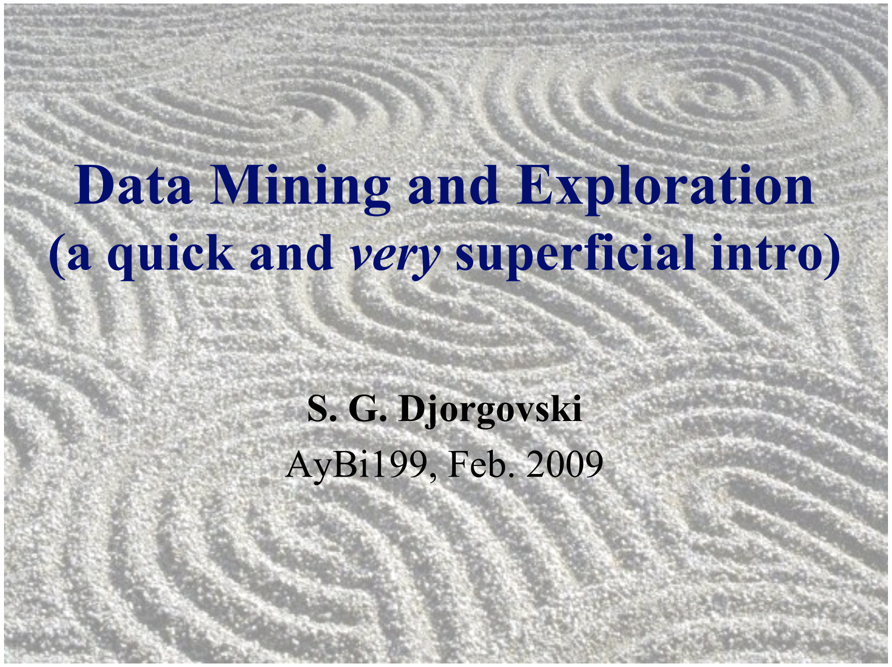


Data Mining and Exploration (a quick and *very* superficial intro)

S. G. Djorgovski

AyBi199, Feb. 2009

A Quick Overview Today

- A general intro to data mining
  - What is it, and what for?
- Clustering and classification
  - An example from astronomy: star-galaxy separation
- Exploratory statistics
  - An example from multivariate statistics: Principal Component Analysis (PCA) and multivariate correlations
- Second part of the lecture:

  Some examples and demos, by Ciro Donalek

Note: This is just a very modest start!
We posted some web links for you to explore, and go from there.

What is Data Mining (DM)?
(or: KDD = Knowledge Discovery in Databases)

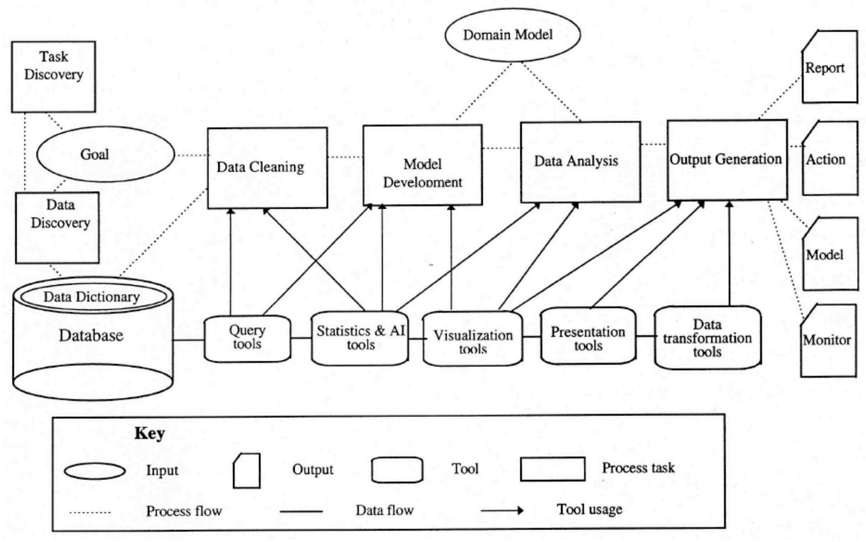
- Many different things, but generally what the name KDD says
  - It includes data discovery, cleaning, and preparation
  - Visualization is a *key* component (and can be very problematic)
  - It often involves a search for patterns, correlations, etc.; and automated and objective classification
  - It includes data modeling and testing of the models
- It depends a lot on the type of data, the study domain (science, commerce, …), the nature of the problem, etc., etc.
- Generally, DM algorithms are computational embodiments of statistics

This is a **Huge**, **HUGE**, field! Lots of literature, lectures, software… And yet, lots of unsolved applied CS research problems

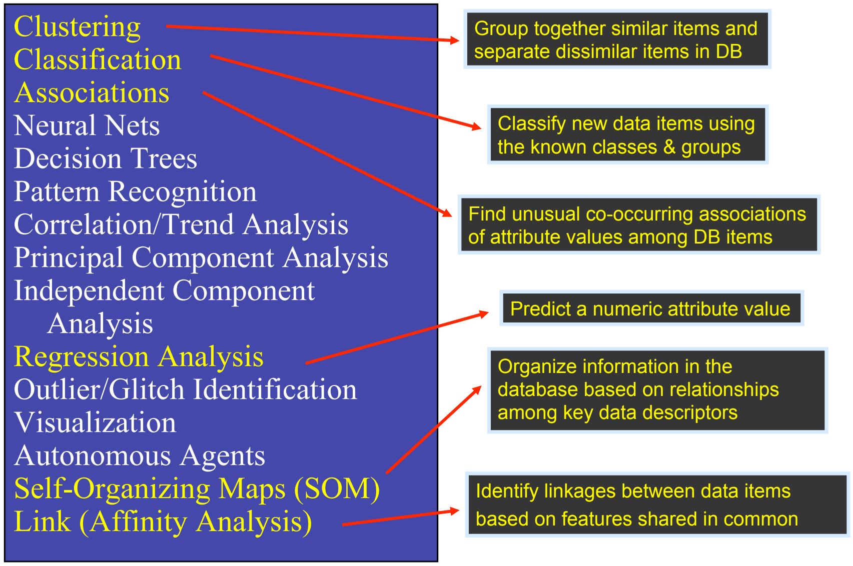
So what is Data Mining (DM)?

