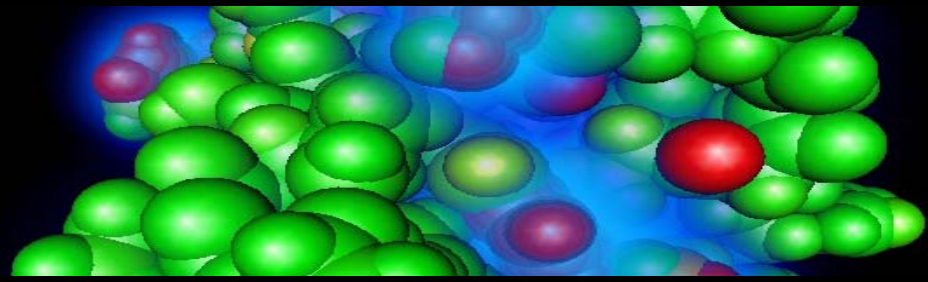


Query-Driven Visualization

April 18, 2005

E. Wes Bethel

with help from Friends at

Lawrence Berkeley National Laboratory

Office of Science
U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

Problem Statement

We live in an information dominant age.

Office of Science
U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

Problem Statement

- **Information management is a limiting factor in many sciences and endeavors:**
  - Time: You have 20 minutes between tokomak experiments to analyze results from previous run and set parameters for next one.
    - Did the magnetic field lines stabilize in the last run?
    - What happened in that other experiment?

Problem Statement

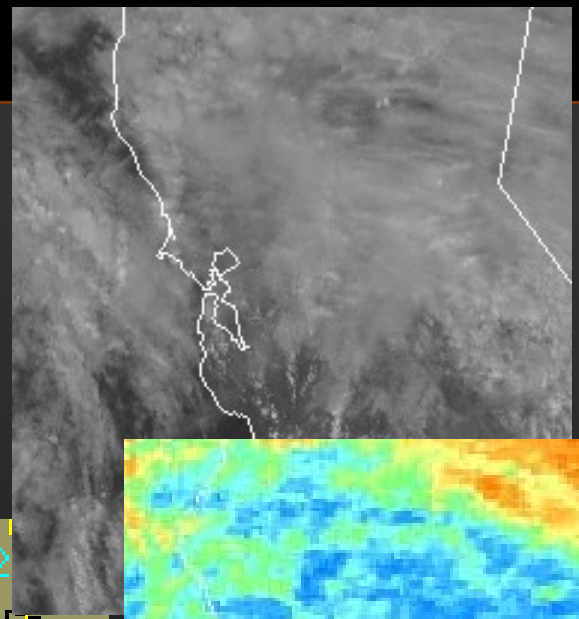
- **Simple questions give rise to startling complexity.**
  - Will a new malaria vaccine be effective?

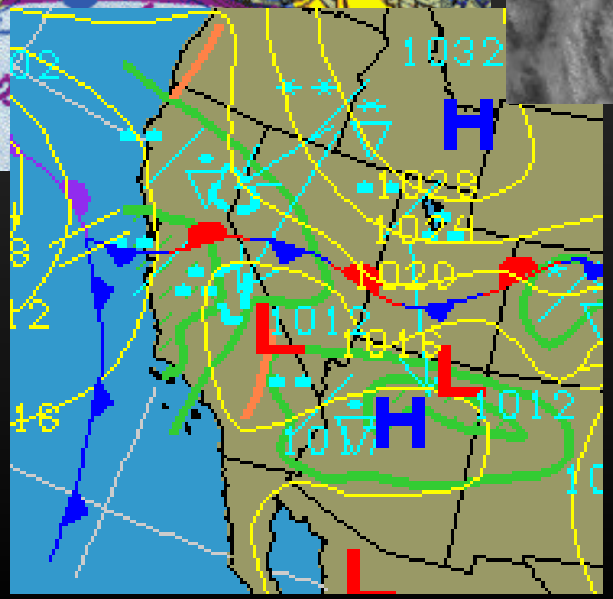
    Genome dbase, metabolic pathway dbases, prioritization, compare against human genes.
  - What is a flame front?
  - Should I fly today? (When should we launch the shuttle or schedule a landing?)

