms that are not motivated by a single specific science driver, we emphasize that any particular application of the methodology must be motivated in such terms, either a specific goal, or a general well-stated purpose such as exploration of new parameter space with the expectation of serendipitous discovery. Such a conclusion was also reached in a recent review of the subject by Ball & Brunner (2010). It is important to note that *astroinformatics is not just the Virtual Observatory.* The VO, while important, is much closer to what CADC already does, a provider of data to be exploited, not the exploitation itself.

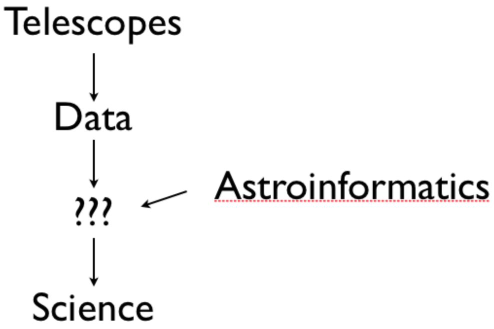

Figure 1. Schematic representation of astronomy, showing the context of astroinformatics.

2. WHY IS IT IMPORTANT?

It is well known that the amount of available data is increasing exponentially. This is a rather predictable trend and we can confidently state that this will continue for at least the 2010–2020 timeframe. More significantly, the data are not just increasing in size, but in complexity and dimensionality. It will thus become commonplace that *the analysis of a dataset will be intractable without a significant component of automation.* Given this complexity, such automated analysis will *require* highly advanced algorithms that are capable of dealing with such data. In many cases, such algorithms already exist in the computer science, data mining, and statistics communities, but have not been utilized in astronomy due to the amount of learning required. Also, readily available new data have yielded many new discoveries through the application of the existing methods, and this has left less incentive to try and extract further information using more modern techniques. However, when the data become such that the use of existing methods is intractable, astroinformatics will be required to enable advances.

Historical precedent has shown time and again in astrophysics that when one accesses a new region of parameter space, such as fainter objects, higher resolution, or shorter timescales, that exciting and unexpected discoveries are made. Astroinformatics will ensure that the parameter space opened up by modern datasets can continue to be similarly exploited.

Electronic address: Nick.Ball@nrc-cnrc.gc.ca
[1] Herzberg Institute of Astrophysics, 5071 West Saanich Road, Victoria, BC V9E 2E7

Examples of astroinformatics algorithms[2] are machine learning methods such as artificial neural network, support vector machine, decision tree, k nearest neighbour, other statistical methods such as kernel density estimation, k-means clustering, mixture modelling, expectation maximization, self-organizing maps, independent component analysis, and well-known data mining algorithms as yet little used in astronomy but with much potential to enable useful science in the field, such as genetic algorithms, the information bottleneck method, Bayesian networks, and semi-supervised learning. Also within the field are semantic processes, scientific visualization, and the development of new statistics.

Common applications of astroinformatics algorithms within astronomy could include image registration, stacking, deconvolution, statistics of images, object detection, classification, removal of contaminants, cross-matching, discovery of rare objects, and photometric redshifts. The algorithms are applicable to multi-wavelength data, the time domain, and simulations, as well as more traditional data. Hence the purview extends across both the pixel-space domain, and the domain of object catalogues. The use of these applications for science is encapsulated in the well-known and widely used methodology, *knowledge discovery in databases.*

3. WORK IN THE PAST DECADE

The previous LRP recommended CADC as the main focus for such work in Canada for 2000–2010, and several advances have been made. It was stated that “The LRPP strongly recommends that NRC’s outstanding CADC develop its ability to manage archives of data from upcoming space and ground-based observatories. Funding should be provided to develop innovative data mining techniques that maximize the scientific usefulness of multi-wavelength observations in astronomy. This should be one of the highest priorities among the computational projects.”

The LRP mid-term review then states that “In accordance with the LRP recommendations, NRC-HIA has invested heavily in its CADC facility which provides archival storage and retrieval of the large amounts of data produced by the major observatories, including data mining tools. The CADC has now largely met the goals set out in the LRP.”

While the CADC is well placed to further the goals of the last LRP, we recommend in §6 below that, in the coming decade, astroinformatics be seen in more general terms as something to be adopted by the community, and not as a desire to expand CADC.

