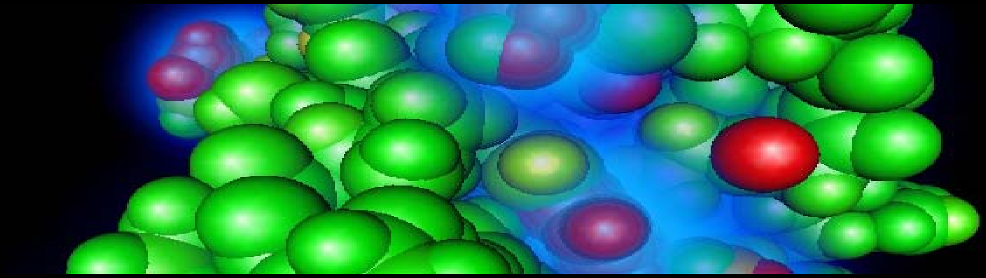


Visualization and SDM

Match Made in Heaven? Or ….

March 02, 2005

SDM All Hands Meeting

Salt Lake City, UT

E. Wes Bethel

with help from Friends at

Lawrence Berkeley National Laboratory

The Central Questions

- Does "it" save me time?
- Does "it" really help me do my job better?
- Am I really better off than without "it?"

Presentation Themes

- **“Data Google” – Query Driven Visualization and Analysis.**
- **Leveraging Human Intuition to Accelerate Scientific Discovery**
  - Why Vis is not a box on the end of the workflow diagram.
- **Parallel visualization retrospective: a decade of hits – 1995-2005.**
  - “Why don’t we all just use parallel VTK?”

Problem Statement

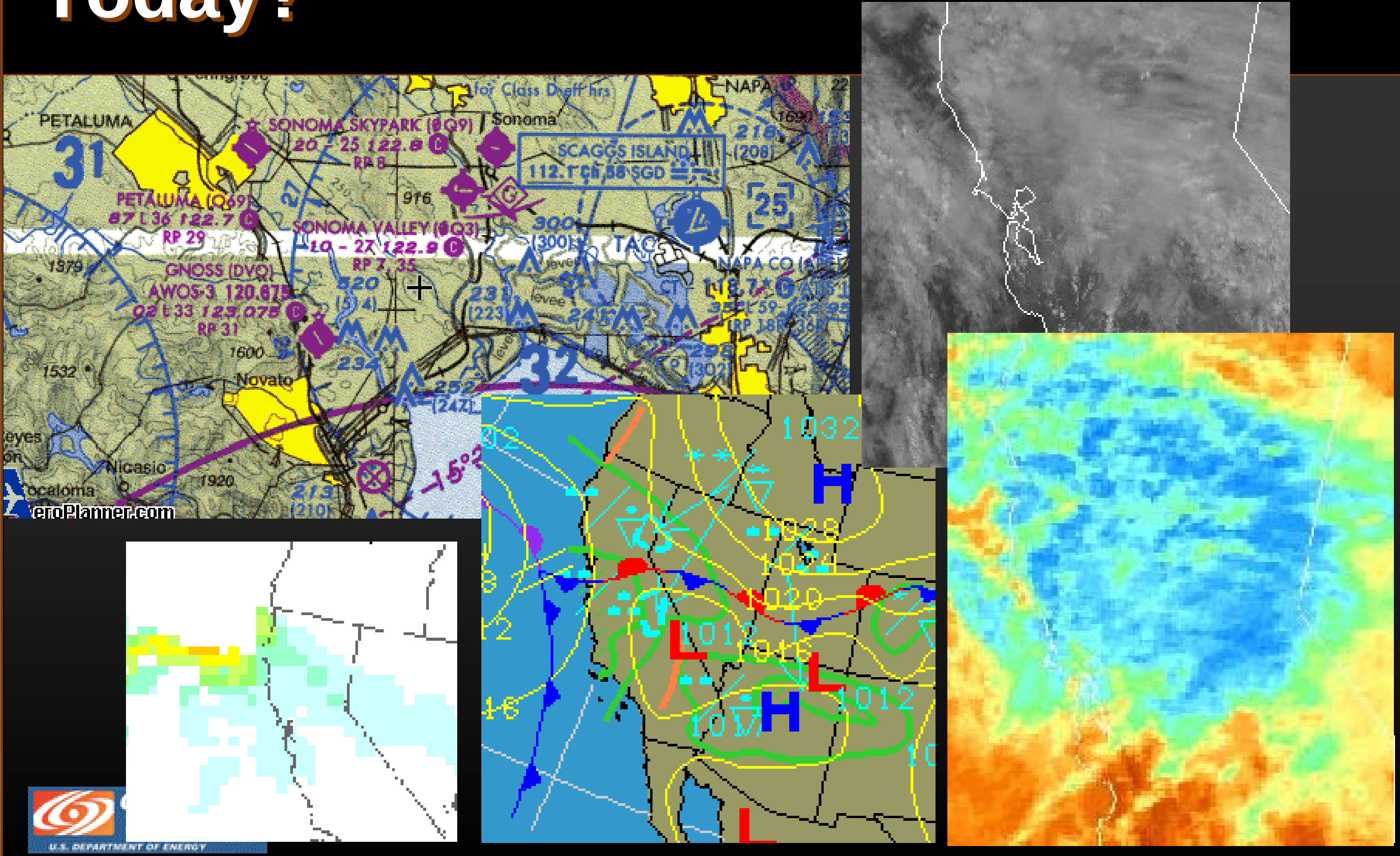
- **Information management is a limiting factor in many sciences and endeavors:**
  - Time: You have 20 minutes between tokomak experiments to analyze results from previous run and set parameters for next one.
    - Did the magnetic field lines stabilize in the last run?
    - What happened in that other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

Problem Statement

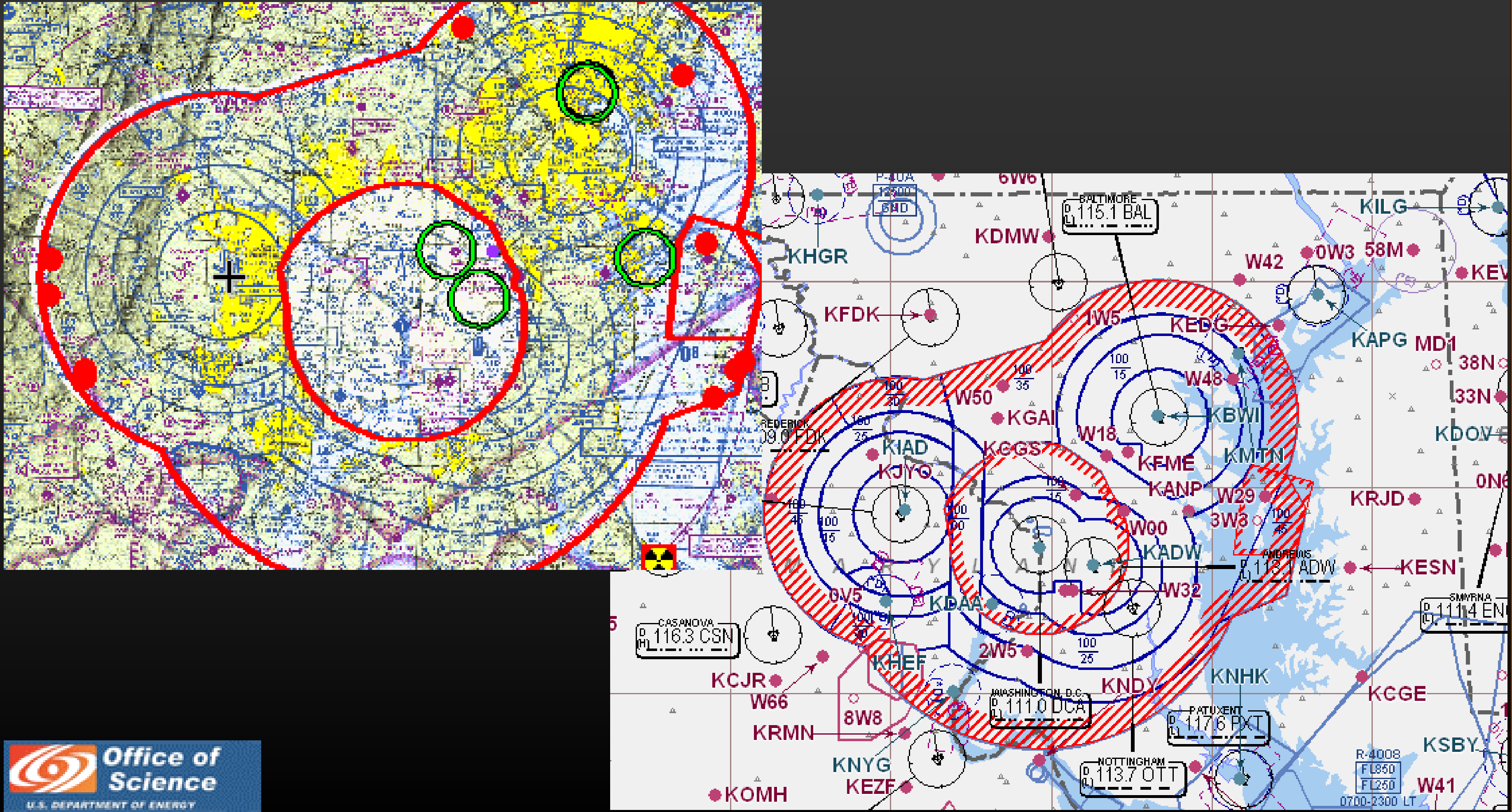
- **Simple questions give rise to startling complexity.**
  - Will a new malaria vaccine be effective?

