Data Mining and Machine Learning in Astronomy: What Next

Nick Ball
Herzberg Institute of Astrophysics
Victoria, BC, Canada

PDFs

- Use PDFs for probabilistic classification, photo-z, etc.
- Avoid a priori cuts
- Improve the signal-to-noise from a given dataset

Machine Learning

- There are a lot of algorithms and tools available
- Select the correct one for the job
- Time domain brings new challenges
- But the data will make more difference than the algorithm

Distributed Data

- Data will be too large to download
- Hence take the analysis to the data
- But it will be distributed over different sites

Make the Data Usefully Available

- Virtual Observatory
- Standard storage schema to make datasets interoperable (e.g. HIA's CAOM)
- High level tools but also accessible to ones own code

Performance Will be I/O Limited

- CPU performance scales to the petascale, unless algorithm worse than n log n
- But I/O will not without substantial infrastructure
- 'Novel' supercomputing hardware: GPGPU, FPGA, Cell
- Parallel programming

Make the ‘Fourth Paradigm’ a Viable Career Focus

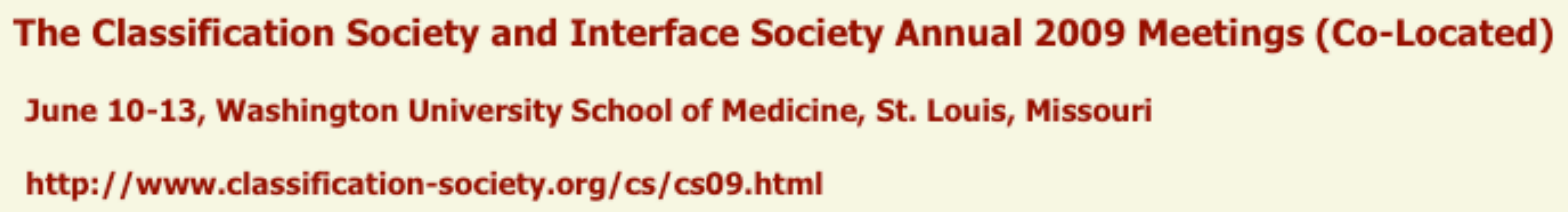
- Jim Gray’s ‘fourth paradigm’: observer, theorist, simulation, now data mining
- Now such that one can specialize in it and produce good science
- Want large scale data mining to be fundable, like sims

Collaboration

- Database experts
- Statisticians
- Hardware
- Software

E.g., Classification Society Conference

- Session on 'Classification in Astronomy'
- They want to increase collaboration with astronomers
- Am organizing similar session next year



The Classification Society and Interface Society Annual 2009 Meetings (Co-Located)

June 10-13, Washington University School of Medicine, St. Louis, Missouri

http://www.classification-society.org/cs/cs09.html

Summary

- PDFs
- Machine learning
- I/O limited
- Distributed data
- Data usefully available
- Fourth paradigm
- Collaboration
- Classification Society conference

arXiv/0906.2173

nick.ball@nrc-cnrc.gc.ca