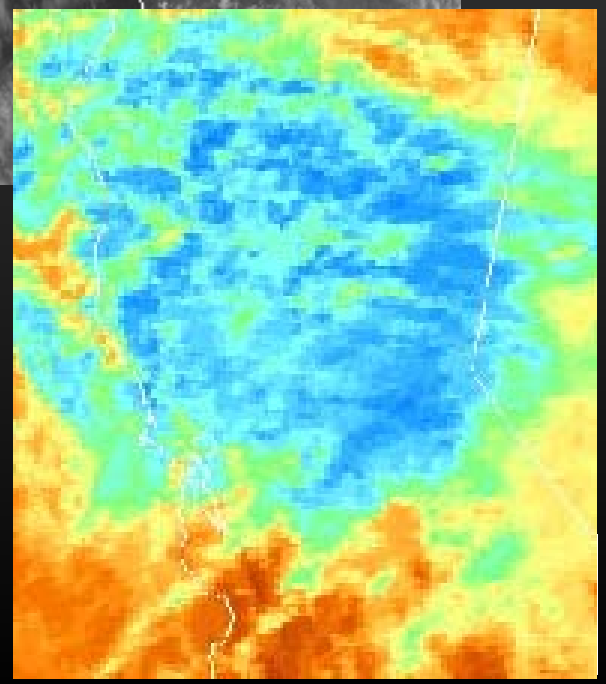
Office of Science
U.S. DEPARTMENT OF ENERGY

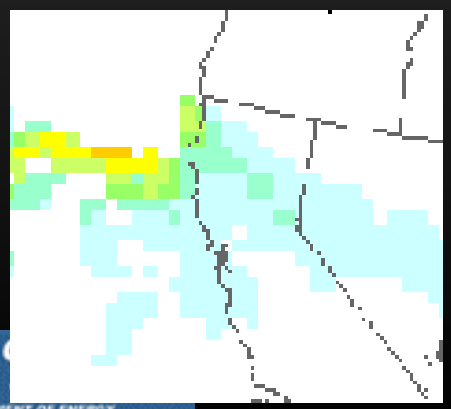
BERKELEY LAB

A Simple Question: Should I Fly Today?



