Lecture 1

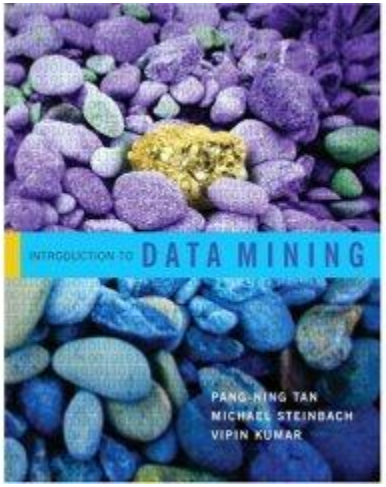
What is (Astronomical) Data Mining

Giuseppe Longo
University Federico II in Napoli – Italy
Visiting faculty – California Institute of Technology

Massimo Brescia
INAF-Capodimonte - Italy

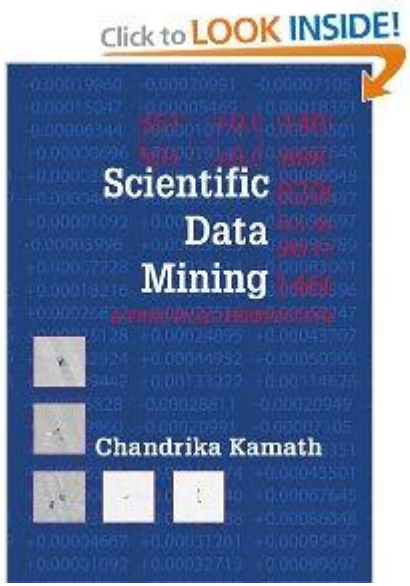
precari inside

A large part of this course was extracted from these excellent books:



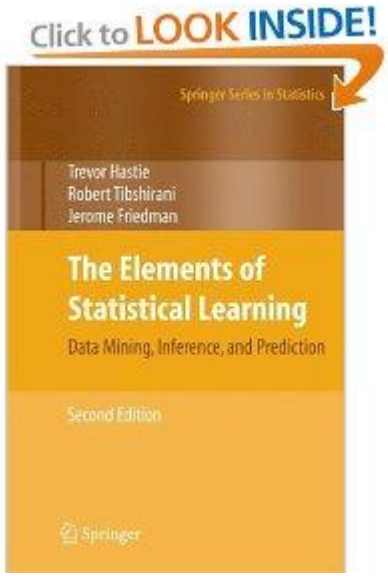
Introduction to Data Mining

Pang-Ning Tan, Michael Steinbach, Vipin Kumar, University of Minnesota



Scientific Data Mining

C. Kamath, SIAM publisher 2009



The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics) by Trevor Hastie, Robert Tibshirani and Jerome Friedman (2009) , Springer

Five slides on what is Data Mining. I

*Data mining (the analysis step of the Knowledge Discovery in Databases process, or KDD), a relatively **young and interdisciplinary** field of computer science, is the process of extracting patterns from **large data sets** by combining methods from statistics and artificial intelligence with database management ….*

***With recent technical advances in processing power, storage capacity, and inter-connectivity of computer technology**, data mining is an increasingly important tool by modern **business** to transform unprecedented quantities of digital data into business intelligence giving an **informational advantage**.*

*The growing consensus that data mining can bring real value has led to an explosion in demand for **novel data mining** technologies….*

From Wikipedia

… Excusatio non petita, accusatio manifesta …

- **There is a lot of confusion which can discourage people.**

  Initially part of KDD (Knowledge Discovery in Databases) together with data preparation, data presentation and data interpretation, DM has encountered a lot of difficulties in defining precise boundaries…

  In 1999 the NASA panel on the application of data mining to scientific problems concluded that: "*it was difficult to arrive at a consensus for the definition of data mining… apart from the clear importance of scalability as an underlying issue*".

- people who work in machine learning, pattern recognition or exploratory data analysis, often (and erroneously) view it as an **extension of what they have been doing for many years**…

- **DM inherited some bad reputation from initial applications.** Data Mining and Data dredging (data fishing, data snooping, etc…) were used to sample parts of a larger population data set that were too small for reliable statistical inferences to be made about the validity of any patterns

  For instance, till few years ago, statisticians considered DM methods as an unacceptable oversimplification

  People also wrongly believe that DM methods are a sort of black box completely out of control…

DATA MINING: my definition

Data Mining is the process concerned with automatically uncovering patterns, associations, anomalies, and statistically significant structures **in large and/or complex** data sets

Therefore it includes all those disciplines which can be used to uncover useful information in the data

What is new is the confluence of the most mature offshoots of many disciplines with technological advances

As such, its contents are «user defined» and more than a new discipline it is an ensemble of different methodologies originated in different fields

D. Rumsfeld on DM functionalities…

There are known knowns,
There are known unknowns,
and
There are unknown unknowns

Classification
Morphological classification of galaxies
Star/galaxy separation, etc.

Regression
Photometric redshifts

Clustering
Search for peculiar and rare objects,
Etc.

Courtesy of S.G. Djorgovski

Is Data Mining useful?

- **Can it ensure the accuracy required by scientific applications?**

  Finding the optimal route for planes, Stock market, Genomics, Tele-medicine and remote diagnosis, environmental risk assessment, etc**… HENCE…. Very likely yes**

- **Is it an easy task to be used in everyday applications (small data sets, routine work, etc.)?**

  NO!!

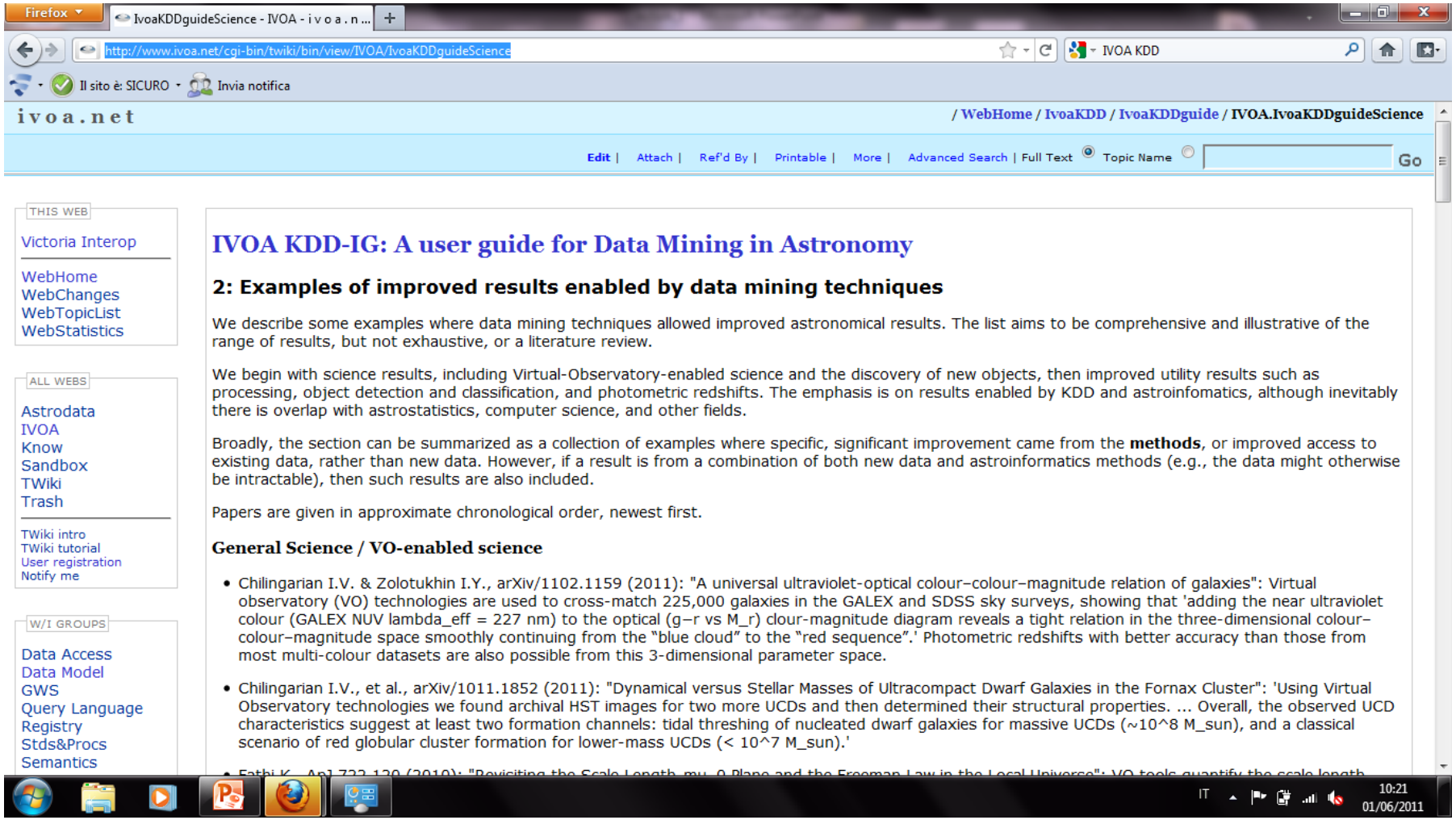
- **Can it work without a deep knowledge of the data models and of the DM algorithms/models?**

  NO!!

- **Can we do without it?**

  On large and complex data sets (TB-PB domain), NO!!!

http://www.ivoa.net/cgi-bin/twiki/bin/view/IVOA/IvoaKDDguideScience



Prepared and Mantained by N. Ball at the IVOA – IG-KDD pages

Scalability of some algorithms relevant to astronomy

- **Querying:** spherical range-search O(N), orthogonal range-search O(N), spatial join $O(N^2)$, nearest-neighbor O(N), all-nearest-neighbors $O(N^2)$
- **Density estimation:** mixture of Gaussians, kernel density estimation $O(N^2)$, kernel conditional density estimation $O(N^3)$
- **Regression:** linear regression, kernel regression $O(N^2)$, Gaussian process regression $O(N^3)$
- **Classification:** decision tree, nearest-neighbor classifier $O(N^2)$, nonparametric Bayes classifier $O(N^2)$, support vector machine $O(N^3)$
- **Dimension reduction:** principal component analysis, non-negative matrix factorization, kernel PCA $O(N^3)$, maximum variance unfolding $O(N^3)$
- **Outlier detection:** by density estimation or dimension reduction $O(N^3)$
- **Clustering**: by density estimation or dimension reduction, k-means, meanshift segmentation $O(N^2)$, hierarchical (FoF) clustering $O(N^3)$
- **Time series analysis:** Kalman filter, hidden Markov model, trajectory tracking $O(N^n)$
- **Feature selection and causality:** LASSO, L1 SVM, Gaussian graphical models, discrete graphical models
- **2-sample testing and testing and matching:** bipartite matching $O(N^3)$, n-point correlation $O(N^n)$

Courtesy of A. Gray – Astroinformatics 2010

Other relevant parameters

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

K-means: $K \times N \times I \times D$

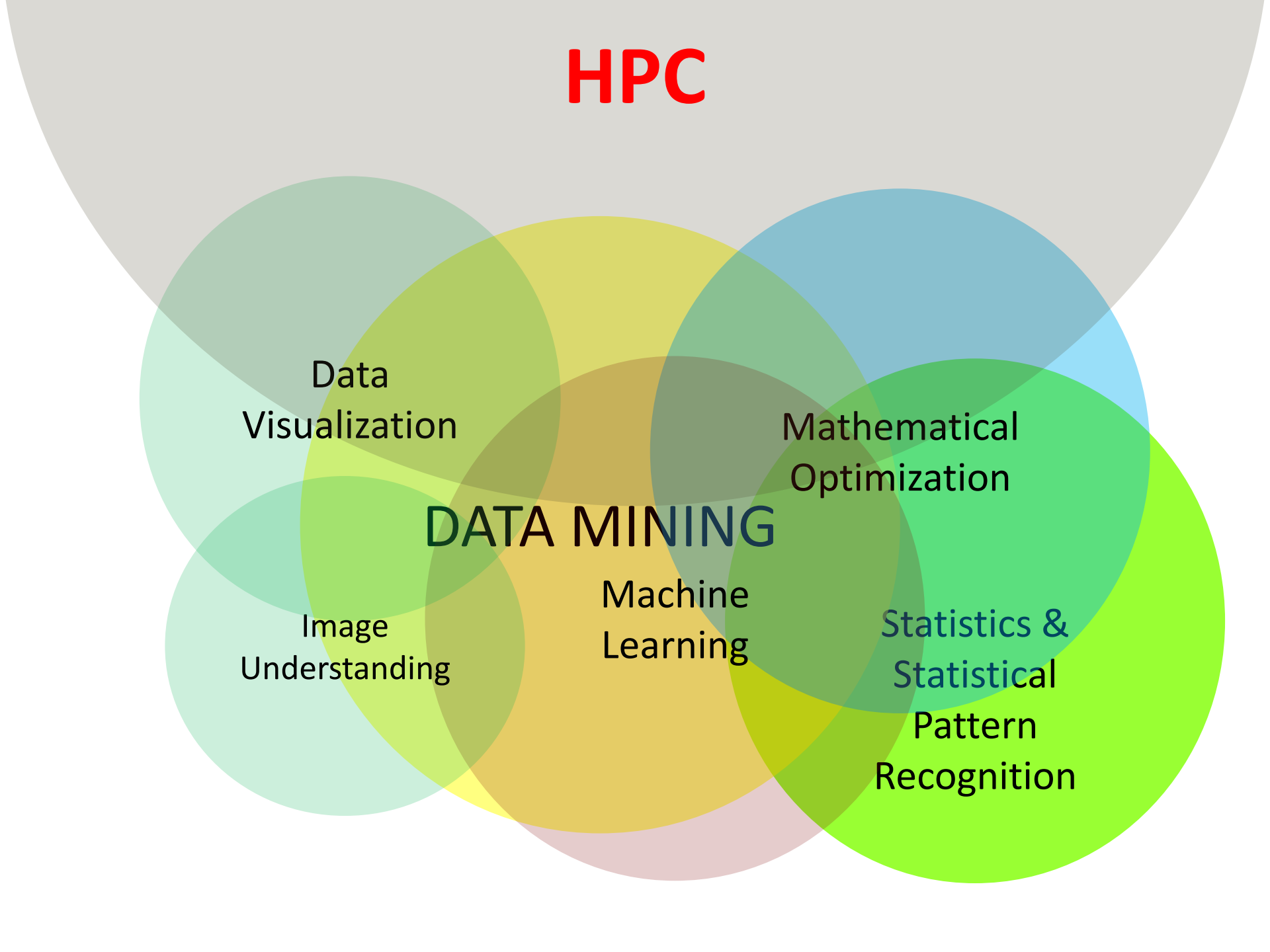
Expectation Maximisation: $K \times N \times I \times D^2$

