CPSC 340: Machine Learning and Data Mining

Outlier Detection

Fall 2015

Admin

- Midterm on Friday.
  - Assignment 3 solutions posted.
  - Practice midterm posted (fixed typos in Q1 and Q2 solutions).
  - List of topics posted.
  - In class, 55 minutes, closed-book, cheat sheet: 2-pages each double-sided. (you will get 4 pages for the final, so you can keep your midterm pages)
- Assignment 4 out on Monday.
  - Will be due November 13.

Last time: Principal Component Analysis

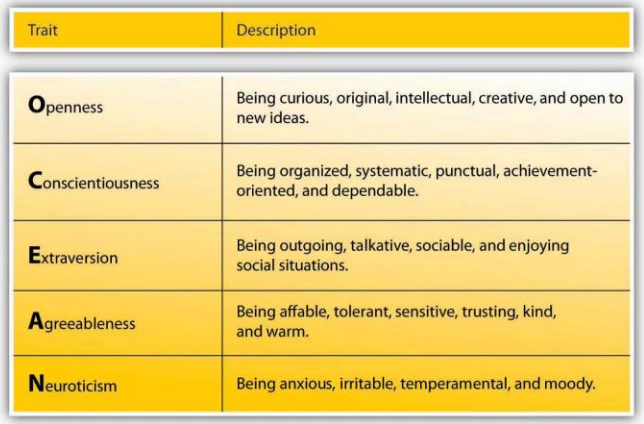
- PCA represents x_i as linear combination of factors:

$$f(W,Z)=\sum_{i=1}^{n}\sum_{j=1}^{d}(x_{ij}-w_j^T z_i)^2$$

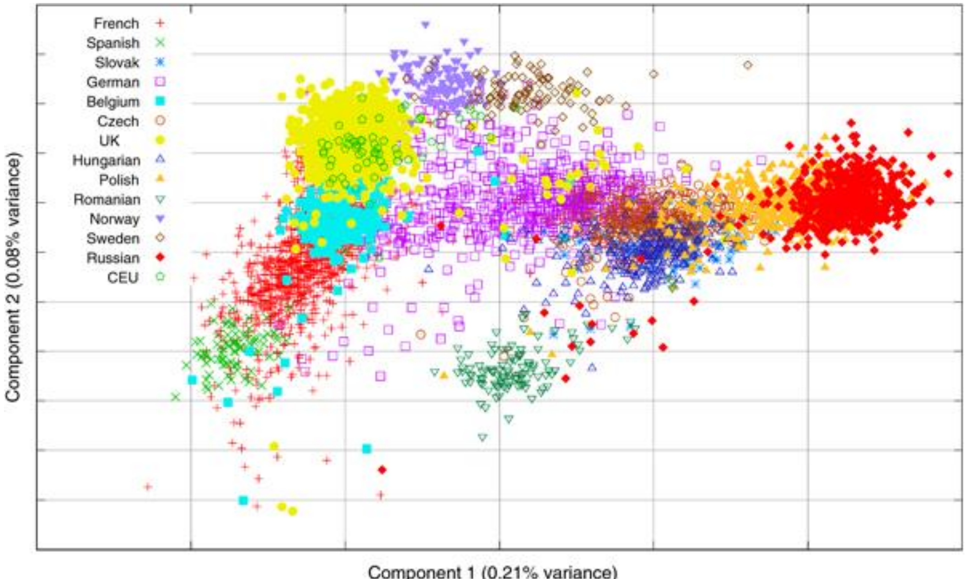
w_c: "principal component"

z_i: low-dimensional representation of x_i

- The w_c have a norm of 1, are orthogonal, and are fit consecutively.
- Gives a low-dimensional approximation of high-dimensional data.



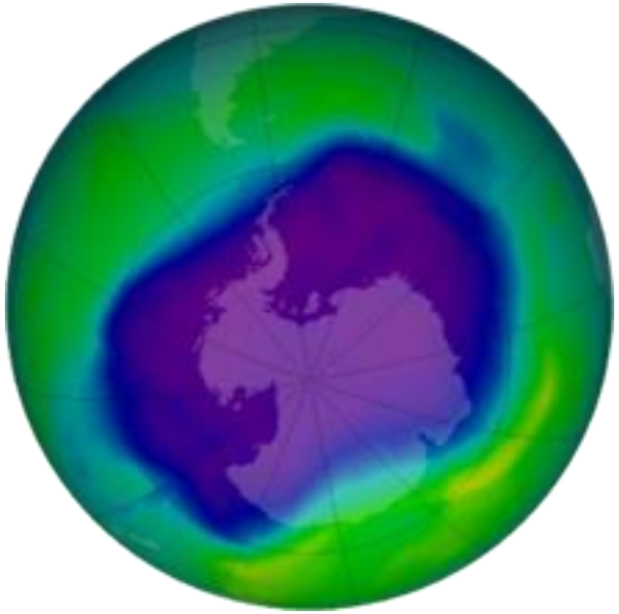
Trait	Description
Openness	Being curious, original, intellectual, creative, and open to new ideas.
Conscientiousness	Being organized, systematic, punctual, achievement-oriented, and dependable.
Extraversion	Being outgoing, talkative, sociable, and enjoying social situations.
Agreeableness	Being affable, tolerant, sensitive, trusting, kind, and warm.
Neuroticism	Being anxious, irritable, temperamental, and moody.



http://infoproc.blogspot.ca/2008/11/european-genetic-substructure.html
https://new.edu/resources/big-5-personality-traits
http://mikedusenberry.com/on-eigenfaces/

Motivating Example: Finding Holes in Ozone Layer

- The huge Antarctic ozone hole was “discovered” in 1985.

