

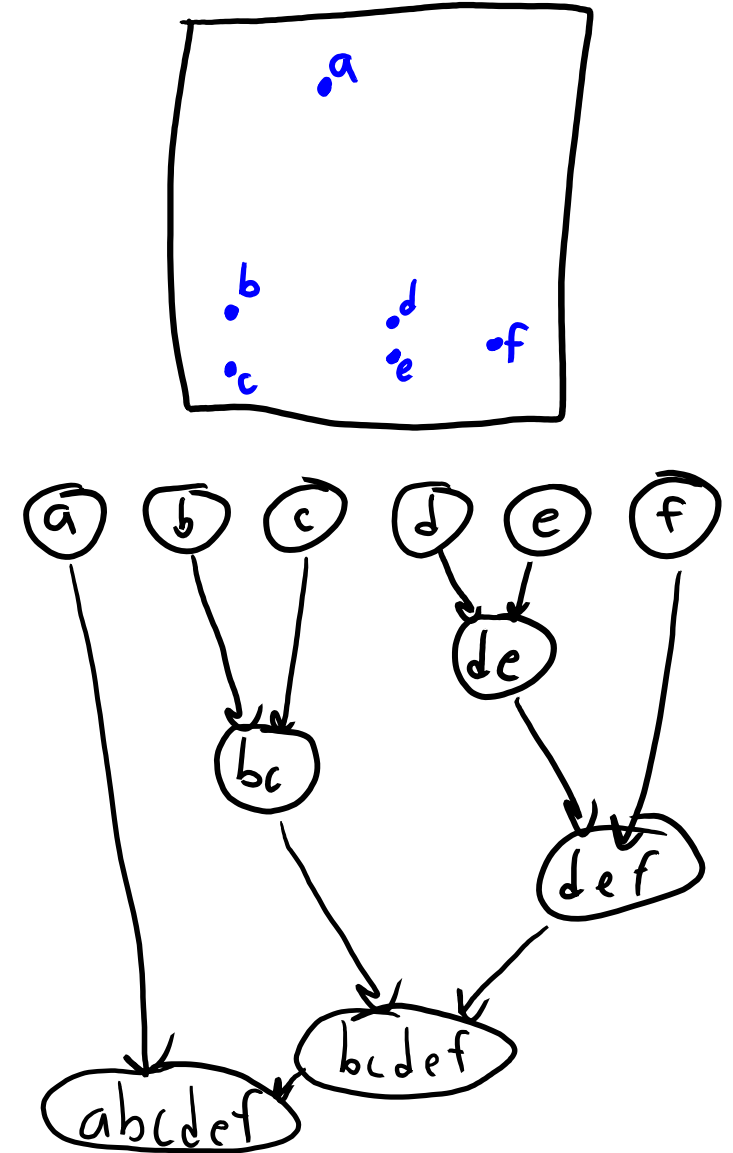
CPSC 340: Machine Learning and Data Mining

