

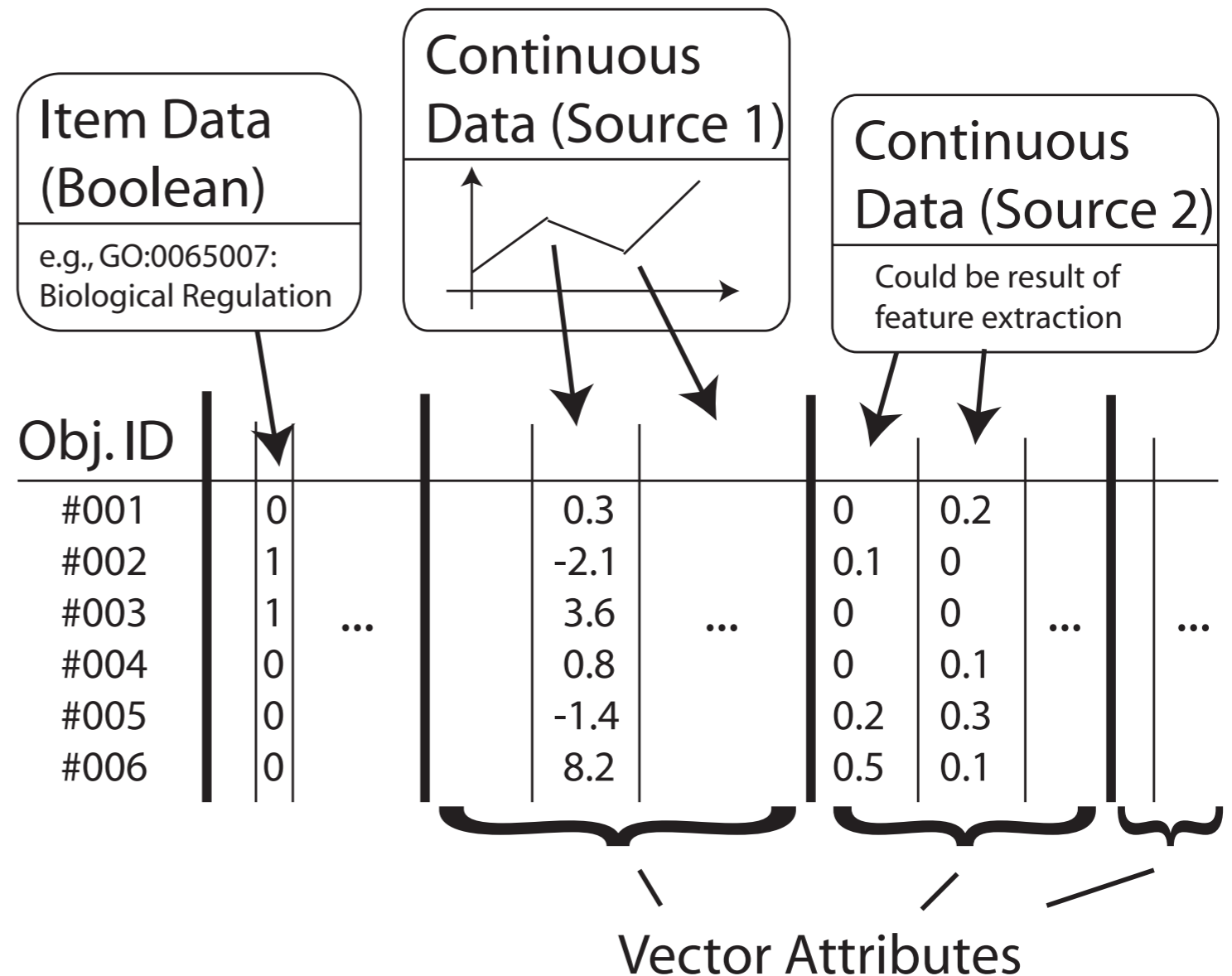
Point-Distribution Algorithm for Mining Vector-Item Patterns

Anne M. Denton, Jianfei Wu
Department of Computer Science
North Dakota State University, Fargo, ND

Dietmar H. Dorr
Research and Development
Thomson Reuters, St. Paul, MN

Motivation: Increasing Diversity of Data

- Vector data
 - Groups of continuous data from multiple sources
 - Could be result of feature extraction
- Item data
 - Binary with presence less frequent than absence
 - Could be item sets

