

# ***Geometrical Complexity of Classification Problems***

*Tin Kam Ho*

Bell Laboratories

Lucent Technologies

# Our Goals

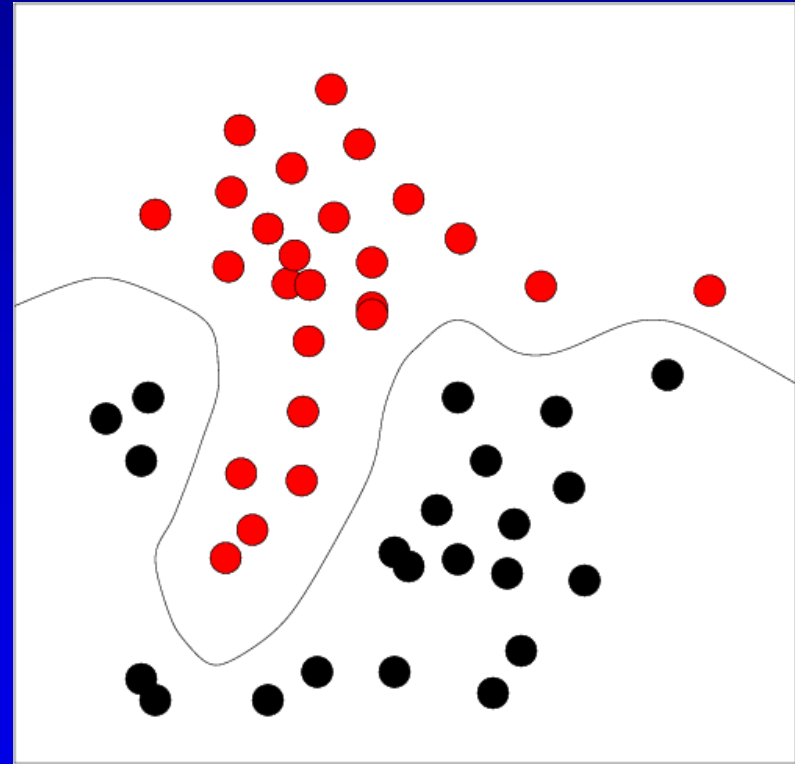
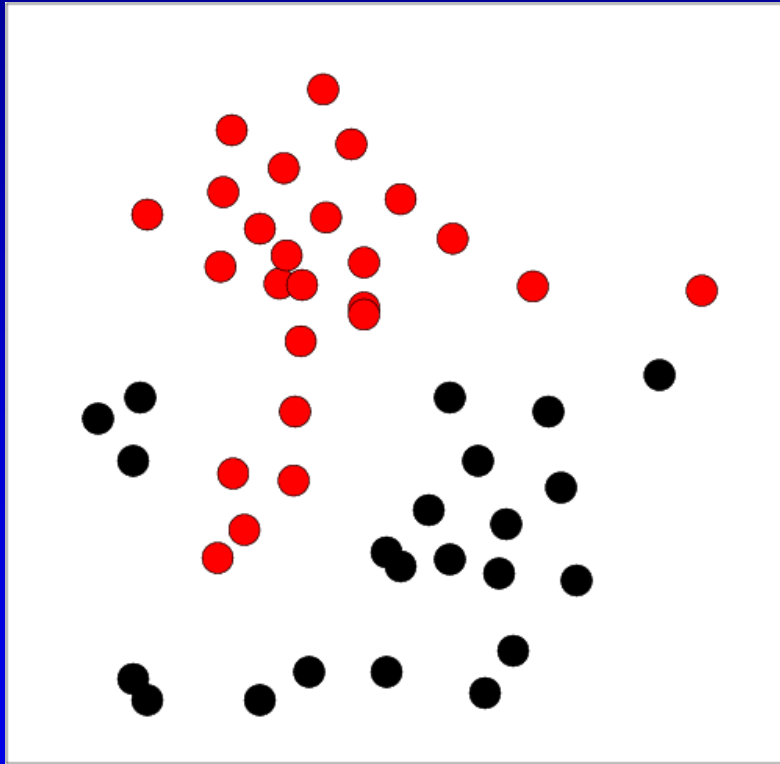
Better understand

- Geometry and topology of point sets in high-dimensional spaces
- Preservation of such characteristics under feature transformations and sampling processes
- Their interaction with geometrical models used in classifiers

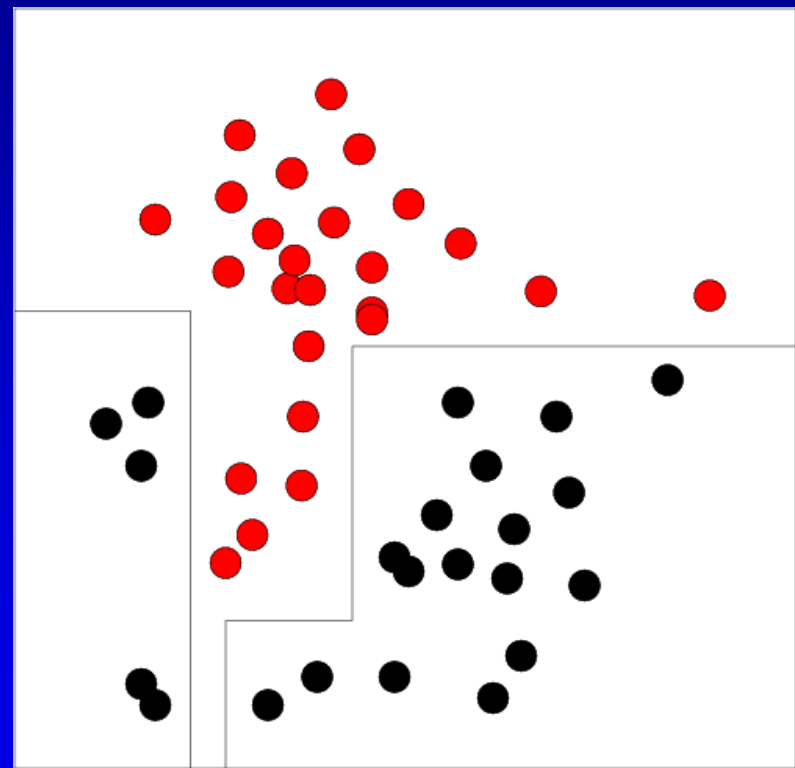
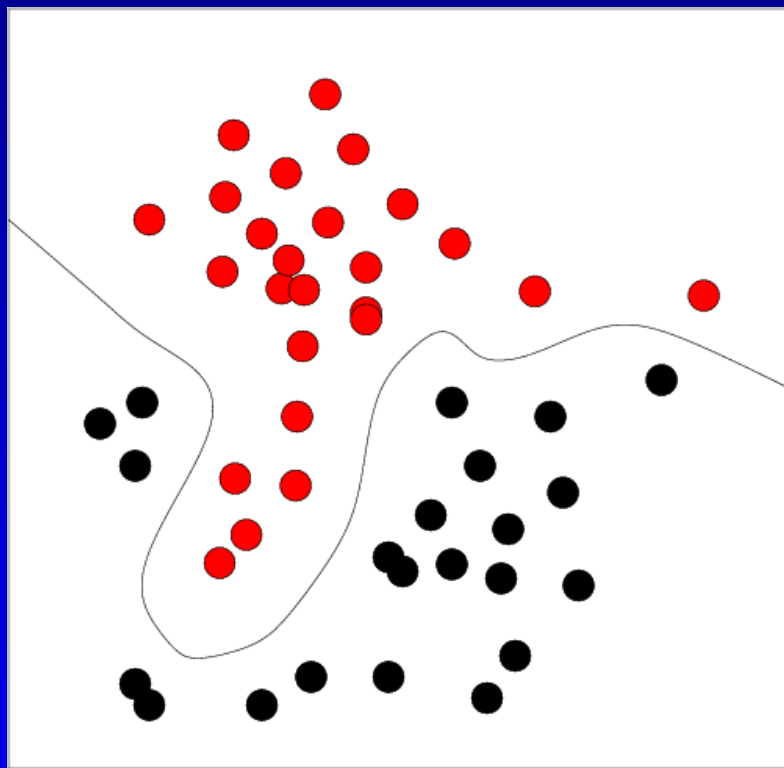
# Geometrical Complexity of Classification

- Data sets:
  - length of class boundary
  - fragmentation of classes / existence of subclasses
  - global or local linear separability
  - convexity and smoothness of boundaries
  - intrinsic / extrinsic dimensionality
  - stability of these characteristics as sampling rate changes
- Classifier models:
  - polygons, hyperspheres, Gaussian kernels, axis-parallel cuts, piece-wise linear surfaces, polynomial surfaces, their unions or intersections, ...

# Supervised Classification -- Discrimination Problems



# An Ill-Posed Problem



# Where Were We in the Late 1990's?

- Statistical Methods
  - Bayesian classifiers, polynomial discriminators, nearest-neighbors, decision trees, neural networks, support vector machines, ...
- Syntactic Methods
  - regular grammars, context-free grammars, attributed grammars, stochastic grammars, ...
- Structural Methods
  - graph matching, elastic matching, rule-based systems, ...

# Classifiers

- Competition among different ...
  - choices of features
  - feature representations
  - classifier designs
- Chosen by heuristic judgements
- No clear winners

# Classifier Combination Methods

- Decision optimization methods
  - find consensus from a given set of classifiers
  - majority/plurality vote, sum/product rule
  - probability models, Bayesian approaches
  - logistic regression on ranks or scores
  - classifiers trained on confidence scores



# Classifier Combination Methods

- Coverage optimization methods
  - subsampling methods:  
stacking, bagging, boosting
  - subspace methods:  
random subspace projection, localized selection
  - superclass/subclass methods:  
mixture of experts, error-correcting output codes
  - perturbation in training

# Layers of Choices



**Best Features?**

**Best Classifier?**

**Best Combination Method?**

**Best (combination of)\*  
combination methods?**

# More Questions



- How do confidence scores differ from feature values?
- Is combination a convenience or a necessity?
- What are common among various combination methods?
- When should the combination hierarchy terminate?

# Difficulties in Classifier Comination

- Many theories have inadequate assumptions
- Geometry and probability lack connection
- Combinatorics defies detailed modeling
- Attempt to cover all cases gives weak results
  
- Empirical results overly specific to problems
- Lack of systematic organization of evidences

# Data Dependent Behavior of Classifiers



- Different classifiers excel in different problems
- So do combined systems
- This complicates theories and interpretation of observations

## Questions to ask:

- Does this method work for all problems?
- Does this method work for this problem?
- Does this method work for **this type of problems?**

**Study the **interaction** of  
data and classifiers**

**Characterization of**  
**Data and Classifier Behavior**  
**in a Common Language**

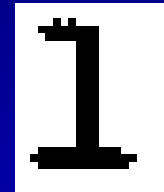
# Sources of Difficulty in Classification

- Class ambiguity
- Boundary complexity
- Sample size and dimensionality



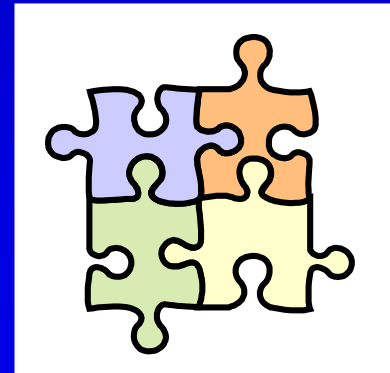
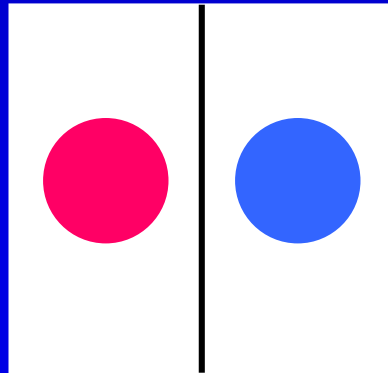
# Class Ambiguity

- Is the problem intrinsically ambiguous?
- Are the classes well defined?
- What is the information content of the features?
- Are the features sufficient for discrimination?

