Computational and Data Sciences

Combining science and computing to meet human needs ...

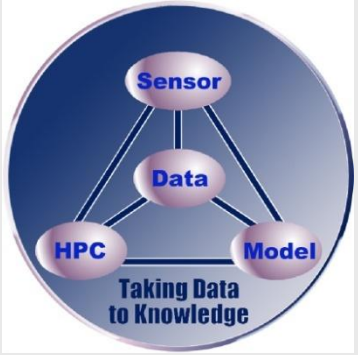
Computational Science and Informatics (Data Science) Programs at GMU

Kirk Borne

George Mason University

School of Physics, Astronomy, & Computational Sciences

http://spacs.gmu.edu/



Outline

- Graduate Program at GMU
- Undergraduate Program at GMU
- Challenge Areas and Reflections

CSI Graduate Program at GMU

- http://spacs.gmu.edu/content/academic-programs
- CSI = Computational Science & Informatics
  - CSI graduate program has existed at GMU since 1991
  - Over 200 PhD's graduated in past 20 years
  - Approximately 90 students currently enrolled
  - About 10% of students end with M.S. (Masters in Computational Science)
  - We have a Graduate Certificate in Computational Techniques and Applications (non-degree professional certification program)
  - We are developing online Distance Education courses

CSI Graduate Program at GMU

- http://spacs.gmu.edu/content/academic-programs
- CSI = Computational Science & Informatics
  - Students can choose a concentration from several choices:
    - Computational Astrophysics
    - Space Science
    - Computational Physics
    - Computational Fluid Dynamics
    - Computational Statistics
    - Computational Learning
    - Computational Mathematics
    - Computational Materials Science (Physical Chemistry)
  - A student may "create" their own concentration, such as one of these previously approved concentrations:
    - Computational Finance
    - Remote Sensing
    - Computational Economics

CSI Graduate Program at GMU

- http://spacs.gmu.edu/content/academic-programs
- CSI = Computational Science & Informatics
  - Note that there is no specific X-Informatics concentration in CSI.
  - However, GMU students can enroll in other departments for specific X-Informatics disciplines:
    - Geoinformatics (including Geospatial Intelligence)
    - Bioinformatics
    - Health Informatics
  - We have discussed adding an AstroInformatics concentration in CSI, but it is not really necessary since interested students are already doing it – i.e., "creating" their own concentration and working with the resident AstroInformatics person (K.B.)

CSI Graduate Program at GMU

- http://spacs.gmu.edu/content/academic-programs
- CSI = Computational Science & Informatics

  Students must complete 4 core courses from this set of 5:
  - CSI 700 Numerical Methods
  - CSI 701 Foundations of Computational Science
  - CSI 702 High Performance Computing
  - CSI 703 Scientific and Statistical Visualization
  - CSI 710 Scientific Databases

  There are also many electives (at least 5 additional CSI courses are required, plus concentration science electives).

  For example:
  - Data Mining, Knowledge Mining, Computational Learning, Statistical Learning, Computational Statistics, Statistical Graphics, Data Exploration, etc.

CSI Graduate Program at GMU

- http://spacs.gmu.edu/content/academic-programs
- CSI = Computational Science & Informatics

  Students must complete 4 core courses from this set of 5:
  - CSI 700 Numerical Methods
  - CSI 701 Foundations of Computational Science
  - CSI 702 High Performance Computing
  - CSI 703 Scientific and Statistical Visualization
  - CSI 710 Scientific Databases

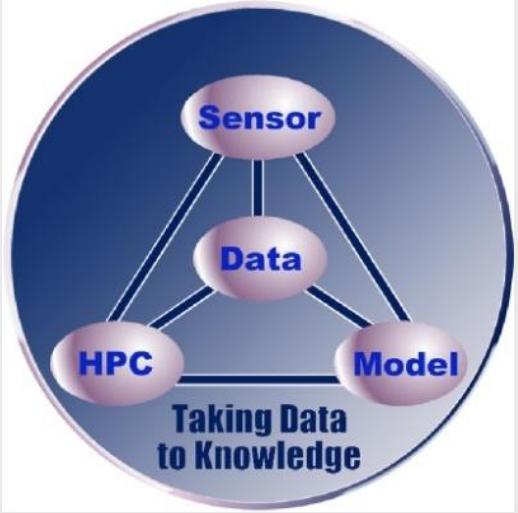
  There are also many electives (at least 5 additional CSI courses are required, plus concentration science electives).

  For example:
  - Data Mining, Knowledge Mining, Computational Learning, Statistical Learning, Computational Statistics, Statistical Graphics, Data Exploration, etc.

