# Introduction to Scientific Data and Workflow Management

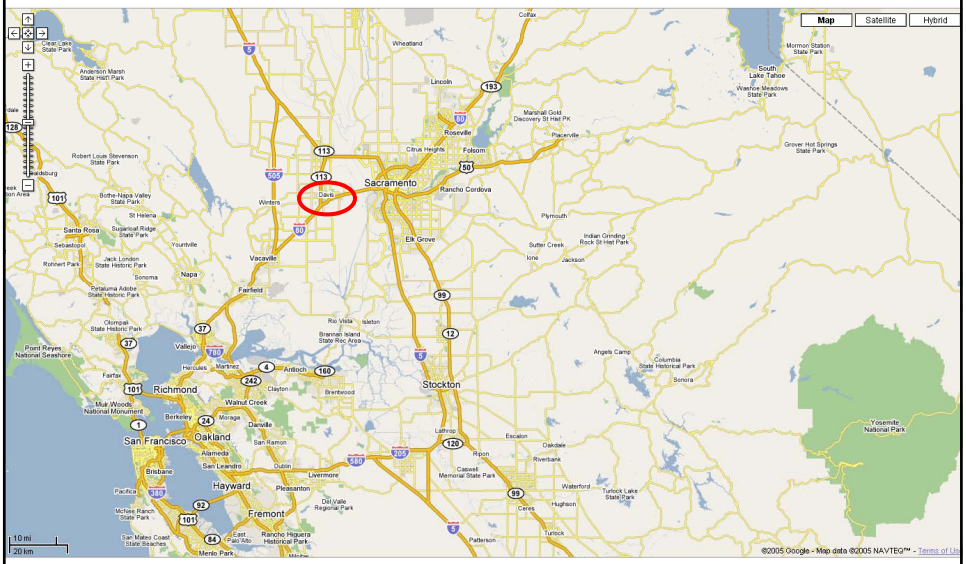

Michael Gertz
**gertz@ucdavis.edu**

Bertram Ludäscher
**ludaesch@ucdavis.edu**

Department of Computer Science
University of California at Davis



## Where we are …



## Outline of the Tutorial Modules

### I. Overview on Scientific Data Management

(13:30-14:15, Gertz)

- What is Scientific Data Management (SDM)
- Typical SDM processes
- SDM domains: Earth Sciences, Astrophysics, Life Sciences

### II. From Conventional to Scientific Data Integration

(14:15—15:00, Ludäscher)

- Conventional Data Integration (Mediation/Schema-based)
- Ontology-based Extensions to Data Integration
- The Role of Metadata
- Link-based “Integration”



## Outline of the Tutorial Modules

### III. From Scientific Data Formats to Data Stream Processing

(15:30—16:15, Gertz)

- Scientific Data Formats and Data Models
- What are Recent Trends ? What about XML ?
- Data Processing Pipelines
- Data Stream Processing

### IV. Introduction to Scientific Workflows

(16:15-17:00, Ludäscher)

- Workflows in e-Science and Cyberinfrastructure
- Scientific Workflows vs Business Workflows
- Features of a Scientific Workflow system (Kepler)
- Flow-based Programming and Scientific Workflow Design
- Semantic Extensions

## Focus of this tutorial

- **Broad overview on/introduction to …**
  - scientific application domains
  - how scientists do data management in these domains
  - open problems where database skills might help
  - opportunities to apply database models and techniques
  - opportunities to develop new database models and techniques
  - building bridges to other science communities (CS+X, DB+X)
- **Some specific details, e.g., on …**
  - scientific data formats
  - semantic extensions for scientific data integration
  - data stream processing framework
  - scientific workflows modeling and design



## Data Management

**An attempt at a definition...**

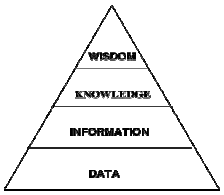
*"Data management encompasses all the disciplines related to managing data as a valuable resource."*
(Wikipedia)

*"The control of data handling operations, such as acquisition, analysis, translation, coding, storage, retrieval, and distribution of data, but not necessarily the generation and use of data".*
(Alliance for Telecommunication Industry Solutions, ATIS)

"*Data Resource Management is the development and execution of architectures, policies, practices and procedures that properly manage the full data lifecycle needs of an enterprise."*
(Data Management Association, DAMA)

