GEORGE MASON UNIVERSITY | Computational and Data Sciences

CSI 695: Scientific Databases

Fall Term 2017

Lecture 2: Introduction to Scientific Databases

Lectures: Prof. Dr. Matthias Renz

Why Scientific Databases

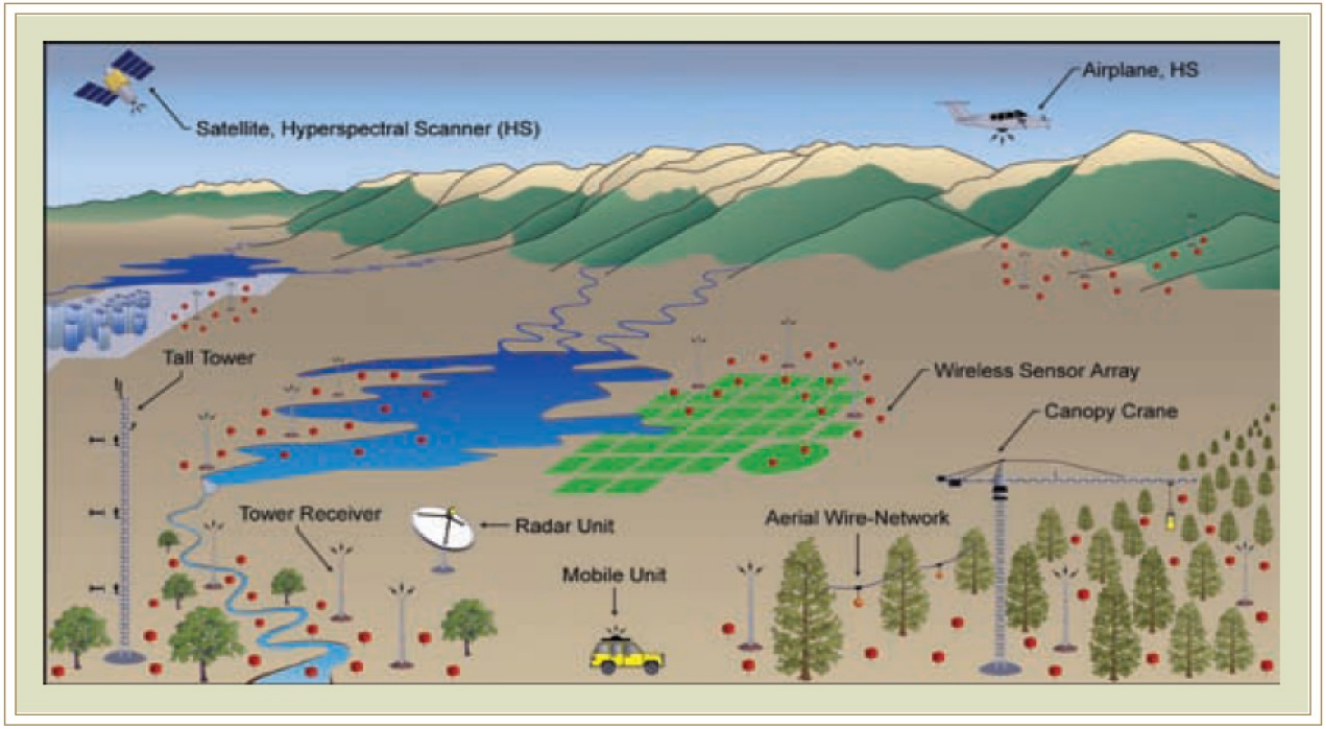
- Why Scientific Data?
  - Irreversible trend in (computational) science → learning from data
  - Applicability of Scientific Data is growing enormously
  - Scientific data is rich, it contains a lot of (useful) information in high quality
  - Scientific data is (can be) complex and allows us to model complex instances of and relationships between entities of our real (scientific) life.

- Why Database Technologies for Scientific Data
  - Availability of scientific data is huge and growing very fast
  - Scientific data is very space intensive
  - Many scientific applications require concurrent multi-user access to the data
  - Scientific data have to be searched in an efficient and effective way

Why Scientific Databases

- What can we get from standard databasesystems?
  - Consistency preserving multi-user access to data
    - Note: In the context of Big Data, modern database technologies relax on consistency to improve accessibility
  - Physical and logical data independency
  - Efficient access through (built-in) index methods (a.k.a. access methods)
  - Support of transactions:
    - Concurrency: Isolation of concurrent updates of multiple users (ACID)
    - Recovery: Consistent recovery from failures
    - Monitoring/control of data integrity
  - Data integrity and privacy preserving

