Scientific Data Analysis via Statistical Learning

Raquel Romano

romano at hpcrd dot lbl dot gov

November 2006

Abstract

The emerging field of scientific data mining addresses the need for scientists to search for, quantify, and visualize information from massive data sets generated by large scale observations and simulations. Statistical machine learning algorithms have enormous potential to provide data-driven solutions to fundamental scientific questions. We present results from ongoing projects in astrophysics and climate modeling at LBNL, including the search for Type Ia supernovae from astronomical observations, and the analysis of hurricanes and tropical storms in climate simulations.

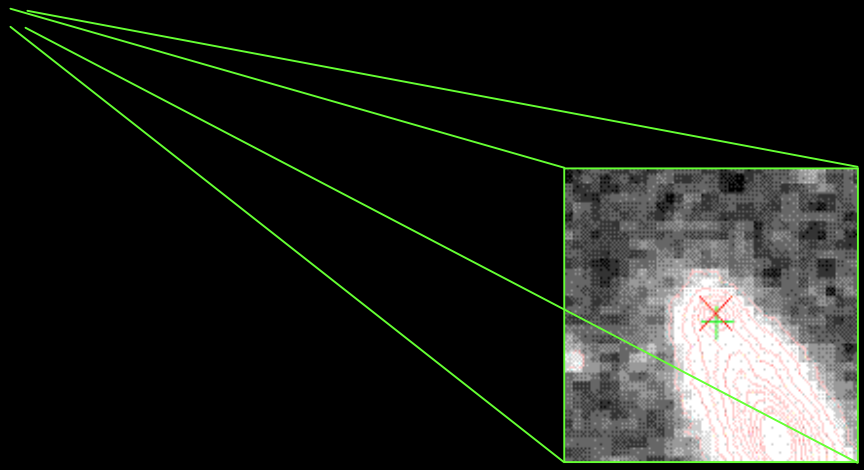
Supervised Learning for Supernova Recognition

Supervised learning techniques, also known as classification methods, offer powerful, adaptive methods for the automatic detection and recognition of rare objects from large scale digital sky surveys. We have integrated state-of-the-art classifiers (Support Vector Machines, boosted decision trees) into the nightly pipeline of the Nearby Supernova Factory at Lawrence Berkeley Laboratory, to quickly and automatically rank more than 500,000 subimages found nightly in order of confidence that they contain supernovae.

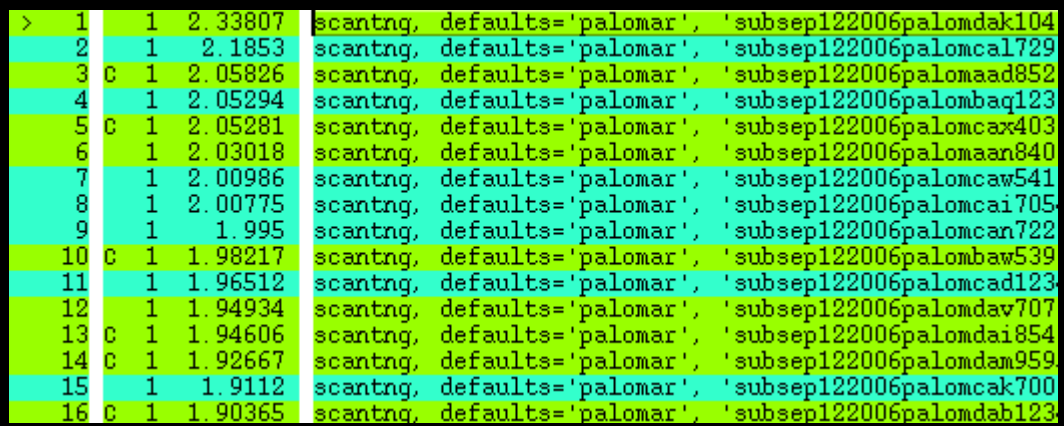
large sets of astronomical imagery

potential supernova on galaxy

measured features (geometry, brightness, signal-to-noise, …)