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

Correlations ~ N log N or N^2, ~ D^k ($k \geq 1$)

Likelihood, Bayesian ~ N^m ($m \geq 3$), ~ D^k ($k \geq 1$)

SVM > ~ $(NxD)^3$


HPC
Data Visualization
Mathematical Optimization
DATA MINING
Machine Learning
Image Understanding
Statistics & Statistical Pattern Recognition

Use cases and domain knowledge….

… define workflows of functionalities

- Dim. reduction
- Regression
- Clustering
- Classification

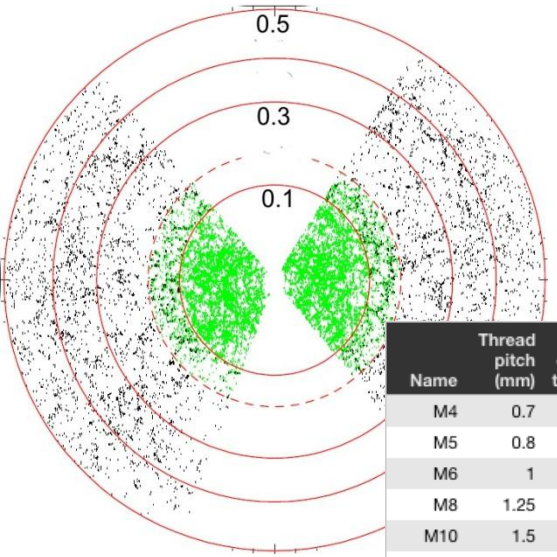
Modes

- supervised
- Unsupervised
- hybrid

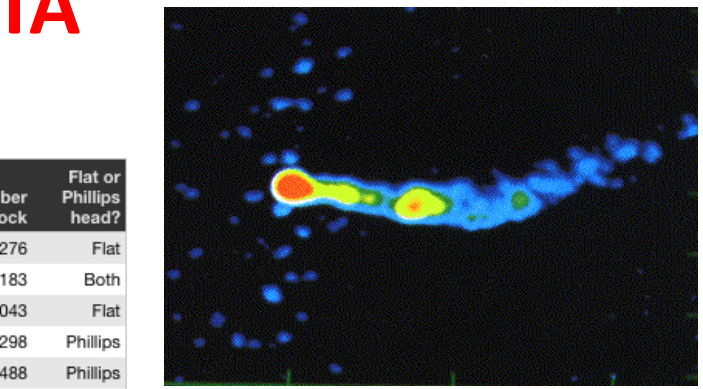
… which are implemented by specific models and algorithms

- Neural Networks (MLPs, MLP-GA, RBF, etc.)
- Support Vector Machines &SVM-C
- Decision trees
- K-D trees
- PPS
- Genetic algorithms
- Bayesian networks
- Etc…

STARTING POINT: THE DATA



Name	Thread pitch (mm)	Minor diameter tolerance	Nominal diameter (mm)	Head shape	Price for 50 screws	Available at factory outlet?	Number in stock	Flat or Phillips head?
M4	0.7	4g	4	Pan	$10.08	Yes	276	Flat
M5	0.8	4g	5	Round	$13.89	Yes	183	Both
M6	1	5g	6	Button	$10.42	Yes	1043	Flat
M8	1.25	5g	8	Pan	$11.98	No	298	Phillips
M10	1.5	6g	10	Round	$16.74	Yes	488	Phillips
M12	1.75	7g	12	Pan	$18.26	No	998	Flat
M14	2	7g	14	Round	$21.19	No	235	Phillips
M16	2	8g	16	Button	$23.57	Yes	292	Both
M18	2.1	8g	18	Button	$25.87	No	664	Both
			20	Pan	$29.09	Yes	486	Both
			24	Round	$33.01	Yes	982	Phillips
			28	Button	$35.66	No	1067	Phillips
			36	Pan	$41.32	No	434	Both
			50	Pan	$44.72	No	740	Flat



Some considerations on the Data

Data set: collection of data objects and their attributes

Data Object: a collection of objects. Also known as record, point, case, sample, entity, or instance

Attributes: a property or a characteristic of the objects. Also called: variables, feature, field, characteristic

Attribute values:are numbers or symbols assigned to an attribute

The same attribute can be mapped to different attribute values
Magnitudes or fluxes

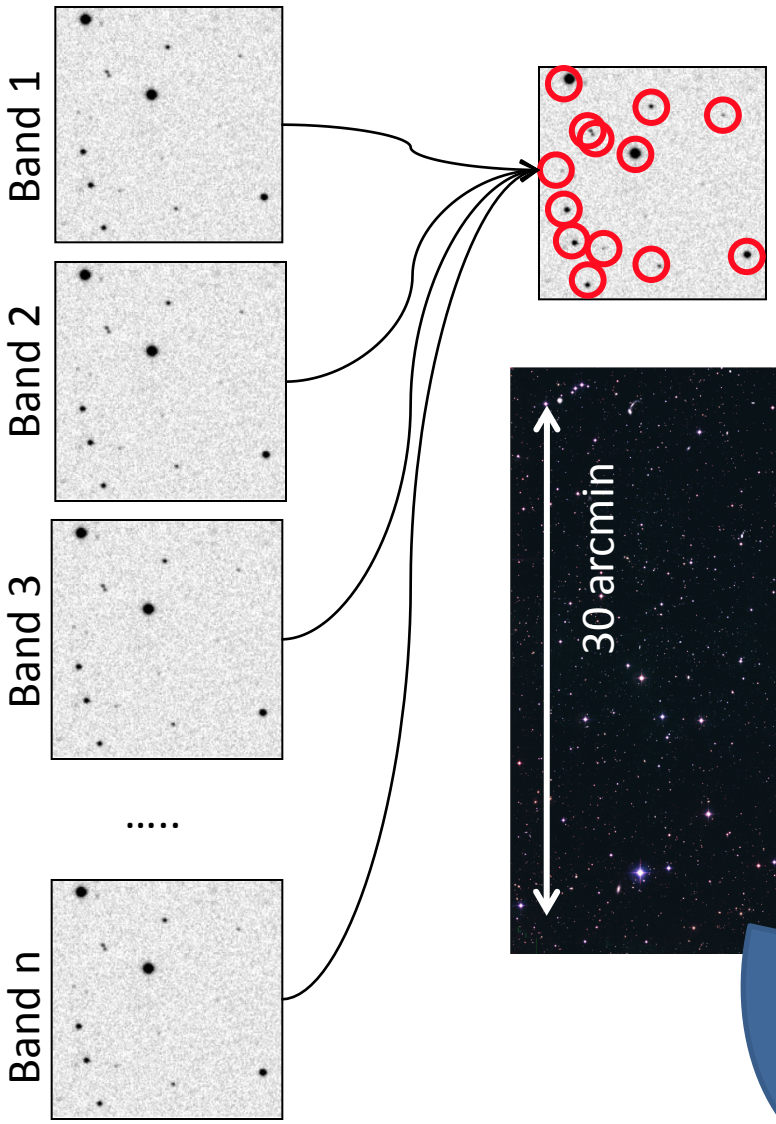
DATA SET: HCG90

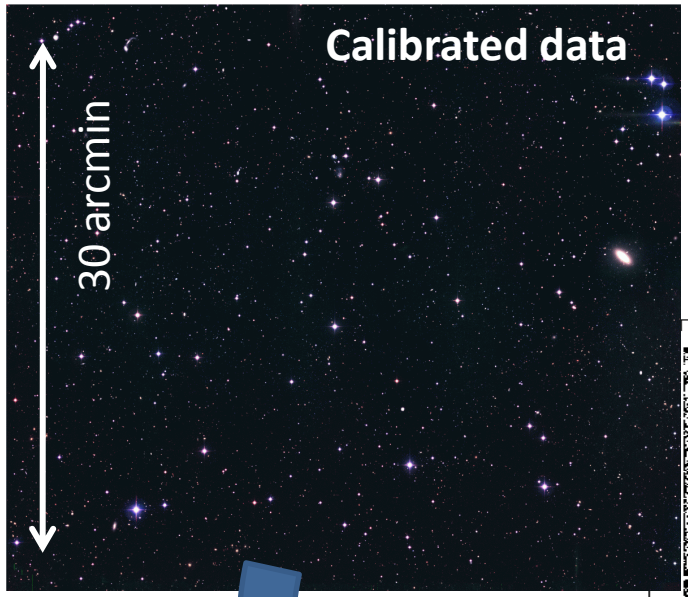
objects

ID	RA	DEC	z	B	Etc.
NGC7172	22h02m01.9s	-31d52m11s	0.008683	12.85	…
NGC7173	22h02m03.2s	-31d58m25s	0.008329	13.08	…
NGC7174	22h02m06.4s	-31d59m35s	0.008869	14.23	…
NGC7176	22h02m08.4s	-31d59m23s	0.008376	12.34	

attributes

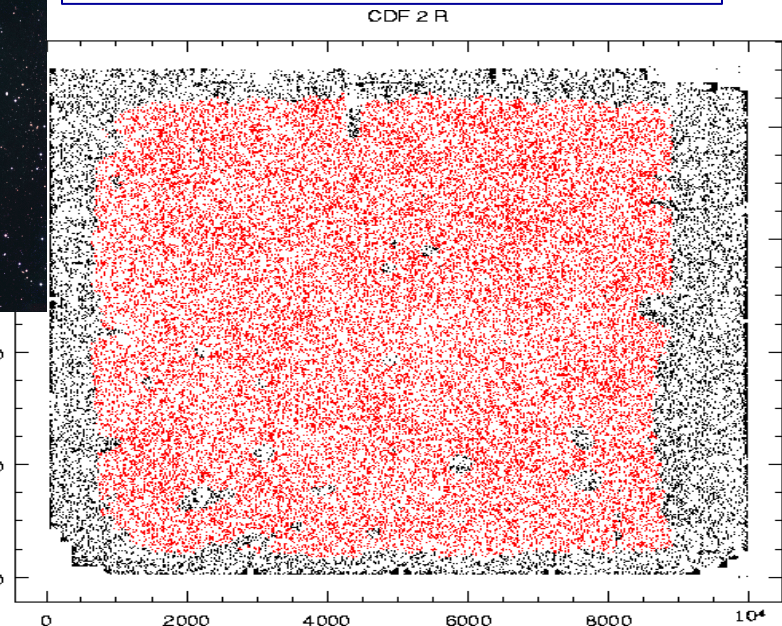
The universe is densely packed



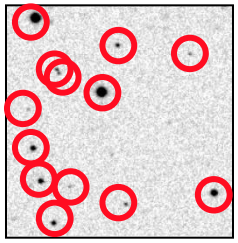


1/160.000 of the sky, moderately deep (25.0 in r)

55.000 detected sources
(0.75 mag above m lim)