    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 landeing?)

A Simple Question: Should I Fly Today?



The Simple Question Becomes More Complex When Considering All Available Data

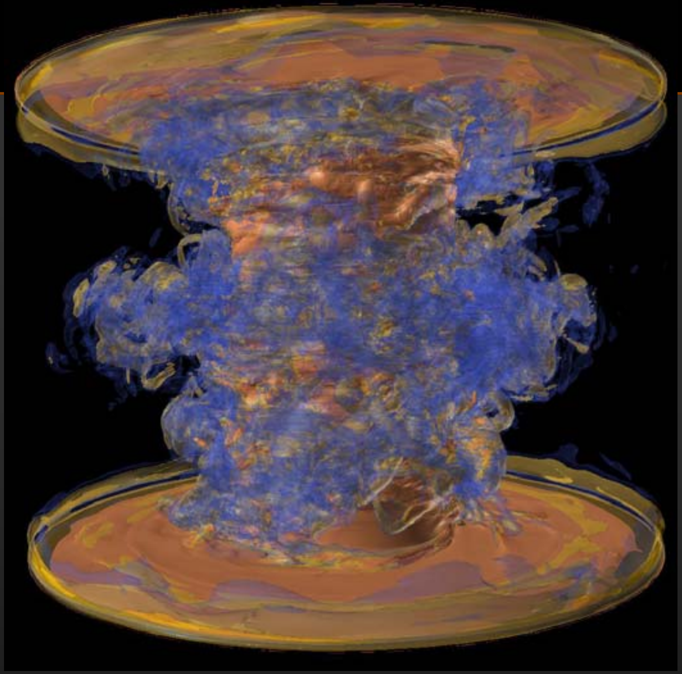


Office of Science
U.S. DEPARTMENT OF ENERGY

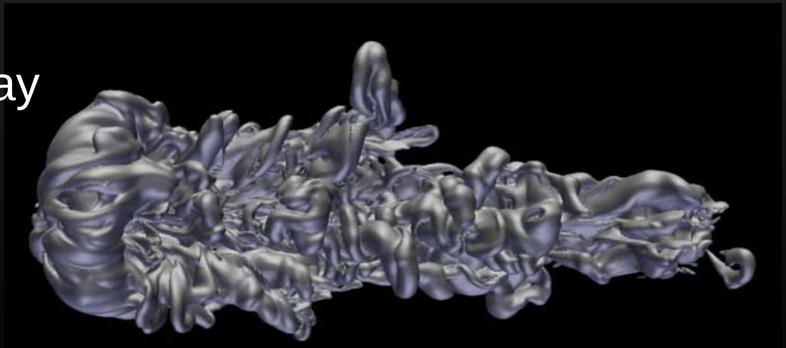
Dimensions of the Problem

- **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, …**

One “Bigger Data” Solution: Use A Bigger Hammer



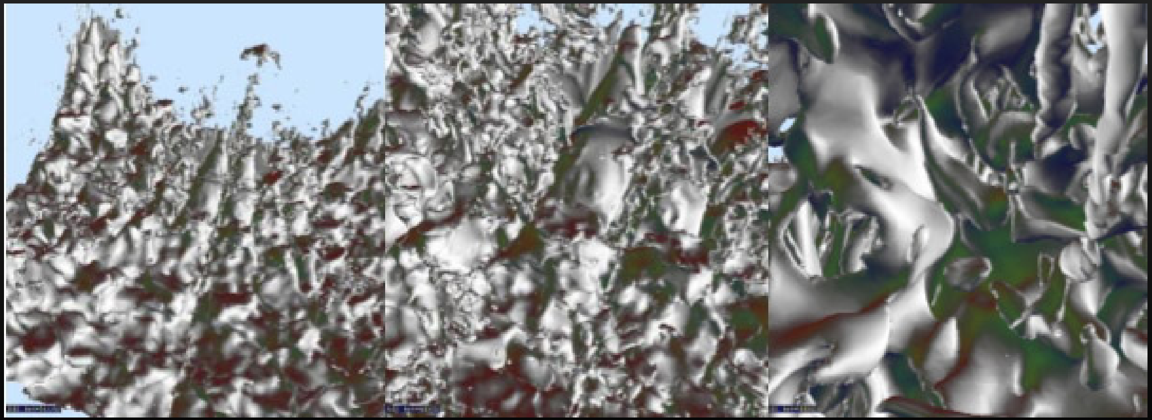
- **Scalable solutions for processing larger data using existing algorithms.**
  - Faster computers, scalable tools produce increased capacity – humans ought to be able to visually process the increased load.
- **Some known problems:**
  - Doesn’t really solve the “overwhelmed with data” problem.
  -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

Office of Science
U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

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.**

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*

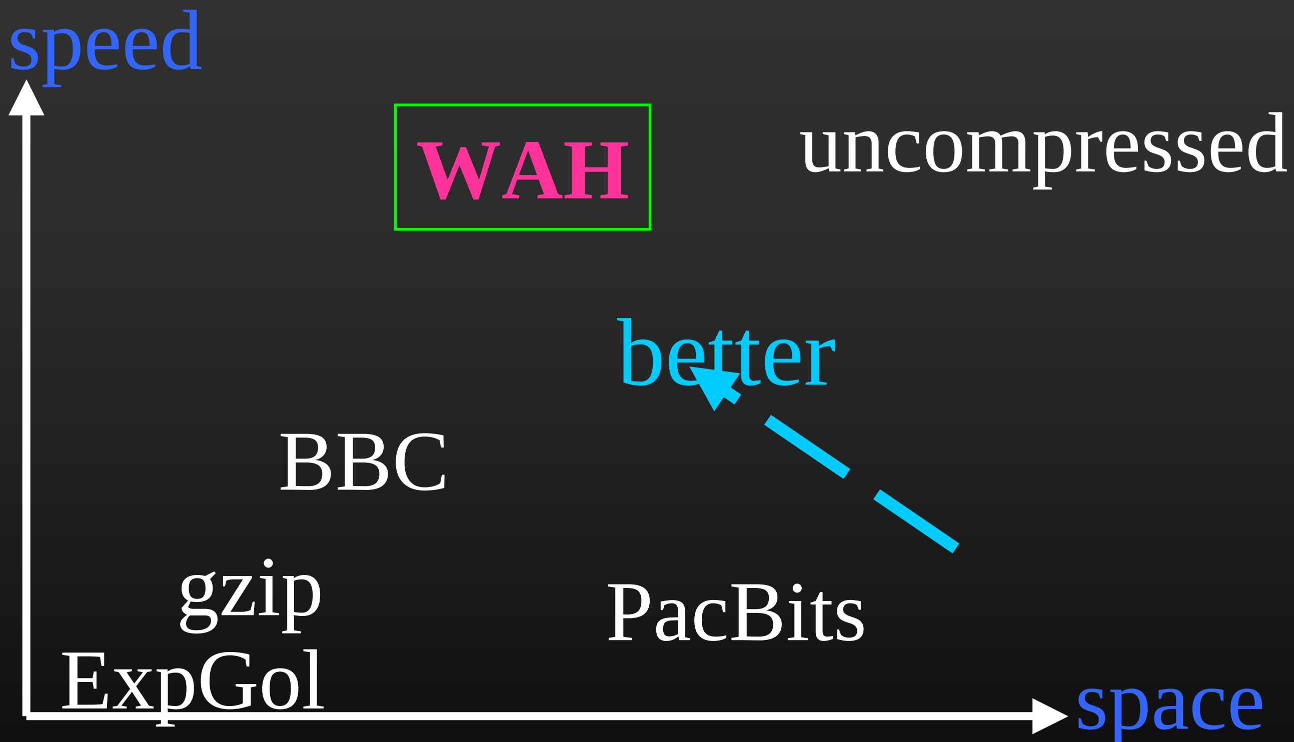
- **Compact: one bit per distinct value per object.**
- **Easy to build: faster than common B-tree**
- **Efficient to query: use bitwise logical operations.**
  - (A < 2) AND (b_0 OR b_1)
- **Efficient for multi-dimensional queries.**
  - Use bitwise operations to combine the partial results
- **What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

