OUTLIER MINING IN HIGH DIMENSIONAL DATASETS

DATA MINING DISCUSSION GROUP

OUTLINE

- □ MOTIVATION
- □OUTLIERS IN MULTIVARIATE DATA
- □ OUTLIERS IN HIGH DIMENSIONAL DATA
  - □ Distribution-based
  - □ Distance-based
    - □ NN-based
    - □ Density-based
    - □Clustering-based
  - □ Depth-based methods
- □ CONCLUSION

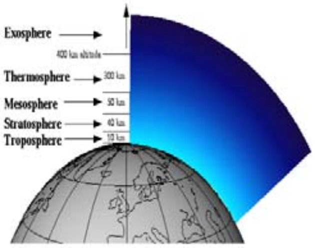
MOTIVATION:

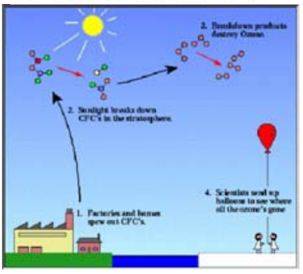
CFC has ability to breakdown ozone

- Estimation in 75's: 7% drop within 60yrs.

- In 1985, British Antarctic Survey showed that ozone levels had dropped to 10% below

- WHY?

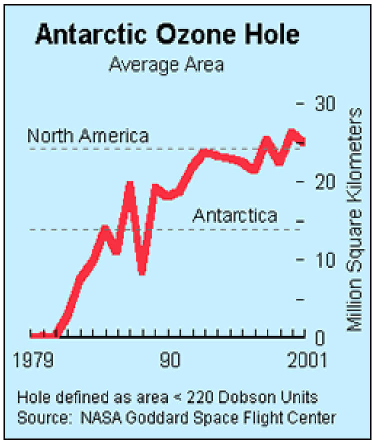



3. Breakdown products destroy Ozone
2. Sunlight breaks down CFC's in the stratosphere
1. Factories and homes spew out CFC's
4. Scientists send up balloons to see where all the ozone's gone

What was wrong?

- Ozone hole had been covered up by a computer-program

- Evidence of the ozone-hole was seen as far back as 1976.

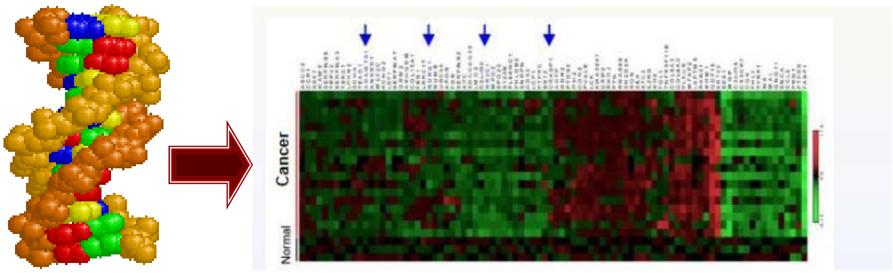


Sources:
http://exploringdata.cqu.edu.au/ozone.html
http://www.epa.gov/ozone/science/hole/size.html

MOTIVATION

Colon Data (Alon et al, 1999)
X : 62×2000, gene expression data,

The colon cancer dataset is known to be heterogeneous because the tissue samples contain a mixture of cell types!!!

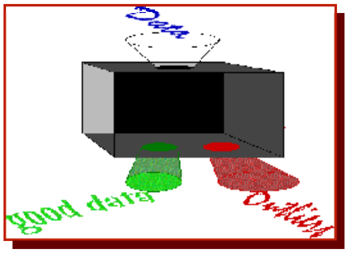


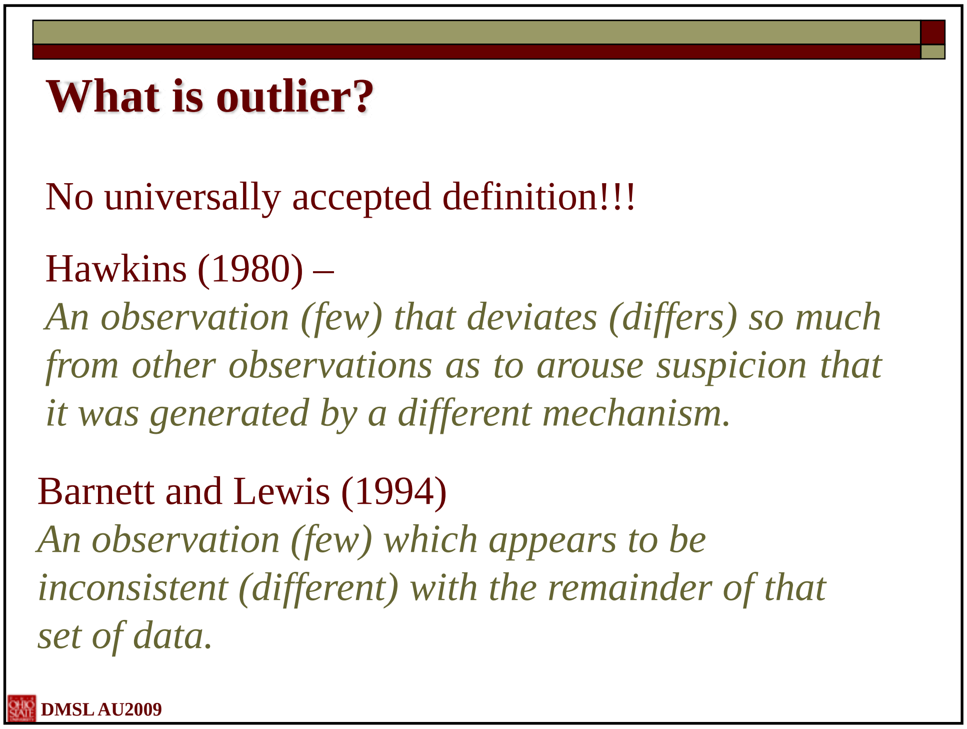
MOTIVATION

- Data often (always) contain outliers.

