Research Statement for Kirk Borne *(updated November 16, 2010)*

Introduction – Galaxies Research

I have 30 years of professional experience in astronomical and astrophysical research, both observational and theoretical. My main astronomical research interests are in the dynamics and evolution of galaxies and groups of galaxies. I have been studying the effects of tidal encounters on the structure and evolution of such systems, with particular attention given to galaxy mergers, the progenitors of mergers, the galactic-scale consequences of merger episodes, the numerical simulation of merger events, the time scales for merging, and the properties of merger remnants. The end-goal of these studies is to increase our understanding of *rate, state, and fate*: *i.e.*, the galaxy hierarchical assembly rate, the current state of galaxies in group environments, and their ultimate fate. Because we believe that the dynamical evolution of the entire Universe is reflected in the assembly of present-day large galaxies from smaller units in the past (through collisions and mergers in the hierarchical formation cold dark matter model), my research has consequences for the biggest cosmological questions in astronomy: Where did all this come from? When did it form? How is it changing? and How will it end? Several of my colleagues in this field point out that I was several years ahead of the astronomical community in my pursuit of this line of research – long before it became fashionable to study colliding and merging galaxies (which is definitely true these days), I was developing detailed numerical models and collecting a wide range of multi-wavelength data (both imaging and spectroscopy). Even my professors in graduate school (at Caltech) said to me when I started my thesis: *"if you want to waste your time in this field, it's your choice."* Just a few years later (in 1985), when NASA's IRAS satellite discovered that that the most luminous and dynamic galaxies in the Universe are in fact colliding and merging galaxies (going through a collision-induced massive burst of new star formation), these same scientists jumped on the bandwagon of colliding galaxy research. These days, a large double-digit percentage of all astronomers in the world now work in this research area.

I have carried out supporting observations for my astronomy research using numerous observatory facilities, both ground-based (Palomar, KPNO, CTIO, Calar Alto, VLA, and others) and space-based (IUE, HST, ROSAT, Chandra, and Spitzer). Two of my HST (Hubble Space Telescope) projects led to "15 minutes of fame." In the first case, it was actually quite a lot more than "15 minutes" – the colorful and dramatic imagery of the Cartwheel ring galaxy has appeared in countless magazines, news stories, websites, and textbooks. My conference presentation of our research results in 1995 yielded the "Best Poster" prize at a major international conference in Puebla, Mexico. In the second case, we discovered conclusive evidence that some of the most dramatic colliding galaxies (the ULIRGs; see below) are in fact the result of multiple mergers – *i.e.*, several galaxies are merging simultaneously, which may indicate that these galaxies are the end-state of the evolution of compact groups of galaxies (which was a long-standing unresolved puzzle). In addition, I have augmented these activities through archival research, which has been facilitated by many online astronomical databases (particularly the Sloan Digital Sky Survey, 2-Micron All-Sky Survey, and the GALEX ultraviolet sky survey), while also making plans for the largest scientific database ever assembled – the Large Synoptic Survey Telescope (**LSST**) 20-Petabyte object database (and 100-Petabyte image archive) that will come online in the next decade.

For the past several years, my research efforts have focused on how gas-rich galaxy-galaxy collisions induce and affect the phenomenally strong starbursts seen in the incredible sample of IRAS-detected ultraluminous IR galaxies (ULIRGs). The ULIRGs may represent the missing link between the very

active quasars and normal quiescent elliptical galaxies. They also present extraordinarily high rates of star formation that accompany strong signatures of tidal interaction. It has been determined that a substantial fraction of the old stars in elliptical galaxies and most of the metals in the intergalactic medium were formed in ULIRG-type starburst episodes. My research on ULIRGs has included two large HST surveys (optical and near-IR; for both which I was the Principal Investigator), Chandra X-ray observations, ground-based imaging and spectroscopy, radio continuum imaging, numerical simulations, and correlative data mining analysis within several very large survey databases. As indicated above, the broad goal of this research is to obtain a global, objective determination of the rate and significance of interactions and mergers in the overall scheme of galaxy formation and evolution. Toward this end, my research has focused on the fact that activity in galaxies (*e.g.*, starbursts, quasars, active galactic nuclei = AGN) is often associated with tidal interactions. This strong physical association has led to various studies of the corresponding interaction-activity connection. It is now believed that several important cosmological populations are related to (if not equivalent to) the IR-luminous galaxy population: high-redshift sub-mm (SCUBA) sources, IR-selected AGN, gamma-ray burst host galaxies, and constituents of the cosmological infrared background radiation. Given that the most luminous (optical, radio, and IR) objects in the universe are such "interactive" systems, studies of ULIRGs are therefore particularly important in the quest to understand the birth rate, current state, and ultimate fate of galaxies in the cosmological setting.

