Exploration of Parameter Spaces in the Catalog Domain

* * *

Clustering, Classification, and Other Data Exploration Tools

S. George Djorgovski

Lecture 3

Inaugural BRAVO Lecture Series,
São José dos Campos, July 2007

An Overview

- Making discoveries in VO: how and where?
- Exploration of parameter spaces: a generic VO science case
  - Discovery of rare types of objects (QSOs, BDs, peculiar objects…)
- Clustering analysis: general issues
  - Automated and objective object classification
  - Star-galaxy separation
    - ✧ Especially in synoptic sky surveys
  - Scalability of clustering algorithms
- Statistics for data exploration
  - The VOStat package
  - Correlation searches and multivariate statistics
- Visualization!
- Some useful websites and resources

How and Where are Discoveries Made?

- **Conceptual Discoveries:** e.g., Relativity, QM, Strings/Branes, Inflation … *Theoretical, may be inspired by observations*
- **Phenomenological Discoveries:** e.g., Dark Matter, QSOs, GRBs, CMBR, Extrasolar Planets, Obscured Universe …

Empirical, inspire theories, can be motivated by them

New Technical Capabilities → (IT/VO) Observational Discoveries ↔ (VO) Theory

Phenomenological Discoveries:

- Pushing along some parameter space axis ← VO useful
- Making new connections (e.g., multi-λ) ← **VO critical!**
- Finding rare instances of new phenomena ← VO very useful

Understanding of complex astrophysical phenomena requires complex, information-rich data (and simulations?)

VO Science: Some General Features

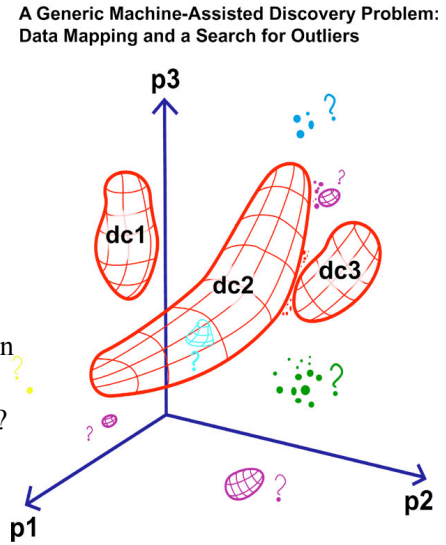
- Data fusion tends to reveal new knowledge, connections
- Massive data sets naturally lead to statistical approaches
  - Data mining algorithms are often just algorithmic expressions of proper statistics
  - Large data sets can enable discovery of rare (new?) instances
- Almost always the problem is reduced to data exploration in some parameter space of source attributes
  - Data mining in image domain is possible - but in most cases one ends up with some kind of image segmentation (e.g., object detection) and parametrization, so the problem reduces to the catalog domain
- Good visualization has to be a key part of the data mining process - it connects the data with our intuition, understanding
  - Effective hyper-dimensional visualization is a *huge* problem

Exploration of Parameter Spaces

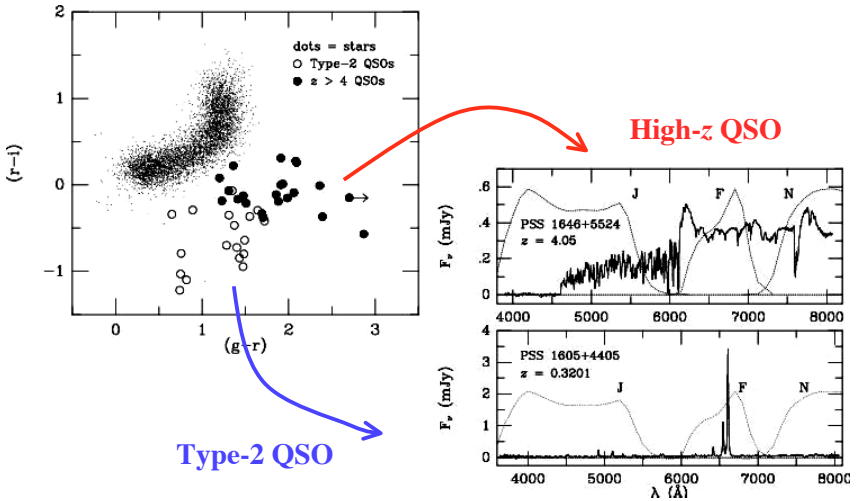
- In most surveys, image/pixel data are reduced to ***catalogs***, using some kind of a processing pipeline, which detects sources and measures their ***attributes/parameters*** on the basis of position, flux, and light distribution
  - Nowadays typically we measure up to a few hundred parameters per source per survey
  - Typically we detect ~ 10^8 - 10^9 sources per survey
  - Data federation from different surveys increases these numbers
- Data are then ***vectors in parameter spaces*** of hundreds of dimensions
- They generally do not populate this space uniformly, but define ***clusters and correlations***
- Their description and analysis leads to ***scientific discoveries***, and is also useful for ***quality control*** purposes

Parameter Spaces: Clustering Analysis

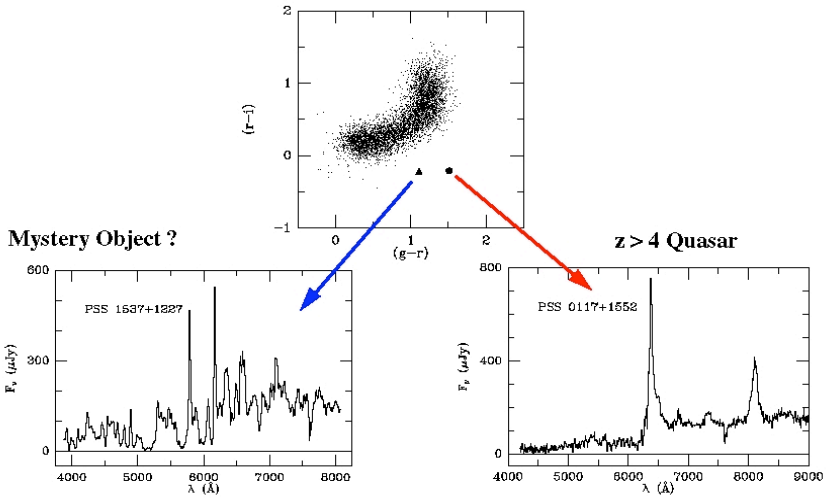
- How many different types of objects are there?
- Which ones are identifiable with known, physically distinct types (e.g., stars, galaxies, quasars at different redshifts, etc.)?
- Are there rare and/or previously unknown classes, seen as outliers or distinct classes?
- Are there intermediate or transition types?
- Are there gaps (negative clusters)?
- Anomalies possibly indicative of problems with the data?