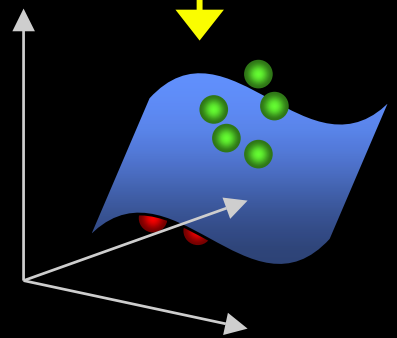


$$\begin{bmatrix} 11.88 \\ 7.847 \\ -0.117 \\ 3.269 \\ 3.22 \\ 10.36 \\ 9.692 \\ -0.1523 \\ 0.7023 \\ \vdots \\ 0.008165 \\ 0.05337 \end{bmatrix}$$



>	1		1	2.33807	scantng, defaults='palomar', 'subsep122006palomdak104
	2		1	2.1853	scantng, defaults='palomar', 'subsep122006palomcal729
	3	C	1	2.05826	scantng, defaults='palomar', 'subsep122006palomaad852
	4		1	2.05294	scantng, defaults='palomar', 'subsep122006palombaq123
	5	C	1	2.05281	scantng, defaults='palomar', 'subsep122006palomcax403
	6		1	2.03018	scantng, defaults='palomar', 'subsep122006palomaan840
	7		1	2.00986	scantng, defaults='palomar', 'subsep122006palomcaw541
	8		1	2.00775	scantng, defaults='palomar', 'subsep122006palomcai705
	9		1	1.995	scantng, defaults='palomar', 'subsep122006palomcan722
	10	C	1	1.98217	scantng, defaults='palomar', 'subsep122006palombaw539
	11		1	1.96512	scantng, defaults='palomar', 'subsep122006palomcad123
	12		1	1.94934	scantng, defaults='palomar', 'subsep122006palomdav707
	13	C	1	1.94606	scantng, defaults='palomar', 'subsep122006palomdai854
	14	C	1	1.92667	scantng, defaults='palomar', 'subsep122006palomdam959
	15		1	1.9112	scantng, defaults='palomar', 'subsep122006palomcak700
	16	C	1	1.90365	scantng, defaults='palomar', 'subsep122006palomdab123