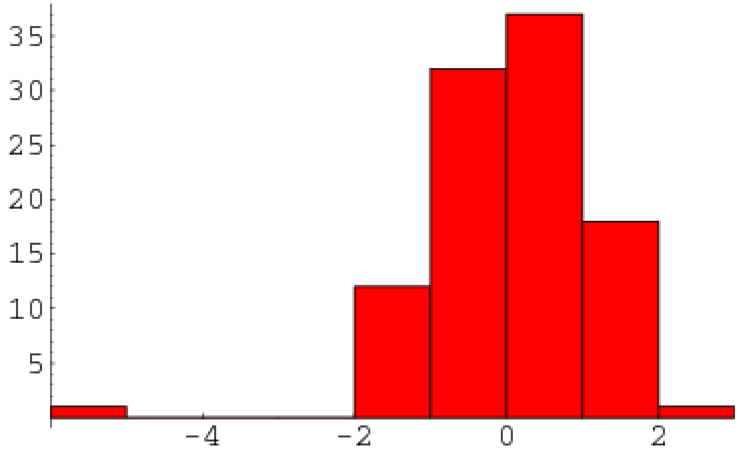


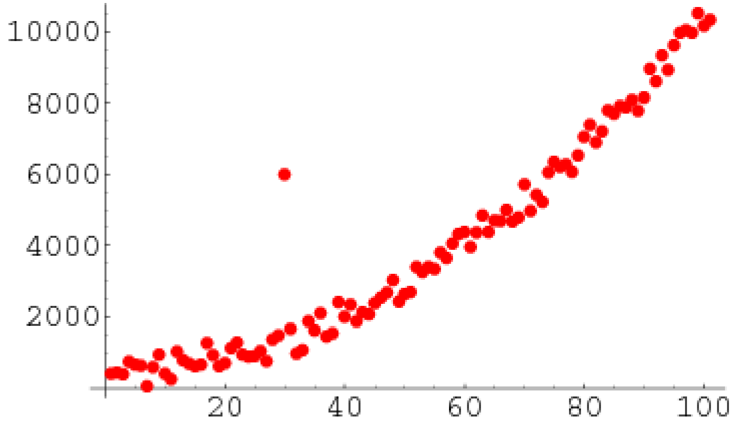
- It had been in satellite data since 1976:
  - But it was flagged and filtered out by quality-control algorithm.

https://en.wikipedia.org/wiki/Ozone_depletion

Outlier Detection

- Outlier detection:
  - find observations that are unusually different from the others.



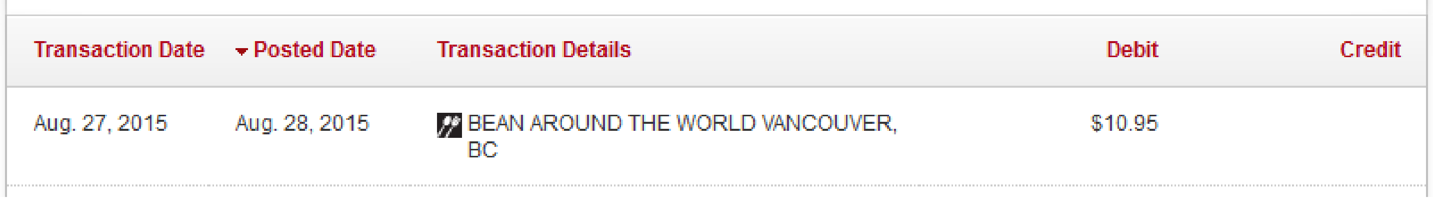


- Some sources of outliers:
  - Errors, contamination of data from different distribution, rare events.
- May want to remove outliers, or interested in the outliers themselves.

http://mathworld.wolfram.com/Outlier.html

Applications of Outlier Detection

- Data cleaning.
- Security and fault detection (network intrusion, DOS attacks).
- Fraud detection (credit cards, stocks, voting irregularities).



Transaction Date	▾ Posted Date	Transaction Details	Debit	Credit
Aug. 27, 2015	Aug. 28, 2015	BEAN AROUND THE WORLD VANCOUVER, BC	$10.95	

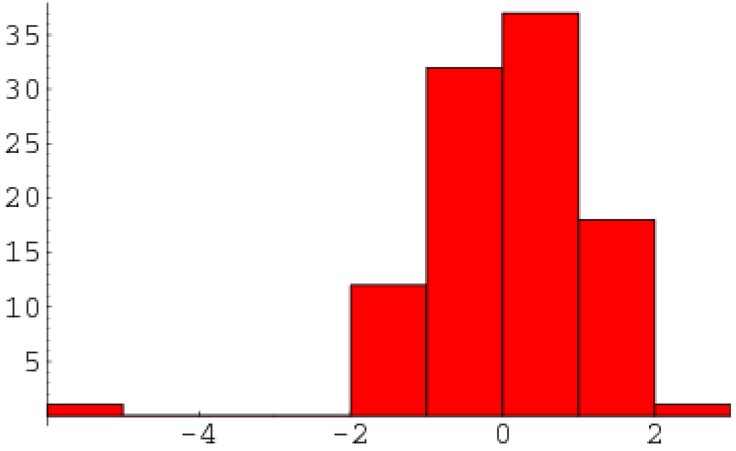
- Detecting natural disasters (earthquakes, particularly underwater).
- Astronomy (find new classes of stars/planets).
- Genetics (identifying individuals with new/ancient genes).

Classes of Methods for Outlier Detection

1. Model-based methods.
2. Graphical approaches.
3. Cluster-based methods.
4. Distance-based methods.

Model-Based Outlier Detection

- Model-based outlier detection:
  1. Fit a probabilistic model.
  2. Outliers are examples with low probability.



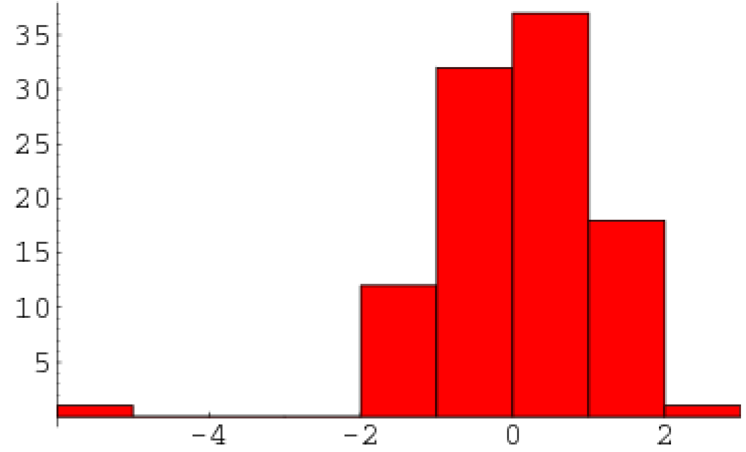
$$z_i = \frac{x_i - \mu}{\sigma}$$

- Simplest approach is z-score:
  - If z > 3, 97% of data is closer to mean?
- Another variation: return big z_i after running PCA.

http://mathworld.wolfram.com/Outlier.html

Problems with Z-Score

- The z-score relies on mean and standard deviation:
  - These measure are sensitive to outliers.

