Scientific Data Management

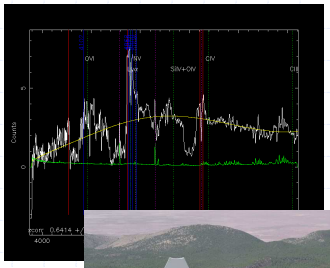
Dr. Laura Bright
Bill Howe

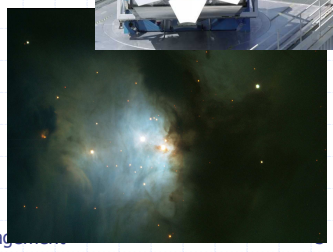
Biology

- Old way:
  - Wet lab chemistry
- New way:
  - Microarray
  - Search GenBank, Ensembl, GDB, SwissProt, Entrez using BLAST, FASTA, GCG, EMBOSS

Astronomy

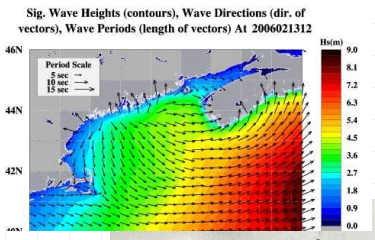
- Old way:
  - Sign up for telescope time
- New way:
  - Sloan Digital Sky Survey
    - Systematically mapping ¼ of the entire sky
    - 12 TB to date,
    - 15 TB final in 2007





Oceanography

- Old way:
  - Field work
  - Simplified Calculations
- New way:
  - Finite Element Analysis
  - In situ sensors
  - CODAR



Science is Changing

- Old Science: "Query the world"
  - *Data acquisition is the dominant cost*
- New Science: "Download the world"
  - *Data analysis is the dominant cost*

Course Structure

- 10% In-class exercises
- 10% Study Questions
- 40% Homework Assignments
- 15% Mini-project
- 25% Short Paper (3 pages)
- No exams

Short Paper Assignment (1/2)

- To be completed individually!
- Compare/Contrast a pair of papers
  - We provide a list to choose from

Short Paper Assignment (2/2)

- 25% = 3 milestones + final paper
  - 2 points: select paper pair. (~ week 3)
  - 5 points: a half-page summary of each paper; one page total. (~ week 5)
  - 3 points: a list of 3 points of contrast/comparison, in complete sentences. (~ week 7)
  - 15 points: Final paper (~ week 11)
  - Both content and mechanics matter!

Study Questions

- Covers the readings
- Discussion ok, but write up your own answers
  - Dr. Bright's "Pizza rule"
  - Try to keep the discussion on the list
- 3-4 questions per set, about 1 set per week
- Details:
  - About a paragraph; use complete sentences
  - Feel free to use diagrams or figures when appropriate!
  - Due at the beginning of class on the due date

Homework Assignments

- Covers Tools (rather than readings)
- To be completed individually!
- Send questions to the instructors rather than the list

Late work

- **Prior** approval is necessary, but not always sufficient

Course Web Page

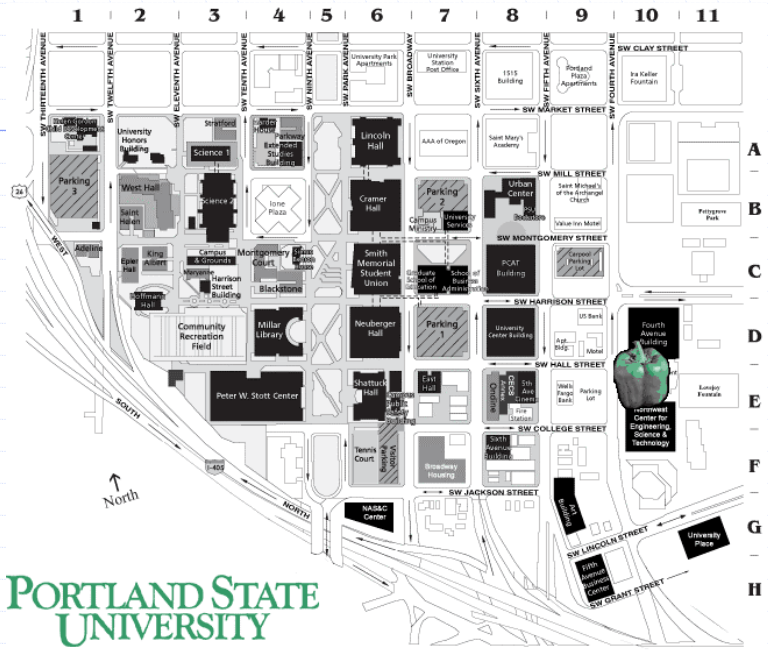
- http://www.cs.pdx.edu/~howe/cs410

material	web page	hard copy
lectures	Yes	No
readings available online	Yes	No
copy-sensitive readings	No	Yes
study questions	Yes	Yes
homework	Yes	Yes

