Data mining opportunities

Charles Elkan

University of California,
San Diego

The charge given to us

* “Highlight methods /models /techniques which may be useful in the analysis of climate data and understanding climate change.”

What can computer science offer climate science?

Virtual and Artificial, but 58,000 Want Course

By JOHN MARKOFF
Published: August 15, 2011

PALO ALTO, Calif. — A free online course at Stanford University on artificial intelligence, to be taught this fall by two leading experts from Silicon Valley, has attracted more than 58,000 students around the globe — a class nearly four times the size of Stanford's entire student body.

The two additional courses will be an introductory course on database software, taught by Jennifer Widom, chairwoman of the computer science department, and an introduction to machine learning, taught by Andrew Ng.

* Students
* New thinking

What can climate science offer computer science?

05/20/2005 - Updated 11:09 AM ET

Predicting hurricanes

Each December, Colorado State University's tropical storm research team forecasts the coming year's Atlantic hurricane season, which is June 1 through Nov. 30.

	Named storms (Winds of at least 39 mph)		Hurricanes (Winds of at least 74 mph)		Major hurricanes (111 mph or more winds)	
	Forecast	**Actual**	**Forecast**	**Actual**	**Forecast**	**Actual**
1998	9	14	5	9	2	3
1999	14	12	9	8	4	5
2000	11	14	7	8	3	3
2001	9	15	5	9	2	4
2002	13	12	8	4	4	2
2003	12	-	8	-	3	-

Source: Department of Atmospheric Science, Colorado State University

(b) Each year, a research team at Colorado State University predicts how many hurricanes (of three different intensity levels) will strike the eastern U.S. the next year. The predicted and actual numbers are attached. Explain how to use your answers to Problem 3 and part (a) to decide whether or not the forecasts have any value.

* “[In 2005] forecast 15 named storms, eight of them hurricanes.” [USA Today, 2008]
  * Actual: 28 storms and 15 hurricanes, including Katrina.

* “After the 2005 Atlantic hurricane season, Gray announced that he was stepping back from the primary authorship of CSU’s tropical cyclone probability forecasts.” [Wikipedia].

* “… forecasts that 16 named storms will form in the Atlantic … 72% chance of a major hurricane striking land.” [USA Today, May 2011]

* “Gray … does not attribute global warming to anthropogenic causes, and is critical of those who do.” [Wikipedia]

The charge given to us

* “Highlight methods /models /techniques which may be useful in the analysis of climate data and understanding climate change.”

Standard versus new techniques

* Standard: clustering
  * Each point is mapped to a single cluster center.
* New: topic modeling
  * Each point is generated by *multiple* centroids.

* Standard: principal components analysis
  * Linear method to reduce dimensionality.
* New: link prediction and collaborative filtering
  * Hidden dimensions are inferred to *predict* outcomes.

