

# Complexity Computational Environment: Data Assimilation SERVVO Grid



UC DAVIS



Andrea Donnellan  
ESTO 2004 Annual Meeting

# Objective

Develop the first real-time, large-scale, data assimilation grid implementation for the study of earthquakes that will:

- Assimilate distributed data sources and complex models into a parallel high-performance earthquake simulation and forecasting system
- Simplify data discovery, access, and usage from the scientific user point of view
- Provide capabilities for efficient data mining
- Act as the first step in the Solid Earth Research Virtual Observatory (SERVO)

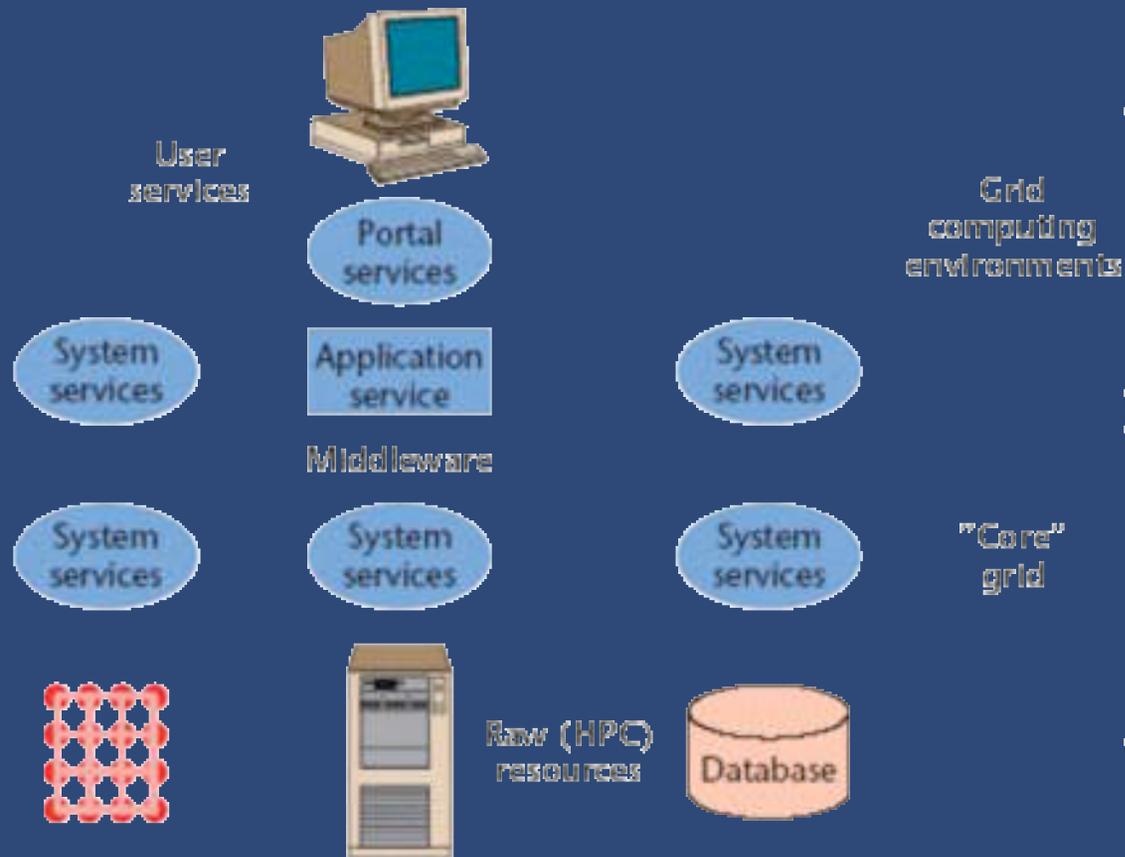
GRAND CHALLENGES  
IN EARTH SYSTEM  
MODELING

ILLUMINATING THE EARTH'S  
INTERIOR THROUGH ADVANCED  
COMPUTING

*Today's computational strategies for modeling the Earth's interior structure and dynamics come from high-performance computing systems in the US and on ones such as the Japanese Earth Simulator. Modeling efforts currently underway focus on problems such as geodynamo and earthquake modeling.*



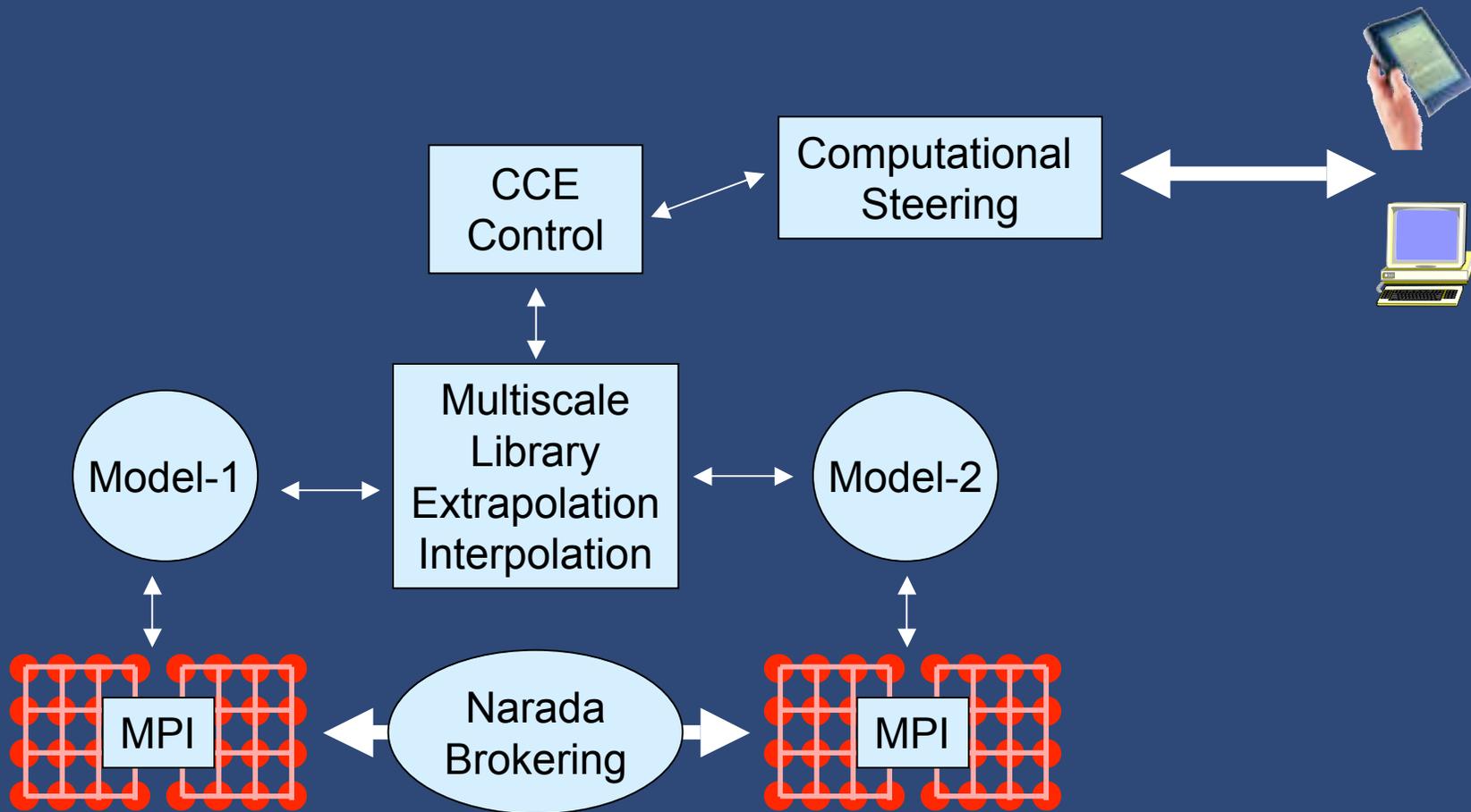
# Three-Tiered Architecture



Three-tiered approach isolates the user from the computational resources

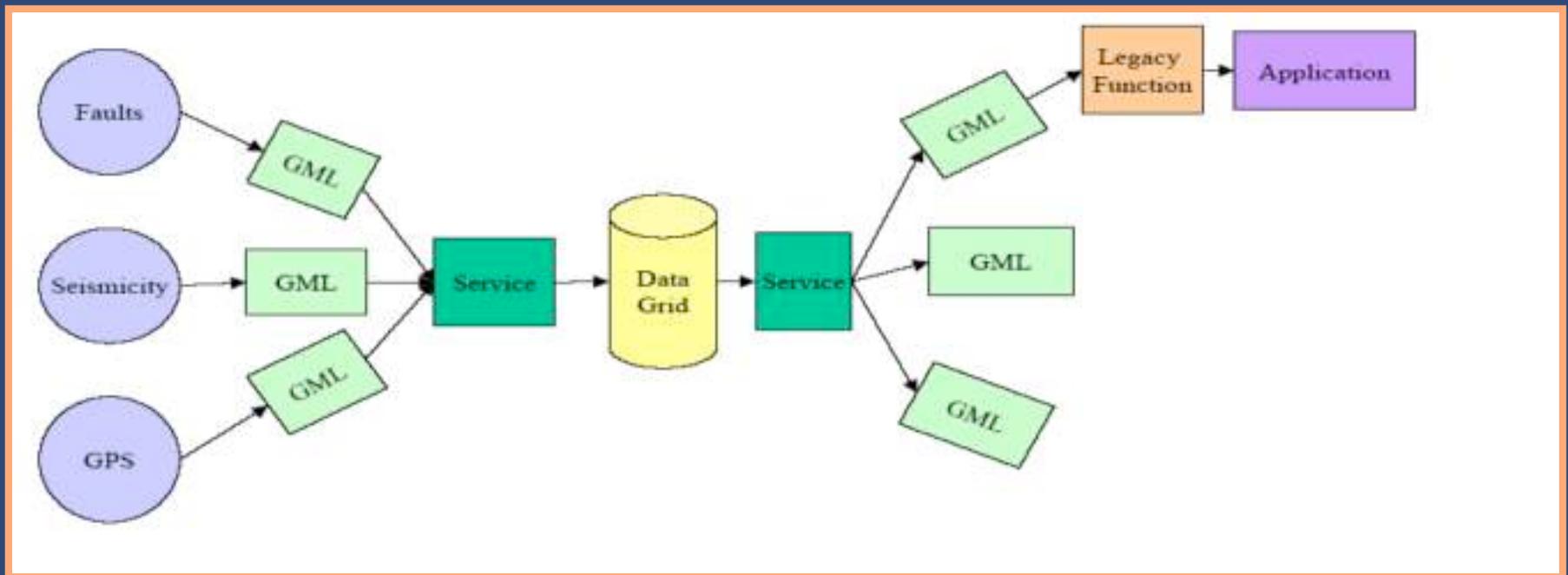
# CCE Architecture in Development

## Complexity Computational Environment



# GML Schemas as Data Models for Services

- Fault and GPS Schemas are based on GML-Feature object.
- Seismicity Schema is based on GML-Observation object.
- Working schema available from <http://grids.ucs.indiana.edu/~gaydin/schemas/>



# Scientific Importance

A solid Earth research environment is required to better understand earthquake processes, which cause an annualized U.S. loss of \$4.4 billion / year. This creates the necessary research infrastructure to efficiently model complex earthquake systems.