The exploding parameter space…



p={isophotal, petrosian, aperture magnitudes concentration indexes, shape parameters, etc.}

$$p^1 = \left\{RA^1, \delta^1, t, \left\{\lambda_1, \Delta\lambda_1, f_1^{1,1}, \Delta f_1^{1,1}, \ldots, f_1^{1,m}, \Delta f_1^{1,m}\right\}, \ldots, \left\{\lambda_n, \Delta\lambda_n, f_n^{1,1}, \Delta f_n^{1,1}, \ldots, f_n^{1,m}, \Delta f_n^{1,m}\right\}\right\}$$

$$p^2 = \left\{RA^2, \delta^2, t, \left\{\lambda_1, \Delta\lambda_1, f_1^{2,1}, \Delta f_1^{2,1}, \ldots, f_1^{2,m}, \Delta f_1^{2,m}\right\}, \ldots, \left\{\lambda_n, \Delta\lambda_n, f_n^{2,1}, \Delta f_n^{2,1}, \ldots, f_n^{2,m}, \Delta f_n^{2,m}\right\}\right\}$$

$$\ldots\ldots\ldots\ldots\ldots\ldots$$

$$p^N = \left\{RA^N, \delta^N, t, \left\{\lambda_1, \Delta\lambda_1, f_1^{N,1}, \Delta f_1^{N,1}, \ldots, f_1^{N,m}, \Delta f_1^{N,m}\right\}, \ldots\right\}$$

$$D = 3 + m \times n$$

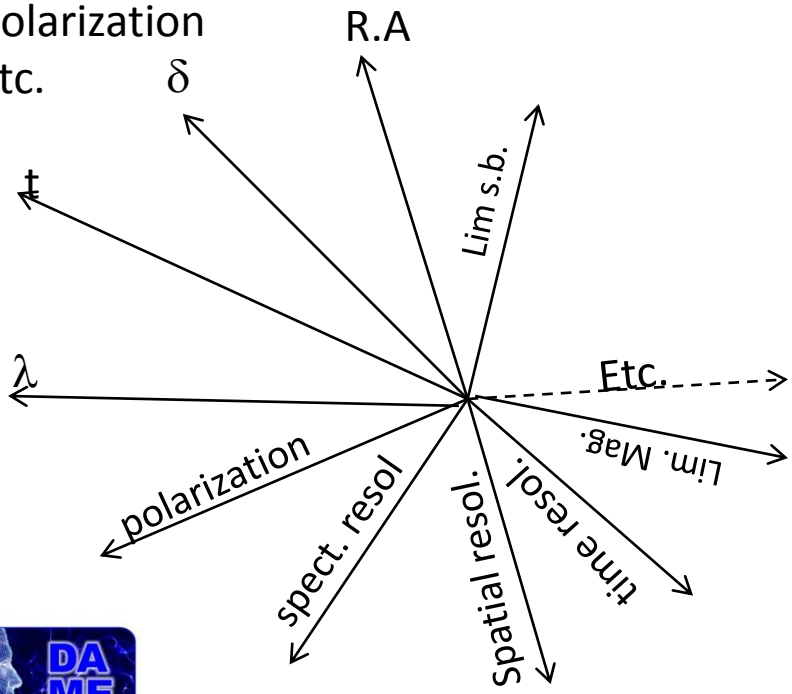
The scientific exploitation of a multi band, multiepoch (K epochs) universe implies to search for patterns, trends, etc. among N points in a DxK dimensional parameter space:

$N > 10^9$, D>>100, K>10

The parameter space

Any observed (simulated) datum *p* defines a point (region) in a subset of R^N. Es:

- RA and dec
- time
- λ
- experimental setup (spatial and spectral resolution, limiting mag, limiting surface brightness, etc.) parameters
- fluxes
- polarization
- Etc.



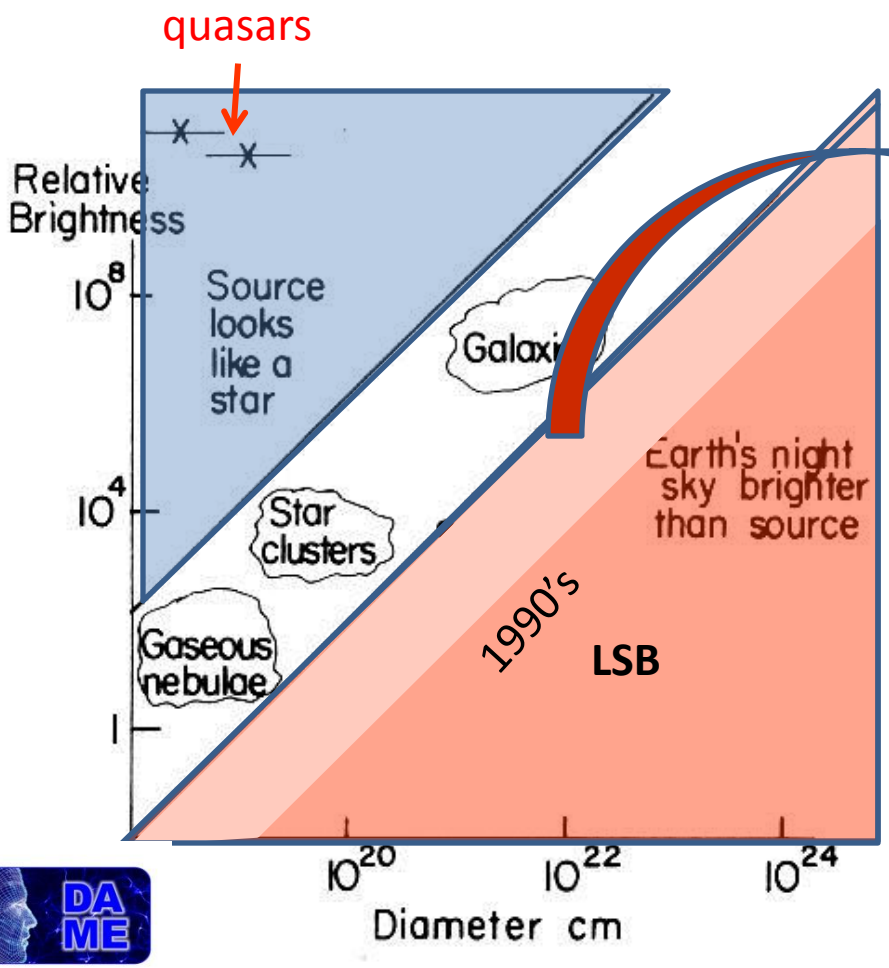
$$p \in \Re^N \qquad N >> 100$$

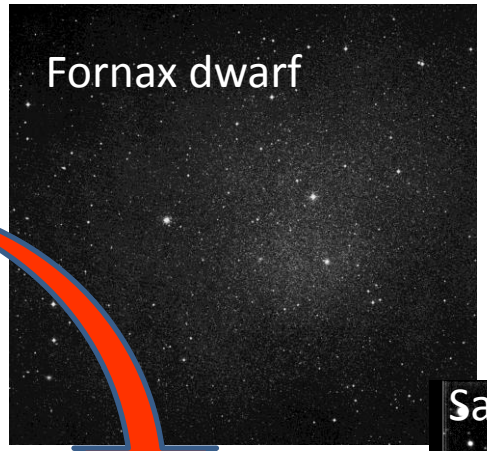
The parameter space concept is crucial to:

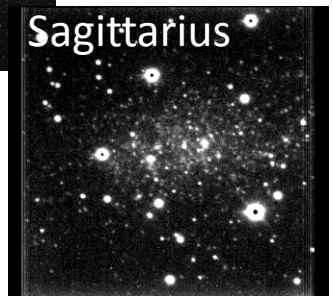
1. Guide the quest for new discoveries (observations can be guided to explore poorly known regions), …
2. Find new physical laws (patterns)
3. Etc,

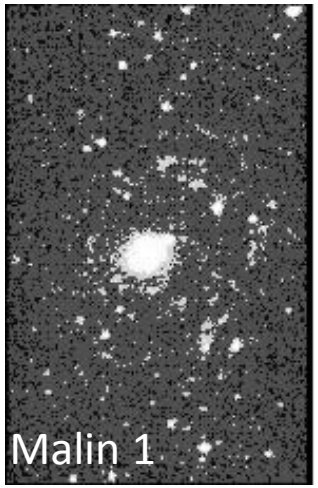
Every time you improve the coverage of the PS….

Every time a new technology enlarges the parameter space or allows a better sampling of it, new discoveries are bound to take place



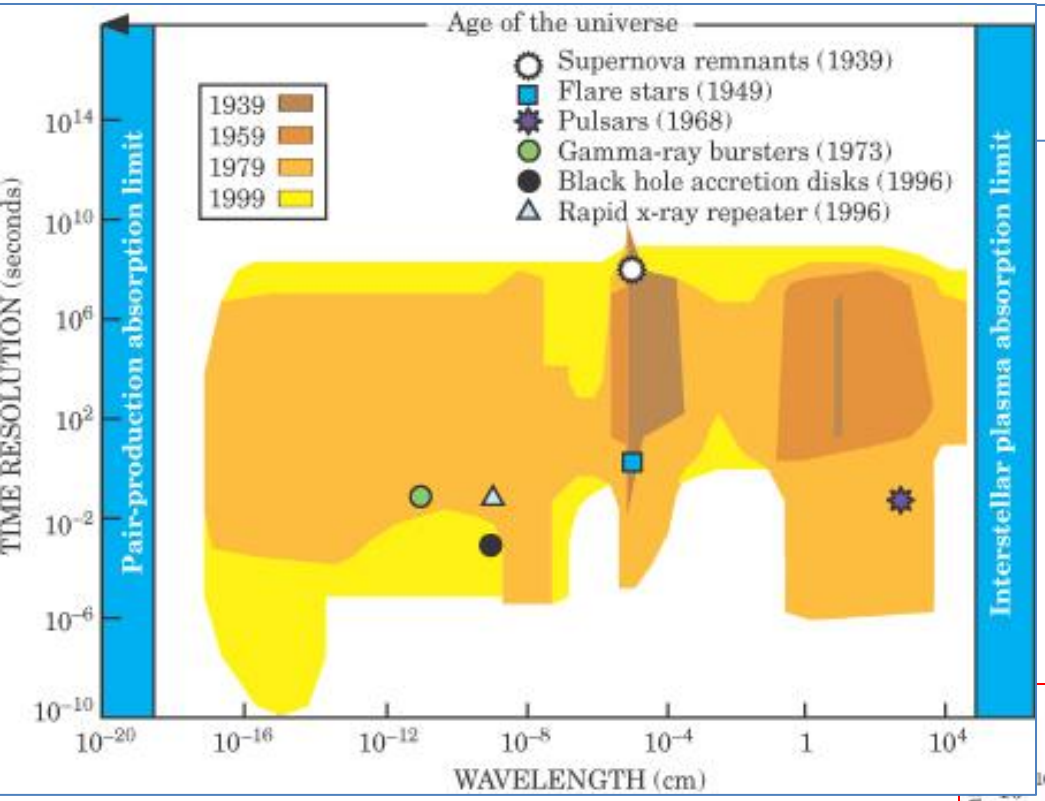




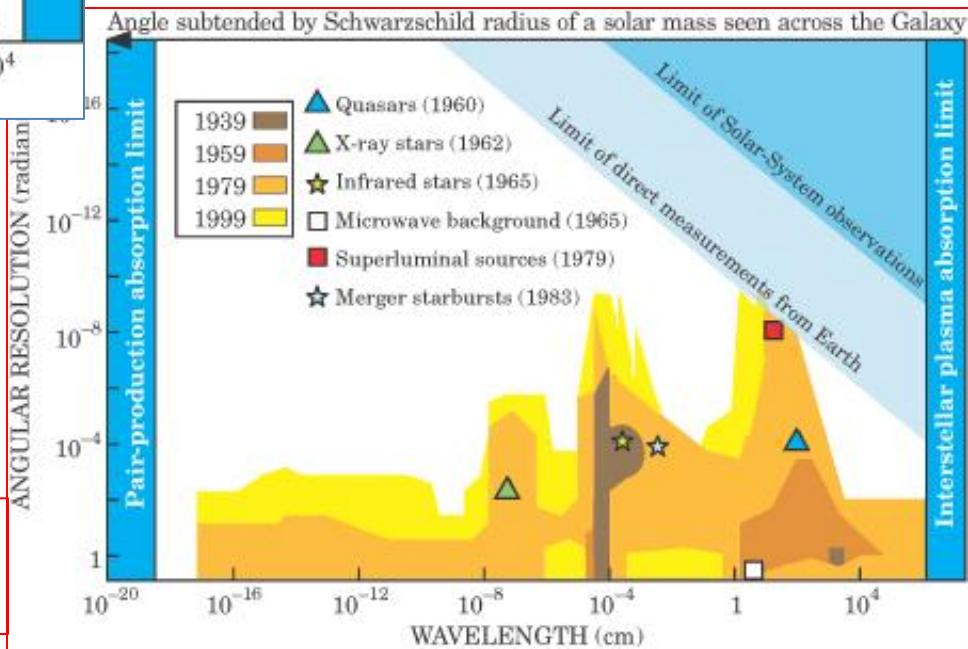


Discovery of Low surface brightness Universe

Improving coverage of the Parameter space - II



Projection of parameter space along (time resolution & wavelength)



Projection of parameter space along (angular resolution & wavelength)

Types of Attributes

Attribute Type	Description	Examples	Operations
Nominal	The values of a nominal attribute are just different names, i.e., nominal attributes provide only enough information to distinguish one object from another. (=, ≠)	NGC number, SDSS ID numbers, spectral type, etc.)	mode, entropy, contingency correlation, χ^2 test
Ordinal	The values of an ordinal attribute provide enough information to order objects. (<, >)	Morphological classification, spectral classification ??	median, percentiles, rank correlation, run tests, sign tests
Interval	For interval attributes, the differences between values are meaningful, i.e., a unit of measurement exists. (+, -)	calendar dates, temperature in Celsius or Fahrenheit	mean, standard deviation, Pearson's correlation, *t* and *F* tests
Ratio	For ratio variables, both differences and ratios are meaningful. (*, /)	temperature in Kelvin, monetary quantities, counts, age, mass, length, electrical current	geometric mean, harmonic mean, percent variation

Attribute Level	Transformation	Comments
Nominal	Any permutation of values	If all NGC numbers were reassigned, would it make any difference?
Ordinal	An order preserving change of values, i.e., *new_value = f(old_value)* where *f* is a monotonic function.	An attribute encompassing the notion of good, better best can be represented equally well by the values {1, 2, 3} or by { 0.5, 1, 10}.
Interval	*new_value =a * old_value + b* where a and b are constants	Thus, the Fahrenheit and Celsius temperature scales differ in terms of where their zero value is and the size of a unit (degree).
Ratio	*new_value = a * old_value*	Length can be measured in meters or feet.