## Scientific Data Management

- **Scientific Data Management** (SDM) employs data management concepts and techniques and tailors them to ***domain-specific research goals*** and scientific data resource management:
  - SDM is not only about managing (raw) data but also about managing and exploiting ***information*** and ***knowledge***
  - SDM often involves ***very large data sets***




  - Scientific data is often more ***context-dependant*** and not targeted towards an enterprise or business
  - Scientific data is typically more ***complex*** (in structure and semantics) than business data, i.e., not just tuples or XML documents
  - Scientific data require ***operations*** that go far beyond standard data processing and query techniques known from databases
  - ***Complex processes*** are associated with generating, analyzing, and disseminating scientific data



## Scientific Data Management (2)

**Notions and concepts associated with SDM …**

*Science objectives*
*Hypothesis*
*Knowledge & Information*

*Simulation*
*Experimentation*
*Observation*

***Scientific Data Management***

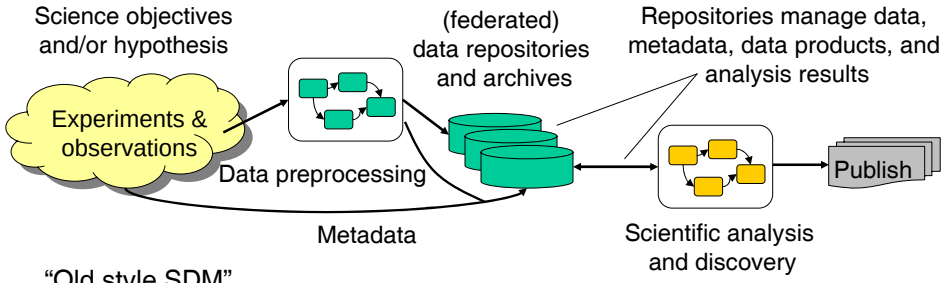
*Data archives*
*Grid computing*
*Data ownership/ stewardship*

*Scientific data formats*
*Metadata*
*Data provenance & lineage*
*Data processing pipelines*

*Scientific models*
*Data analysis*
*Data exploration*
*Data visualization*

# Scientific Data Management (3)

**The Scientific Data Management Process**




"Old style SDM"

1. formulate hypothesis
2. design experiment
3. run experiment
4. analyze result
5. evaluate hypothesis

- little/no data sharing/reuse
- single-PI centric

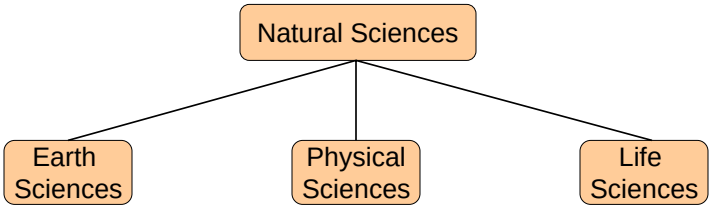
Trend

1. formulate hypothesis
2. **lookup and explore data**
3. evaluate hypothesis

- sharing/reuse of data products
- community-oriented



# Science Domains




Geology
- Mineralogy
- Paleontology
Soil Sciences
Geodesy & Geoph.
Oceanography & Hydrology
Atmospheric Sciences
- Climatology
- Meteorology
.....

**Observational Sciences**

Astronomy
- Astrophysics
- Cosmology
Chemistry
- (In)organic chemistry
- Analytical chemistry
Physics
- Classical mechanics
- Thermodynamics
- Quantum mechanics
.....

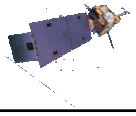
Botany
Zoology
Anthropology
Molecular biology
Cell biology
Biochemistry
Genetics
.....

# Earth Sciences

Space Act of 1958 established NASA, with an objective being "*the expansion of human knowledge of the Earth and of phenomena in the atmosphere and space...*"

- Studies all aspects of land and atmosphere using orbiting, aircraft-based, and in-situ sensors
- United States Geological Survey (USGS) and National Oceanic and Atmospheric Administration (NOAA) have somewhat overlapping mission with NASA.
- NOAA was given the responsibility of being the primary archive for Earth Observation System (EOS) data.

NOAA and NASA operate numerous satellites with Earth Science focus areas being *Weather, Earth Surface and Interior, Climate Variability and Change, Carbon Cycle and Ecosystems, Atmospheric Composition, Water and Energy Cycle*, and *Sun-Earth Connection*.