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 costs 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



speed

WAH

uncompressed

better

BBC

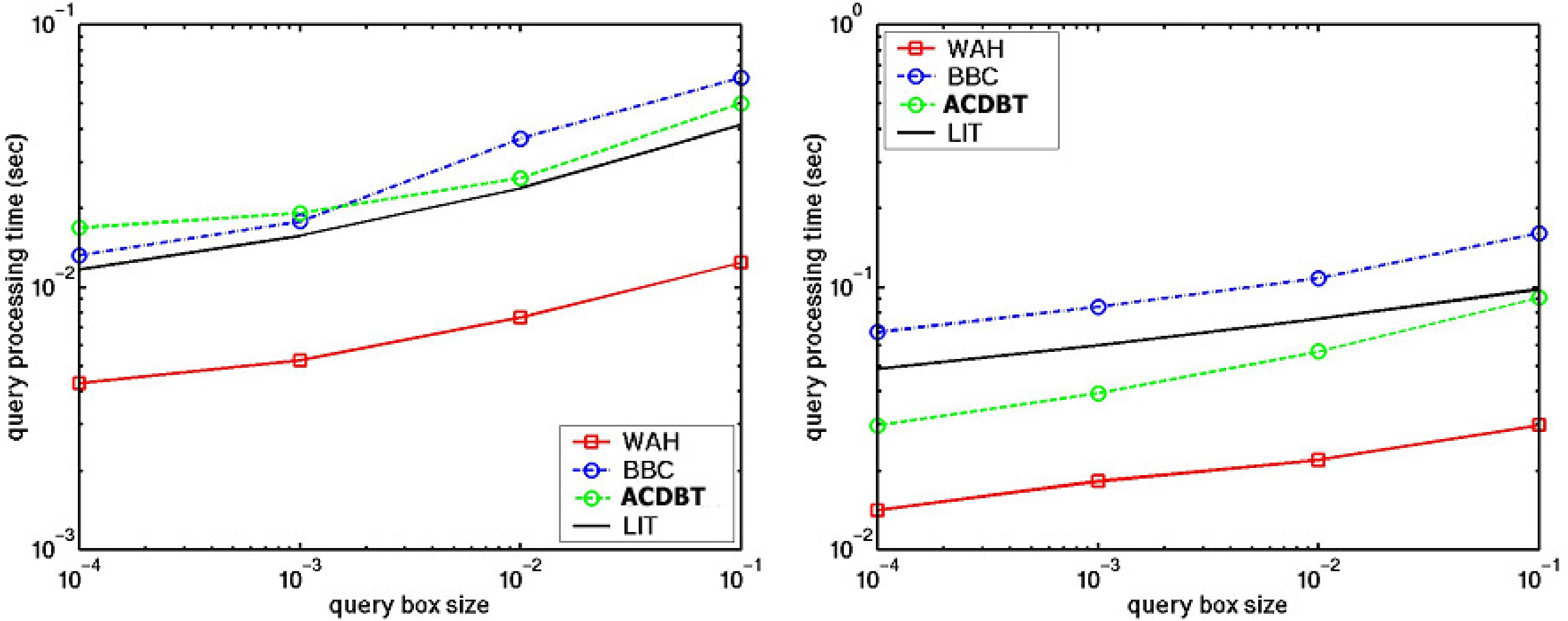
gzip

PacBits

ExpGol

space

WAH Query Performance



Office of Science
U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

What Does This All Mean for Scientific Research?

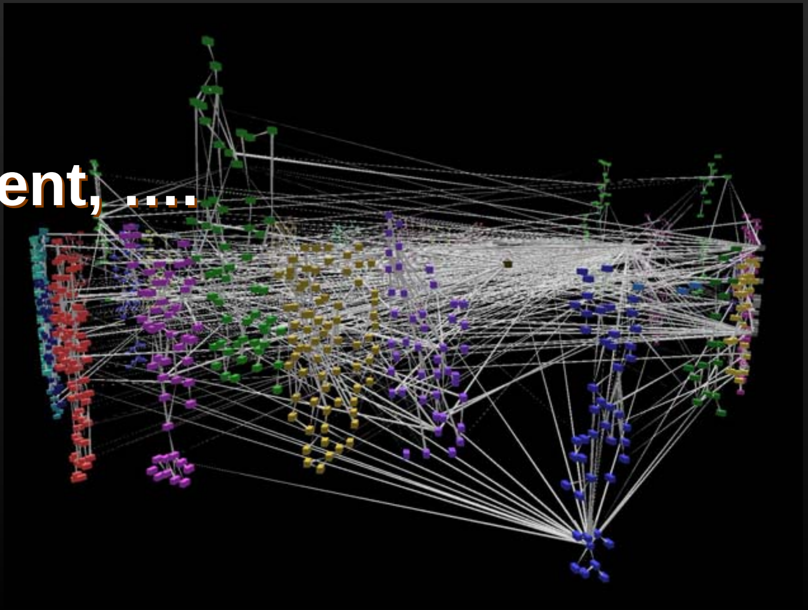
- **More productive science:**
  - E.g.; Locate regions of data relevant to line of scientific inquiry and focus processing/analysis on “interesting regions.”
- **Through new analysis capabilities:**
  - Traditional visualization tools (slice, crop, isosurface) fall short of meeting current scientific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

Some Potential Uses

Multidimensional "Data Google"

- Not only data values, but relationships between data elements.

Image courtesy of S. Smith, LANL/LBNL



- Scientific: physics, astronomy, biology, ...
- Economic: Credit risk assessment, ...
- Cybersecurity: internet traffic analysis, ...

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.

(Next, Predicting Protein Structure)

Predicting Protein Structure

- **Grand challenge in computational biology:**
  - Function follows from form:
    - Hemoglobin’s shape allows it to carry oxygen.
    - Collagen’s shape is ideal for connective tissue.
- **Why predict protein structure?**
  - Knowing shape is critical for designing therapeutic drugs.
  - Crystallization/x-ray diffraction is time-consuming, expensive and not always possible.
  - NMR methods don’t work well with large proteins.
  - The simulate vs. experiment argument.