- The job of science is ***Knowledge Discovery***; data are incidental to this process, representing the empirical foundations, but not the understanding *per se*
  - A lot of this process is pattern recognition (including discovery of correlations, clustering/classification), discovery of outliers or anomalies, etc.
- DM is ***Knowledge Discovery in Databases (KDD)***
- DM is defined as ***"an information extraction activity whose goal is to discover hidden facts contained in (large) databases"***
- ***Machine Learning*** (ML) is the field of Computer Science research that focuses on algorithms that learn from data
- DM is the application of ML algorithms to large databases
  - And these algorithms are generally computational representations of some statistical methods

A Schematic View of KDD



Data Mining Methods and Some Examples



Clustering
Classification
Associations
Neural Nets
Decision Trees
Pattern Recognition
Correlation/Trend Analysis
Principal Component Analysis
Independent Component Analysis
Regression Analysis
Outlier/Glitch Identification
Visualization
Autonomous Agents
Self-Organizing Maps (SOM)
Link (Affinity Analysis)

Group together similar items and separate dissimilar items in DB

Classify new data items using the known classes & groups

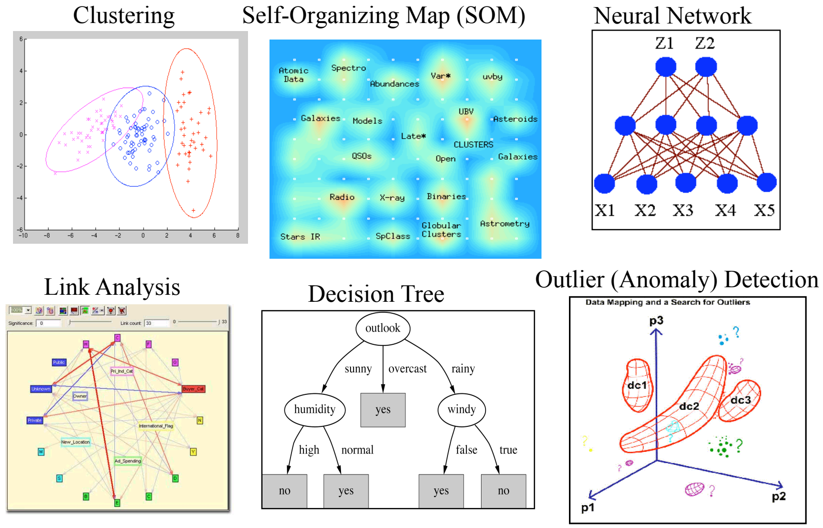
Find unusual co-occurring associations of attribute values among DB items

Predict a numeric attribute value

Organize information in the database based on relationships among key data descriptors

Identify linkages between data items based on features shared in common

Some Data Mining Techniques Graphically Represented



Here we show selected Kirk Borne's slides from the NVO Summer School 2008,

http://nvo-twiki.stsci.edu/twiki/bin/view/Main/NVOSS2008Sched

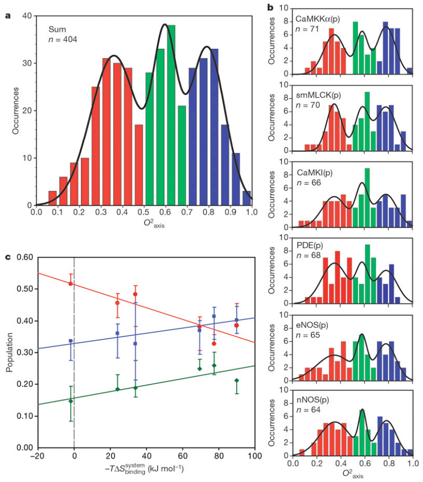
Clustering and Classification

- Answering the questions like:
  - How many statistically distinct kinds of things are there in my data, and which data object belongs to which class?
  - Are there anomalies/outliers? (e.g., extremely rare classes)
  - I know the classes present in the data, but would like to classify efficiently all of my data objects
- Clustering can be:

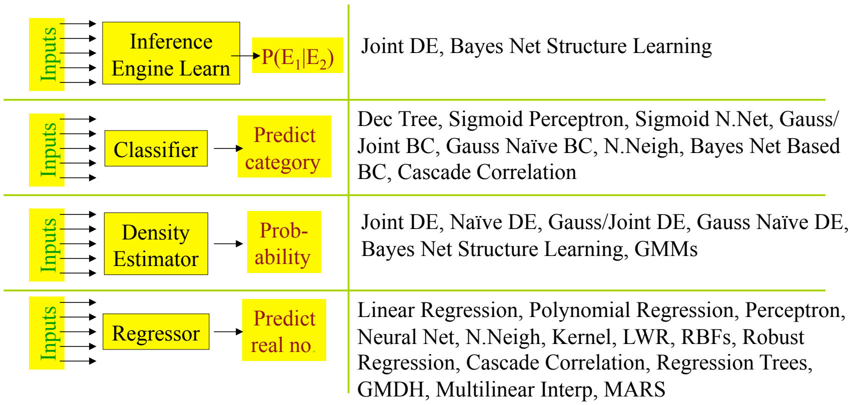
1. **Supervised**: a known set of data objects (“ground truth”) can be used to train and test a classifier
   - Examples: Artificial Neural Nets (ANN), Decision Trees (DT)
2. **Unsupervised:** the class membership (and the number of classes) is not known a priori; the program should find them
   - Examples: Kohonen Nets (SOM), Expectation Maximization (EM), various Bayesian methods…

