SciDB Testbed @ NERSC

- Analyze Terabytes of Data with Ease

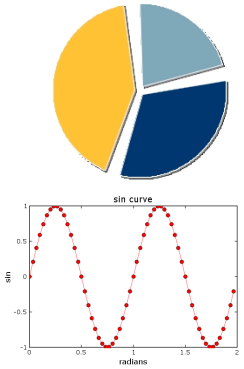
Yushu Yao, NUG 2013

Quickly Converge on the right analysis and turn large dataset to scientific insight

Terabytes of Data

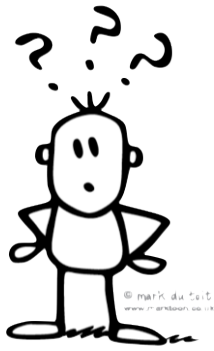
Analysis

Refined/New Analysis



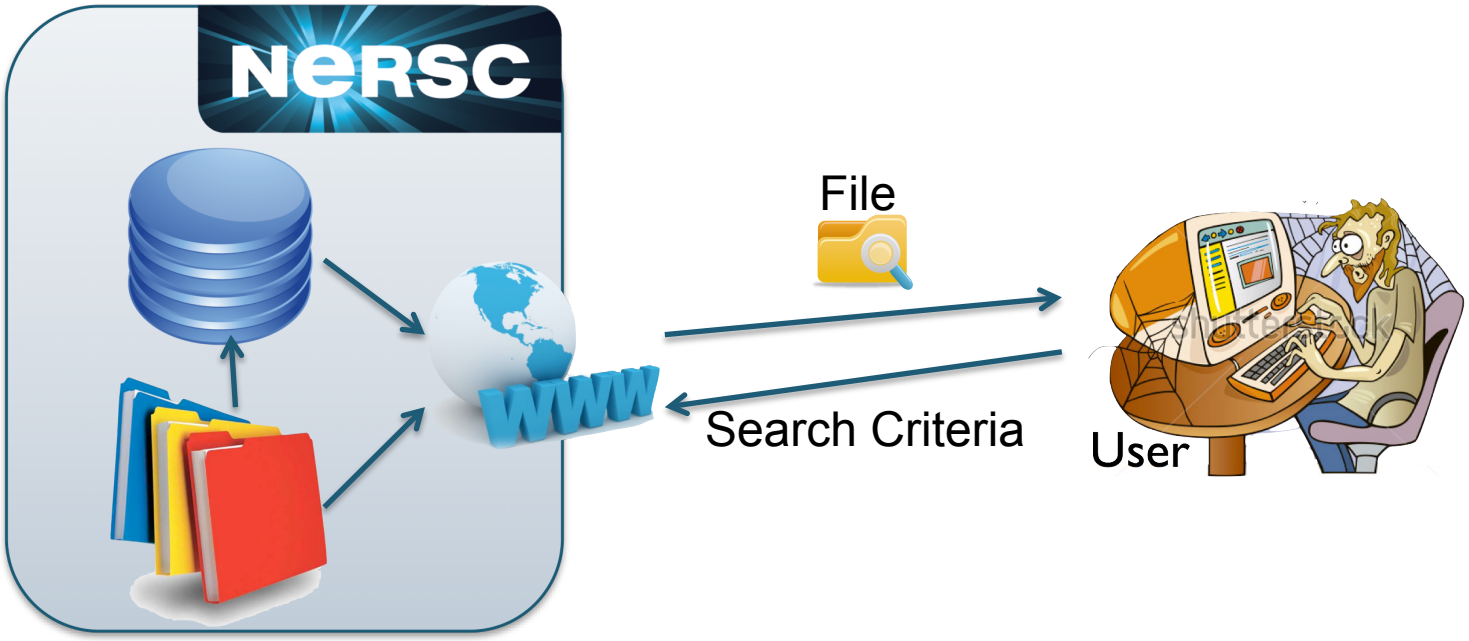
Scientific Insight

I have Terabytes of data, I need an easy, fast, interactive way to filter, aggregate, …