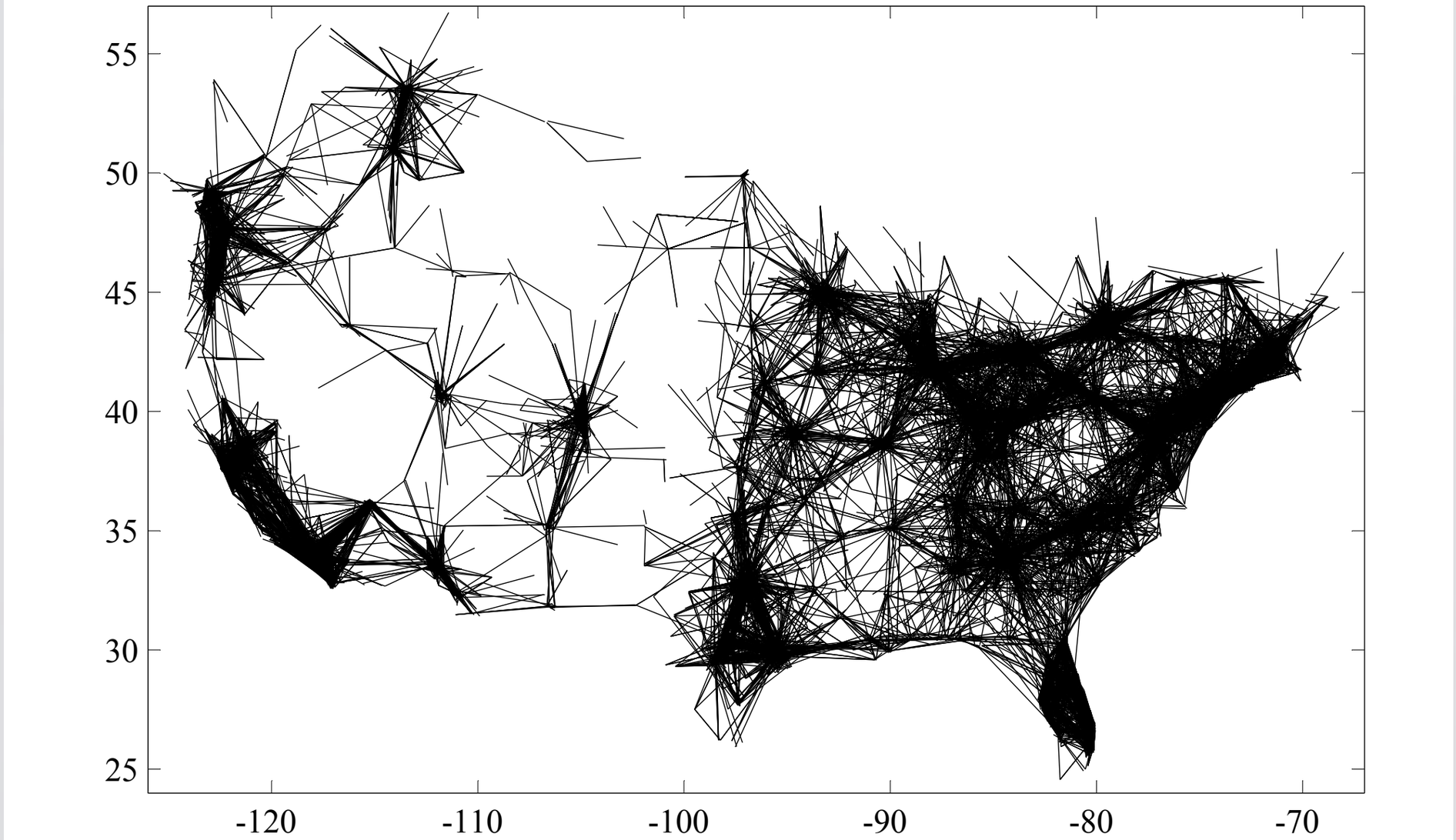
Topic modeling

- In a topic model, each data point is a *combination* of prototypes.
  - Nonlinear, applicable to both discrete and continuous data.

Topic 1 $\bar{\theta}_z = .021$	Topic 2 $\bar{\theta}_z = .019$	Topic 3 $\bar{\theta}_z = .017$	Topic 4 $\bar{\theta}_z = .017$
Southwestern Energy	Penneys	Capital One	Simon Property
Range Resources	Macys	BNY Mellon	Kimco Realty
Cabot Oil & Gas	Kohls	Discover	Equity Residential
EOG Resources	Nordstrom	Northern Trust	AvalonBay Communities
Chesapeake Energy	Target	Janus	Apartment Investment
Pioneer Resources	Limited	JPMorgan Chase	Vornado Realty Trust
Devon Energy	Lowes	State Street	Boston Properties
Peabody Energy	Home Depot	Wells Fargo	Public Storage
Anadarko Petroleum	American Express	PPL	Host Hotels
Massey Energy	Abercrombie	T. Rowe Price	HCP Inc.