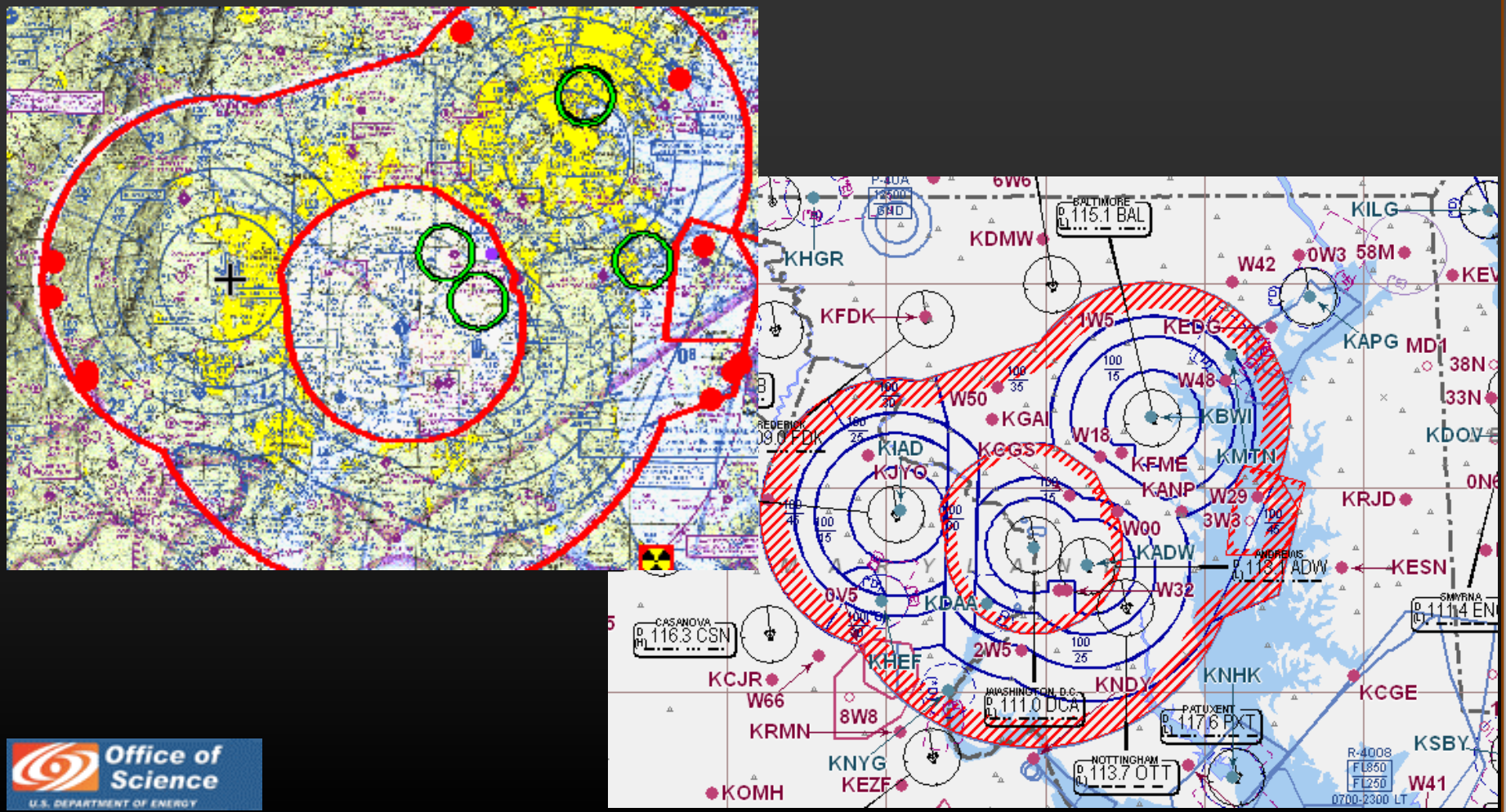






U.S. DEPARTMENT OF ENERGY

The Simple Question Becomes More Complex When Considering All Available Data



Office of Science
U.S. DEPARTMENT OF ENERGY

Dimensions of the Problem

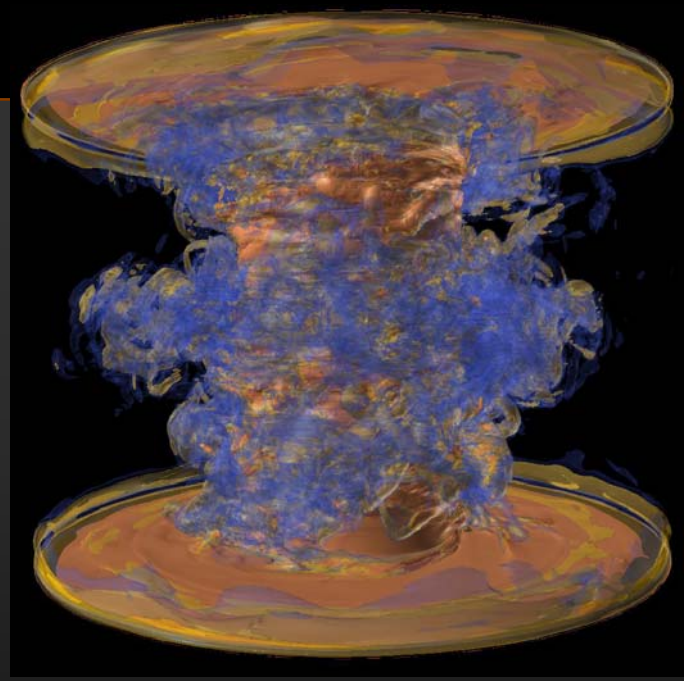
- **Data size and complexity.**
  - Where to store it? How to access it?
  - “I’m spending nearly all my time, finding, processing, organizing, and moving data—and it’s going to get much worse.”
- **N-body problem.**
  - Multiple research groups within one discipline.
  - Migration of data between disciplines.
- **Other problems: metadata management, workflows, federated data, distributed data, data analysis, …**