Discrete and Continuous Attributes

- **Discrete Attribute**
  - Has only a finite or countably infinite set of values
  - Examples: SDSS IDs, zip codes, counts, or the set of words in a collection of documents
  - Often represented as integer variables.
  - Note: binary attributes **(flags)** are a special case of discrete attributes

- **Continuous Attribute**
  - Has real numbers as attribute values
  - Examples: fluxes,
  - Practically, real values can only be measured and represented using a finite number of digits.
  - Continuous attributes are typically represented as floating-point variables.

LAST TYPE: Ordered Data

Data where the position in a sequence matters:

Es. Genomic sequences
Es. Metereological data
Es. Light curves

```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

Ordered Data

- Genomic sequence data

```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

Data Quality

- What kinds of data quality problems?
- How can we detect problems with the data?
- What can we do about these problems?

- Examples of data quality problems:
  - Noise and outliers
  - duplicate data
  - missing values

Missing Values

- Reasons for missing values
  - Information is not collected (e.g., instrument/pipeline failure)
  - Attributes may not be applicable to all cases (e.g. no HI profile in E type galaxies)

- Handling missing values
  - Eliminate Data Objects
  - Estimate Missing Values (for instance upper limits)
  - Ignore the Missing Value During Analysis (if method allows it)
  - Replace with all possible values (weighted by their probabilities)

Data Preprocessing

- Aggregation
- Sampling
- Dimensionality Reduction
- Feature subset selection
- Feature creation
- Discretization and Binarization
- Attribute Transformation

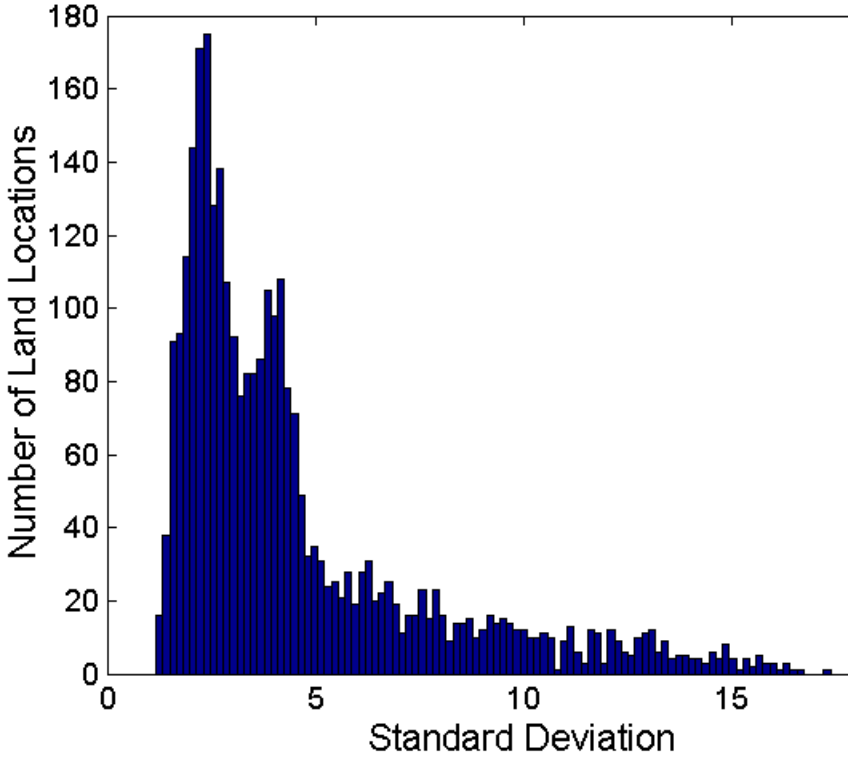
Aggregation

- Combining two or more attributes (or objects) into a single attribute (or object)

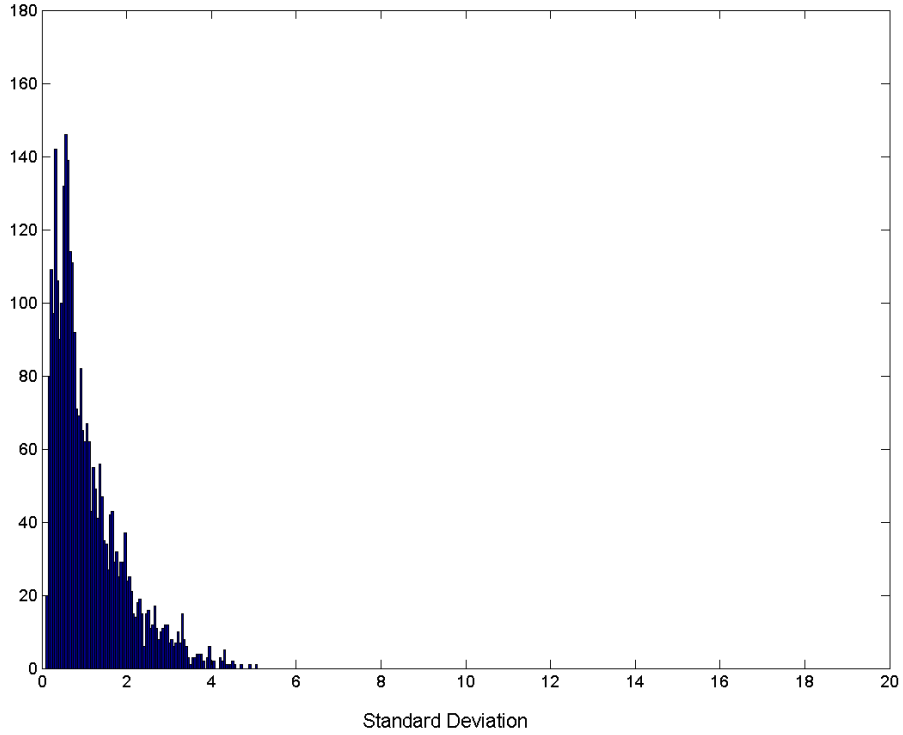
- Purpose
  - Data reduction
    - Reduce the number of attributes or objects
  - Change of scale
    - Cities aggregated into regions, states, countries, etc
  - More “stable” data
    - Aggregated data tends to have less variability

Aggregation

Variation of Precipitation in Australia



Standard Deviation of Average Monthly Precipitation



Standard Deviation of Average Yearly Precipitation

Sampling

- **Sampling is the main technique employed for data selection.**
  - **It is often used for both the preliminary investigation of the data and the final data analysis.**

- **Statisticians sample because obtaining the entire set of data of interest is too expensive or time consuming.**

- **Sampling is used in data mining because processing the entire set of data of interest is too expensive or time consuming.**

Sampling …

- The key principle for effective sampling is the following:

  - **using a sample will work almost as well as using the entire data sets, if the sample is representative**

  (remember this when we shall talk about phot-z's)

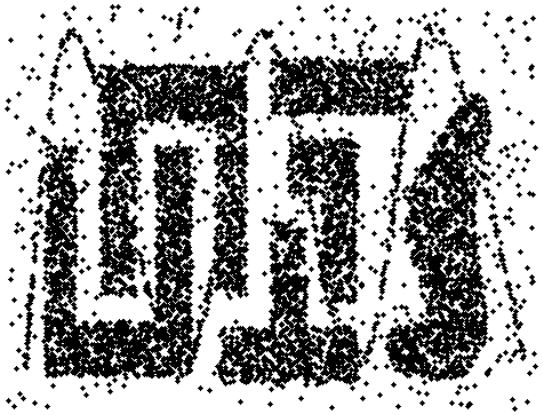
  - A sample is representative if it has approximately the same property (of interest) as the original set of data

    (sometimes this may be verified only a posteriori)

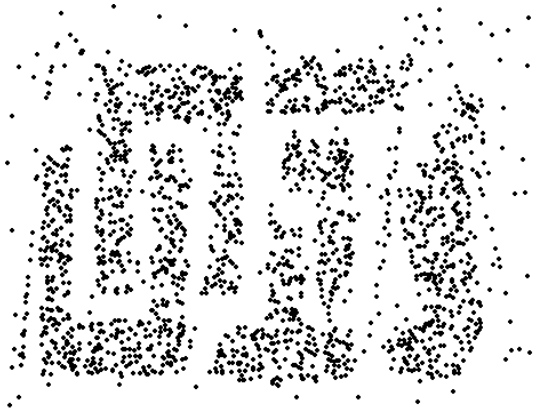
Types of Sampling

- **Simple Random Sampling**
  - There is an equal probability of selecting any particular item
- **Sampling without replacement**
  - As each item is selected, it is removed from the population
- Sampling with replacement
  - Objects are not removed from the population as they are selected for the sample.
    - In sampling with replacement, the same object can be picked up more than once
- **Stratified sampling**
  - Split the data into several partitions; then draw random samples from each partition