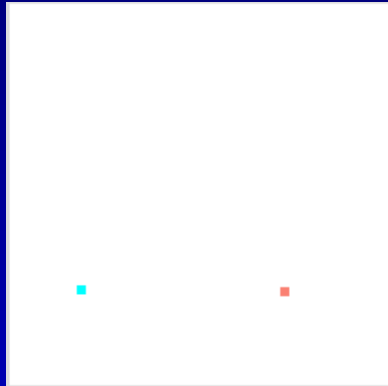


# Boundary Complexity

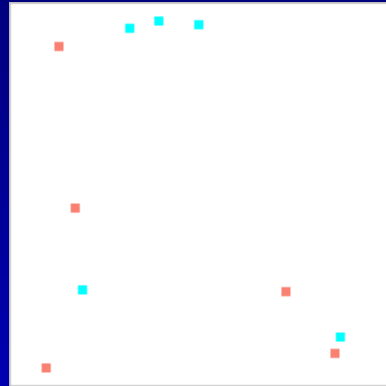
- Kolmogorov complexity
- Length may be exponential in dimensionality
- Trivial description: list all points, class labels
- Is there a shorter description?



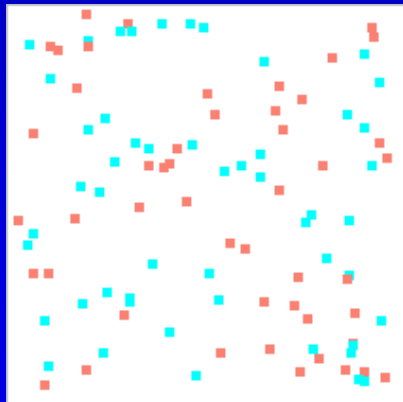
# Sampling Density



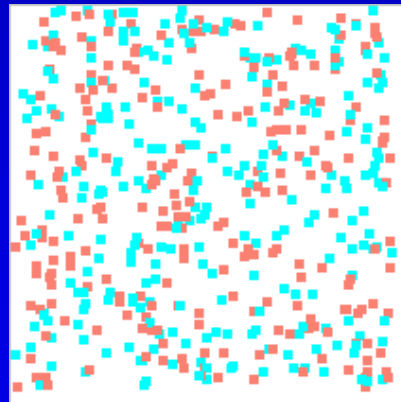
$N = 2$



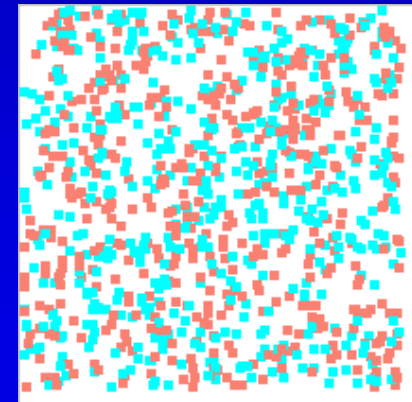
$N = 10$



$N = 100$



$N = 500$



$N = 1000$



# Sample Size & Dimensionality

- Problem may appear deceptively simple or complex with small samples
- Large degree of freedom in high-dim. spaces
- Representativeness of samples vs. generalization ability of classifiers

# Mixture of Effects

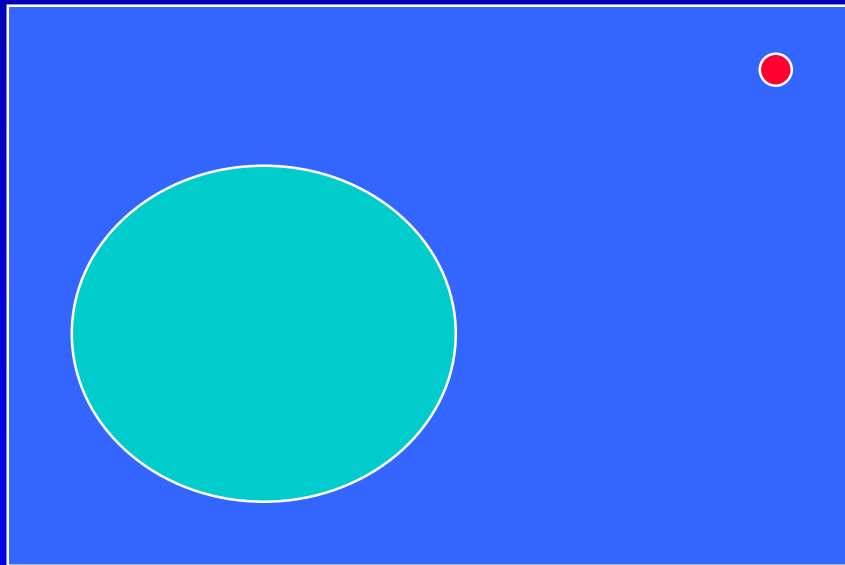
- Real problems often have mixed effects of
  - class ambiguity
  - boundary complexity
  - sample size & dimensionality
- Geometrical complexity of class manifolds
  - coupled with probabilistic sampling process

# Geometry vs. Probability

- Geometry of classifiers determines the **rule of generalization** to unseen samples
- Assumption of representative samples  
 Optimistic error bounds
- Distribution-free arguments  
 Pessimistic error bounds

# Geometry vs. Probability

- Difficult by probability: detecting 1 disease case from 1,000,000 normal ones
- Not necessarily difficult by geometry:



# Geometrical Complexity of Classification Problems

- Study geometry of data sets
- Study geometry of decision regions
- Develop a **language** for describing geometrical properties of point sets in high-dimensional spaces
- Develop **tools** for understanding data and decision geometry

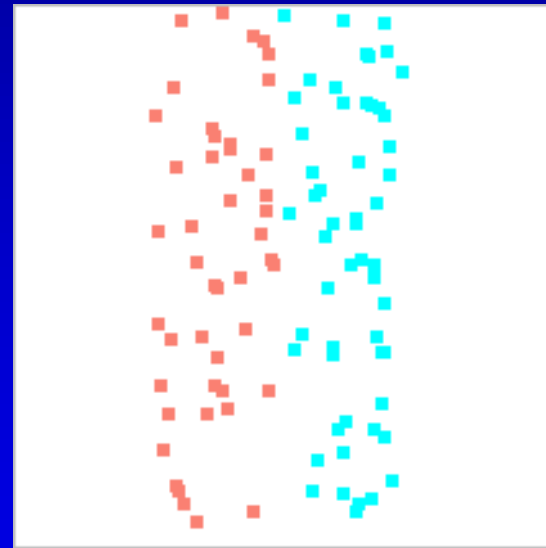
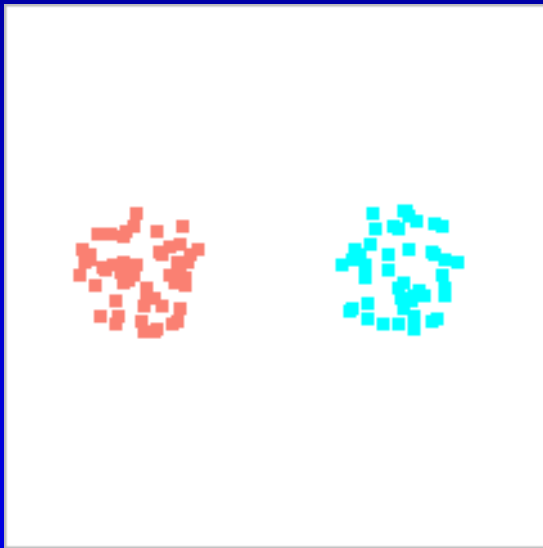


# Building up the **Language**

- Identify **key features** of data geometry that are relevant for classification
- Develop **algorithms** to extract such features from a dataset
- Describe **patterns** of classifier behavior in terms of geometrical features
  - ... pattern recognition in pattern recognition problems
  - ... classification of classification problems