- Statistical methods are severely affected by outliers.

- We have to identify OUTLIER (s) accurately!!!



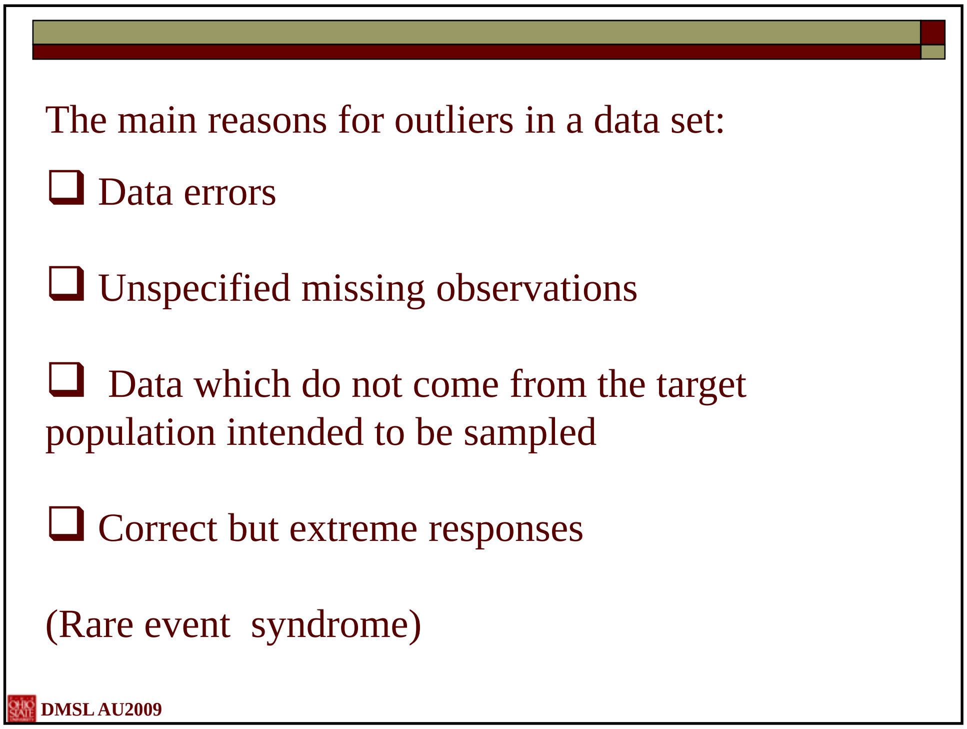


What is outlier?

No universally accepted definition!!!

Hawkins (1980) –
An observation (few) that deviates (differs) so much from other observations as to arouse suspicion that it was generated by a different mechanism.

Barnett and Lewis (1994)
An observation (few) which appears to be inconsistent (different) with the remainder of that set of data.



The main reasons for outliers in a data set:

❑ Data errors

❑ Unspecified missing observations

❑ Data which do not come from the target population intended to be sampled

❑ Correct but extreme responses

(Rare event syndrome)

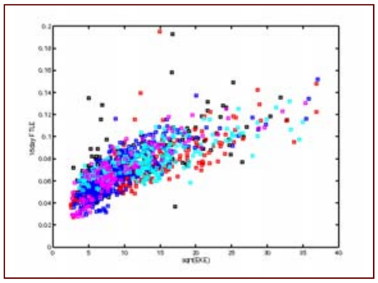
A few applications of outlier detection:

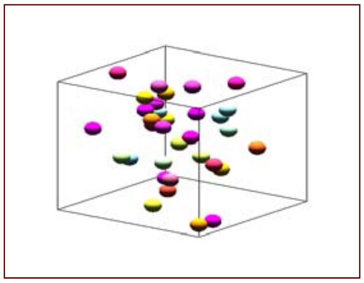
- Fraud detection
- Network intrusion detection
- Satellite image analysis
- Structural defect detection
- Loan application processing
- Discovery of astronomical objects
- Motion segmentation
- Detection of unexpected entries in databases
- And many more…

OUTLIERS IN MULTIVARIATE DATA

Visual tools

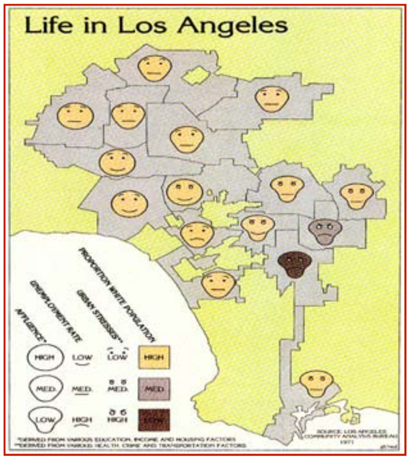
Scatter plots and 3D scatter plots

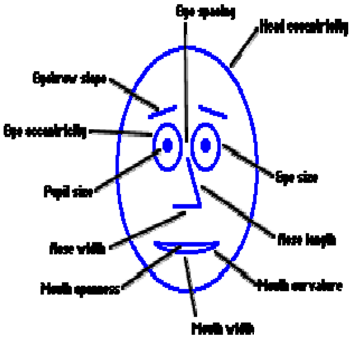




Higher dimensions???