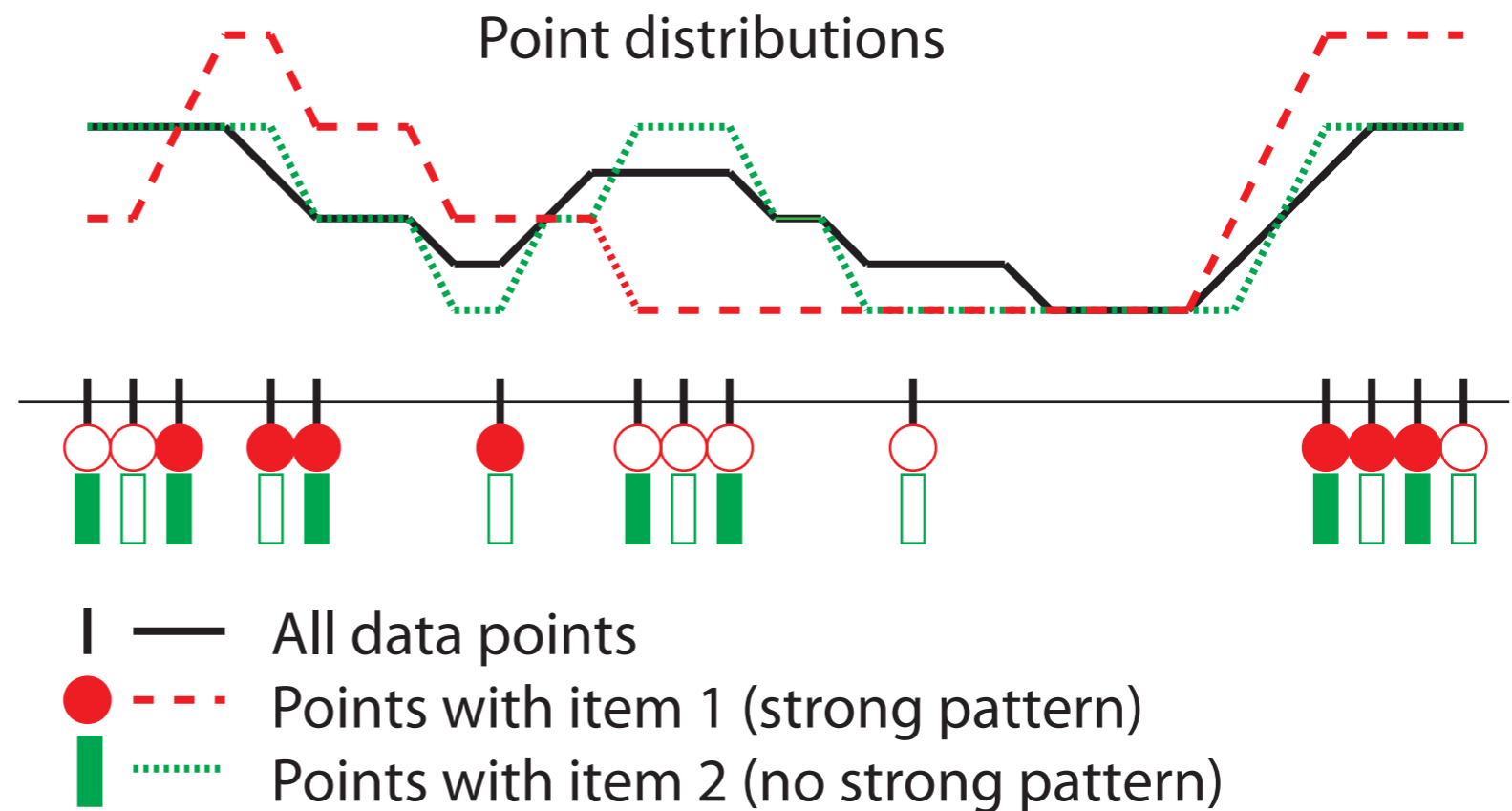


“Patterns of Usefulness”

- Supervised learning as well as some pattern mining approaches assume relationships exist
- Find which vector attribute is most relevant to which item: Multi-dimensional feature selection
- Find item sets that result in the clearest patterns (design of coatings)
- Establish relationship: Multi-dimensional hypothesis testing