new supernova candidates ranked by confidence

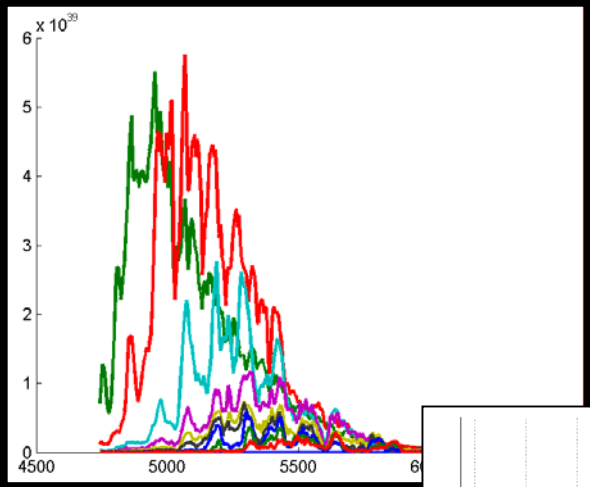


learned decision boundary identifies supernovae

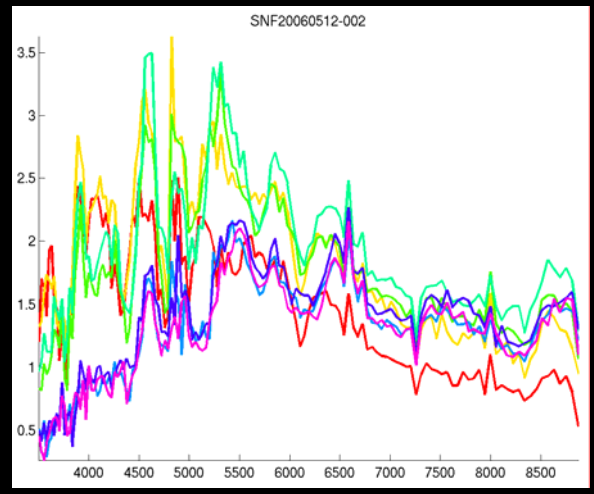
Dimensionality Reduction of Supernova Spectra

Dimensionality reduction techniques decompose data into a small number of basis elements that capture useful attributes of the original signals. Spectra measured from observed supernovae are projected onto a lower dimensional subspace or manifold whose dimensions are associated with physical properties such as luminosity and phase. These decompositions are useful for validating synthetic spectra generated from simulations, classifying observed spectra by luminosity and phase, and discovering new classes of supernovae.