The "Tonnabytes" Data Challenge – Data Science Research

In conjunction with my two decades of scientific support for various NASA space science missions and their science data systems, I have developed a strong interest in the application of scientific databases and information technologies to scientific research problems. This interest has been stimulated by the fact that the astronomical research community is about to be deluged with multi-petabyte databases from numerous projects. The rich and diverse content of this "virtual sky" will far exceed existing data sets and information resources by orders of magnitude! By "virtual sky", we mean that the "instrument" that we will use to study the sky will be the distributed online databases (which have been aggregated from large sky survey telescope projects). It is for this reason that the NVO (National Virtual Observatory) was initiated worldwide in 2001. I was a co-Investigator on the U.S. NVO project, which was a large NSF-funded Information Technology Research (ITR) project to build the framework for the NVO. My contributions were in these areas: distributed data mining, science user scenarios, user requirements analysis, semantic e-Science, astronomical event (alert) classification, and education/public outreach.

My focus on data-intensive astronomy in the last 10 years has pioneered new research programs in Astronomy within the following Data Science areas: data mining, machine learning algorithms, statistics, information visualization, distributed data systems, semantic e-Science, XML technologies, grid and cloud computing, genetic algorithms, large scientific database management, sensor networks, robotic telescopes, and information retrieval systems. The goals of this research include the following: to maximize the value, long-term use, and scientific impact of astronomical data through cross-discipline correlative research; to implement seamless access across distributed data sources; and to address the real problem, which is not a lack of data, but how to discover, integrate, and mobilize the relevant data for scientific discovery. These research areas include data exploration and data exploitation, for the benefit of science. I am therefore pursuing research projects today that exploit several leading-edge information/database technologies in this area, including the application of supercomputers to massively parallel data mining of terabyte (and

eventually, petabyte) science databases. I am also focusing on the application of similar data science methodologies to autonomous spacecraft operations: mining and analysis of large volumes of time-series data (engineering, telemetry, and science streams) for decision-making (about spacecraft operations, trajectory maneuvers, and science observations, respectively) in deep space with minimal human intervention.

I define the above data challenges as the **"Tonnabytes Challenge"**, since scientists in different disciplines run into data-volume challenges at different levels (gigabytes, terabytes, petabytes, or exabytes) – each one of these scientists is facing the challenge of a "ton of bytes" (= "tonnabytes").

Computational Astrophysics Research

For my research programs in the areas described above, I have received funding from NASA and NSF. I have been P.I. on several grants, a co-I on the NSF NVO project, a co-I on the NASA SIM (Space Interferometry Mission) Science Team, a PI and co-I on several data mining research grants, and PI and co-I on several NSF and NASA Education projects. I am also currently a sponsored (funded) member of the LSST project team (see below). I fully anticipate continued success in obtaining grant funding to support investigations in these active research areas.

By way of example, I will describe one particular research program that was started several years ago, and is now in full blossom. I submitted a proposal to NASA in 2004 to pursue this project – *"Simulating a Million Interacting Galaxies"* – the proposal received scores of "Excellent", but it was not funded due to insufficient funds in that particular NASA research program. The project is an application of parallel computing and Genetic Algorithms to the computational study of colliding and merging galaxies. To quantify objectively the role of galaxy interactions and mergers within the context of galaxy formation and evolution, it is critical to know which structures are indeed produced by galaxy-galaxy collisions and which structures are not. An historically productive test of the tidal interaction hypothesis has been the derivation of collision model solutions that can consistently explain the observed peculiarities, disturbances, and exotic phenomena within the corresponding galaxies – this was a field of research that I pioneered (Borne et al. 1994, ApJ, 435, 79; and references therein), and which is now very active (http://www.etsu.edu/physics/wars/wars.html). Such tests and validations of gravitational interaction physics as the main cause for observed galaxy structures also contribute to the overall verification of our understanding of the Universe, hierarchical mass assembly, galaxy formation, and galaxy evolution. Unfortunately, there has been a significant bottleneck in these collision scenario tests – the enormously large volume of parameter space that must be explored in order to find the best-fit (preferably unique) collision encounter solution has permitted only a handful of interesting low-redshift cases to be modeled to the necessary level of detail. Even more worrisome is the fact that the uniqueness of such solutions is hard to prove and is frequently a premature claim.