- We hope to post class materials at least an hour before class (no promises)
- Extra copies of printed material will be available outside Dr. Bright's office (FAB 310-24)

Office Hours

- Howe: FAB 310-C
  - Monday 4-6 (or by appointment)
- Bright: FAB 310-24
  - Thursday 1-3 (or by appointment)



Course Email List

- "scidata"

https://webmail.cecs.pdx.edu/mailman/listinfo.cgi/scidata

- Ok to discuss study questions
- Not ok to discuss homework answers
- Send HW Questions to instructors

Academic Integrity

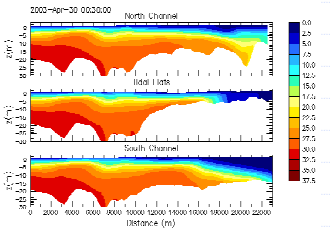
- 2004-2005 PSU Catalog pages 29-30
- Posted on the web page

A First Class Exercise

- **1) Name** (feel free to add pronounciation hints!)
- **2) Email** you wish to use for this class
- **3) How much experience with RDBMS**?
  - (A) What's an RDBMS?
  - (B) I've taken CS 386, but that's it
  - (C) I've used an RDBMS on a few projects
  - (D) I write SQL semi-daily
  - (E) I'm a DBA
- **4)** How might Scientific Data Management be different than "regular" data management?

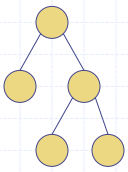
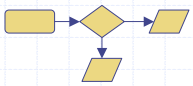
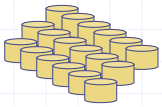
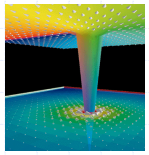
(Scientific Data) Management

- Interesting data types
  - Gene sequences, spatio-temporal objects,
  - scalars, vectors, tensors
  - map layers, images, meshes
  - unstructured metadata
- Interesting Scale
  - Terabytes becoming Petabytes
- Interesting Access patterns
  - Data "products"
  - Data "releases"



Scientific (Data Management)

- Readings drawn from database literature
- We will consider:
  - Conventional technology
    - Relational databases
    - Web Services/XML
  - Specialized technology
    - GIS
    - Grid
    - Workflow
    - Visualization
- Emphasis on Case Studies

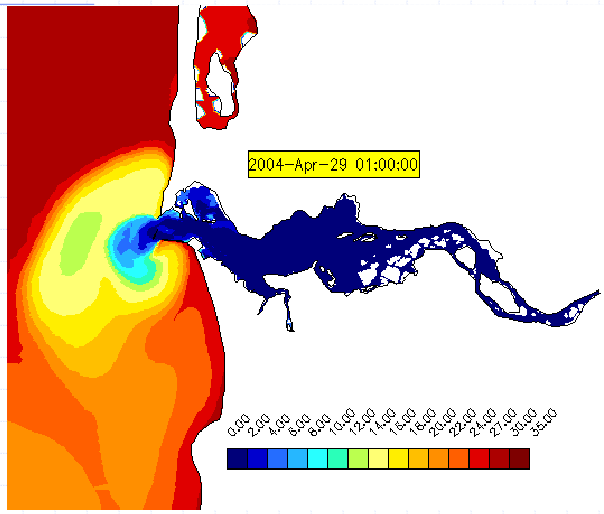
Characterizing SDMS (1/3)

- What logical data types are involved?
  - DNA sequences,
  - maps of the earth, rivers, lakes
  - maps of the sky, galaxies, stars
  - Particle trajectories
- What physical data types are involved?
  - Multimedia?
  - Multidimensional arrays?
  - Spatio-temporal objects?
  - "ordinary" tuples?