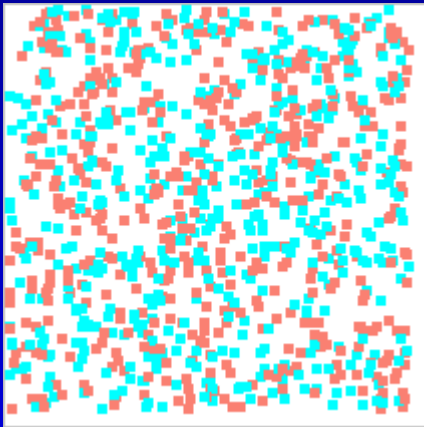
# Easy or Difficult Problems

- Linearly separable problems

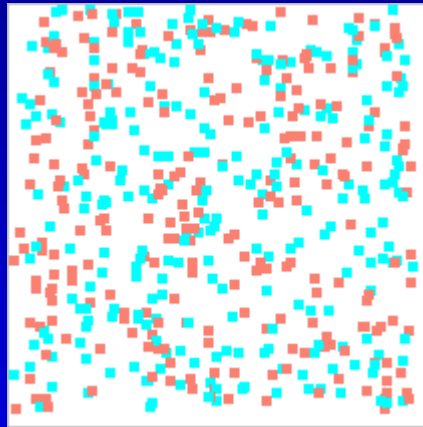


# Easy or Difficult Problems

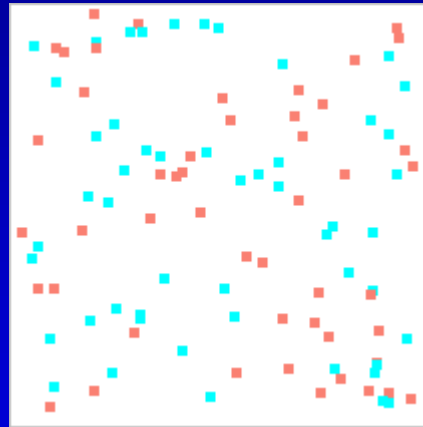
- Random noise



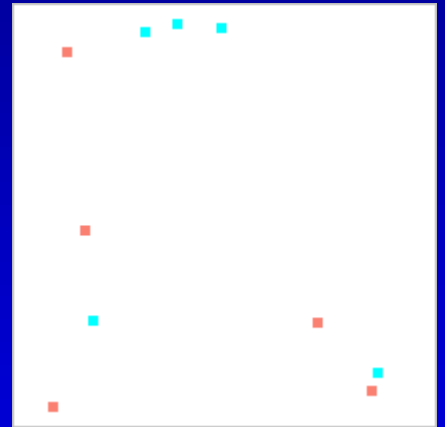
1000 points



500 points



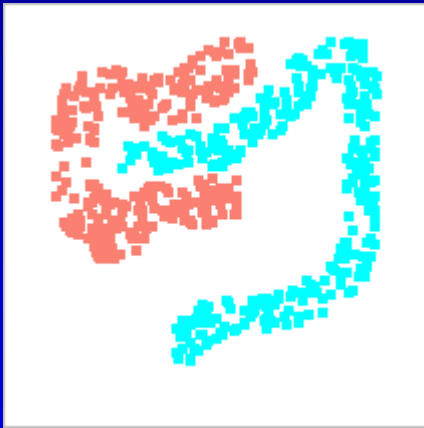
100 points



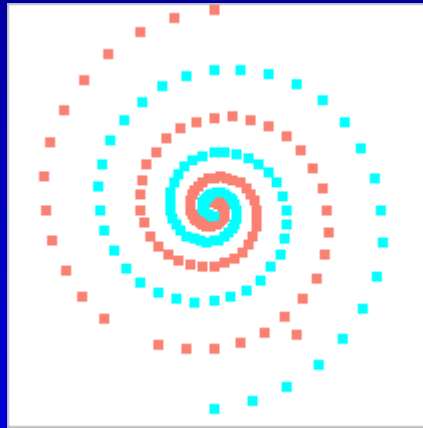
10 points

# Easy or Difficult Problems

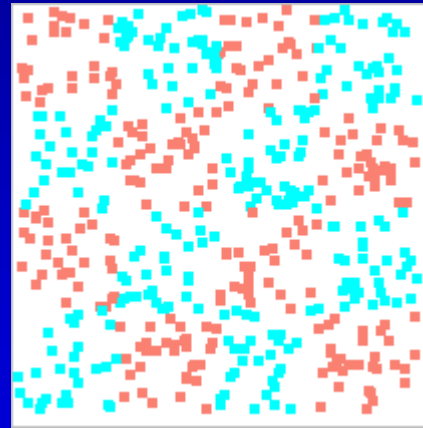
- Others



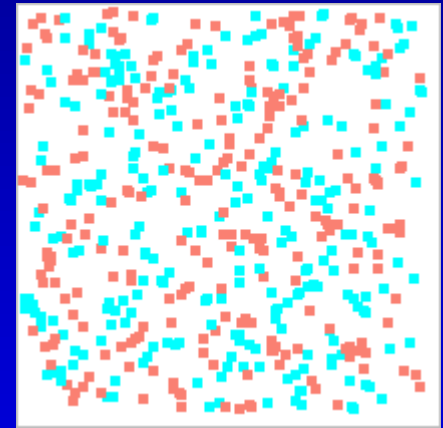
Nonlinear  
boundary



Spirals



4x4  
checkerboard



10x10  
checkerboard

# Description of Complexity

- What are real-world problems like?
- Need a description of complexity to
  - set expectation on recognition accuracy
  - characterize behavior of classifiers
- Apparent or true complexity?

# Possible Measures

- Separability of classes
  - linear separability
  - length of class boundary
  - intra / inter class scatter and distances
- Discriminating power of features
  - Fisher's discriminant ratio
  - overlap of feature values
  - feature efficiency

# Possible Measures

- Geometry, topology, clustering effects
  - curvature of boundaries
  - overlap of convex hulls
  - packing of points in regular shapes
  - intrinsic dimensionality
  - density variations

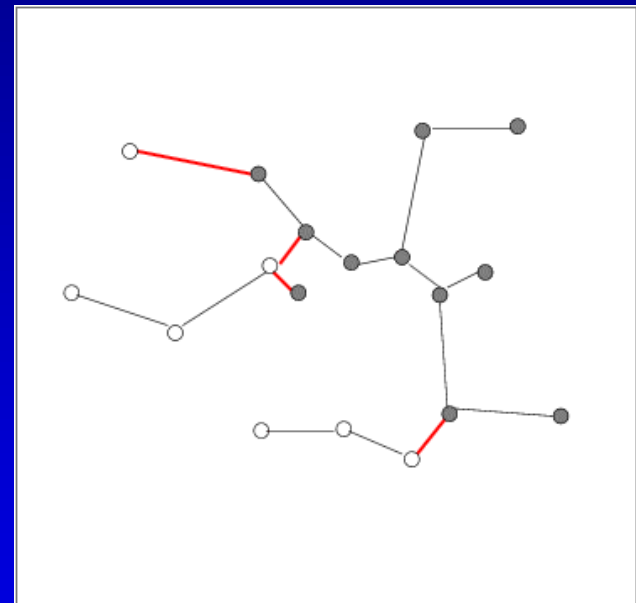
# Linear Separability

- Intensively studied in early literature
- Many algorithms only stop with positive conclusions
  - Perceptrons, Perceptron Cycling Theorem, 1962
  - Fractional Correction Rule, 1954
  - Widrow-Hoff Delta Rule, 1960
  - Ho-Kashyap algorithm, 1965
  - Linear programming, 1968



# Length of Class Boundary

- Friedman & Rafsky 1979
  - Find MST (minimum spanning tree) connecting all points regardless of class
  - Count edges joining opposite classes
  - Sensitive to separability and clustering effects



# Fisher's Discriminant Ratio

- Defined for one feature:

$$f = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2}$$

$$\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$$

means, variances of classes 1,2

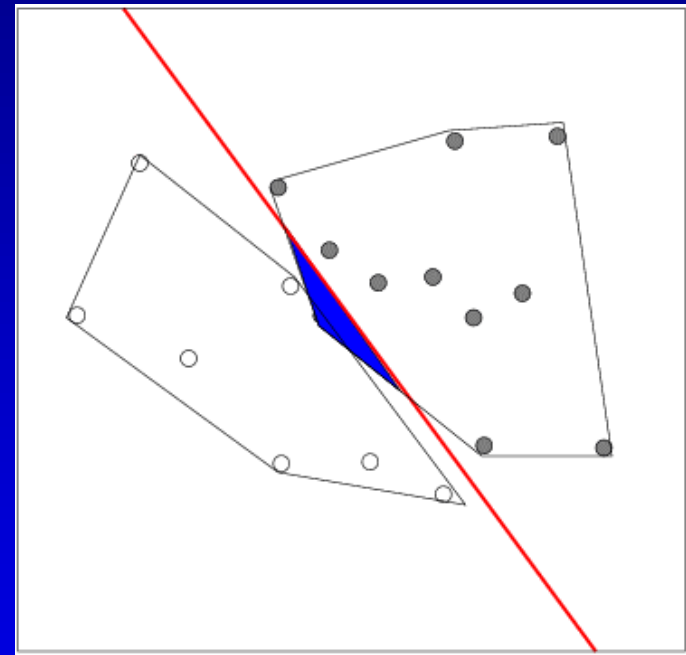
- One good feature makes a problem easy
- Take maximum over all features

# Volume of Overlap Region

- Overlap of class manifolds
- Overlap region of each dimension as a fraction of range spanned by the two classes
- Multiply fractions to estimate volume
- Zero if no overlap

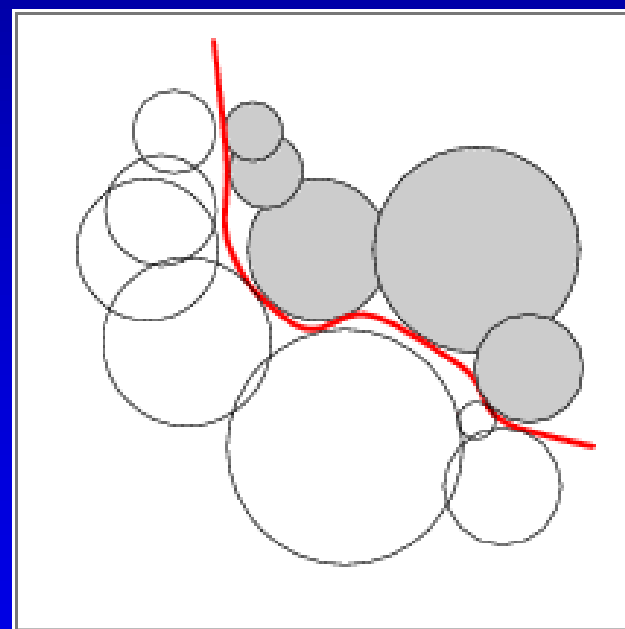
# Convex Hulls & Decision Regions

- Hoekstra & Duin 1996
- Measure **nonlinearity** of a classifier w.r.t. a given dataset
- Sensitive to smoothness of decision boundaries



# Shapes of Class Manifolds

- Lebourgeois & Emptoz 1996
- Packing of same-class points in hyperspheres
- Thick and spherical, or thin and elongated manifolds



# Measures of Geometrical Complexity

F1	maximum Fisher's discriminant ratio
F2	volume of overlap region
F3	maximum (individual) feature efficiency
L1	minimized error by linear programming (LP)
L2	error rate of linear classifier by LP
L3	nonlinearity of linear classifier by LP
N1	fraction of points on boundary (MST method)
N2	ratio of average intra/inter class NN distance
N3	error rate of 1NN classifier
N4	nonlinearity of 1NN classifier
T1	fraction of points with associated adherence subsets retained
T2	average number of points per dimension

# Space of Complexity Measures

- Single measure may not suffice
- Make a measurement space
- See where datasets are in this space
- Look for a continuum of difficulty:




Easiest Cases




Most difficult cases

# Data Sets: UCI

- UC-Irvine collection
- 14 datasets (no missing values, > 500 pts)
- 844 two-class problems
- 452 linearly separable 
- 392 linearly nonseparable
- 2 - 4648 points each
- 8 - 480 dimensional feature spaces