4. CURRENT STATE OF THE FIELD

In addition to the above, the mid-term review also states that: “To meet the increasing challenges created by the rapid growth of the data volumes worldwide, it is now timely to consider how best to plan for increased effectiveness of the CADC and to ensure that Canadian scientists will have the tools to retrieve and analyze their data from LRP and other large facilities in an effective manner.” We still consider this latter statement to be true, and such data analysis tools fall under the purview of astroinformatics.

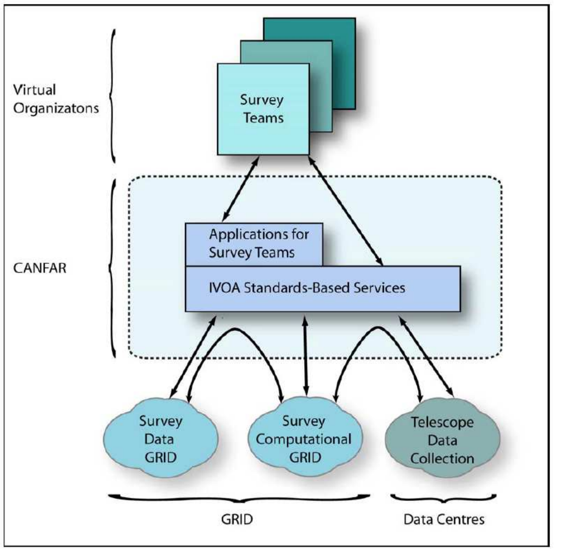
Virtual Organizatons
Survey Teams
CANFAR
Applications for Survey Teams
IVOA Standards-Based Services
Survey Data GRID
Survey Computational GRID
Telescope Data Collection
GRID
Data Centres

Figure 2. CANFAR infrastructure, showing the environment provided for researchers. While many different actual computers are involved in the setup, an individual or collaboration is able to set up and maintain a single *virtual* machine, that is in effect like running just one computer. The resources provided include the ability of the user to install and run software, substantial processing power, accessed as needed via the CADC or the Compute Canada grid, temporary or permanent storage of public or private data via files or databases, and access to existing data via the standard International Virtual Observatory protocols. From Pritchet et al. (2008).

The current CADC infrastructure enabling the ongoing achievement of the LRP goals is now implemented as the Canadian Advanced Network for Astronomical Research project (CANFAR, Pritchet et al. 2008). A schematic representation of what the CANFAR infrastructure provides is shown in Figure 2.

5. EXAMPLES

We give two examples of existing applications using methods from astroinformatics, each enabling new science.

5.1. *MACHO Reprocessing*

Data from the Massive Compact Halo Objects Project (MACHO, Alcock et al. 1997) are being reprocessed on the CANFAR infrastructure to provide significantly improved photometry. This will enable the comparison of light curves for long period variable stars to radiative transfer models to significantly improve constraints on such models describing stellar atmospheres. The data processing involves the ingestion of 7.6 terabytes of data, calculation of PSFs and photometry, and the production of catalogues. The total estimated processing time on one processor is 13,444 days, or 36.8 years. While clearly impractical on any desktop setup, the provision of several thousand computer cores by CANFAR, starting at HIA and the University of Victoria, but extensible to the national Compute Canada infrastructure as needed, renders the task completable in a matter of days. Hence, science is enabled by the CANFAR infrastructure that would not otherwise have been carried out.

5.2. *Photometric Redshift Probability Density Functions*

[2] We do not necessarily expect the reader to be familiar with all of these: the take-home message is that there are a lot available, with a lot of potential.

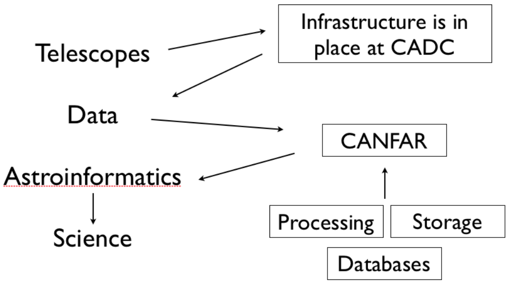

Figure 3. Schematic representation of astroinformatics and the CADC/CANFAR infrastructure.