OUTLIERS IN MULTIVARIATE DATA


Life in Los Angeles
PROPORTION WHITE POPULATION
URBAN STRESSES**
UNEMPLOYMENT RATE
AFFLUENCE*
HIGH
LOW
LOW
HIGH
MED
MED
MED
MED
LOW
HIGH
HIGH
LOW
*DERIVED FROM VARIOUS EDUCATION, INCOME AND HOUSING FACTORS
**DERIVED FROM VARIOUS HEALTH, CRIME AND TRANSPORTATION FACTORS
SOURCE: LOS ANGELES COMMUNITY ANALYSIS BUREAU 1971

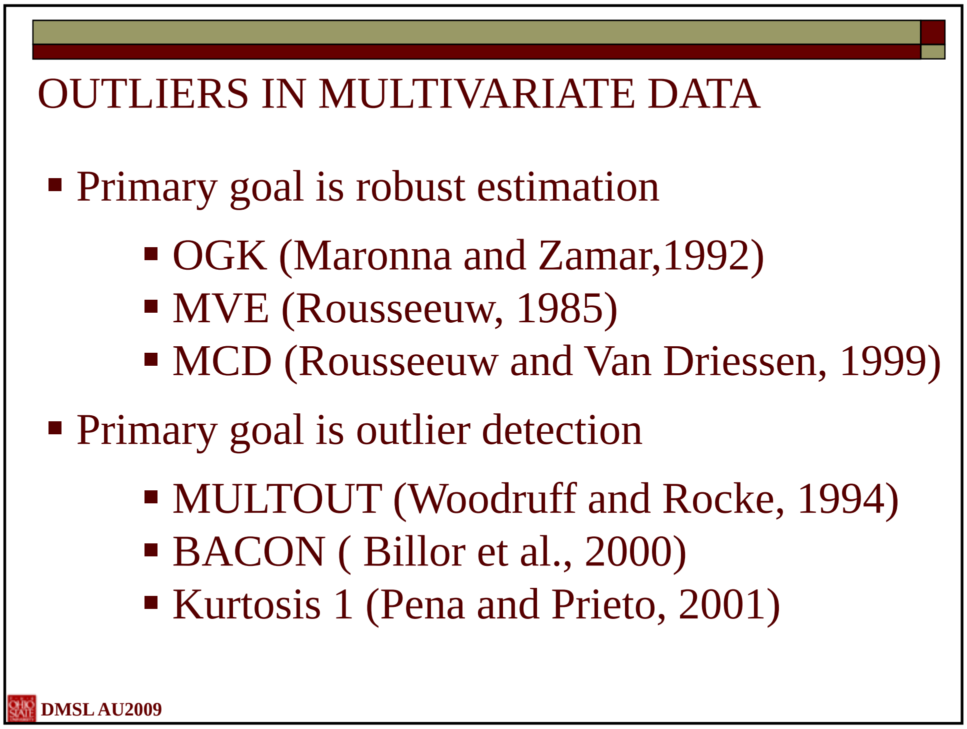


OUTLIERS IN MULTIVARIATE DATA

Numerical methods aim to detect outliers by computing a measure of how far a particular point is from the center of the data.

The usual measure of “outlyingness” for a data point, x_i *is the Mahalanobis distance:*

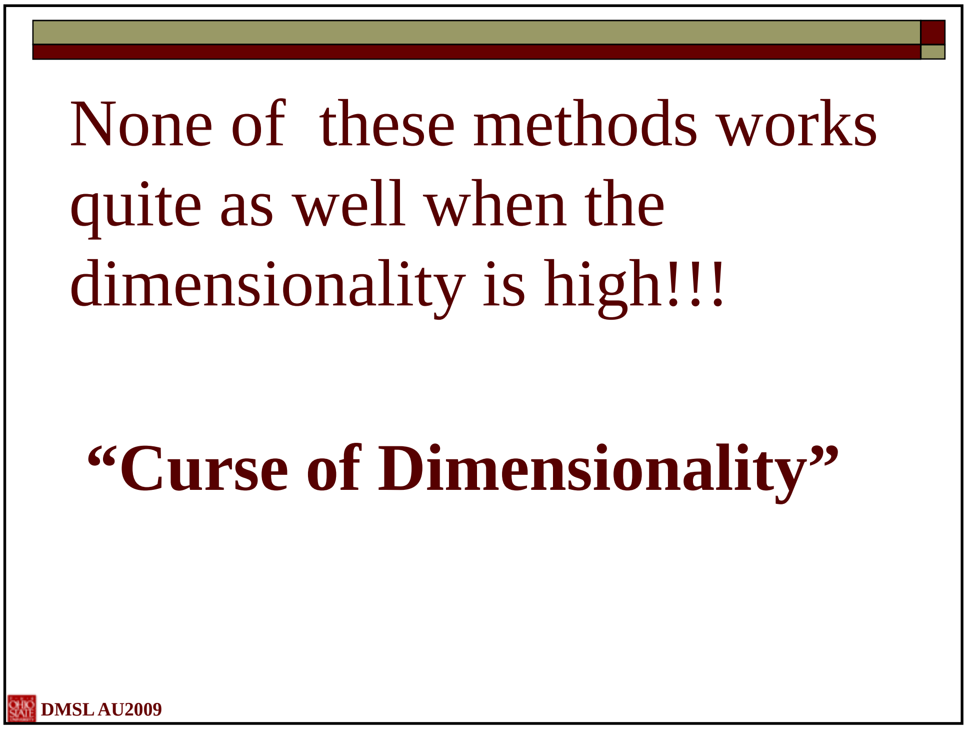
$$\boxed{D_i = \sqrt{(x_i - \overline{X})' S^{-1} (x_i - \overline{X})}, \ \mathrm{i} = 1,2,\ldots,\mathrm{N}}$$



OUTLIERS IN MULTIVARIATE DATA

- Primary goal is robust estimation
  - OGK (Maronna and Zamar,1992)
  - MVE (Rousseeuw, 1985)
  - MCD (Rousseeuw and Van Driessen, 1999)
- Primary goal is outlier detection
  - MULTOUT (Woodruff and Rocke, 1994)
  - BACON (Billor et al., 2000)
  - Kurtosis 1 (Pena and Prieto, 2001)

DMSL AU2009



None of these methods works quite as well when the dimensionality is high!!!