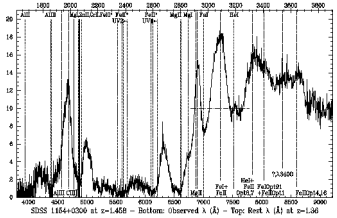
An Example: Discoveries of High-Redshift Quasars and Type-2 Quasars in DPOSS



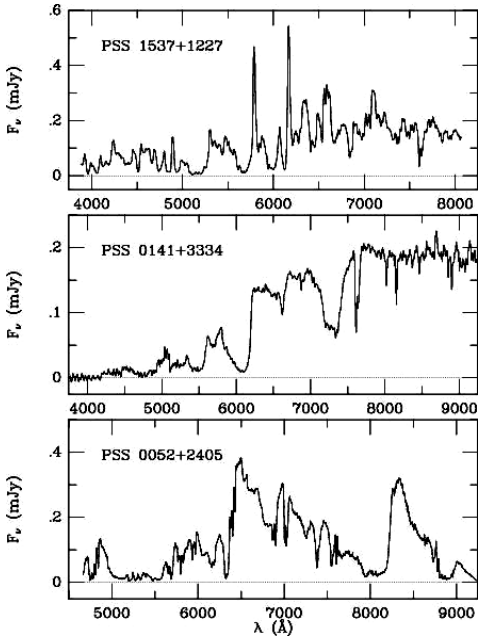
But Sometimes You Find a Surprise…



Spectra of **Peculiar Lo-BAL (Fe) QSOs** Discovered in DPOSS (also FIRST, SDSS):



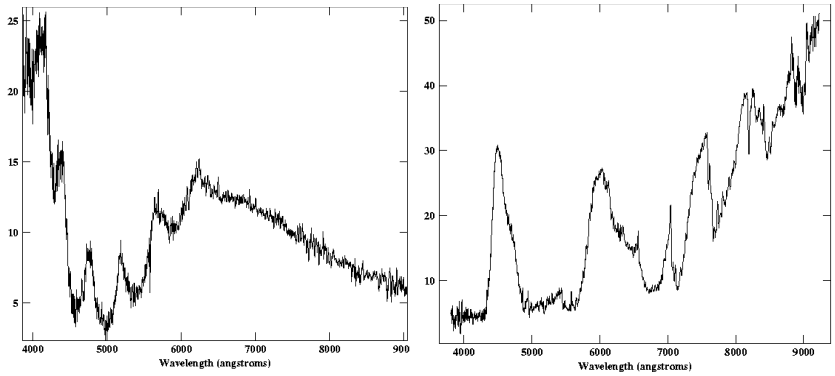
No longer a mystery, but a rare subspecies



More Peculiar Objects From SDSS

DQ White Dwarf

Highly peculiar CV

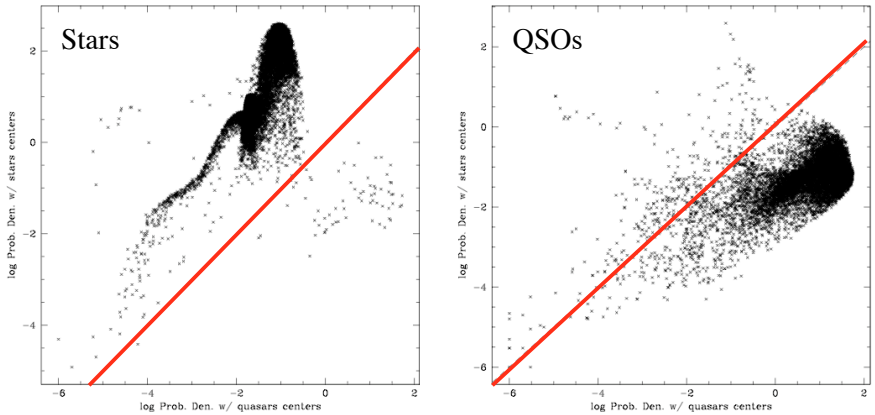


(Fan et al.)

An Improved Star-QSO Classifier

(Nichol, Connolly, et al. 2001)

Evaluate the probability density functions for known stars and known QSOs in the SDSS ugriz 4-D color space, using a multi-Gaussian mixture model; use it to evaluate likelihood that a given object is a star or a QSO.

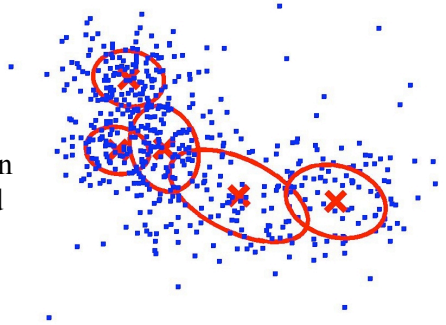


Clustering and Classification

- This is a generic problem for VO analysis in the catalog domain
- Many (many!) statistical and machine-learning methods exist

→ ***Make friends among the statisticians and computer scientists!***

- Generally, this is a very non-trivial task, especially with noisy, missing, or heterogeneous data
- The goal is to associate a probability for each data point belonging to any given cluster or class
- In other words, one seeks an optimal (in some statistical sense) multi-component probability density distribution which describes the data - and its functional form may be unknown or undefined



Clustering analysis or automated classification is a key task for VO science. There are many good tools out there, but most need to be developed further for our needs (mainly the scalability issues).

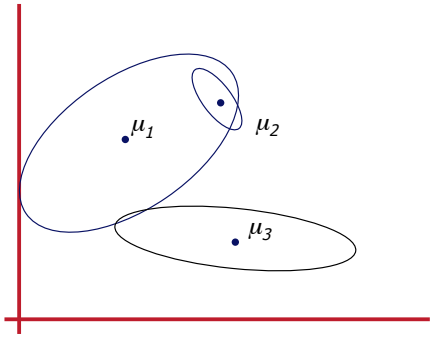
Inputs → Inference Engine Learn	$P(E_1\|E_2)$	Joint DE, Bayes Net Structure Learning
Inputs → Classifier	Predict category	Dec Tree, Sigmoid Perceptron, Sigmoid N.Net, Gauss/Joint BC, Gauss Naïve BC, N.Neigh, Bayes Net Based BC, Cascade Correlation
Inputs → Density Estimator	Prob-ability	Joint DE, Naïve DE, Gauss/Joint DE, Gauss Naïve DE, Bayes Net Structure Learning, GMMs
Inputs → Regressor	Predict real no.	Linear Regression, Polynomial Regression, Perceptron, Neural Net, N.Neigh, Kernel, LWR, RBFs, Robust Regression, Cascade Correlation, Regression Trees, GMDH, Multilinear Interp, MARS

(from Moore 2002)

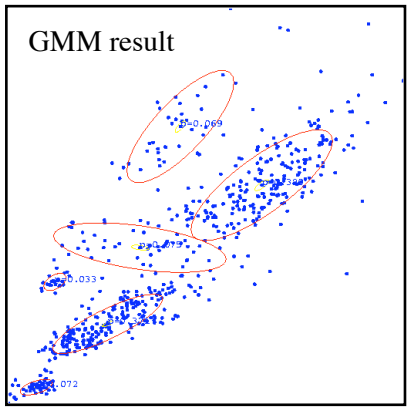
Simple Clustering Analysis:
Gaussian Mixture Modeling

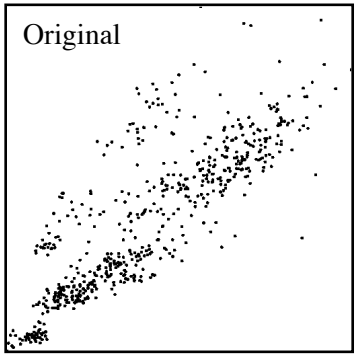
Data points are distributed in some *N*-dimensional parameter space, x_j , $j = 1, \ldots N$

- There are k clusters, w_i , $i = 1, \ldots, k$, where the **number of clusters, *k*,** may be either given by the scientist, or derived from the data themselves
- Each cluster can be **modeled as an *N*-variate Gaussian** with mean μ_i and covariance matrix S_i (NB: in the real life, things are seldom Gaussian…)
- Each data point has an association probability of belonging to each of the clusters, P_i

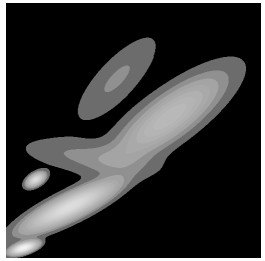


An Example
(from Moore et al.)