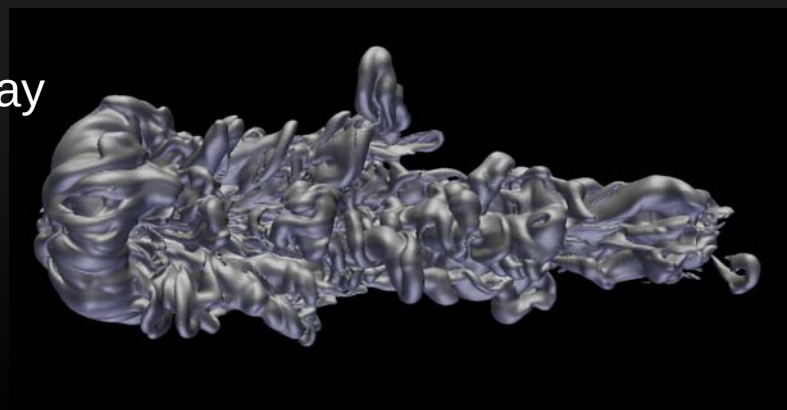
Office of Science
U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

One “Bigger Data” Solution: Use A Bigger Hammer

- **Scalable solutions for processing larger data using existing algorithms.**
  - Faster computers, scalable tools produce increased capacity – humans ought to be able to visually process the increased load.
- **Some known problems:**
  - Doesn’t really solve the “overwhelmed with data” problem.
  - Increasing the amount of visible data may result in *less* comprehension.

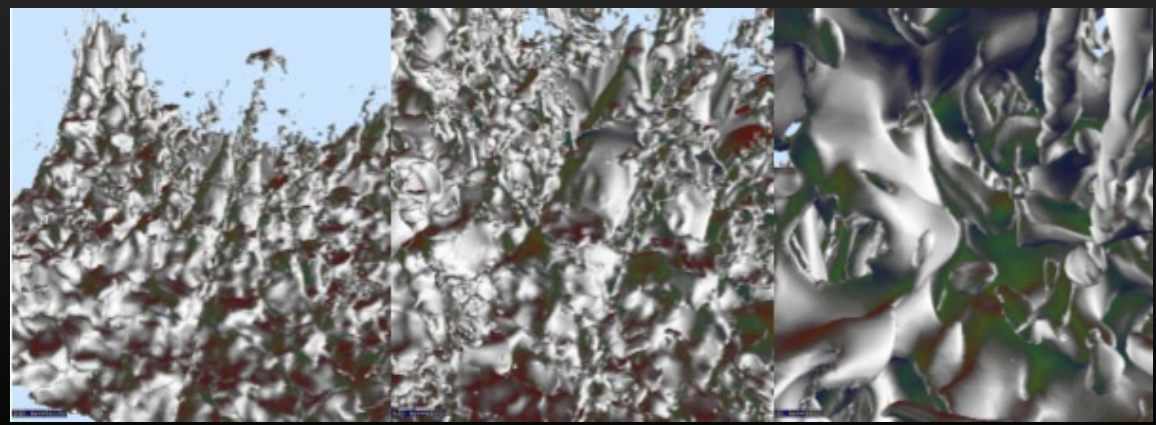




Office of Science
U.S. DEPARTMENT OF ENERGY

Another “Big Data” Solution: Save and Analyze only Interesting Data

- A researcher is focusing effort on a specific line of inquiry. Engineering vs. scientific discovery.
- Large, parallel simulation includes some visualization processing code.
- “Throwing away data” has an opportunity cost.



(Image from ASCI TSB project)

Office of Science
U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

Alternative: Query-Driven Analysis

- Combines scientific data management and visualization/analysis technology.
- Quickly locate scientifically interesting or relevant data from a larger, complete collection (don’t throw data away).
- Limit processing in downstream analysis pipeline to smaller-sized data subset.
- This approach adaptable to many different deployment alternatives: big hammer, specialized hammer, etc.

Query-Driven Visualization and Analysis

- **New capability: Bitmap Indices – find data records/cells that meet search criteria.**
  - (500<temp<1000) && (pressure<10.0mb) && (CH4>10ppm)
- **New capability: For spatial data, generate connected regions from records/cells returned by search.**
- **Exceptional performance:**
  - Searches evaluated in linear time proportional to number of hits as opposed to number of data records/points.
- **Widely applicable: Search results are input to visualization or analysis tools.**

Office of Science
U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

What is a Bitmap Index?