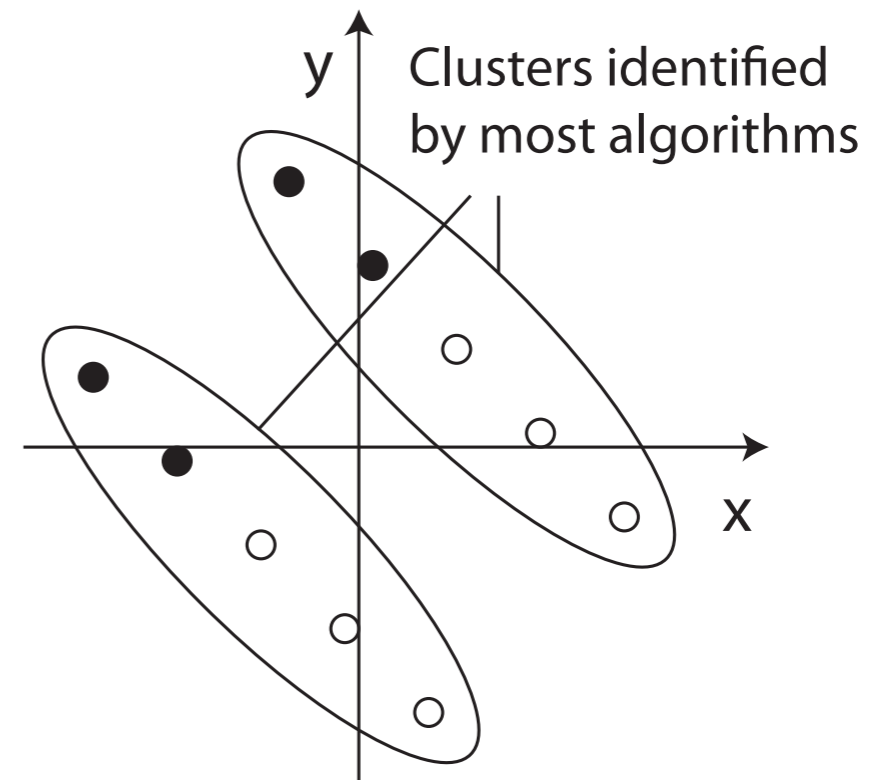
Problem Statement

- Identify items for which distribution of points with item differs significantly from overall distribution (figure 1-d)
- Related work: Is classification significant?



Common Approach in Bioinformatics

- Cluster, then look for enrichment of clusters
- Can miss significant relationships



Approach

- Define density, using kernel function (uniform kernel)
- Compare densities of points with item to densities of all points
- Previous approach used histograms
- Kullback-Leibler divergence quantifies difference between distributions directly

$$D_{\text{KL}}^{(d)}(P||Q) = \int_{-\infty}^{\infty} p(\mathbf{x}) \log \frac{p(\mathbf{x})}{q(\mathbf{x})} dx_1, \dots, dx_d$$

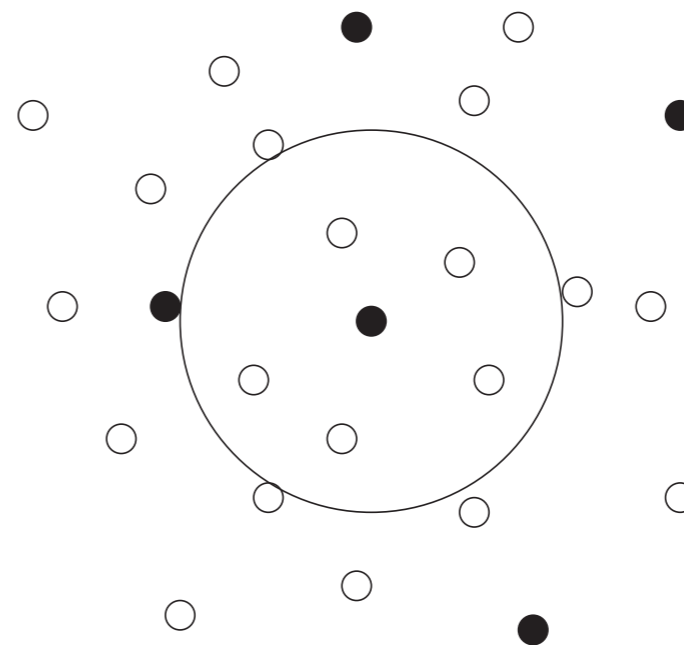
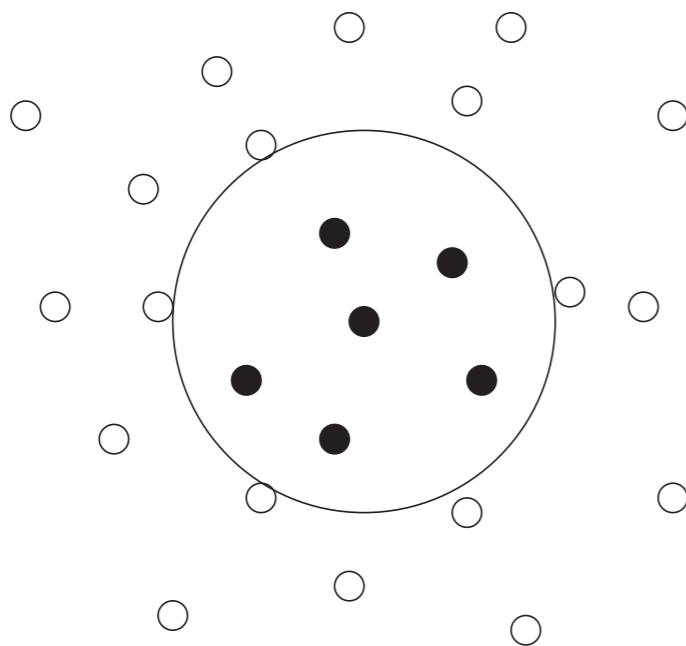
$$D_{\text{KL}}^{(d)}(P||Q) = \frac{1}{n} \sum_{i=1}^n \log \frac{p(\mathbf{x}^{(i)})}{q(\mathbf{x}^{(i)})}$$