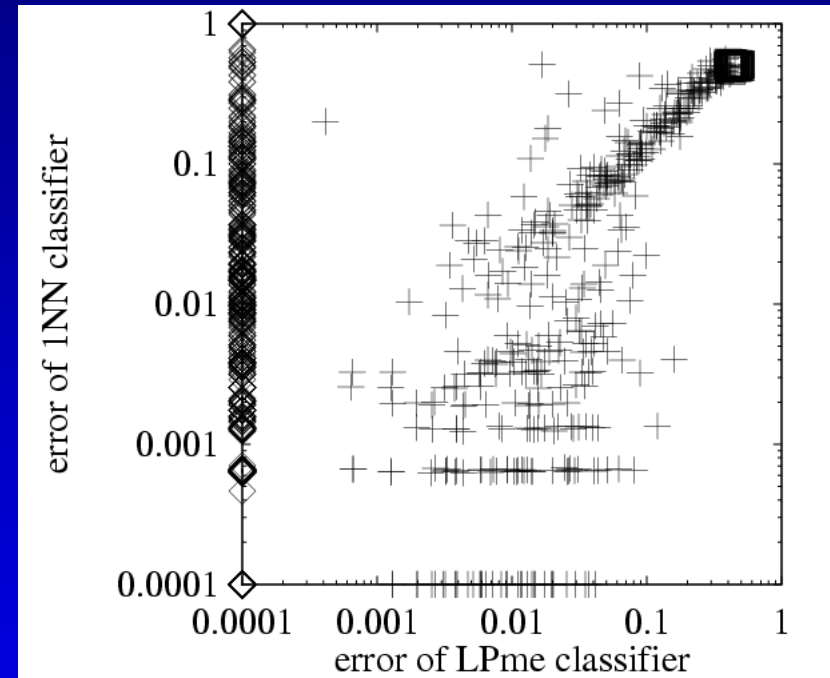
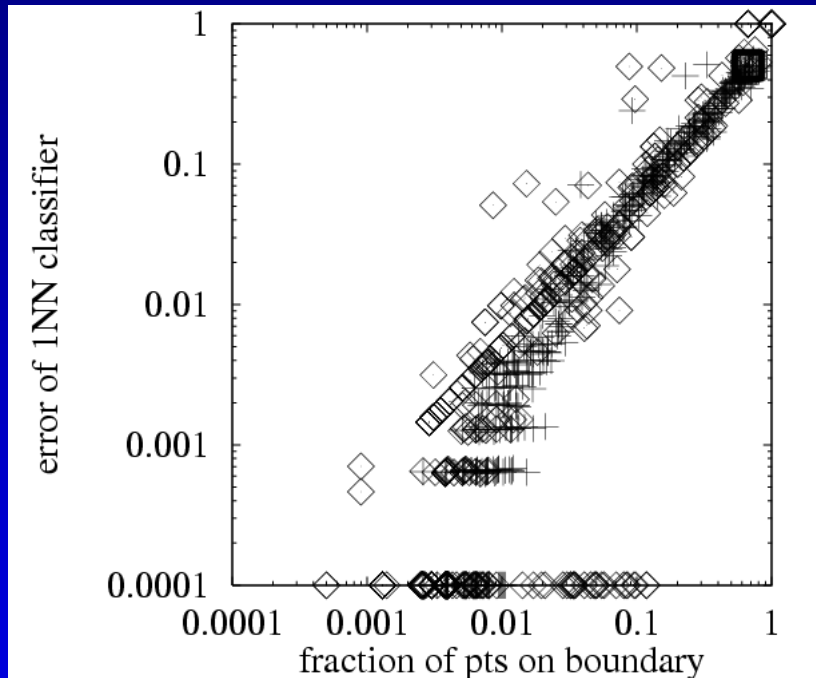


# Data Sets: Random Noise

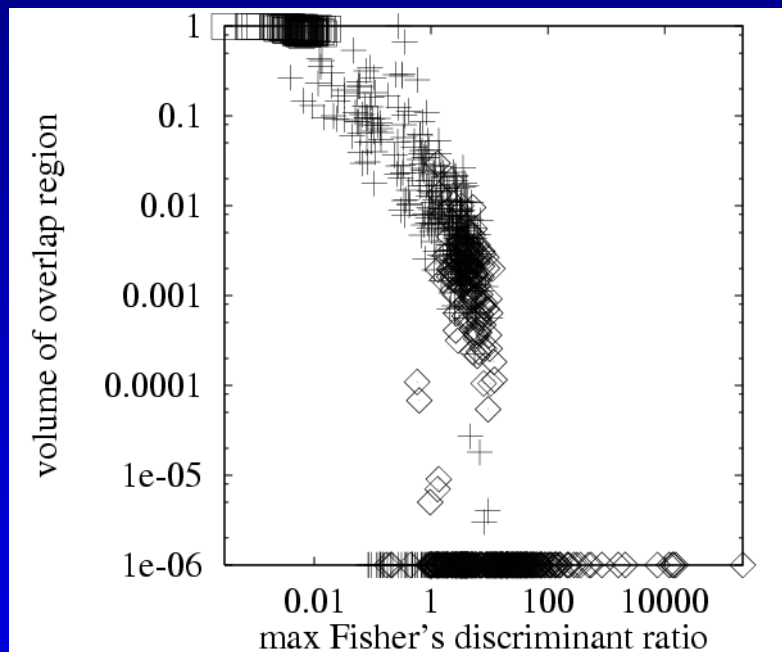
- Randomly located and labeled points 
- 100 artificial problems
- 1 to 100 dimensional feature spaces
- 2 classes, 1000 points per class



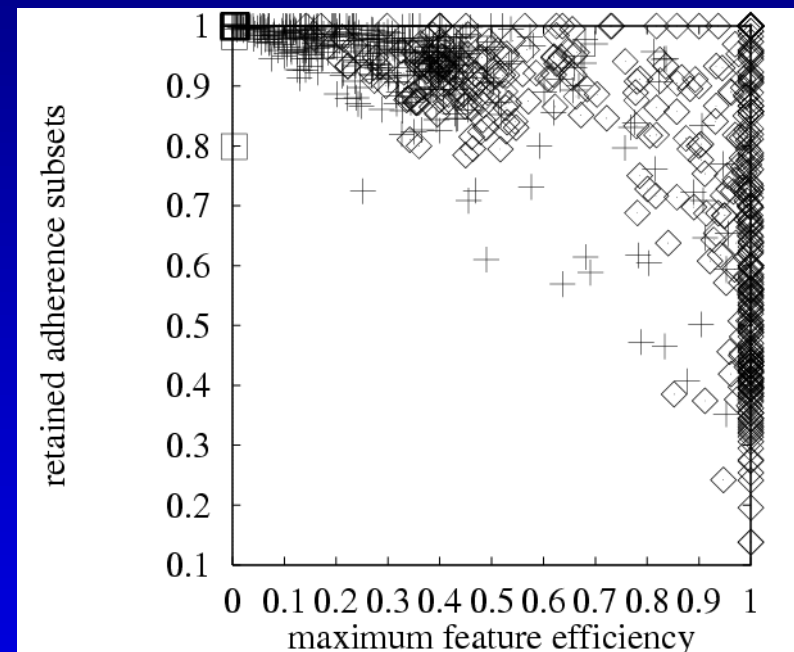
# Correlated or Uncorrelated Measures



# Separation



# Separation + Scatter



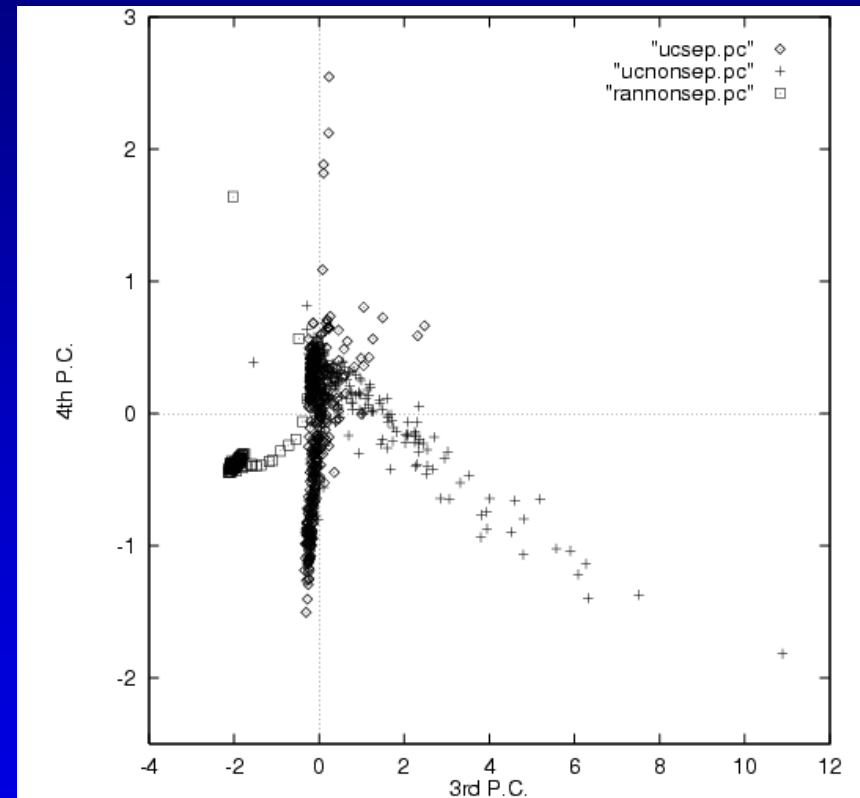
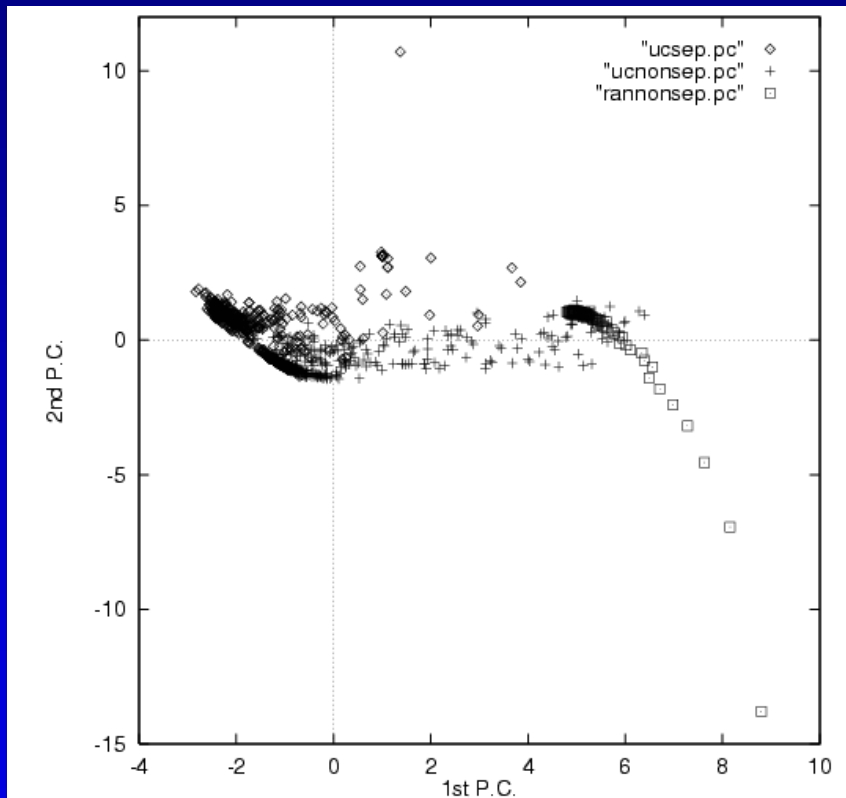
# Observations

- Noise sets and linearly separable sets occupy **opposite ends** in many dimensions
- In-between positions tell **relative difficulty**
- **Fan-like structure** in most plots
- At least **2 independent factors**, **joint effects**
- Noise sets are **far** from real data
- Ranges of noise sets: **apparent complexity**

# Principle Component Analysis

Component	C1	C2	C3	C4	C5	C6
Prop. of Var.	0.5033	0.1162	0.1064	0.0859	0.0761	0.0521
Cum. Prop.	0.5033	0.6195	0.7259	0.8118	0.8879	0.9400
Loadings						
F1	0.01	0.26	0.03	0.86	-0.26	-0.33
F2	0.33	0.08	-0.43	-0.12	-0.09	-0.20
F3	-0.29	0.42	0.03	-0.11	-0.32	0.29
L1	0.17	0.08	0.68	-0.15	0.00	-0.36
L2	0.38	0.04	-0.15	-0.14	-0.10	-0.24
L3	0.38	0.05	-0.16	-0.14	-0.12	-0.23
N1	0.36	0.30	0.04	0.01	-0.04	0.36
N2	0.37	-0.02	0.02	0.03	0.12	-0.03
N3	0.32	0.36	0.00	0.11	-0.03	0.49
N4	0.24	-0.20	0.52	-0.04	-0.35	0.16
T1	0.23	-0.32	0.07	0.37	0.57	0.28
T2	0.08	-0.61	-0.15	0.13	-0.58	0.22