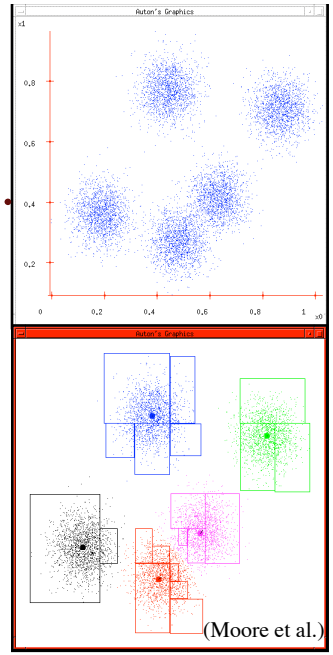


Model density distribution ⇒



A Popular Technique: K-Means

- Start with ***k*** random cluster centers
- Assume a data model (e.g., Gaussian)
  - In principle, it can be some other type of a distribution
- Iterate until it converges
  - There are many techniques; Expectation Maximization (EM) is very popular; multi-resolution *kd*-trees are great (Moore, Nichol, Connolly, et al.)
- Repeat for a different ***k*** if needed
- Determine the optimal ***k*** :
  - Monte-Carlo Cross-Validation
  - Akaike Information Criterion (AIC)
  - Bayesian Information Criterion (BIC)



In VO data sets: $D_D >> 1, D_S >> 1$

Data Complexity ➞ Multidimensionality ➞ Discoveries

But the bad news is …

The computational cost of clustering analysis:

K-means: $K \times N \times I \times \mathbf{D}$

Expectation Maximisation: $K \times N \times I \times \mathbf{D^2}$

Monte Carlo Cross-Validation: $M \times K_{max}^2 \times N \times I \times \mathbf{D^2}$

N = no. of data vectors, D = no. of data dimensions
K = no. of clusters chosen, K_{max} = max no. of clusters tried
I = no. of iterations, M = no. of Monte Carlo trials/partitions

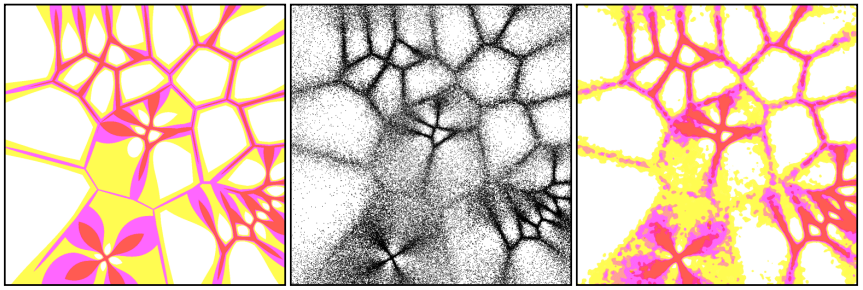
⇨ *Terascale (Petascale?) computing and/or **better algorithms***

Some dimensionality reduction methods do exist (e.g., PCA, class prototypes, hierarchical methods, etc.), but more work is needed

Probably the best (fastest) GMM techniques to date:

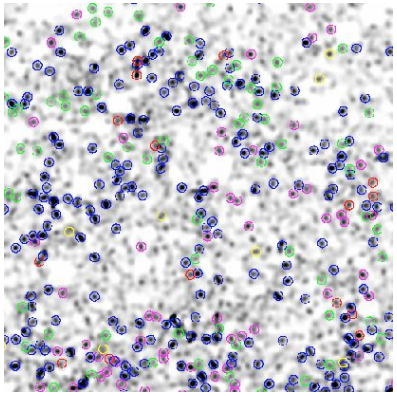
Multi-resolution *kd*-tree (*mrkd*) implementation of the EM method by Moore, Nichol, Connolly, et al. (PICA group) (see, e.g., astro-ph/0012333, 0007404, 0008187)

Voronoi Foam Model | Random data realization | *mrkd* (EM+AIC) restored



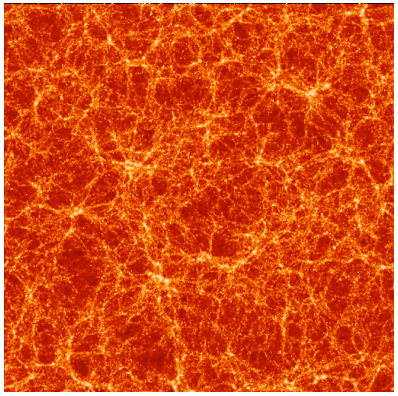
Examples of Challenges for Clustering Analysis from "Standard" Astronomical Clustering/LSS Analysis:

Clustering on a clustered background



DPOSS Clusters (Gal et al.)

Clustering with a nontrivial topology



LSS Numerical Simulation (VIRGO)

Then: Selection effects, missing data, non-Gaussianity…

Exploration of Parameter Spaces in the Catalog Domain (Source Attributes)

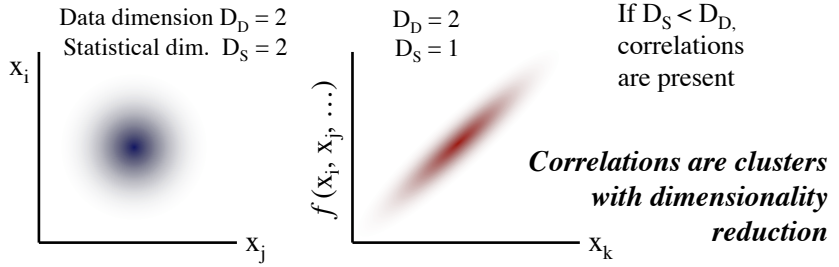
- Clustering Analysis (supervised and unsupervised):
  - How many different types of objects are there?
  - Are there any rare or new types, outliers?
- Multivariate Correlation Search:
  - Are there significant, nontrivial correlations present in the data?