Algorithm (simplified)

- One parameter: Similarity threshold *thresh* (next slide)
- For each item
 - For all points with item
 - Find number of neighbors with item closer than *thresh* and divide by overall support of item: $p(x)$
 - Find total number of neighbors closer than *thresh* and divide by total number of points: $q(x)$
 - Calculate Kullback-Leibler divergence and compare with distribution of K-L divergences for random data

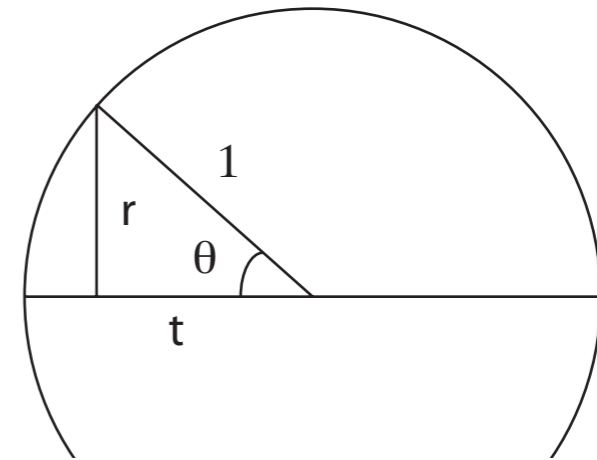
Parameter Choice

- Only parameter: Expected number of neighbors
- Choice of one intuitive (can be justified mathematically)
- Confirmed by experiments



Calculation of *thresh*

- Calculate surface of cap of hypersphere



$$S_{d-1}^{(\text{cap})} = \int_0^{\theta_t} S_{d-2} r^{d-2} d\theta = \int (d-1) C_{d-1} r^{d-2} d\theta$$

- Gives expression of support as function of *thresh* that can be used for lookup

$$\frac{1}{\text{support}} = \frac{\int_0^{\sqrt{1-\text{thresh}^2}} \frac{r^{d-2}}{\sqrt{1-r^2}} dr}{2 \int_0^1 \frac{r^{d-2}}{\sqrt{1-r^2}} dr}$$

Genomics Application

- Finding protein domains that are related to a set of experiments in yeast
- Protein domains
 - Binary attributes
 - From Interpro database
- Gene expression data
 - All come from cell cycle experiments and are expected to represent related information
 - Four time series, each one consisting of 14 - 24 experiments

Results for Gene Expression Data

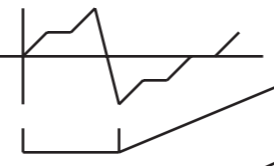
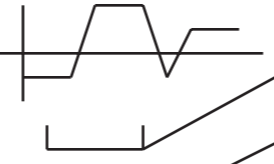
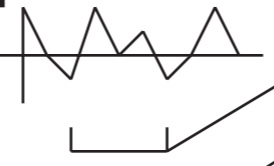
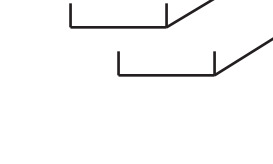
- Significance not known independently
- Results should be consistent over comparable experiments
 - Top right: Overlap
 - Bottom left: Significance that results are related