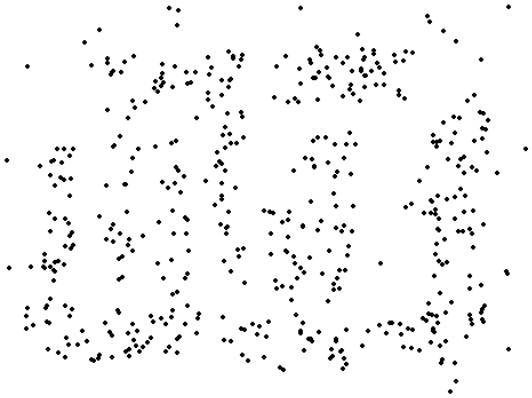
Sample Size matters



8000 points



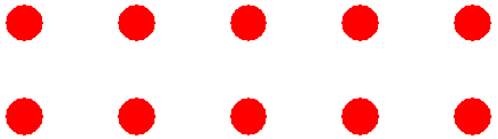
2000 Points

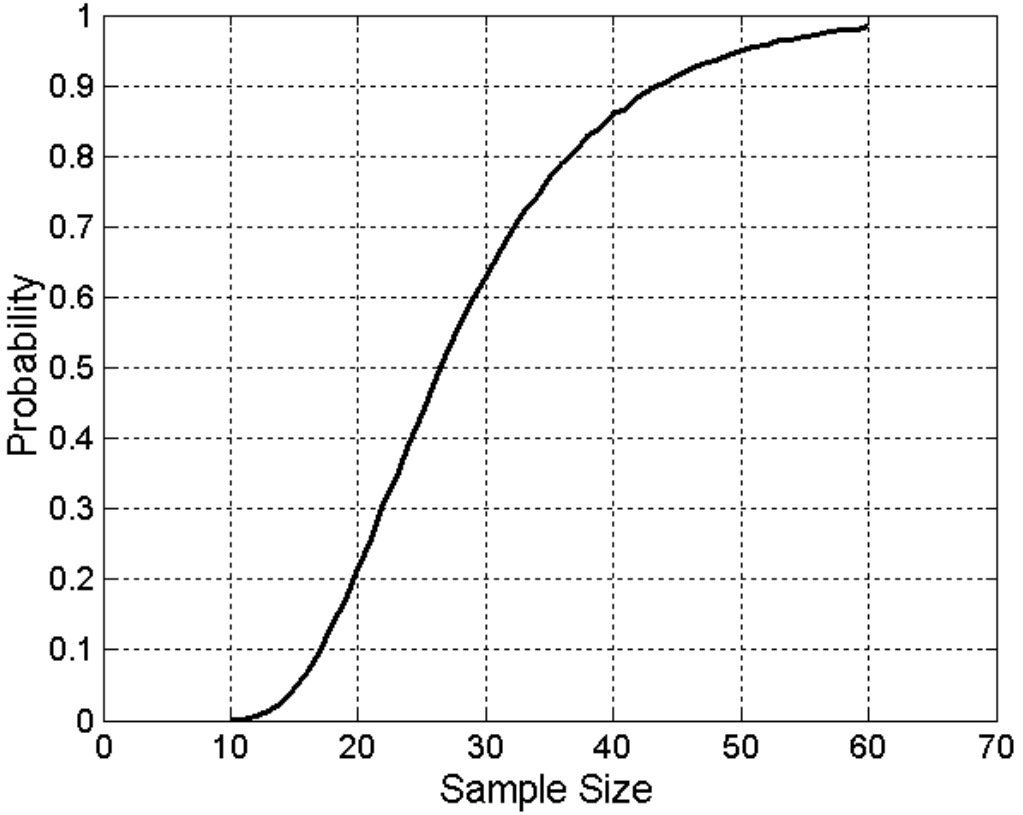


500 Points

Sample Size

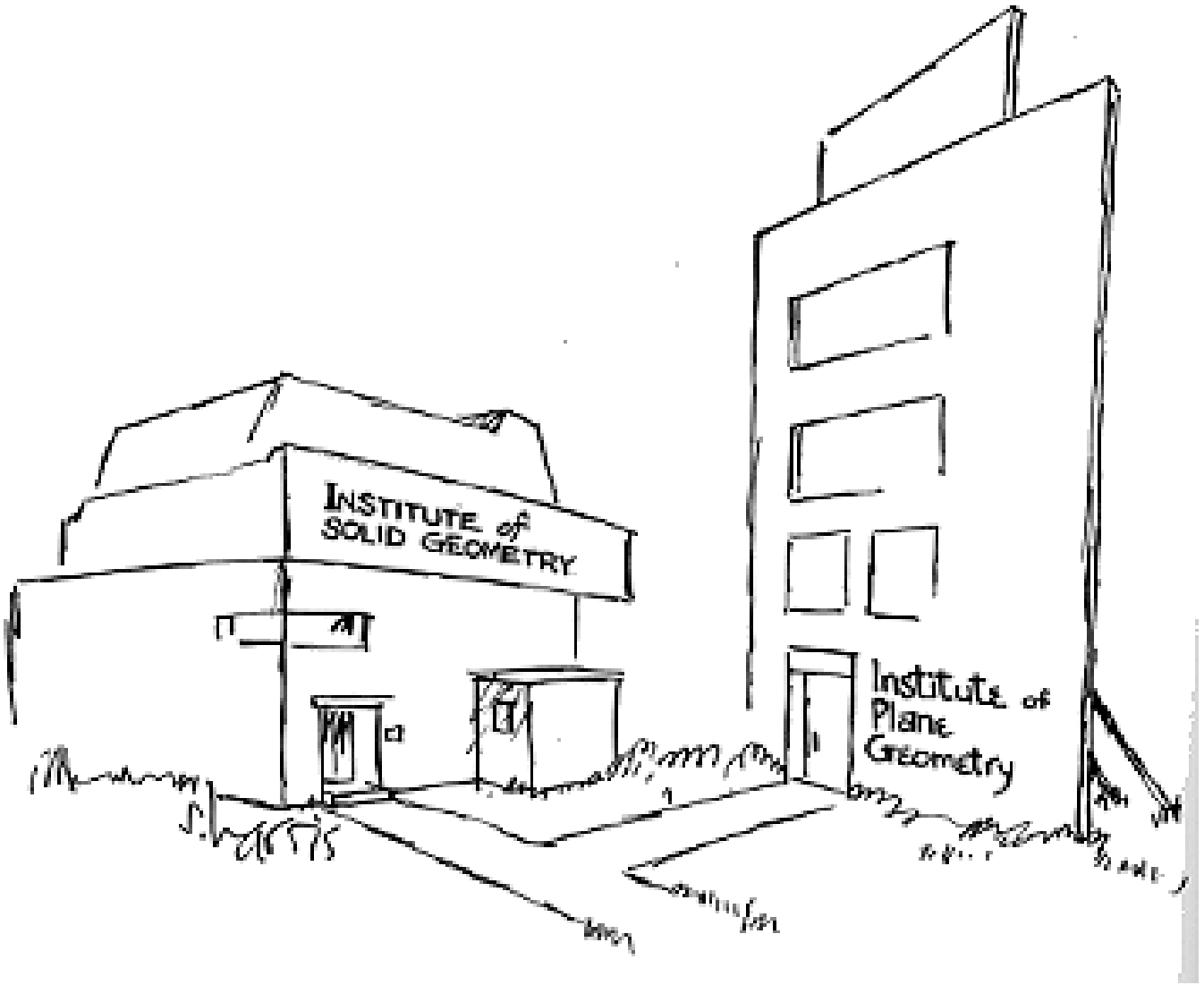
- **What sample size is necessary to get at least one object from each of 10 groups.**





3-D is always better than 2-D

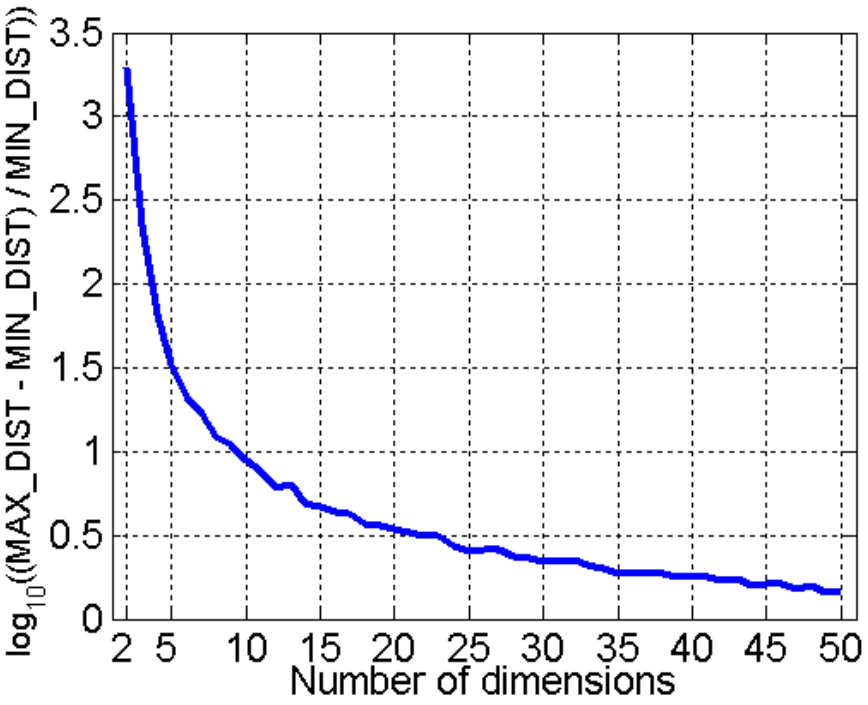
N-D is not always better than (N-1)-D



Curse of Dimensionality (part – II)

- When dimensionality increases (es. Adding more parameters), **data becomes increasingly sparse** in the space that it occupies

- Definitions of density and distance between points, which is critical for clustering and outlier detection, become less meaningful



- Randomly generate 500 points
- Compute difference between max and min distance between any pair of points

Dimensionality Reduction

- Purpose:
  - Avoid curse of dimensionality
  - Reduce amount of time and memory required by data mining algorithms
  - Allow data to be more easily visualized
  - May help to eliminate irrelevant features or reduce noise
- Some Common Techniques
  - Principle Component Analysis
  - Singular Value Decomposition
  - Others: supervised and non-linear techniques

Feature Subset Selection

First way to reduce the dimensionality of data

Redundant features

duplicate much or all of the information contained in one or more other attributes

Example: 3 magnitudes and 2 colors can be represented as 1 magnitude and 2 colors

Irrelevant features

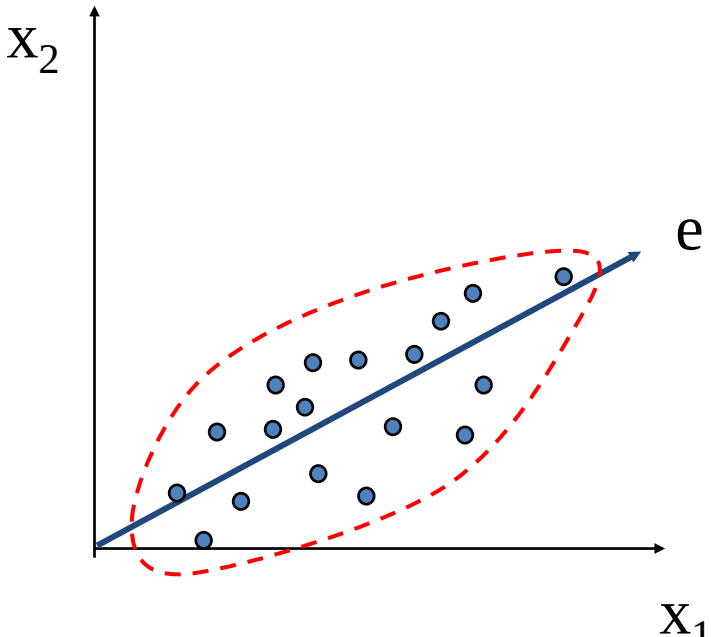
contain no information that is useful for the data mining task at hand … Example: ID is irrelevant to the task of deriving photometric redshifts

Exploratory Data Analysis is crucial.

Refer to the book by Kumar et al.

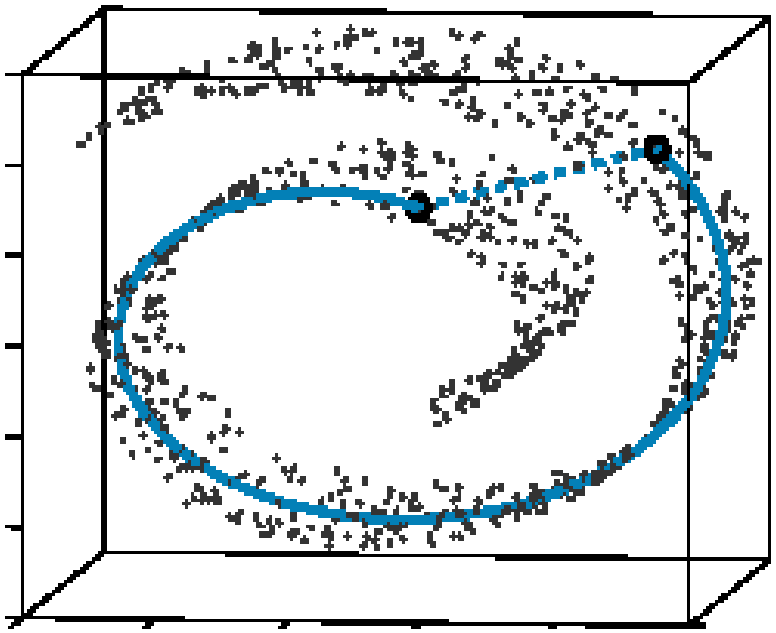
Dimensionality Reduction: PCA

- Find the eigenvectors of the covariance matrix
- The eigenvectors define the new space of lower dimensionality
- Project the data onto this new space



Dimensionality Reduction: ISOMAP

By: Tenenbaum, de Silva, Langford (2000)

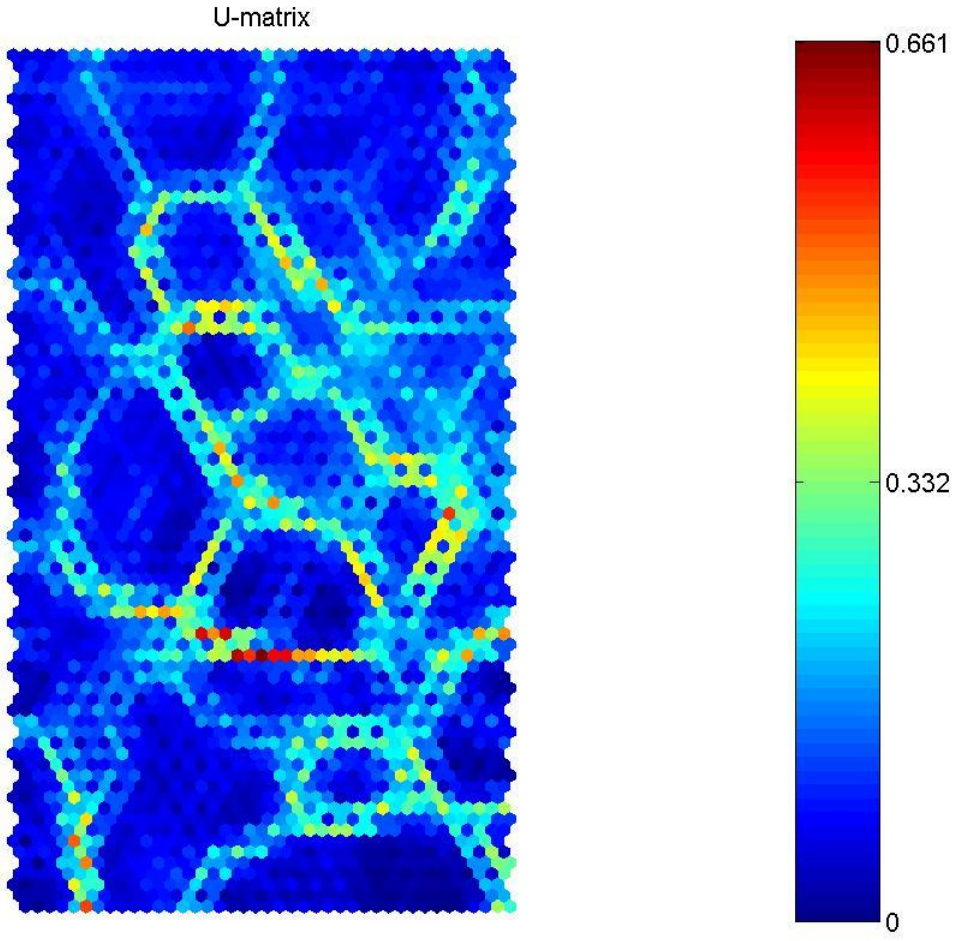


- Construct a neighbourhood graph
- For each pair of points in the graph, compute the shortest path distances – geodesic distances