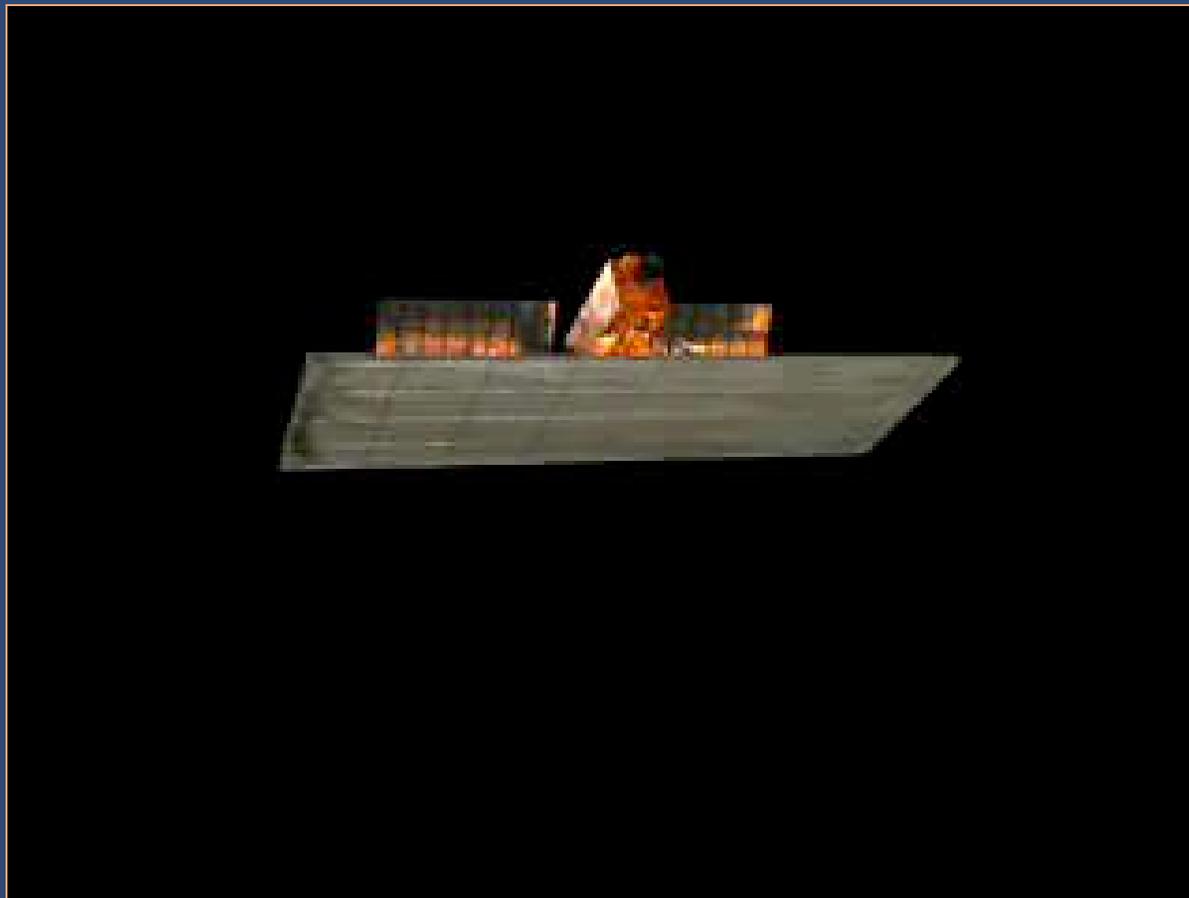
Surface deformation measurements fill a critical gap for understanding earthquake processes, and such diverse data sets must be incorporated into high performance models and analytical tools.

The models and tools elucidate the unobservable or subtle underlying physics of earthquake processes.

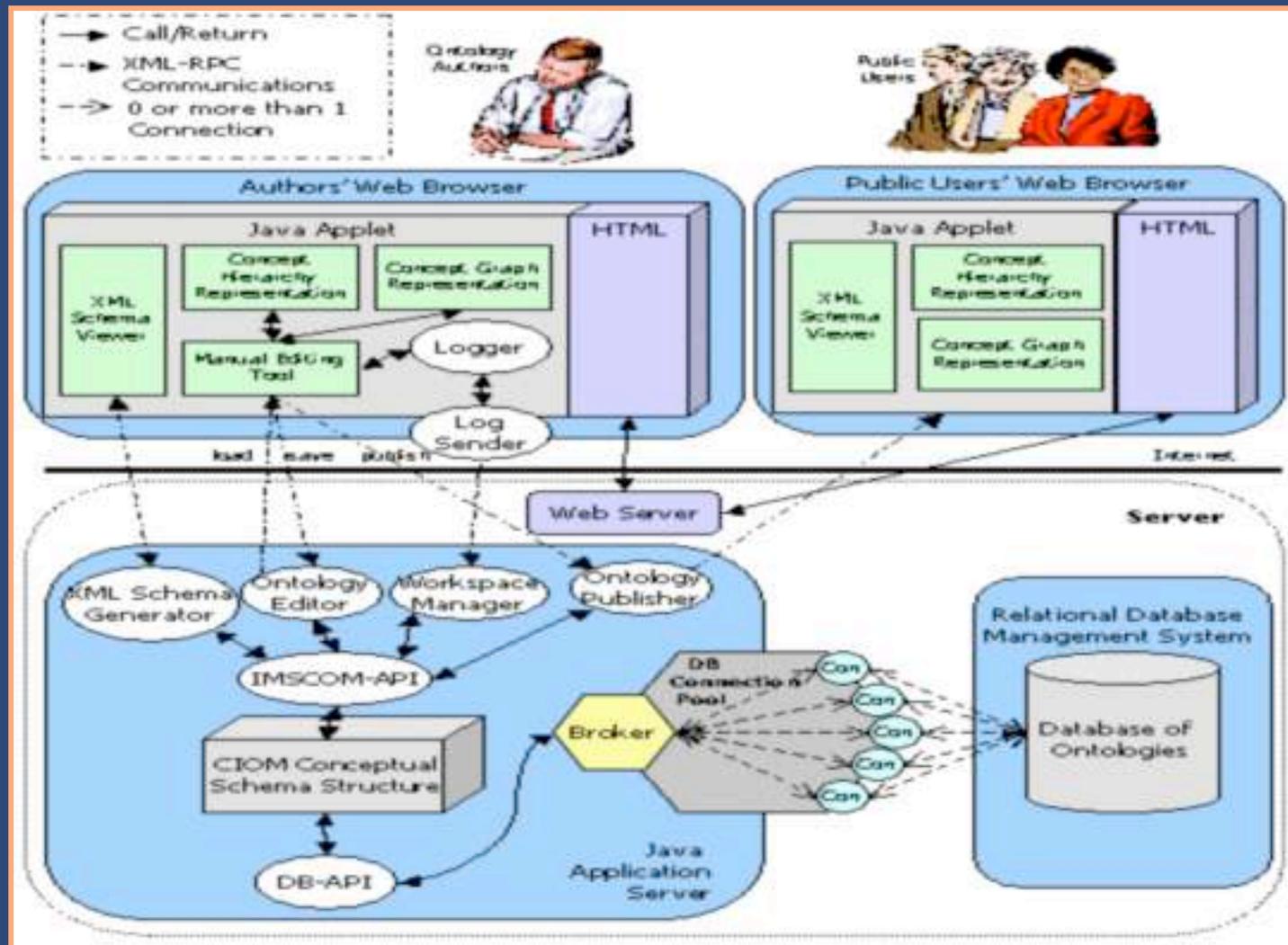


# Earth Science Community

This work will develop the necessary infrastructure for future gravity and InSAR missions



# System Architecture



# Progress to date



- Developed XML schemas to support services and data structures of complexity computational environment (CCE).
- Developed geophysics meta-ontology.
- Completed stand-alone modules supporting database extraction and allowing direct integration into modeling simulation codes.
- Assessed techniques using limited datasets and determined appropriate resolution scaling.