Characterizing SDMS (2/3)

- Who are the Customers?
  - Other Researchers
  - General Public
  - Policy Makers
  - Emergency Workers
  - Commercial

Customers?



Characterizing SDMS (3/3)

- What is the Architecture?
  - Pipeline (Workflow)
  - Archive (Database)
  - Clearinghouse (Portal)
- What Interfaces are supported?
  - Browse
  - Query
  - Upload
  - Derive
  - Script (Web Services)

More Examples

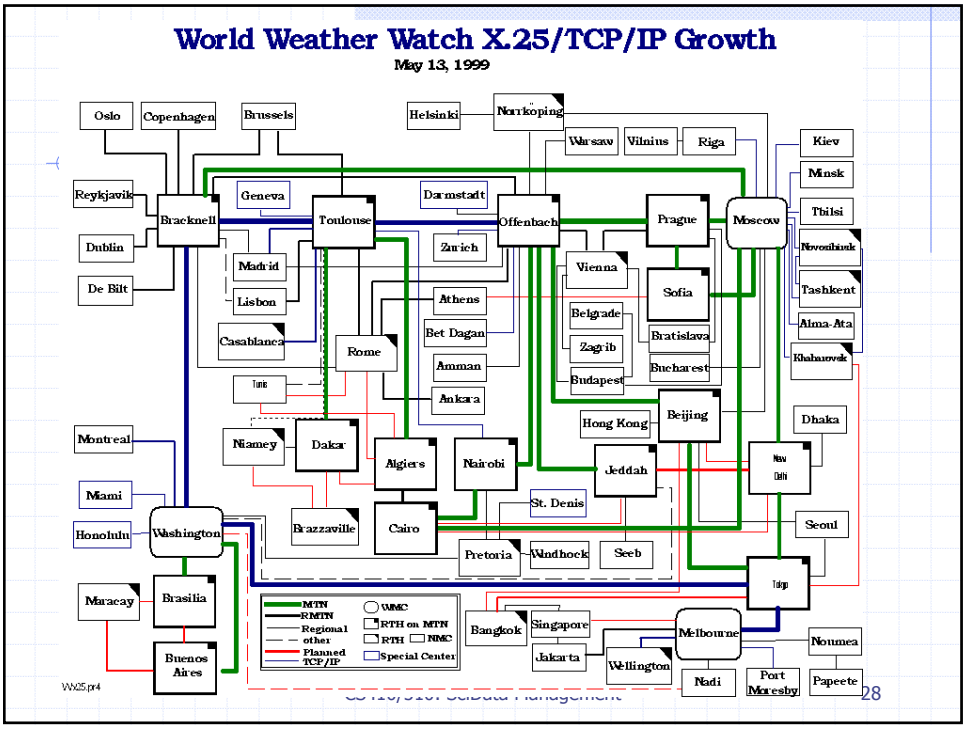
- geodata.gov
  - governmental GIS clearinghouse
- EOSDIS
  - NASA's satellite image repository
- IOOS
  - Ocean measurement and forecasting
- Others?

National Weather Service: Timeline

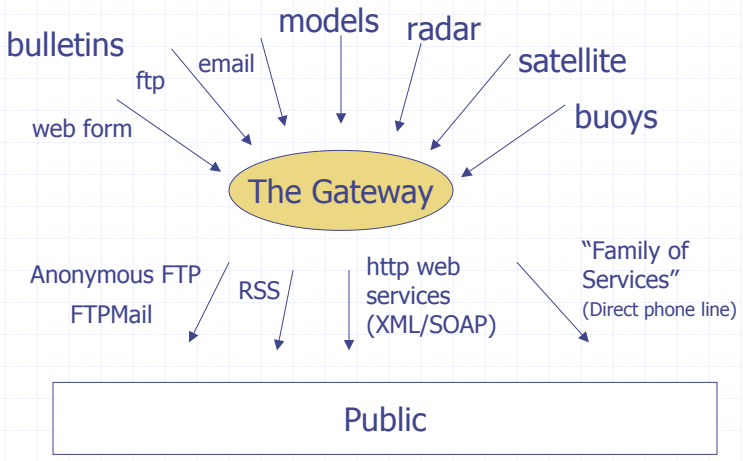
- **1849:** Smithsonian Institution provides weather instruments to telegraph operators
- **1900:** Galveston Hurricane
- **1935:** Long range forecasts; buoys
- **1955-1960:** Computer forecasts scheduled regularly; weather satellite TIROS I launched.
- **1979**: AFOS Computer system is deployed, connecting all Weather Service forecast offices.
- **1988:** Weather Service mobilizes local forecasting operation to assist in fighting week-long wildfire in Yellowstone park
- **1990:** NEXRAD Radar deployment project; a Cray supercomputer deployed