The redshift of an object such as a galaxy or quasar can be estimated from the broad band colours derived from photometry. Such photometric redshifts are becoming increasingly common. While enabling a much greater abundance of objects than those with spectroscopy to be utilized, the resulting redshifts are less accurate. Nevertheless, many studies to-date have been carried out with these redshifts. However, much more useful than a single scalar-valued redshift, is the assignation of a full probability density function (PDF) in redshift. This substantially increases the amount of information available to the next stage of the analysis. For example, Myers, White & Ball (2009) showed that by utilizing the full PDFs of quasars, in combination with an appropriately sophisticated estimator, the signal to noise of the clustering signal could be improved such that *this was equivalent to a survey 4-5 times the size.* This will enable them to break the so-called redshift-luminosity degeneracy for active galactic nuclei, a limitation to-date in constraining models of their physics. Hence, a more sophisticated algorithm of the type enabled by increased skills in astroinformatics, directly improves the science that can be done with a given dataset.

6. RECOMMENDATIONS

The most appropriate way to leverage the capabilities of astroinformatics within the Canadian community is to utilize the fact that, as shown above in Figure 2, *the required fundamental infrastructure is already in place at CADC.* The nature of this infrastructure, and of astroinformatics, does not require the researcher to be at the same institution as, or to expend significant time in collaboration with, CADC personnel[3]. The existing infrastructure can be straightforwardly extended from the current regime of storage, processing, and databases, to provide full astroinformatics capabilities of the scope described in section 2, for both public data, and proprietary data within a given project. The current infrastructure is provided by CANFAR, and although this particular project is not funded indefinitely, we hope that the community will endorse the importance of continuing the CADC infrastructure developed over the past 24 years. The relationship of the existing infrastructure to astroinformatics is shown in Figure 3.

Our specific recommendations, in order of importance, are:

- *The existing infrastructure at CADC should be maintained and utilized as the basis for enabling astroinformatics methodologies to be widely adopted within the Canadian community. Given implementations of the methodology should remain science-driven.* The infrastructure is straightforwardly and logically extensible to allow this, and we strongly urge that this infrastructure should remain in place and be built upon for the foreseeable future. Note that this recommendation is *not* to expand CADC, but to enable the full potential of astroinformatics implemented on its infrastructure to provide exciting science.

- *The established services should not attempt to duplicate the in-house efforts of large international collaborations such as LSST*, but could serve as leverage to joining such collaborations, or provide complimentary or extra science-driven services.

- *The full exploitation of the potential of astroinformatics for new discoveries should be furthered by the hiring of specialists in this area, especially at the postdoctoral level.* Such hiring might be funded as part of more general community-wide mechanism for increasing the number of postdoctoral and other researchers, or via the formal establishment of astroinformatics as a route for funding in its own right, as has been recommended in the US decadal survey (Borne et al. 2009).

- *The skills in astroinformatics demonstrated by such specialists should play a strong role, alongside their science, in the assessment of such specialists in their astronomical careers*, making this approach a viable career path. This viability should lead to more permanent positions in which astroinformatics expertise plays a substantial role.

- If possible, collaboration with computer scientists and statisticians should be encouraged, so that innovative new astroinformatics methodologies can be developed or adopted.

REFERENCES

Alcock, C. et al. 1997, *The MACHO Project Large Magellanic Cloud Microlensing Results from the First Two Years and the Nature of the Galactic Dark Halo*, ApJ, 486, 697

Ball, N. M. & Brunner, R. J. 2010, *Data Mining and Machine Learning in Astronomy*, invited review, International Journal of Modern Physics D, submitted, arXiv/0906.2173

Borne, K. et al. 2009, *Astroinformatics: A 21st Century Approach to Astronomy*, white paper, Astro2010: The US Astronomy and Astrophysics Decadal Survey

Myers, A. D., White, M., & Ball, N. M. 2009, *Incorporating Photometric Redshift Probability Density Information into Real-Space Clustering Measurements*, MNRAS, 399, 2279

Pritchet, C. et al. 2008, *Canadian Advanced Network for Astronomical Research (CANFAR)*, CANARIE Proposal NEP-39 Statement of Work

[3] It is likely that some projects will retain more direct involvement, as is currently the case for several surveys, especially in the earlier years of the 2010–2020 timeframe. The surveys already involved with CANFAR represent a significant fraction of the Canadian community.