Classification ~ Mixture Modeling

- A lot of DM involves automated classification or mixture modeling
  - How many kinds of data objects are there in my data set?
  - Which object belongs to which class with what probability?
- Different classes often follow different correlations
  - Or, correlations may define the classes which follow them
- Classes/clusters are defined by their probability density distributions in a parameter space



There are many good tools out there, but you need to choose the right ones for your needs



Joint DE, Bayes Net Structure Learning

Dec Tree, Sigmoid Perceptron, Sigmoid N.Net, Gauss/Joint BC, Gauss Naïve BC, N.Neigh, Bayes Net Based BC, Cascade Correlation

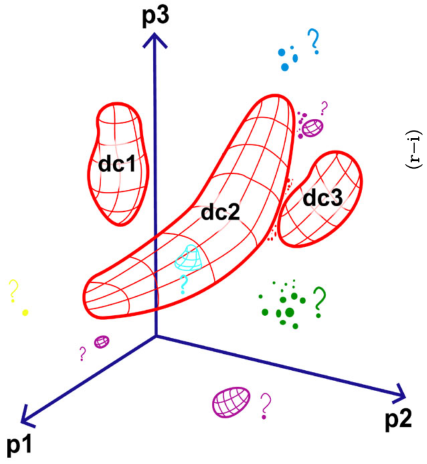
Joint DE, Naïve DE, Gauss/Joint DE, Gauss Naïve DE, Bayes Net Structure Learning, GMMs

Linear Regression, Polynomial Regression, Perceptron, Neural Net, N.Neigh, Kernel, LWR, RBFs, Robust Regression, Cascade Correlation, Regression Trees, GMDH, Multilinear Interp, MARS

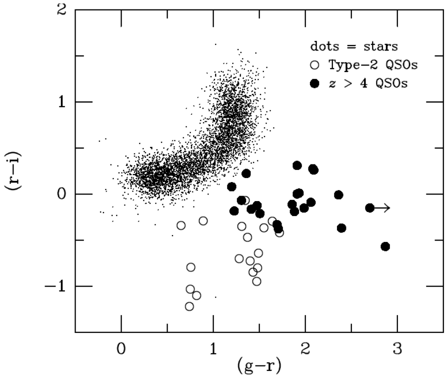
(from Moore 2002)

Exploration of observable parameter spaces and searches for rare or new types of objects

A Generic Machine-Assisted Discovery Problem: Data Mapping and a Search for Outliers



A simple, real-life example:



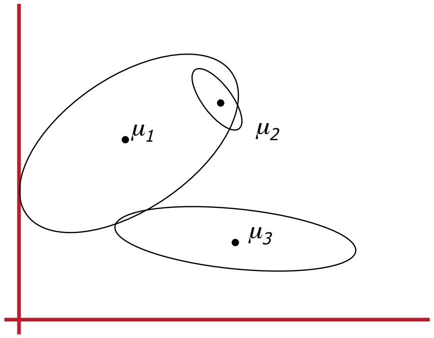
Now consider ~ 10^9 data vectors in ~ 10^2 - 10^3 dimensions …

Gaussian Mixture Modeling

- Data points are distributed in some *N*-dimensional parameter space,

  x_j, $j = 1, \ldots N$

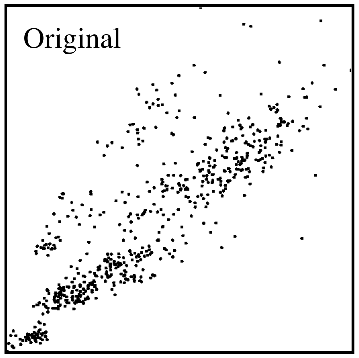
- There are *k* clusters, w_i, $i = 1, \ldots, k$, where the **number of clusters, *k*,** may be either given by the scientist, or derived from the data themselves

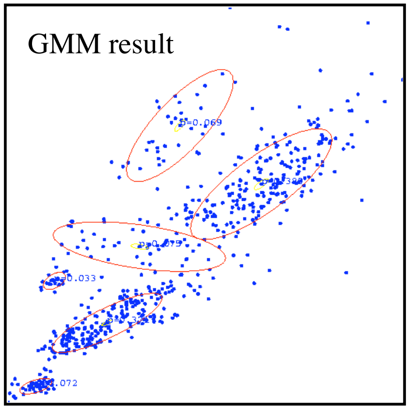


- Each cluster can be **modeled as an *N*-variate Gaussian** with mean μ_i and covariance matrix S_i
- Each data point has an association probability of belonging to each of the clusters, P_i

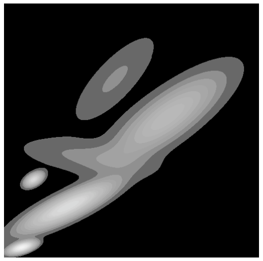
An Example

(from Moore et al.)



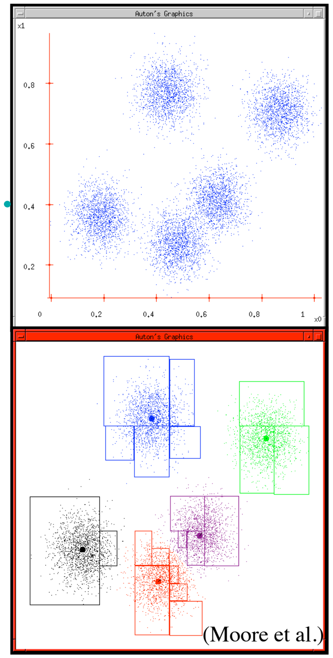


Model density distribution ⇒



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)



(Moore et al.)

