The VO and Large Surveys: What more do we need?

Kirk Borne

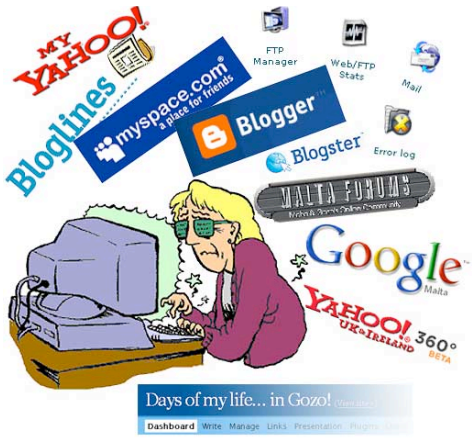
George Mason University

The Changing Landscape of Astronomical Research

- Astronomical data are now accessible from federated distributed heterogeneous sources: **the Virtual Observatory**.
- Astronomy is now a data-intensive science.
- Astronomy will become even more data-intensive in the coming decade with the growth of **massive data-producing sky surveys**.
- **Astroinformatics** (data-intensive astronomical research) will become a stand-alone scientific research discipline (similar to Bioinformatics, Geoinformatics, Cheminformatics, and many others).
  - Informatics is *the discipline of organizing, accessing, mining, & analyzing information describing* ***complex systems*** *(e.g., the human genome, or the Universe).*
  - X-informatics is a key enabler of scientific discovery in the era of data-intensive science. (X = Bio, Geo, Astro, ...) (Jim Gray, KDD-2003)
- Community-based semantics-enabled intelligent data discovery, browse, integration, and mining tools will enable exponential knowledge discovery within exponentially growing data collections.

"If you are not overwhelmed by the rush of new ideas, then you are not paying attention."

- Frank Wilczek (SciFoo 2007)



Summary of Topics

- **VO**: provides the infrastructure and protocols for data discovery & access – DM/ML tools are part of application layer, which is just beginning to blossom – distributed data mining (DDM) is hard work (DMD, not MDD)
  (DMD = Distributed Mining of Data; MDD = Mining of Distributed Data)

- **Large Surveys**: lots of data, with lots of project introspection – lots of project self-focus on the "other DM" = Data Management. The focus on big data should give way to focus on big science (enabling petascale query processing, data manipulation, information retrieval, knowledge discovery)
  … large surveys are producing a Tonnabytes!

- **Informatics/ML/AI/Data Science approach**: the 4th leg of science – the three most important aspects are metadata, metadata, and metadata – for integration, self-learning, recommendations, self-discovery, relevance analysis, dimension reduction, feature selection, constraint-based mining, and more
  … which leads to the main message …

...

Party with your data !

Four categories of ML/AI applications in astrophysics, and some outstanding and upcoming challenges:

1. Cultural challenges
2. Astronomy ML applications
3. (Astro)informatics applications
4. Data Science research challenges

References:

a) arXiv/0906.2173 – *"Data Mining and Machine Learning in Astronomy"* (Ball & Brunner 2009)

b) "Scientific Data Mining in Astronomy", in Next Generation of Data Mining (CRC Press), pp. 91-114 (Borne 2009)

c) Astro2010 Decadal Survey ~~papers~~ manifestos (described later)

Cultural issue #1 that needs to be acknowledged:
Astronomers have been and always will be Data Miners

- Astronomers love to classify things ... (Supervised Learning. e.g., classification)
- Astronomers love to characterize things ...
  (Unsupervised Learning. e.g., clustering)
- And we love to discover new things ...
  (Semi-supervised Learning. e.g., outlier detection)

Cultural issue #2 that needs to be acknowledged:
Astronomical Data Mining is Research

- Obtaining funding for ML/AI in Astronomy has thus far met with mixed success (or worse):
  - *Astronomical data mining is **not research** for the computer science community – it is considered to be an application*
  - *Astronomical data mining is **not research** for the astronomical research community – it is considered to be computing infrastructure*

Astronomical use cases in large surveys that can be advanced with Petascale Machine Learning:

- The distance problem
  - Which combinations (linear or non-linear functions) of observed parameters correlate most strongly with distance (i.e., what is the most accurate estimator)? (e.g., Photo-z)
- Class Discovery among billions of samples
  - Finding new classes (and sub-classes) of objects and behaviors
- Outlier Detection
  - Finding rare, one-in-a-million(billion)(trillion) objects and events
- Novelty Discovery
  - Finding the unknown unknowns
- Correlation Discovery across hundreds of dimensions
  - Finding new patterns and dependencies, which reveal new physics or new scientific principles
- Dimension Reduction
  - Finding principal components, fundamental planes, and condensed representations among hundreds of attributes
- The superposition / decomposition problem
  - Finding distinct clusters (Classes of Object) among 10^{10} objects that overlap in a 10^3-D parameter space (e.g., analogous to star-galaxy separation)

Some Informatics Use Cases that I would like to see:

- Query-By-Example (QBE) science data systems:
  1. "Find more database entries that are similar to this one"
  2. "Find another object that is like this one"
- Automated Recommendation (Filtering) Systems:
  1. "Other users who examined these data also retrieved the following..."
  2. "Other data that are relevant to these data include..."
- Information Retrieval Metrics for Scientific Databases:
  1. Precision = "How much of the retrieved data is relevant to my query?"
  2. Recall = "How much of the relevant data did my query retrieve?"
- Semantic Annotation (Tagging) Services:
  - Report discoveries back to the science database for community reuse
- Transparent reuse and analysis of scientific data:
  - Enabling authentic science experiences in the classroom through inquiry-based learning (http://serc.carleton.edu/usingdata/)
- Key concepts that need defining (by community consensus):
  Similarity, Relevance, the Semantics (dictionaries, ontologies)

Data Science Research Challenge Areas to be addressed the next 10 years

- Distributed Data Mining = algorithms for distributed mining of data
- Scalability of statistical, computational, & data mining algorithms to multi-petabyte scales
- Algorithms for optimization of simultaneous multi-point fitting across massive multi-dimensional data cubes
- Multi-resolution, multi-pole, fractal, hierarchical methods and structures for exploration of condensed representations of petascale databases
- Petascale analytics for visual exploratory data analysis of massive databases (including feature detection, pattern recognition, correlation analysis, clustering, decomposition, eigenvector discovery, dimension reduction)
- Indexing and associative memory techniques (trees, graphs, networks) for highly-dimensional petabyte databases
- Rapid query and search algorithms for petabyte databases

Astroinformatics paper available !

Addresses the data science challenges, research agenda, application areas, use cases, and recommendations for the new science of *Astroinformatics*.

http://mason.gmu.edu/~kborne/Borne_astroinformatics_CDH_FFP_APP.pdf
http://www8.nationalacademies.org/astro2010/publicview.aspx

State of the Profession position paper, submitted to the Astro2010 Decadal Survey
3/15/2009

Astroinformatics: A 21st Century Approach to Astronomy

Authorship: This Position Paper was prepared and endorsed by the following team of 91 astronomers and information scientists (listed separately). The lead author is Kirk D. Borne (Dept. of Computational and Data Sciences, George Mason University, kborne@gmu.edu). The team maintains a web site that hosts information about the authors (including email addresses and links to web sites) and supporting information for this document: http://inference.astro.cornell.edu/Astro2010/ .

Data Science paper available !

State of the Profession position paper, submitted to the Astro2010 Decadal Survey
3/15/2009

http://mason.gmu.edu/~kborne/Borne_data_sciences_education_CDH_EPO.pdf
http://www8.nationalacademies.org/astro2010/publicview.aspx

The Revolution in Astronomy Education

Data Science for the Masses

Authorship: This Position Paper was prepared and endorsed by the following team of astronomers, educators, and information scientists. The lead authors are Kirk D. Borne (Dept. of Computational and Data Sciences, George Mason University, kborne@gmu.edu) and Suzanne Jacoby (LSST Education and Public Outreach, sjacoby@lsst.org).

Summary of Topics: The VO and Large Surveys: What more do we need?

- **VO**: provides the infrastructure and protocols for data discovery & access – DM/ML tools are part of application layer, which is just beginning to blossom – distributed data mining (DDM) is hard work (DMD, not MDD)
  The VO is essential, but alone it is not enough!

- **Large Surveys**: lots of data, with lots of project introspection – lots of project self-focus on the "other DM" = Data Management. The focus on big data should give way to focus on big science (enabling petascale query processing, data manipulation, information retrieval, knowledge discovery)
  Large Surveys are essential, but alone with VO they are not enough!

- **Informatics/ML/AI/Data Science approach**: the 4th leg of science – the three most important aspects are metadata, metadata, and metadata – for integration, self-learning, recommendations, self-discovery, relevance analysis, dimension reduction, feature selection, constraint-based mining, and more
  We need more work here – Send More Money!