“Curse of Dimensionality”

DMSL AU2009

OUTLIERS IN HIGH DIMENSIONAL DATA

(Hodge et al., 2004 ; Lazarevic et al., 2000)

- Distribution-based
- Distance-based
  - NN based
  - Density-based
  - Cluster-based
- Depth-Based

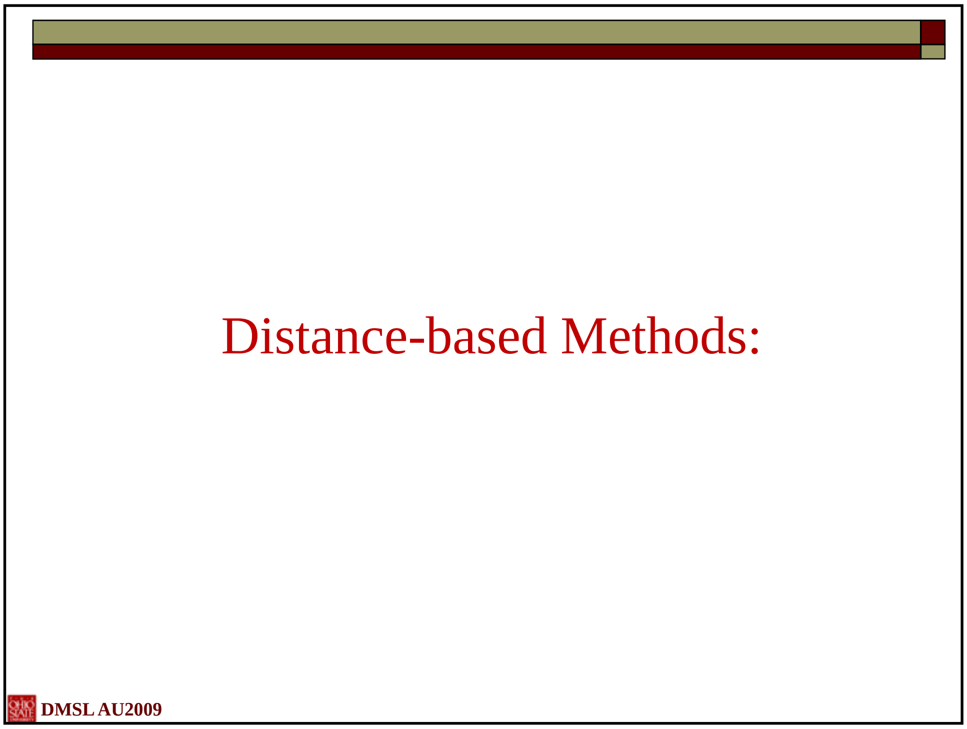
Distribution-based approaches

(Barnett&Lewis, 1994; Hawkins, 1980)

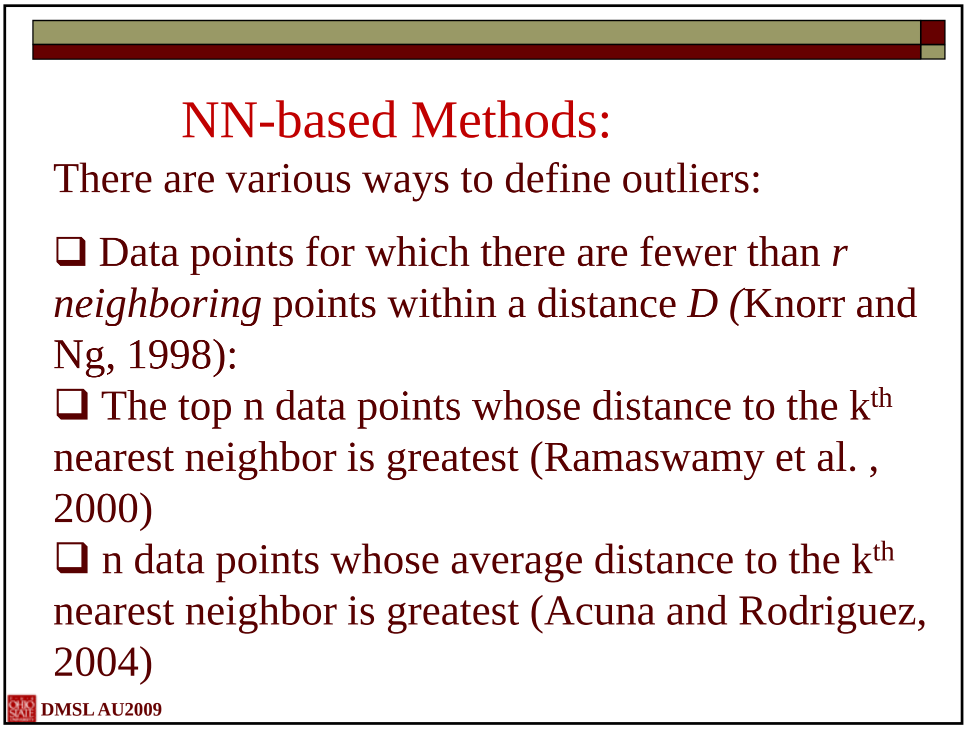
- Data points are modeled using a stochastic distribution
- Outliers are observations which deviate from the given distribution.

Drawbacks:

- Unsuitable even for moderately high-dimensional data sets.
- Perform expensive tests to determine which model fits the data best, if any!