Feature Subset Selection

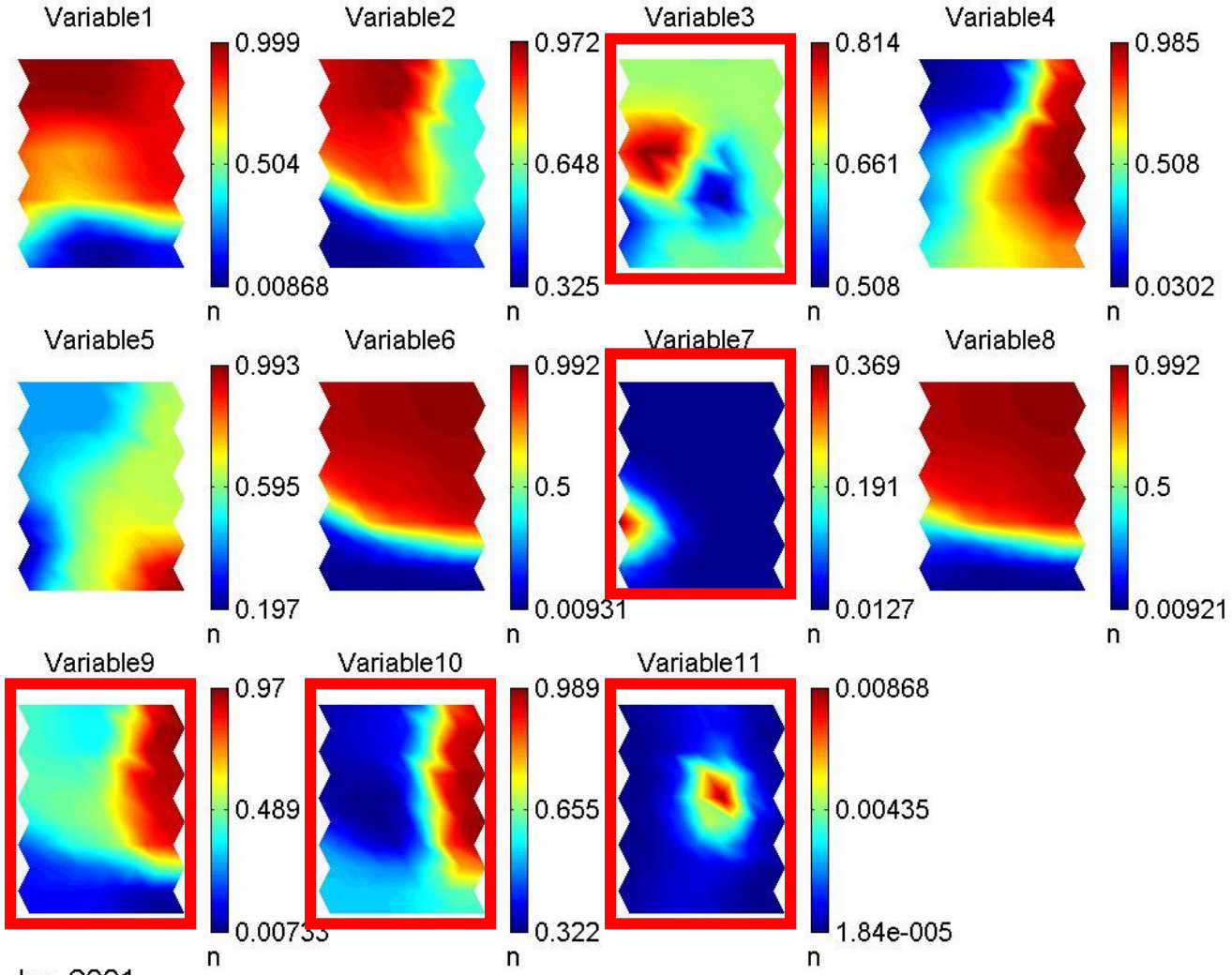
- Techniques:
  - Brute-force approch:
    - Try all possible feature subsets as input to data mining algorithm (backwards elimination strategy)
  - Embedded approaches:
    - Feature selection occurs naturally as part of the data mining algorithm (E.G. SOM)
  - Filter approaches:
    - Features are selected before data mining algorithm is run

SOME DM methods have built in capabilities to operate feature selection



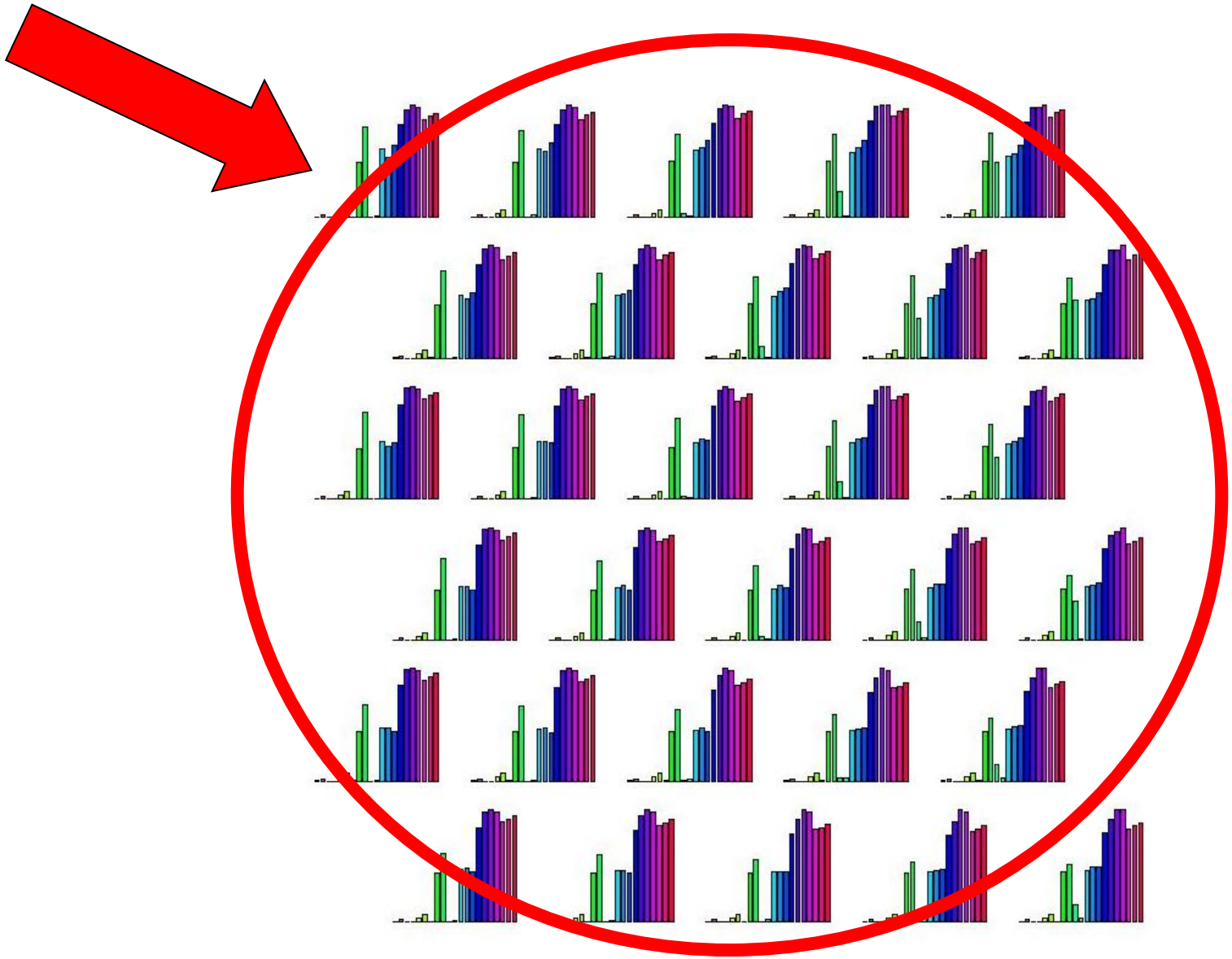
- Regions of low values (blue color) represent clusters themselves
- Regions of high values (red color) represent cluster borders

SOM: U-Matrix



SOM 14-Jun-2001

… bar charts



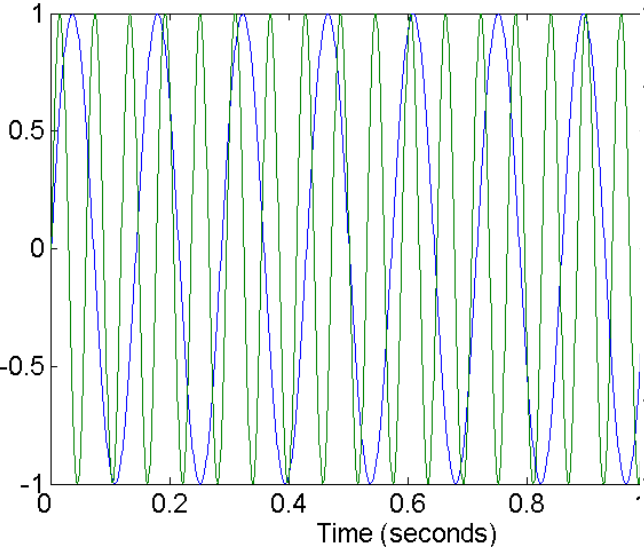
Feature Creation

- Create new attributes that can capture the important information in a data set much more efficiently than the original attributes

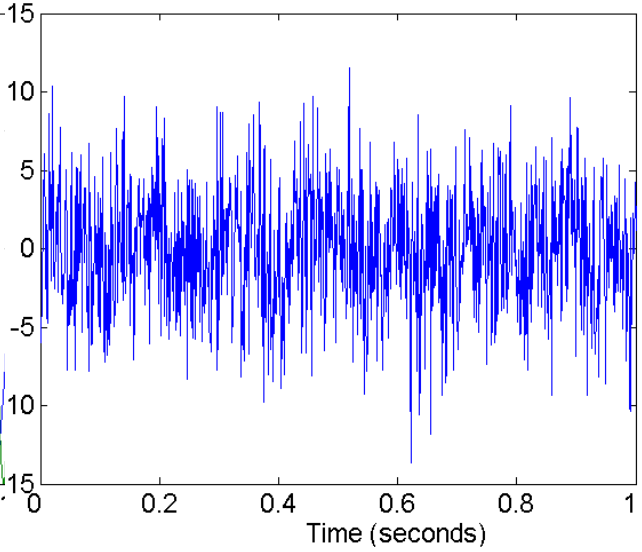
- Three general methodologies:
  - Feature Extraction
    - domain-specific
  - Mapping Data to New Space
  - Feature Construction
    - combining features