Informatics Courses !

Example Course Syllabus: CSI 710 – Scientific Databases

- CSI 710 Scientific Databases (taught by K.B.) – lectures include:
  - Relational Databases: Modeling, Schemas, Normalization, SQL
  - Scientific Databases, Big Data in Science, The 4th Paradigm
  - E-Science, Ontologies, Semantic E-Science, X-Informatics
  - Distributed Data, Federated Data, Virtual Observatories
  - Citizen Science with Big Data
  - Scientific Data Mining I
  - Scientific Data Mining II
  - Astroinformatics and Astro databases
  - Bioinformatics and Bio databases
  - Geoinformatics and Geo databases
  - Health Informatics
  - Online Science (Jim Gray's KDD-2003 lecture)
  - Intelligent Archives of the Future



Outline

- Graduate Program at GMU
- Undergraduate Program at GMU
- Challenge Areas and Reflections

CDS Undergraduate Program at GMU

- http://spacs.gmu.edu/content/academic-programs
- CDS = Computational and Data Sciences
  - Undergraduate B.S. degree program at GMU since 2008
  - Graduating ~2 students per year
  - 10 students currently enrolled
  - … this is considered "under-performing"
  - … we need strategies for increasing enrollment.

CDS Undergraduate Program at GMU

- http://spacs.gmu.edu/content/academic-programs
- CDS = Computational and Data Sciences
  - Undergraduate B.S. degree program at GMU since 2008
  - Graduating ~2 students per year
  - 10 students currently enrolled
  - … this is considered "under-performing"
  - Strategies for increasing enrollment:
    - Target marketing – send email to all "local" high school students interested in science *and* interested in GMU College of Science programs already
    - Advertising materials posted on campus and emailed to existing GMU undergraduates
    - Modify the curriculum to satisfy student interests (e.g., "Computing for Scientists"**)
    - Future idea: find a sponsor for a CDS scholarship program
    - REU = Research Experiences for Undergraduates: students want this experience!
    - … or … wait and see what happens, because …
    - Enrollment is already increasing this year due to a new course** that is very exciting for the students and is attracting ~100 students each semester, some of whom have already converted to CDS and several others are thinking about it.

CDS Undergraduate Program at GMU

- http://spacs.gmu.edu/content/academic-programs
- CDS = Computational and Data Sciences
  - Students can choose the general CDS degree, or else choose one of these concentrations:
    - Physics
    - Chemistry
    - Biology
  - A student may "create" their own concentration. For example:
    - Environmental Science
  - There is no AstroInformatics option, but one course is being modified to include this material, plus students can take Independent Study and Independent Research courses that cover this material.

CDS Undergraduate Program at GMU

- http://spacs.gmu.edu/content/academic-programs
- CDS = Computational and Data Sciences
  - The DATA SCIENCE component of the curriculum was developed with the support of a grant from the NSF (National Science Foundation):
    - CUPIDS = ***C****urriculum for an* ***U****ndergraduate* ***P****rogram* ***I****n* ***D****ata* ***S****ciences*
  - Primary Goal: *to increase student's understanding of the role that data plays across the sciences as well as to increase the student's ability to use the technologies associated with data acquisition, mining, analysis, and visualization.*
  - Objectives – students are trained:
  - … to access large distributed data repositories
  - … to conduct meaningful inquiries into the data
  - … to mine, visualize, and analyze the data
  - … to make objective data-driven inferences, discoveries, and decisions

CDS Undergraduate Program at GMU

- http://spacs.gmu.edu/content/academic-programs
- CDS = Computational and Data Sciences

Students must complete 6 core courses from this set of 8 courses:

- CDS 101 – Introduction to Computational and Data Sciences
- CDS 130 – Computing for Scientists
- CDS 301 – Scientific Information and Data Visualization
- CDS 302 – Scientific Data and Databases
- CDS 351 – Introduction to Scientific Programming
- CDS 401 – Scientific Data Mining
- CDS 410 – Modeling and Simulations I
- CDS 411 – Modeling and Simulations II

Additional required courses include Math, Statistics, Computer Science, Physics I and II, plus courses in student's chosen science concentration

CDS Undergraduate Program at GMU

- http://spacs.gmu.edu/content/academic-programs
- CDS = Computational and Data Sciences

Students must complete 6 core courses from this set of 8 courses:

- CDS 101 – Introduction to Computational and Data Sciences
- **CDS 130 – Computing for Scientists – this is new & very popular!**
- CDS 301 – Scientific Information and Data Visualization
- CDS 302 – Scientific Data and Databases
- CDS 351 – Introduction to Scientific Programming
- CDS 401 – Scientific Data Mining
- CDS 410 – Modeling and Simulations I
- CDS 411 – Modeling and Simulations II

Additional required courses include Math, Statistics, Computer Science, Physics I and II, plus courses in student's chosen science concentration

CDS Undergraduate Program at GMU

- http://spacs.gmu.edu/content/academic-programs
- CDS = Computational and Data Sciences

Students must complete 6 core courses from this set of 8 courses:

- CDS 101 – Introduction to Computational and Data Sciences

An accompanying "science lab" course is needed, to go along with CDS 101 … this course CDS 102 will be taught starting Spring 2012.