Data values	b_0	b_1	b_2	b_3	b_4	b_5
0	1	0	0	0	0	0
1	0	1	0	0	0	0
5	0	0	0	0	0	1
3	0	0	0	1	0	0
1	0	1	0	0	0	0
2	0	0	1	0	0	0
0	1	0	0	0	0	0
4	0	0	0	0	1	0
1	0	1	0	0	0	0
	=0	=1	=2	=3	=4	=5

- **Compact: one bit per distinct value per object.**
- **Easy to build: faster than common B-tree**
- **Efficient to query: use bitwise logical operations.**
  - (A < 2) AND (b_0 OR b_1)
- **Efficient for multi-dimensional queries.**
  - Use bitwise operations to combine the partial results
- **What about floating point data?**

Office of Science U.S. Department of Energy

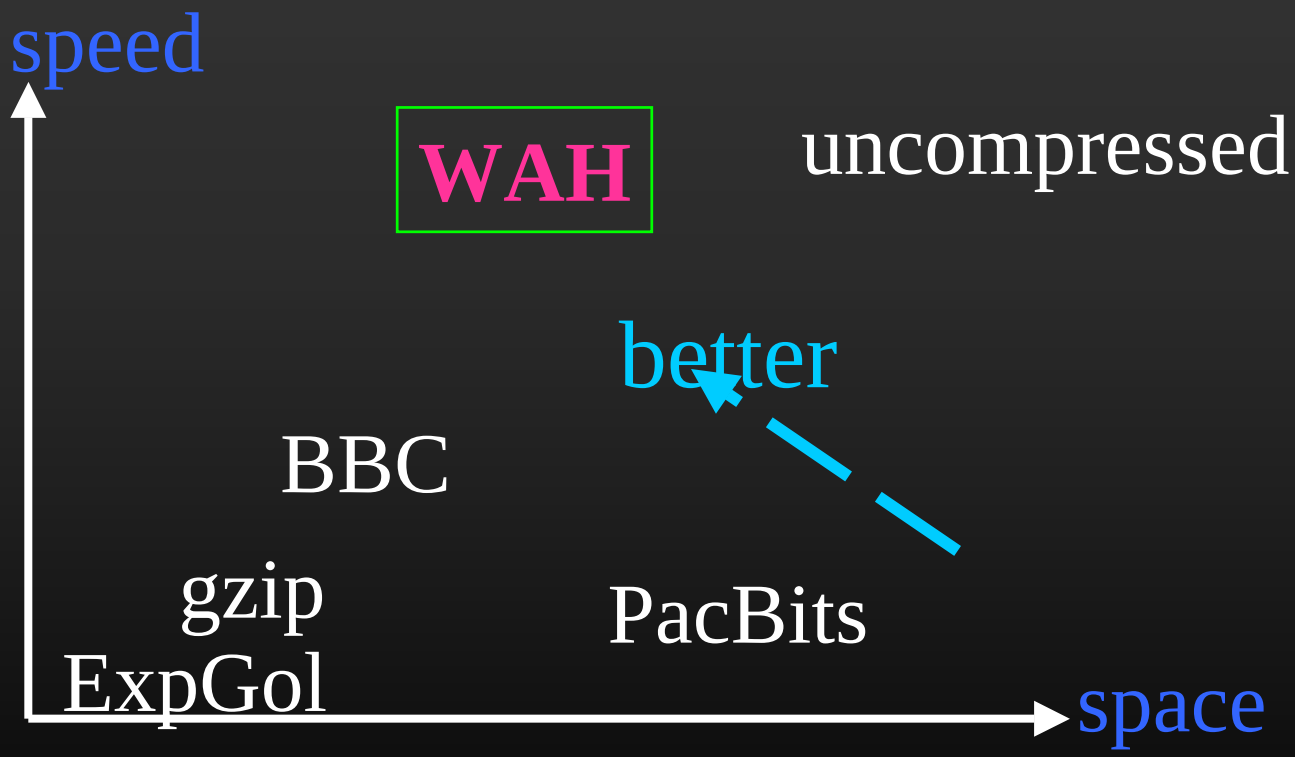
Berkeley Lab

Bitmap Index Compression

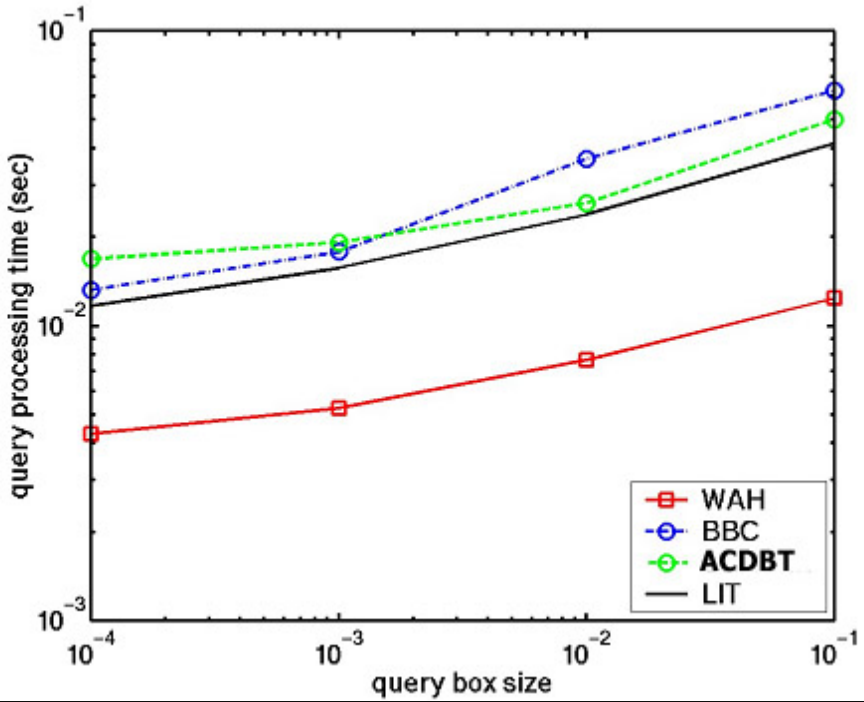
- Let N denote the number of objects and H denote the number of hits of a condition
- Using uncompressed bitmap indices, search time is O(N)
- With a good compression scheme, the search time is O(H) – the theoretical optimum.

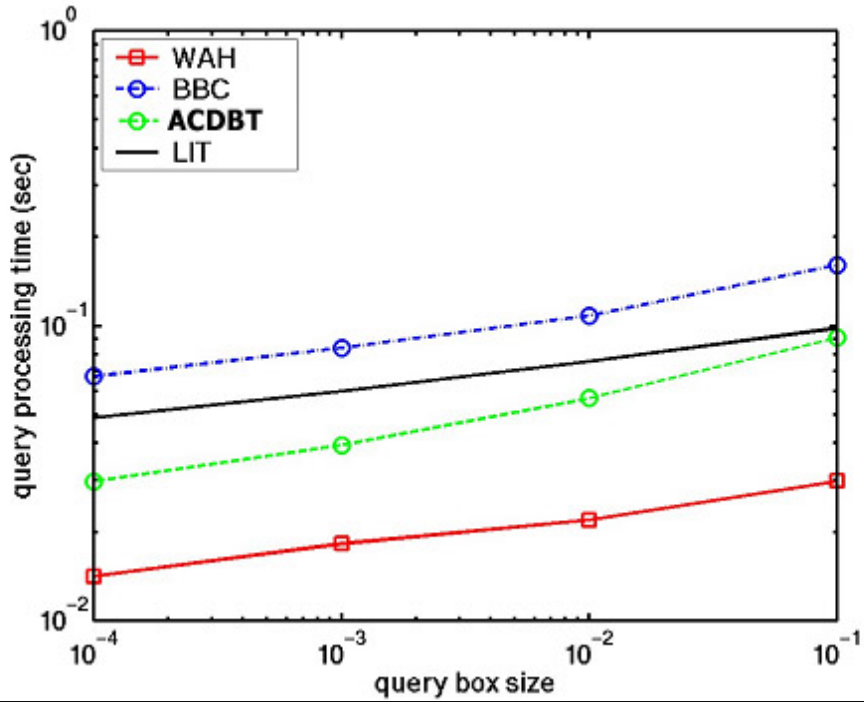
- In the worst case (completely random data), the bitmap index requires about 2x in data size.
- On the average, we've seen a cost of 1/10th the size of the original data.