Examples of Scientific Databases and Data Sources

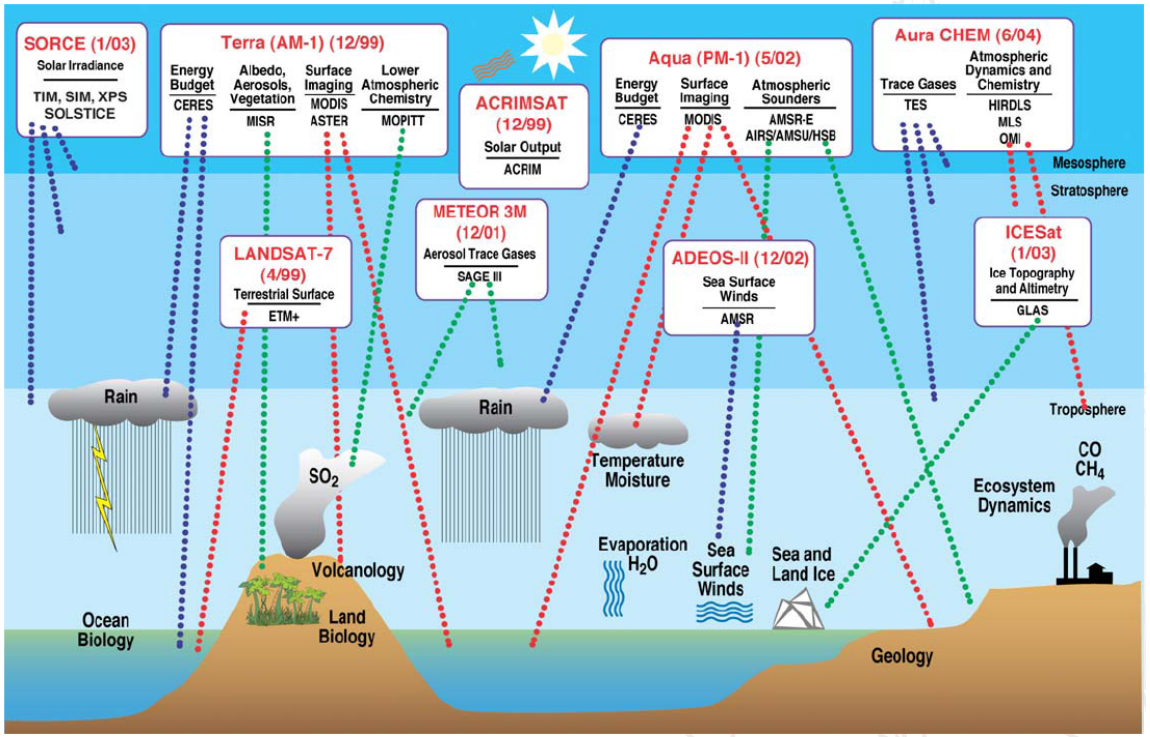


- From the NSF (2007) report "Cyberinfrastructure Vision for 21st Century Discovery"
- NEON: National Ecological Observatory Network – http://www.neoninc.org/

[Material: Dr. Kirk D. Borne, CDS @ GMU]

Examples of Scientific Databases and Data Sources

- EOS
  Earth Observing System
  http://eos.nasa.gov

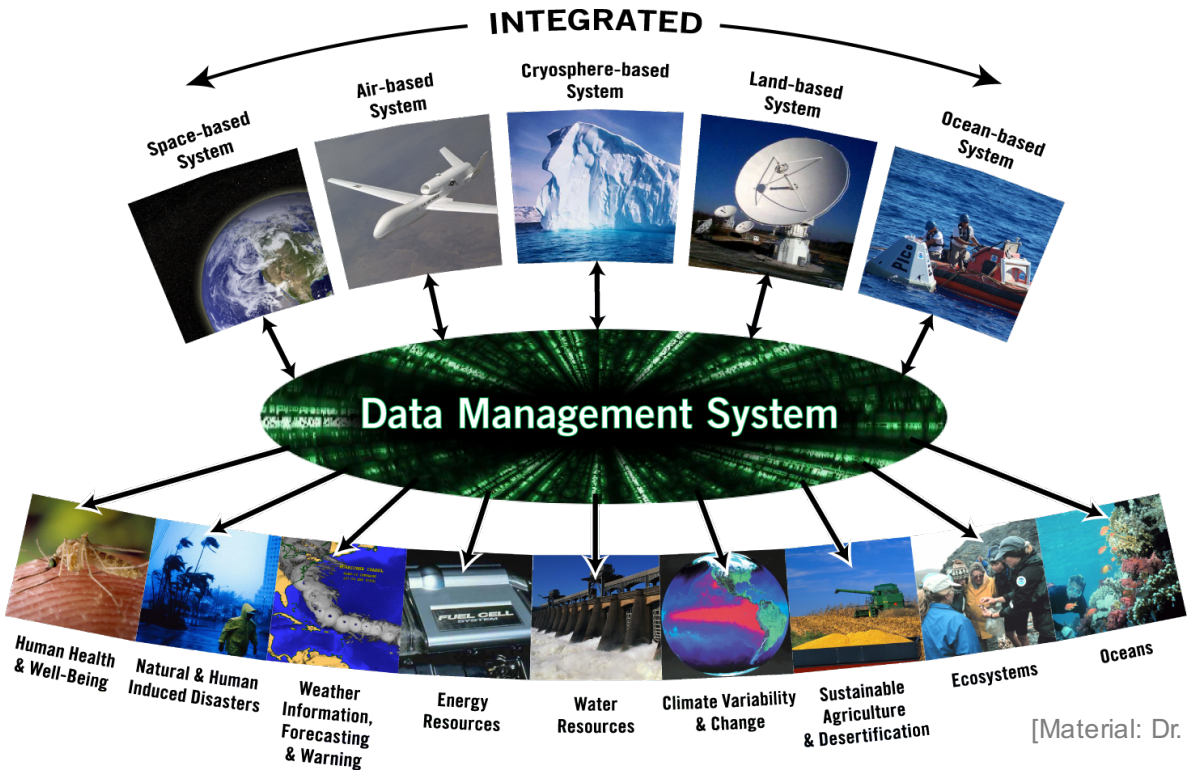


- 1-10 TB/day from large suite of remote sensing and satellite instruments.
- By 2017, total data collected will be >140 Petabytes.
- Provides access to more than 3,000 types of Earth science data products and specialized services for interdisciplinary studies.