In the recent past (*e.g.*, in our 2004 NASA proposal), two computational techniques were envisioned to attack this problem (*i.e.*, to conduct a rapid search, in a very high-dimensional parameter space, for the optimal collision model that matches a given set of observational data). These techniques were the application of: (a) Genetic Algorithms (for rapid optimization) and (b) Beowulf Linux Clusters (for rapid parallel computations). The application of genetic algorithms (GAs) allows a rapid search through a high-dimensional parameter space, avoiding local (false) extrema in the optimization function, and propagation of the best parameters from one generation of models to the next ("survival of the fittest"). GA research has shown that these techniques find optimal solutions

100-1000 times faster than trial-and-error searches, and also yield true (global) extrema in the optimization function (unlike traditional hill-climbing algorithms, such as gradient-descent). If one then adds to this rapid model-fitting the ability to run ∼1000 galaxy-galaxy collision scenarios in parallel (through the application of the Beowulf parallel computing environment), one is thus able to run millions of merger simulations for about the same wall clock time as one previously needed for 10-100 runs. With this fantastic speed-up and robust solution-finding algorithm, one can assess the state of galaxy mergers, mass assembly, and hierarchical galaxy formation in the Universe by applying our computational method to large data sets of imaging observations of galaxies (from the Hubble Deep Fields, from the Sloan Digital Sky Survey, from other sky surveys, from the Spitzer Space Telescope, and eventually from the Large Synoptic Survey Telescope, which will image billions and billions of galaxies, of which approximately one billion will be involved in some sort of collision/merger episode).

Citizen Science and U-Science Research

In the past year, the development of Citizen Science has transformed the above problem from a massively parallel computing problem with exotic hard-coded Genetic Algorithms into a completely manageable "solved" problem. Citizen Science refers to the involvement of (non-PhD) members of the general public in the science process – performing data characterization, pattern recognition, feature selection, anomaly detection, and discovery. GMU is now one of the world leaders in this field, through the efforts of John Wallin (now at MTSU) and myself (through the NSF-funded Zooniverse grant at Mason: zooniverse.org). Our deployment of the Galaxy Mergers Zoo (mergers.galaxyzoo.org) in November 2008 has led to Citizen Scientists running and viewing millions(!) of galaxy collision simulations, and then scoring them on the likelihood that a given numerical simulation model actually "looks like" the observed galaxy collision seen in the astronomical images. The power of human cognition is the "genetic algorithm" and the army of 300,000 Galaxy Zoo volunteers is our "parallel computer." The end-result will ultimately be a collection of best-fit numerical models for thousands of colliding galaxies, which will then be fed into state-of-the-art galaxy simulations (adaptive mesh tree codes, with gravitational and various gas dynamical processes included) for further physics-based refinement.

My current research now includes studies of Citizen Science and how it helps science, how the volunteers learn science, and which type of volunteer science experiences are most engaging. I have coined a name for this new field of human-centered computing for science: **U-Science**. U-Science is derived from the concept of e-Science. The emergence of e-Science over the past decade as a paradigm for Internet-based science was an inevitable evolution of science that built upon the web protocols and access patterns that were prevalent at that time, including Web Services, XML-based information exchange, machine-to-machine communication, service registries, the Grid, and distributed data. We now see a major shift in web behavior patterns to social networks, user-provided content (e.g., tags and annotations), ubiquitous devices, user-centric experiences, and user-led activities. The inevitable accrual of these social networking patterns and protocols by scientists and science projects leads to U-Science as a new paradigm for online scientific research (i.e., ubiquitous, user-led, untethered, You-centered science). U-Science applications include components from semantic e-science (ontologies, taxonomies, folksonomies, tagging, annotations, and classification systems), which are applied within a social networking context. Citizen science is an example, but other examples include more hard-core science (*e.g.*, biodas.org; and AstroDAS – see Borne, "A machine learning classification broker for the LSST transient database," Astronomische Nachrichten, 329, 255, 2008).

Astroinformatics Research

At the leading edge of all of my research energies right now is the application of data mining, statistics, and machine learning algorithms (*i.e.*, Informatics, Data Science) to scientific discovery from large databases. In parallel with several colleagues, I coined the term **"Astroinformatics"** to describe data-oriented astronomical research and education. I have written "the paper" on Astroinformatics and I have authored a national study report on the subject (with over 90 co-signers) that was submitted to the National Academies as part of the recent Decadal Survey of Astronomy and Astrophysics. My study of informatics and data science has led me to collaborate and consult with scientists in numerous disciplines outside of astronomy, including agriculture, remote sensing, drug safety, national intelligence, digital libraries, computer science, and health informatics. I will continue to pursue research along many of these paths in the years to come. Of special interest is the field of Health Informatics, for which Mason has started a new center in this domain. I am already in discussions with Mason scientists working in this area. Health Informatics focuses on data science and information technologies for healthcare, including portable electronic health records, health database modeling and management, dashboards (for information visualization and change detection), evidence-based medicine, drug discovery, insurance fraud detection, diagnosis and predictive analytics, and more. There is an enormous (multiple tens of billions of dollars) investment in this area nationally, and it is one of the top 3 challenges identified by the present White House administration. I have learned through my decade of research in the fields of informatics, data mining, and data science that the algorithms and methodologies of one field (such as Astroinformatics) are transferrable to numerous other fields (such as Health Informatics). I have co-hosted several informatics workshops and meetings in the fields of Earth and Space Science informatics, and I plan to organize a multi-disciplinary Discovery Informatics workshop at Mason with representation from nearly two dozen X-informatics science disciplines (such as Geo-informatics, Bio-informatics, Health Informatics, Materials Informatics, Biodiversity Informatics, etc.).