Outlier Detection

Fall 2019

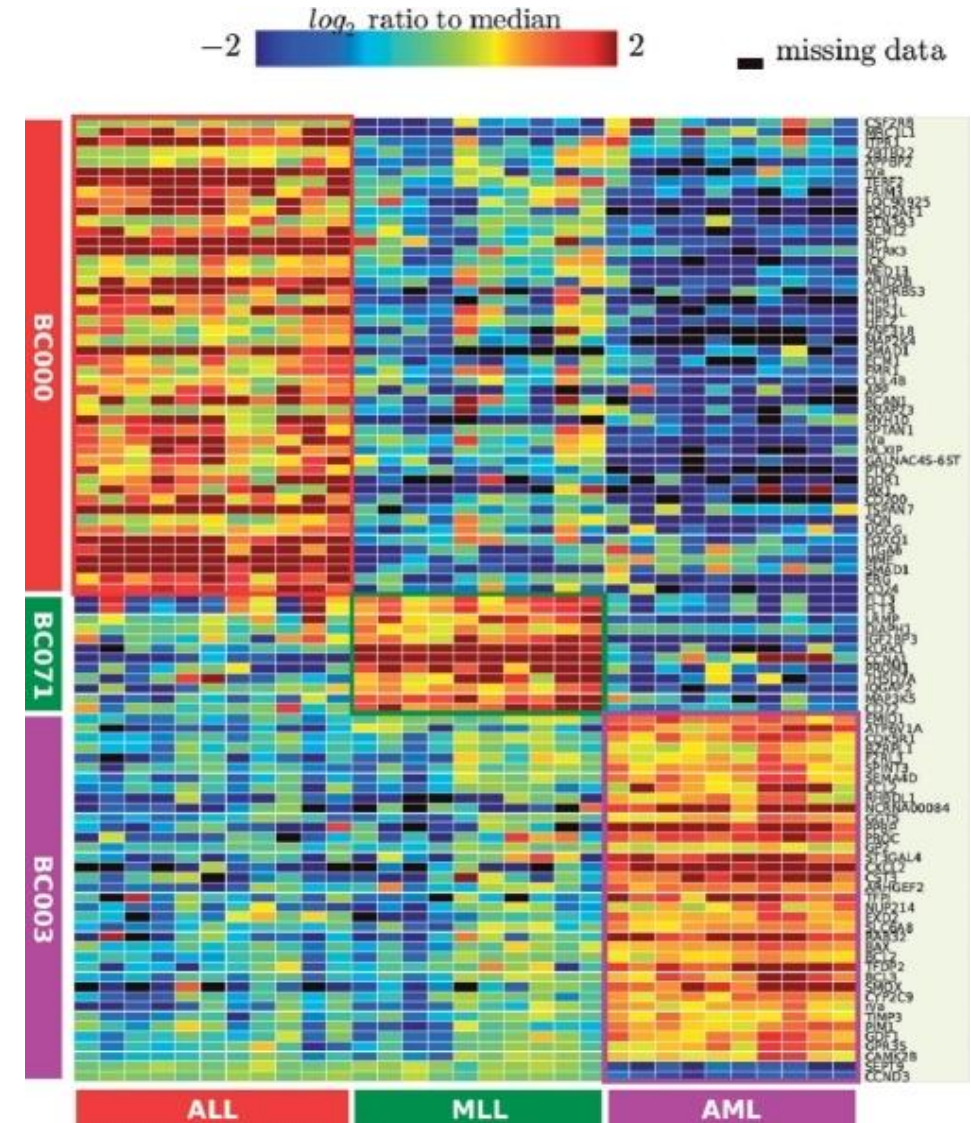
Last Time: Hierarchical Clustering

- We discussed **hierarchical clustering**:
 - Performs **clustering at multiple scales**.
 - Output is usually a **tree diagram** (“dendrogram”).
 - Reveals much more structure in data.
 - Usually non-parametric:
 - At finest scale, every point is its own clusters.
- We discussed some application areas:
 - Animals (phylogenetics).
 - Languages.
 - Stories.
 - Fashion.