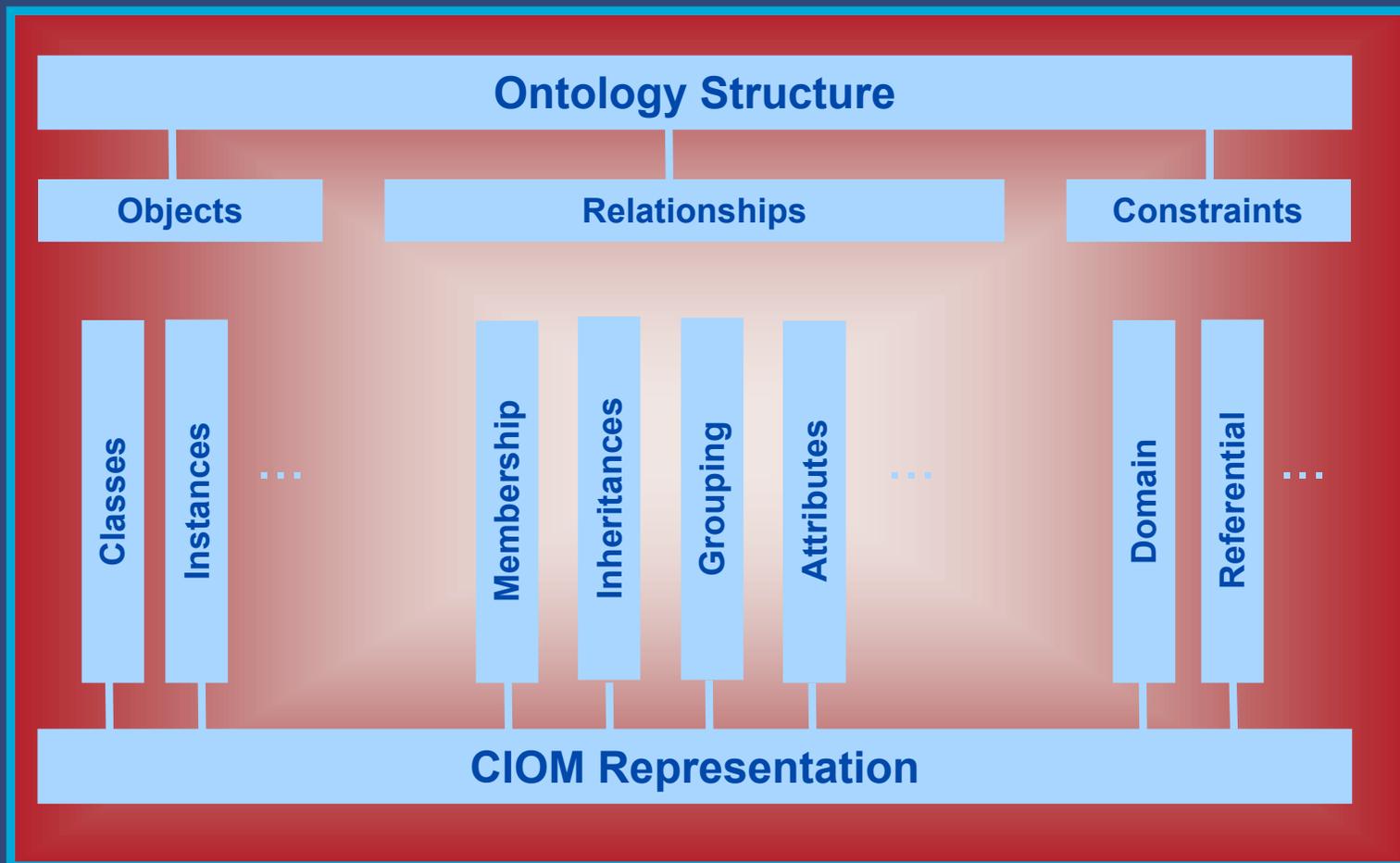
# Ontology Manager



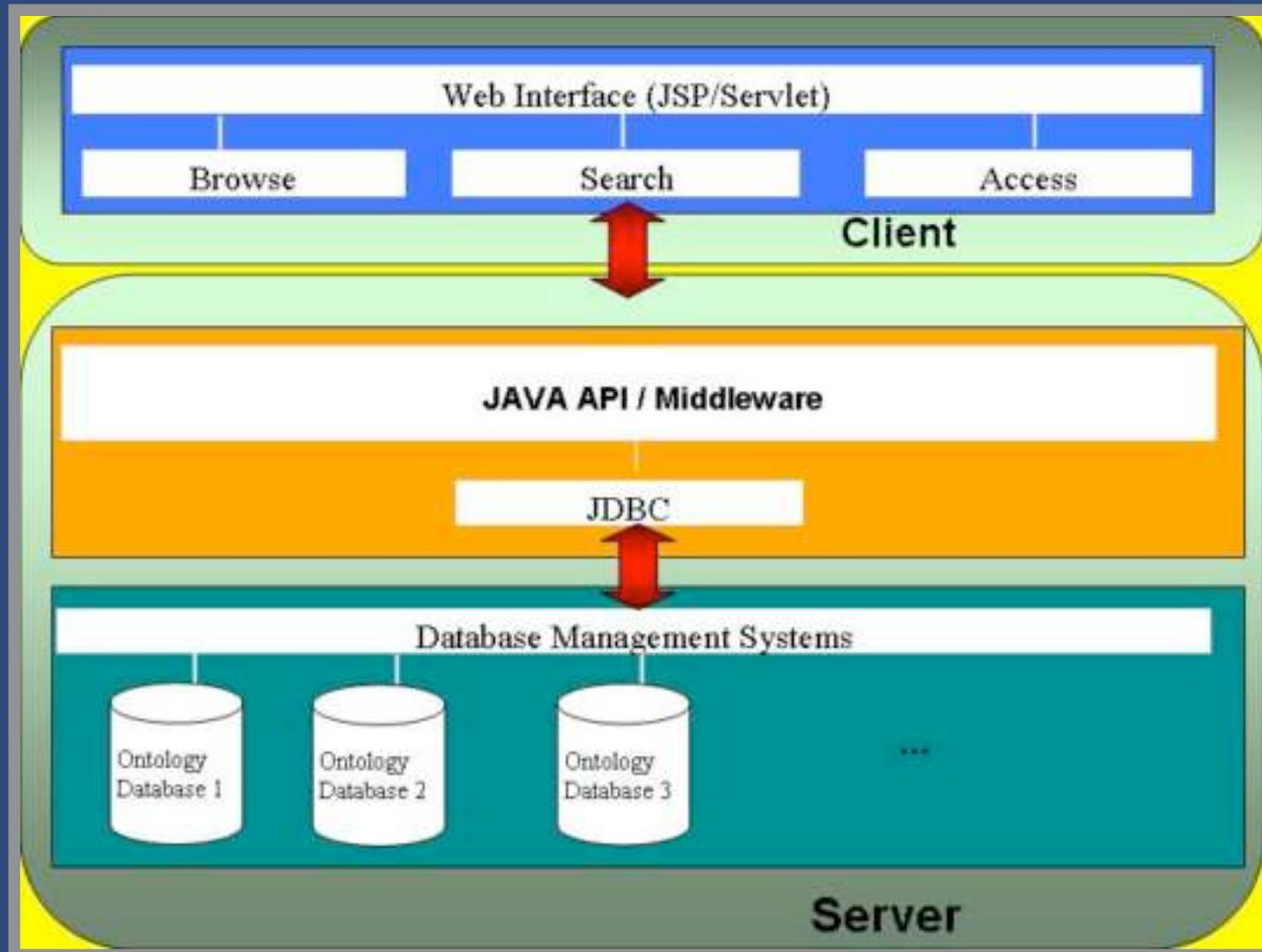
- Facilitate domain ontology creation and update.
- Associate the ontology/ metadata with observational and hypothetical data.
- Learn new concepts, relationships, and patterns among the metadata and data.
- Support user (scientist) data and meta-data discovery/search.
- Provide the base for the semantic wrapping of information sources.