Astroinformatics is needed in order to address the flood of new data in astronomy. The most significant data-producing astronomical research project in the years ahead will be the LSST (Large Synoptic Survey Telescope). The $600-million LSST project will produce 30 terabytes of data daily for 10 years, resulting in a 100-petabyte final image data archive and a 20-petabyte final catalog (metadata) database. This large telescope will begin operations in 2018 at Cerro Pachon in Chile. It will operate with the largest camera in use in astronomical research: 3 gigapixels, covering 10 square degrees, roughly 1000 times the coverage of one Hubble Space Telescope image. Two pairs of 6-gigabyte images will be acquired, processed, and ingested every 40 seconds, for 10 years. Each spot on the available sky will be re-imaged approximately every 3 days, resulting in about 2000 images per sky location after 10 years of operations. This enormous data collection will provide vast astronomical scientific research and discovery potential. I am currently a sponsored member of the LSST team, providing contributions in the areas of galaxy research, data mining research, scientific data management, and education & public outreach. In addition, I am the national Chairman of the LSST Informatics and Statistics Science Collaboration Team (with nearly 40 members), and I am the GMU representative to the LSST Board of Directors. It was through my efforts that GMU became an institutional member of the LSST consortium in summer 2010. The Informatics and Statistics Science Collaboration Team that I am chairing (which consists of astronomers, computer scientists, data mining experts, and statisticians) is specifically devoted to finding efficient algorithms and effective methods for extracting the most scientific knowledge from the enormous LSST science data collection = KDD (Knowledge Discovery from Data)!

Concluding Remarks

There are many more details of my research activities and long-term goals that I could describe here, but I will conclude by pointing out that all of these areas are relevant to the research and teaching goals of Mason's COS Department of Computational and Data Sciences and Department of Physics and Astronomy. I will devote much of my research effort in the coming years to LSST-related science – this will include graduate and undergraduate student involvement, will include collaborations across numerous Mason departments (*e.g.*, Physics & Astronomy, Applied Mathematics, Computer Science, Geoinformatics, Statistics, Education, and Health), and will include the development of curricula and course materials related to data-intensive science and scientific discovery from large databases (*i.e.*, Data Science = Discovery Informatics).

Appendix

Listed here are several items that illustrate my most current research activities:

- Journal articles submitted (still in review):
  - Borne, K., "Effective Outlier Detection using K-Nearest Neighbor Data Distributions: Unsupervised Exploratory Mining of Non-Stationarity in Data Streams," submitted, Machine Learning Journal (2010).
  - Bhaduri, K., Das, K., Borne, K., Giannella, C., Mahule, T., & Kargupta, H., "Distributed Change Point Detection for Mining Astronomy Data Streams," submitted, Journal of Statistical Analysis and Data Mining (2010).
- Invited journal articles (in preparation):
  - "Scientific Metadata for Semantic Annotation," Journal of Library Metadata.
  - "U-Science," Journal of Earth Science Informatics.
  - "Tonnabyte Datasets and Citizen Science," Wiley Interdisciplinary Reviews: Computational Statistics .
- Invited book chapters (in preparation):
  - Borne, K., "The Virtual Observatories and Astronomical Data Mining," in Planets, Stars, and Stellar Systems[1] (Springer), in preparation (2010).
  - Borne, K., "Data Mining in the Virtual Observatory," in Advances in Machine Learning and Data Mining for Astronomy[2] (Taylor & Francis), in preparation (2010).
  - Borne, K., & Tyson, J. A., "Surprise Detection in Future Sky Surveys," in Advances in Machine Learning and Data Mining for Astronomy[8] (Taylor & Francis), in preparation (2010).
- Invited talks (not yet presented):
  - January 2011 – e-Science Institute on Next Generation Astronomy (University of Edinburgh, UK) – talk title TBD
  - Spring 2011 – Univ. Pittsburgh Department of Physics and Astronomy – talk title TBD
  - June 2011 – Statistical Challenges in Modern Astronomy (Penn State) – invited conference talk and invited 2-day tutorial on Astroinformatics and scientific data mining in astronomy
  - (date TBD) – King Abdullah University of Science and Technology (Saudi Arabia) – title TBD

[1] http://www.springer.com/astronomy/book/978-90-481-8818-5
[2] http://www.giss.nasa.gov/staff/mway/book/