- CDS 130 – Computing for Scientists
- CDS 301 – Scientific Information and Data Visualization
- CDS 302 – Scientific Data and Databases
- CDS 351 – Introduction to Scientific Programming
- CDS 401 – Scientific Data Mining
- CDS 410 – Modeling and Simulations I
- CDS 411 – Modeling and Simulations II

CDS Undergraduate Program at GMU

- http://spacs.gmu.edu/content/academic-programs
- CDS = Computational and Data Sciences

Students must complete 6 core courses from this set of 8 courses:

- CDS 101 – Introduction to Computational and Data Sciences
- CDS 130 – Computing for Scientists
- ???? 201
- ???? 202
- CDS 301 – Scientific Information and Data Visualization
- CDS 302 – Scientific Data and Databases
- CDS 351 – Introduction to Scientific Programming
- CDS 401 – Scientific Data Mining
- CDS 410 – Modeling and Simulations I
- CDS 411 – Modeling and Simulations II

Notice the "sophomore gap". There is something missing here that we hope to add next year … Calculus-based CDS, similar to Physics.

CDS Undergraduate Program at GMU

- http://spacs.gmu.edu/content/academic-programs
- CDS = Computational and Data Sciences

  We are also "infiltrating" the entire undergraduate program at GMU through 3 of our courses that satisfy university General Education graduation requirements **for all students at the university**:

- CDS 101 – Introduction to Computational and Data Sciences
  - **Satisfies GMU's Natural Science requirement**
- CDS 130 – Computing for Scientists
  - **Satisfies GMU's I.T. requirement**
- CDS 151 – Data Ethics
  - **Satisfies GMU's Ethics requirement**

Example of Learning Objectives: CDS 401 – Scientific Data Mining

- Be able to explain the role of data mining within scientific knowledge discovery.
- Be able to describe the most well known data mining algorithms and correctly use data mining terminology.
- Be able to express the application of statistics, similarity measures, and indexing to data mining tasks.
- Identify appropriate techniques for classification and clustering applications.
- Determine approaches used for mining large scientific databases (e.g., genomics, virtual observatories).
- Recognize techniques used for spatial and temporal data mining applications.
- Express the steps in a data mining project (e.g., cleaning, transforming, indexing, mining, analysis).
- Analyze classic data mining examples and use cases, and assess the applicatio of different data mining techniques.
- Effectively prepare data for mining.
- Effectively use software packages for data exploration, visualization, and mining.

Outline

- Graduate Program at GMU
- Undergraduate Program at GMU
- Challenge Areas and Reflections

Challenge Areas
(primarily for the undergraduate program, since the graduate program is doing very well)

- Attracting students: nobody says *"I want to be a computational and data scientist when I grow up."*
- Visibility: most other science departments are not aware of the importance of our courses for their majors. (Note: Biology and Neuroscience now require our Science Computing course.)
- Retention: after Year 1 courses, there are no other CDS courses until junior year (Year 3)
- Lab course: in order to attract sufficient enrollment, the CDS 101 General Education Natural Science course should be accompanied with a lab section.

- Scientific computing course: this was identified 2 years ago as a necessary course to attract students to CDS … we now have this course, and it is very popular (nearly 100 students each semester, and growing … up to 400+)

Reflections

- Students with a broad interest in computers and sciences will benefit from these types of programs: Computational **and** Data Sciences.
  - Actual quote from high school senior visiting the university:

    "I plan to major in biology, but I wish I could do something with computing also."
- Students graduating with a traditional discipline-based bachelors degree in biology, chemistry, or physics generally do not have the required computational background necessary to participate as productive members of modern interdisciplinary scientific research teams, which are becoming increasingly computational- and data-intensive.
- The motivating theme and goal of these programs should be to train the next-generation scientists in the tools and techniques of *cyber-enabled science (e-Science) to prepare them to* confront the emerging petascale challenges of data-intensive science.
- It is also good for society in general that all members of the 21st century workforce are trained in computational and data science skills – i.e., computational literacy and data literacy are critical for all.