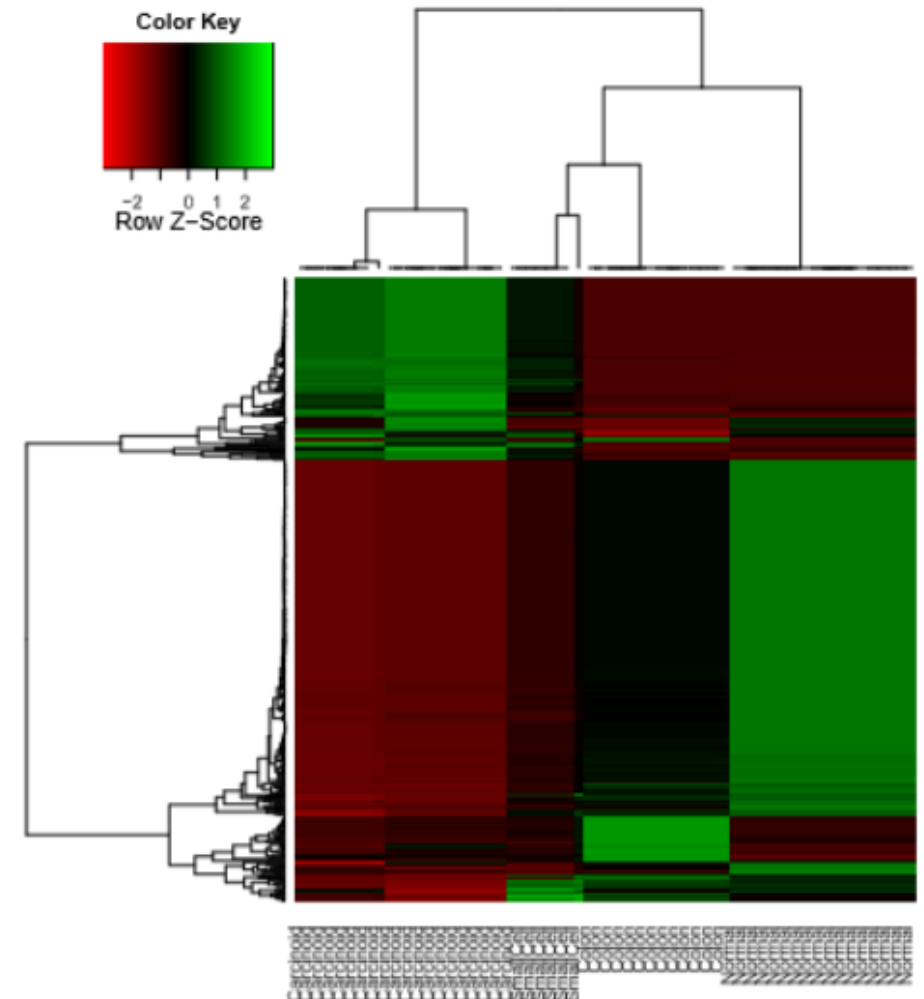
Biclustering

- Biclustering:
 - Cluster the training examples and features.
 - Also gives feature relationship information.
- Simplest and most popular method:
 - Run clustering method on 'X' (examples).
 - Run clustering method on 'X^T' (features).
- Often plotted with 'X' as a heatmap.
 - Where rows/columns arranged by clusters.
 - Helps you 'see' why things are clustered.



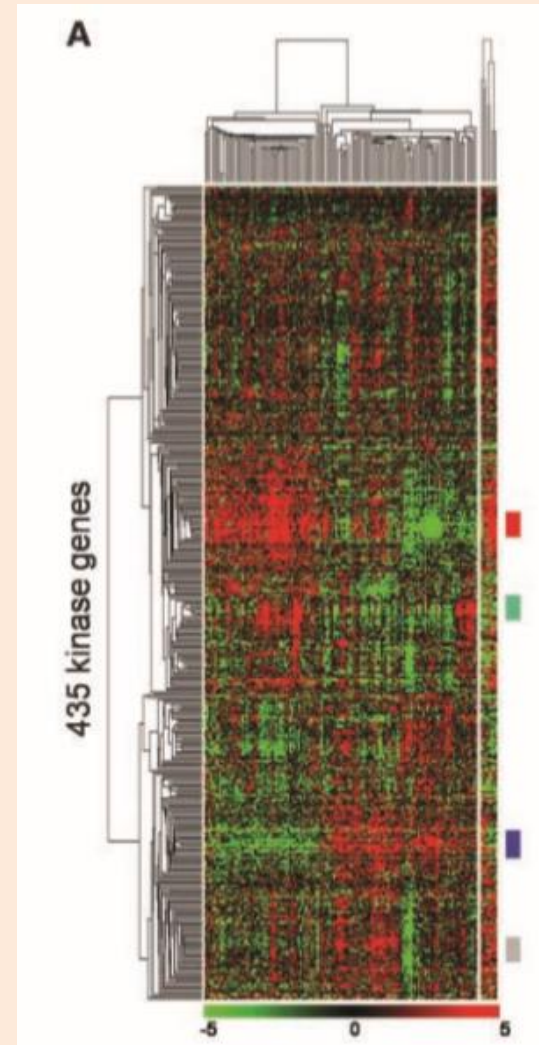
Biclustering

- Visualization: **hierarchical biclustering + heatmap + dendrograms.**
 - Popular in biology/medicine.



Application: Medical data

- Hierarchical clustering is very common in **medical data analysis**.
 - Biclustering different samples of breast cancer:

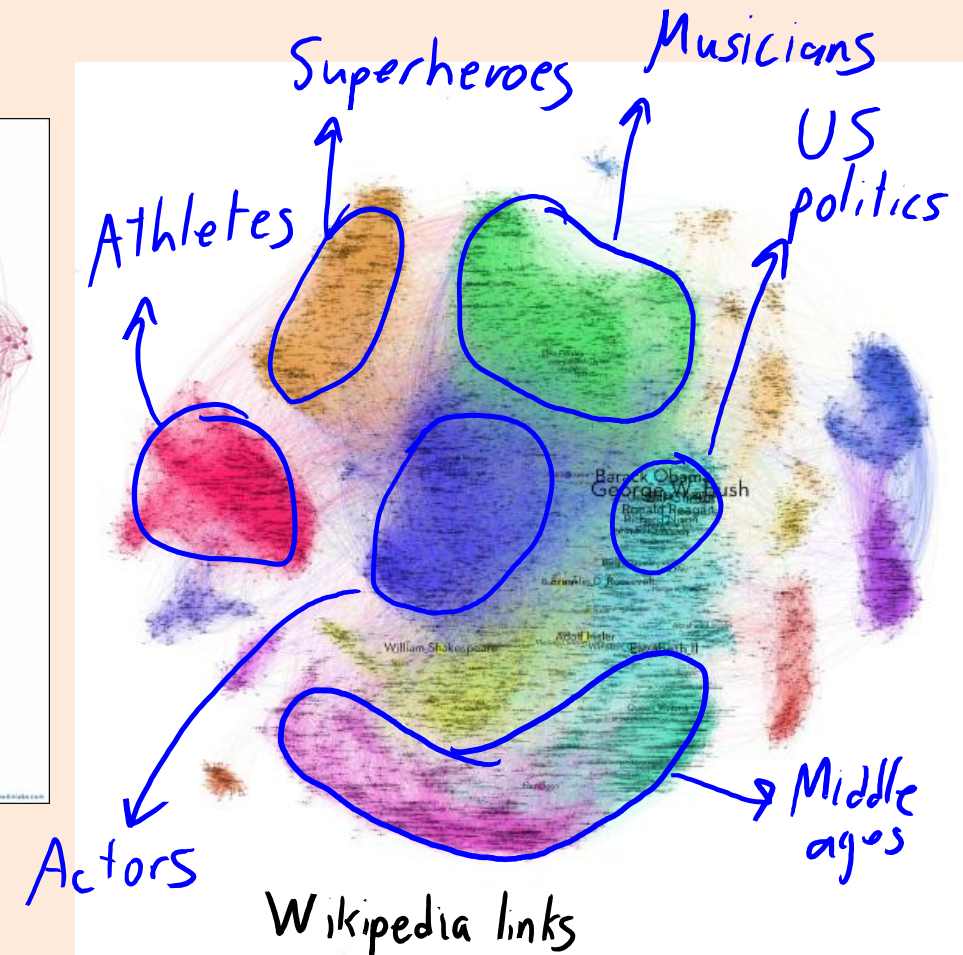
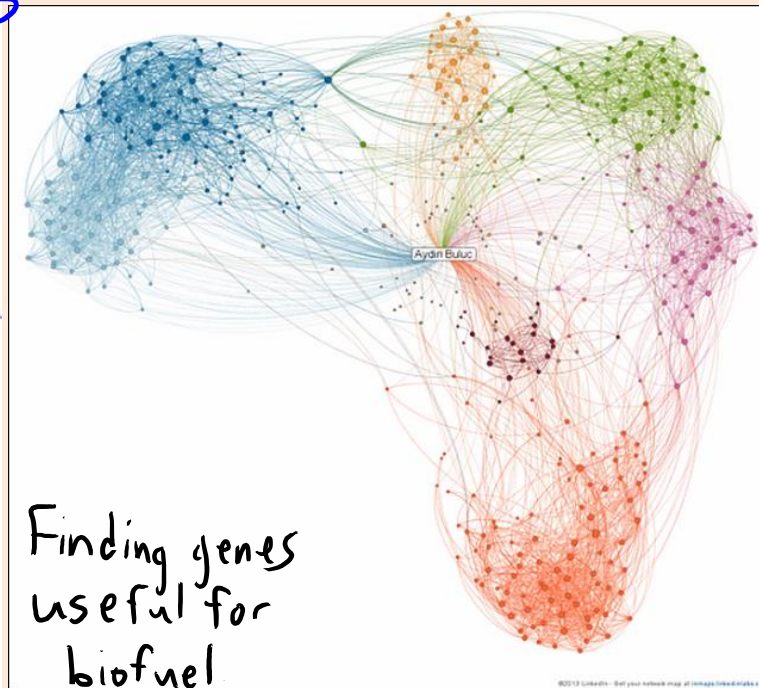
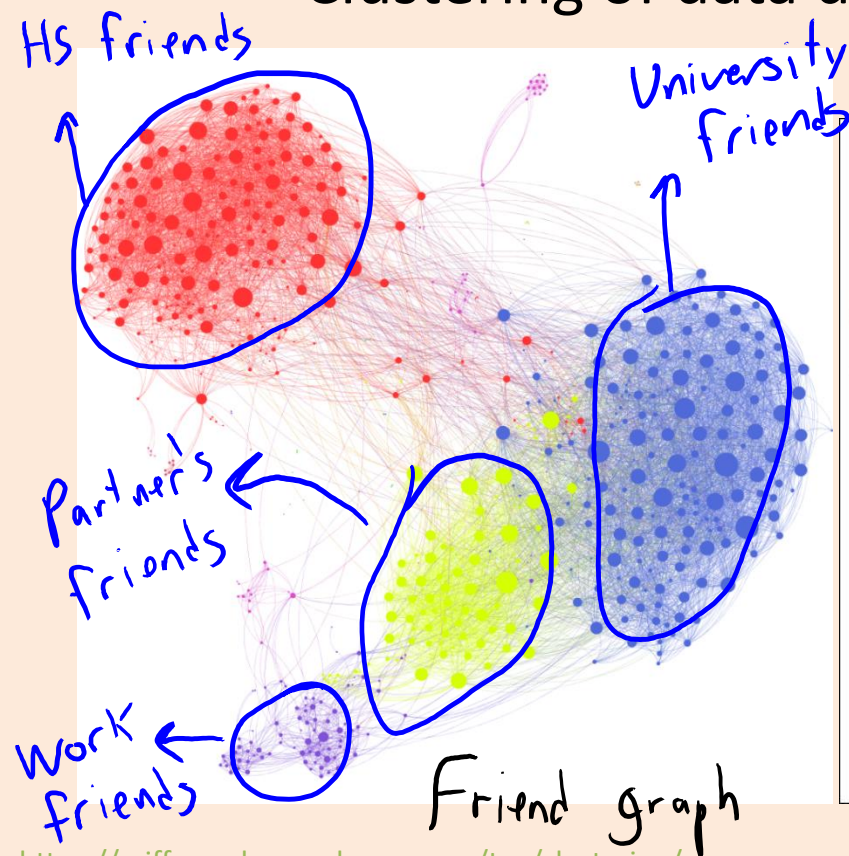