# Earth Sciences – The Science Drivers

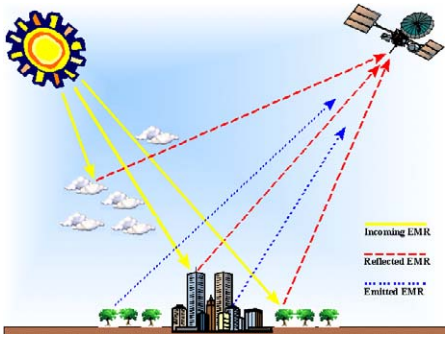
| Variability | Forcing | Response | Consequence | Prediction |
|---|---|---|---|---|
| Precipitation, evaporation & cycling of water changing? | Atmospheric constituents & solar radiation on climate? | Clouds & surface hydrological processes on climate? | Weather variation related to climate variation? | Weather forecasting improvement? |
| Global ocean circulation varying? | Changes in land cover & land use? | Ecosystems, land cover & biogeochemical cycles? | Consequences of land cover & land use change? | Improve prediction of climate variability & change? |
| Global ecosystems changing? | Motions of the Earth & Earth's interior? | Changes in global ocean circulation? | Coastal region impacts? | Ozone, climate & air quality impacts of atmospheric composition? |
| Atmospheric composition changing? | | Sea level affected by Earth system change? | Regional air quality impacts? | Carbon cycle & ecosystem change? |
| Ice cover mass changing? | | | | Predict & mitigate natural hazards from Earth surface change? |

Climate Variability and Change
Carbon Cycle and Ecosystems
Water and Energy Cycle
Atmospheric Composition
Weather
Earth Surface and Interior

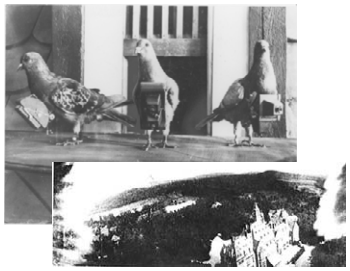
Src: "Drivers and Challenges for NASA's Earth Science Data Holdings", Presentation by K. Fountaine, F. Lindsay, and M. Maiden, PV 2005

## Remote Sensing

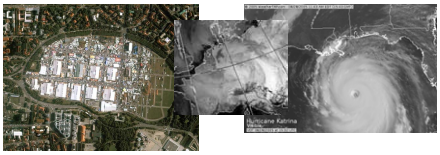
**Remote Sensing** is the science of measurement from a distance




Remote Sensing techniques make use of emitted or reflected electromagnetic radiation (EMR).



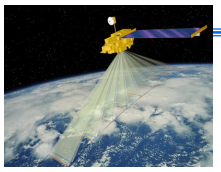
Key technology for a variety of Earth observations, such as meteorology, environmental monitoring, map making, and reconnaissance.



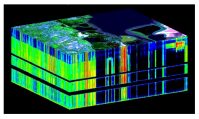


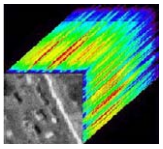
## Remote Sensing (2)

How does data management for remotely-sensed data typically look like?
→ processing of streaming satellite data in a file-based fashion...



All this is typically done using a file-based approach in which individual scenes described by multi- or hyper-spectral data are processed in separate steps.





time-tagged & geo-referenced

Converted to Bio-Geophysical Variables

Sensor Data Records (SDRs)

Homogenization and Calibration

Environmental Data Records (EDRs)

Fundamental Climate Data Records (FCDRs)

Converted to Bio-Geophysical Variables

Climate data Records or Homogenized Time Series

Thematic Climate Data Records (TCDRs)

## Remote Sensing (3)

***Data processing levels***, as used in the NASA Earth Observing System Program

- *Level 0*: reconstructed, unprocessed instrument data; all communication artifacts are removed (e.g., duplicate data)
- *Level 1A*: unprocessed instrument data at full resolution and time-referenced, annotated with radiometric and geometric calibration coefficients, and geo-referencing parameters, but not applied to level 0 data
- *Level 1B*: Level 1A data that have been processed to sensor units by applying radiometric corrections
- *Level 2*: Derived geophysical variables at the same level and resolution as Level 1 data; geophysical variables include sea surface temperature, soil moisture etc.

---

- *Level 3*: Derived geophysical variables mapped to uniform space-time grid
- *Level 4*: Model output or results from analyses of lower level data, e.g., variables derived from multiple measurements.



