eScience and Computer Science

Jim Gray
Microsoft Research

Sensor Networks Summer School @ Semiahmoo
eScience: What is it?

- Synthesis of information technology and science.
- Science methods are changing.
- Science is being codified/objectified. How represent scientific information and knowledge in computers?
- Science faces a data deluge. How to manage and analyze information?
- Scientific communication changing.
How We Engage With An Area

- eScience is inter-disciplinary
- We bring informatics expertise

Process:
1. Find someone who is desperate.
2. Start with requirements: 20 questions
3. Help build systems to:
   - Answer those questions faster
   - Answer new questions.
4. Long-term and deep collaborations
Water is Different

• Water crosses disciplines
  – Engineering, geology, natural resources, physical science, climate,
  – USBR, USGS, Army Corps, DWR, 200+ local water boards
  – “we only see water through the salmon lense”

• Water data is widely distributed
  – All of the above agencies take some sort of data (including overlap, including cross hiring)
  – Lots of small (perhaps cleaned) data sets on the internet

• Water data may not be “data”
  – Water planning done at many levels with different inputs, outputs, and data terms – comparisons aren’t easy
  – Existing data often piecemeal
    – nice pictures, but not science
Water Collaborators

• Collaborating with Berkeley Water Center
  – Prototyping portal for Ameriflux carbon flux data
  – Now selecting water data for second generation portal

• Emerging collaboration with CUASHI HODM
  – Leverage their thinking
  – Can we come up with a common schema to allow datasets to be curated and merged at will?
What’s Wrong In These Pictures?

• Visualization allows simple quality assurance sanity checks
  – We started with “clean” data after all

• Shopping for other visualization tools now to address initial science questions
From Raw Data to Useful Quantities

Calibrations in the Lab

- Temperature sensor A/D units
- Reference voltage A/D units
- Moisture sensor A/D units
- Light Intensity A/D units
- Air Temperature A/D units
- CPU clock
- UTC DateTime

Voltage

Resistance

Soil Temperature Calibration

Moisture Sensor Calibration

Water Deficit Soil Matrix Potential

Soil Water Potential -> Volumetric Conversion

Water Content Volumetric

Temperature Conversion

Celsius
to improved surface and groundwater models [Carter2008], improve better irrigation plans [USCC], resource management. Soil ecologists will be able to better predict where and when the mineral invertebrate activity occur. This activity is tightly coupled with soil respiration, which is an increasingly unknown component of the global carbon cycle. Continuous in-situ monitoring of these soil respiration estimates the contribution of the soil to these large-scale processes.

Figure 1. Air and soil temperatures over a period of a year. Each point represents six hour averages. Top: soil temperature at 10 cm depth. Bottom: air temperature at soil surface. Black and Blue: maximum and minimum temperatures, respectively, for the Baltimore Metropolitan Area.

Figure 2. Soil moisture readings over six weeks recorded by six nodes. Each point represents six hour averages. Bars on the bottom indicate precipitation events in the Baltimore Metropolitan Area. Highest column (Feb 25) corresponds to 25.4 mm rain.
Sensor Dimension Model

Temperature

- all
- year
- season
- week
- day

Sensor Dimension Model

- all
- category
- depth
- hour
- tenMinute

Sensor Dimension Model

- all
- sensor

Temperature
Standard Sensor Queries

1. Temperature (average, min, max, standard deviation, standard error, count-missing)
   for a particular time or interval (e.g. when animal samples are taken)
   10-minute, hour, day, week, season, year, all
   (s1) for one sensor
   (s2) for a patch (sensor and its immediate neighbors)
   (s3) all sensors at the site
   (s4) all sites
   vs depth (air, surface, 10cm, 20cm, 50cm)
   vs time
   (t1) 10 minute
   (t2) hour
   (t3) day
   (t4) week
   (t5) season
   (t6) year
   (t7) all time
   vs category
   land use,
   land cover,
   age of vegetation,
   crop management type,
   upslope, downslope,
   etc

2. Look for unusual patterns, outliers: a mote behaving differently, unusual spike, etc.

3. Look extreme events: e.g. rainstorm, people watering their lawns, etc.
   And show data in time-after-event coordinates

4. Correlate with another dataset (e.g. with weather data, the CO2 flux tower data, runoff data, etc)

5. Visualize the habitat heterogeneity, preferentially in 3-D integrated with maps (e.g. 3-D with LIDAR maps, 2-D with vegetation data, animal density data, etc.)

6. Notify me if the data has unexpected values (this is the “real time” thing),
sensors might be damaged, and need to be checked, etc.