Computer Science Challenges at LLNL

Terence Critchlow

Center for Applied Scientific Computing
Lawrence Livermore National Laboratory

critchlow@llnl.gov
www.llnl.gov/casc/people/critchlow

Outline

- Lab mission
- Computational Challenges
- The DataFoundry Project
- Conclusions

LLNL is a multi-programmatic, multi-disciplinary, national security lab.

There are major projects in each focus area.

There are major projects in each focus area.

LLNL mission areas

Academic Collaborations

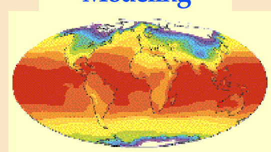
Environmental and Energy Sciences

Earthquake simulation
Climate Modeling

...gical ...ces

National Security

Global Climate Modeling



Lasers Optics

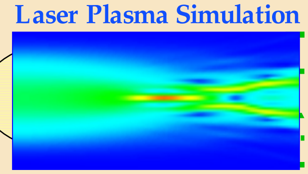
Industrial Partnerships

There are major projects in each focus area.

LLNL mission areas

Academic Collaborations

Laser Plasma Simulation



Environmental and Energy Sciences

National Security

Lasers Optics

National Ignition Facility

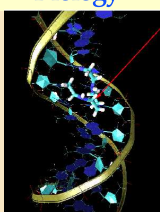
...erials ...ence

... Partnerships

There are major projects in each focus area.

LLNL mission areas

Computational Biology



Envir... and ... Sc...

Human Genome Project
Joint Genome Institute

Biological Sciences

National Security

Lasers Optics

Materials Science

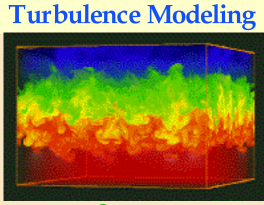
...orations

Industrial Partnerships

There are major projects in each focus area.

LLNL mission areas

Turbulence Modeling



...mental ...nergy ...nces

Biological Sciences

National Security

Laser... Optic...

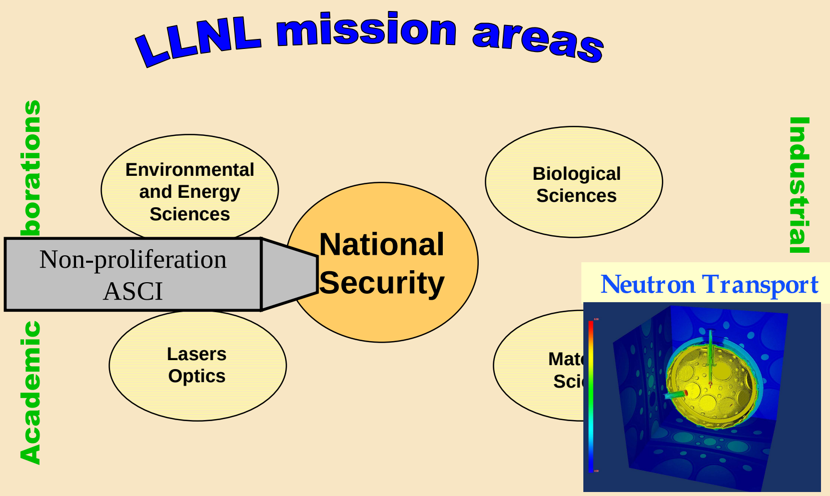
Material Modeling

Materials Science

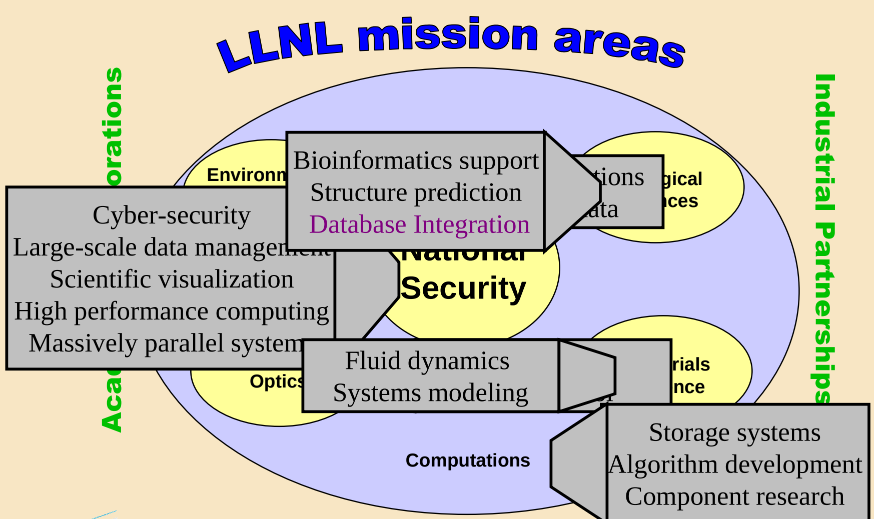
Academic Co...