Table 1: Results for Gene Expression Data

All	Alpha	Cdc15	Cdc28	Elu	
259	114	117	160	166	All
	119	67	85	79	Alpha
0		134	86	72	Cdc15
7E-15	3E-12		173	107	Cdc28
0	2E-16	7E-12		198	Elu
0	1E-7	0.027	5E-8		

Labeled Data From Time Series

- Construct labeled data from time series sub-sequences
- Item data: membership in time series
- Noisy data (intentionally chosen)
- Allows varying item support by adding random walk data

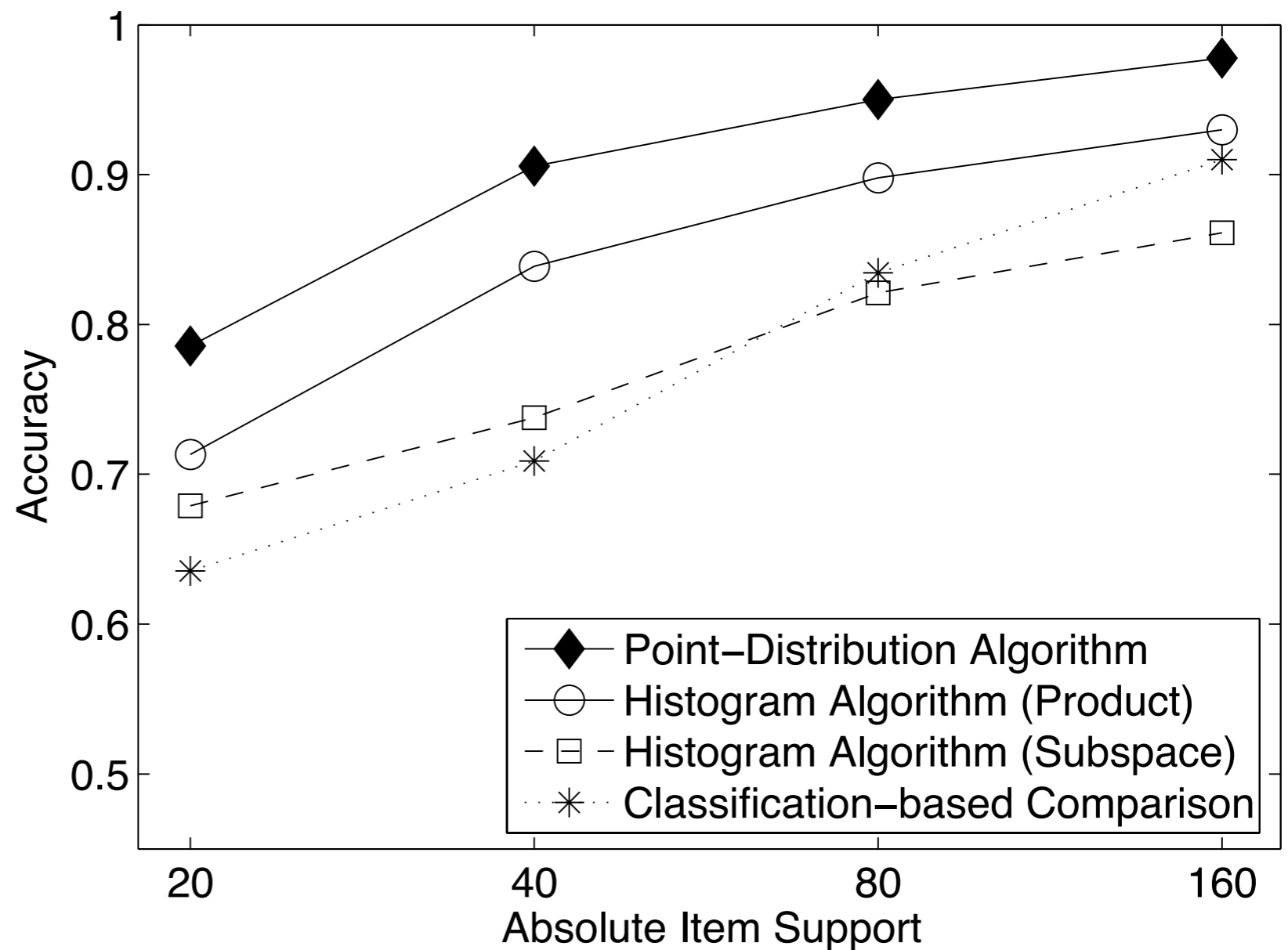
		Vector Data	Time Series Items	Randomized Items
Time-series 1		1 0 1 -4 -4 1 0 1 0 3 0 0	1 0 1 0 0 1	0 0 0 1 1 0
Time-series 2		3 0 0 -3 3 -2 1 -2	0 1 0 1 0 0	0 1 0 1 0 0
Random Walk		1 -2 1 2	0 0 0 0	1 0 0 0
Time-series				