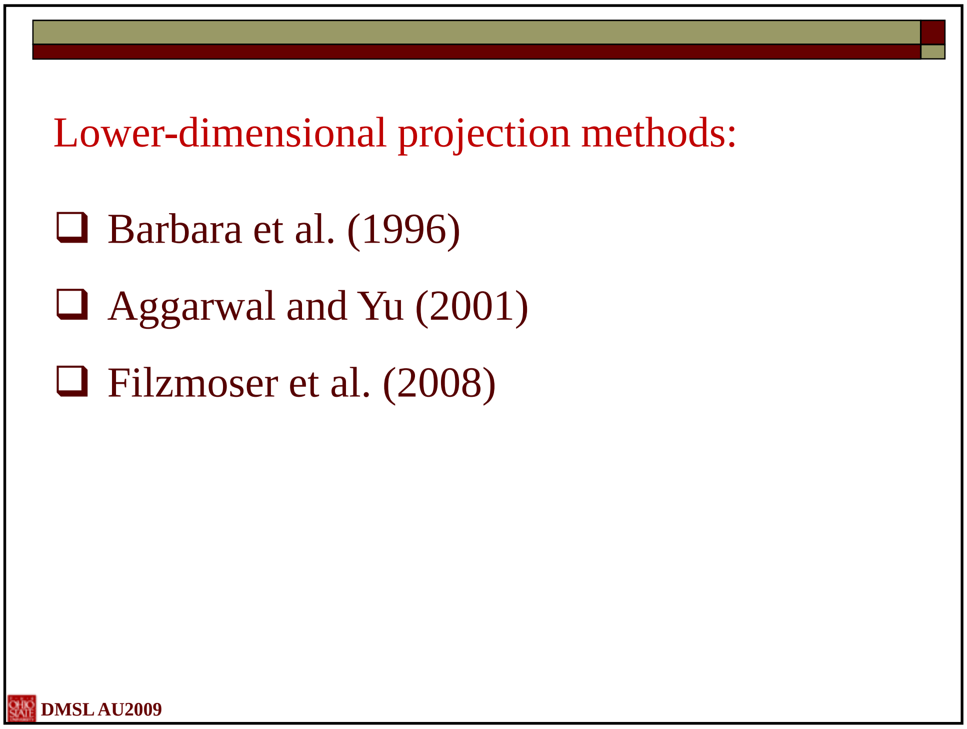
Distance-based Methods:



NN-based Methods:

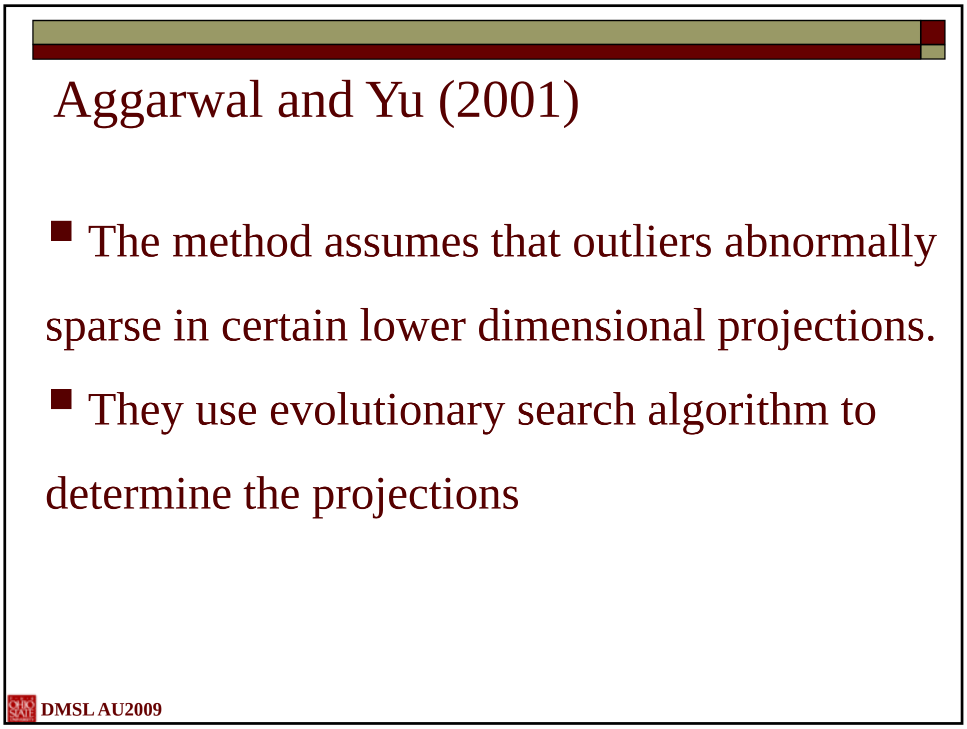
There are various ways to define outliers:

- ❑ Data points for which there are fewer than *r neighboring* points within a distance *D (*Knorr and Ng, 1998):
- ❑ The top n data points whose distance to the kth nearest neighbor is greatest (Ramaswamy et al. , 2000)
- ❑ n data points whose average distance to the kth nearest neighbor is greatest (Acuna and Rodriguez, 2004)