# Ontologies using CIOM

*Structuring Ontologies with CIOM  
(Concept Interrelationship Object Model)*



# Retrieval in the Ontology Manager

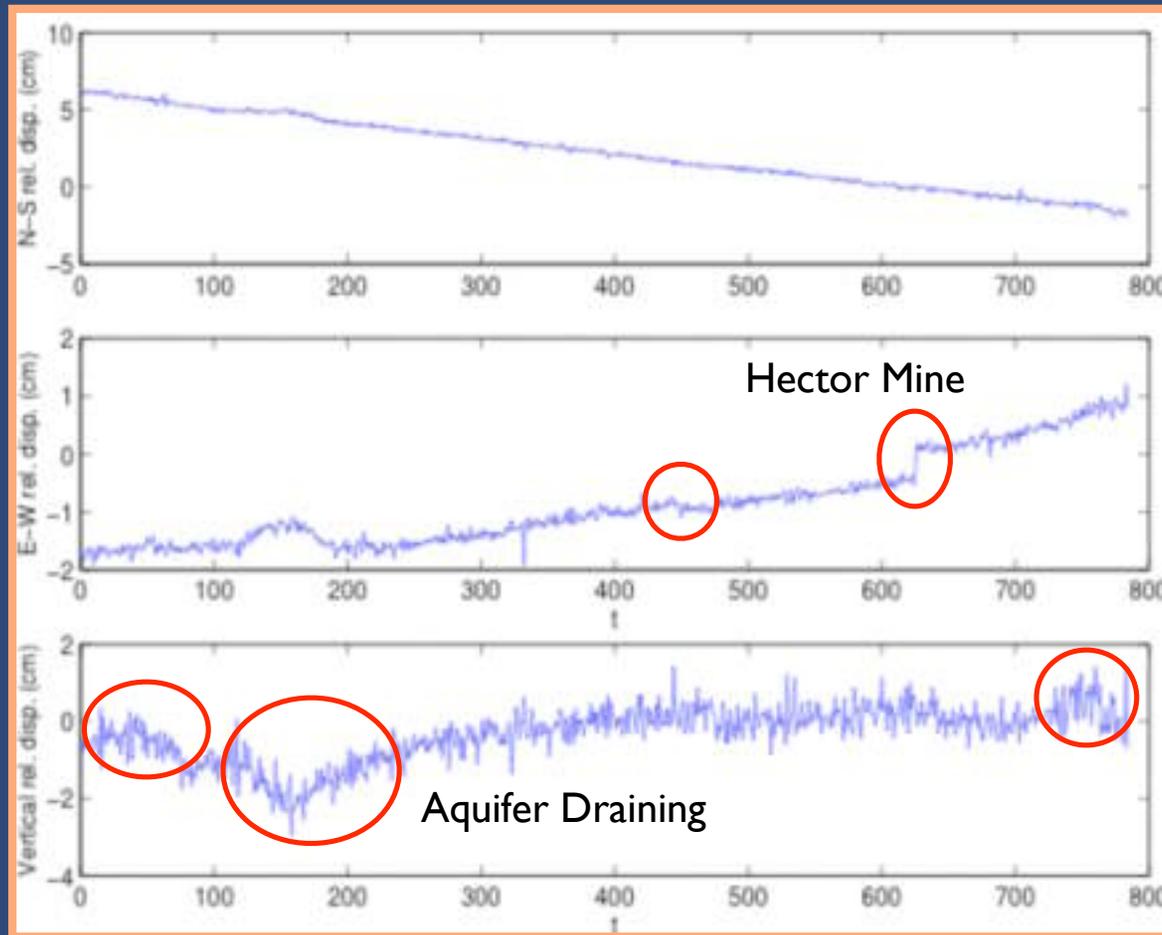


# Time Series Analysis



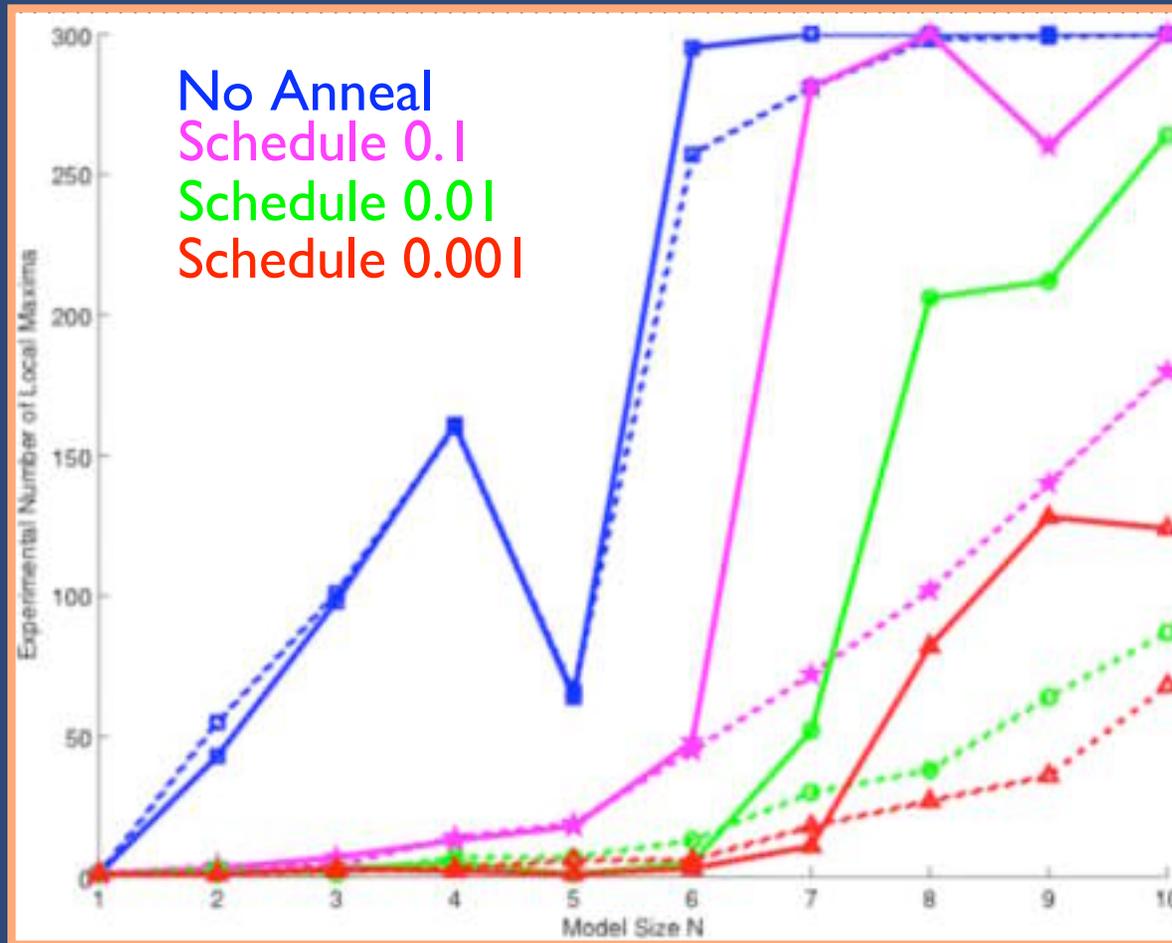
- Regularized Deterministic Annealing Hidden Markov Models (RDAHMM) have been developed to understand geophysical time series data.
- The RDAHMM can be used to train HMMs on geophysics data without domain-based constraints.
- Can be used to identify regional events using GPS displacement data from multiple stations.
- Can identify evidence of long-range fault interactions in seismicity data.
- Can learn signatures of aseismic signals from waveform spectrograms for cataloging.

# Claremont GPS Data



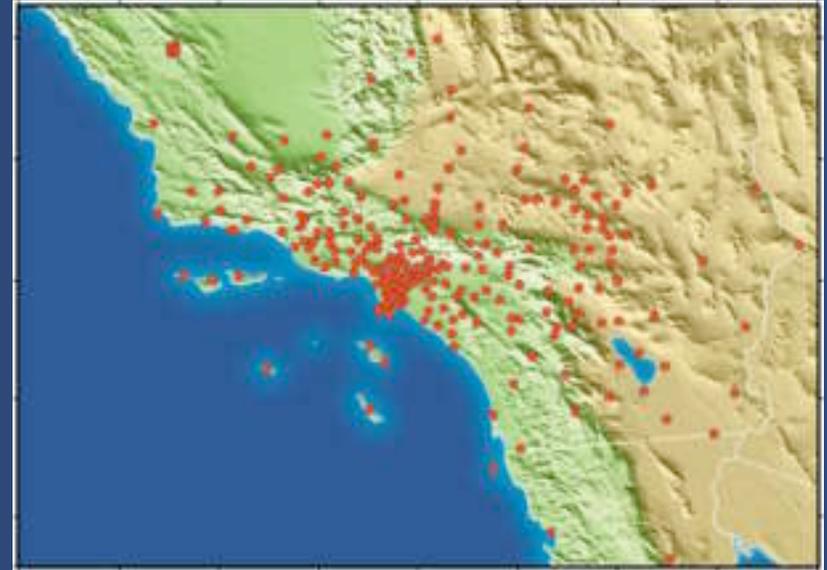
Daily displacement data from Claremont, CA

# (R)DAHMM Results



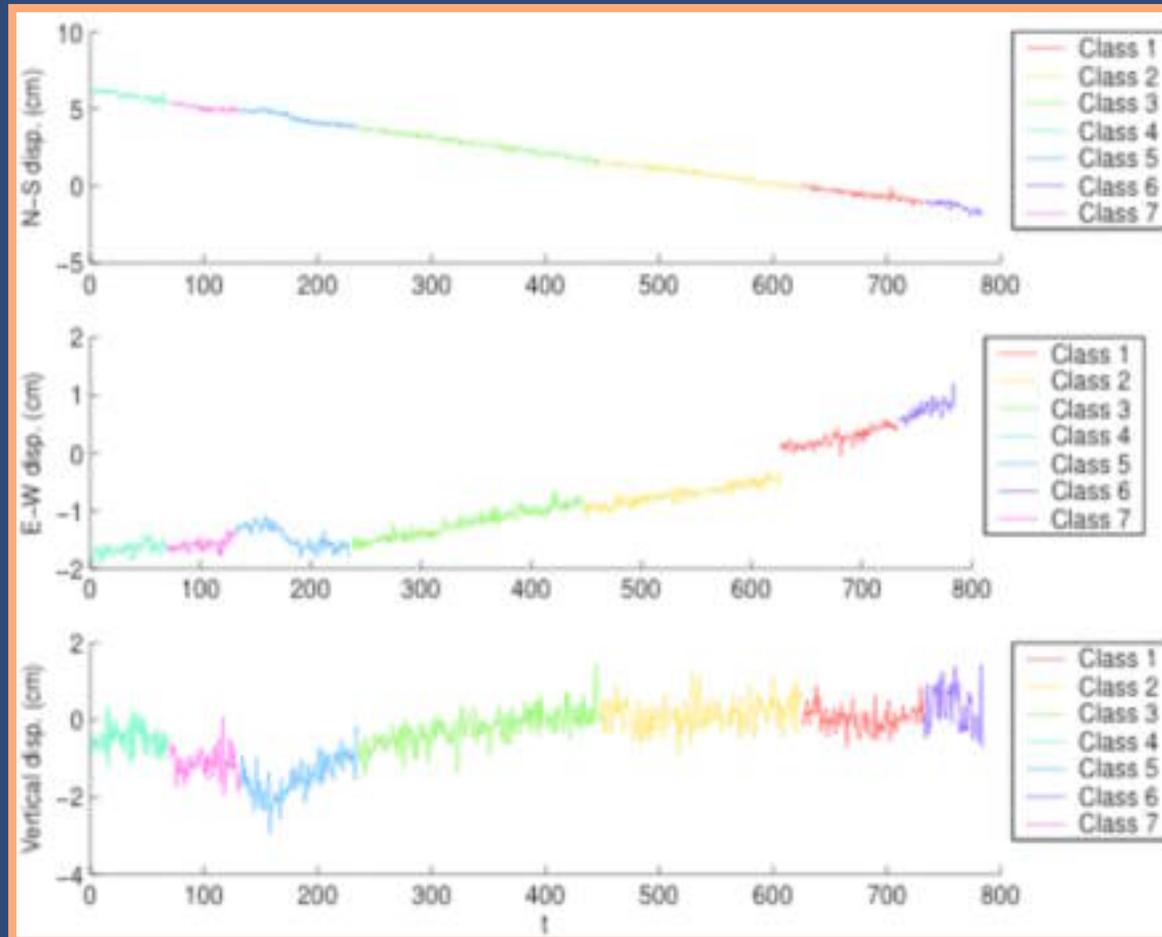
Solid: Regularized; Dashed: Unregularized.

# SCIGN Data



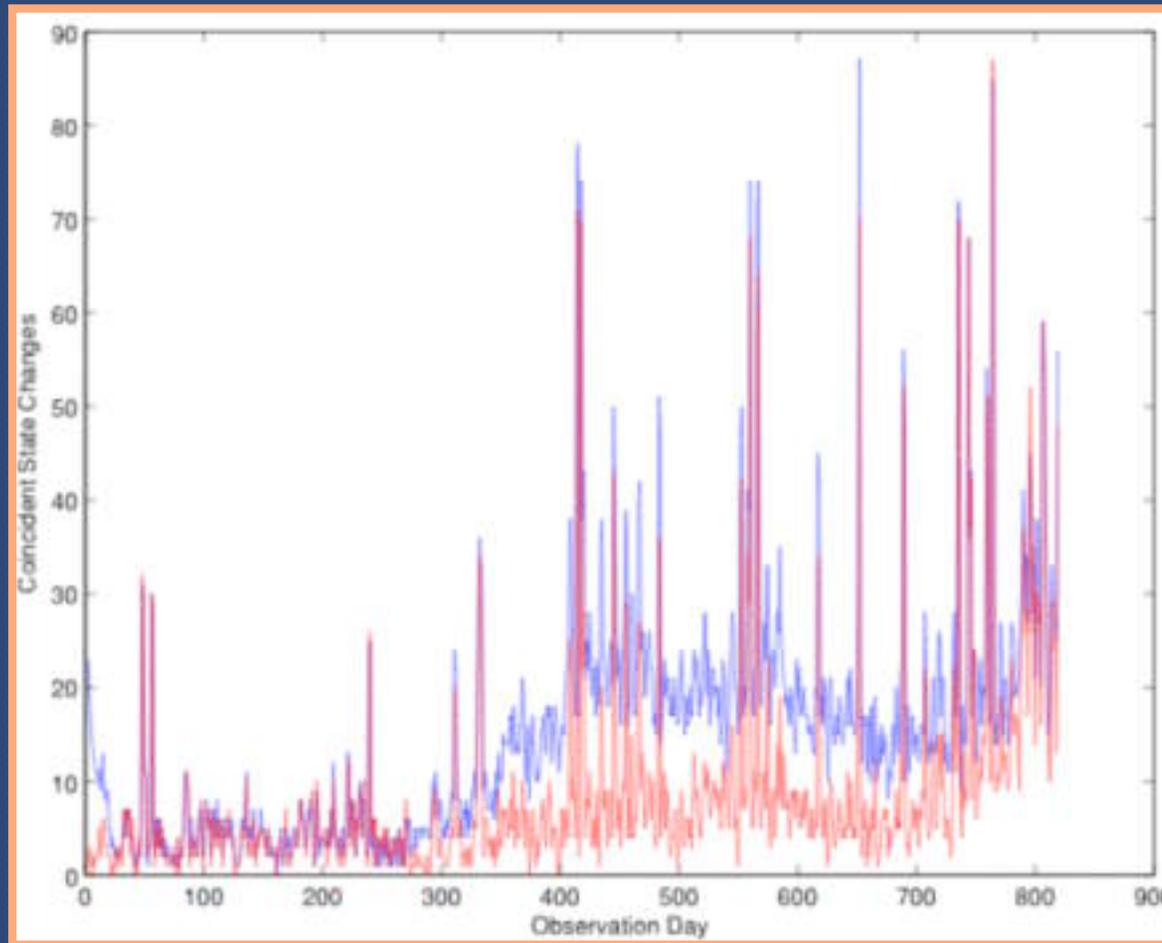
- GPS-integrated measurements of 3-D surface displacement calculated daily.
- Data collected over 800 days starting Jan 1, 1998 from 127 stations.
- Trained 6-state HMMs separately on each data sequence. Classified observations based on trained HMMs.
- Identified correlated state changes across different stations. Correlated state changes indicate regional events.