Predicting Protein Structure – Workflow

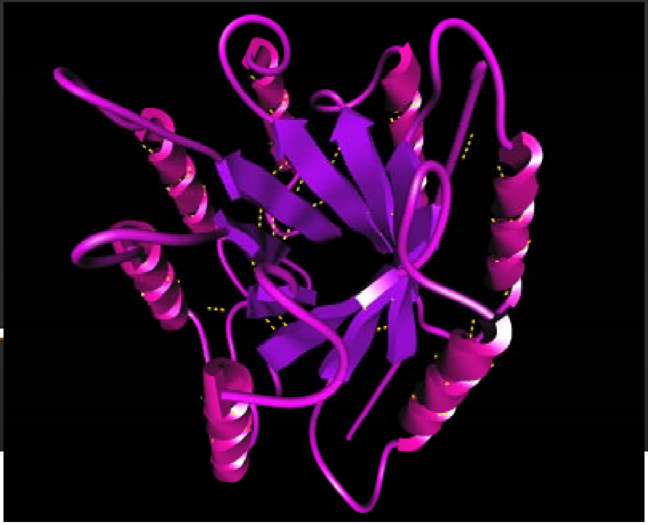
Given: an amino acid sequence,

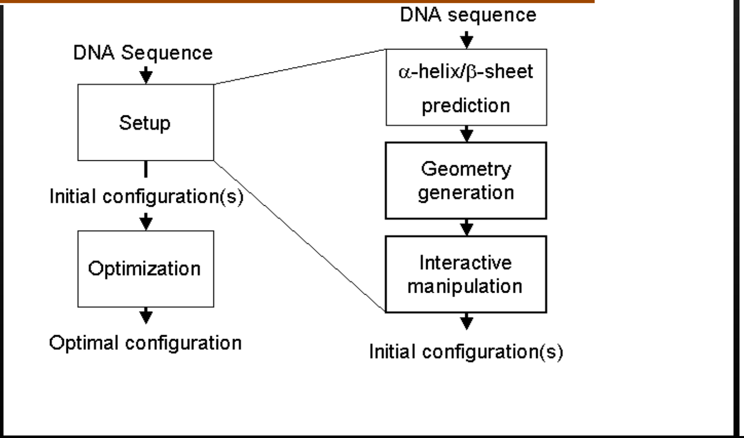
Find: an optimal protein conformation.

Many approaches are possible.

- Physics-based, knowledge-based, …

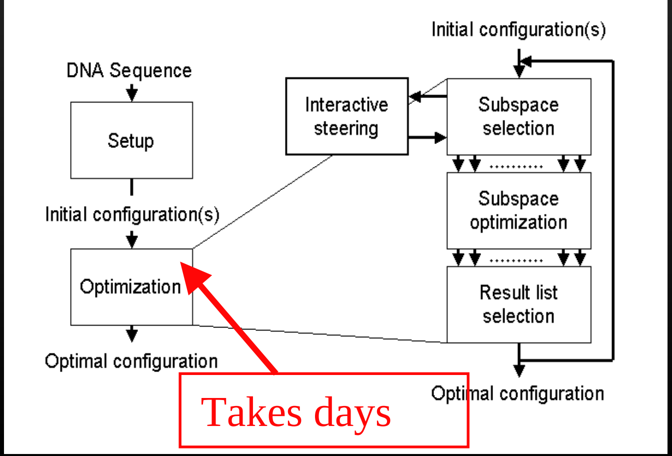
```
Conf: 99999999999999999999999999999999999
Pred: HHHHHHHCCCEEEEEEECCCEEEEEEEECCCCCCC
AA:   FKQYANDNGVDGVWTYDDATKTFTVTEMVTEVPVA
```





DNA Sequence
Setup
Initial configuration(s)
Optimization
Optimal configuration

DNA sequence
α-helix/β-sheet prediction
Geometry generation
Interactive manipulation
Initial configuration(s)



DNA Sequence
Setup
Initial configuration(s)
Optimization
Optimal configuration

Initial configuration(s)
Interactive steering
Subspace selection
Subspace optimization
Result list selection
Optimal configuration

Takes days

Predicting Protein Structure – Physics-Based Approach

- Given sequence of amino acids, the primary structure,
- Classify groups of them into secondary structures (alpha helices, beta strands),
- Arrange secondary structures into some 3D arrangement (tertiary structure).
- Adjust dihedral angles of atoms comprising protein chain.
- When minimal energy level is discovered, you’re done!

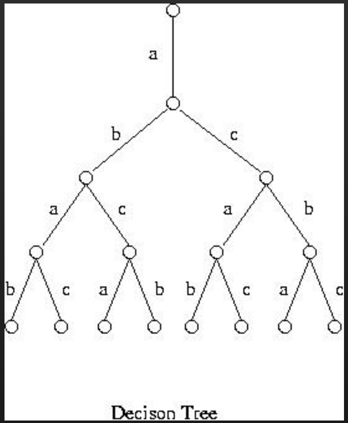
- Problem: finding minimum energy conformation can take days on large supercomputers.

Predicting Protein Structure – Energy Optimization

“Find Optimal Conformation”

- **Problem:** **what is the minimal-energy structure of a sequence of amino acids?**
- **How to compute:**

1. Begin with molecule in some configuration.
2. Compute total internal energy for that configuration.
3. Is energy minimal? If so, stop.
4. If not, adjust dihedral angles of molecule to create a new conformation.
5. Proceed to step 2 and repeat.



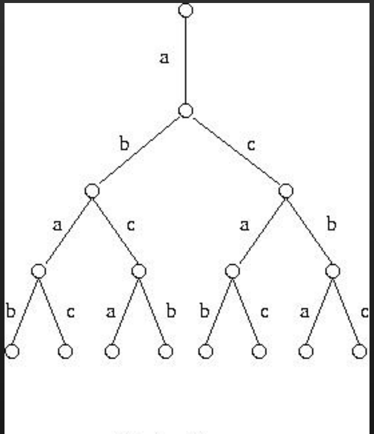
Decison Tree

Energy Optimization Significance and Complexity

- **Theory – a protein's "final" shape is one in which it has minimum internal energy.**
  - 1972 Nobel Prize in Chemistry, Anfinsen, NIH.
- **Confirmed by comparing theoretical and experimental results.**
- **Energy Optimization is not known to be solvable:**
  - Combinatorial optimization problem.
  - Because search space is not discrete (dihedral angles), the complexity is (theoretically) not bounded by problem size.

Quick Overview of Software Complexity

- **Deterministic algorithms: well-defined sequence of steps.**
- **P-problems: complexity is bounded by problem size.**
- **NP problems (Non-deterministic, polynomial): e.g., combinatorial optimization.**
    - Traveling salesman: must try all possible combinations.
    - Complexity bounded by P, the problem size.
- **Energy optimization falls into a category of problems where complexity is not bounded by problem size!!**