Industrial Partnerships

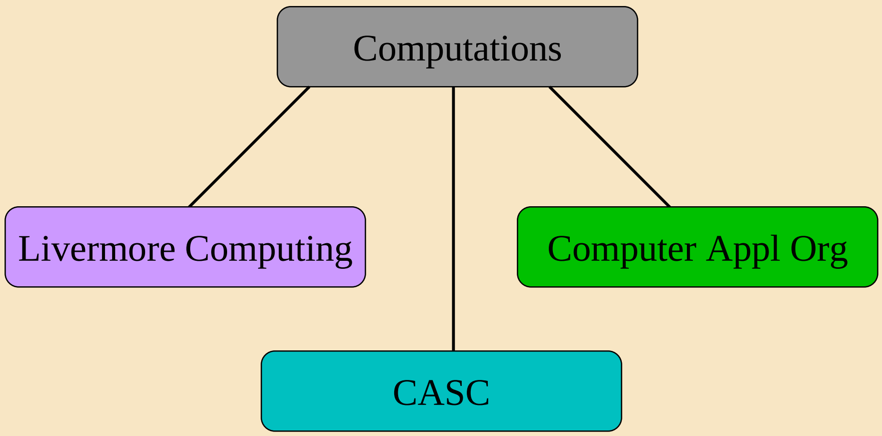
There are major projects in each focus area.



The Computations directorate supports all of the programs.



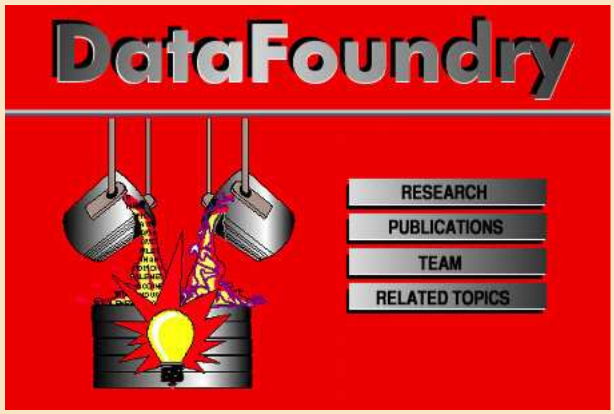
Computations has three major groups.



Computer and computational scientists go where they are needed.

- **People of all levels hired into Computations**
  - **Over 500 people (B.S.s to Ph.D.s)**
  - **Mostly computer science but also math and physics**
- **People assigned to specific projects**
  - **move between assignments as needed / wanted**
  - **lots of opportunities to pick up new skills**
- **Training to meet needs of new assignments**
  - **on-site classes**
  - **televised classrooms**
  - **support for continuing education**

I work on the DataFoundry project.



A brief history of DataFoundry.

Small, internal research effort started in 1997.

Goal
Improve scientists interaction with data.

Approach
Reduce cost of data warehousing to provide intuitive access to consistent data.

DataFoundry is a collaboration between CASC and the Biology directorate.