[Material: Dr. Kirk D. Borne, CDS @ GMU]

Examples of Scientific Databases and Data Sources

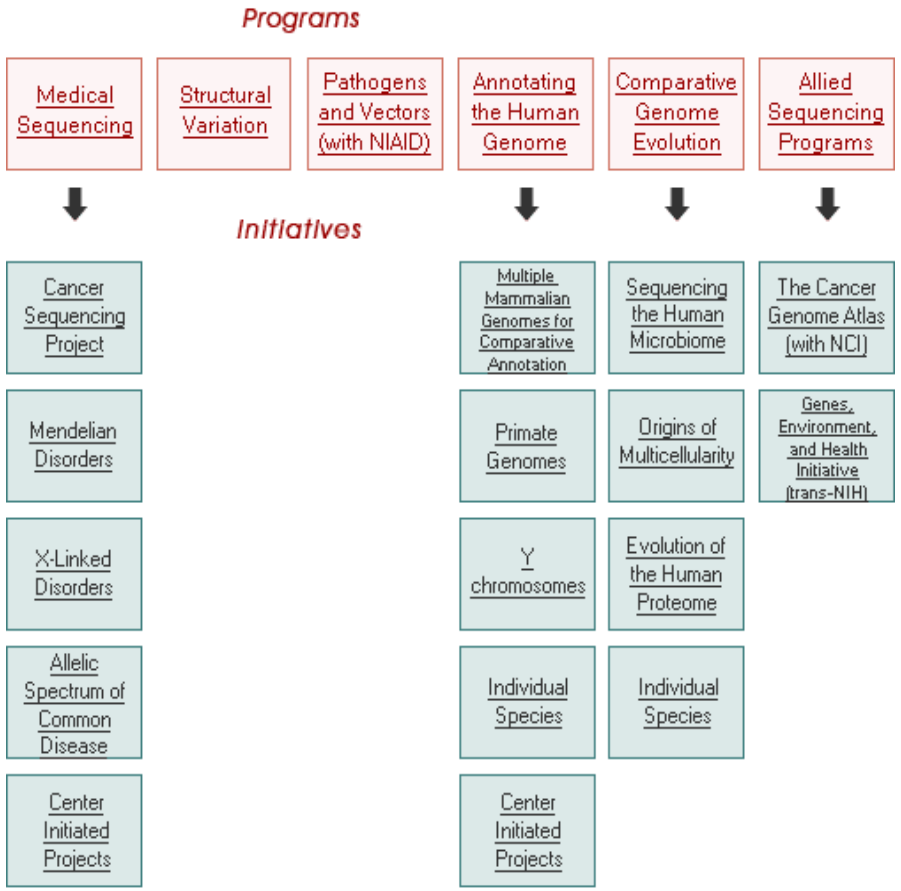
- GEOSS: Global Earth Observation System of Systems
- http://www.epa.gov/geoss/ or http://earthobservations.org/



[Material: Dr. Kirk D. Borne, CDS @ GMU]

Examples of Scientific Databases and Data Sources

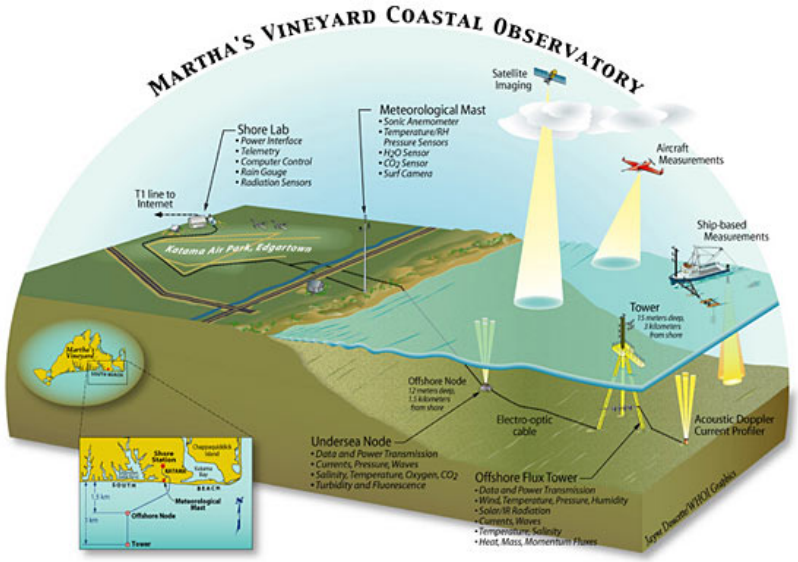
- Human Genome Project
  http://www.genome.gov/ and
  http://www.1000genomes.org/
  - The Human Genome Project undertakes large-scale sequencing projects to provide critical genomic information with significant value in areas of very broad scientific interest.
    - Over 150 billion raw Q20 base pairs (accuracy = 99%) are sequenced each year.
  - The 1000 Genomes Project was announced in January 2008:
    - 8.2 billion bases sequenced per day
    - Over 1000 complete human genomes to be mapped
    - 6 trillion DNA bases to be mapped – 60 times more data than compiled within all DNA databases over the past 25 yrs