In modern 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 × N × I × **D**

Expectation Maximization: K × N × I × $\mathbf{D^2}$

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

N = no. of data vectors, D = no. of data dimensions

K = no. of clusters chosen, K_{max} = max no. of clusters tried

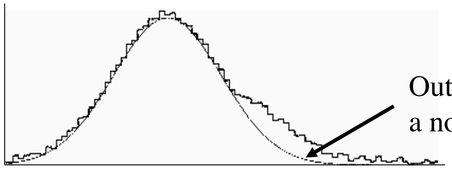
I = no. of iterations, M = no. of Monte Carlo trials/partitions

⇨ *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

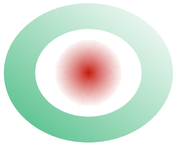
Some Practical and Theoretical Problems in Clustering Analysis

- Data heterogeneity, biases, selection effects …
- Non-Gaussianity of clusters (data models)



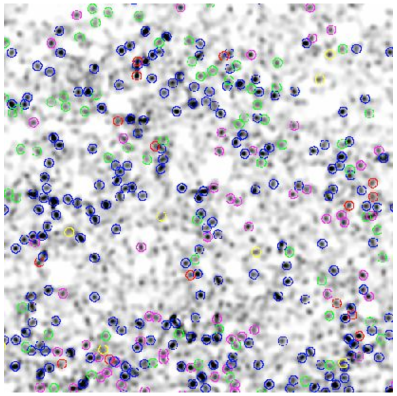
Outlier population, or a non-Gaussian tail?

- Missing data, upper and lower limits
- Non-Gaussian (or non-Poissonian) noise
- Non-trivial topology of clustering
- Useful vs. “useless” parameters …



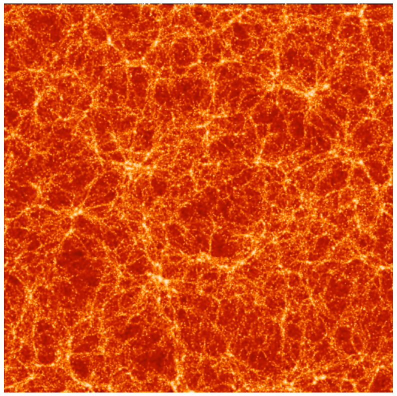
Some Simple Examples of Challenges for Clustering Analysis from “Standard” Astronomical Galaxy Clustering Analysis

Clustering on a clustered background



DPOSS Clusters (Gal et al.)

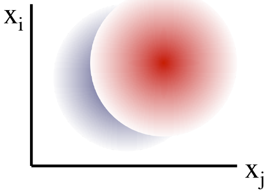
Clustering with a nontrivial topology

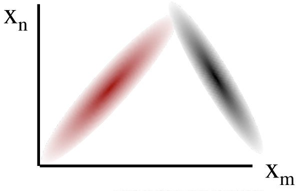


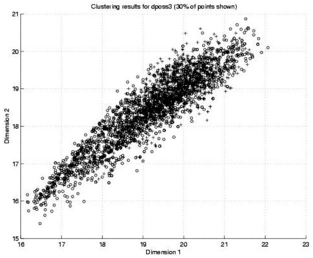
LSS Numerical Simulation (VIRGO)

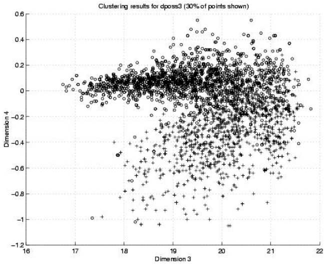
Useful *vs.* “Useless” Parameters:

Clusters (classes) and correlations may exist/separate in some parameter subspaces, but not in others





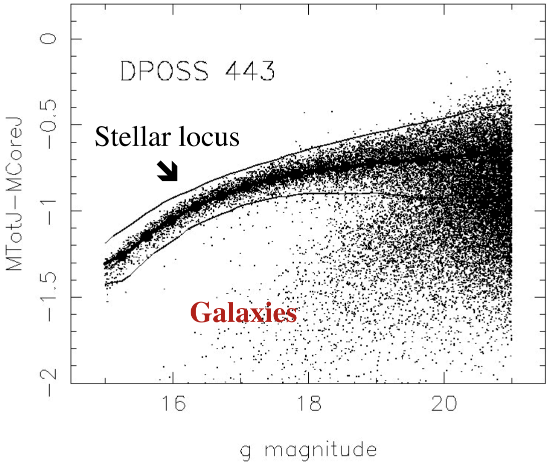




A Relatively Simple Classification Problem: Star-Galaxy Separation

- 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***

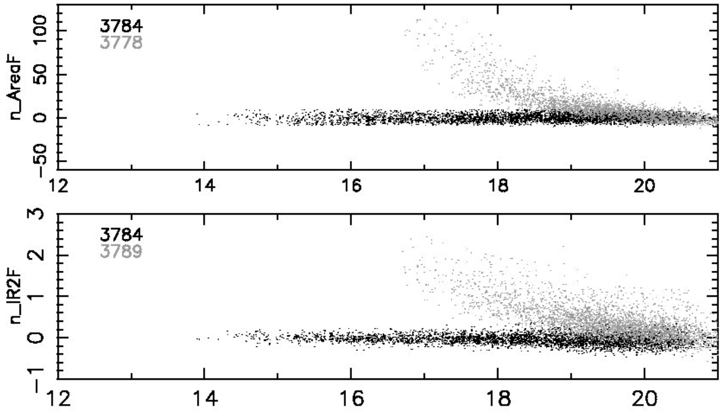
Typical Parameter Space for S/G Classif.



(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

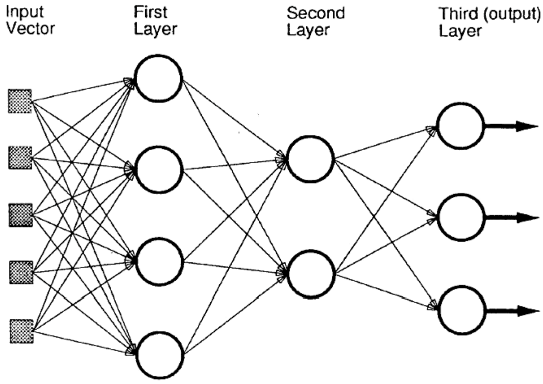
More S/G Classification Parameter Spaces: Normalized By The Stellar Locus



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)

Input: various image shape parameters.