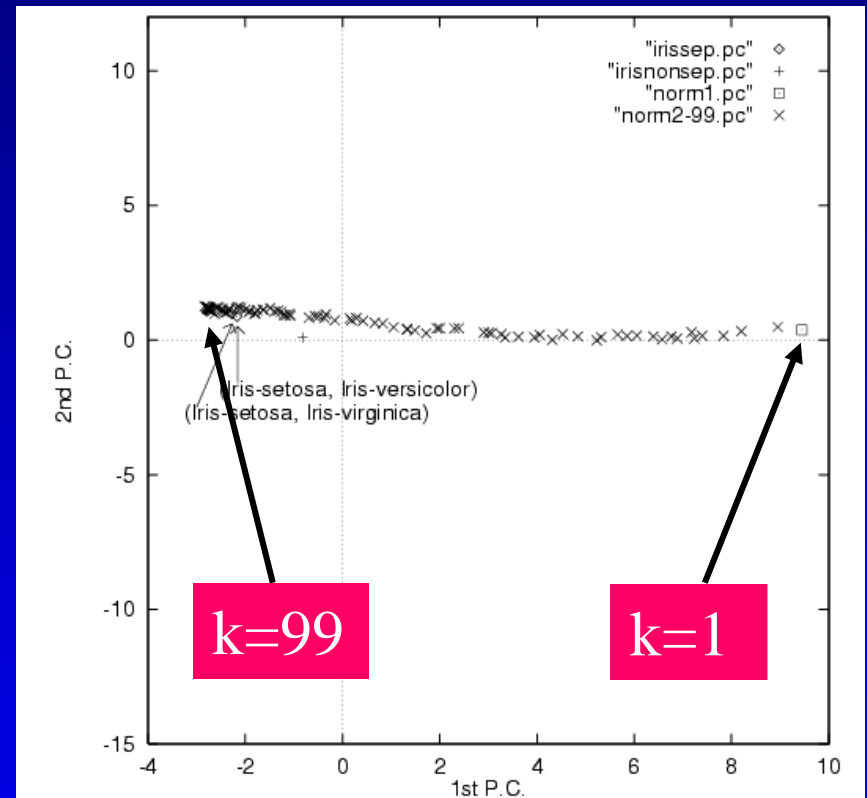
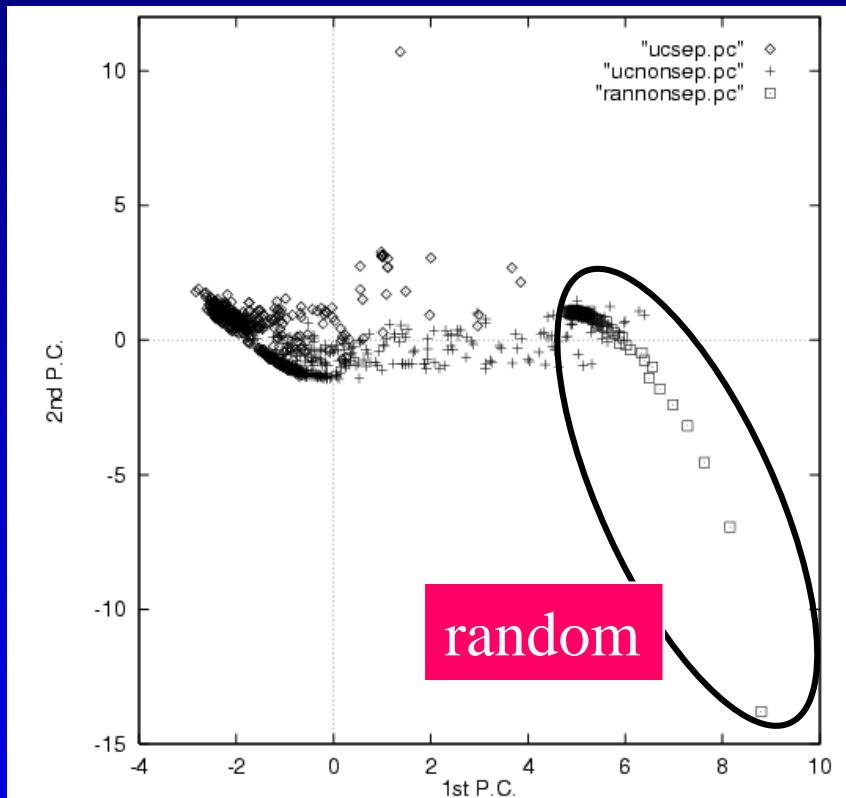
# Principle Component Analysis



# A Trajectory of Difficulty

1-dim, 2 classes, 100 pts/class,

Normal dist. stddev=30, mean= +k, -k





# What Else Can We Do?

Study effectiveness of these measures

Interpret problem distributions

- Find **clusters** in this space
- Determine **intrinsic dimensionality**

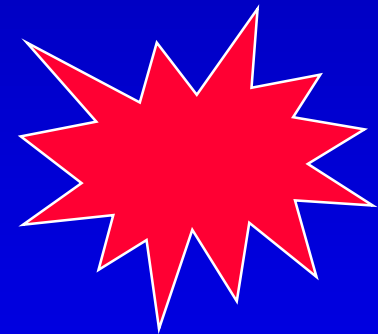
# What Else Can We Do?

Apply these measures to more problems

- Study **specific domains** with these measures
- Study **alternative formulations, sub-problems** induced by localization, projection, transformation

# What Else Can We Do?

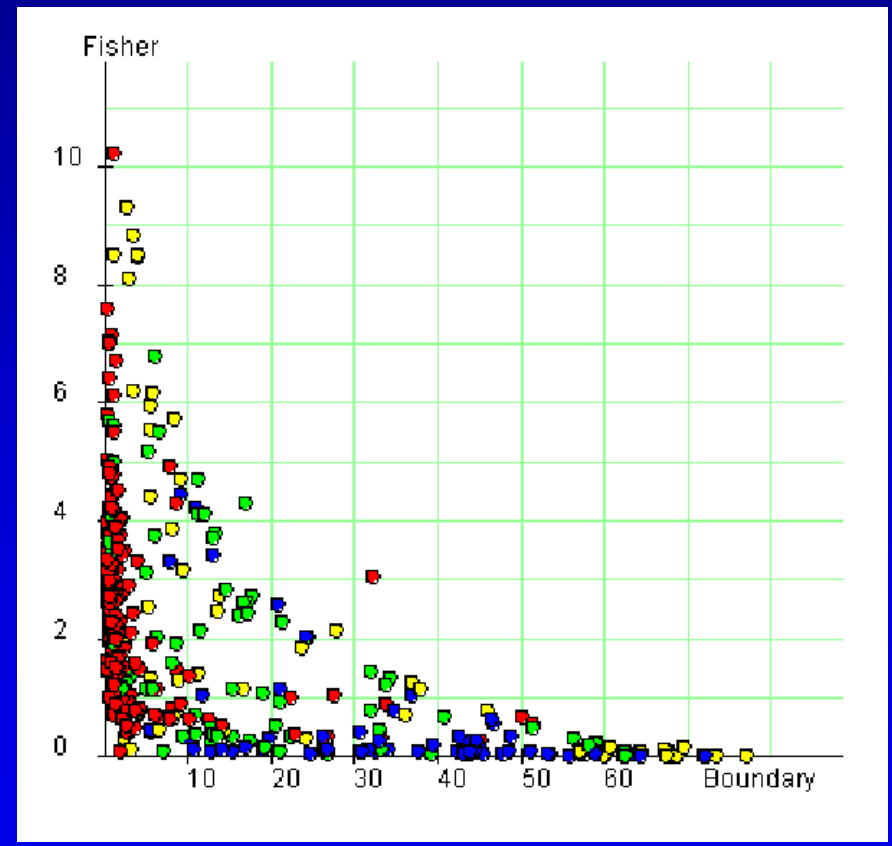
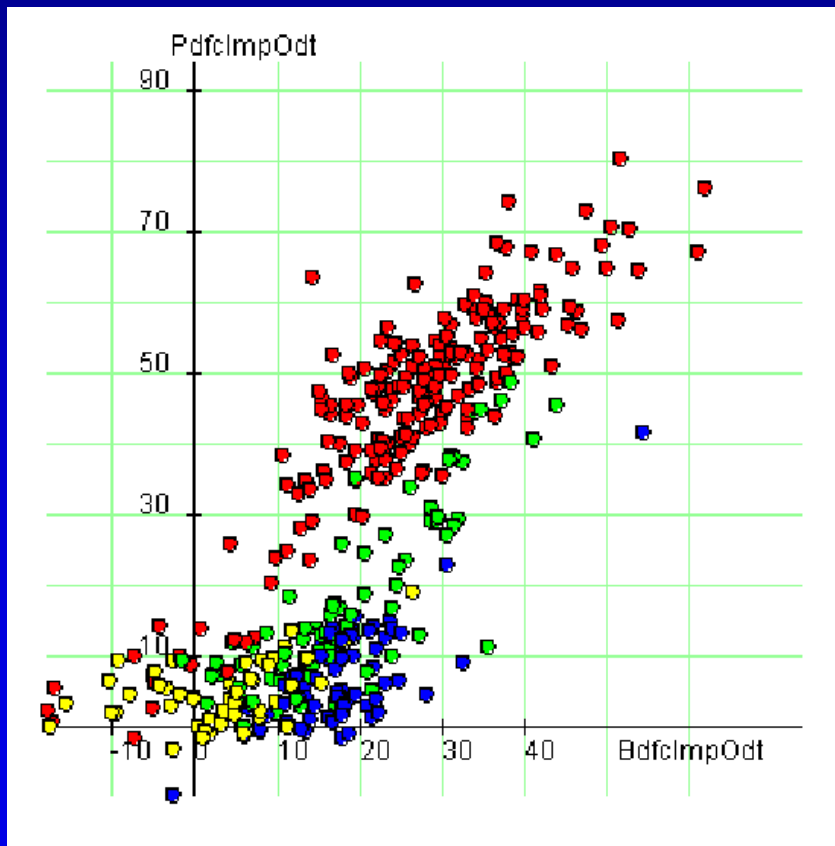
Relate complexity measures to  
**classifier behavior**