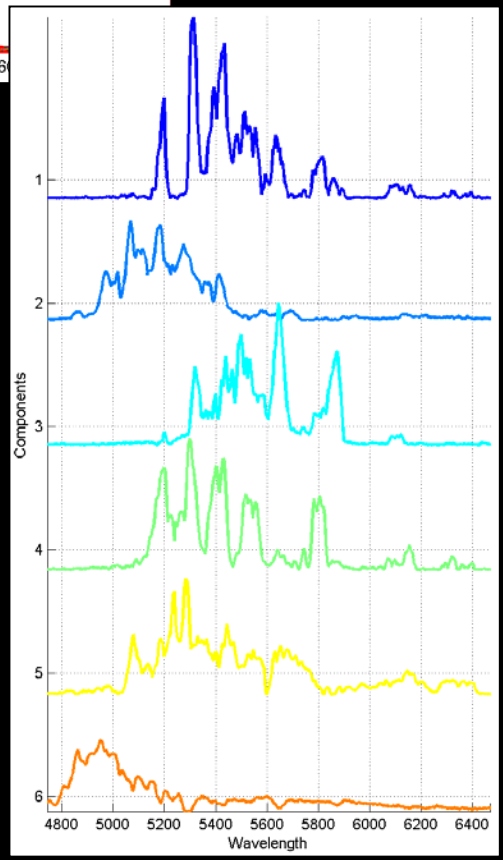
synthetic time-varying spectra



observed time-varying spectra

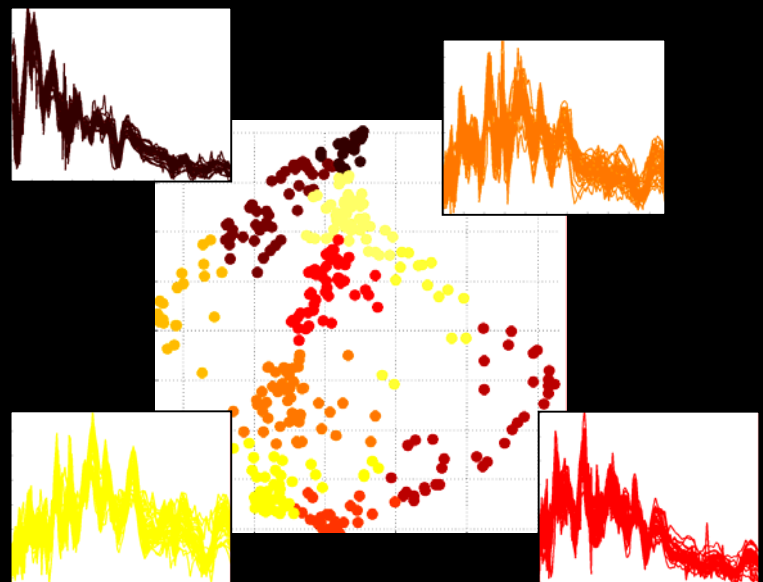


low-dimensional projection encodes parameters such as phase or luminosity



basis spectra

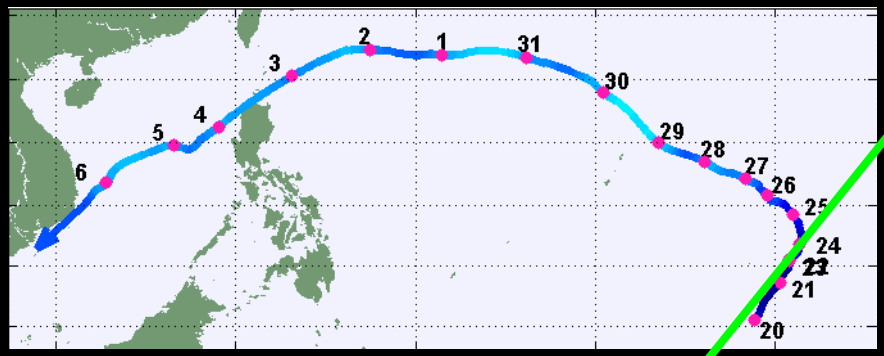
clustering and matching in low-dimensional space

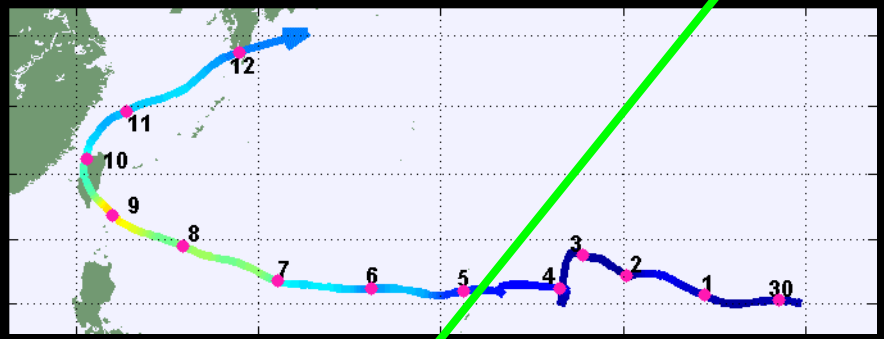


Spatiotemporal Analysis of Climate Data

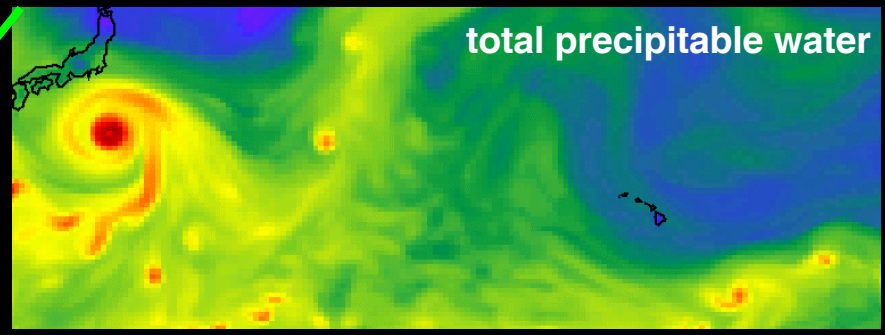
Statistical modeling of high-dimensional, time-varying, non-stationary time series enable the discovery of temporal, spatial, and multivariate statistical dependencies in large-scale climate simulations. Using latent variable models of tropical storm trajectories, we can infer the statistical relationships between space-time-varying atmospheric variables along a storm trajectory and the likelihood of the storm evolving into an intense hurricane. Hierarchical, stochastic models can predict the influence of global-scale, long-term climate patterns on such short-term, local events.

collection of tropical storm tracks





multivariate time series for each track



large-scale, high-resolution simulations yield many examples for probabilistic modeling