Massive Data Discussions Summary

thrust leaders: Ken Joy, Tamara Munzner

Mathematical Foundations of Scientific Visualization, Computer Graphics, and Massive Data Exploration (@ BIRS)

27 May 2004

Features

- the Hard Part
- "what do you want to know?"
- domain-dependent semantics
- finding the right abstraction
- even more critical for massive data
- infovis language: visual encodings

Defining "massive"

- ratio between features and data
  - is overview so simplified as to be useless?
- uses lots of computational resources
- multivalued at each point
- time-varying
- complexity of algorithms you can afford
  - only linear? n log n? below n^2!
- doesn't all fit into:
  - local disk
  - main memory
  - screen
  - human mind

Current practice

- store
- ignore
- dig through manually
- computational steering during run
  - see problem, kill run early, restart
  - now happening at LLNL, sim outputs data in viz format
- computational steering between runs
  - "hero run": simulate for two weeks
  - look at results with visualization (mpeg movie)
  - find interesting few frames
  - do next run of just that period, with smaller timesteps

Driving problems: collecting data

- simulation
  - CFD, engr. analysis, high-energy physics, microprocessor design
- sensors
  - MRI, visible human, human genome project, satellites, sky surveys
  - future scenario: globelog
- logging
  - telephone networks, web traffic, network traffic

Driving problems: using data

validation

debugging
confirming hypotheses
monitoring

exploration

few/no a priori ideas

exposition

illustrate known things for others

- "overview, zoom and filter, details on demand"
  - infovis mantra, Shneiderman 1996

Conferences/Journals

- Vis, TVCG, InfoVis, SIGGRAPH
  - SIGMOD (Management of Data), VLDB (Very Large Databases), SoCG (Symp on Computational Geometry)
- VisSym, VolVis, PVG, Supercomputing, IEEE CG&A, Graph Drawing, I3D (Symp. on Interactive 3D Graphics), UIST (User Interface Software and Technologies), InfoVis Journal
  - IEEE CS&E (Computational Science and Engineering)
- VDA (Visual Data Analysis), CGIM (Computer Graphics and Imaging), CGI (Computer Graphics International), Journal of Visualization and Animation

Online resources – few!

- software
  - sourceforge.net
- data
  - visible human, www.nlm.nih.gov/research/visible/
  - www.cs.umd.edu/hcil/InfovisRepository
  - KL Ma's dataset collection
    not available on web for outsiders
  - raw data not enough
    need tasks/metadata/problem description
  - often datasets not ours to release

Other useful areas

- software engineering
- mathematics
- numerical methods
- statistics, data analysis
- databases
- vision/image processing
- cognitive and perceptual psychology
- data mining
- human-computer interaction
  - user-centered design, ethnography
- expansionists vs. isolationists
  - which stuff is our problem?