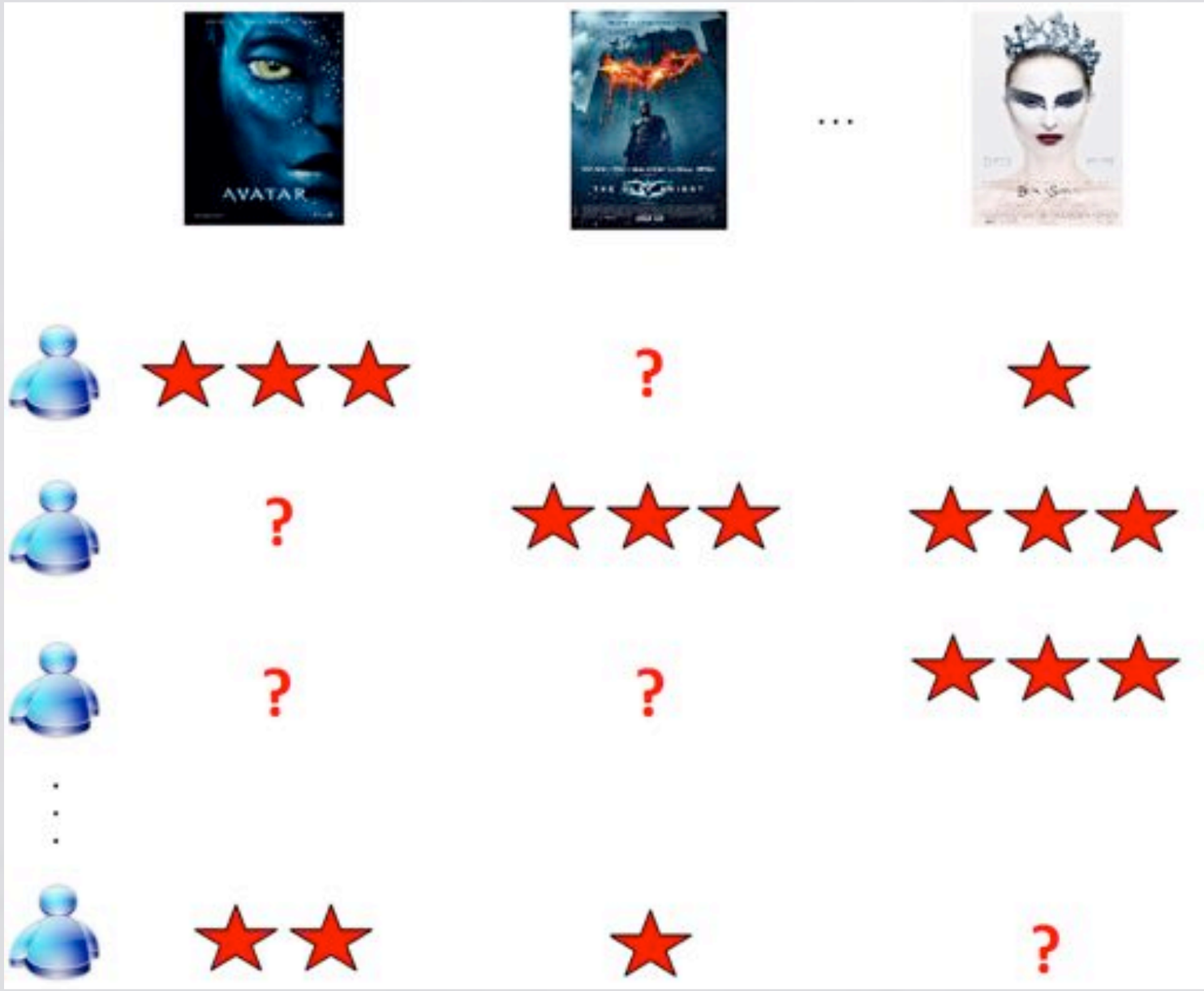
- *Financial topic models*, G. Doyle and C. Elkan, Proc. NIPS Workshop on Applications for Topic Models: Text and Beyond, 2009.
- *Accounting for burstiness in topic models*, G. Doyle and C. Elkan, Proc. Intl Conference on Machine Learning, 2009.

Link prediction



* A. K. Menon and C. Elkan, *Link prediction via matrix factorization*, Proc. ECML 2011.

Collaborative filtering



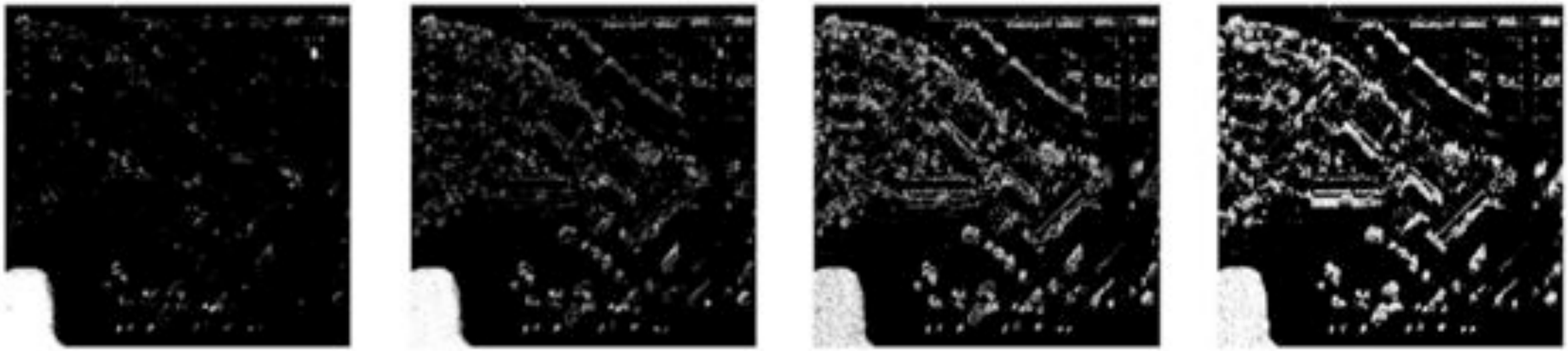
* A. K. Menon and C. Elkan, *Predicting labels for dyadic data*. Data Mining and Knowledge Discovery (2):327-343, 2010.