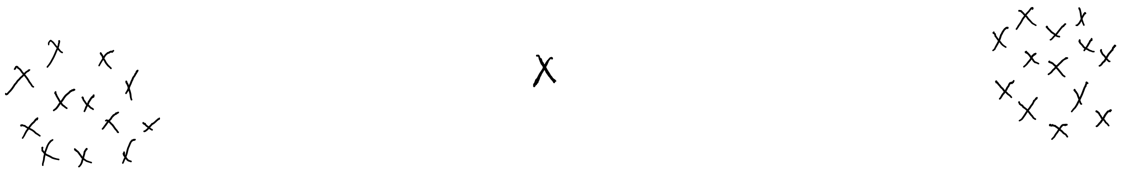


  - Possible fixes: use quantiles, or sequentially remove worse outlier.
- The z-score assumes that data is uni-modal…

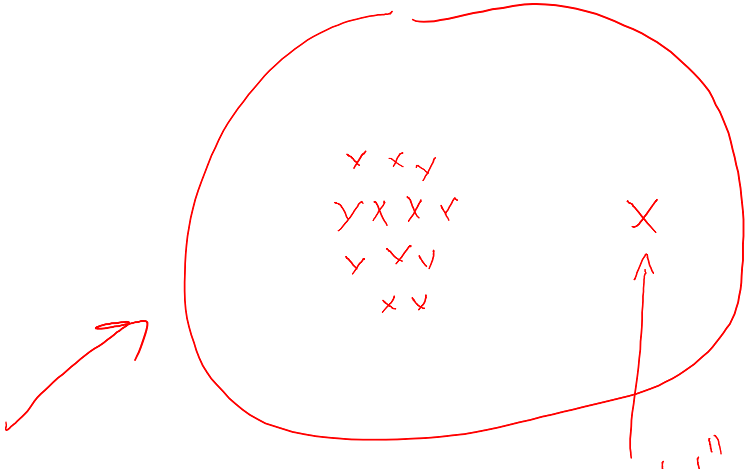
http://mathworld.wolfram.com/Outlier.html

Global vs. Local Outliers

- Is the middle point an outlier?



- Middle point has the lowest z-score.
  - It’s not a ‘global’ outlier, but is a clear ‘local’ outlier.
- In general, hard to give precise definition of ‘outliers’
  - What about outlier groups?

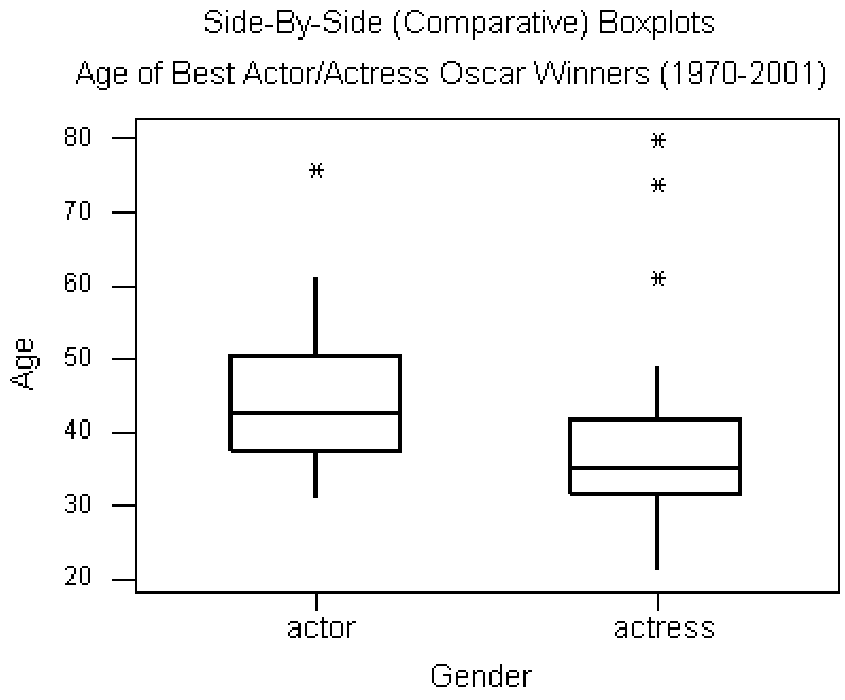


"local" outlier: normal range but far from data.

"global" outlier: outside range of data.

Graphical Outlier Detection

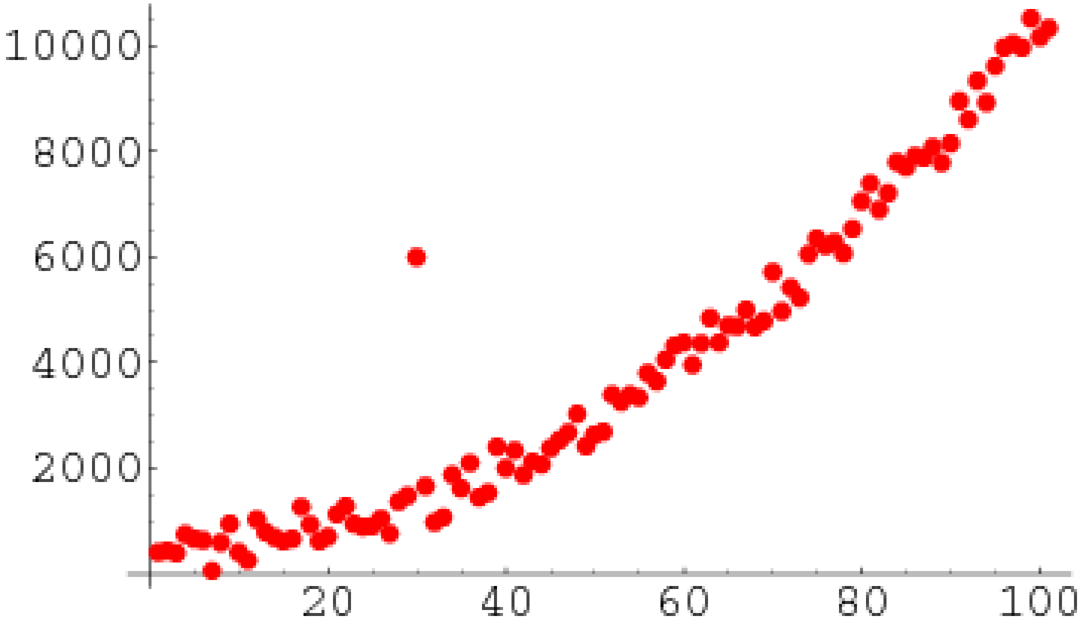
- Graphical approach to outlier detection:
  1. Look at a plot of the data.
  2. Human decides if data is an outlier.
- Examples:
  1. Box plot:
     - Visualization of quantiles/outliers.
     - Only 1 variable at a time.