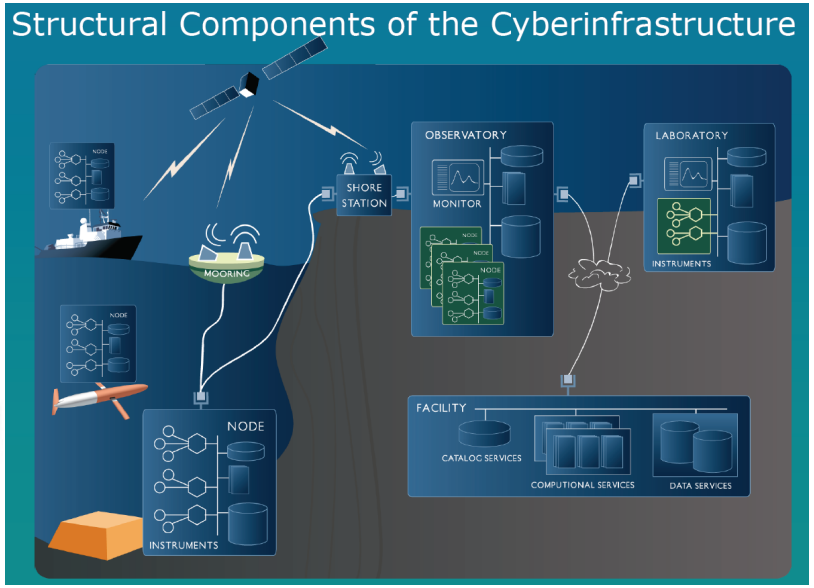


[Material: Dr. Kirk D. Borne, CDS @ GMU]

Examples of Scientific Databases and Data Sources

- Ocean Observatories Initiative (OOI) -- http://www.oceanleadership.org/
  (now called The Consortium for Ocean Leadership)





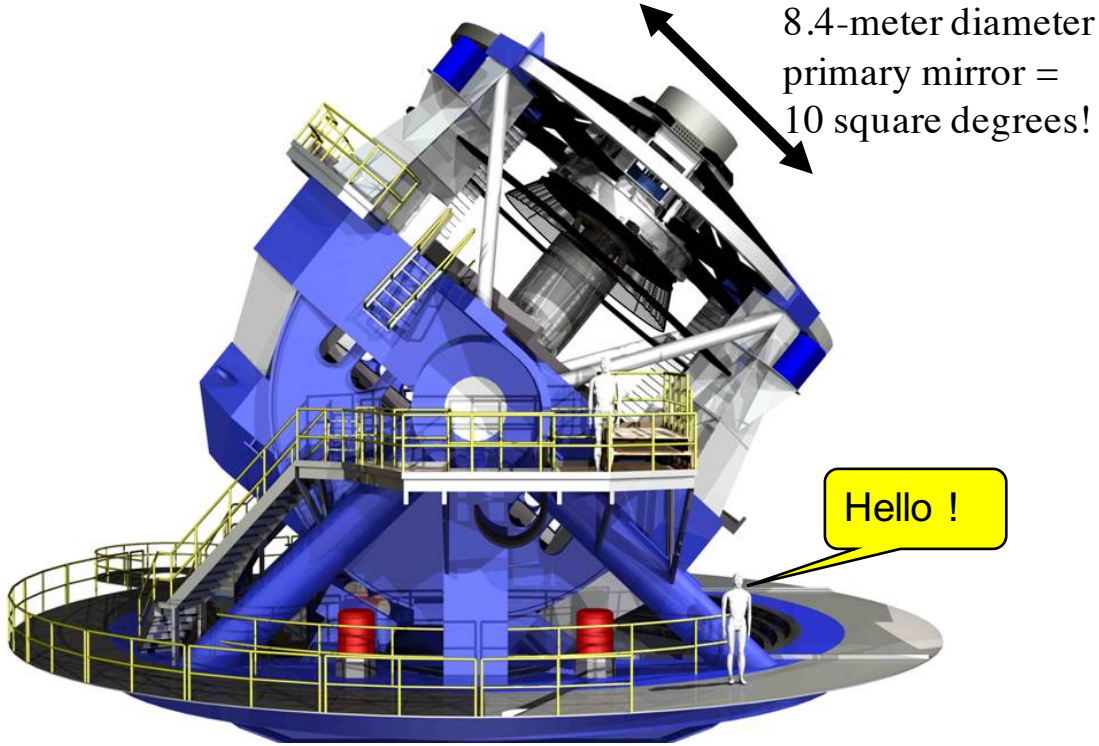
- OOI will construct a networked infrastructure of science-driven sensor systems to measure the physical, chemical, geological and biological variables in the ocean and seafloor.
- Will collect ~1 Petabyte of new experimental data per year.

[Material: Dr. Kirk D. Borne, CDS @ GMU]

Examples of Scientific Databases and Data Sources

- LSST = Large Synoptic Survey Telescope -- http://www.lsst.org/

- "**Data Management:**
  Software is one of the **most challenging** aspects of the LSST, as more than 30 terabytes of data must be processed and stored each night in producing the largest non-proprietary data set in the world."