Clusters vs. Correlations:

Astrophysics ➞ Correlations

Correlations ➞ reduction of the statistical dimensionality

Correlation Searches in Attribute Space

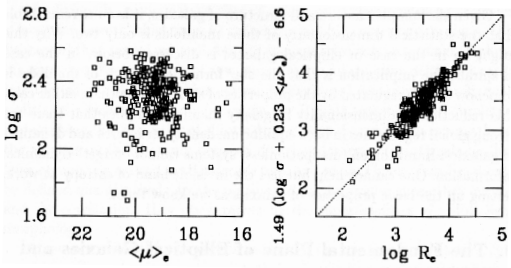


If $D_S < D_D$, correlations are present

Correlations are clusters with dimensionality reduction

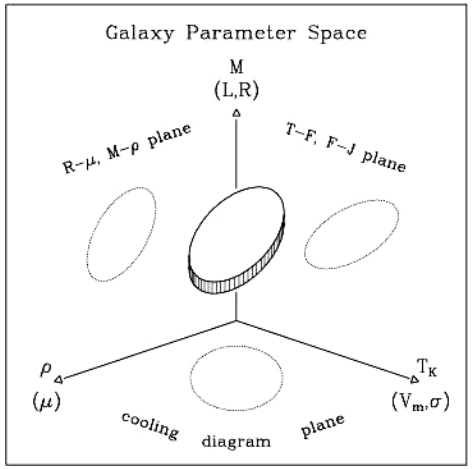
A real-life example:

"Fundamental Plane" of elliptical galaxies, a set of bivariate scaling relations in a parameter space of ~ 10 dimensions, containing valuable insights into their physics and evolution



Galaxy Parameter Space

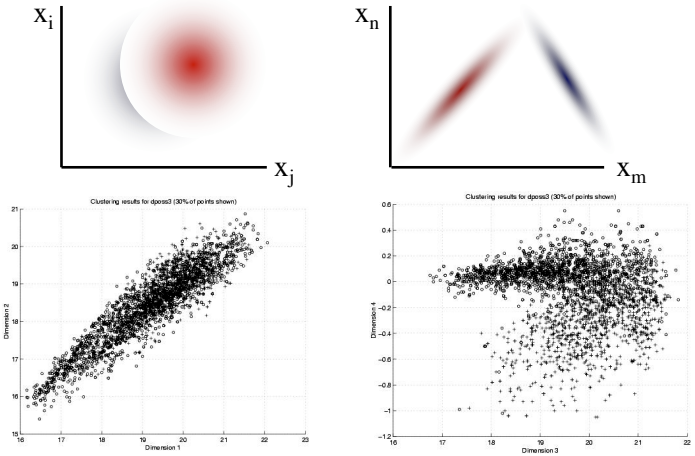
Most galaxy properties are connected by bivariate (D_s=2) scaling relations in a $D_p \geq 3$ parameter space of independent observables or physical properties (e.g., measures of "size", "density", and "temperature"). Projections onto the individual coordinate planes produce correlations with a large intrinsic scatter ➟ loss of information.



However, some properties do not participate in these correlations, retaining a high D_s: no correlations ➟ no physical insight.

Useful *vs.* "Useless" Parameters:

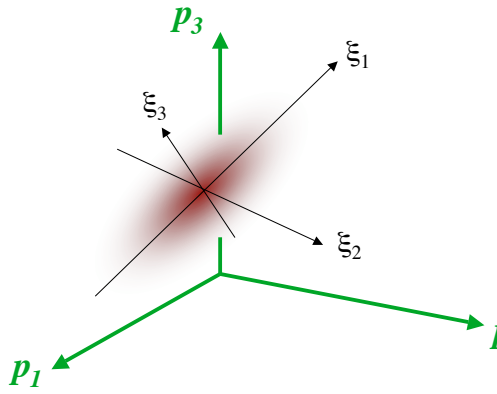
Clusters (classes) and correlations may exist/separate in some parameter subspaces, but not in others - the trick is to find which ones …



Principal Component Analysis

An essential dimensionality reduction tool

Solving the eigen-problem of the data hyperellipsoid in the parameter space of measured attributes



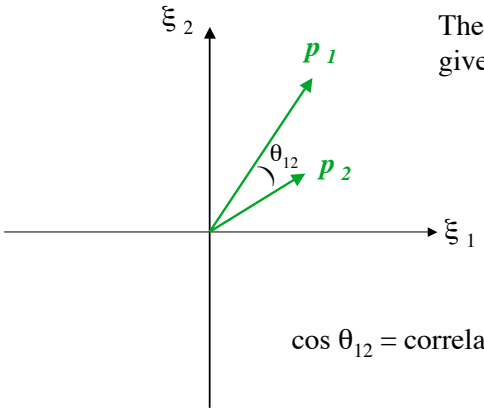
p_i = observables ($i = 1, \ldots D_{data}$)

ξ_j = eigenvectors, or principal axes of the data hyperellipsoid

e_j = eigenvalues, or amplitudes of ξ_j ($j = 1, \ldots D_{stat}$)

Correlation Vector Diagrams:

Projections of the data and observable axes onto the planes defined by the eigenvectors



The natural coordinate system given by the data themselves

$$\xi_i = a_{i1}\, p_1 + a_{i2}\, p_2 + \ldots$$
$$p_i = b_{i1}\, \xi_1 + b_{i2}\, \xi_2 + \ldots$$

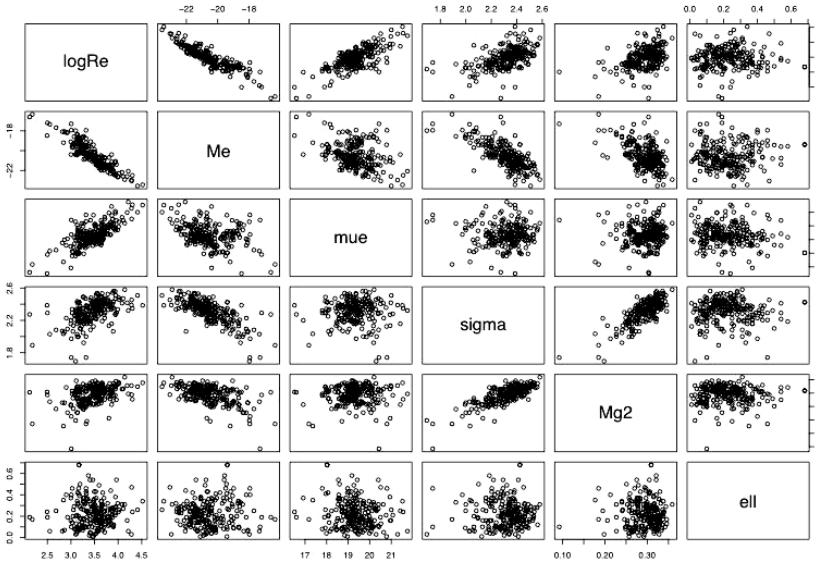
$\cos\theta_{12}$ = correlation coef. of p_1 and p_2