http://bolt.mph.ufl.edu/6050-6052/unit-1/one-quantitative-variable-introduction/boxplot/

Graphical Outlier Detection

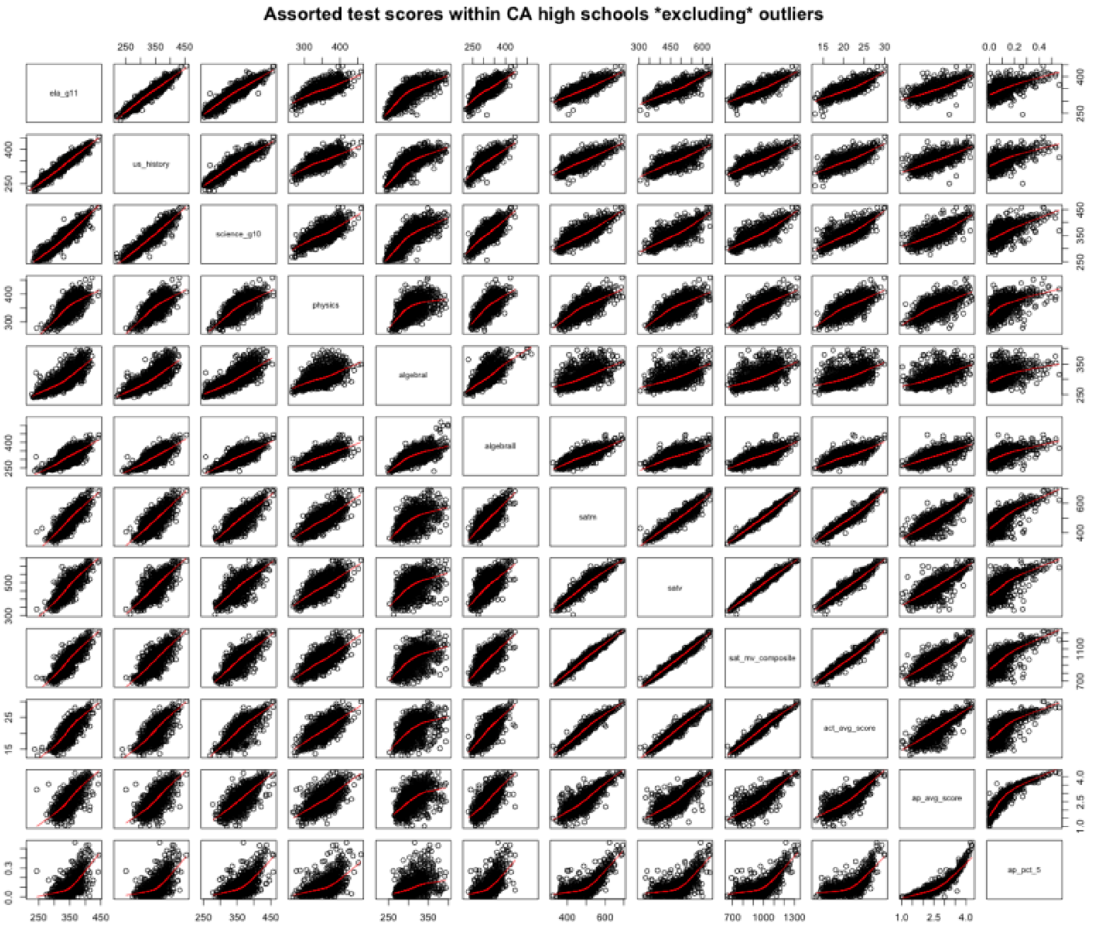
- Graphical approach to outlier detection:
  1. Look at a plot of the data.
  2. Human decides if data is an outlier.
- Examples:
  1. Box plot.
  2. Scatterplot:
     - Can detect complex patterns.
     - Only 2 variables at a time.