# Bagging vs Random Subspaces for Decision Forests

Subspaces better      same imp  
Subsampling better    no imp

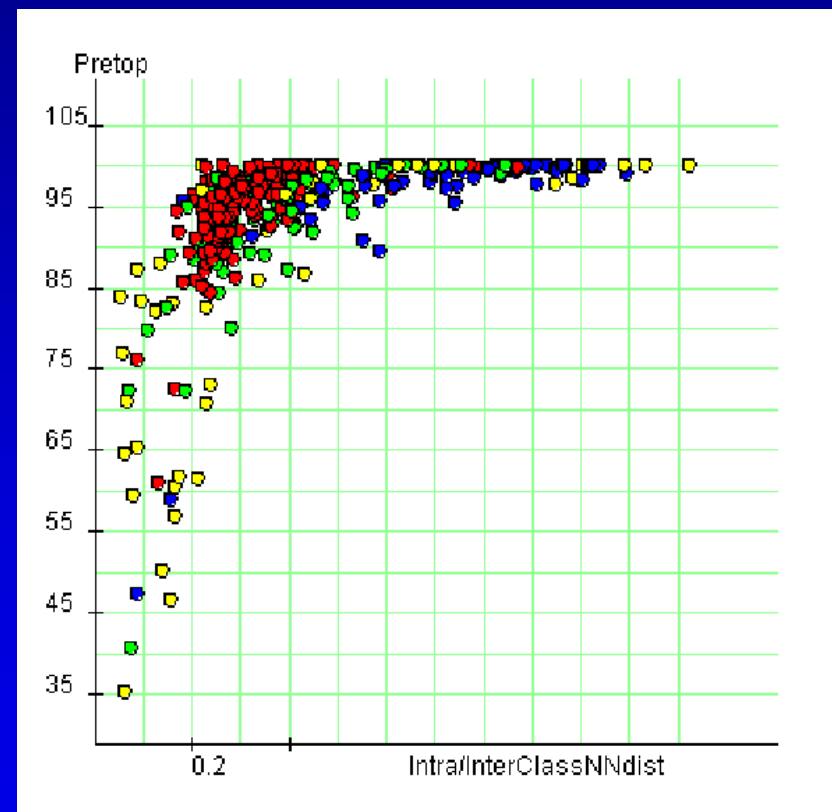
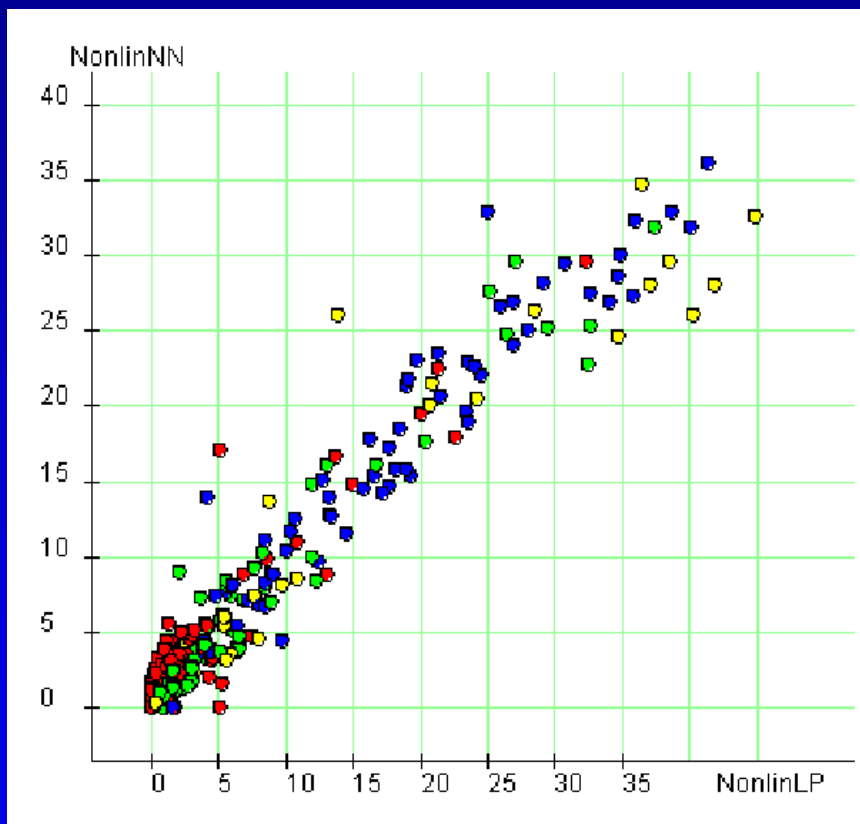
Fisher's discriminant ratio  
vs. length of class boundary



# Bagging vs Random Subspaces for Decision Forests

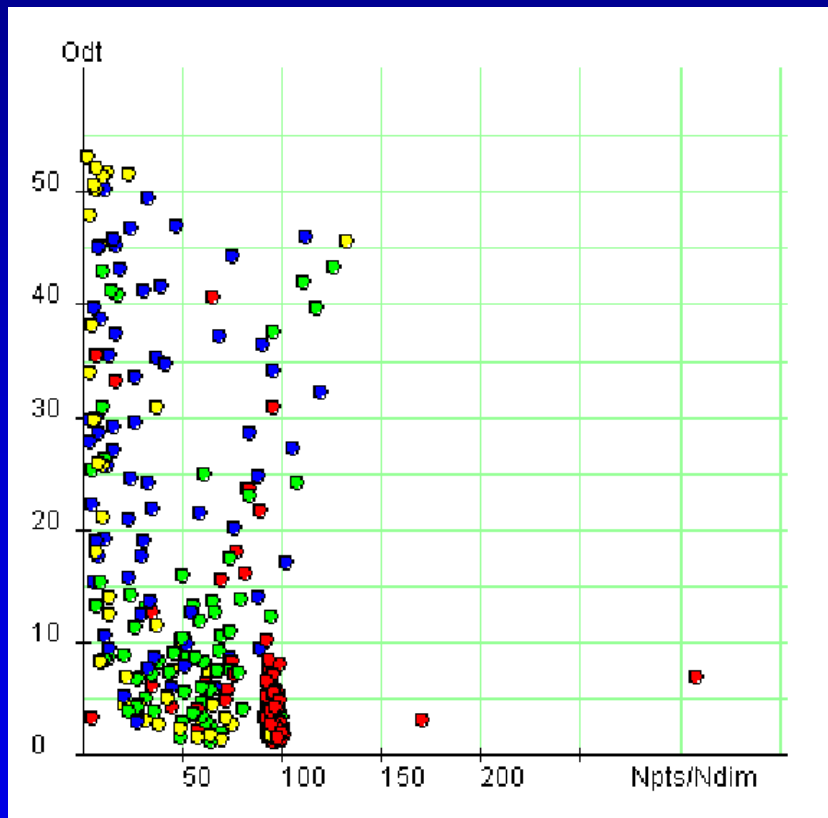
Nonlinearity, nearest neighbors  
vs. linear classifier

% Retained adherence subsets,  
vs. intra/inter class NN distances

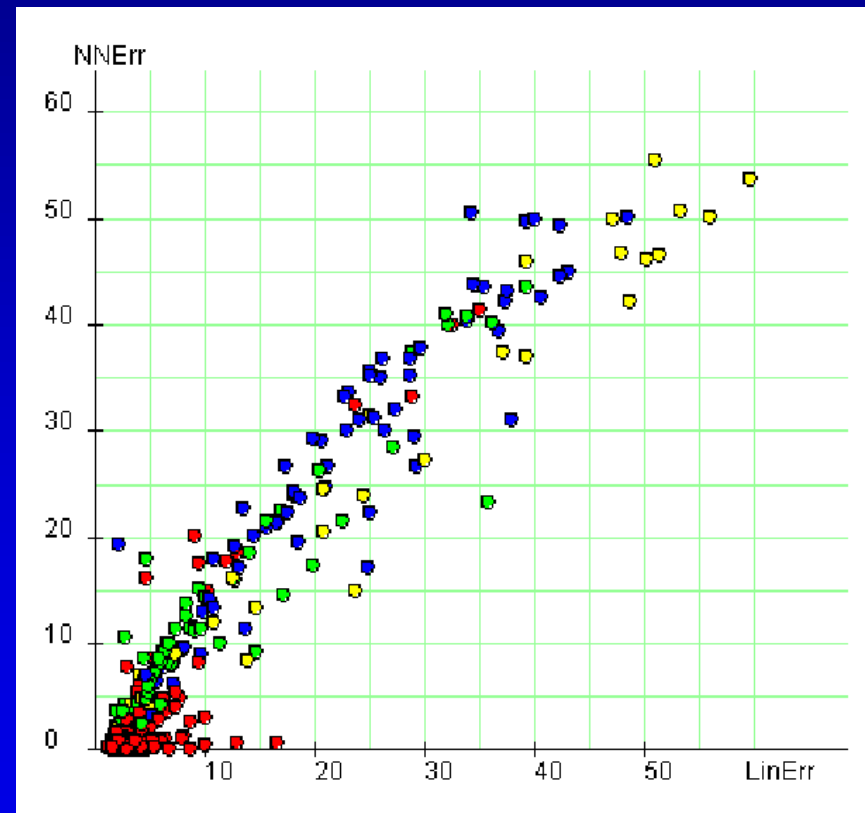


# Error Rates of Individual Classifiers

Error rate, single trees  
vs. sampling density



Error rate, nearest neighbors  
vs. linear classifier



# Observations

- Both types of forests are good for problems of various degrees of difficulty
- Neither is good for extremely difficult cases
  - many points on boundary
  - ratio of intra/inter class NN dist. close to 1
  - low Fisher's discriminant ratio
  - high nonlinearity of NN or LP classifiers
- Subsampling is preferable for sparse samples
- Subspaces is preferable for smooth boundaries

# Summary

- Real-world problems have different types of geometric characteristics
- Relevant measures can be related to classifier accuracies
- Data complexity analysis improves understanding of classifier or combination behavior
- Helpful for combination theories and practices



# Exploratory **Tools** Needed

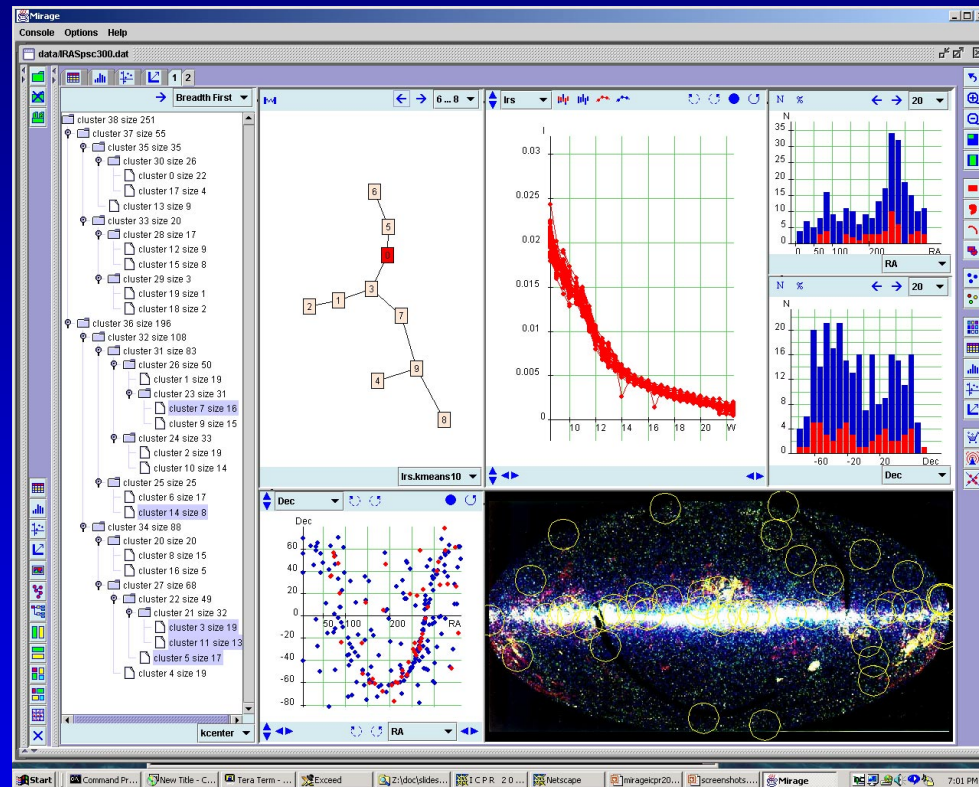
To study data or classifier geometry

To study

- correlations
- proximity structures between points
- correlations between proximity structures

# Exploratory Analysis of Proximities and Correlations Using **Mirage**

A walk on a cluster graph being tracked in other views.