## Remote Sensing (4)

- **Major issues in SDM for remote sensing:**
  - Huge number of remote-sensing equipment out there
  - New instruments deliver even more data at steadily increasing spatial, temporal, and spectral resolution.
  - According to unofficial sources: only 4-8% of the data is used !
  - Users obtain data (levels 0, 1A, 1B) and data products (level 2) from the various NASA, NOAA, and USGS data centers.
  - Centers provide excellent services, but users have to compute and derive specialized data products on their own.
  - How to share such specialized data products with other users?

- **Ideally, users should be able to formulate queries like**

  *"Give me data product P for region R at resolution Z in real-time."*

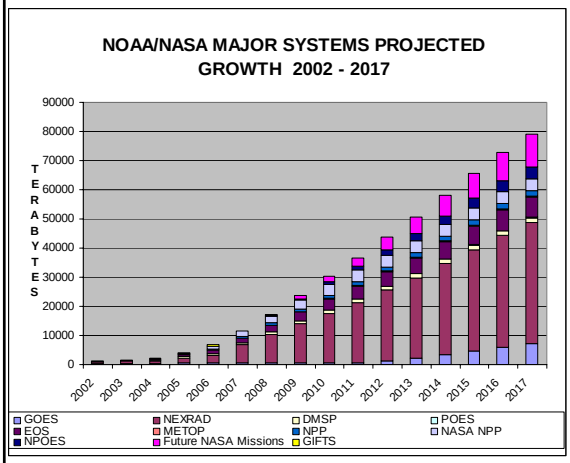
  → Requires a ***real-time stream*** of a data product (→ Module 3)

  → Requires ***ad-hoc data integration*** and ***fusion*** (e.g., whatever satellite has the most recent data for region R)

  → Requires sophisticated ***query processing capabilities*** at data centers

# Remote Sensing (5)

**Some notes on the amount of remote-sensing data**




Src: John Bates, Chief, NOAA Nat. Climate Center

Some example stream/data rates:
- Landsat ETM: 300Mb/sec – 3TB/day
- GOES 2.1 Mb/sec – 22GB/day
- Terra 150Mb/sec – 194 GB/day

In 2004, NOAA recorded more data in one year than in all years up to 1998.

NOAA's cumulative digital archive grew to 130 TB from 1978-1990
- grew another 130 TB from 1990-1995
- grew another 130 TB in 1996 alone
- approximately 900 TB in 2005.

NOAA maintains ~1300 "databases" containing ~2500 environmental variables.

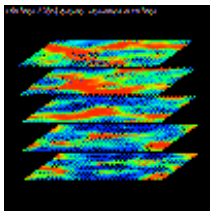
**… moreover, there are many military satellites ...**



# Cosmology/Astrophysics

## Astronomy data

- It has no commercial value whatsoever, no privacy or security concerns; can be freely shared
- Obtained from many different instruments (telescopes and satellites), resulting in different sky surveys and archives.
- High-dimensional data (temporal and spatial) that is well document and real; basically 2D images at different wavelengths



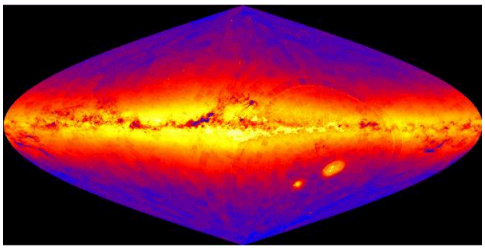
- Catalogs as part of the surveys are based on image processing and extracting object parameters.
- Spectra yield more object properties, such as clues to physical state, formation history, and 3D maps.
- In general, people are looking for new kinds of objects and interesting objects (e.g., quasars, supernovae, dwarfs, asteroids)
- Data mining is key! Objects can have 400 and more attributes!

## Cosmology/Astrophysics (2)

### How much data are we talking about?

The size of the archived data for an all sky survey (40,000 sq deg) is about two trillion pixels

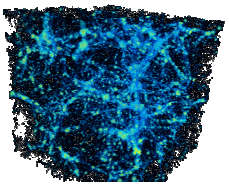
- One band: ~ 4TB
- Multiple wavelengths: 10-100TB
- Incl. time dimension: 10PB



All-sky distribution of 526,280,881 stars from the MACHO survey

The Large-aperture Synoptic Survey Telescope (LSST) is expected to collect about 7-10TB/night and 10-15PB/year. Too much data to easily move around.