http://mathworld.wolfram.com/Outlier.html

Graphical Outlier Detection

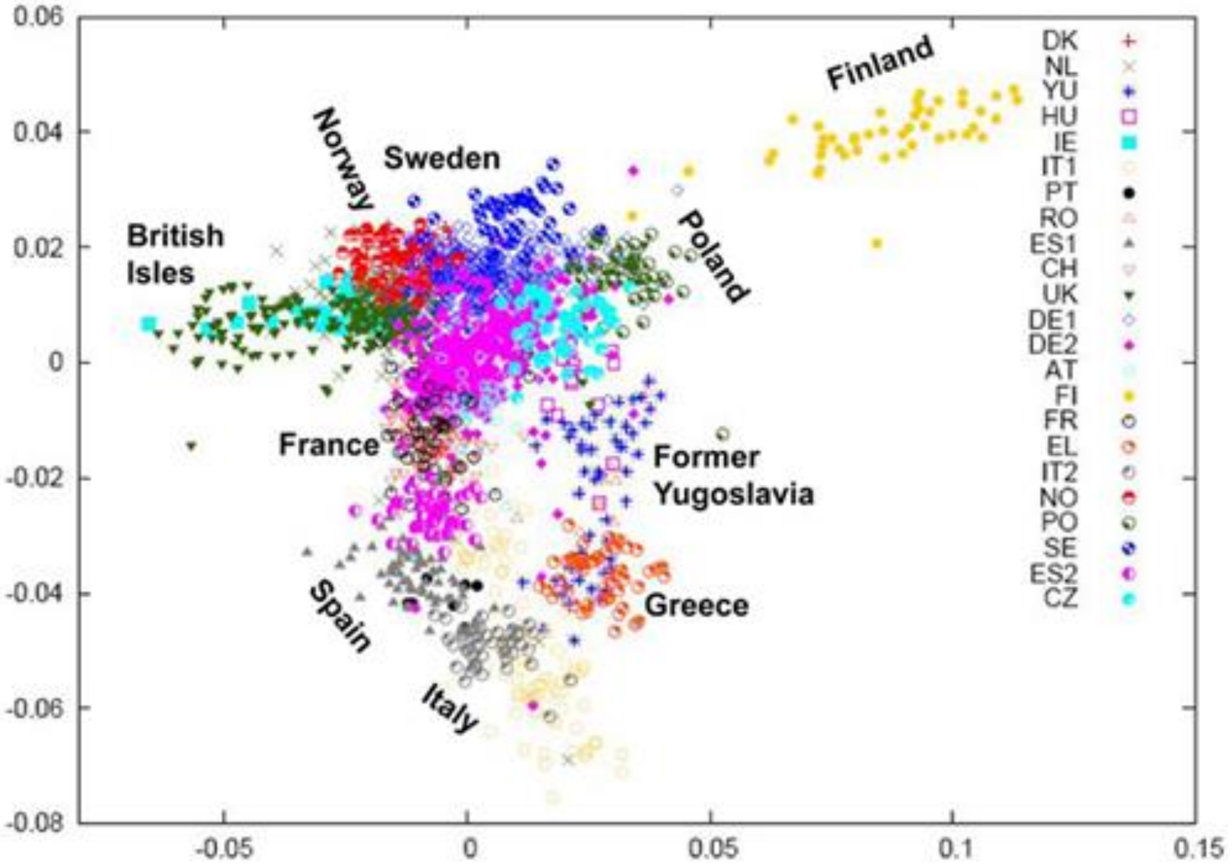
- Graphical approach to outlier detection:
  1. Look at a plot of the data.
  2. Human decides if data is an outlier.
- Examples:
  1. Box plot.
  2. Scatterplot.
  3. Scatterplot array:
     - Look at all combinations of variables.
     - But laborious in high-dimensions.
     - Still only 2 variables at a time.



https://randomcriticalanalysis.wordpress.com/2015/05/25/standardized-tests-correlations-within-and-between-california-public-schools/

Graphical Outlier Detection

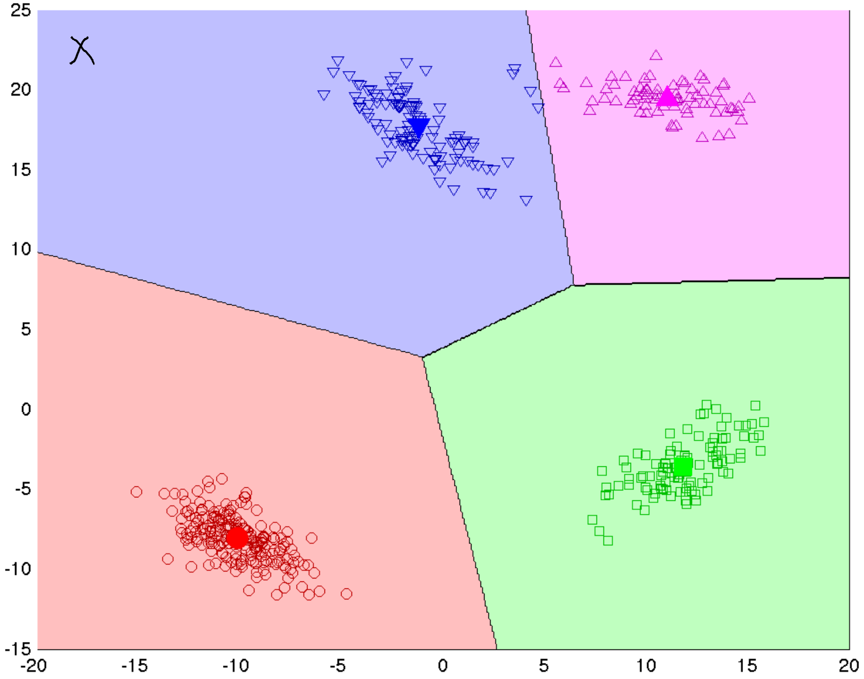
- Graphical approach to outlier detection:
  1. Look at a plot of the data.
  2. Human decides if data is an outlier
- Examples:
  1. Box plot.
  2. Scatterplot.
  3. Scatterplot array.
  4. Scatterplot of 2-dimensional PCA:
     - ‘See’ high-dimensional structure.
     - But PCA is sensitive to outliers.
     - There might be info in higher PCs.