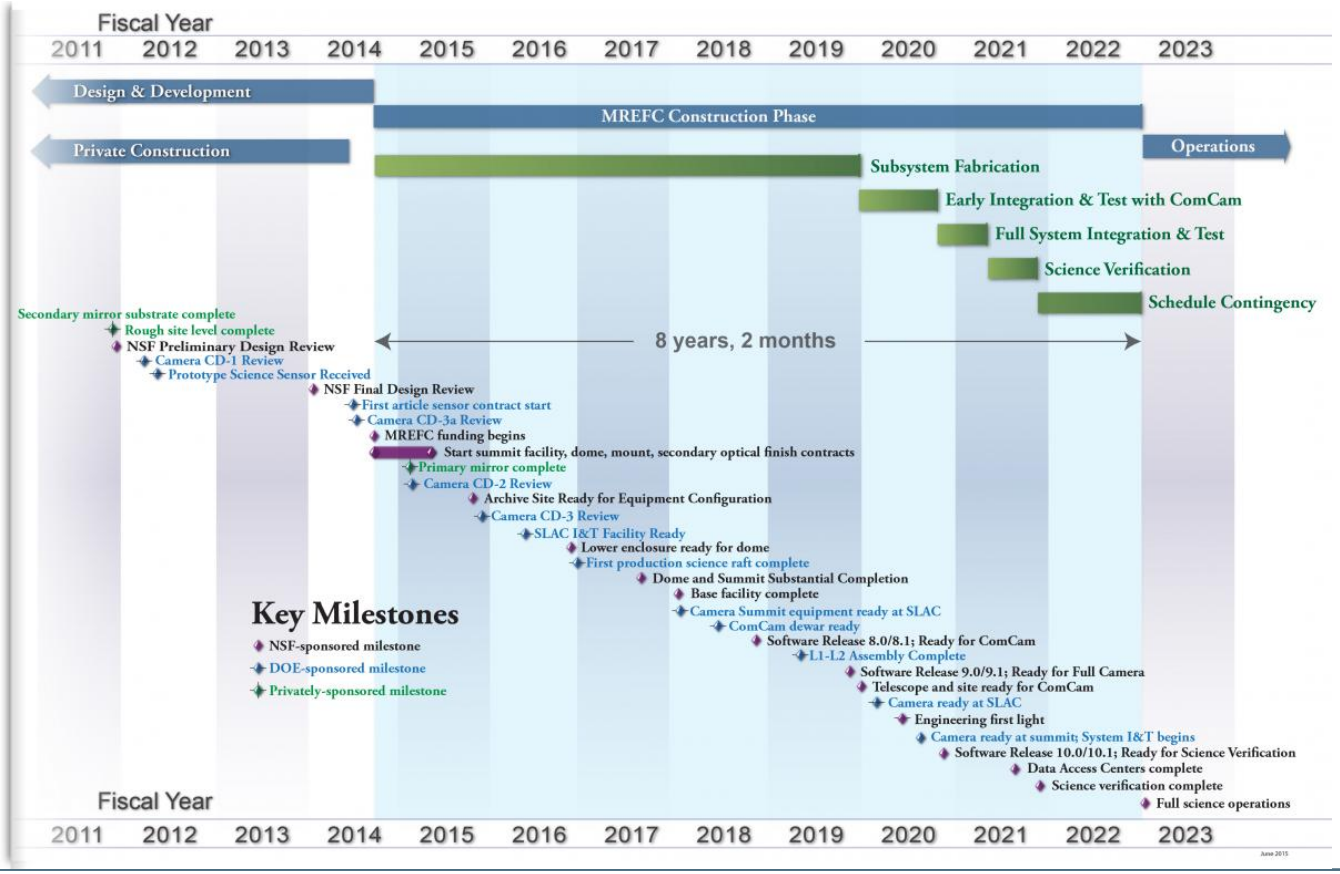


- 100-200 Petabyte image archive
- 20-40 Petabyte database catalog

[Material: Dr. Kirk D. Borne, CDS @ GMU]

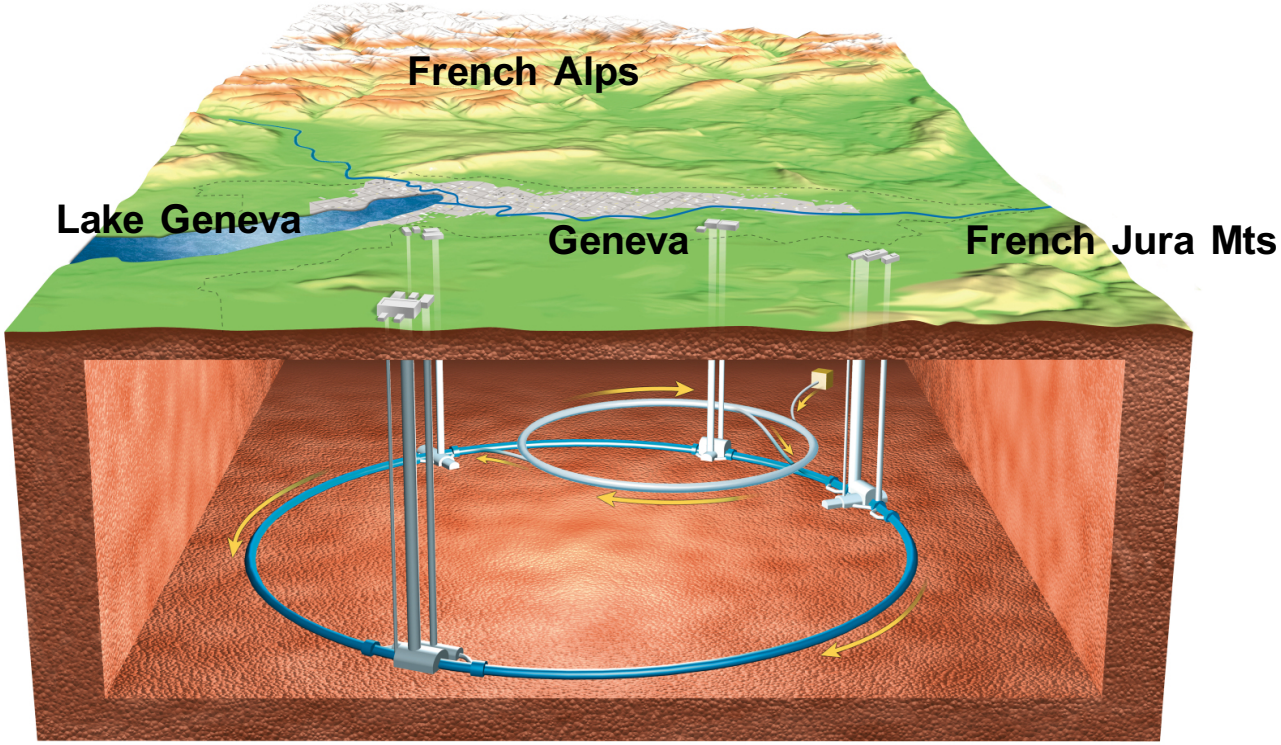
Examples of Scientific Databases and Data Sources