# Regularized Classification: Claremont



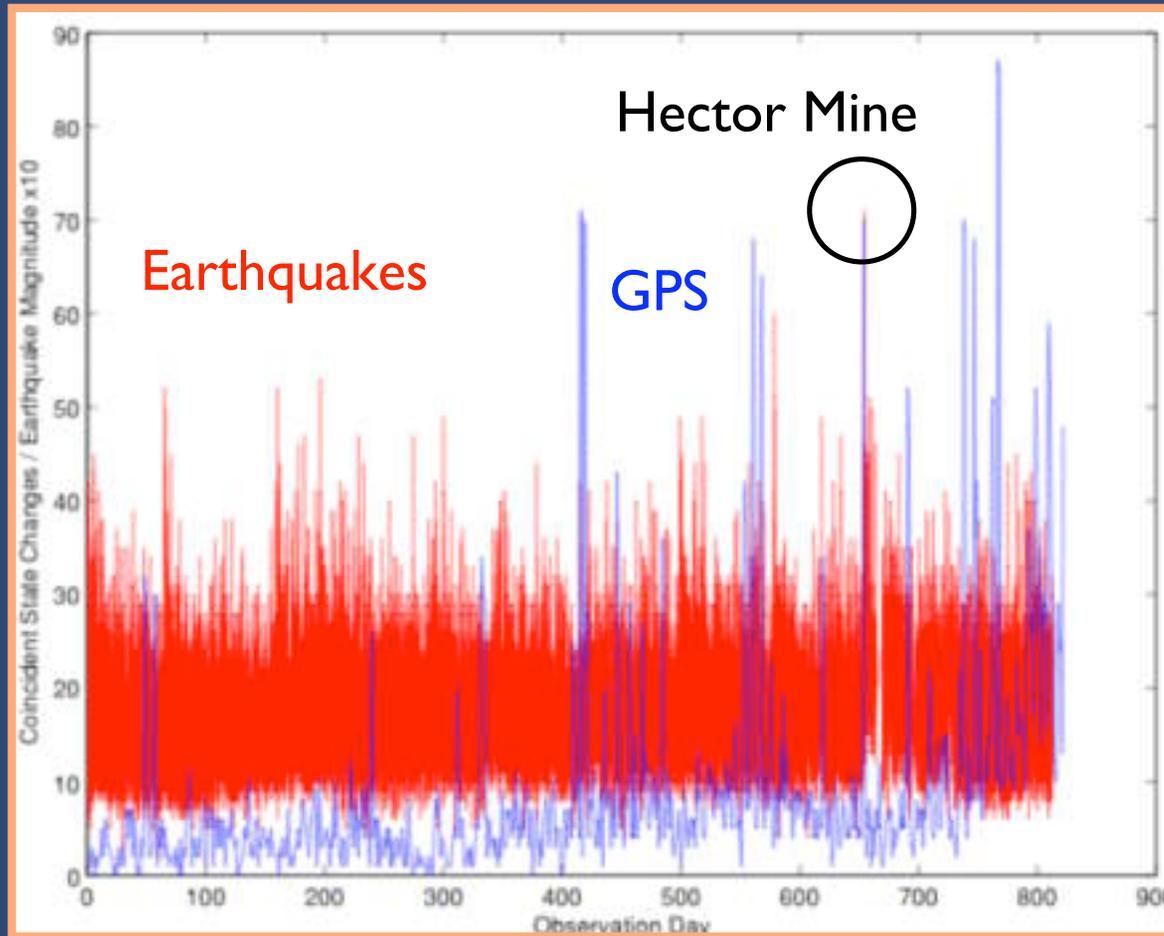
Majority Solution

# SCIGN: EM and RDAEM Comparison



RDAEM (red) has greatly reduced noise

# SCIGN: Seismic Record Comparison

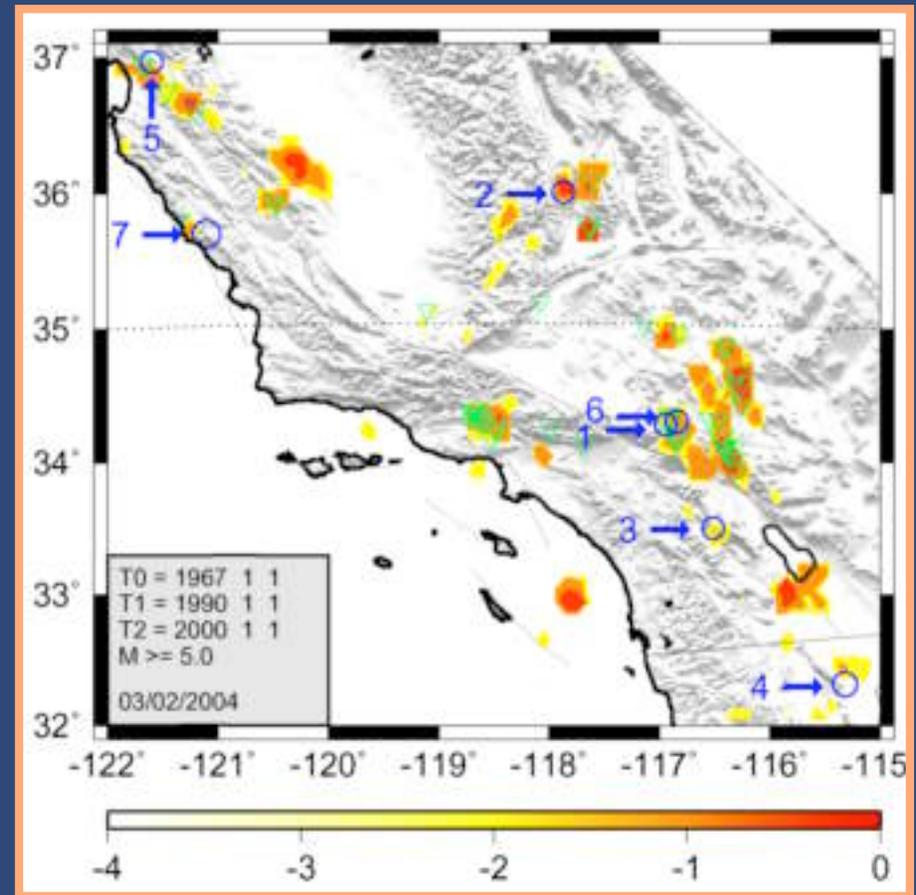


Only the Hector Mine is correlated with a SCIGN regional activity spike - others aseismic activity?

# Data Assimilation Component

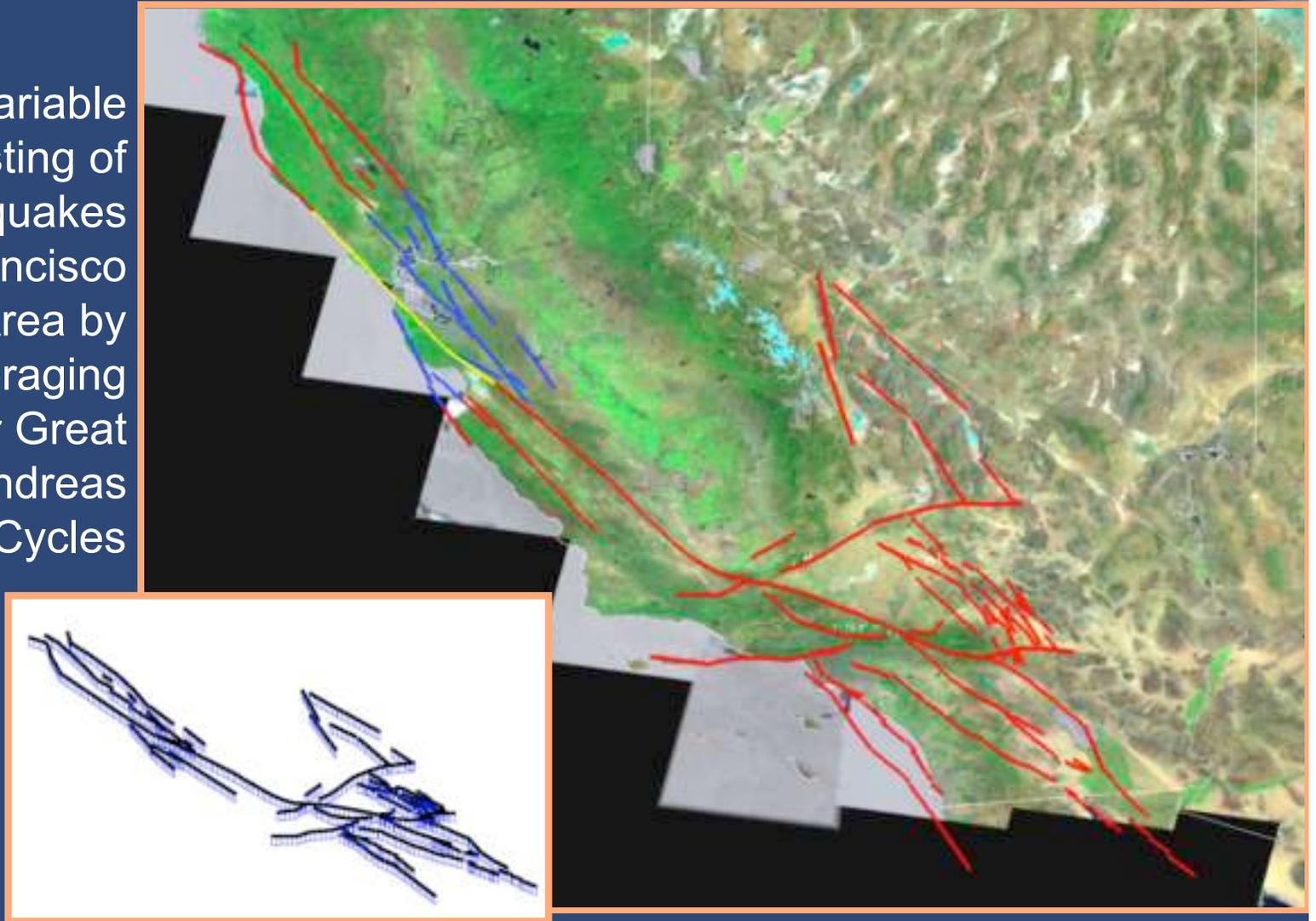
Current data mining approaches for solid earth system data are based on a variety of algorithms, including Pattern Informatics (see figure at right), Karhunen-Loeve methods, Wavelet-based methods, and other somewhat ad-hoc (although successful!) methods