Decison Tree

Office of Science
U.S. DEPARTMENT OF ENERGY

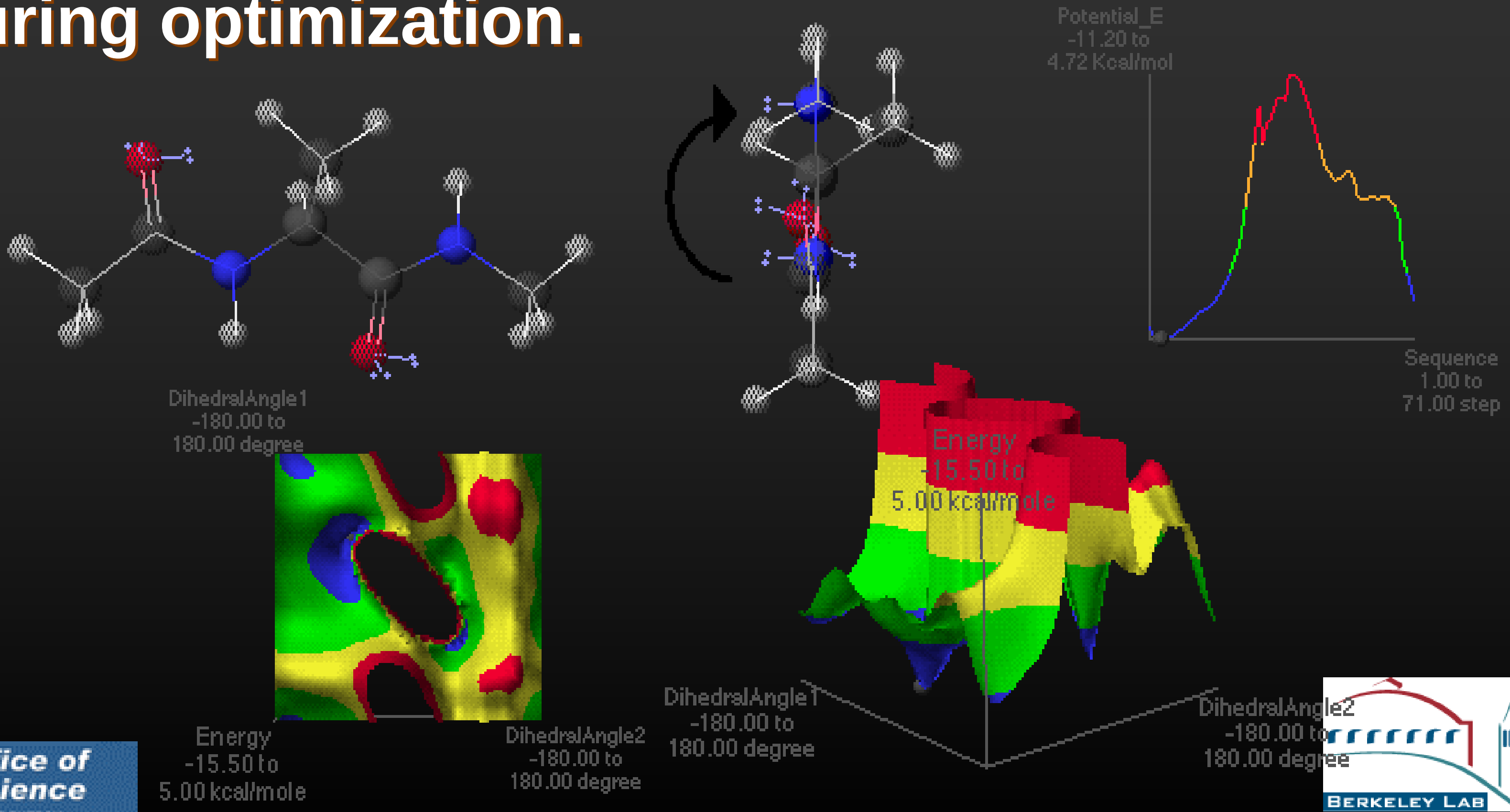
BERKELEY LAB

Our Approach: ProteinShop

- Reduce size of search space by focusing processing “most promising” candidates.
- Use human intuition to define the group of most promising candidates.
- Use human intervention during optimization to further refine the group of “most promising candidates.”

What is a Good Initial Candidate?

- One that won’t “get stuck” in a local minimum during optimization.

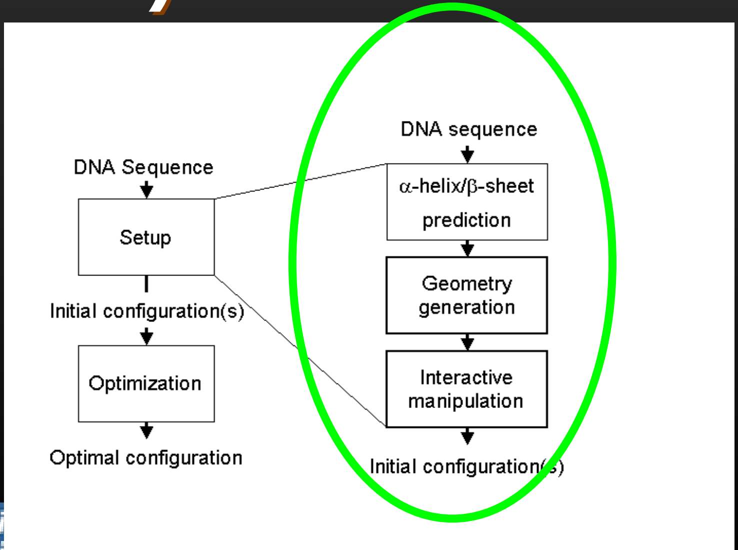


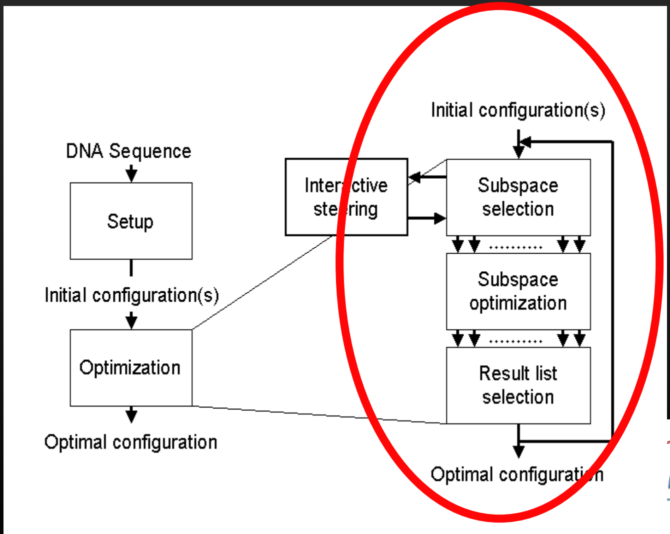
Office of Science
U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

Good Initial Configurations Reduce Time to Solution

- Better **initial configurations** (minutes, hours) reduce time-to-solution in **optimization phase** (from days or weeks to hours).





Why Domain-Specific Visualization and Interaction is Crucial

- Without constraints, it is possible to generate configurations that are not chemically meaningful (images produced with PyMol).

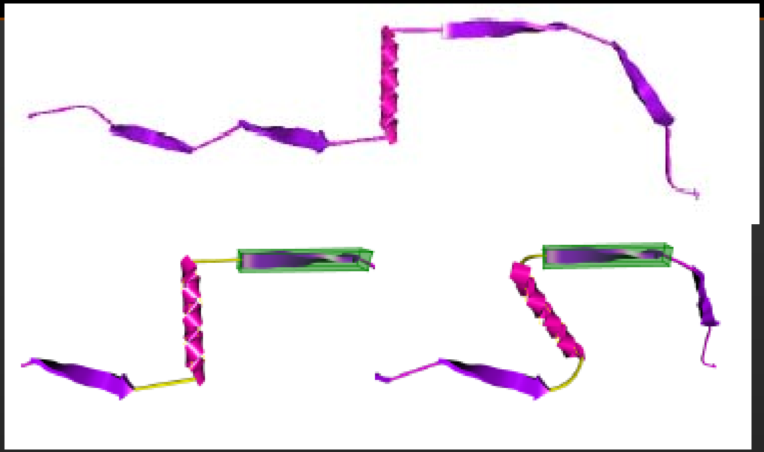
ProteinShop Interactive Manipulation

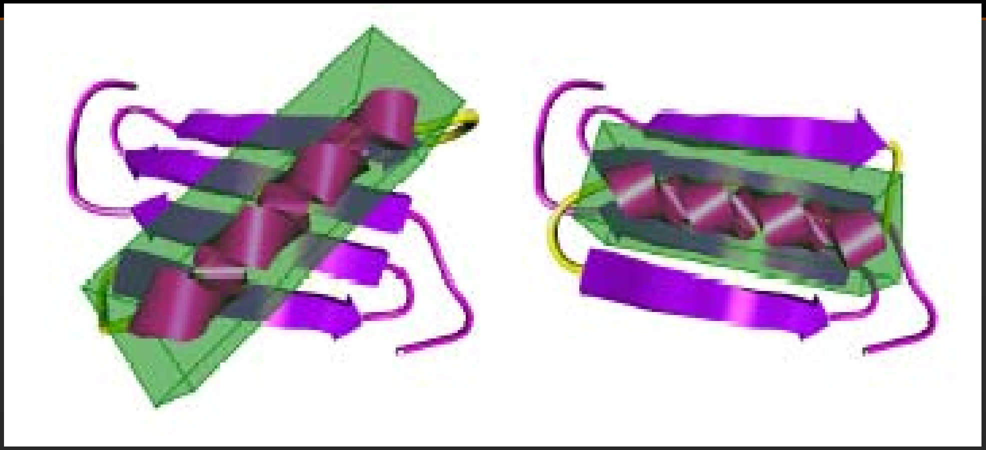
- **Inverse kinematics algorithm (from robotics) used to constrain transforms.**
  - While not chemically valid, it is physically valid, and has proven to be quite close to a chemically valid solution.
- **Interaction “handles” are secondary structures familiar to biochemists:**
  - Alpha helices, beta strands and coils.
  - Each secondary structure behaves differently during transformation.
  - Substantial improvement over other molecular visualization and manipulation tools.
- **“Guides” to help with “protein sculpting.”**