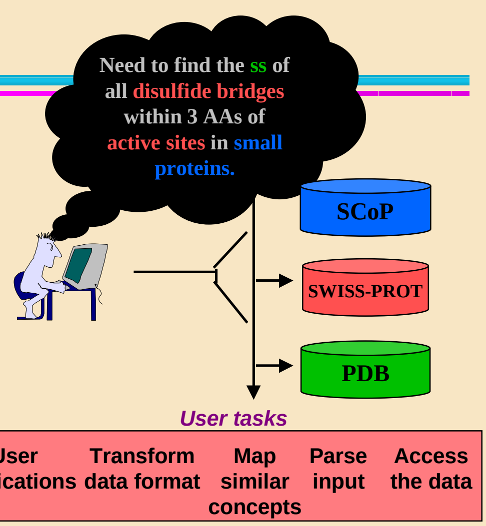
- **Team members:**
  - Ron Musick (CASC)
  - Tom Slezak (CAO/ BBRP)
  - Krzystof Fidelis (BBRP)

  - Terence Critchlow (CASC)
  - Madhavan Ganesh (CASC)
  - Ida Lozares (CAO)

  - Several summer students and visiting faculty

Current environment

- **Data is**
  - Hard to find
  - Hard to understand
  - Hard to reconcile
  - Hard to analyze

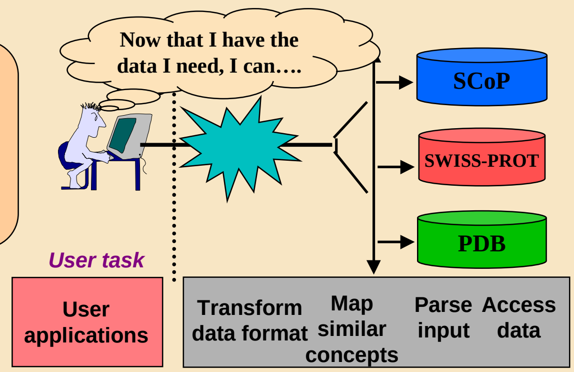


Scientists waste time and energy doing data management.

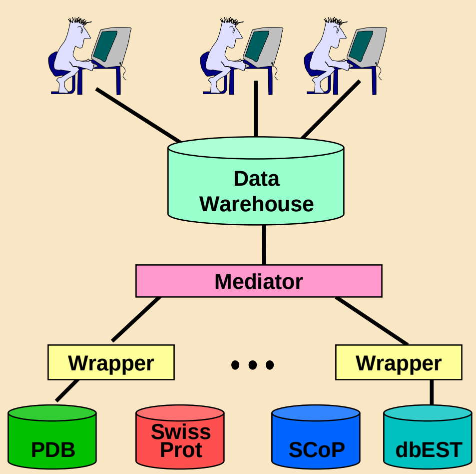
What is our ideal environment?

A *single* location that provides *effective* access to a *consistent* view of data from *many* sources through an *intuitive and useful* interface.

Businesses use data warehouses to accomplish this.

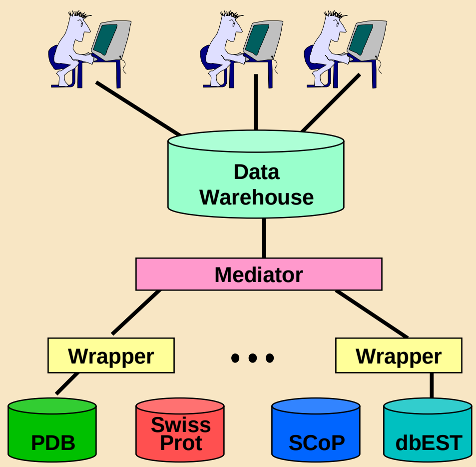


Data warehouses



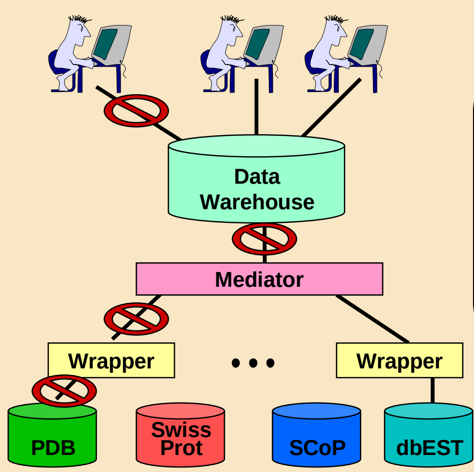
A data warehouse is a repository that provides a single access point to a collection of data obtained from a set of distributed, heterogeneous sources.