Distributed at  
[www.bell-labs.com/project/mirage](http://www.bell-labs.com/project/mirage)

# Proximity Structures

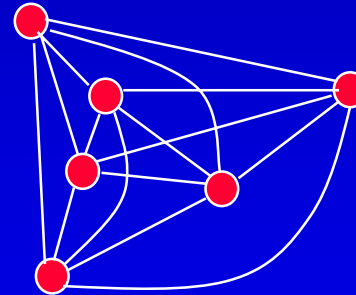
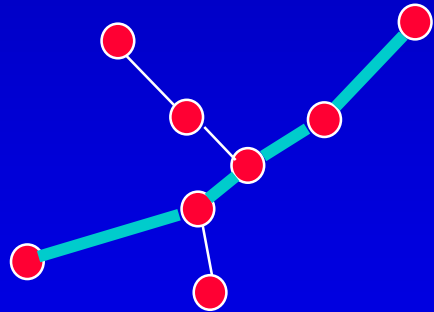
- $P = (S, G)$ 
  - $S$ : a set of subsets in dataset  $D$
  - $G$ : a weighted graph, weights represent proximity
- 1. Partitional structures:
  - $S$  partitions  $D$ ,  $G=(S,E)$
- 2. Hierarchical structures:
  - $G$  is a tree that splits  $D$ ,  $S$  contains all the nodes
- Traversals of the structures

# Correlation of Proximity Structures

**Continuity, Monotonicity, Linearity**

of dependencies, and

**Connectedness, Intrinsic dimensionality** of the changes



# Other Types of Proximity Structures

Proximity structures not resulting from clustering:

- Trivial structures: singletons, distances
- Degenerate structures: categorical features
- Structures correlated by construction:
  - e.g. CART & partitional structure on class labels

# Observation

Study of correlation between different proximity structures is fundamental to data analysis

This includes proximity between points, point sets, projected to different subspaces

# Addressing Curse of Dimensionality

No. of variables (NOT no. of objects) determines the mathematical difficulty of modeling

Combinatorial difficulty scales exponentially with no. of variables, but only linearly with no. of objects

Clustering in subspaces helps by divide-and-conquer

# Mirage

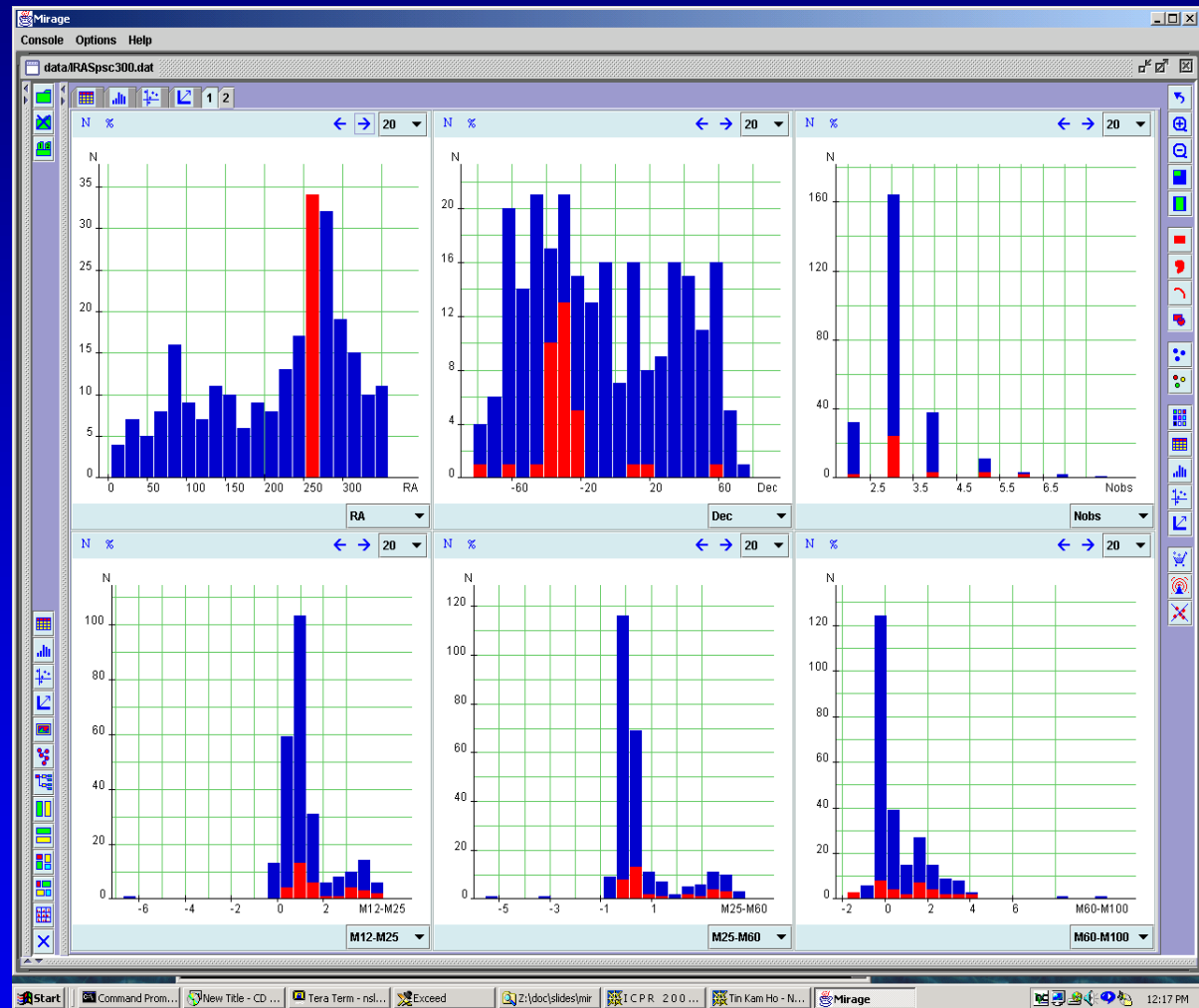
- Software tool for studying proximity in a data set, especially for measurements of multiple types
- Different treatments of individual subspaces
- Examination of data as isolated subsets or in context
- Heavy emphasis on interaction and intuitive manipulations



# Traversal of Partitional Structures

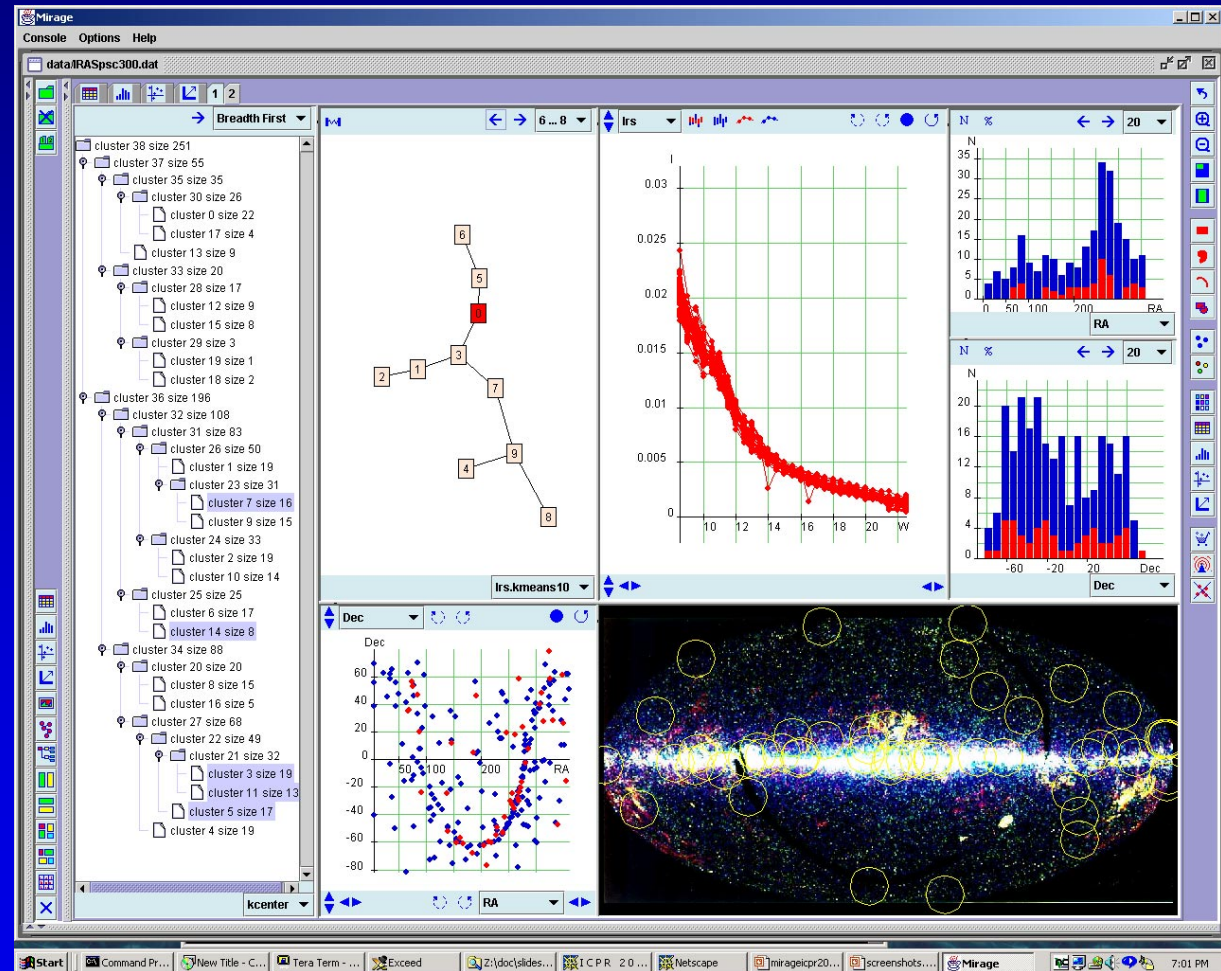
A walk on one histogram being tracked in the others.

Bins in a histogram give a simple partitional structure.