Output:

Star, *p(s)*

Galaxy, *p(g)*

Other, *p(o)*

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)

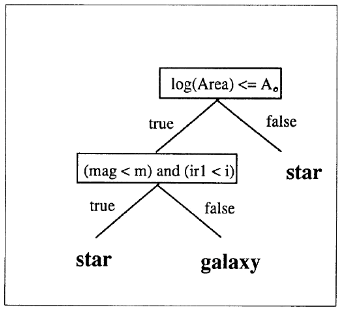


FIG. 1. In this sample decision tree, one starts at the top node(root), following the appropriate path to a final leaf (class) based upon the truth of the assertion at each node.

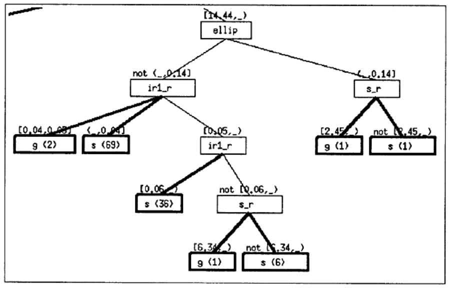
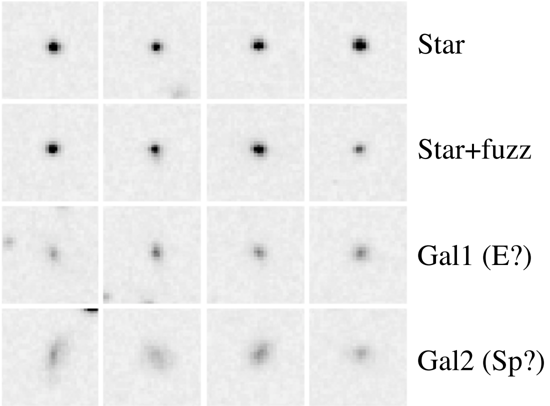


FIG. 2. A portion of a much larger actual decision tree generated by the O-Btree algorithm for performing star/galaxy classification. The interval appearing above each node indicates the range in value of the attribute specified in the node above that an object must meet for it to pass along that branch. 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.

(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.

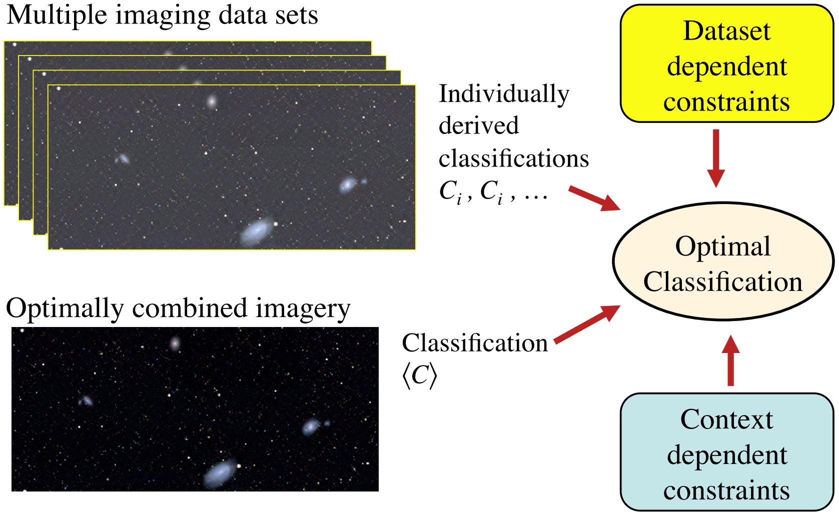


An example: AutoClass *(Cheeseman et al.)* Uses Bayesian approach in machine learning (ML).

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

Star-Galaxy Classification:

The Next Generation



Multiple imaging data sets

Individually derived classifications $C_i , C_i , \ldots$

Dataset dependent constraints

Optimal Classification

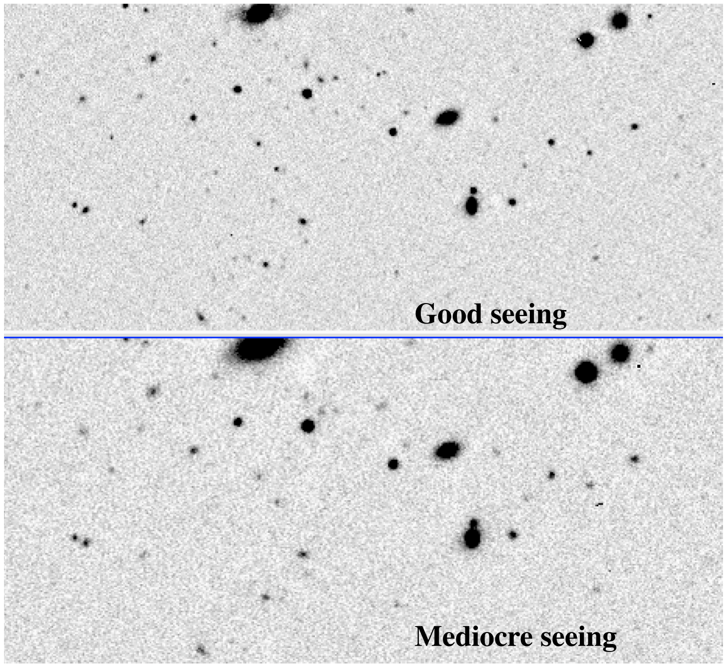
Optimally combined imagery

Classification $\langle C \rangle$

Context dependent constraints

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

Object dependent

Dataset dependent

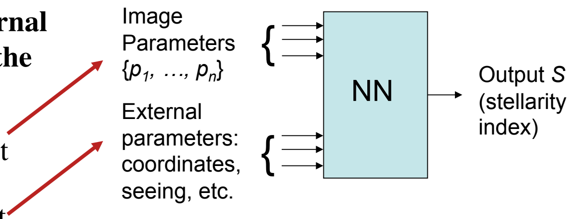
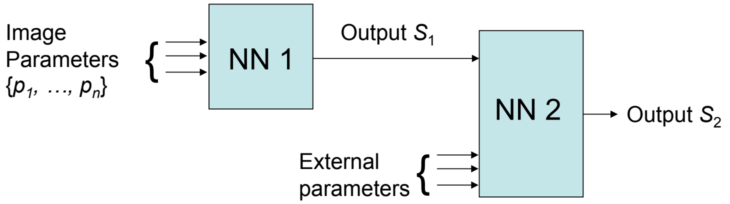
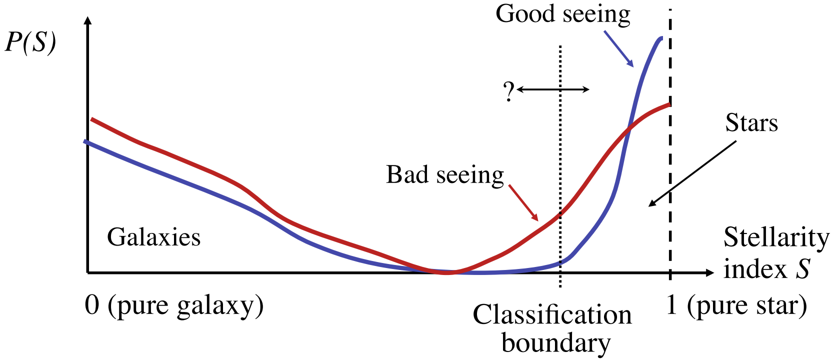


Image Parameters $\{p_1, \ldots, p_n\}$
External parameters: coordinates, seeing, etc.
NN
Output S (stellarity index)

2. A two-step classification:



Classification Bias and Accuracy



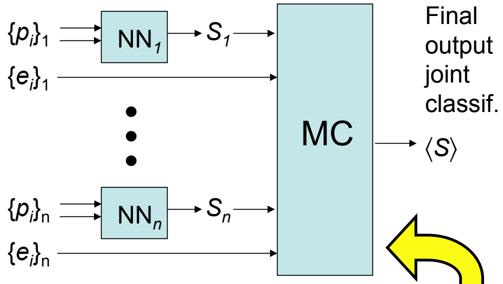
Assuming a classification boundary divider (stars/galaxies) derived from good quality data, and applying it to poorer quality data, would lead to a ***purer, but biased sample***, as some stars will be misclassified as galaxies.

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?

Measured attributes and classifications from individual (independent) passes



Final output joint classif.

Note: individual classifiers ⤴ may be optimized or trained differently

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

The (Proper) Uses of Statistics

- Hypothesis testing
- Model fitting
- **Data exploration:**
  - Multivariate analysis (MVA) and correlation search
  - Clustering analysis and classification
  - Image processing and deconvolutions
- **Data Mining** (or KDD):
  - Computational/algorithmic implementations of statistical tools

NB: Statistical Significance ≠ Scientific Relevance!

BAD uses of statistics:

- As a substitute for data (quantity or quality)
- To justify something *a posteriori*

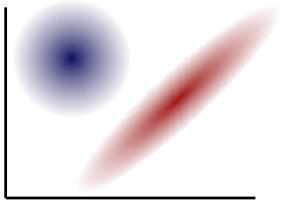
Multivariate Analysis (MVA)

Multivariate Correlation Search:

- Are there significant, nontrivial correlations present in the data?
- Simple monovariate correlations are rare: multivariate data sets can contain more complex correlations
- What is the statistical dimensionality of the data?

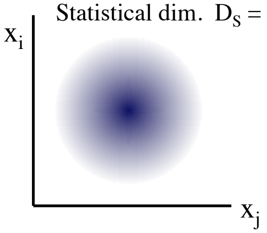
Clusters vs. Correlations:

"Physics" → Correlations
Correlations → reduction of the statistical dimensionality

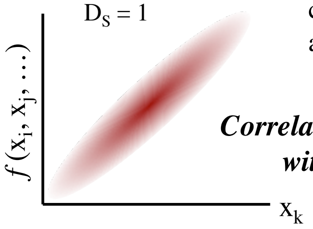


Correlation Searches in Attribute Space

Data dimension $D_D = 2$
Statistical dim. $D_S = 2$



$D_D = 2$
$D_S = 1$

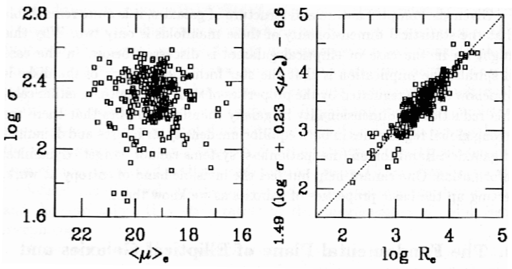


If $D_S < D_D$, correlations are present

Correlations are clusters with dimensionality reduction

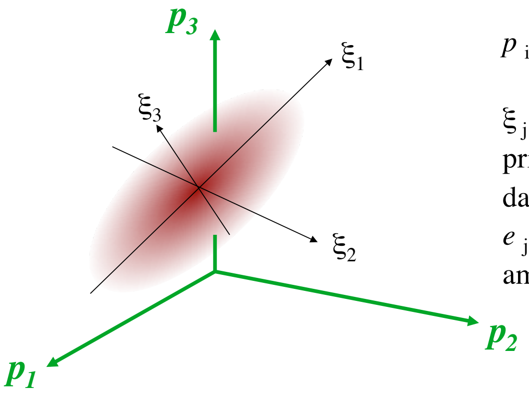
A real-life example:

"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



Principal Component Analysis

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



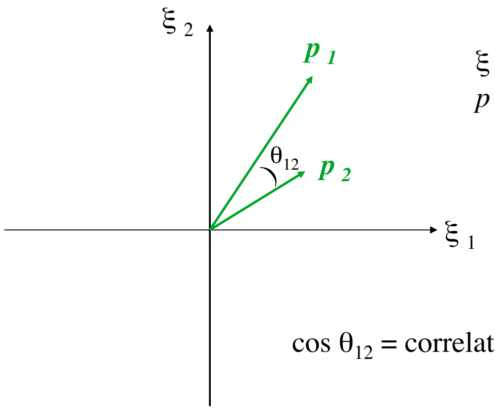
p_i = observables ($i = 1, \ldots D_{data}$)

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

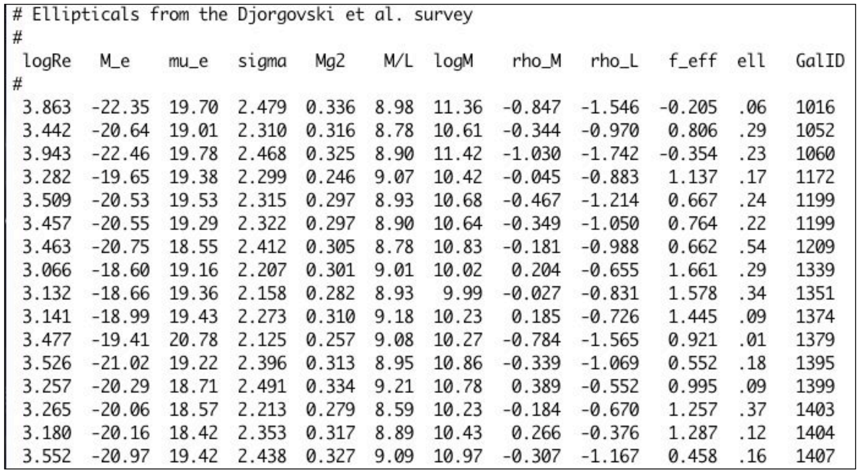


$\xi_i = a_{i1}\, p_1 + a_{i2}\, p_2 + \ldots$
$p_i = b_{i1}\, \xi_1 + b_{i2}\, \xi_2 + \ldots$

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

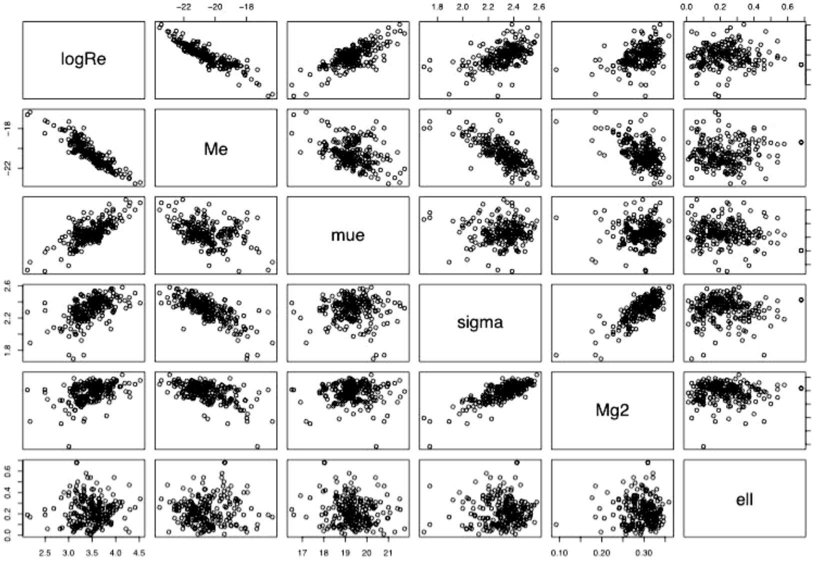
An Example, Using VOStat

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

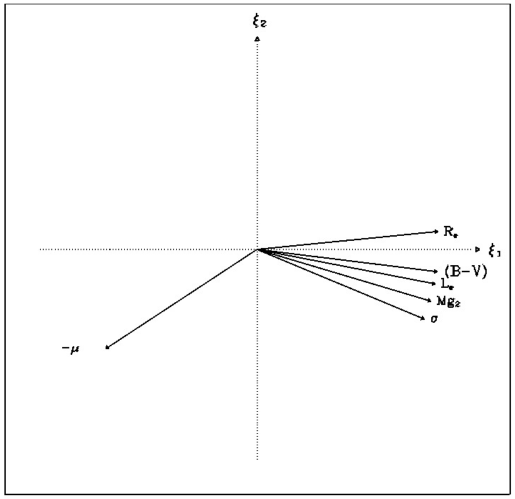
(This from a slightly different data set …)

```
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:

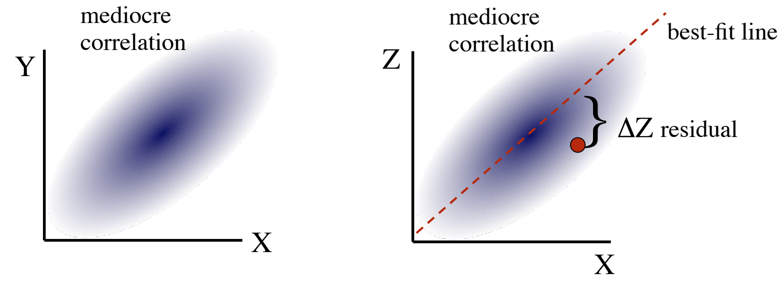
```
 VBLE.    EV-1    EV-2    EV-3    EV-4    EV-5
------  ------  ------  ------  ------  ------
logRe  -0.5119  0.3443  0.1649  0.1563  0.7535
 M_e     0.5291 -0.0310 -0.5158 -0.3689  0.5630
<mu>e  -0.2764  0.6991 -0.4679 -0.3181 -0.3388
sigma  -0.4614 -0.4399  0.1187 -0.7610  0.0194
 Mg2   -0.4108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