- LSST = Large Synoptic Survey Telescope -- http://www.lsst.org/
- Project Schedule:



Examples of Scientific Databases and Data Sources

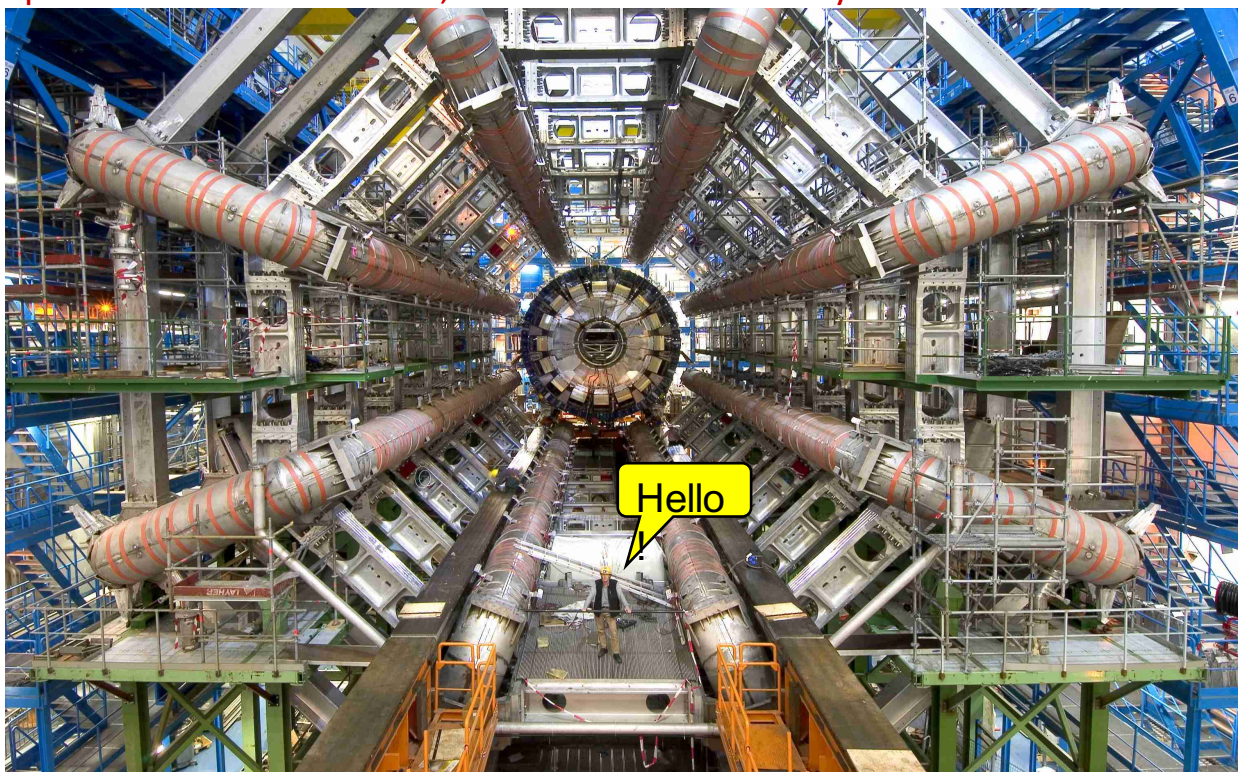
- **Large Hadron Collider (LHC):** 27km diameter particle accelerator ring, 100m underground, below the border of Switzerland and France, started experiments in 2010 – 100,000 DVDs of data each year
  < http://www.uslhc.us/ >
- High-energy particle collisions allow scientists to study the properties of particles such as the Higgs boson and search for new fundamental laws and phenomena.