An Example, Using VOStat

Here is a data file, with 6 observed and 5 derived quantities (columns) for a few hundred elliptical galaxies (rows, data vectors):

```
# Ellipticals from the Djorgovski et al. survey
#
 logRe    M_e    mu_e   sigma    Mg2    M/L   logM    rho_M   rho_L   f_eff  ell   GalID
#
 3.863  -22.35  19.70  2.479  0.336  8.98  11.36  -0.847  -1.546  -0.205  .06   1016
 3.442  -20.64  19.01  2.310  0.316  8.78  10.61  -0.344  -0.970   0.806  .29   1052
 3.943  -22.46  19.78  2.468  0.325  8.90  11.42  -1.030  -1.742  -0.354  .23   1060
 3.282  -19.65  19.38  2.299  0.246  9.07  10.42  -0.045  -0.883   1.137  .17   1172
 3.509  -20.53  19.53  2.315  0.297  8.93  10.68  -0.467  -1.214   0.667  .24   1199
 3.457  -20.55  19.29  2.322  0.297  8.90  10.64  -0.349  -1.050   0.764  .22   1199
 3.463  -20.75  18.55  2.412  0.305  8.78  10.83  -0.181  -0.988   0.662  .54   1209
 3.066  -18.60  19.16  2.207  0.301  9.01  10.02   0.204  -0.655   1.661  .29   1339
 3.132  -18.66  19.36  2.158  0.282  8.93   9.99  -0.027  -0.831   1.578  .34   1351
 3.141  -18.99  19.43  2.273  0.310  9.18  10.23   0.185  -0.726   1.445  .09   1374
 3.477  -19.41  20.78  2.125  0.257  9.08  10.27  -0.784  -1.565   0.921  .01   1379
 3.526  -21.02  19.22  2.396  0.313  8.95  10.86  -0.339  -1.069   0.552  .18   1395
 3.257  -20.29  18.71  2.491  0.334  9.21  10.78   0.389  -0.552   0.995  .09   1399
 3.265  -20.06  18.57  2.213  0.279  8.59  10.23  -0.184  -0.670   1.257  .37   1403
 3.180  -20.16  18.42  2.353  0.317  8.89  10.43   0.266  -0.376   1.287  .12   1404
 3.552  -20.97  19.42  2.438  0.327  9.09  10.97  -0.307  -1.167   0.458  .16   1407
```

Pairwise Plots for Independent Observables



Their Correlation Matrix:

	logRe	Me	mue	sigma	Mg2	ell
logRe	1.00	-0.90	0.73	0.53	0.41	0.03
Me	-0.90	1.00	-0.38	-0.74	-0.54	0.03
mue	0.73	-0.38	1.00	-0.01	0.04	-0.13
sigma	0.53	-0.74	-0.01	1.00	0.79	-0.01
Mg2	0.41	-0.54	0.04	0.79	1.00	0.00
ell	0.03	0.03	-0.13	-0.01	0.00	1.00

You can learn a lot just from the inspection of this matrix, and comparison with the pairwise (bivariate) plots …

Now Let's Do the Principal Component Analysis (PCA):

Principal Component Analysis(m) for logRe M_e mu_e sigma Mg2 :

```
Importance of components:
                          PC1 PC2   PC3    PC4    PC5
Standard deviation        1.4 0.8 0.090 4e-02 2e-02
Proportion of Variance    0.8 0.2 0.003 6e-04 2e-04
Cumulative Proportion     0.8 1.0 0.999 1e-00 1e+00
```

5 independent observables, but only **2** significant dimensions: the first 2 components account for all of the sample variance! The data sit on a plane in a 5-dim. parameter space: this is the Fundamental Plane of elliptical galaxies. Any one variable can be expressed as a combination of any 2 others, within errors.

PCA Results in More Detail

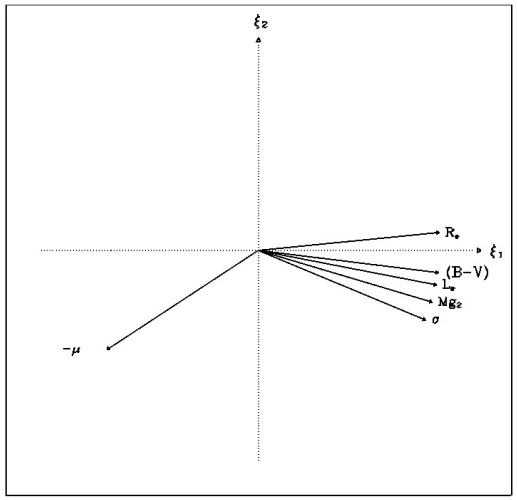
(This from a slightly different data set …)

Eigenvalues	As Percentages	Cumul. Percentages
3.1359	62.7189	62.7189
1.3574	27.1482	89.8671
0.3883	7.7670	97.6341
0.1110	2.2199	99.8540
0.0073	0.1460	100.0000

Eigenvectors and projections of parameter axes:

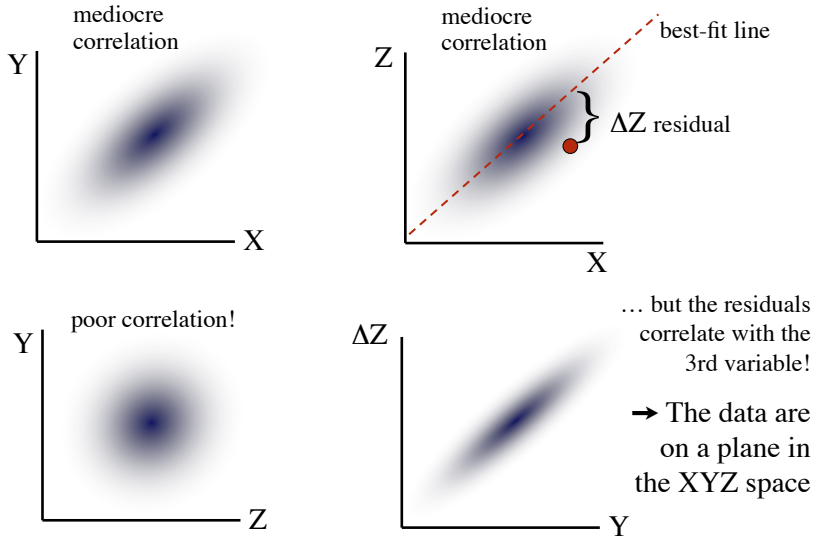
VBLE.	EV-1	EV-2	EV-3	EV-4	EV-5
logRe	-0.5119	0.3443	0.1649	0.1563	0.7535
M_e	0.5291	-0.0310	-0.5158	-0.3689	0.5630
<mu>e	-0.2764	0.6991	-0.4679	-0.3181	-0.3388
sigma	-0.4614	-0.4399	0.1187	-0.7610	0.0194
Mg2	-0.4108	-0.4453	-0.6883	0.3989	-0.0077

Now Project the Observable Axes Onto the Plane Defined by the Principal Eigenvectors:



Compare with the correlation matrix: Cosines of angles between parameter axes give the correlation coefficients.

Another Approach: Correlated Residuals

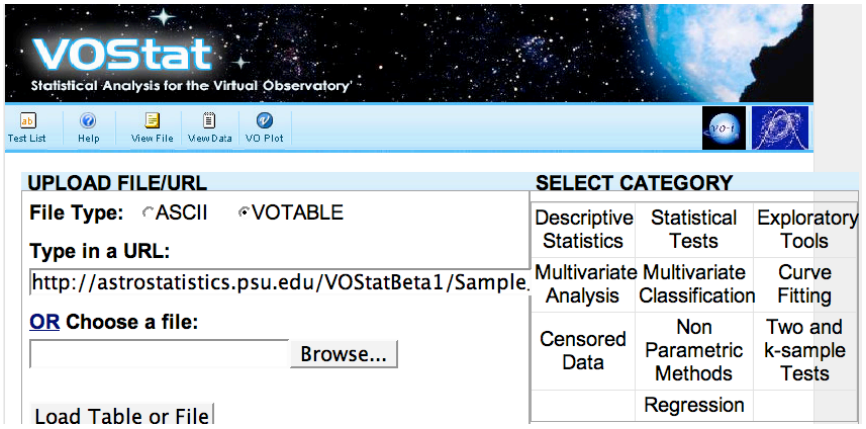


… but the residuals correlate with the 3rd variable!