http://scienceblogs.com/gnxp/2008/08/14/the-genetic-map-of-europe/

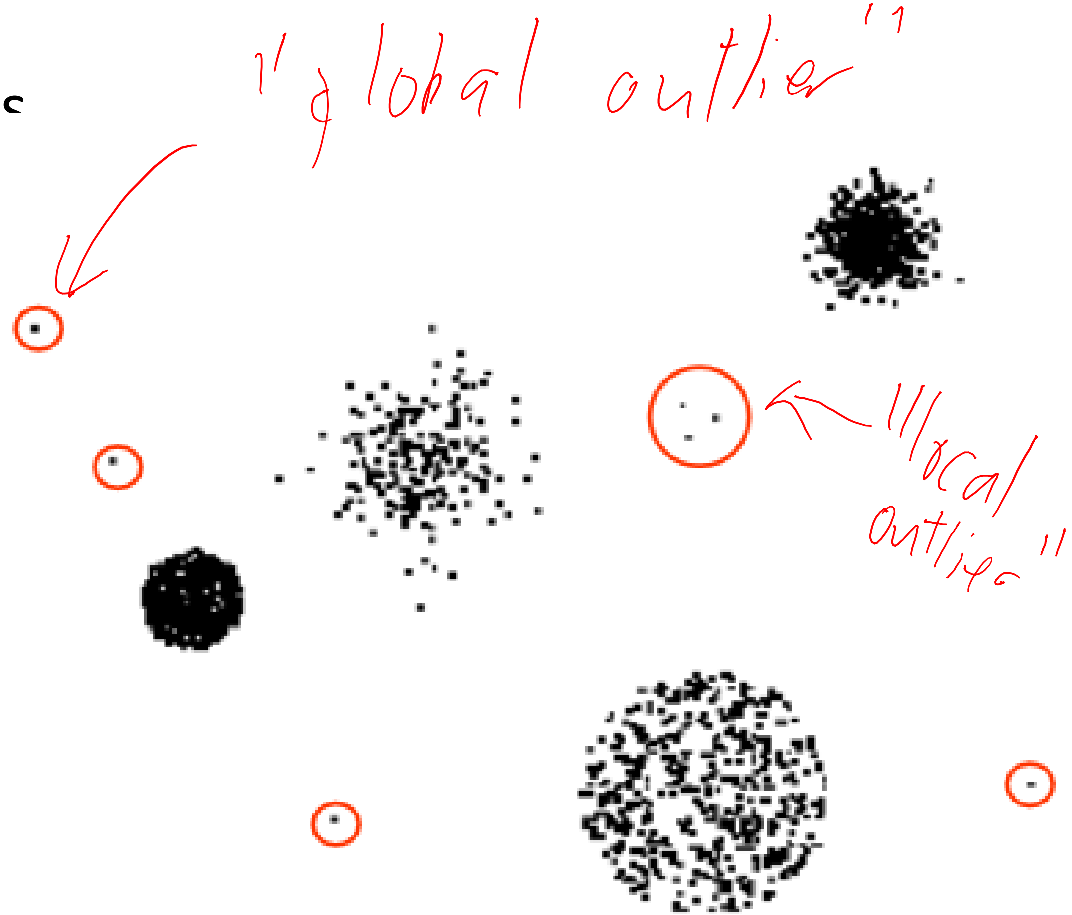
Cluster-Based Outlier Detection

- Detect outliers based on clustering:
  1. Cluster the data.
  2. Find points that don't belong to clusters.
- Examples:
  1. K-means:
     - Find points that are far away from any mean.
     - Find clusters with a small number of points.



Cluster-Based Outlier Detection

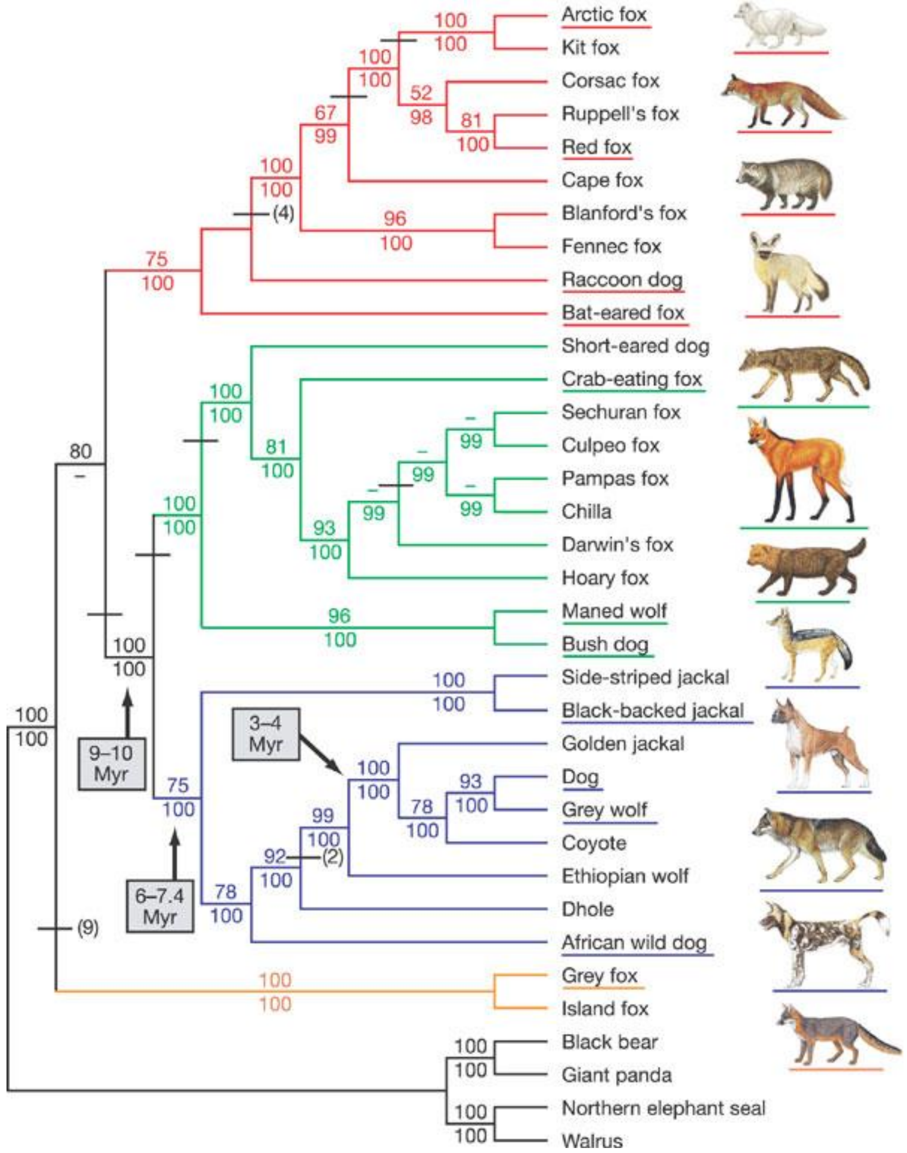
- Detect outliers based on clustering:
  1. Cluster the data.
  2. Find points that don’t belong to clusters
- Examples:
  1. K-means.
  2. Density-based clustering:
     - Outliers are points not assigned to cluster.



http://www-users.cs.umn.edu/~kumar/dmbook/dmslides/chap10_anomaly_detection.pdf

Cluster-Based Outlier Detection

- Detect outliers based on clustering:
  1. Cluster the data.
  2. Find points that don’t belong to clusters.
- Examples:
  1. K-means.
  2. Density-based clustering.
  3. Hierarchical clustering:
     - Outliers take longer to join other groups.
     - Also good for outlier groups.