Other Clustering Methods

- **Mixture models:**
 - Probabilistic clustering.
- **Mean-shift clustering:**
 - Finds local “modes” in density of points.
 - Alternative approach to vector quantization.
- **Bayesian clustering:**
 - A variant on ensemble methods.
 - Averages over models/clusterings, weighted by “prior” belief in the model/clustering.

Graph-Based Clustering

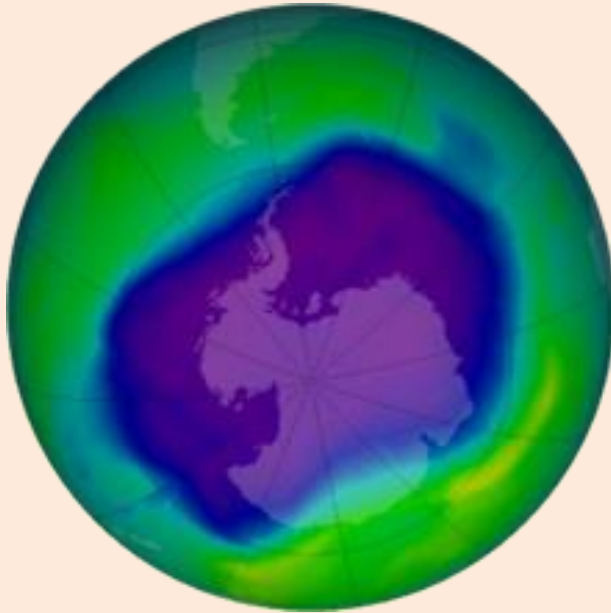
- Spectral clustering and graph-based clustering:
 - Clustering of data described by graphs.