Word-Aligned Hybrid Codes – Fast and Compact



WAH Query Performance





Office of Science
U.S. DEPARTMENT OF ENERGY

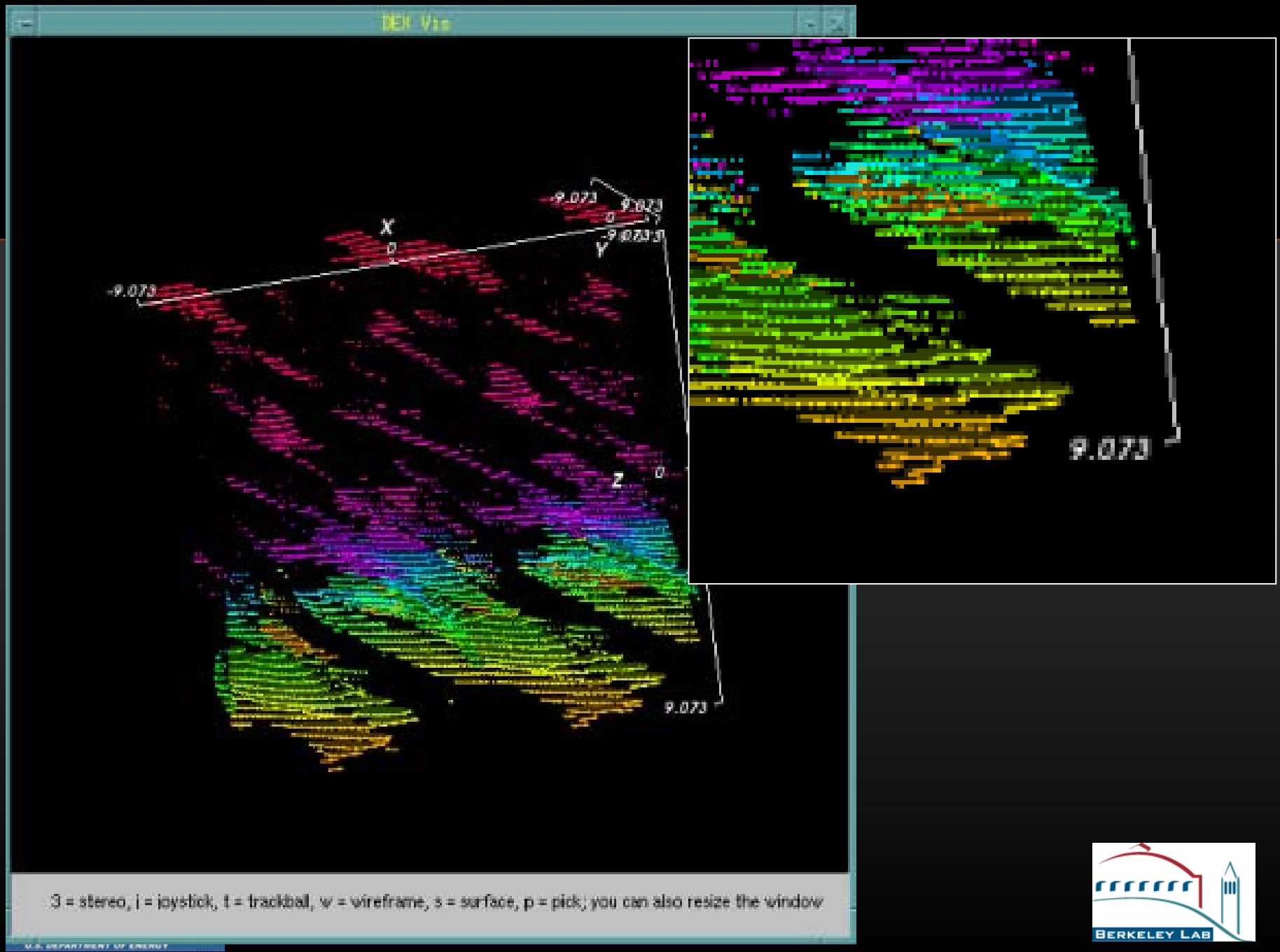
BERKELEY LAB

What Does This All Mean for Scientific Research?