http://www.nature.com/nature/journal/v438/n7069/fig_tab/nature04338_F10.html

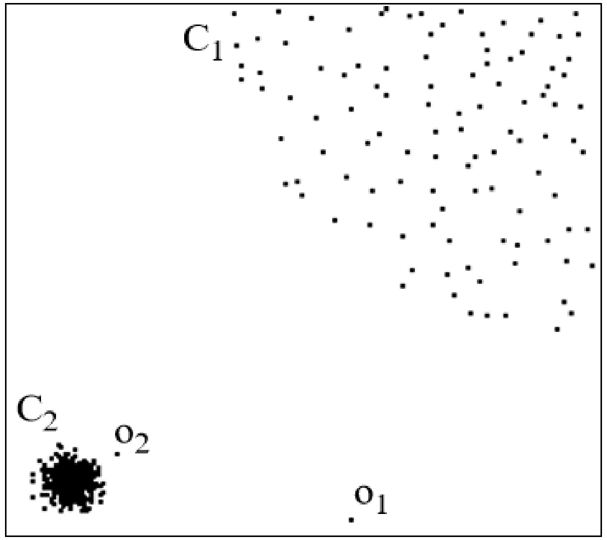
Distance-Based Outlier Detection

- Most of these approaches are based on distances.
- Can we skip the models/plot/clusters and directly use distances?
- Distance-based outlier detection:
  - Use some measure of how close objects are to their neighbours.
- Examples:
  - How many points lie in a radius ‘r’?
  - What is distance to kth nearest neighbour?

https://en.wikipedia.org/wiki/Local_outlier_factor

Distance-Based Outlier Detection

- As with density-based clustering, problem with differing densities:



- Outlier o_2 has similar density as elements of cluster C_1.
- ‘Local outlier factor’ and variations:
  - Is point further away from its neighbours, then they are from each other?

http://www.dbs.ifi.lmu.de/Publikationen/Papers/LOF.pdf

Outlierness Ratio

- Let $N_k(x_i)$ be the k-nearest neighbours of x_i.
- Let $D_k(x_i)$ be the average distance of xi to its k-nearest neighbours:

$$D_k(x_i) = \frac{1}{k} \sum_{j \in N_k(x_i)} \| x_i - x_j \|$$

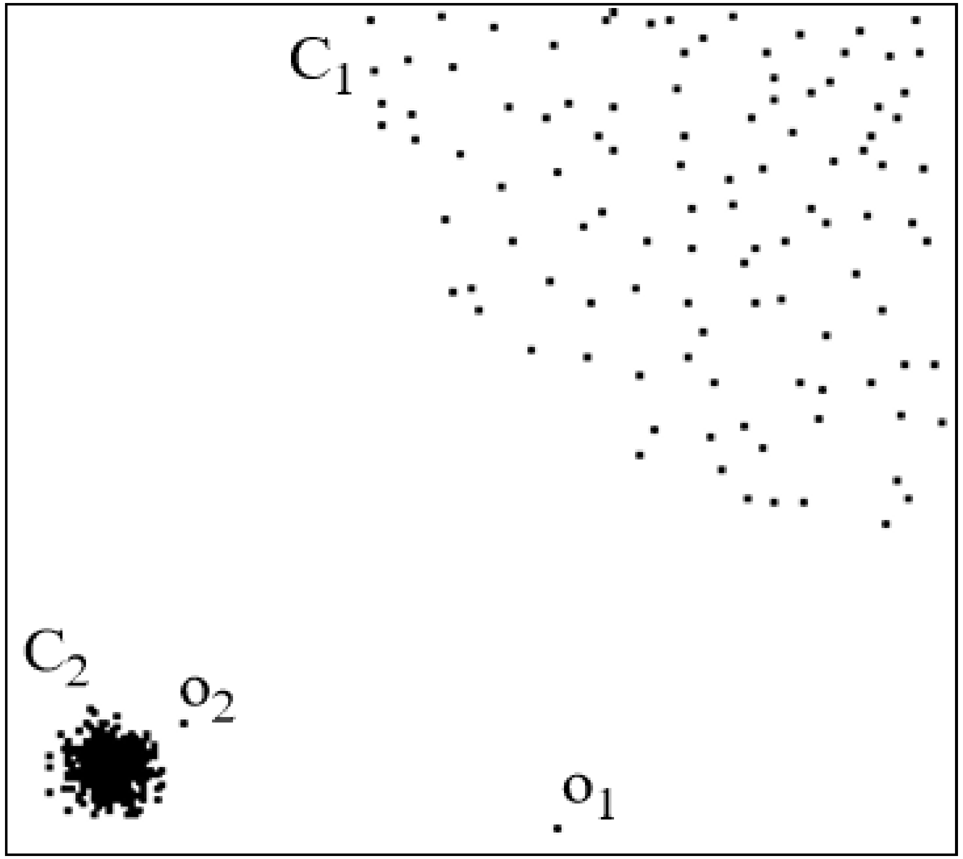
- '**Outlierness**' is ratio of $D_k(x_i)$ to average $D_k(x_j)$ for its neighbours 'j':