Our plan is to develop a more systematic approach to this problem using model-based ensemble forecasting methods, combined with data assimilation into the simulations



# Example

Time-Variable  
Forecasting of  
Large Earthquakes  
in the San Francisco  
Bay Area by  
Ensemble Averaging  
Over Great  
San Andreas  
Earthquake Cycles

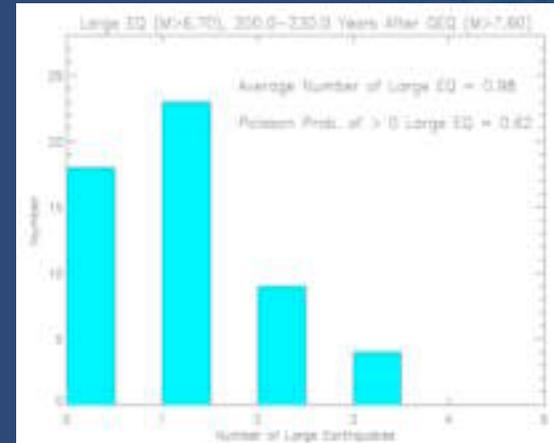


# Unconditional Probabilities from a Model

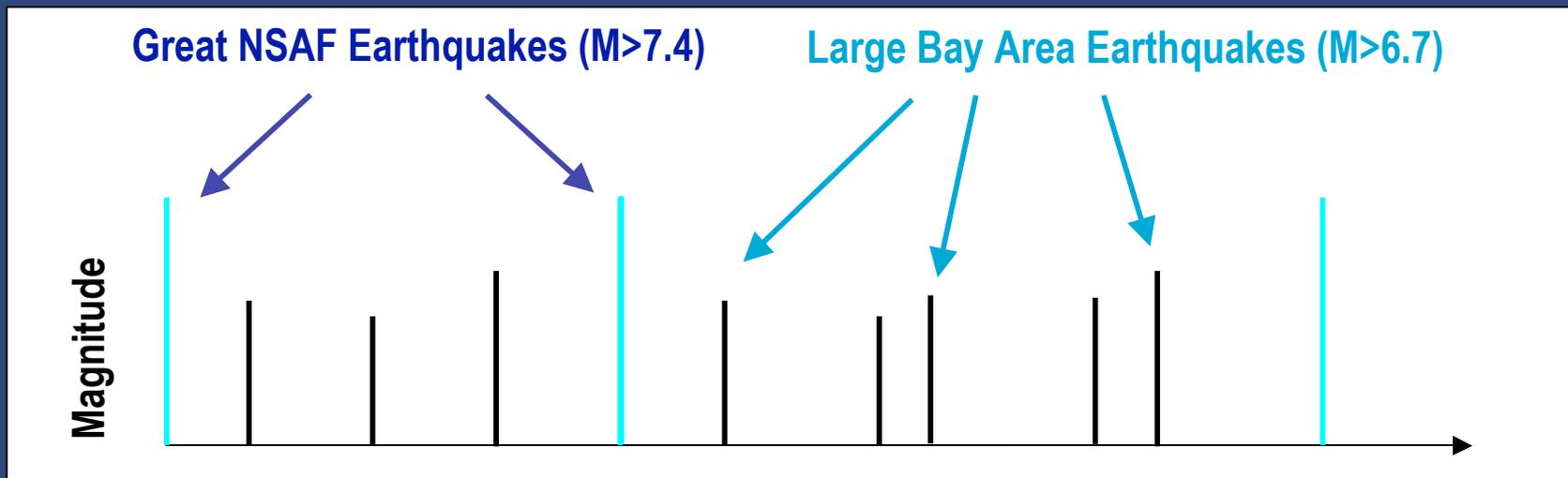


*We compute the histogram of Large Bay Area earthquakes occurring in a 30-year window at a time  $T$  prior to and after a “great” northern San Andreas earthquake via an ensemble method*

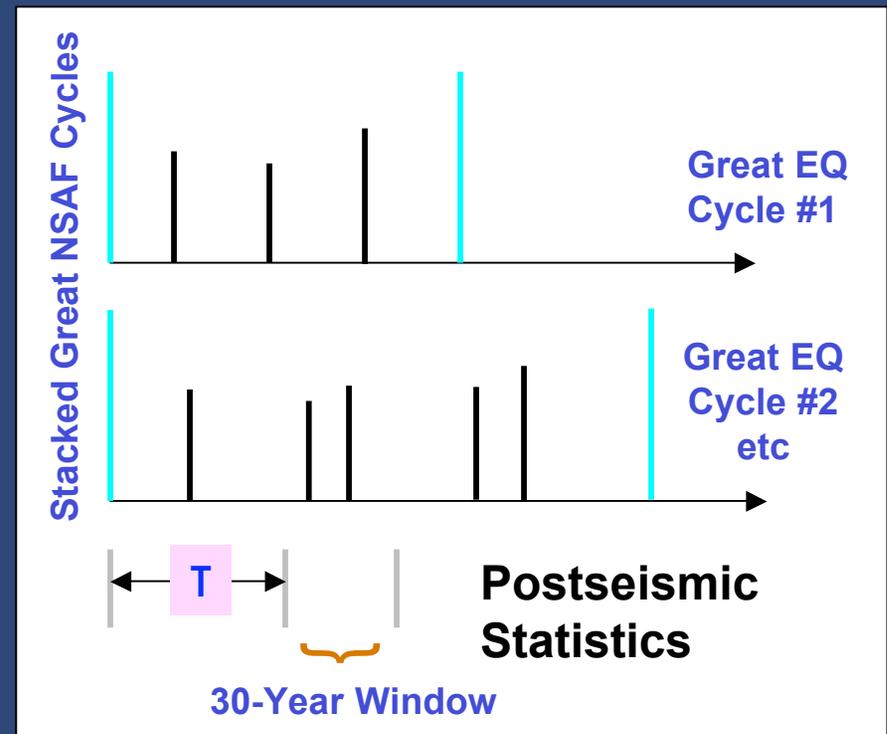
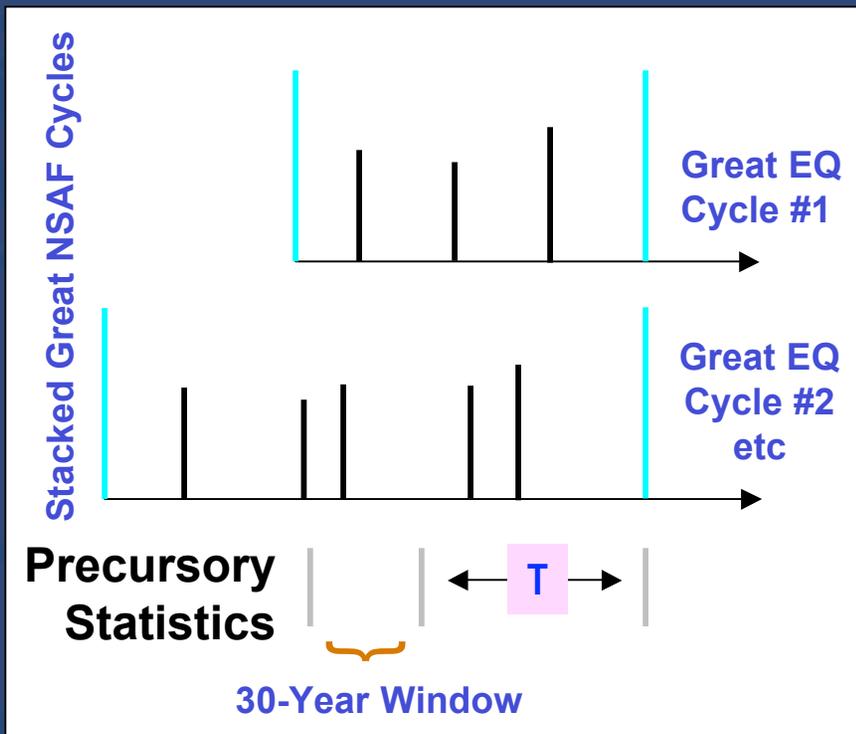
Number of Time Windows