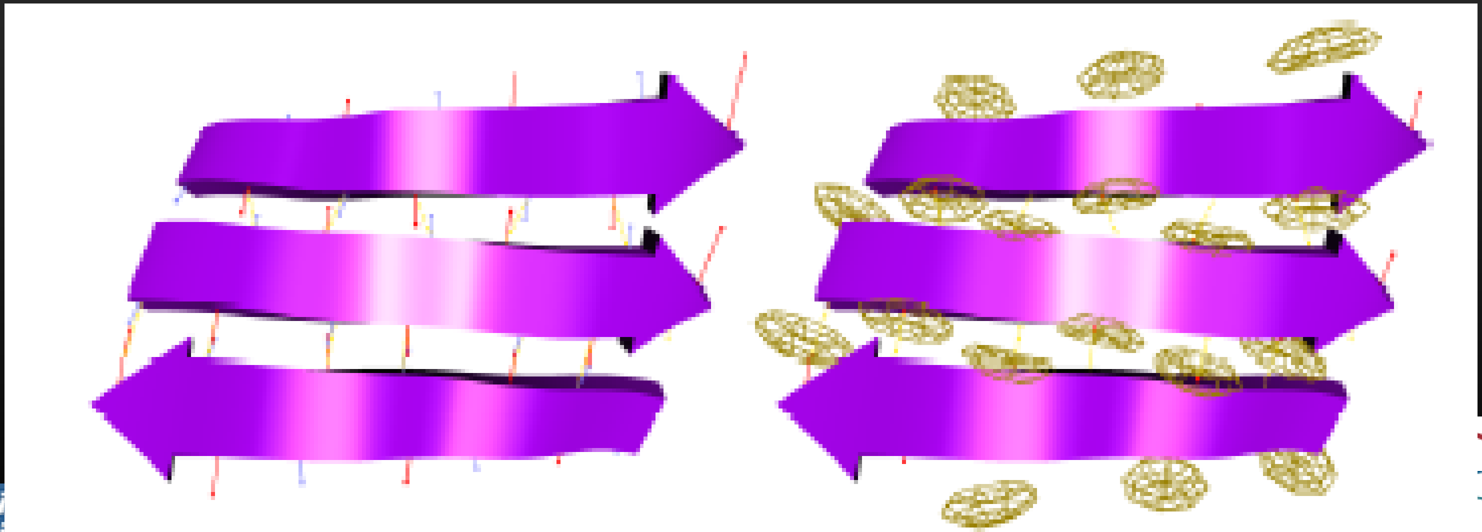
Office of Science
U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