→ The data are on a plane in the XYZ space

VOStat: Interactive Statistics Package for VO Applications

http://vostat.org



Automated Classification Techniques

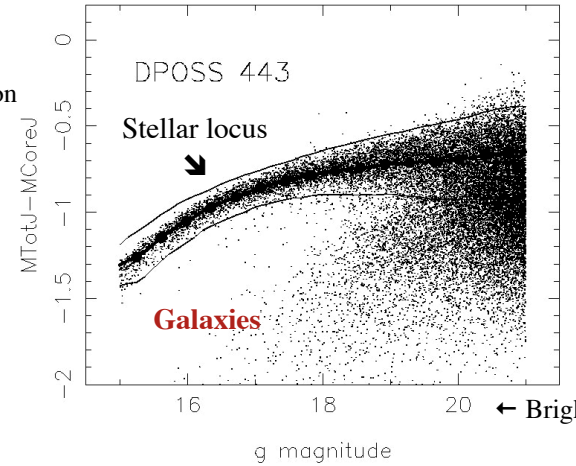
- Implementation of clustering algorithms in a machine-learning (ML) or AI setting
  - Examples: star-galaxy separation, automated galaxy morphology classification, stellar or galaxy spectral types, etc., etc.
- ***Supervised classifiers***: a set of learning examples is provided; the number of possible classes is known
  - Examples: Artificial Neural Nets (ANN), Decision Trees (DT), Support Vector Machines (SVM)…
- ***Unsupervised classifiers***: the program decides how many classes are needed to account for the diversity of the data, and classifies on the basis of the data

A Relatively Simple Classification Problem: Star-Galaxy Separation

- Important, since for most astronomical studies you want either stars (~ quasars), or galaxies; the depth to which a reliable classification can be done is the effective limiting depth of your catalog - not the detection depth
  - There is generally more to measure for a non-PSF object
- You'd like to have an automated and objective process, with some estimate of the accuracy as a *f (mag)*
  - Generally classification fails at the faint end
- Most methods use some measures of light concentration vs. magnitude (perhaps more than one), and/or some measure of the PSF fit quality (e.g., χ^2)
- For more advanced approaches, use some ***machine learning method, e.g., neural nets or decision trees***

Typical Parameter Space for S/G Classif.

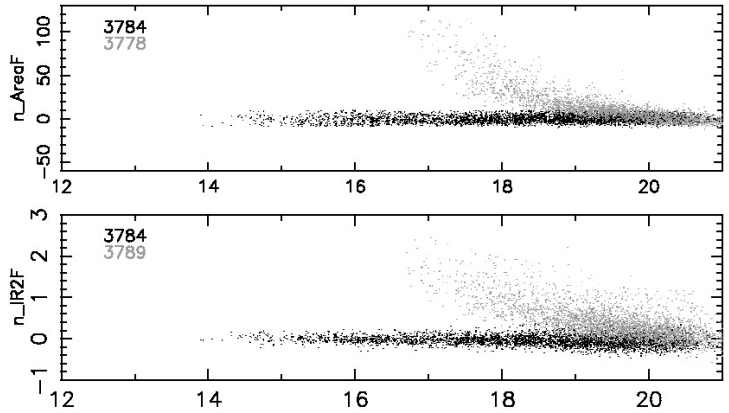
A measure of image concentration



(From DPOSS)

A set of such parameters can be fed into an automated classifier (ANN, DT, …) which can be trained with a "ground truth" sample

Star/Galaxy Classification Parameter Spaces: Normalized By The Stellar Locus



Then a set of such parameters can be fed into an automated classifier (ANN, DT, …) which can be trained with a "ground truth" sample

Automated Star-Galaxy Classification: Artificial Neural Nets (ANN)

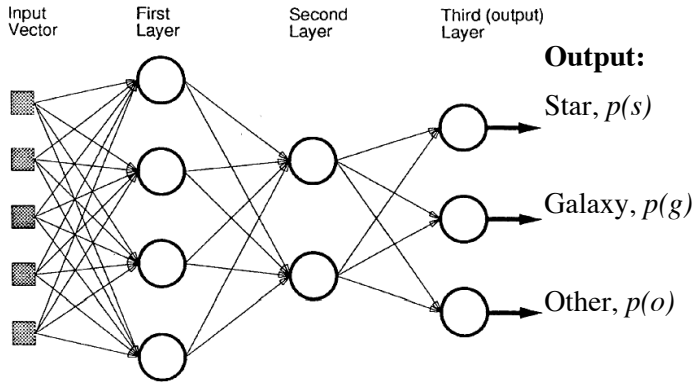
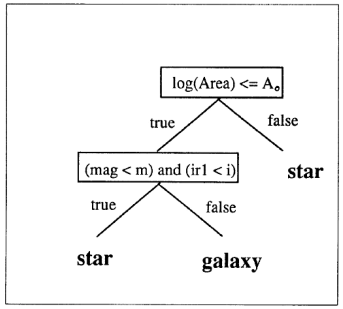


FIG. 6. Schematic illustration of a network with an input vector of length five, four nodes in the first layer, two nodes in the second layer, and three in the output layer. As a shorthand, such a network can be written as (5:4,2,3).

(Odewahn et al. 1992)

Automated Star-Galaxy Classification: Decision Trees (DTs)



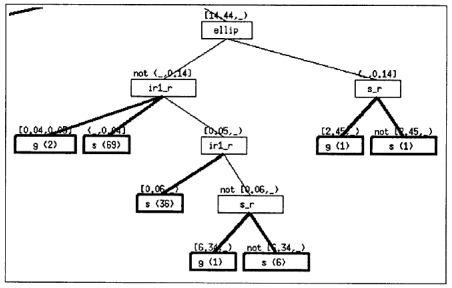


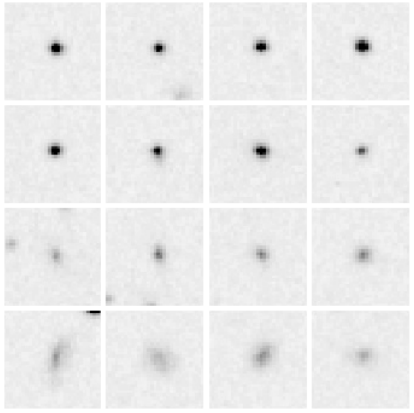
FIG. 2. A portion of a much larger actual decision tree generated by the O-Btree algorithm for performing star/galaxy classification. The interval appearing above each node indicates the range in value of the attribute specified in the node above that an object must meet for it to pass along that branch. The dark branches lead to actual classifications. The number in parentheses within each leaf indicates the number of training examples classified correctly at that node.

FIG. 1. In this sample decision tree, one starts at the top node(root), following the appropriate path to a final leaf (class) based upon the truth of the assertion at each node.