Data warehouses



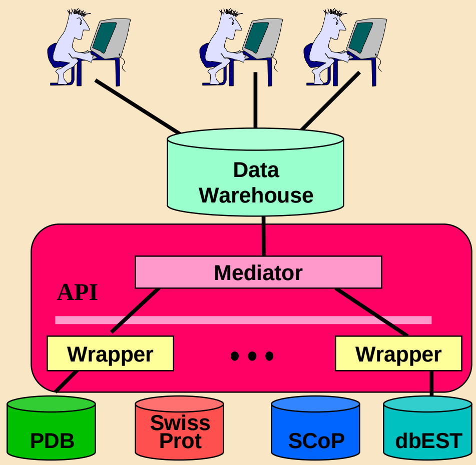
- Interfaces
  - provide intuitive access to the data
  - possibly change data format to meet user expectations
- Warehouse
  - stores a consistent view of data in a local repository
- Mediator
  - transform data from source format to warehouse format
- Wrappers
  - read data from source into internal representation

Warehouses don't work in dynamic domains



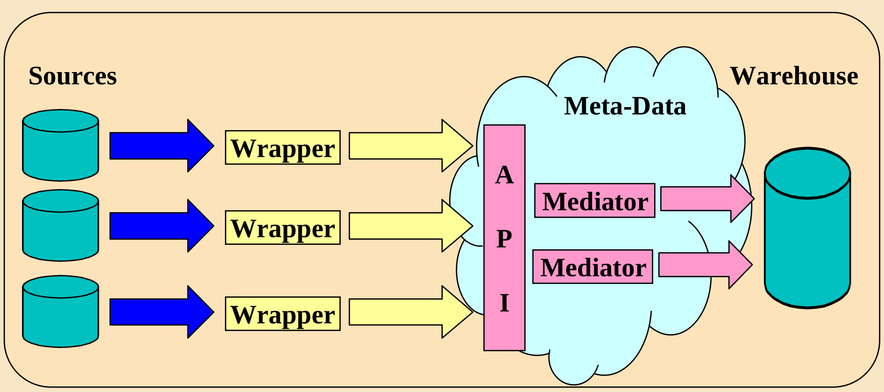
When schemata are modified, or new sources are added, wrappers and mediators break.

Key insight

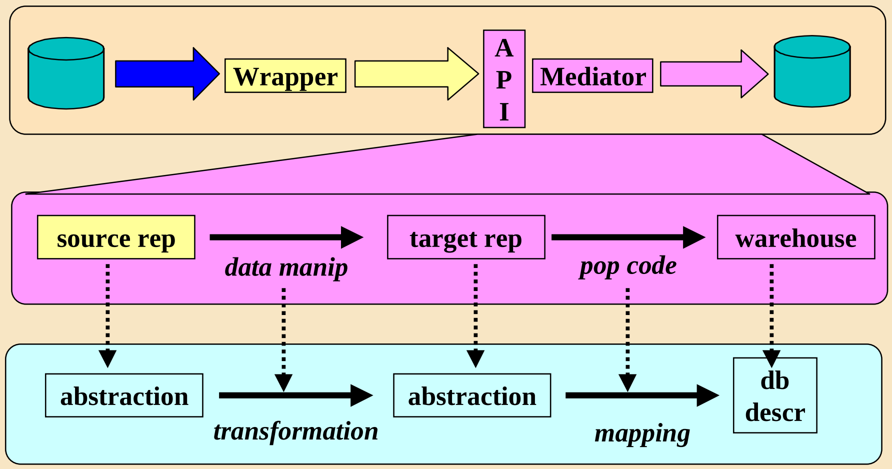


Extensive use of meta-data can dramatically reduce maintenance costs.