(pause)

Motivating Example: Finding Holes in Ozone Layer

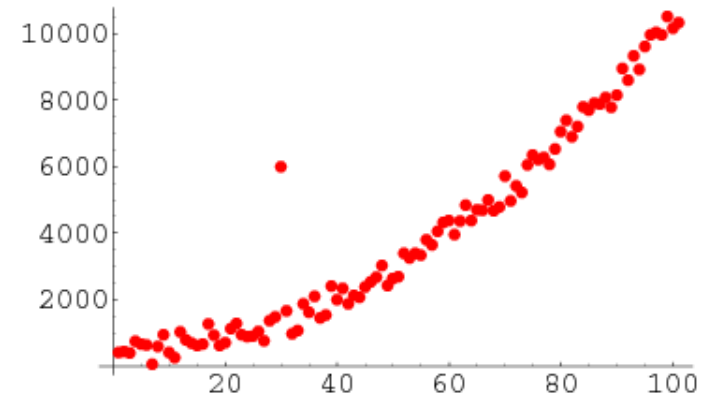
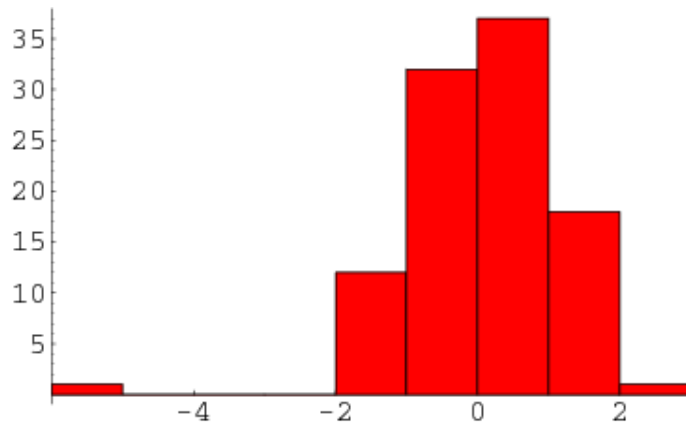
- The huge Antarctic ozone hole was “discovered” in 1985.



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

Outlier Detection


- **Outlier detection:**
 - Find observations that are “unusually different” from the others.
 - Also known as “anomaly detection”.
 - May want to remove outliers, or be interested in the outliers themselves (security).



- **Some sources of outliers:**
 - Measurement errors.
 - Data entry errors.
 - Contamination of data from different sources.
 - Rare events.

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 (underwater earthquakes).
- 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.
 5. Supervised-learning methods.
- Warning: this is the topic with the most ambiguous “solutions”.

But first...

- Usually it's good to do some **basic sanity checking**...

Egg	Milk	Fish	Wheat	Shellfish	Peanuts	Peanuts	Sick?
0	0.7	0	0.3	0	0	0	1
0.3	0.7	0	0.6	-1	3	3	1
0	0	0	"sick"	0	1	1	0
0.3	0.7	1.2	0	0.10	0	0	2
900	0	1.2	0.3	0.10	0	0	1

- Would any values in the column cause a Python/Julia **"type" error**?
- What is the **range of numerical features**?
- What are the **unique entries for a categorical feature**?
- Does it look like parts of the table are **duplicated**?
- These types of simple errors are VERY common in real data.

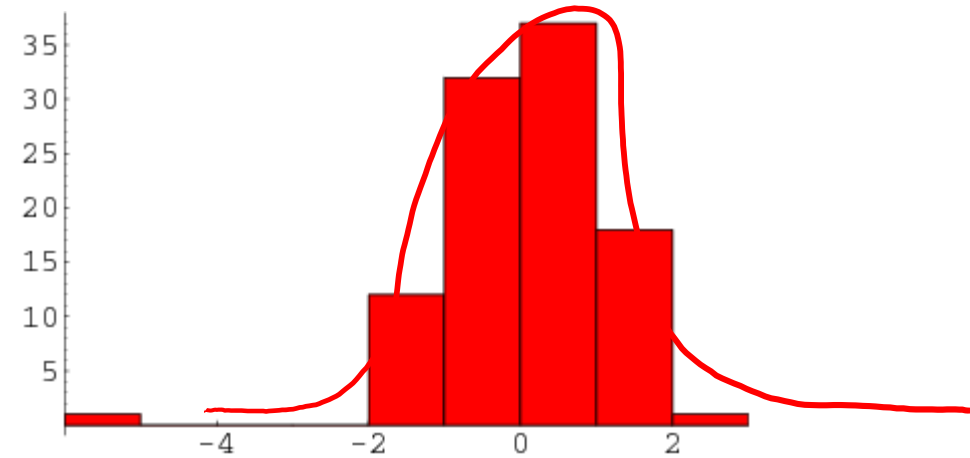
Model-Based Outlier Detection

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

- Example:
 - Assume data follows normal distribution.
 - The z-score for 1D data is given by:

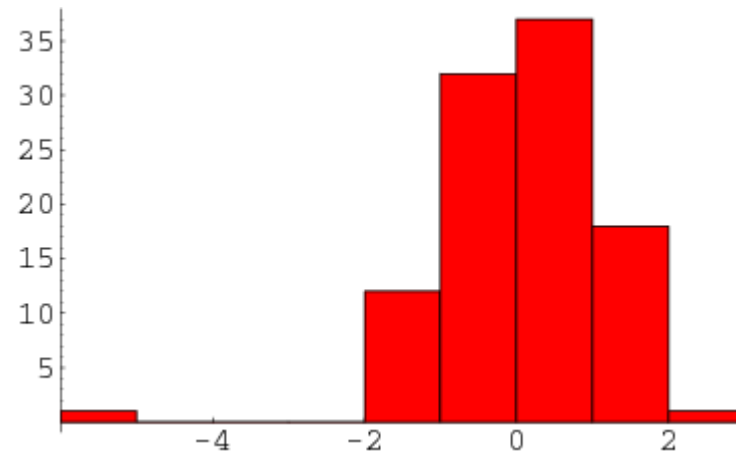
$$z_i = \frac{x_i - \mu}{\sigma} \quad \text{where } \mu = \frac{1}{n} \sum_{i=1}^n x_i \quad \text{and } \sigma = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2}$$

- “Number of standard deviations away from the mean”.
 - Say “outlier” if $|z| > 4$, or some other threshold.

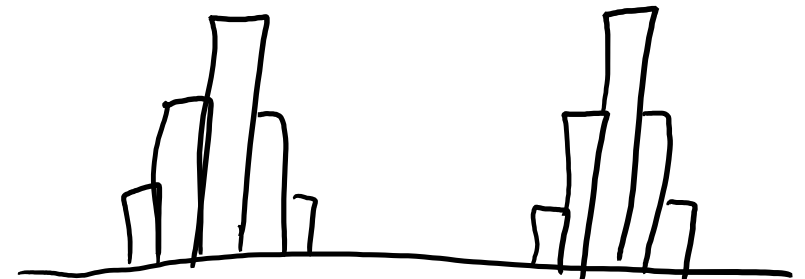


Problems with Z-Score

- Unfortunately, the **mean and variance are sensitive to outliers**.



- Possible fixes: **use quantiles, or sequentially remove worse outlier**.
- The z-score also assumes that data is “uni-modal”.
 - Data is concentrated around the mean.



Global vs. Local Outliers

- Is the **red point** an outlier?



Global vs. Local Outliers

- Is the **red point** an outlier? What if we add the **blue points**?



Global vs. Local Outliers

- Is the **red point** an outlier? What if we add the **blue points**?



- Red point has the **lowest z-score**.
 - In the first case it was a “**global**” outlier.
 - In this second case it’s a “**local**” outlier:
 - Within normal data range, but **far from other points**.
- It’s hard to precisely define “outliers”.

Global vs. Local Outliers

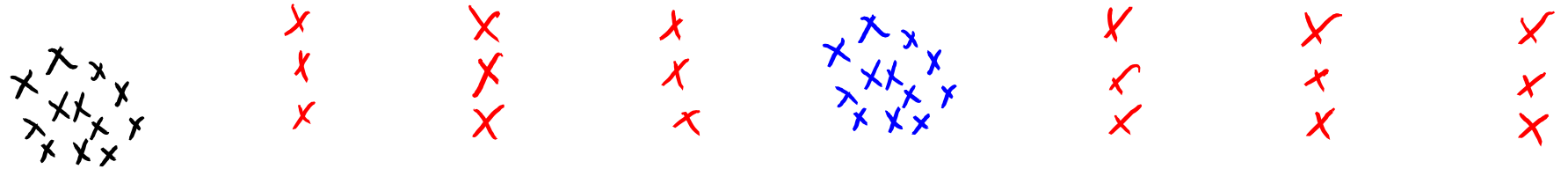
- Is the **red point** an outlier? What if we add the **blue points**?



- Red point has the **lowest z-score**.
 - In the first case it was a “**global**” outlier.
 - In this second case it’s a “**local**” outlier:
 - Within normal data range, but **far from other points**.
- It’s hard to precisely define “outliers”.
 - Can we have **outlier groups**?

Global vs. Local Outliers

- Is the **red point** an outlier? What if we add the **blue points**?