(Weir et al. 1995)

Automated Star-Galaxy Classification: Unsupervised Classifiers

No training data set - the program decides on the number of classes present in the data, and partitions the data set accordingly.



Star

Star+fuzz

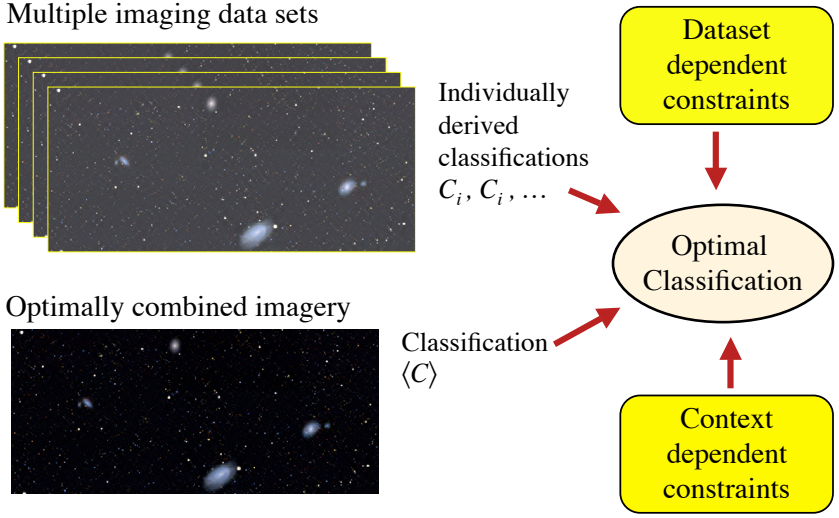
Gal1 (E?)

Gal2 (Sp?)

An example: AutoClass *(Cheeseman et al.)* Uses Bayesian approach in machine learning (ML).

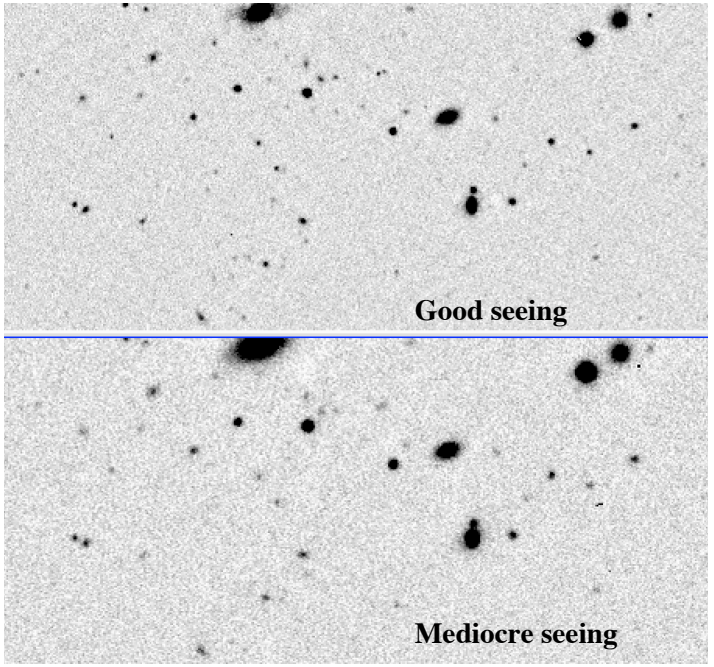
This application from DPOSS *(Weir et al. 1995)*

Star-Galaxy Classification:
The Next Generation



One key external constraint is the "seeing" quality for multiple imaging passes

(quantifiable e.g., as the PSF FWHM)

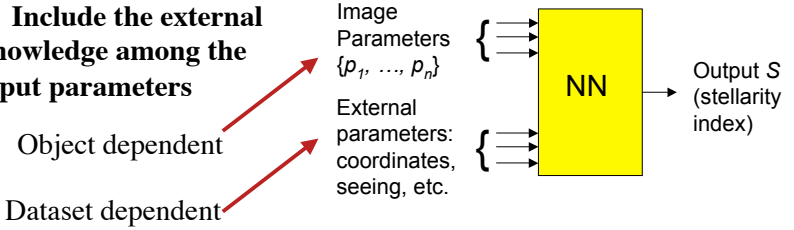


How to Incorporate the External or A Priori (Contextual) Knowledge?

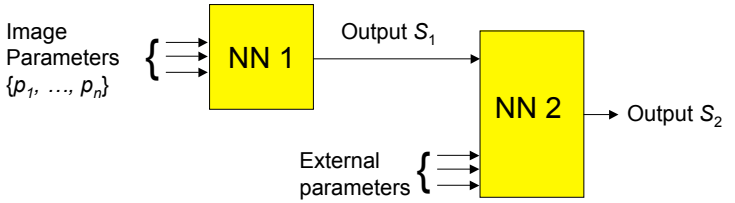
- Examples: seeing and transparency for a given night; direction on the sky, in Galactic coordinates; continuity in the star/galaxy fraction along the scan; etc.
- Still an open problem in the machine learning
- In principle, it should lead to an improved classification
- The problem occurs both in a "single pass" classification, and in combining of multiple passes
- In machine learning approaches, must somehow convert the external or a priori knowledge into classifier inputs - but the nature of this information is qualitatively different from the usual input (individual measurement vectors)

Two Approaches Using ANN:

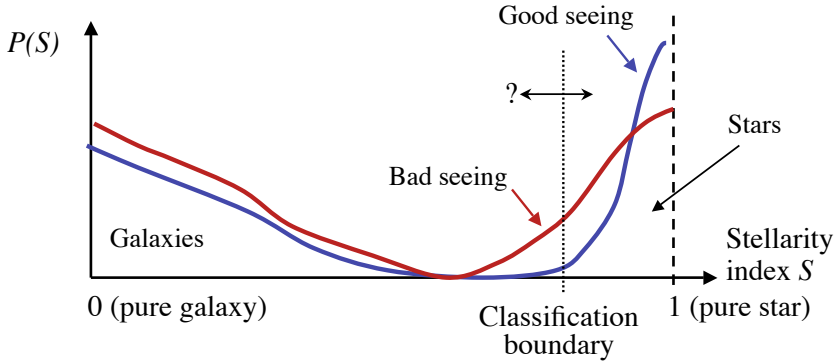
1. Include the external knowledge among the input parameters



2. A two-step classification:



Classification Bias and Accuracy

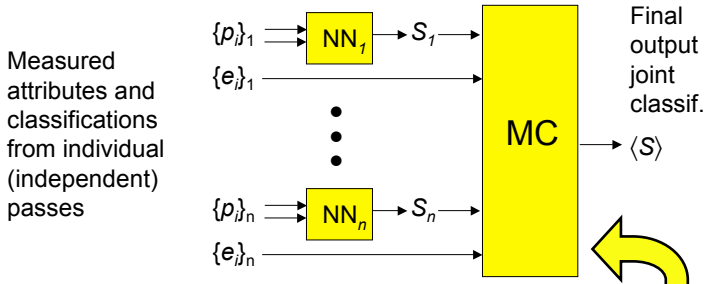


Assuming a classification boundary divider (stars/galaxies) derived from good quality data, and applying it to poorer quality data, would lead to a ***purer, but biased sample***, as some stars will be misclassified as galaxies.