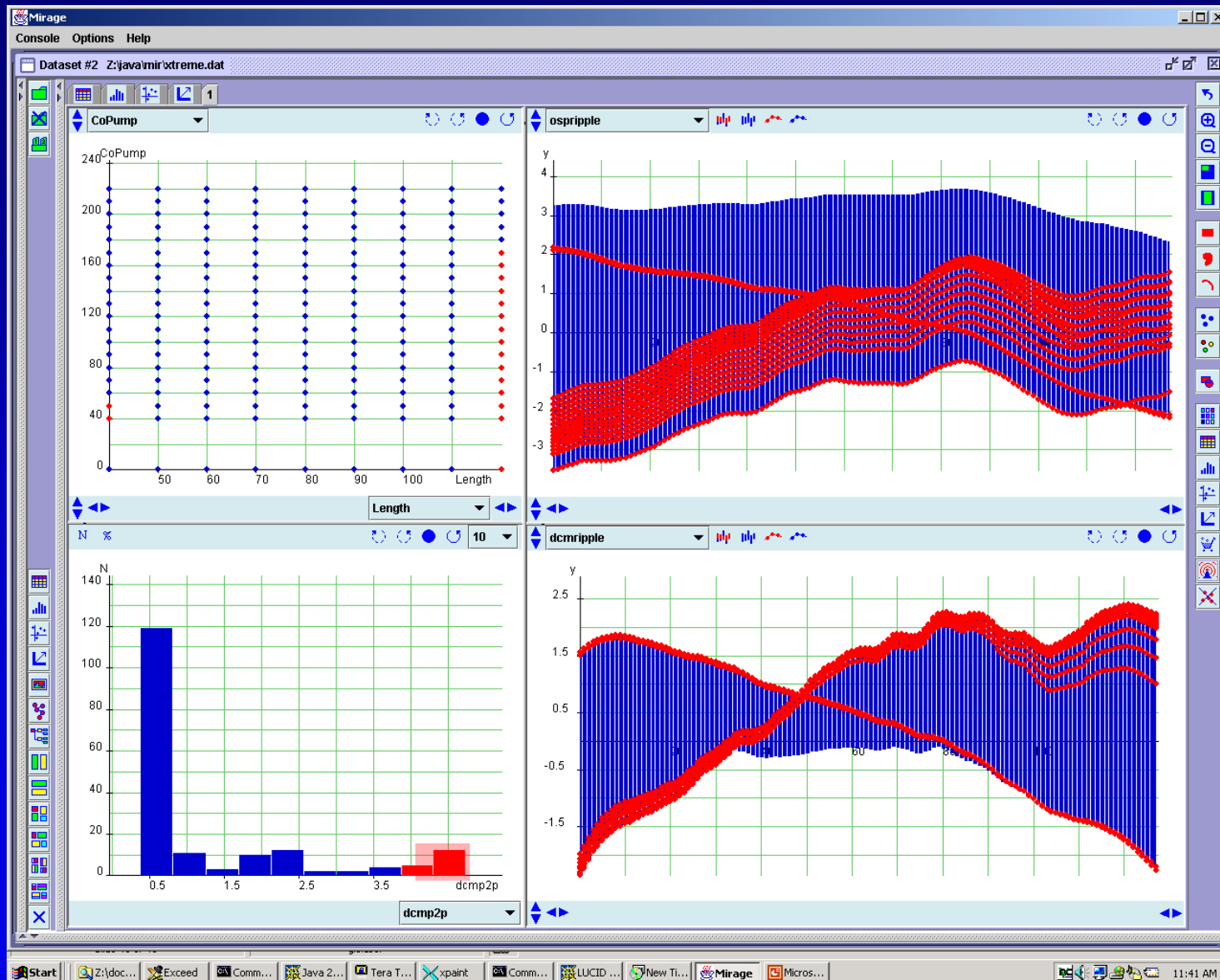


# Correlating with Other Views

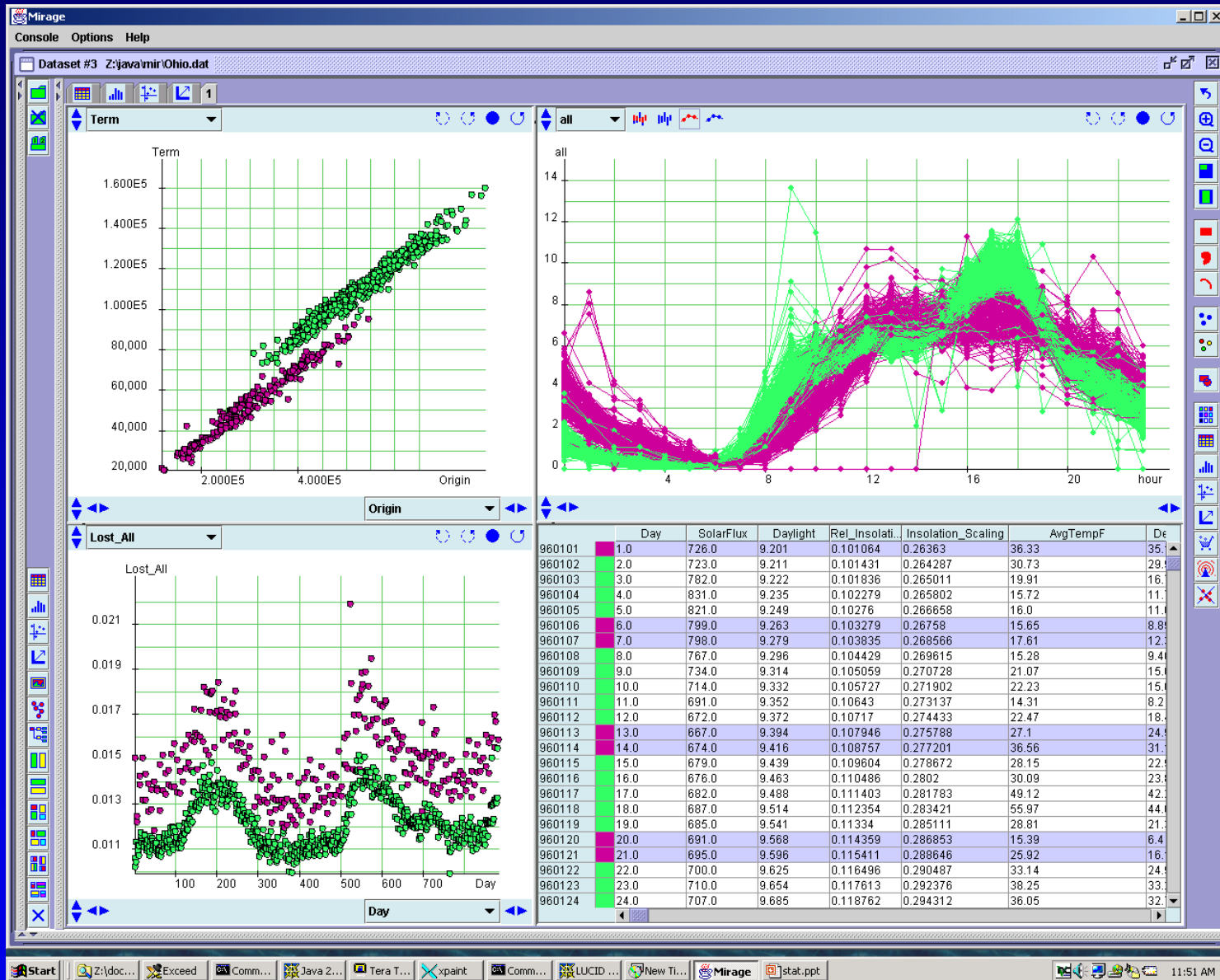
A walk on a cluster graph being tracked in other views.



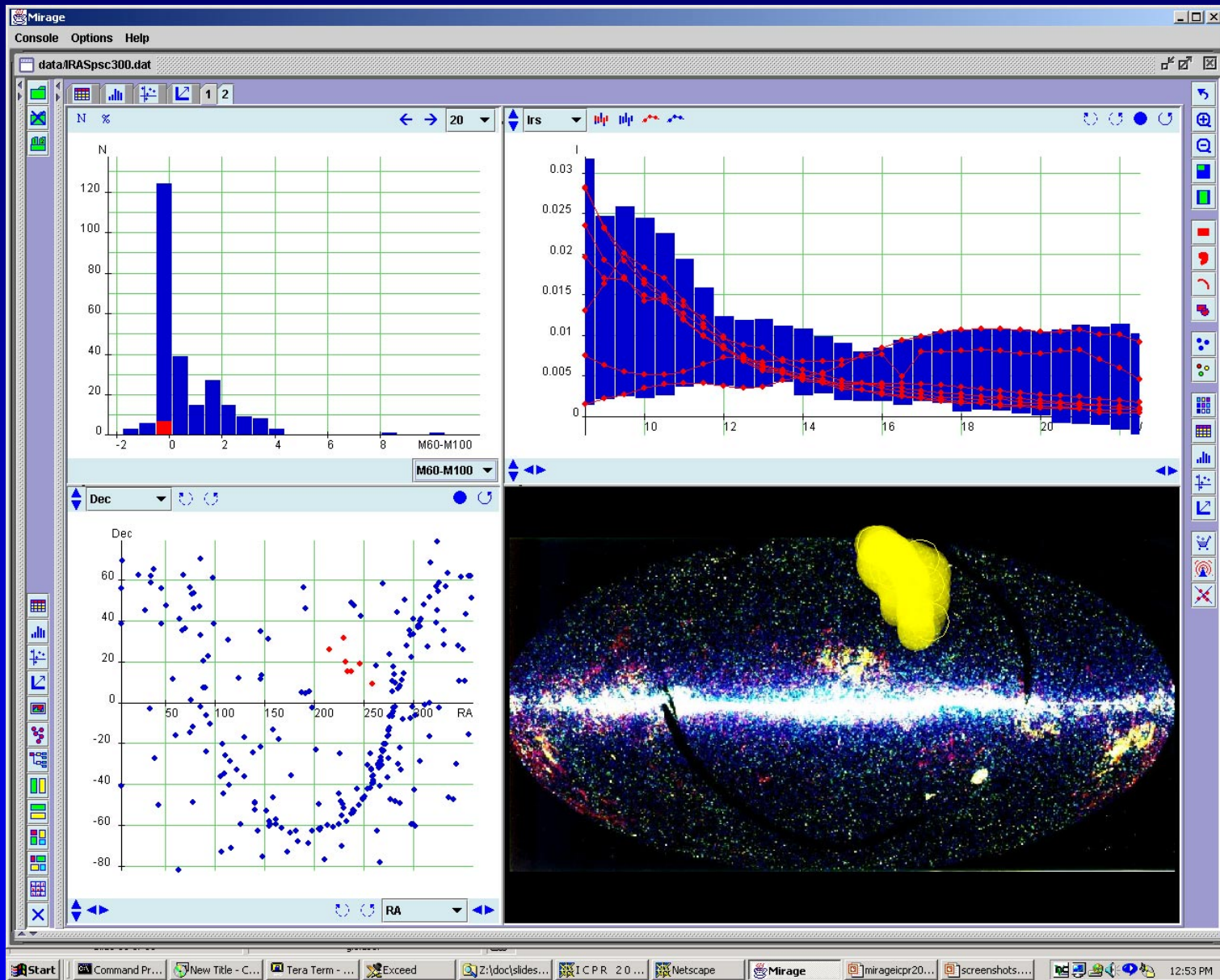
# Parameter exploration in a simulation analysis



# Correlation of clusters in one space with patterns in others



# Examining data located in a neighborhood in an image



# Mirage

Beta-test copy available at

[www.bell-labs.com/project/mirage](http://www.bell-labs.com/project/mirage)