- **More productive science:**
  - E.g.; Locate regions of data relevant to line of scientific inquiry and focus processing/analysis on “interesting regions.”
- **Through new analysis capabilities:**
  - Traditional visualization tools (slice, crop, isosurface) fall short of meeting current scientific needs.
  - Multidimensional queries directly addresses many types of scientific inquiry.
- **With less time-to-solution:**
  - Bitmap index searches are theoretically optimum.

Office of Science
U.S. DEPARTMENT OF ENERGY

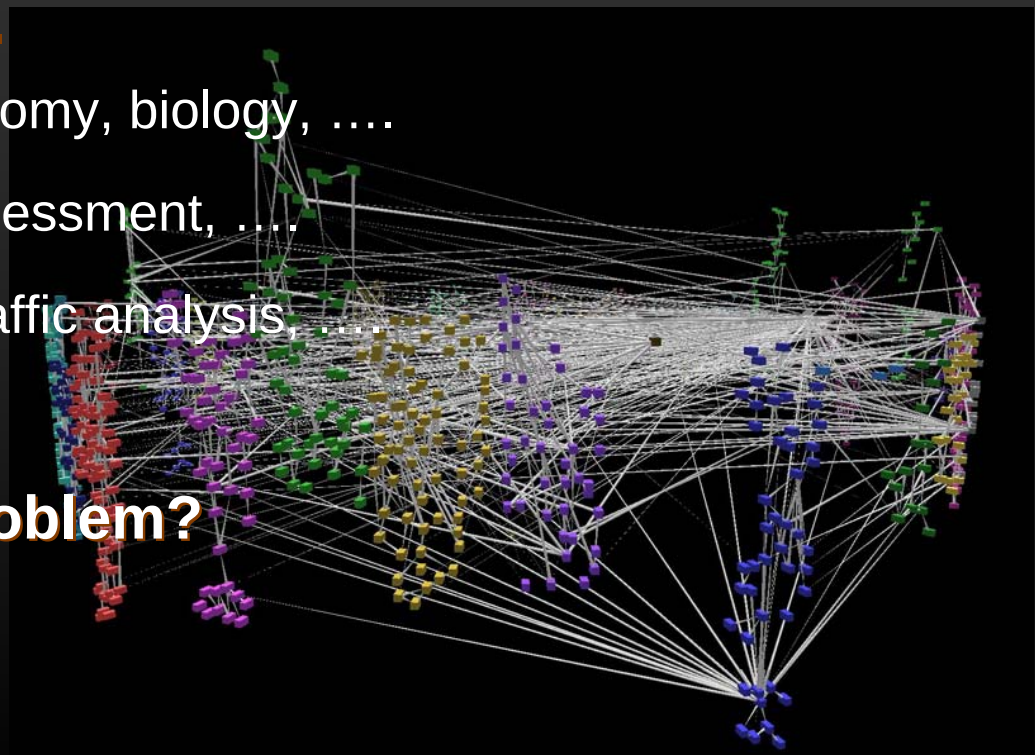
BERKELEY LAB


DEX Vis
X
Y
Z
-9.073
9.073
0
-9.073
9.073
-9.073
0
9.073
0
9.073
3 = stereo, j = joystick, t = trackball, w = wireframe, s = surface, p = pick; you can also resize the window
BERKELEY LAB

Some Potential Uses

Multidimensional "Data Google"

- **Not only data values, but relationships between data elements.**
  - Scientific: physics, astronomy, biology, ….
  - Economic: Credit risk assessment, ….
  - Cybersecurity: internet traffic analysis, ….



Toehold on Data Babel problem?

Office of Science
U.S. DEPARTMENT OF ENERGY

Query-Driven Analysis Themes

- **Human judgment guides how to extract meaningful data from large and complex data collections.**
- **QDVA, when combined with interactive analysis pipelines, accommodates well-known cognitive processes:**
  - Switching between macro and micro views.
  - Data equivalent of motion parallax.
- **A patented, highly efficient data analysis capability.**

Query-Driven Analysis Future

- Multiresolution queries, temporal queries.
- Queries across federated sources.
- “Embedded” bitmap indexing as a filter in real-time, stream-processing applications.
- As the basis for comparative and integrative visual data analysis.

The End