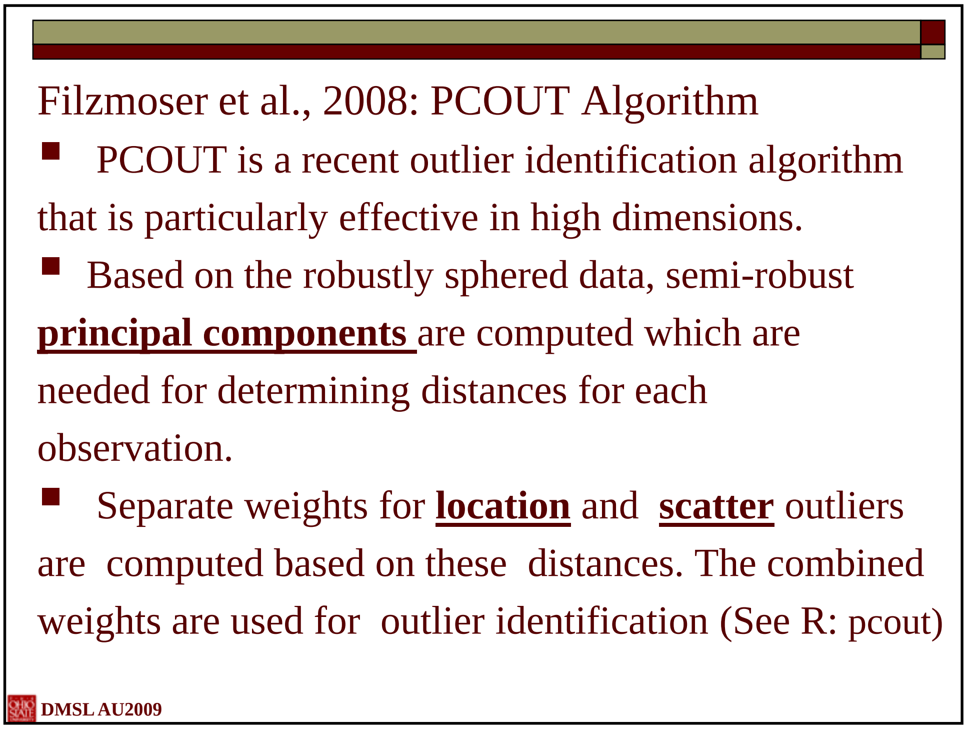
Lower-dimensional projection methods:

- ❑ Barbara et al. (1996)
- ❑ Aggarwal and Yu (2001)
- ❑ Filzmoser et al. (2008)



Aggarwal and Yu (2001)

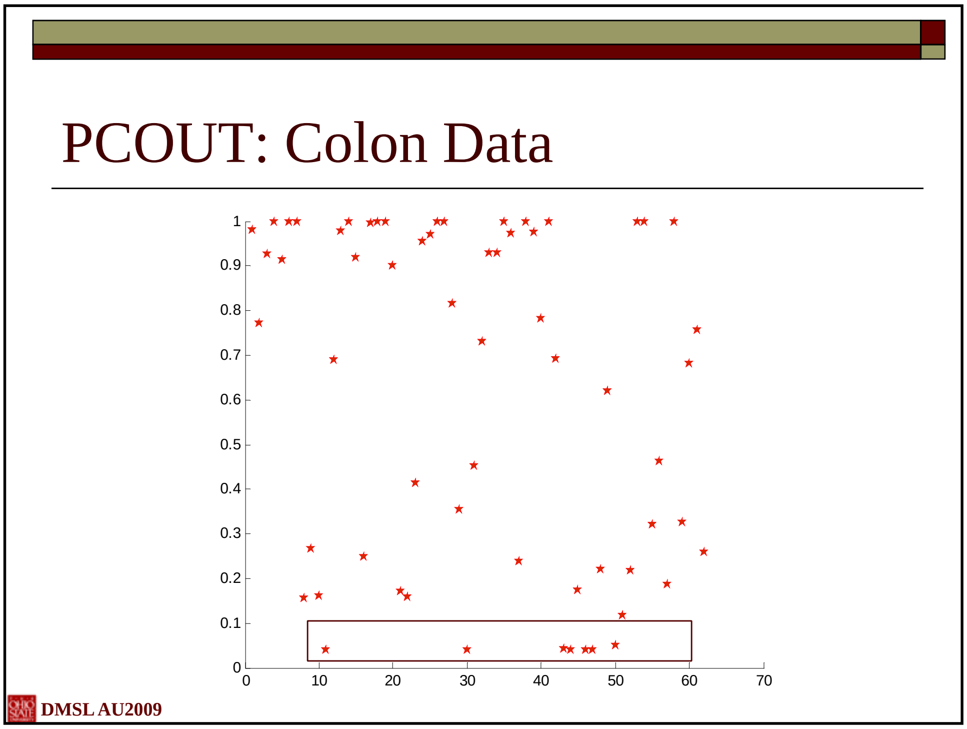
- ■ The method assumes that outliers abnormally sparse in certain lower dimensional projections.
- ■ They use evolutionary search algorithm to determine the projections



Filzmoser et al., 2008: PCOUT Algorithm

- PCOUT is a recent outlier identification algorithm that is particularly effective in high dimensions.
- Based on the robustly sphered data, semi-robust **principal components** are computed which are needed for determining distances for each observation.
- Separate weights for **location** and **scatter** outliers are computed based on these distances. The combined weights are used for outlier identification (See R: pcout)

DMSL AU2009

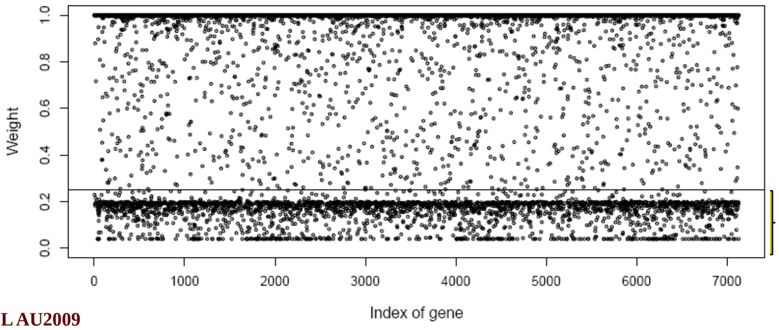


PCOUT: Colon Data

DMSL AU2009

PCOUT: Leukemia Data (72x7129)

We will try to identify multivariate outliers among the 7129 genes, without using the information of the two leukemia types ALL and AML.