Comparison Approaches

- Histogram-based approach from
 - A.M. Denton and J. Wu, KAIS, 2009
 - Summarizes density distributions as histograms
- Classification-based approach
 - Predict each item using classification (tree-based classifier in MATLAB)
 - Make prediction using 2-fold cross-validations
 - Calculate significance based on confusion matrix

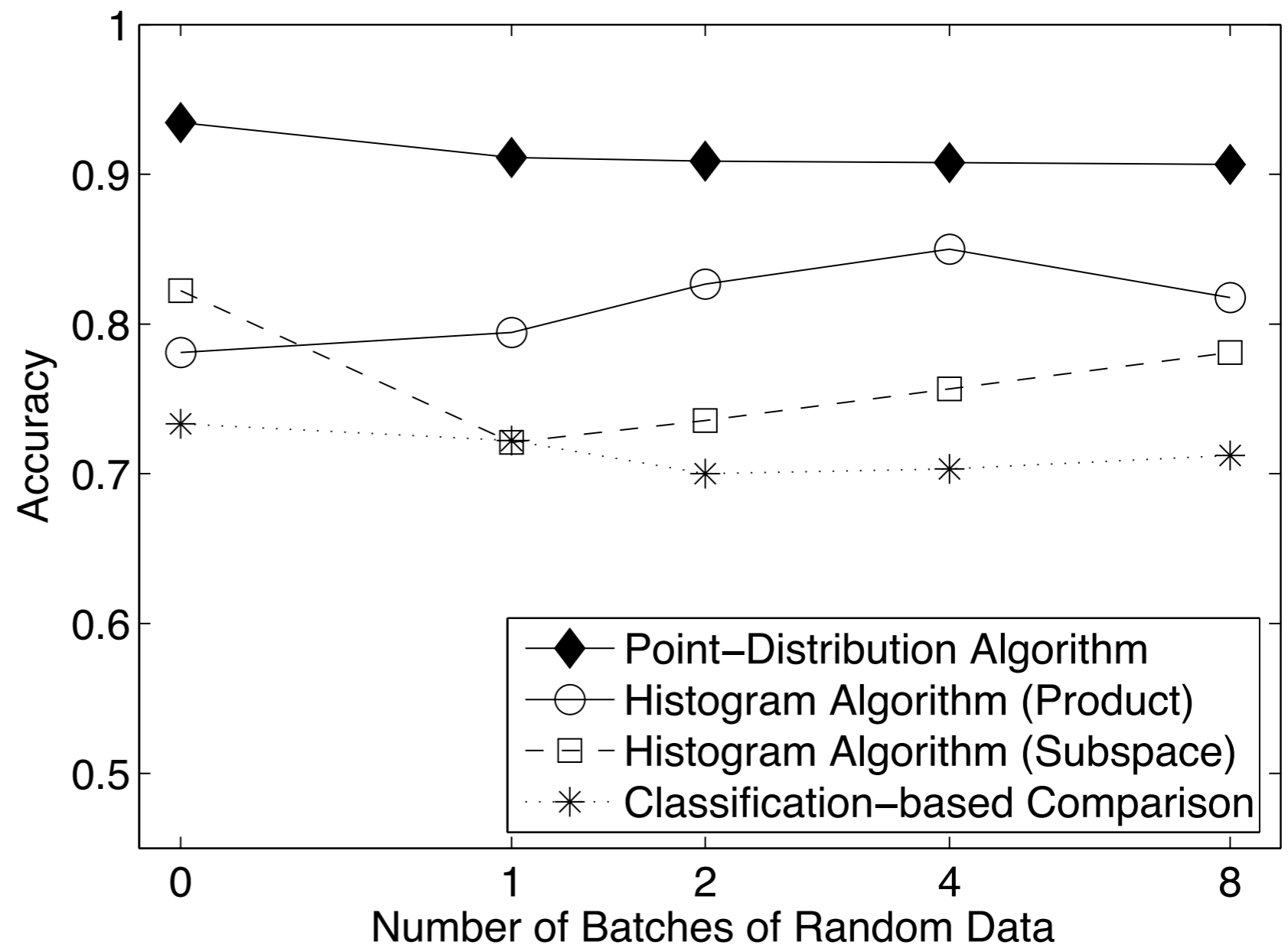
Accuracy Depending on Item Support

- Clearly superior to comparison algorithms
- For very large item support classification may become competitive