Shifting the boundary (e.g., on the basis of external knowledge) would ***diminish the bias, but also degrade the purity.***

Combining Multiple Classifications

Metaclassifier, or a committee of machines with a chairman?

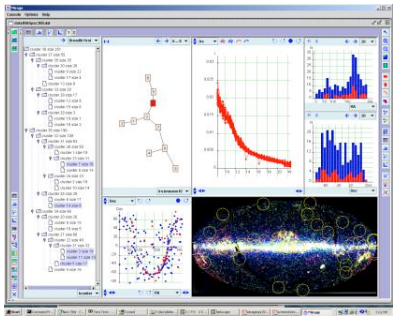


Note: individual classifiers may be optimized or trained differently

Design?
Weighting algorithm?
Training data set?
Validation data set?

Visualization: An Essential Need

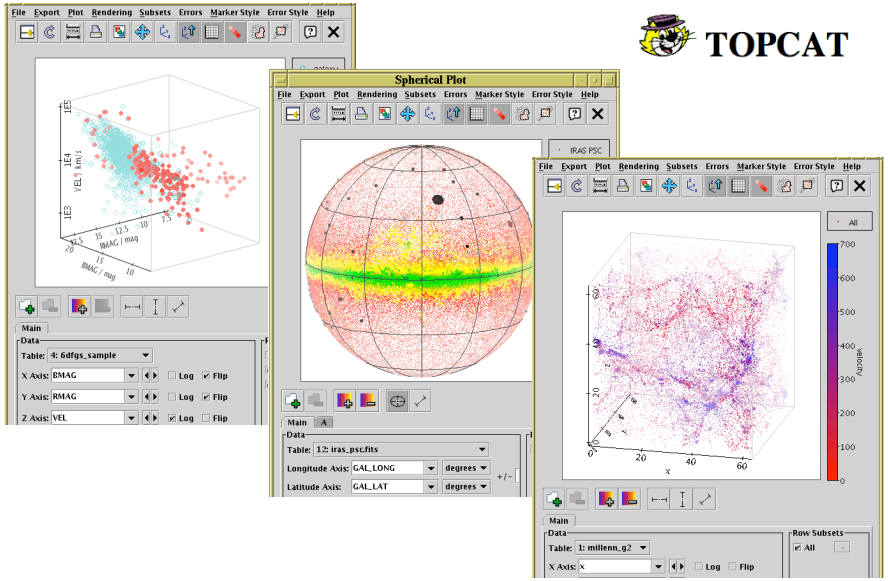
- Visualization is the bridge from the data (and data mining results) to the human intuition and understanding
- It has to be an integral part of the data mining and exploration process
- Many good packages exist, but they generally do not scale well to huge numbers of data points, and to a high dimensionality of data sets



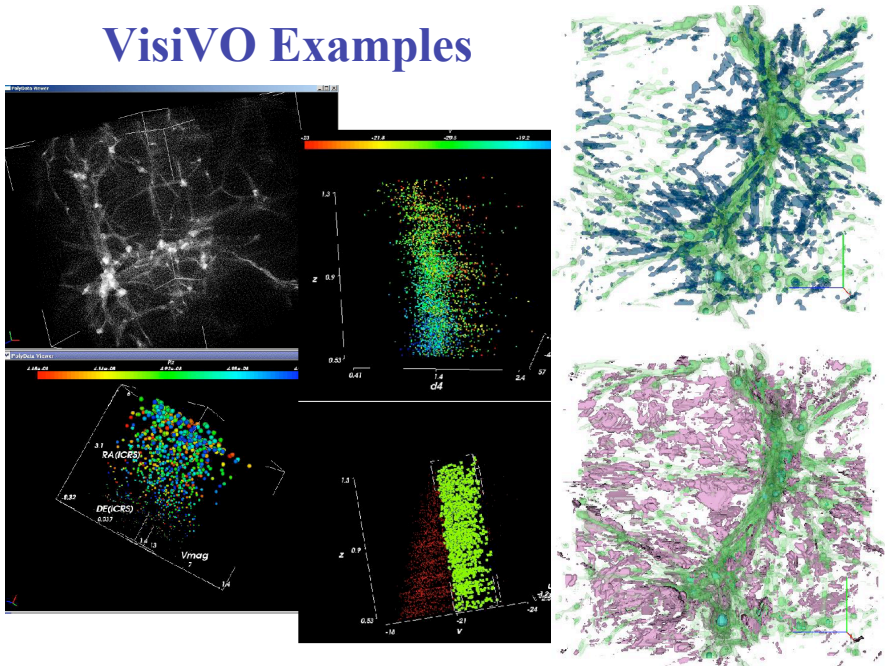
Mirage GUI - Tin Kam Ho

- Some popular and useful VO options include:
  - TopCat: ***http://www.star.bristol.ac.uk/~mbt/topcat/***
  - VisiVO: ***http://visivo.cineca.it/*** (see also astro-ph/0707.2474)
  - Mirage: ***http://cm.bell-labs.com/who/tkh/mirage/***
  - PartiView: ***http://virdir.ncsa.uiuc.edu/partiview/***

TopCat Examples



VisiVO Examples

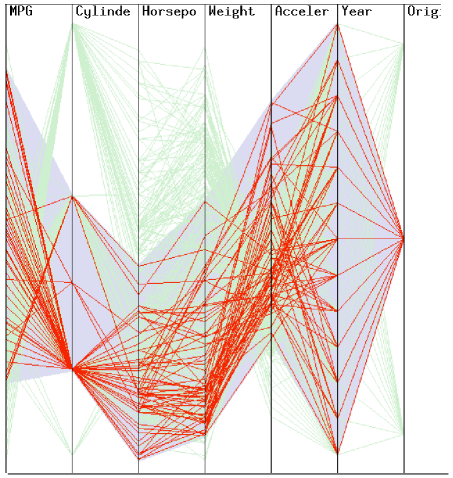


Dealing With Hyper-Dimensionality

- A key problem is visualization of hyper-dimensional (D >> 3) data sets, *without* a loss of information
- We are biologically and evolutionary geared to perceive visual signals in 2-D and 3-D; yet, interesting structures and information may be present in the data in highly-dimensional structures which cannot be projected to 2-D or 3-D spaces
  - We may need to deploy machine intelligence to help us discover, analyze and understand such structures
- This is closely related to the problem of dimensionality reduction in data mining
- Unfortunately, essentially all R&D on visualization now seems to be oriented towards 3-D, driven by the commercial apps.
- This suggests a potentially very important research program

Parallel Coordinates

- One technique to visualize hyper-dimensional data sets
- Each dimension corresponds to an axis, and the N axes are organized as uniformly spaced vertical lines
- A data element in N-dim. space manifests itself as a connected set of points, one on each axis. Points lying on a common line or plane create readily perceived structures in the image



- Not the greatest method - but it may help in some situations

Some Useful Websites

In addition to those already cited, start with:

http://www.astro.caltech.edu/~george/dposs/kdd-links.html

Links on surveys, VO, statistics, etc., at:

http://www.astro.caltech.edu/~george/ay122/

… and follow the links from there