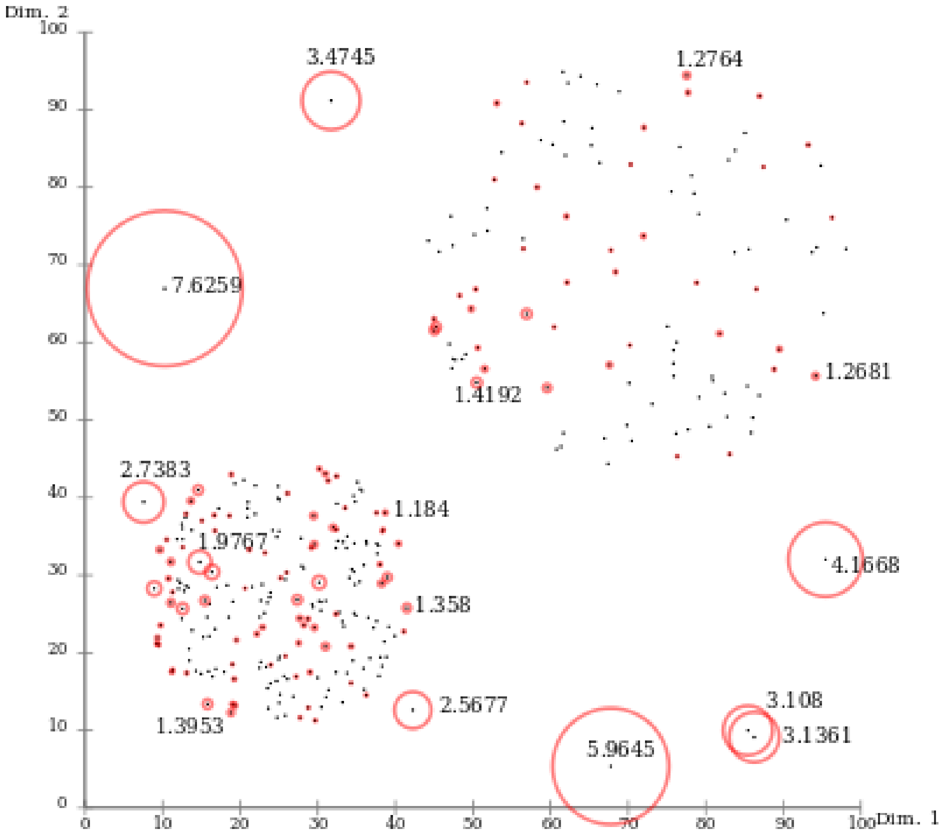
$$\frac{D_k(x_i)}{\frac{1}{k} \sum_{j \in N_k(x_i)} D_k(x_j)}$$

- If outlierness > 1, x_i is further away from neighbours than expected.

Outlierness Ratio

- Outlierness and LOF will find o_1 and o_2.

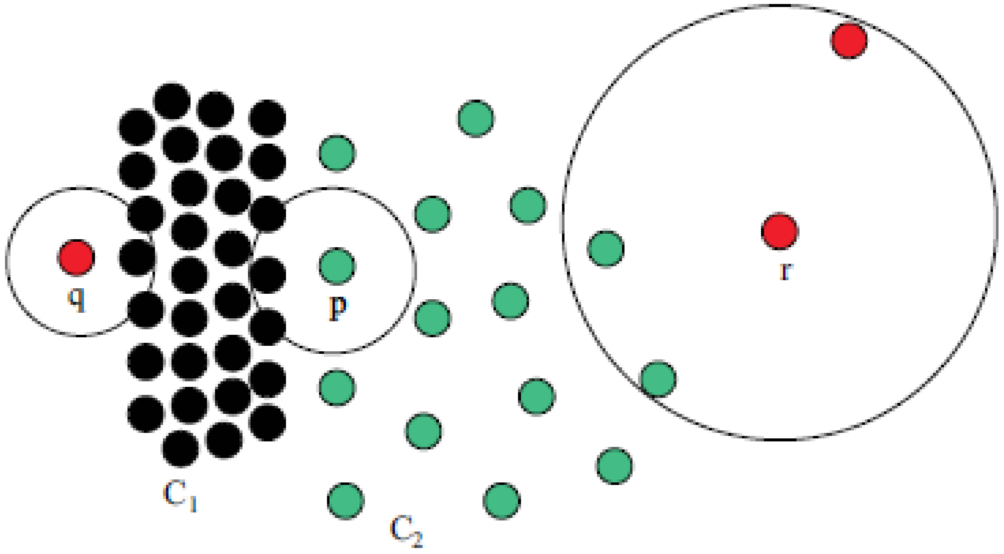




http://www.dbs.ifi.lmu.de/Publikationen/Papers/LOF.pdf
https://en.wikipedia.org/wiki/Local_outlier_factor

Outlierness with Close Clusters

- If clusters are close, outlierness gives unintuitive results:

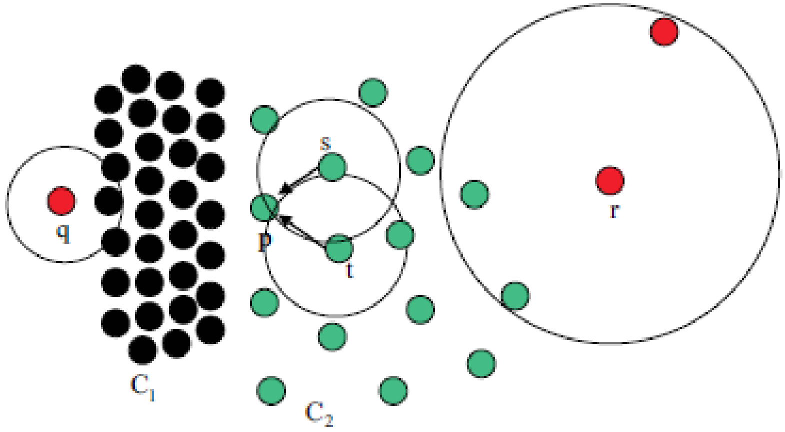


- In this example, ‘p’ has higher outlierness than ‘q’ and ‘r’:
  – The green points are not part of the KNN list of ‘p’ for small ‘k’.

http://www.comp.nus.edu.sg/~atung/publication/pakdd06_outlier.pdf

Outlierness with Close Clusters

- ‘Influenced outlierness’ (INFLO) ratio:
  - Include ‘reverse’ k-nearest neighbours (points that have ‘p’ in KNN list).
    - Included in the average in the denominator of outlierness ratio.
  - Adds ‘s’ and ‘t’ from bigger cluster that includes ‘p’:



- Still not perfect, particularly for hierarchical clusters.
  - You should also try multiple values of ‘k’.

http://www.comp.nus.edu.sg/~atung/publication/pakdd06_outlier.pdf

Summary

- Outlier detection is task of finding unusually different object.
- Model-based methods check if objects are unlikely in fitted model.
- Graphical methods plots data and use human to find outliers.
- Cluster-based methods check whether objects belong to clusters.
- Distance-based methods measure distance to nearby objects.

- Next time: midterm.
  - Then on Monday, changing PCA so it splits faces into ‘eyes’, ‘mouths’, etc.