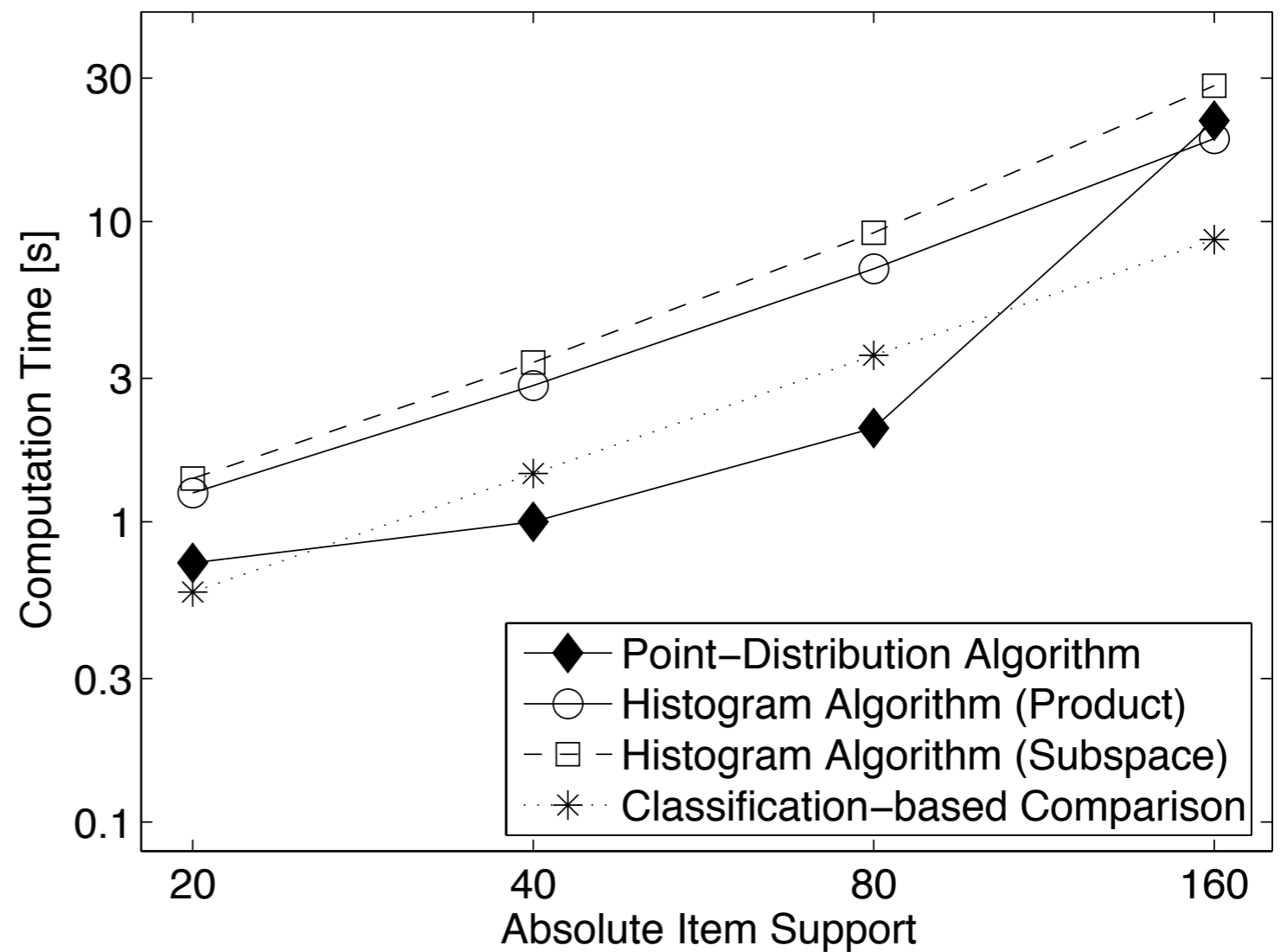
Accuracy Depending on Amount of Noise

- Accuracy superior for all settings
- Not degrade much with added noise



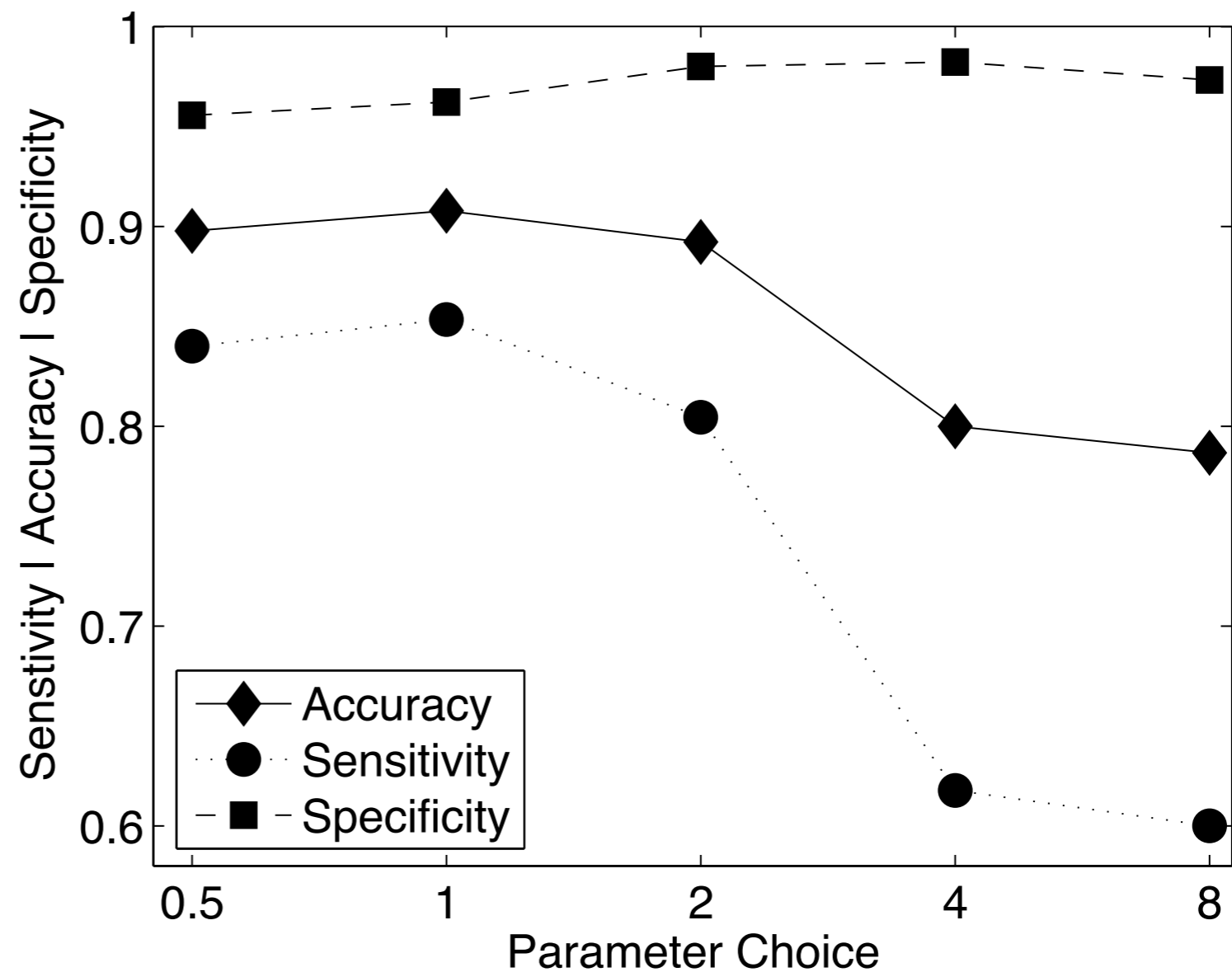
Performance

- Speed comparable to other algorithms
- Scaling with item support poorer
- Accuracy main motivation
- Algorithm most important for small item support



Parameter Choice

- Single parameter (expected # of neighbors) set to one
- Experiments confirm choice



Conclusions

- Solves an important problem
 - Finding significant relationships between vectors and items or item sets
- Use of Kullback-Leibler divergence better justified theoretically than histograms
- Application to genomic data gives consistent results
- Accuracy much improved on semi-artificial data (constructed from real time series)