[Material: Dr. Kirk D. Borne, CDS @ GMU]

Examples of Scientific Databases and Data Sources

- **Large Hadron Collider (LHC):** 27km diameter particle accelerator ring, 100m underground, below the border of Switzerland and France, started experiments in 2010 – 100,000 DVDs of data each year
  < http://www.uslhc.us/ >
- The Large Hadron Collider (LHC) Atlas Experiment – over 1 Petabyte/sec data generated in this huge detector (just one of several LHC experiments) < http://atlas.ch/ >

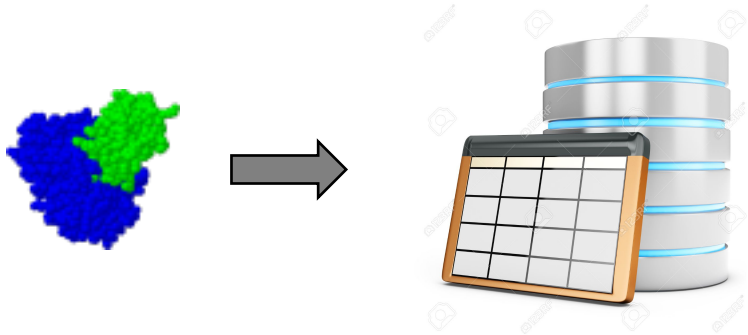


[Material: Dr. Kirk D. Borne, CDS @ GMU]

Scientific vs Standard Databases

- Managing scientific data usually requires methods that go beyond the capabilities of standard database management systems.
- Scientific data often involves complex structured data not well supported by the table schema used in standard databases.
- Exact match queries as provided in standard dabase management systems often do not suffice for searching in scientific data, we need something different!!!

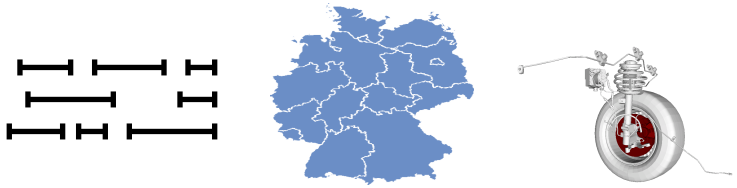
Example:
How would you store/manage molecules in a standard relational database ?



Think about it ...

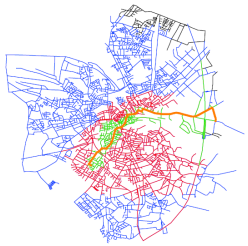
Scientific vs Standard Databases

- Scientific data often consists of complex stuctured data including spatial, temporal, spatio-temporal, and multi-media data?
  - Spatial data: 1D, 2D, 3D



  - Temporal data: time series

  - Spatio-temporal data:
    - objects moving in a given space



  - Multi-media data:
    - audio sequences, video sequences

Scientific vs Standard Databases

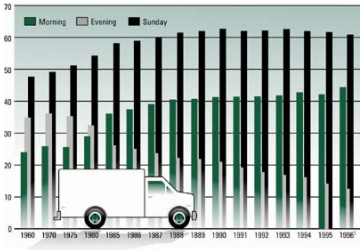
- A motivating example
  - Given an archive with 2,000,000 images (2D objects)
  - Is a given image included in that archive?

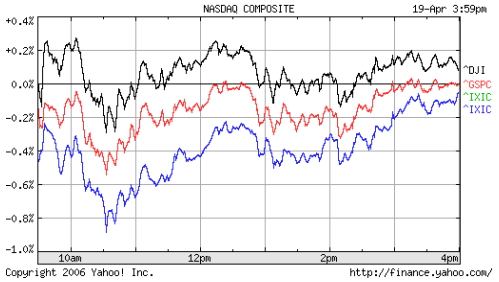
  - Challenges
    - "included" does not necessarily mean "identical binary representation!

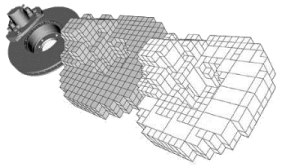
- Images may vary in
  - Size (scaling, resolution)
  - Perspective (reflection, …)
  - Coloring, shading
  - Clipping, cutting
  - Add-ons (border, annotation, …)

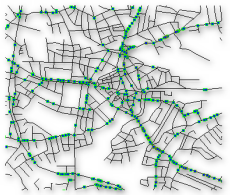
Scientific vs Standard Databases

- A motivating example
  - Similar problems with other complex objects
    - Temporal data
      - Different length of sequences
      - Different scaling or shifting
      - …
    - Spatial data
      - Invariance against special transformations
      - Different resolution
      - Query object may not be in the database
      - Query object may only be approximated (sketch)
      - …
    - Spatio-temporal data
      - Objects are moving
      - Location of objects may be uncertain
      - …









Scientific vs Standard Databases

- A motivating example
  - Similar problems with other complex objects (cont.)
    - Multi-media data (audio/video data)
      - Content-based image comparison vs. direct similarity
      - Incomplete sequences
      - Sequences with different lengths
      - Different audio/video formats
      - Different resolution (frames per second, image resolution, …)
      - Query object may only be approximated (query by humming)
      - …
  - ⇒ Instead of searching for

    exact matches (supported by standard databases (relational-, object-relational DBMS)

    we need to search for

    "similar" objects! → **new concepts for managing data required (non-standard databases)**

Scientific vs Standard Databases

- Searching:
  - Searching in traditional databases:
    - A query is specified by a number of attribute values the result objects should directly match (direct match queries)
    - Example: Archeozoological Database
      SELECT *
      FROM bones b
      WHERE b.width = 15 and b.height = 40 and b.weight = 130
    - Result:
      - Only bone b2
      - Bone b3 which is very similar to b2 remains hidden like b1
      - If b2.height would be 40.1, the result would be empty

BID	width	height	weight
b1	14	20	86
b2	15	40	130
b3	15	40.5	131

  - Searching in scientific databases:
    - A query is often specified by
      - an object (query object) provided by the user (e.g. URL, file, etc.)
      - simplified approximation of an object (draft, aggregates, etc.)
      - no concrete query object, but the user can navigate through the set of hierarchically organized objects, where similar objects are hierarchically grouped and reference objects are used as keys representing the groups.