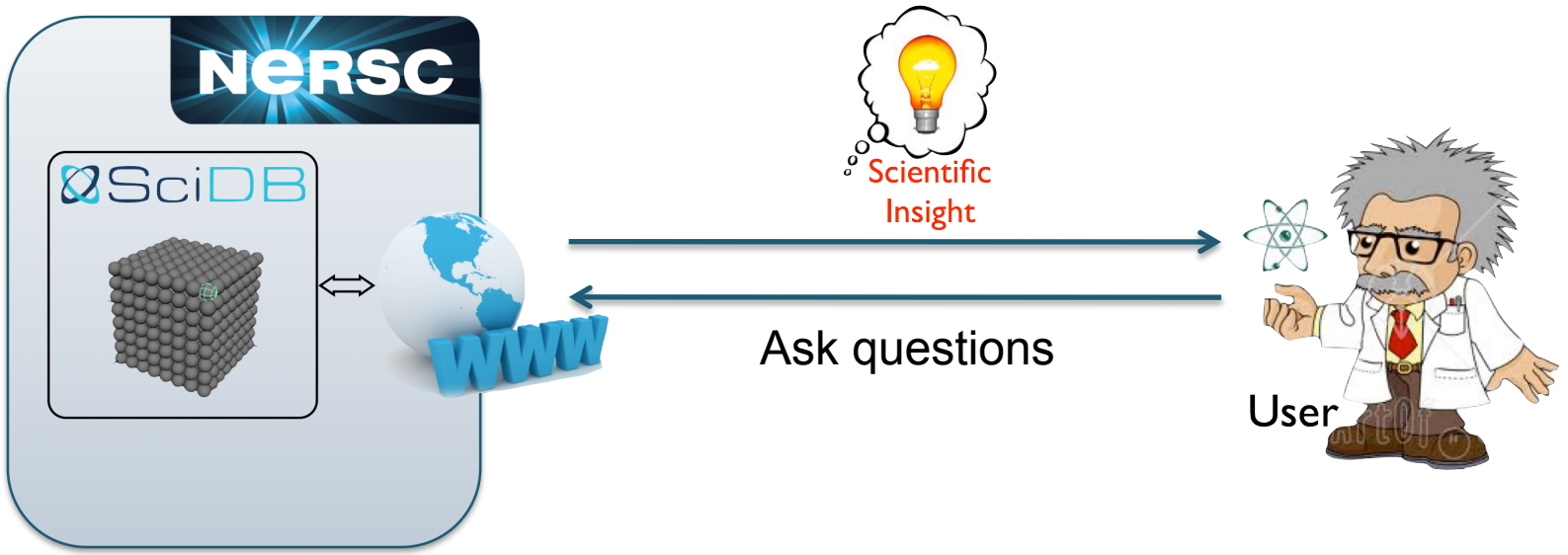


Data is too big. Old way of sharing data over web is no longer efficient (if at all possible)

- **Old way:**
  - Data files are cataloged, catalogs are stored into a database;
  - Web user search through catalog and locate the data file;
  - Data file is downloaded for local analysis