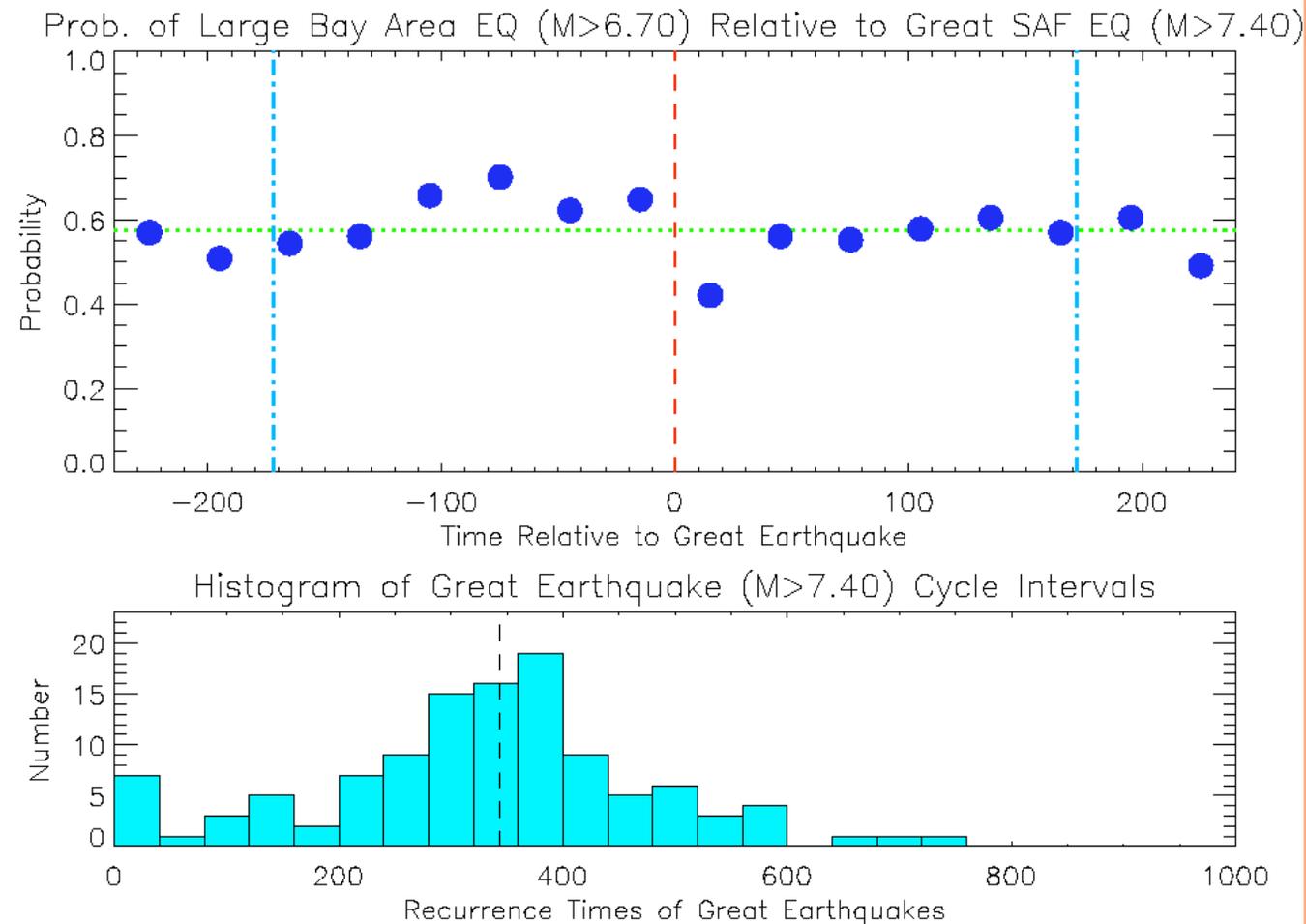
Number of Large SF Bay Area earthquakes in a Time Window



# Unconditional Probabilities from a Model



# Ensemble Forecasts Give Absolute Probabilities



Assume 30-year  
Time-averaging  
Windows.

“Stress Shadow”  
or “Regional  
Stress Recovery”  
effect can be  
seen, as well as  
“Precursory  
Activation”

# Data Assimilation Using Tangent Adjoint Model Compilers



Ensemble forecasting methods are only accurate to the degree that observed data has been assimilated into them, allowing for model updating and model steering.

- Basic Method: Use numerical differentiators such as ADIFOR, which is a *Tangent Adjoint Model Compiler*, for model steering and data assimilation.
- ADIFOR is based on the idea that model follows an evolutionary path through state space, so observations can be used to periodically adjust model parameters, to bring the model path as close as possible to the path represented by the observed system.

# Data Assimilation Using Tangent Adjoint Model Compilers

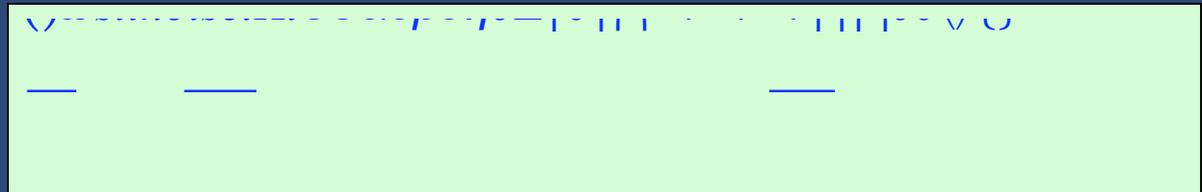


- Adjoint methods are based on the initial definition of a cost function, or fitness function, which defines the misfit between the model path through time and the path represented by the observed data.
- Model tuning occurs by computation of a gradient vector in state space, which specifies the direction in which corrections to the model must be applied in order to return the model evolution towards the path defined by the observations.
- Whereas the forward model specifies the sensitivity of a change in model parameters on data evolution, the adjoint model acts in the reverse direction, specifying the sensitivity of a change in data on the model parameters.
- An equivalent implementation for linear or quasi-linear systems uses Kalman filters, which can be used in a continuous mode to compute corrections to model parameters as new data is acquired.

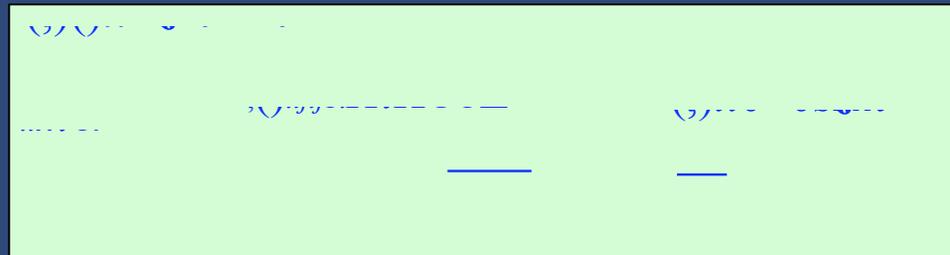
# Example: Steps for Data Assimilation in Virtual California



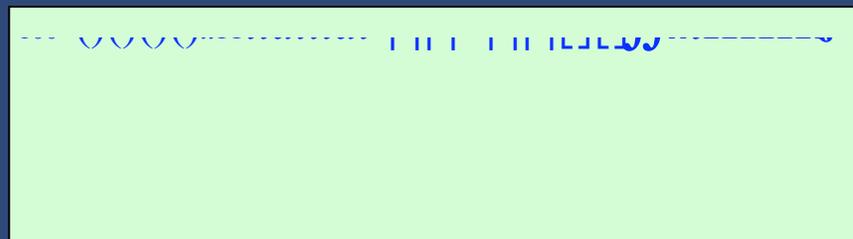
Basic model dynamics



Define linear data assimilation equations

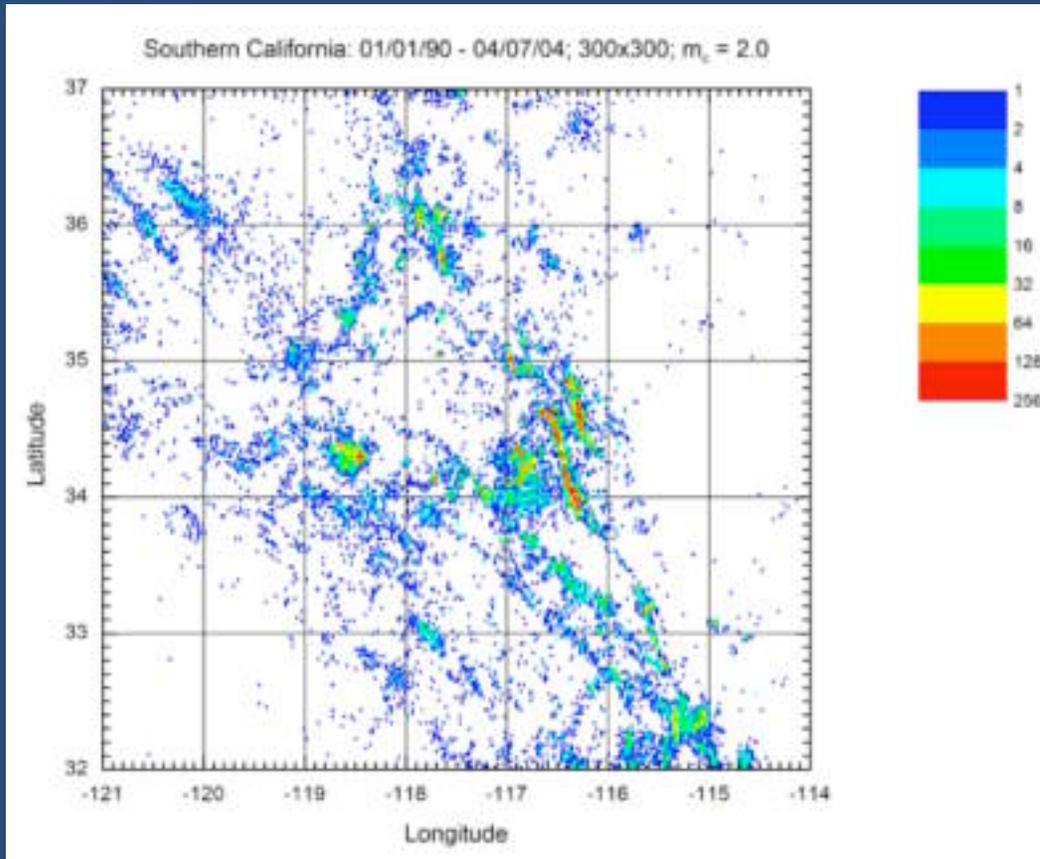


Estimate model parameters using Kalman filter



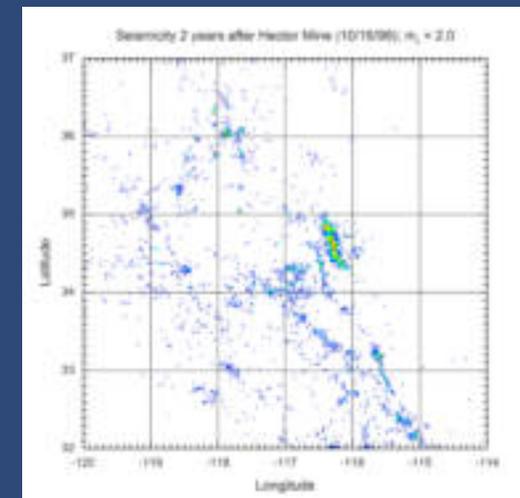
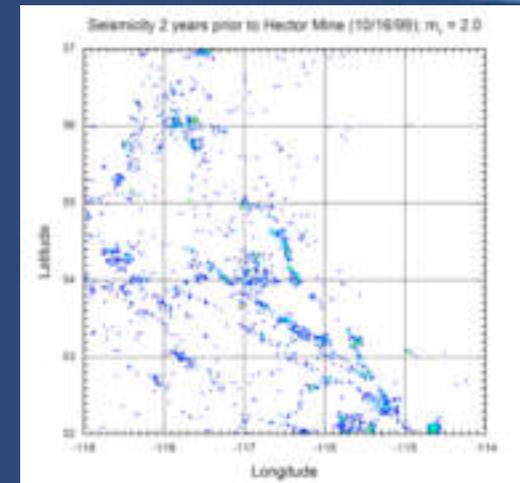
# Example: Assimilating Coarse-Grained Seismicity Data into Simple Model