National Weather Service

- Data Collection
  - Radar
  - Satellite
  - Forecasts
  - Bulletins
- Data Dissemination
  - Radio: aviation, marine, military channels
  - FTP, HTTP, email, RSS: public
- Part of a UN sponsored Gobal network



NWS: Gateway



National Weather Service: Products (1/2)

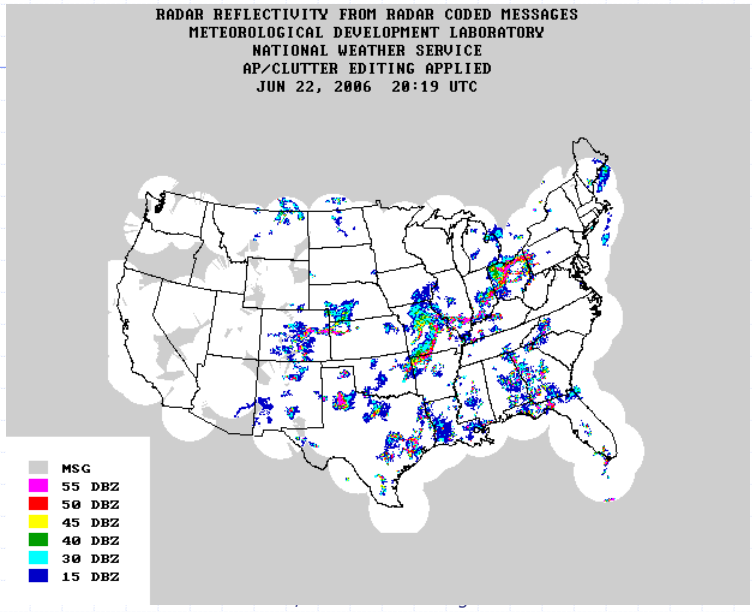
- **Computer Models**
  - GRIB files from 10+ models from regional to global scale
    - Example:

      SL.008001/ST.opnl/MT.ruc_CY.06/RD.20000622/PT.grid_DF.gr1/fh.0003x_tl.press
  - Facsimile/Images
  - Text products derived from models
  - Special products in special formats
- **Text Products** -
  - Warnings, outlooks, advisories, forecast, discussion
  - ~100 different types

National Weather Service: Products (2/2)

- **Observed Data** -
  - kept for 24 hours at least
  - observations from aviation, buoys, ships, balloons
  - special formats, but some have parsed them to XML
- **Radar Products** -
  - Multicast by connecting a router directly to NWS as well as FTP
  - SL.us008001/DF.of/DC.radar/DS.p19r1/SI.kfws/sn.0114
- **Satellite Products** –
  - Cloud Water Vapor, Cloud Liquid Water, Rain Rate, Sea Ice Concentration, Sea Ice Age, Sea Ice Edge, Soil Moisture, Surface Wind, Water Vapor over oceans, Surface Temperature, Snow Water Content, Cloud Amount, and EDR Surface Type

National Weather Service: Radar



National Weather Service: Forecasts (1/3)

- Several Climate Models:
  - Weather Research and Forecast (WRF)
  - Global Forecast System (GFS)
  - North American Mesoscale (NAM)
  - Nested Grid Model (NGM)
- Specialized Models:
  - Fire Weather
  - Hurricane
  - Aviation

National Weather Service: Forecasts (2/3)

- National Digital Forecast Database
  - 3 hr temporal resolution
  - 5km spatial resolution
  - GRIB files, GIS map layers, data products

National Weather Service: Forecasts (3/3)

- Model Output Statistics (MOS)
  - Examples:
    - Max/Min Temperature Forecasts
    - Surface Temp / Dewpoint Forecasts
    - Opaque Cloud Amount
    - Probability of Precipitation
    - Severe weather probabilities
  - MOS products

National Weather Service: Satellites

- Geostationary Operational Environmental Satellites
- Variety of images and products

National Weather Service: Summary

- Domain?
- Customers?
- Architecture?
- Interfaces?