In the future, data will be shared by opening up customizable operations

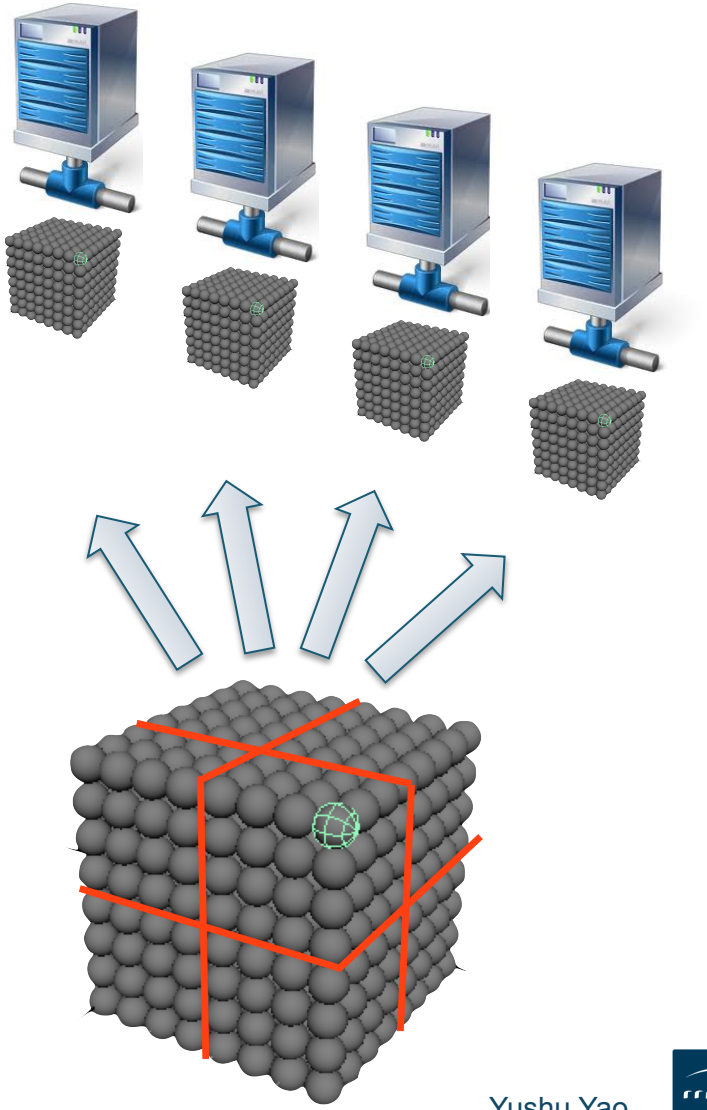


Recap, what are needed?

- **Interactive Analysis on TB of data that:**
  - No need to write complicated parallel IO code
  - No need to wait in the batch queue
  - Gives decent performance
- **Powerful Backend to a Web Gateway that:**
  - Gives near real time response to queries that scan through TB of data
  - Return an “answer”, not a file

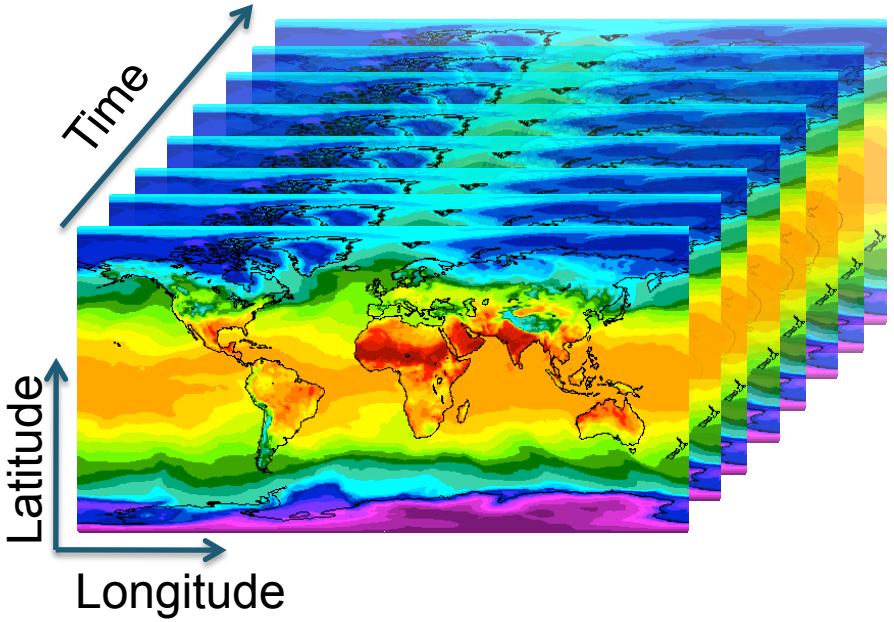
SciDB, parallel processing without parallel programming

- **Michael Stonebreaker**
- **Everything in Arrays**
  - Multi-dim indexing: locate an element at O(constant)
  - Best for machine/ simulation generated structure data
  - Good for metadata too
- **Query-like language, auto-parallelization**