ProteinShop Interactive Manipulation

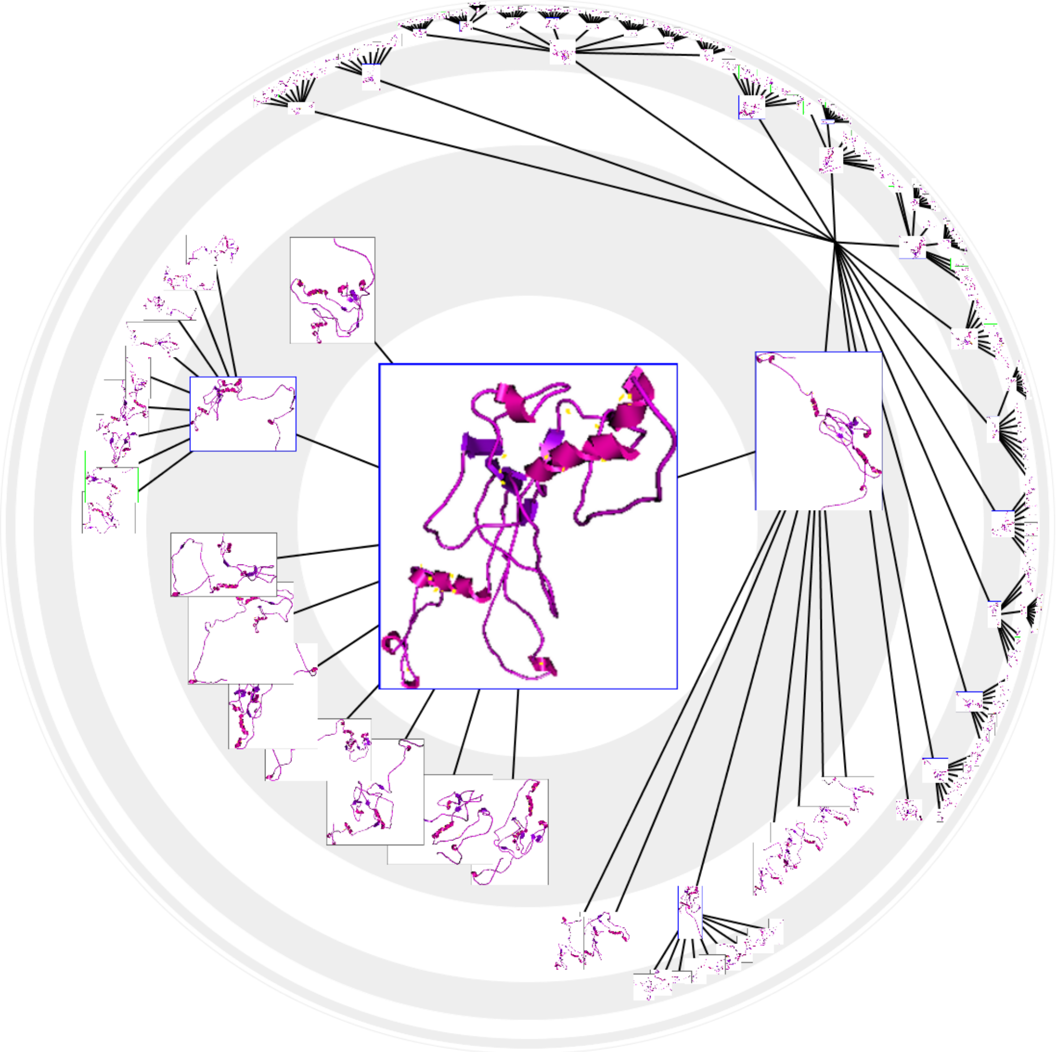






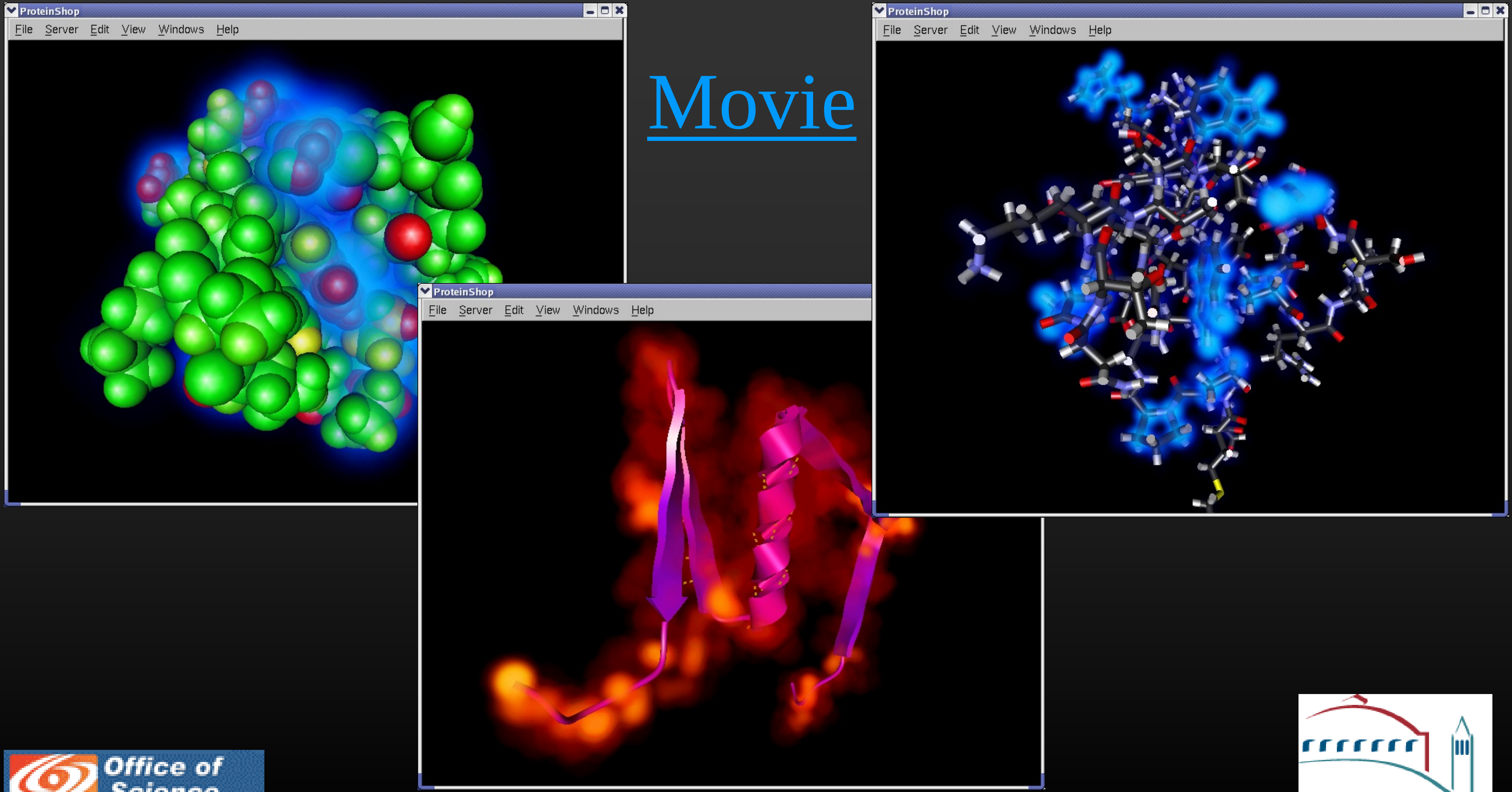
ProteinShop Guided Optimization

- Initial configurations used as “seed points” for optimization.
- Intermediate results – the “search tree” – is displayed for inspection.
- A human may intervene in the optimization and prune the search space.



Energy Visualization (Preliminary Results)

Movie



Office of Science
U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

ProteinShop Results (need final data)

- **CASP4 (2000, no ProteinShop)**
  - Eight targets, max length 240 amino acids.
- **CASP5 (2002, w/ProteinShop)**
  - 20 targets, max target length 417 amino acids.
  - Time to generate initial configurations reduced from days to hours.
- **CASP6 (2004, w/ProteinShop)**
  - Avg 80th percentile in results (very good!).
  - Best score for most difficult target.

Office of Science
U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

Human Intervention Accelerates Protein Structure Prediction

- **Use of biochemical knowledge with domain-specific interactive visual analysis tools to produce good initial configurations.**
  - Reduces time to solution by avoiding local minima.
  - More computational time spent on best candidates.
- **Prune search space during optimization to reject unpromising families of optimization targets.**

Office of Science
U.S. DEPARTMENT OF ENERGY

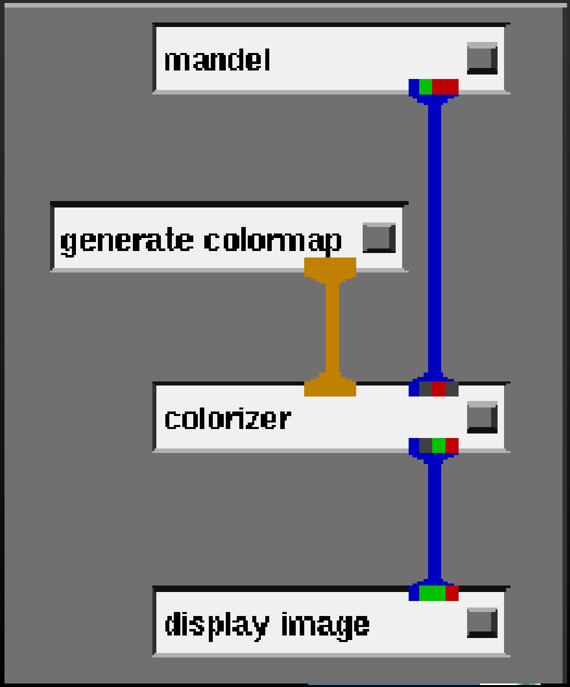
BERKELEY LAB

Parallel Visualization – Decade of Hits and Misses, 1995-2005.

- **(1995-1999) Early efforts: Distributed modules in AVS & Khoros, parallel AVS on the CM5, T3D, etc.**
- **(2000-2001) Point solutions: Visapult.**
- **(2002-present)**
    - Building on VTK: ParaView, VisIt.
    - Parallel AVS redux: Parallel Support Toolkit for AVS/Express
    - Alternative approaches: SCIRun, SCIRun2.
- **(2006-2010) Crystal Ball**

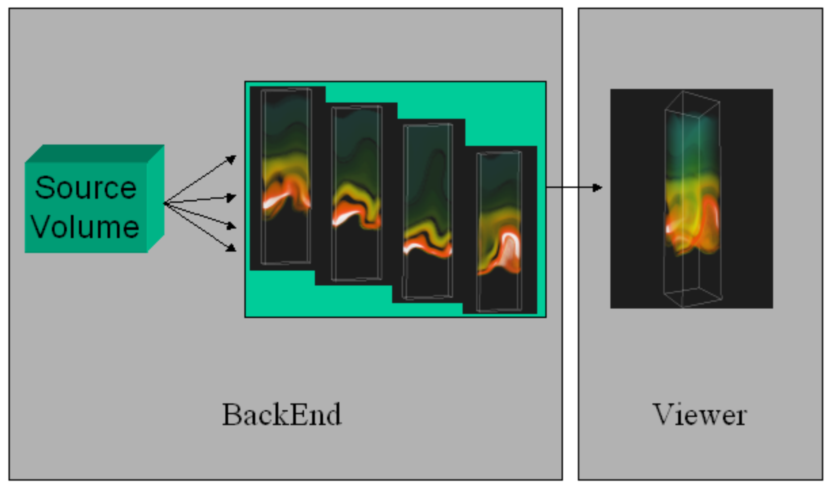
Early Efforts (circa 1995-1999)

- **Parallel AVS Modules for the T3D (http://www.cs.uaf.edu/~roth/t3d_avs/)**
- **Parallel modules play nicely with the rest of AVS, hides complexity of scatter/gather and parallel execution (mostly).**
  - Standard AVS data types.
- **Run-time "binding," stable (often inadequate) data models.**



Point Solutions: Visapult (2000-2001)

- Scalable, latency tolerant.
- “Difficult” to use, user-unfriendly.
- Not based on any “standards.”
- Layer-4 protocol is akin to “run-time” binding.
- Dead project.



More Recently (2002-present)

- **LLNL's VisIt (www.llnl.gov/visit)**
  - Idea: use visualization algorithms from VTK, deploy on parallel platforms to achieve scalability.
  - To make it work:
    - Completely repackage VTK algorithms to
    - Effectively manage execution.
  - Key problem: VTK tightly couples algorithm implementation, data and execution models (compile-time binding).
  - Underlying algorithms completely replaced with non-VTK versions (maintainability, etc.)