⇒ Astrophysics with terabyte data sets and extensive data mining requirements

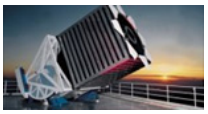


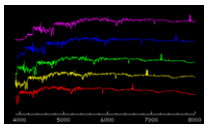


## Cosmology/Astrophysics (3)

Sloan Digital Sky Survey (www.sdss.org) aka "Cosmic Genome Project"

- Goal: create a detailed colored map of the Northern Sky.
- Scientific objectives: What is the origin of fluctuations? What is the topology of distribution? How much dark matter is there? How did galaxies form? What are the highest quasar red-shifts?



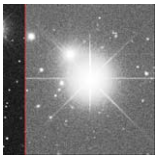


- Over the course of five years, SDSS-I imaged more than 8,000 square degrees of the sky in five bands, detecting nearly 200 million celestial objects, and it measured spectra of more than 675,000 galaxies, 90,000 quasars, and 185,000 stars.
- Catalog data and derived objects are managed in a database:
  - Full Data Collection ~20 TB, object catalog 400 GB (*parameters of >$10^8$ objects),* redshift catalog 1 GB (*parameters of $10^6$ objects),* Atlas Images 1.5 TB (*5 color cutouts of >$10^8$ objects),* spectra 60 GB (*in a one-dimensional form)*

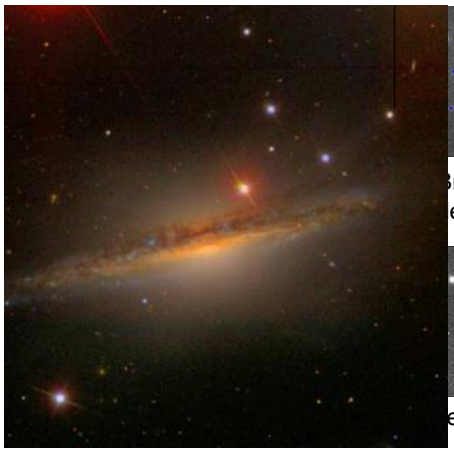
# Cosmology/Astrophysics (4)

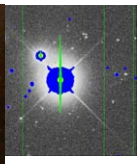
Data processing steps

Src: SDSS Web site

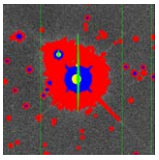


Raw data frame

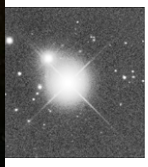




right object
etection



Faint object
detection



econstructed image

SDSS data can be queried through SkyServer (skyserver.sdss.org), which also provides data visualization and analysis tools (SQL queries!)

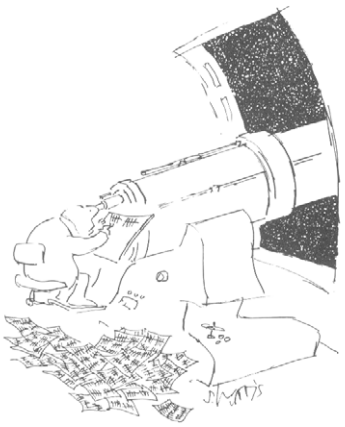


# Cosmology/Astrophysics (5)

## SDM Issues

- Many sky surveys and sky catalogs (MACHO, DLS, …)
- Data growth is dramatic (Petabyte sized data sets are near)
- Data analysis:
  - Hypothesis/model formulation
  - Data lookup and model verification based on the data
  - Heavily relies on data mining and statistical analysis
  - Requires highly scalable analysis algorithms and data management infrastructures (e.g., spatio-temporal indexing)
  - Take analysis to the data
- Efforts toward developing ***virtual observatories***: link geographically distributed astronomical data archives and information resources
  → Middleware standards for data, metadata, resource descriptions, queries, query results, and semantics.
  → Support heterogeneity, interoperability, and federation.

## Cosmology/Astrophysics (6)




Build systems that help scientists
verifying hypothesis



## Summary so far …

- Scientific data collections are increasing in size, heterogeneity, complexity and resolution.
- Scientific data repositories are specialized.
- Scientific data typically do not come in the form of relations or XML documents.
- It is not sufficient to simply make the data Web accessible.
  → Network bandwidth is a limiting factor.
- In the ideal case, have ***virtual repositories*** with sophisticated computation/query front-ends and that make data and data products accessible to anyone, anywhere.
- Metadata play an very important role.
- Don't forget the many other types of "data collectors", such as the huge variety of sensor networks...