mediocre correlation

Y

X

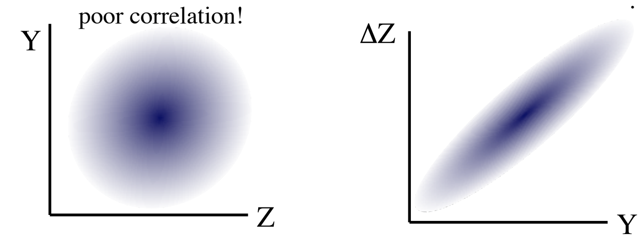
Z

mediocre correlation

best-fit line

ΔZ residual

X


poor correlation!

Y

Z

ΔZ

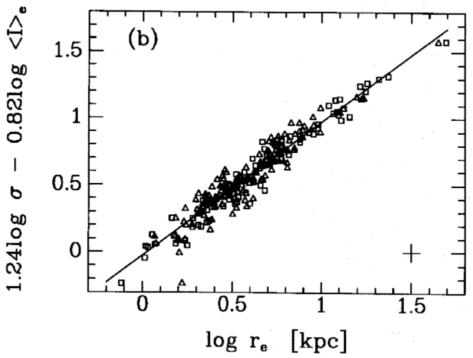
Y

… but the residuals correlate with the 3rd variable!

➞ The data are on a plane in the XYZ space

Bivariate Correlations in Practice

Once the dimensionality has been established from PCA, one can either derive the optimal bivariate combinations of variables from the PCA coefficients, or optimize the mixing ratios for any two variables vs. a third one (for a 2-dimensional manifold; the generalization to higher dimensional manifolds is obvious).



Some Data Mining Software & Projects

General data mining software packages:

- Weka (Java): http://www.cs.waikato.ac.nz/ml/weka/
- Weka4WS (Grid-enabled): http://grid.deis.unical.it/weka4ws/
- RapidMiner: http://www.rapidminer.com/

Astronomy-specific software and/or user clients:

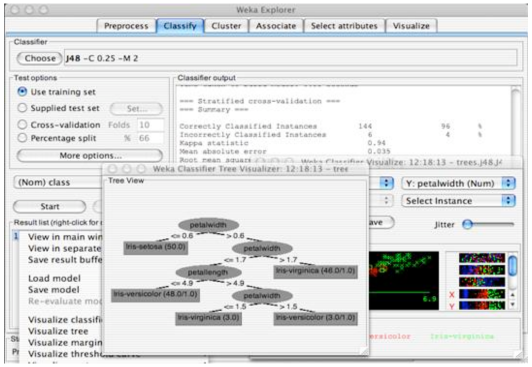
- VO-Neural: http://voneural.na.infn.it/
- AstroWeka: http://astroweka.sourceforge.net/
- OpenSkyQuery: http://www.openskyquery.net/
- ALADIN: http://aladin.u-strasbg.fr/
- MIRAGE: http://cm.bell-labs.com/who/tkh/mirage/
- AstroBox: http://services.china-vo.org/

Astronomical and/or Scientific Data Mining Projects:

- GRIST: http://grist.caltech.edu/
- ClassX: http://heasarc.gsfc.nasa.gov/classx/
- LCDM: http://dposs.ncsa.uiuc.edu/
- F-MASS: http://www.itsc.uah.edu/f-mass/
- NCDM: http://www.ncdm.uic.edu/

Examples of Data Mining Packages: Weka http://www.cs.waikato.ac.nz/ml/weka/

- A collection of open source machine learning algorithms for data mining tasks
- Algorithms can either be applied directly to a dataset or called from your own Java code
- Comes with its own GUI
- Contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization



Classifier: J48 -C 0.25 -M 2

Test options: ● Use training set ○ Supplied test set ○ Cross-validation Folds 10 ○ Percentage split % 66

Classifier output

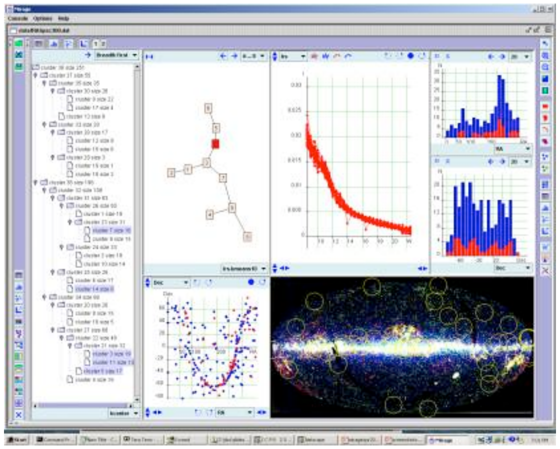
```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         144               96      %
Incorrectly Classified Instances         6                4      %
Kappa statistic                          0.94
Mean absolute error                      0.035
Root mean squar
```

(Nom) class

Y: petalwidth (Num)

Examples of Data Mining Packages: Mirage http://cm.bell-labs.com/who/tkh/mirage/

Java Package for exploratory data analysis (EDA), correlation mining, and interactive pattern discovery.



Here are some useful books:

- P.-N. Tan, M. Steinbach, & V. Kumar, *Introduction to Data Mining*, Addison-Wesley, 2005. ISBN: 9780321321367
- M. Dunham, *Data Mining: Introductory and Advanced Topics*, Prentice-Hall, 2002. ISBN: 9780130888921
- R. J. Roiger & M. W. Geatz, *Data Mining: A Tutorial-Based Primer*, Addison-Wesley, 2002. ISBN: 9780201741285

- ***Lots*** of good links to follow from the class webpage!