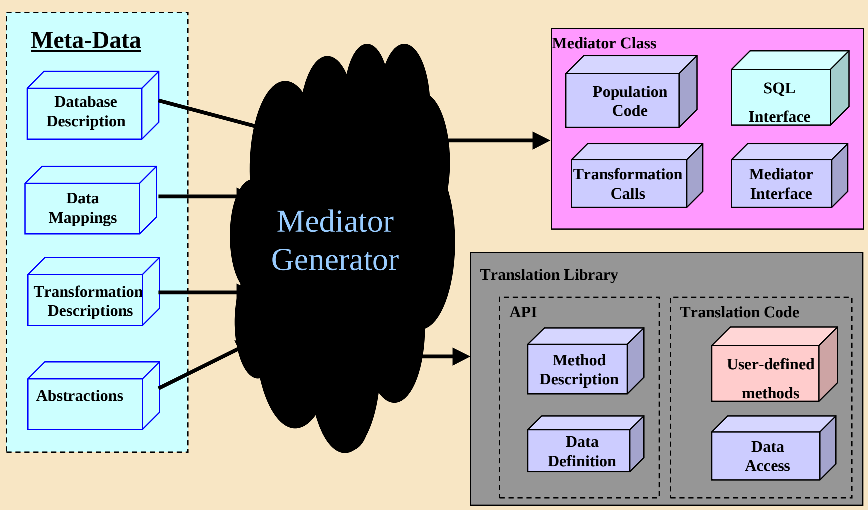
The DataFoundry approach:



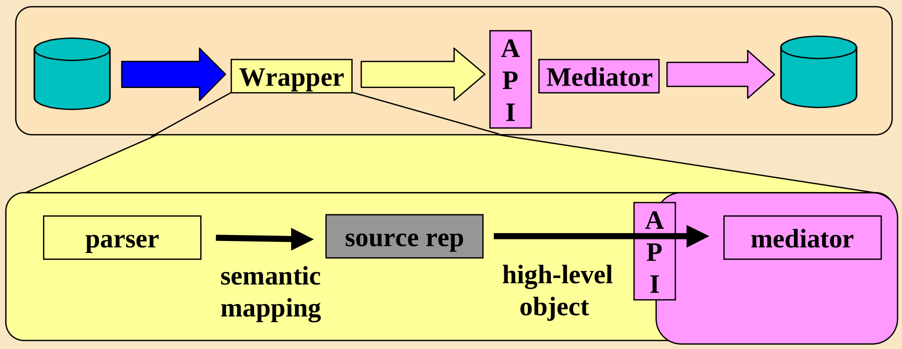
Four types of meta-data are required



Generating the mediators



The translation library and mediator class are used by the wrapper



Results:

Integrating SCoP into warehouse that already contains PDB and SWISS-PROT.

Activity/ integration style	manual	meta-data	diff	%diff
understanding SCOP	2.0	2.0	0.0	0
writing wrapper	4.5	2.5	2.0	44%
modifying schema	0.5	0.5	0.0	0
writing mediator	4.0	0.0	4.0	---
modifying meta-data	0.0	1.0	(1.0)	---
total time in days	11.0	6.0	5.0	45%

Results:

Integrating SCoP into warehouse that already contains PDB and SWISS-PROT.

Activity/ integration style	manual	meta-data	diff	%diff
understanding SCOP	2.0	2.0	0.0	0
writing wrapper	4.5	2.5	2.0	44%
modifying schema	0.5	0.5	0.0	0
writing mediator	4.0	0.0	4.0	---
modifying meta-data	0.0	1.0	(1.0)	---
total time in days	11.0	6.0	5.0	45%

Results:

Integrating SCoP into warehouse that already contains PDB and SWISS-PROT.

Activity/ integration style	manual	meta-data	diff	%diff
understanding SCOP	2.0	2.0	0.0	0
writing wrapper	4.5	2.5	2.0	44%
modifying schema	0.5	0.5	0.0	0
writing mediator	4.0	0.0	4.0	---
modifying meta-data	0.0	1.0	(1.0)	---
total time in days	11.0	6.0	5.0	45%