More Recently (2002-present), ctd.

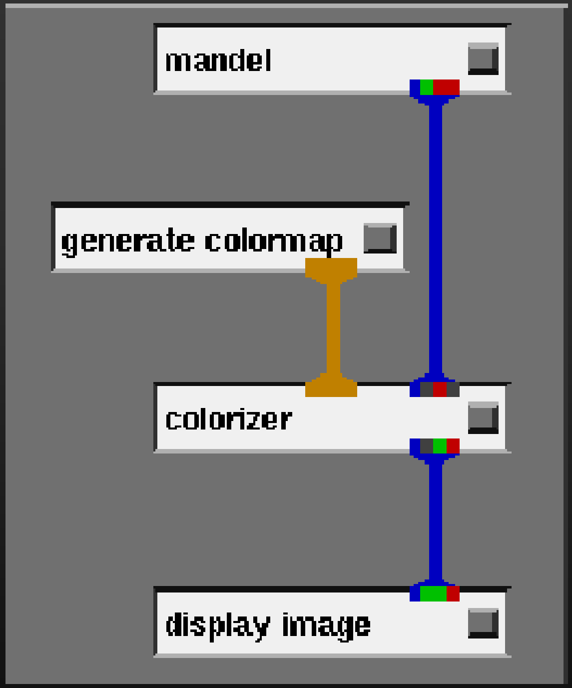
- **ChomboVis, point-n-click AMR vis app for AMR data.**
  http://seesar.lbl.gov/ANAG/chombo/chombovis.html
- **Same theme as VisIt:**
  - Want to use VTK's visualization algorithms,
  - But not their execution model, which doesn't (didn't) work for multi-grid problems.

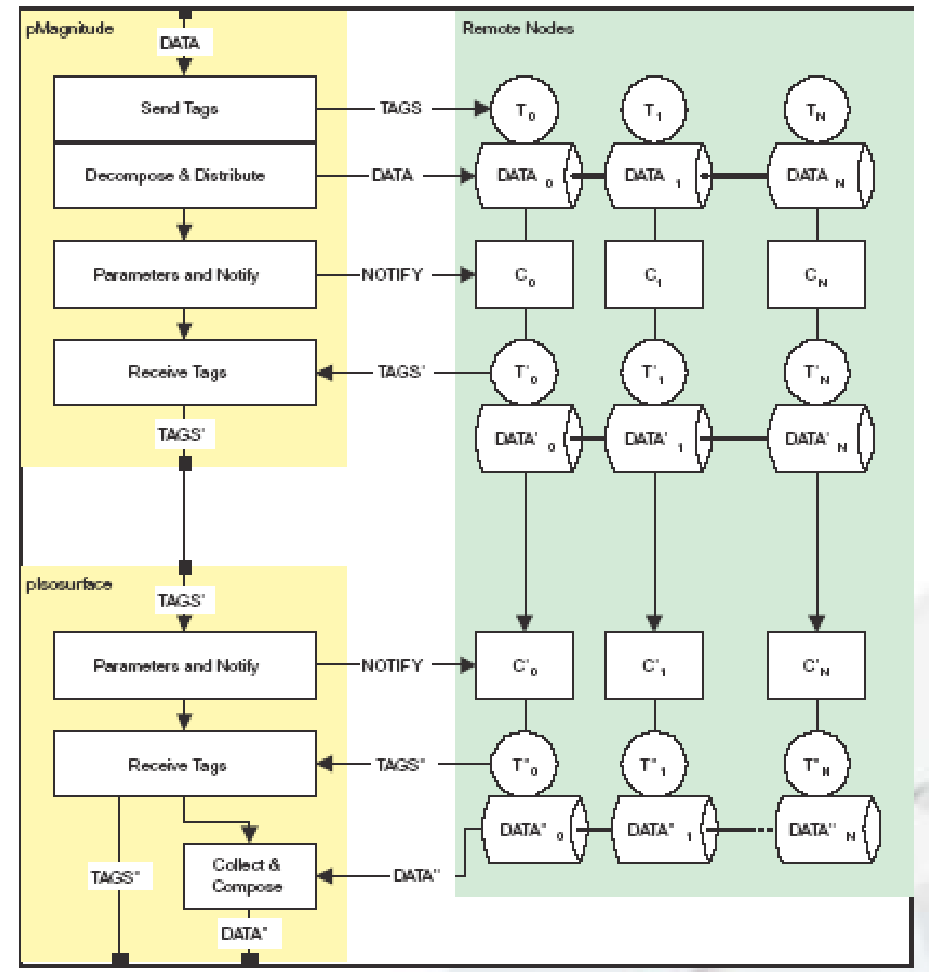
Digression – How Easy Was it to Create VisIt, ChomboVis

- VisIt – nine person-years spread over three people to create initial version.
- ChomboVis – about two person years, but built on ideas and implementation that had been gestating for many years previously.

- It’s not easy.

More Recently – PST for AVS





Office of Science
U.S. DEPARTMENT OF ENERGY

More Recently – PST for AVS

- Parallel Support Toolkit for AVS/Express
  http://www.csar.cfs.ac.uk/about/csarfocus/focus10/pst.pdf
- Runtime bindings, stable data model, looks like it supports old, serial and new, parallel modules.
- Product in 2004? Not visible on avs.com. (Maybe Klasky knows…)

More Recently (2002-present), ctd.

- **SCIRun, SCIRun2.**
  - Run-time bindings (no compile-time bindings).
  - SCIRun (1) – Parallel operation on SMP.
  - SCIRun2 – Adaptors for bolting together different species of components.

Musings on Decade of Hits and Misses

- **VTK has pros and cons.**
  - Pro: largest known collection of open source visualization algorithms.
  - Con: compile-time bindings hinder extensibility, adaptability, maintainability.
- **Fact: two examples of vis apps (VisIt, ChomboVis) went to great lengths to “mine” VTK algorithms.**
  - VisIt has abandoned VTK completely.

Musings on Decade of Hits and Misses, ctd.

- Fact and Opinion: run-time bindings increase productivity, compile-time bindings hinder productivity.
- Corollary: increased productivity results when one is insulated from complexity.
- Have we saved any time?
- Have we had a positive impact on application science?

Conclusion

- **Modern science is dominated (limited?) by information management challenges.**
- **Human intuition can and should play a key role in accelerating understanding in data-intensive scientific research.**
- **The two examples we present offer glimpses into promising avenues of better using human knowledge to accelerate scientific discovery.**

Conclusions

- Query-driven visual data analysis offers new capabilities to scientific researchers aimed at helping reduce “information overload” and information management.
- Human intuition coupled with domain-specific tools accelerate protein structure prediction by reducing time-to-solution and enabling study of larger and more complex problems.

The End

Further Reading

- NSF/NIH Workshop on Biological Data Management. http://pueblo.lbl.gov/~olken/wdmbio/ Feb 2003.
- DOE/OS Data Management Challenge. Sep 2004. http://www-user.slac.stanford.edu/rmount/dm-workshop-04/Final-Report-Work-Area/
- Improved Searching for Spatial Features in Spatio-Temporal Data <http://sdm.lbl.gov/~kurts/research/region-growing-lbnl-sept2004.ps>, Technical Report, LBNL-56376, Berkeley, California, September 2004
- LBNL Visualization Group: http://vis.lbl.gov/

Further Reading, ctd.

- ProteinShop. http://proteinshop.lbl.gov/. Fall 2004.
- Unfolding Proteins, Krell Institute. http://www.krellinst.org/csgf/mag/2003/print.cgi?id=2100
- The Art of Protein Structure Prediction. http://www.eurekalert.org/features/doe/2004-12/ddoe-tao122204.php
- "Interactive Protein Manipulation." In Proceedings of *IEEE Visualization 2003*, October 19-24, 2003, Seattle, Washington, USA, pp 581-588. LBNL-52414 (Best Application Paper Award) http://vis.lbl.gov/Publications/2003/Kreylos-ProteinShop-Vis2003.pdf

Office of Science
U.S. DEPARTMENT OF ENERGY

BERKELEY LAB