Mapping Data to a New Space

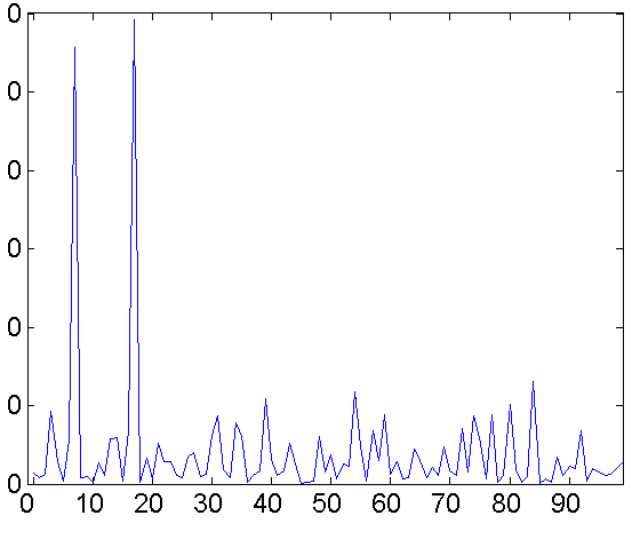
- Fourier transform
- Wavelet transform



Two Sine Waves



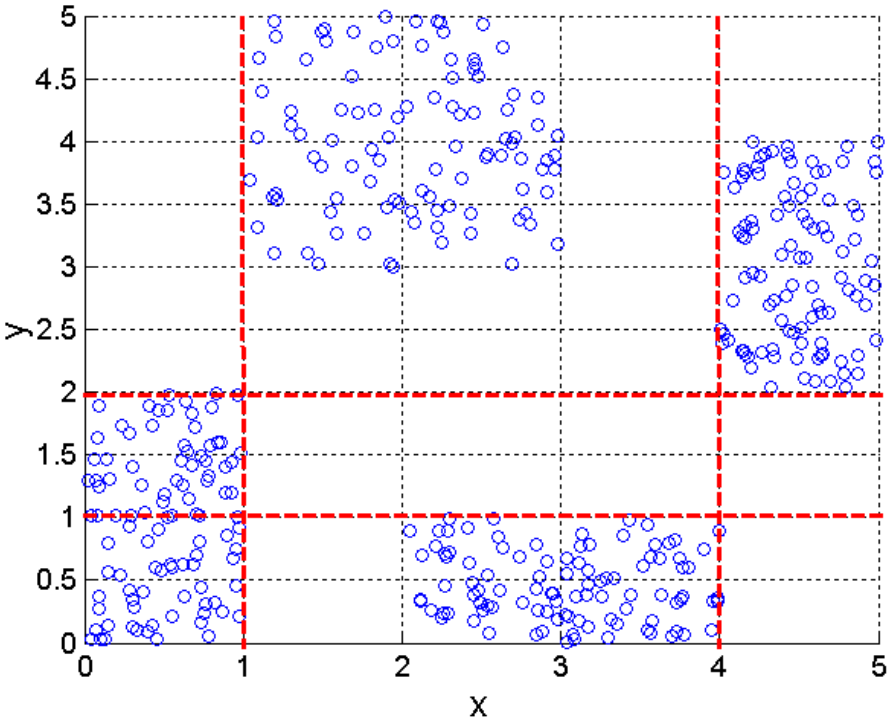
Two Sine Waves + Noise



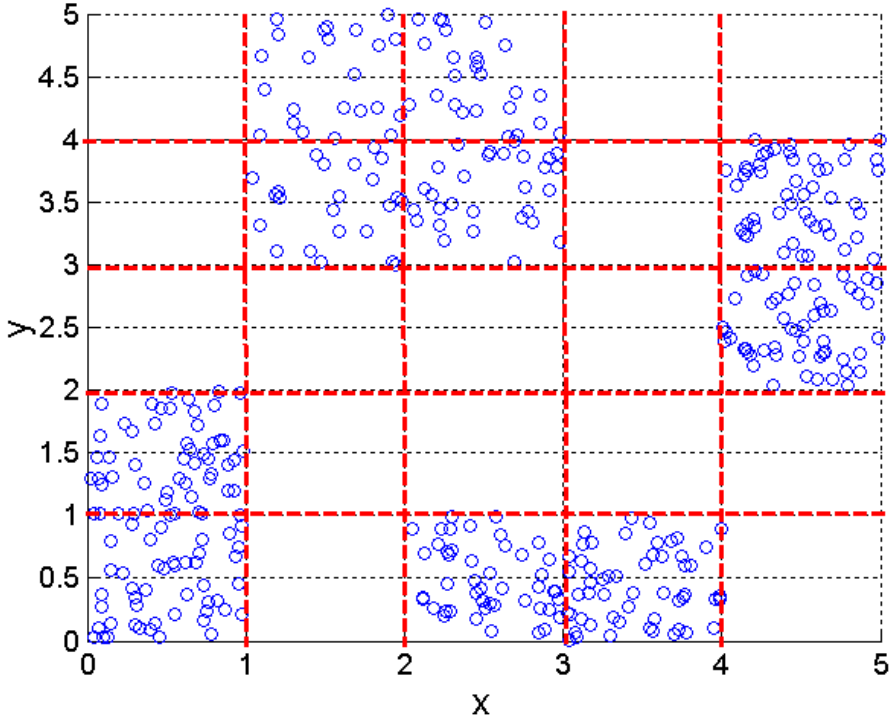
Frequency

Discretization Using Class Labels

- **Entropy based approach (see later in clustering)**



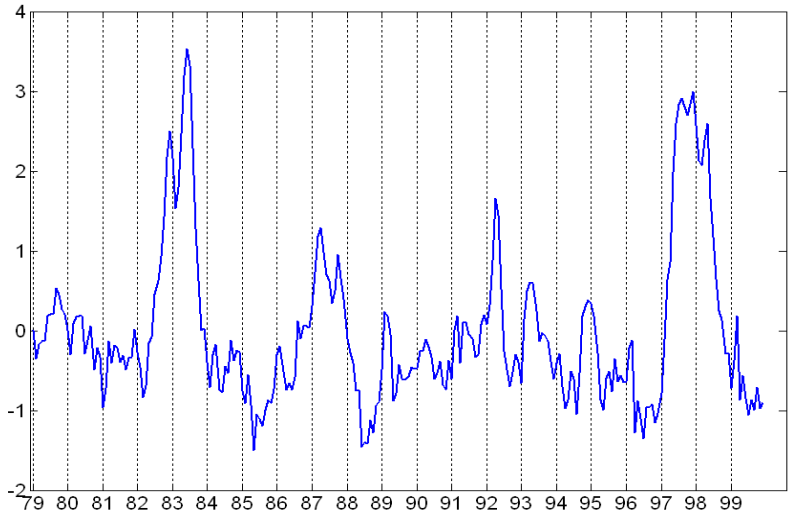
3 categories for both x and y



5 categories for both x and y

Attribute Transformation

- A function that maps the entire set of values of a given attribute to a new set of replacement values such that each old value can be identified with one of the new values
  - Simple functions: x^k, log(x), e^x, |x|
  - Standardization and Normalization



Similarity and Dissimilarity

- Similarity
  - Numerical measure of how alike two data objects are.
  - Is higher when objects are more alike.
  - Often falls in the range [0,1]
- Dissimilarity
  - Numerical measure of how different are two data objects
  - Lower when objects are more alike
  - Minimum dissimilarity is often 0
  - Upper limit varies
- Proximity refers to a similarity or dissimilarity

Similarity/Dissimilarity for Simple Attributes

p and *q* are the attribute values for two data objects.

Attribute Type	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 & \text{if } p = q \\ 1 & \text{if } p \neq q \end{cases}$	$s = \begin{cases} 1 & \text{if } p = q \\ 0 & \text{if } p \neq q \end{cases}$
Ordinal	$d = \frac{\lvert p-q \rvert}{n-1}$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - \frac{\lvert p-q \rvert}{n-1}$
Interval or Ratio	$d = \lvert p - q \rvert$	$s = -d$, $s = \frac{1}{1+d}$ or $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$

Table 5.1. Similarity and dissimilarity for simple attributes

Euclidean Distance

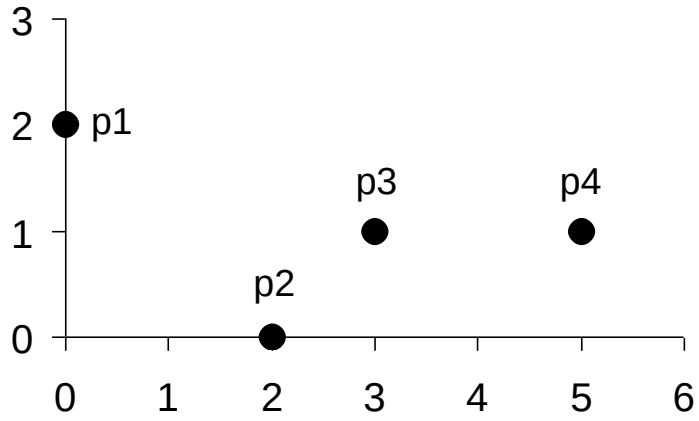
- Euclidean Distance

$$dist = \sqrt{\sum_{k=1}^{n} (p_k - q_k)^2}$$

Where n is the number of dimensions (attributes) and p_k and q_k are, respectively, the kth attributes (components) or data objects p and q.

- Standardization is necessary, if scales differ.

Euclidean Distance



point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

Distance Matrix

Minkowski Distance

- Minkowski Distance is a generalization of Euclidean Distance

$$dist = (\sum_{k=1}^{n} | p_k - q_k |^r)^{\frac{1}{r}}$$

Where r is a parameter, n is the number of dimensions (attributes) and p_k and q_k are, respectively, the kth attributes (components) or data objects p and q.

Minkowski Distance: Examples

- $r = 1$. City block (Manhattan, taxicab, L_1 norm) distance.
  - A common example of this is the Hamming distance, which is just the number of bits that are different between two binary vectors
- $r = 2$. Euclidean distance
- $r \rightarrow \infty$. "supremum" (L_{max} norm, L_∞ norm) distance.
  - This is the maximum difference between any component of the vectors
- Do not confuse *r* with *n*, i.e., all these distances are defined for all numbers of dimensions.

Minkowski Distance

point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

L1	p1	p2	p3	p4
p1	0	4	4	6
p2	4	0	2	4
p3	4	2	0	2
p4	6	4	2	0

L2	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

L_∞	p1	p2	p3	p4
p1	0	2	3	5
p2	2	0	1	3
p3	3	1	0	2
p4	5	3	2	0

Distance Matrix

The drawback is that we assumed that the sample points are distributed isotropically

Were the distribution non-spherical, for instance ellipsoidal, then the probability of the test point belonging to the set depends not only on the distance from the center of mass, but also on the direction.

Putting this on a mathematical basis, in the case of an ellipsoid, the one that best represents the set's probability distribution can be estimated by building the covariance matrix of the samples.

The **Mahalanobis distance** is simply the distance of the test point from the center of mass divided by the width of the ellipsoid in the direction of the test point.

Consider the problem of estimating the probability that a test point in *N*-dimensional Euclidean space belongs to a set, where we are given sample points that definitely belong to that set.

find the average or center of mass of the sample points: the closer the point is to the center of mass, the more likely it is to belong to the set.

However, we also need to know if the set is spread out over a large range or a small range, so that we can decide whether a given distance from the center is noteworthy or not.

The simplistic approach is to estimate the standard deviation of the distances of the sample points from the center of mass.

quantitatively by defining the normalized distance between the test point and the set to be $\frac{x - \mu}{\sigma}$

and plugging this into the normal distribution we can derive the probability of the test point belonging to the set.

Formally, the Mahalanobis distance of a multivariate vector $x = (x_1, x_2, x_3, \ldots, x_N)^T$ from a group of values with mean $\mu = (\mu_1, \mu_2, \mu_3, \ldots, \mu_N)^T$ and covariance matrix S, is defined as:

$$D_M(x) = \sqrt{(x-\mu)^T S^{-1} (x-\mu)}.$$

Mahalanobis distance (or "generalized squared interpoint distance" for its squared value) can also be defined as a dissimilarity measure between two random vectors x and y and of the same distribution with the covariance matrix S:

$$d(\vec{x}, \vec{y}) = \sqrt{(\vec{x}-\vec{y})^T S^{-1} (\vec{x}-\vec{y})}.$$

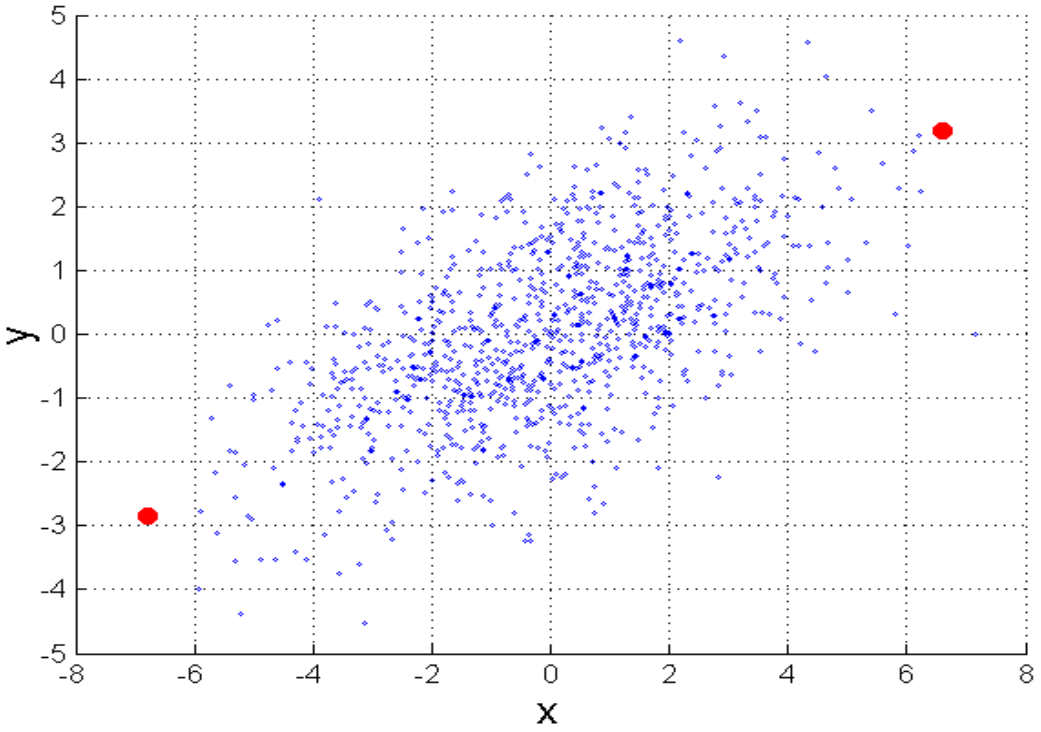
If the covariance matrix is the identity matrix, the Mahalanobis distance reduces to the Euclidean distance. If the covariance matrix is diagonal, then the resulting distance measure is called the *normalized Euclidean distance*:

$$d(\vec{x}, \vec{y}) = \sqrt{\sum_{i=1}^{N} \frac{(x_i - y_i)^2}{\sigma_i^2}},$$

where σ_i is the standard deviation of the x_i over the sample set..