Prototype for Complexity Computational Environment SERVGrid



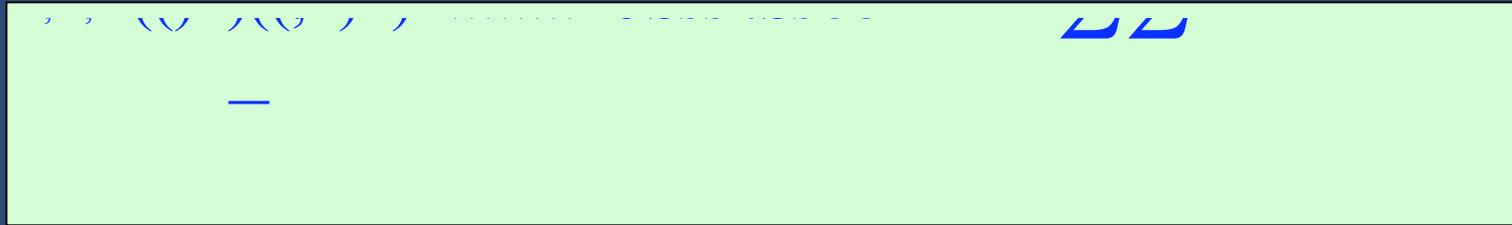
## Steps:

- Coarse grain a spatial region with a spatial grid
- Using models, analyze the earthquake activity time series in each grid box with the idea of using changing space-time seismicity patterns to forecast future activity of large earthquakes

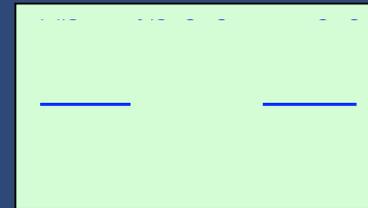


Activity associated with  
October 19, 1999 Hector  
Mine earthquake

# Potts Model for Coarse-Grained Seismicity Forecasting



Maximum Entropy -  
Minimum Free Energy  
Dynamics



Here  $s_k(t)$  can be in any of the states  $s_k(t) = 1, \dots, S$  at time  $t$ ,  $\delta(s_k, s_k')$  is the Dirac delta, and the field  $h_k$  favors box  $k$  to be in the low energy state  $s_k(t) = 1$ . This conceptually simple model is a more general case of the *Ising model* of magnetic systems, in which case  $S = 2$ .

In our case for example, the state variable  $s_k(t)$  could be similarly chosen to represent earthquake seismicity, GPS displacements or velocities, or InSAR fringe values.

# Next Step: Prototyping our Approach



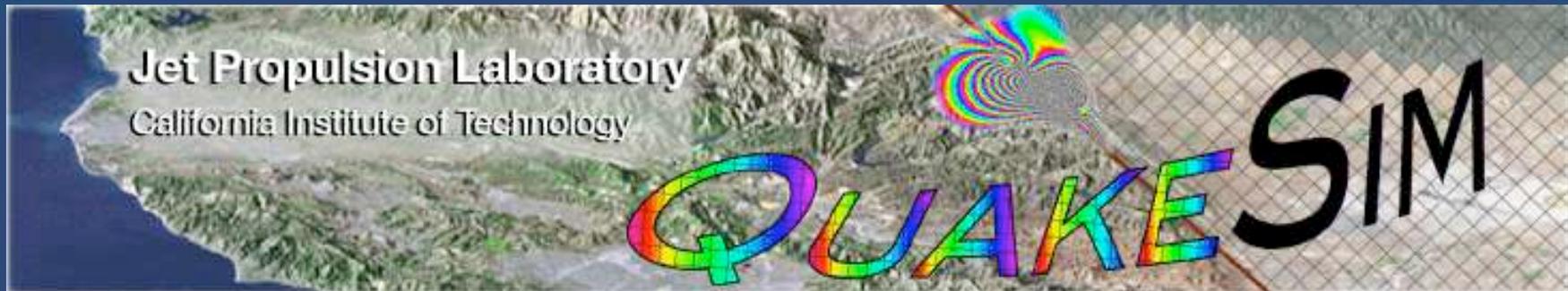
1. Begin with  $q = 2$  state Potts model {same as Ising model  $H(J,h)$  }
2. Erect a simple square lattice  $\mathbf{Z}$  in the plane
3. Adopt a simple dynamics based on maximum entropy (minimum free energy)
4. Pick a set of parameters  $(J,h)$  that give reasonable time series on each grid point, and produce a synthetic data set, both with and without noise
5. Use Kalman filter and back-propagation techniques to optimally fit the time series via model steering, and determine how closely the known values of  $J,h$  can be recovered.
6. Repeat with larger lattices, and in the presence of noise to determine robustness of algorithm, and how it scales with lattice size.

# Cooperation with Others



- Collaboration with international partners from Australia, Japan, and China has led to development of iSERVO (international Solid Earth Research Virtual Observatory). Prototype is under development.
  - See Mora et. al., *The International Solid Earth Research Virtual Observatory Institute Seed Project*, Fall AGU 2003.
- Working with the Open GIS Consortium to develop consistent standards under GML.

# Cooperation with Others

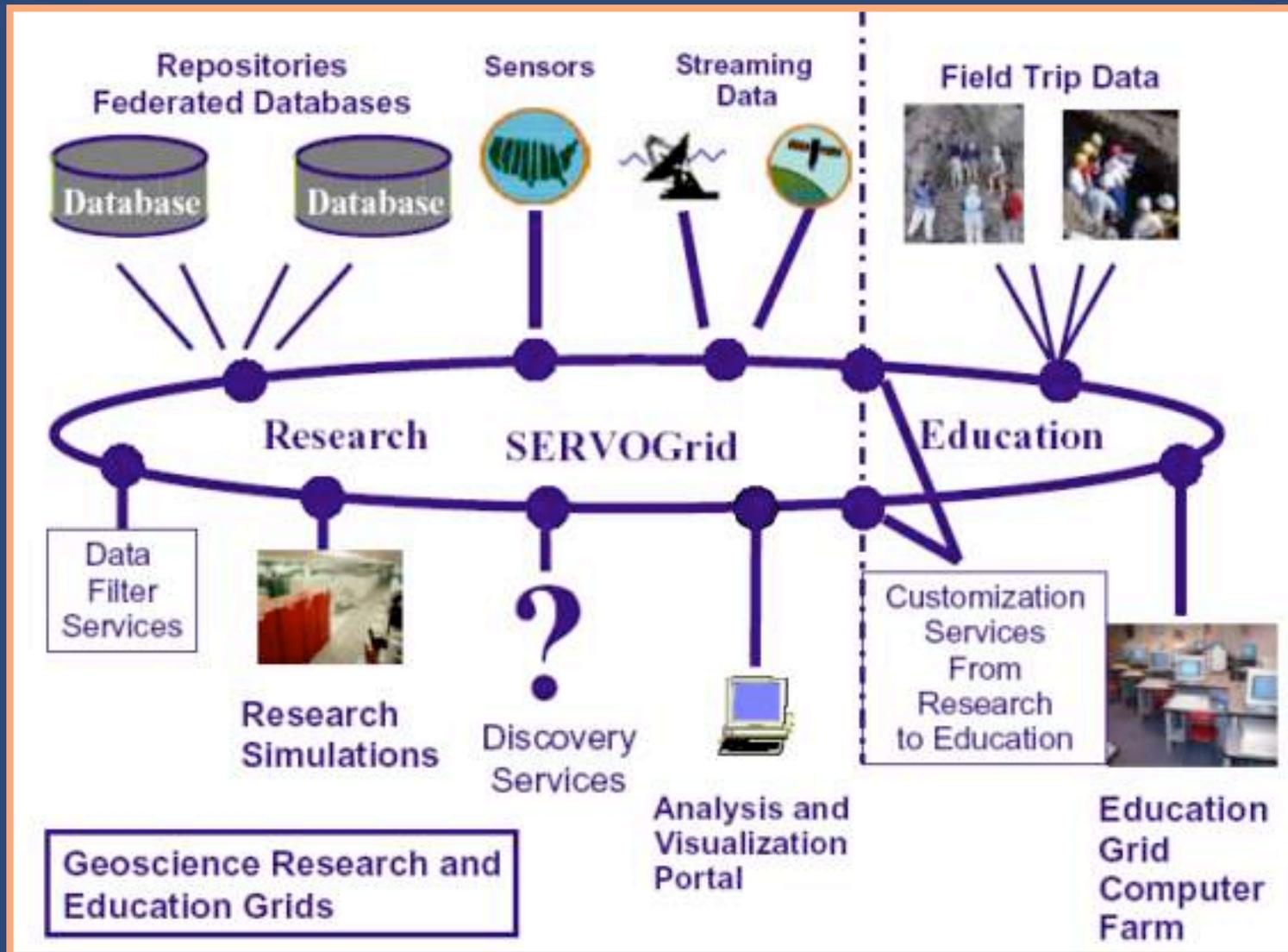


- Working with “active tectonics” (QuakeSim) CT project team members
- NSF Center for Computational Infrastructure for Geodynamics
  - Actively collaborating to develop standards and compatible approaches for planned center



- Southern California Earthquake Center Information Technology Research Project on computational infrastructure. They are primarily focused on earthquake wave propagation (which is complementary to this work).

# SERVOGrid links to Education



# Future Outlook



- Complete exploring coupling methodologies to guide development of cross-scale tools
- Complete detailed design of CCE architecture
- Data assimilation and coarse graining interfaces defined in Web service framework
- Complete prototype data assimilation on SERVO Grid with one application and one coarse graining approach
- Develop meta-query mediator and translator