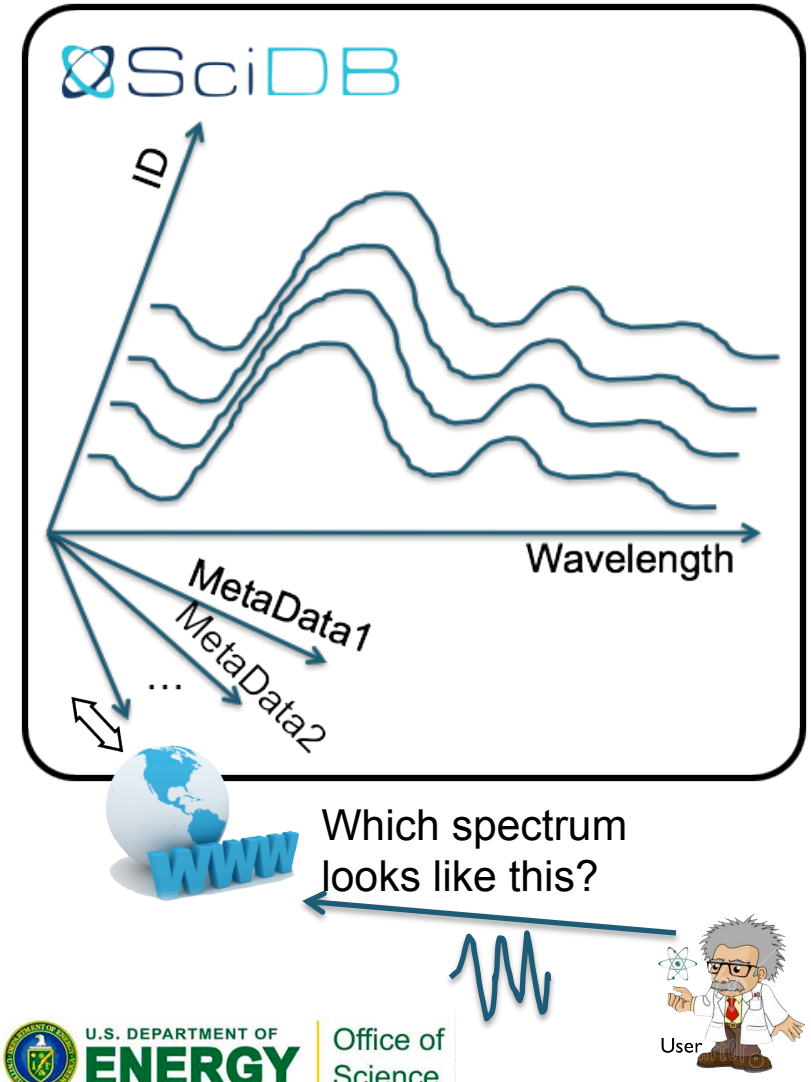


Quick Turnaround Interactive Analysis of Climate Simulation Results

- **Each simulation run can generate Terabytes of data**
- **Operations such as:**
  - Average Temperature
  - Sort/Percentile (e.g. get top 3% hottest days)
  - Temperature change in one area



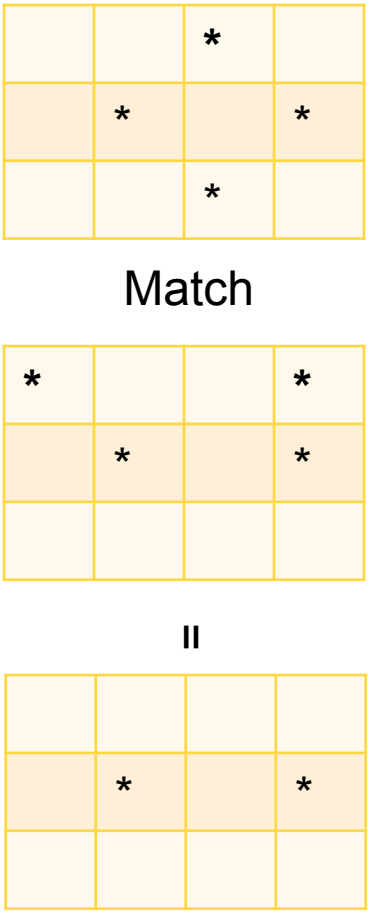
Case Study: Web-based Spectra Matching



- Millions of Spectra stored in SciDB
- Web users can search with metadata criteria, like any other gateway
- **NEW: web user can upload a custom spectrum, and SciDB will find the most "similar" spectrum in the database**
  - Very compute/IO intensive
  - Scan 1Million spectra and return the most "similar" one in 2 min.

Case Study: Cross Matching Catalogs of Stars in the Sky

- Given 2 sets of observed objects, return the objects observed in both sets (~300GB for 1billion stars)
- Spatial query not efficient in SQL
- In SciDB, you can lay out the stars in a 2D table, and overlay the them. In parallel.
- 50 times faster to match 1 billion stars in SciDB (5min) than PostgreSQL (5hr)



Existing Projects in Several Science Areas

- Climate
- Cosmology
- Astronomy
- Bioinformatics
- Bio-imaging

Recap, what can SciDB give us?

- **Interactive Analysis on TB of data that:**
  - No need to write complicated parallel IO code
  - No need to wait in the batch queue
  - Gives very fast response
- **Powerful Backend to a Web Gateway that:**
  - Gives near real time response to queries that scan through TB of data
  - Return an “answer”, not a file

You are welcome to try it out

- **What we need from you:**
  - Data (10+GB to start with, projected data > TB)
  - Key Operations
  - (Even better) a PostDoc/Student who doesn't want to write MPI parallel IO code.
- **What we provide:**
  - Hardware/Software, ready to use
  - Help port-in, and help to implement the key operations
- **Contact: consult@nersc.gov**

National Energy Research Scientific Computing Center