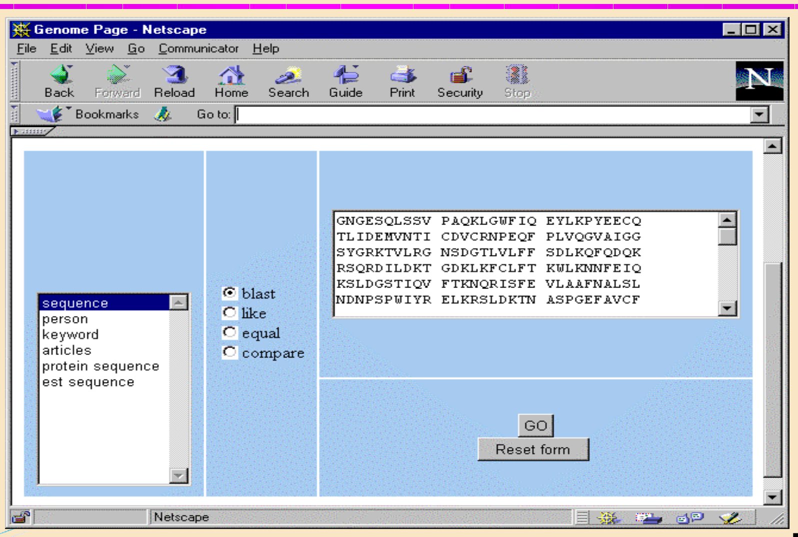
Improving Data Access

Scientists need

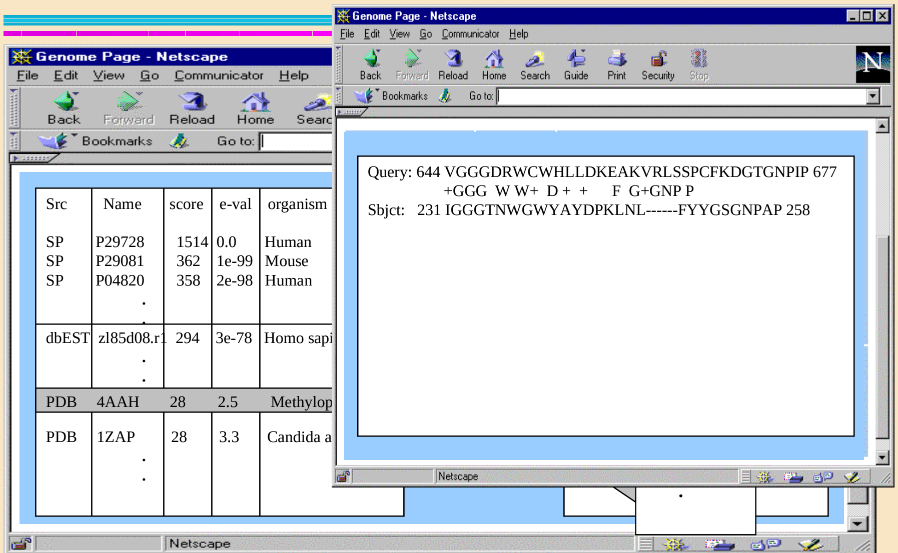
- **Better access to the data**
  - combine data from multiple sources
  - perform complex queries
  - annotate data
- **Better interaction**
  - integrated interface to tools
  - customized notification messages
  - personalized responses to queries

By extending our meta-data representation, we can provide powerful, customizable access to data.

DataFoundry's meta-data driven interface

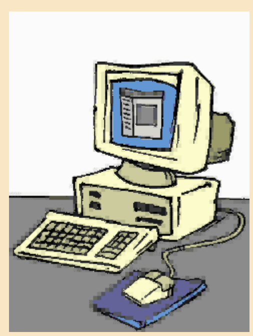


DataFoundry’s meta-data driven interface



Beyond fully integrated data

- There are over 500 genomics data sources available on the web.
- Scientists want as much *relevant* information as possible.
- Integrating data from all of these sources is impossible.



Semantically integrating critical data sources, and providing basic access to others, offers the best possible solution.

Conclusions

- Meta-data provides a way to
  - reduce the cost of integrating new sources
  - reduce the cost of accessing non-integrated sources
  - provide a powerful, and *intuitive*, query mechanism
  - customize the user interface

DataFoundry is building on its meta-data based infrastructure to develop a scalable, flexible, and useable system.

Consider this:

LLNL is always looking for good people.

If you are interested in working at the Lab, either after graduation *or* as an intern please contact me.

Computer Science Challenges at LLNL

Terence Critchlow

Center for Applied Scientific Computing
Lawrence Livermore National Laboratory

critchlow@llnl.gov
www.llnl.gov/casc/people/critchlow

Work performed under the auspices of the U.S. DOE by LLNL under contract No. W-7405-ENG-48. UCRL-JC-134854