Mahalanobis Distance

$$\boldsymbol{mahalanobis}(\boldsymbol{p},\boldsymbol{q}) = (\boldsymbol{p}-\boldsymbol{q})\Sigma^{-1}(\boldsymbol{p}-\boldsymbol{q})^{\boldsymbol{T}}$$

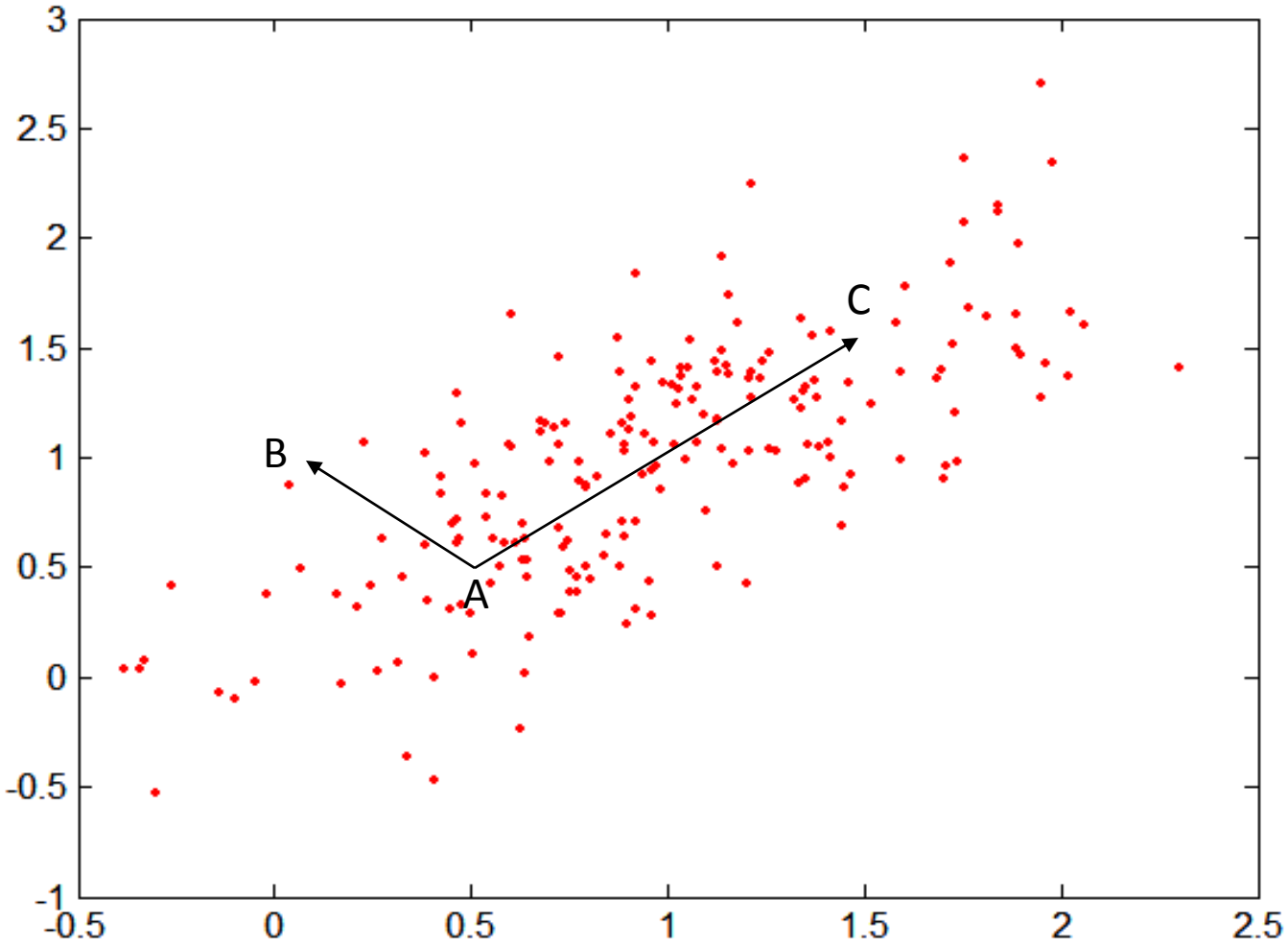


Σ is the covariance matrix of the input data X

$$\Sigma_{j,k} = \frac{1}{n-1}\sum_{i=1}^{n}(X_{ij} - \overline{X}_j)(X_{ik} - \overline{X}_k)$$

For red points, the Euclidean distance is 14.7, Mahalanobis distance is 6.

Mahalanobis Distance



Covariance Matrix:

$$\Sigma = \begin{bmatrix} 0.3 & 0.2 \\ 0.2 & 0.3 \end{bmatrix}$$

A: (0.5, 0.5)

B: (0, 1)

C: (1.5, 1.5)

Mahal(A,B) = 5

Mahal(A,C) = 4

Common Properties of a Distance

- Distances, such as the Euclidean distance, have some well known properties.

1. $d(p, q) \geq 0$ for all p and q and $d(p, q) = 0$ only if $p = q$. (Positive definiteness)
2. $d(p, q) = d(q, p)$ for all p and q. (Symmetry)
3. $d(p, r) \leq d(p, q) + d(q, r)$ for all points p, q, and r. (Triangle Inequality)

where $d(p, q)$ is the distance (dissimilarity) between points (data objects), p and q.

- A distance that satisfies these properties is a metric

Common Properties of a Similarity

- Similarities, also have some well known properties.

1. *s(p, q) = 1* (or maximum similarity) only if *p = q*.

2. *s(p, q) = s(q, p)* for all *p* and *q*. (Symmetry)

where *s(p, q)* is the similarity between points (data objects), *p* and *q*.

Similarity Between Binary Vectors

- Common situation is that objects, *p* and *q*, have only binary attributes
- Compute similarities using the following quantities

  M_{01} = the number of attributes where p was 0 and q was 1
  M_{10} = the number of attributes where p was 1 and q was 0
  M_{00} = the number of attributes where p was 0 and q was 0
  M_{11} = the number of attributes where p was 1 and q was 1
- Simple Matching and Jaccard Coefficients

  SMC = number of matches / number of attributes
  = $(M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00})$

  J = number of 11 matches / number of not-both-zero attributes values
  = $(M_{11}) / (M_{01} + M_{10} + M_{11})$

Cosine Similarity

- If d_1 and d_2 are two document vectors, then

$$\cos(d_1, d_2) = (d_1 \bullet d_2) / ||d_1|| \, ||d_2|| ,$$

where • indicates vector dot product and $||d||$ is the length of vector d.

- Example:

d_1 = **3 2 0 5 0 0 0 2 0 0**

d_2 = **1 0 0 0 0 0 0 1 0 2**

$d_1 \bullet d_2$= 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5

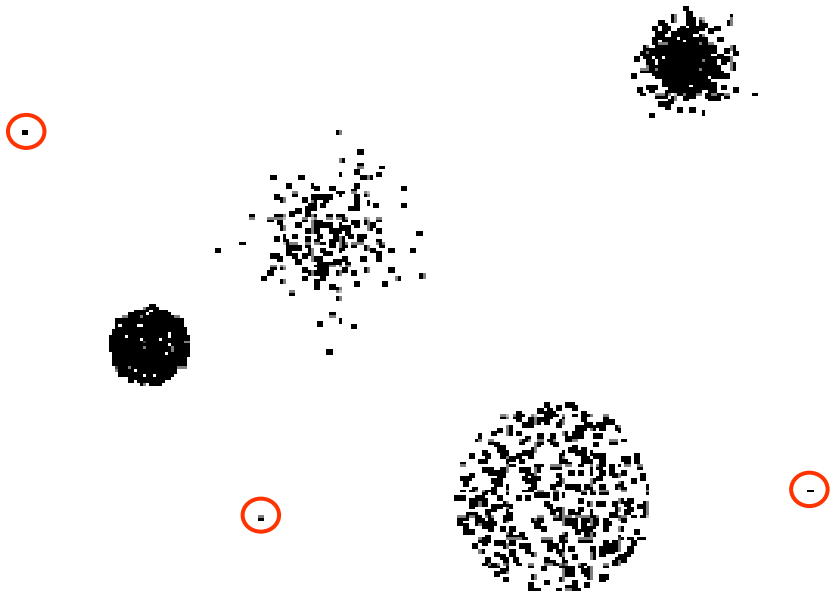
$||d_1||$ = $(3*3+2*2+0*0+5*5+0*0+0*0+0*0+2*2+0*0+0*0)^{0.5}$ = $(42)^{0.5}$ = 6.481

$||d_2||$ = $(1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2)^{0.5}$ = $(6)^{0.5}$ = 2.245

$$\cos(d_1, d_2) = .3150$$

Outliers

- Outliers are data objects with characteristics that are considerably different than most of the other data objects in the data set



Sometimes attributes are of many different types, but an overall similarity is needed.

1. For the k^{th} attribute, compute a similarity, s_k, in the range $[0, 1]$.

2. Define an indicator variable, δ_k, for the k_{th} attribute as follows:

$$\delta_k = \begin{cases} 0 & \text{if the } k^{th} \text{ attribute is a binary asymmetric attribute and both objects have a value of 0, or if one of the objects has a missing values for the } k^{th} \text{ attribute} \\ 1 & \text{otherwise} \end{cases}$$

3. Compute the overall similarity between the two objects using the following formula:

$$similarity(p, q) = \frac{\sum_{k=1}^{n} \delta_k s_k}{\sum_{k=1}^{n} \delta_k}$$

Using Weights to Combine Similarities

- May not want to treat all attributes the same.
  - Use weights w_k which are between 0 and 1 and sum to 1.

$$similarity(p,q) = \frac{\sum_{k=1}^{n} w_k \delta_k s_k}{\sum_{k=1}^{n} \delta_k}$$

$$distance(p,q) = \left(\sum_{k=1}^{n} w_k |p_k - q_k|^r \right)^{1/r}$$

Density

- Density-based clustering require a notion of density

- Examples:
  - Euclidean density
    - Euclidean density = number of points per unit volume
  - Probability density
  - Graph-based density

Euclidean Density – Cell-based

- Simplest approach is to divide region into a number of rectangular cells of equal volume

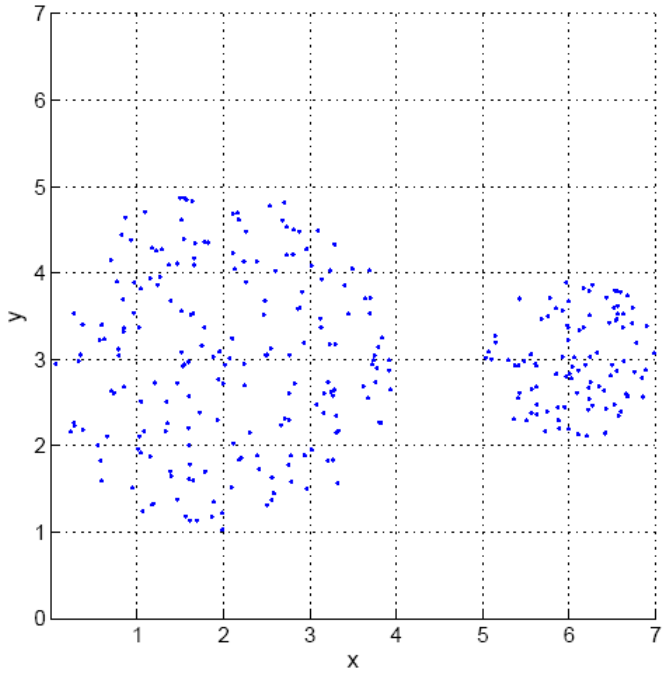


Figure 7.13. Cell-based density.

0	0	0	0	0	0	0
0	0	0	0	0	0	0
4	17	18	6	0	0	0
14	14	13	13	0	18	27
11	18	10	21	0	24	31
3	20	14	4	0	0	0
0	0	0	0	0	0	0

Table 7.6. Point counts for each grid cell.

Euclidean Density – Center-based

- Euclidean density is the number of points within a specified radius of the point

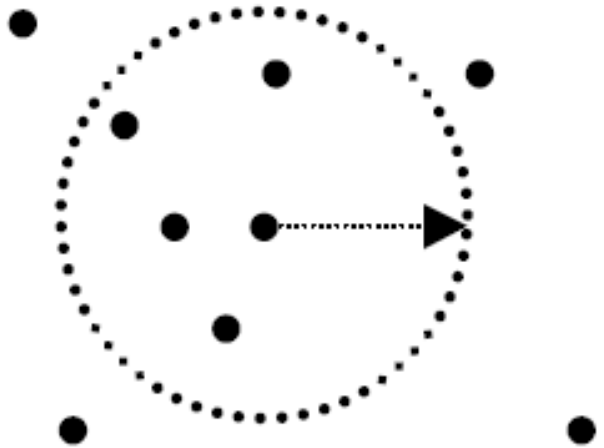


Figure 7.14. Illustration of center-based density.