NISS

Statistical Computing Infrastructure Initiative: SCI^2

Alan Karr
March 4, 2004

The Question

Statistics
+
Cyberinfrastructure
+
Domain science
=
?????

History

- January 2003: Atkins report
- October 2003: CISE reorganization
- December 2003: NISS planning meeting
- February 2004: SCI workshop
- ????: NSF solicitation


Revolutionizing Science and Engineering Through Cyberinfrastructure:

Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure

January 2003

NISS Working Group

- David Banks (Duke)
- Stephen Eick (SSS Research/UIC)
- Todd Graves (LANL)
- Robert Grossman (NCDM/UIC)
- Susan Holmes (Stanford)
- Alan Karr (NISS)
- Diane Lambert (Bell Labs)
- Duncan Temple Lang (UC Davis)
- Padraic Neville (SAS)
- Ashish Sanil (NISS)
- Lee Wilkinson (SPSS)
- Paul Whitney (PNNL)

Some Initial Issues

- Statistical theory, methodology and software for data stored in distributed, relational databases
  - Violations of common statistical assumptions, such as independence, in data stored in RDBMSs
- Explicit means of incorporating domain knowledge (Example: attribute A is known to be of low quality) into large-scale, automated statistical analyses
- Interoperability of statistical software systems
- Inference and graphics for data streams

More Initial Issues

- Seamless links between statistical software systems and tools Example: Excel) employed by consumers of statistics (rather than statistical researchers)
- Natural language tools for translating among software systems
- Business models for statistical software, including usage-based payment.
- Statistical computing infrastructure capable of accommodating multiple user communities (Examples: researchers, business personnel, military intelligence)
  - Conveying uncertainties (possibly with visualizations).
- Grid computing; computing models that "bring the code to the data," rather than vice versa

"Coalesced" Themes—1

- Data and Metadata Issues
  - Divergence between new and emerging forms (and formats of data, as well as the scale of the data, and statistical abstractions, theory and methodology based on flat files
  - Confidentiality and security
- Software Abstractions and Interoperability
  - Recursive descent parsers
  - Universal executor for statistical languages
  - Open source software
  - Selective transparency

"Coalesced" Themes—2

- Graphics and Visualization
  - Information visualization and data visualization are both important
  - Data-driven selection of graphics
  - " Interestingness" of different visualizations; custom visualizations
  - Metadata visualization
- Uncertainties
  - Primary products that distinguish us as statistical scientists
  - Effective tools for communicating and visualizing} them, especially for "consumers of statistics," are lacking
  - Visualization is a natural path to pursue, but past approaches (e.g., blurring or the width of confidence bands as visual metaphors for uncertainty) seem ineffective or non-scalable.
  - Can Bayesian approaches help?

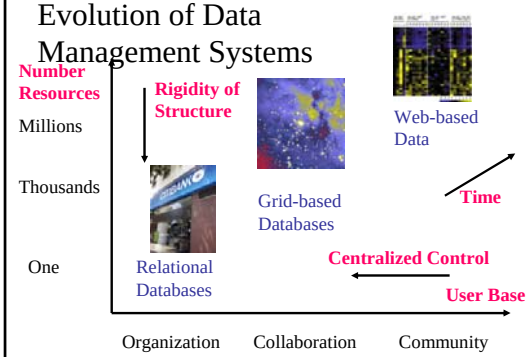
Other Items

- SCI for simulation purposes, such as MCMC
- Interactions with other large-scale computing (Example: complex scientific models)
- Curricular implications
- Cultural issues, in particular, the perceived disconnect between university reward (and support) structures and statistical computing
- NISS or some other organization (ASA?) sponsoring a week-long workshop on statistical computing
- Multi-processor version of R

An Emerging Perspective

- Most of the scientific world cannot afford "grid" computing
  - Too expensive
  - Too customized
  - Too standardized
  - Doesn't meet needs
- For this part of the scientific world, data and metadata are part of the cyberinfrastructure
- If cyberinfrastructure is construed as cycles and pipes, NSF is in danger of leaving this part of the scientific world behind

Evolution of Data Management Systems



A Matrix View

- Processes
  - Data discovery
  - Data cleaning (characterization of DQ)
  - Data integration (fusion)
  - Data exploration
  - Modeling and analysis
- Technologies
  - Search
  - Privacy and confidentiality
  - Metadata generation and model management
  - Visualization
  - Analysis tools

Questions

- Where next?
- Who wants to join?