Standard versus new techniques

* Two more examples

Learning a classifier from only positive examples

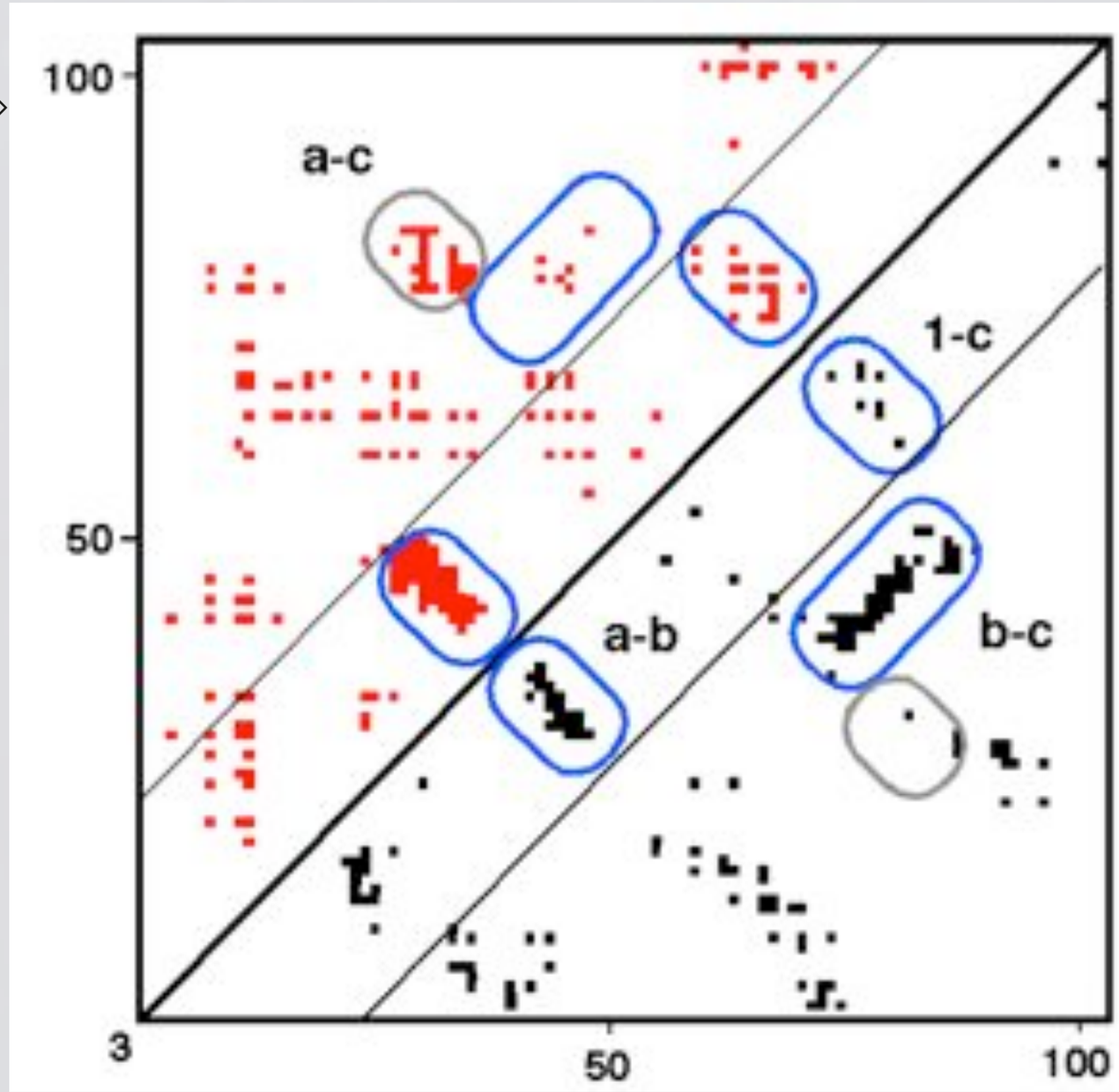
* Training a model to label land cover based on remote-sensing data