Scientific vs Standard Databases

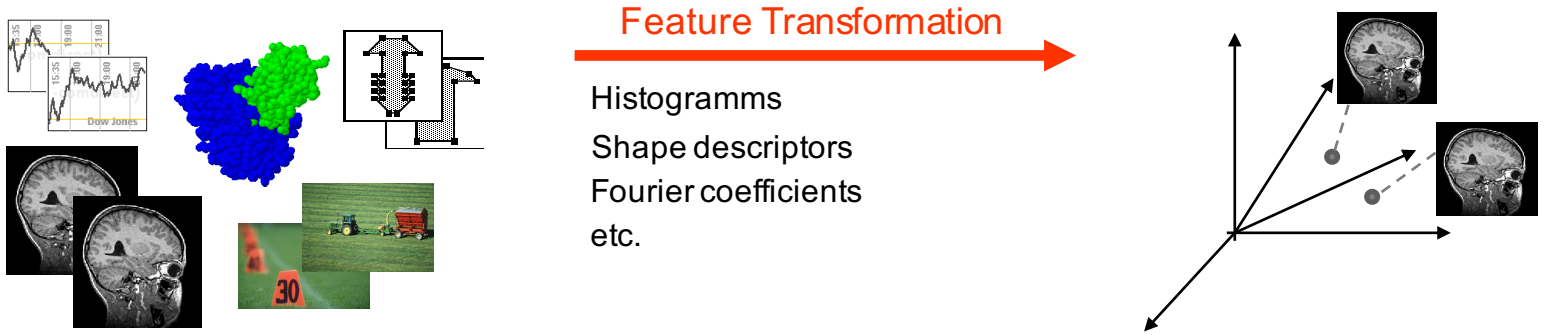
- Searching:
  - General problems when searching for similar objects
    - Informal level
      - Similarity depends on the application
        - Searching for images showing “sunset” => color is important
        - Searching for images showing “animals” => shape is important
      - Similarity depends on the user’s notion
    - Formal level
      - How are the objects represented?
      - How can the similarity between objects be modeled?
    - Pragmatic level
      - Efficient algorithm for computing the similarity
      - Efficient algorithm for searching in a large disc-resident database

Scientific vs Standard Databases

- Searching:
  - Here, we focus on the sub-problem
    - Efficiently searching for similar objects in a large database and to some extend on
    - Efficiently computing the similarity between objects
  - We assume a very common model of similarity: ***Feature-based similarity***
- Feature-based similarity
  - How can we model the similarity between complex objects like images, 3D objects, video sequences, etc.?
  - Considerations
    - Efficiency: Model should allow efficient query processing => use of index structures should be possible
    - Generality: Avoid the necessity to develop algorithms and index structures for each application separately but develop a general way to model similarity
      - We have indexes for
      - Spatial data and multi-dimensional vectors
      - General metric data (objects with a metric distance function)

Scientific vs Standard Databases

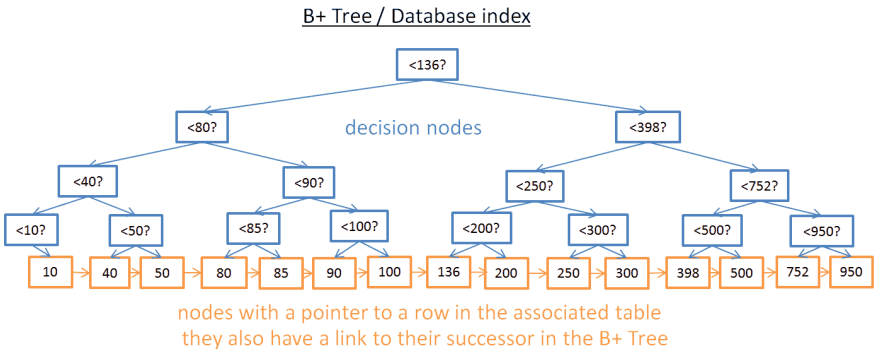
- Feature-based similarity (cont.)

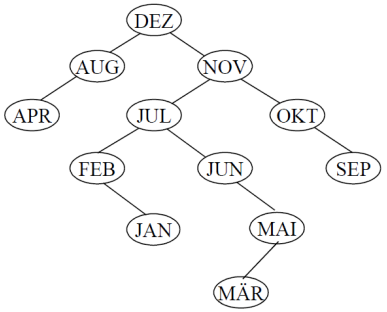


- Objects from real world are transformed into multi-dimensional feature vector points
  1. Identify a set (sequence) of (numerical) features from objects that best describe the objects
  2. Build a multidimensional point vector from the set (sequence) of features
  3. Manage the point vectors where each point vector has a link to the detailed object descriptions
  4. (Object) point vectors are efficiently organized (managed) using appropriate index structures

Scientific vs Standard Databases

- What is an index? (Abstract Definition)
  - An index in the context of a database is a data structure that improves the speed of data retrieval operations on a database at the cost of additional writes and storage space to maintain the index data structure.
  - Indexes are used to quickly locate data without having to search every entry in a database every time a database is accessed.
  - Examples:





Binary search tree