2609 genes

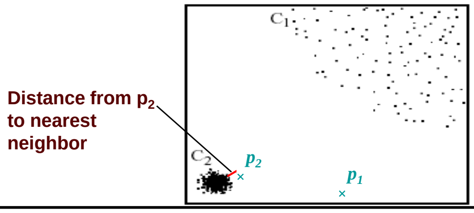
Density based Methods:

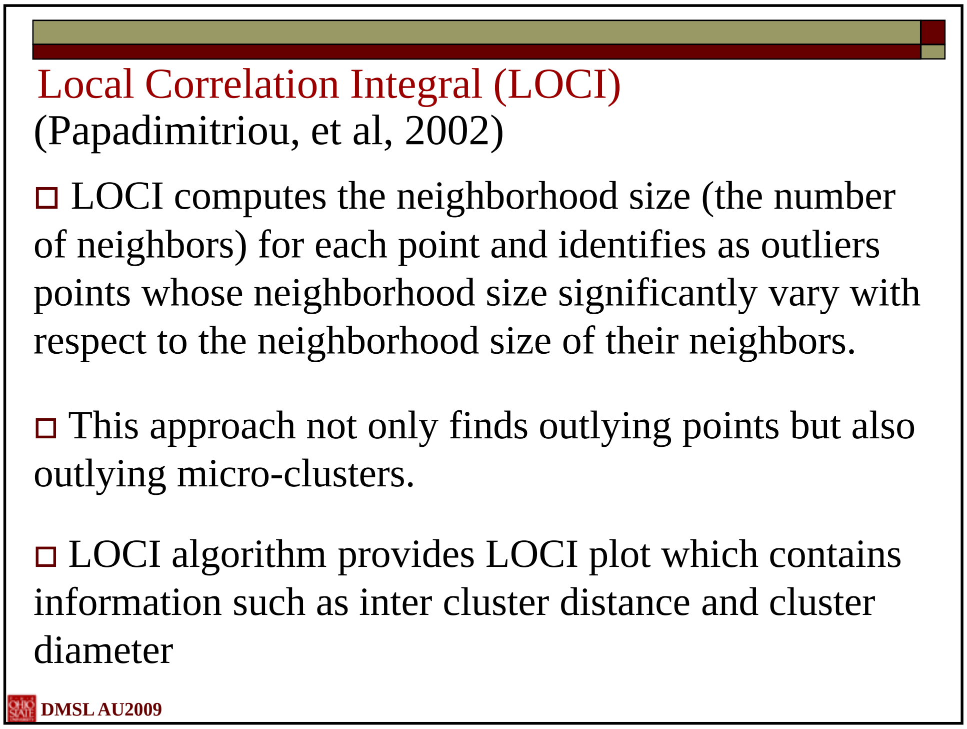
Local Outlying Factor: LOF (Breunig et al., 2000)

❑ For each point, compute the density of its local neighborhood

❑Compute local outlier factor (LOF) of a sample *p as the* average of the ratios of the density of sample *p and the* density of its nearest neighbors

❑Outliers are points with largest LOF value





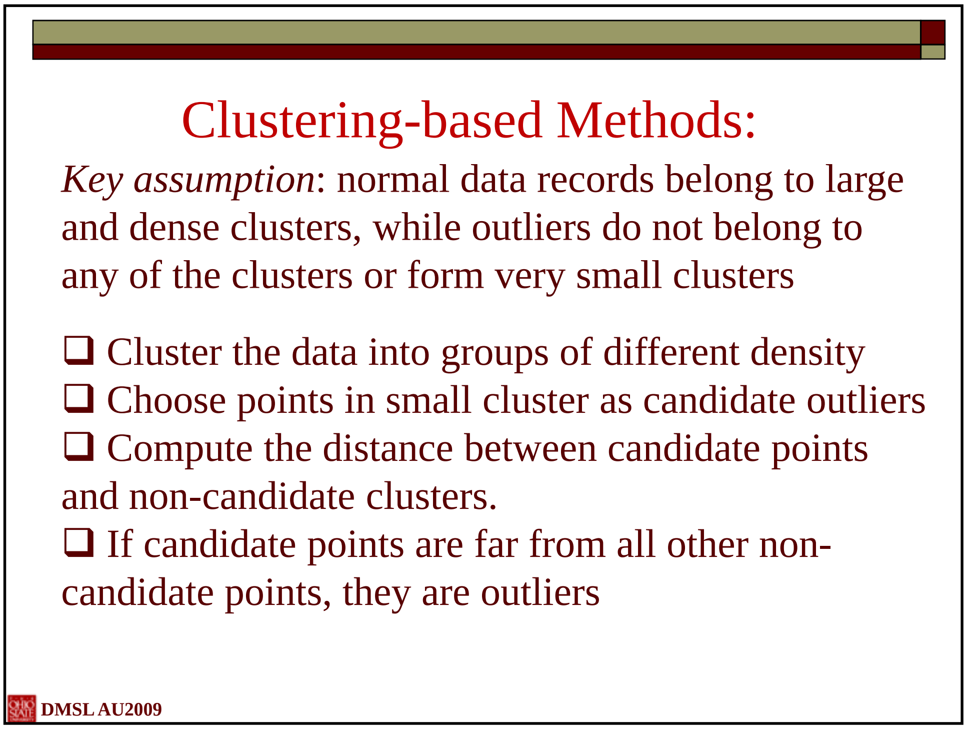
Local Correlation Integral (LOCI)

(Papadimitriou, et al, 2002)

□ LOCI computes the neighborhood size (the number of neighbors) for each point and identifies as outliers points whose neighborhood size significantly vary with respect to the neighborhood size of their neighbors.

□ This approach not only finds outlying points but also outlying micro-clusters.