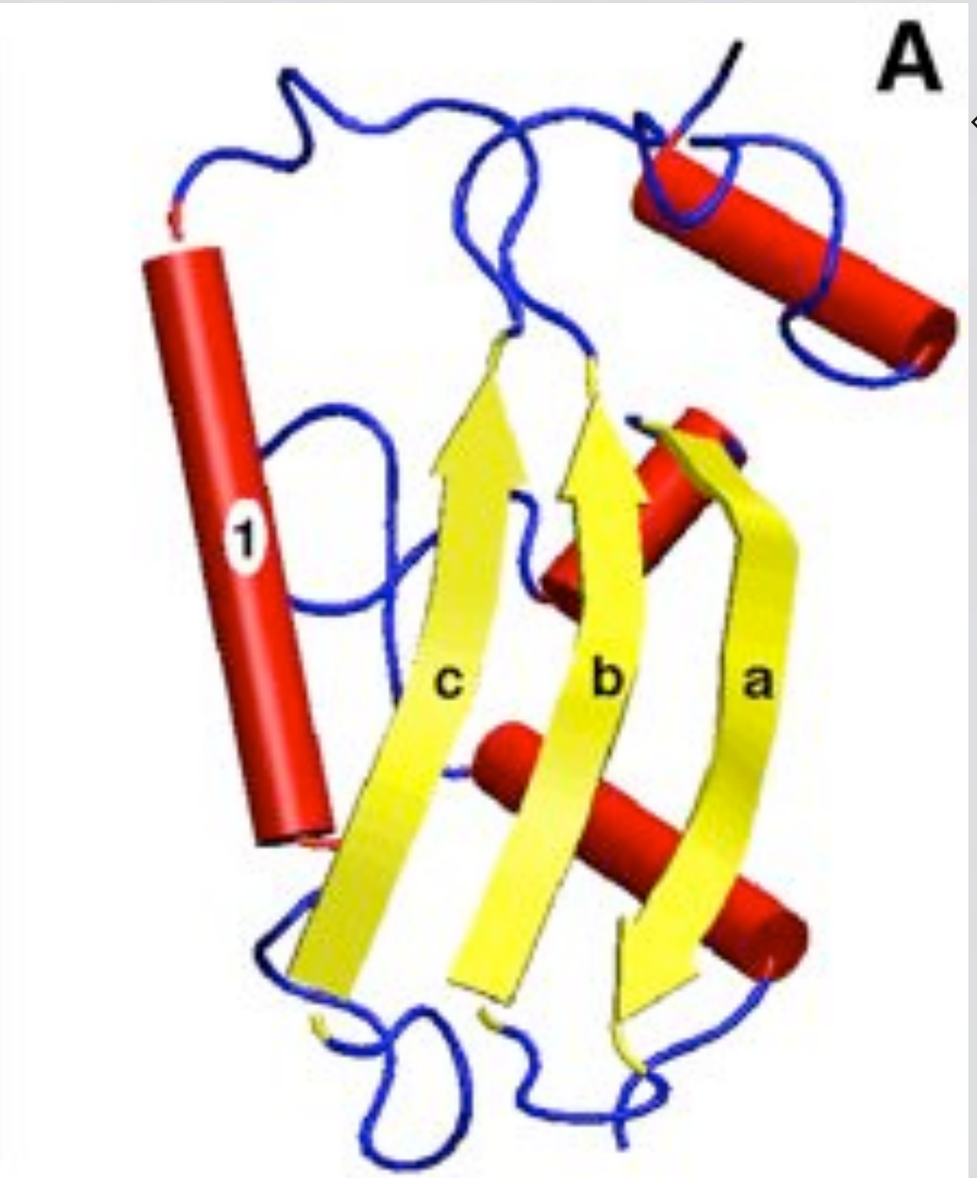


* W. Li, Q. Guo, C. Elkan, *A Positive and Unlabeled Learning Algorithm for One-Class Classification of Remote-Sensing Data*. IEEE Transactions on Geoscience and Remote Sensing 49(2)717-725, 2011.

* C. Elkan, K. Noto, *Learning classifiers from only positive and unlabeled data,* Proc. KDD 2008, pp. 213-220.

Spatial coherence and auto-correlation





* Use two- or three- dimensional conditional random fields