□ LOCI algorithm provides LOCI plot which contains information such as inter cluster distance and cluster diameter



Clustering-based Methods:

Key assumption: normal data records belong to large and dense clusters, while outliers do not belong to any of the clusters or form very small clusters

❑ Cluster the data into groups of different density
❑ Choose points in small cluster as candidate outliers
❑ Compute the distance between candidate points and non-candidate clusters.
❑ If candidate points are far from all other non-candidate points, they are outliers

Clustering-based Methods:

1- CLARANS (Ng and Han, 1994)

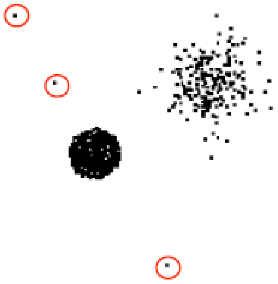
2- BIRCH (Zhang et al., 1996)

3- DBSCAN (Ester et al., 1996)

4- CURE (Guha et al., 1998)

:

:



□Advantages:

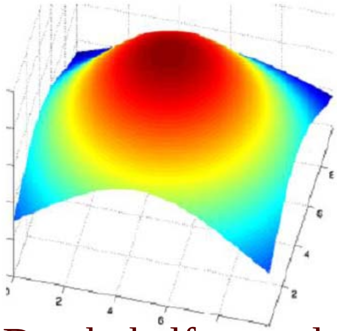
- No need to be supervised
- Easily adaptable to on-line / incremental mode suitable for anomaly detection from temporal data

□ Drawbacks

- Computationally expensive
- If normal points do not create any clusters the techniques may fail
- In high dimensional spaces, data is sparse and distances between any two data records may become quite similar.
- They are not optimized for outlier detection. The outlier detection criteria are implicit and cannot easily be inferred from the clustering procedures (Papadimitriu et al., 2002)

Depth-based Methods:

Depth is is a quantitative measurement of how central a point is with respect to a data set.

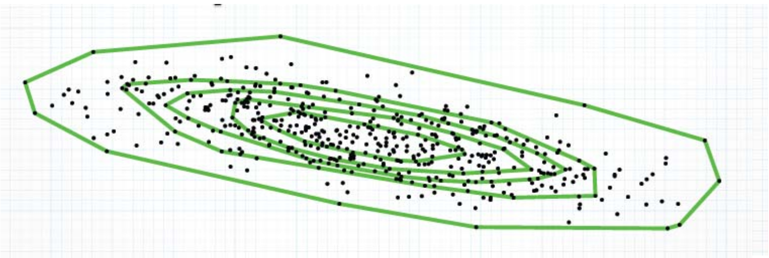


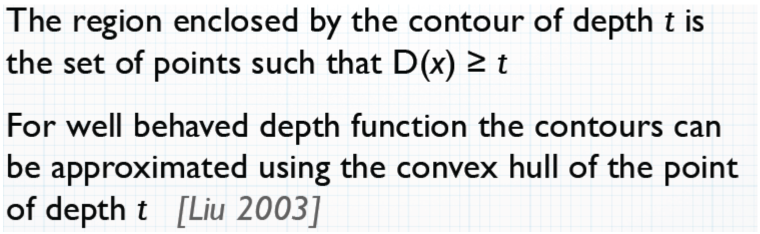
Mahalanobis depth; spatial Depth; halfspace depth; projection depth, zonoid depth,…
(Zuo and Serfling, 2000)

web source: http://www.cs.tufts.edu/research/geometry/research.php

Depth-based Methods:

Preparata and Shamos (1988); Serfling and Wang (2005)

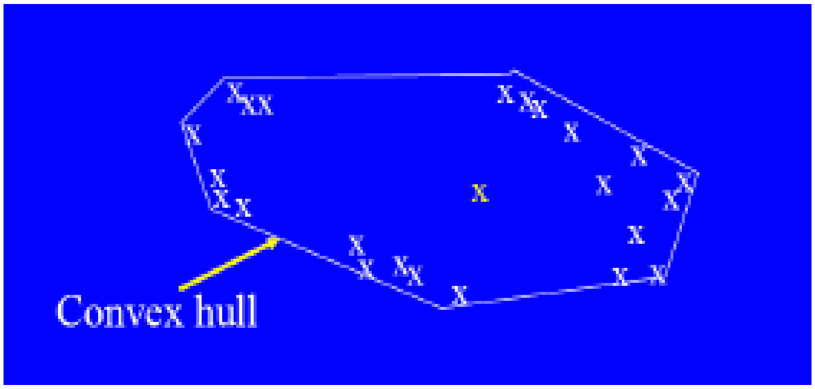




The region enclosed by the contour of depth *t* is the set of points such that D(*x*) ≥ *t*

For well behaved depth function the contours can be approximated using the convex hull of the point of depth *t* *[Liu 2003]*

What if outlier is in the middle of the data?

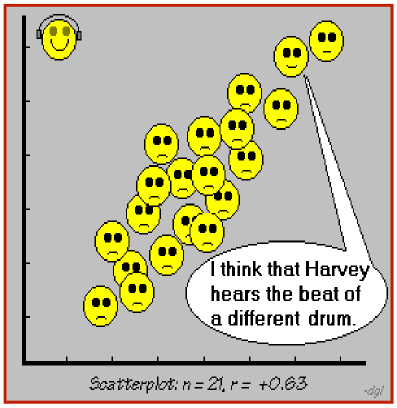


Conclusions

- ❑ Outlier detection can detect critical information in data
- ❑ Highly applicable in various application areas
- ❑ There is no single universally applicable or generic outlier detection approach.
- ❑ Researcher should select an algorithm that is suitable for their data set in terms of the correct distribution model, the correct attribute types, the scalability, the speed, any incremental capabilities.

THANKS!!!

e-mail: turkmen@stat.osu.edu


I think that Harvey hears the beat of a different drum.
Scatterplot: n = 21, r = +0.63
-dgl