- Red point has the **lowest z-score**.
 - In the first case it was a **“global”** outlier.
 - In this second case it’s a **“local”** outlier:
 - Within normal data range, but **far from other points**.
- It’s hard to precisely define “outliers”.
 - Can we have **outlier groups**? What about repeating patterns?

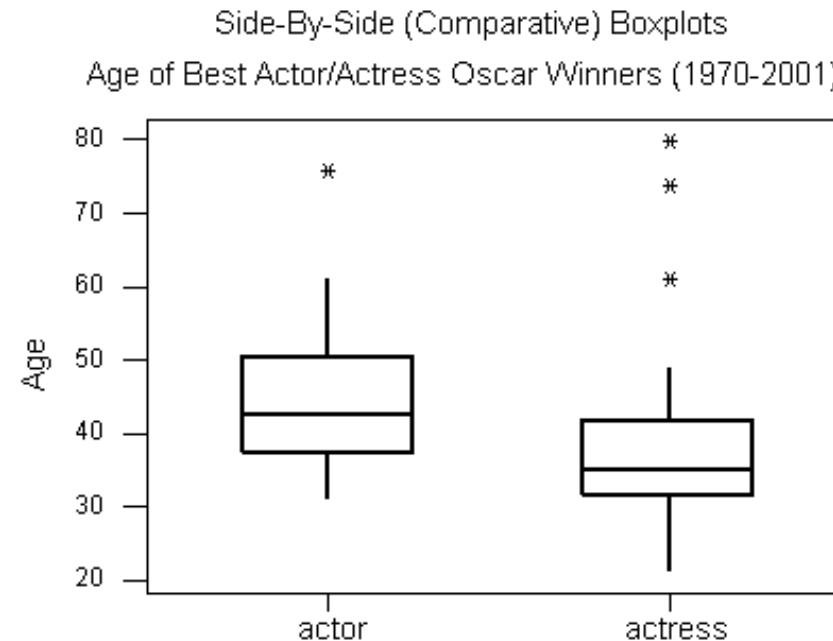
Graphical Outlier Detection

- 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.



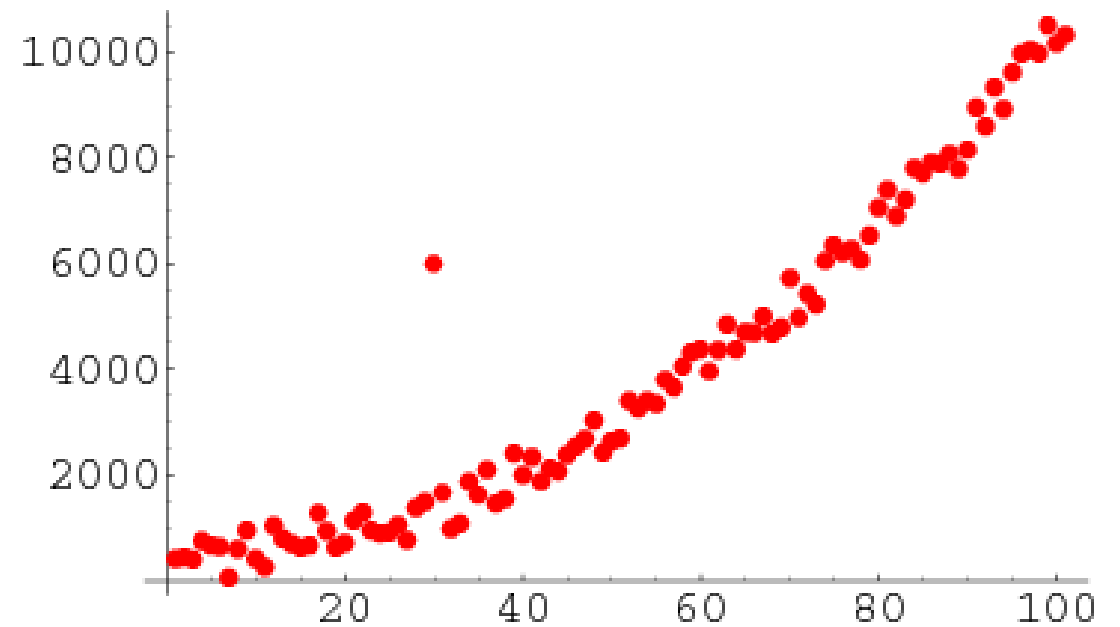
Graphical Outlier Detection

- 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.



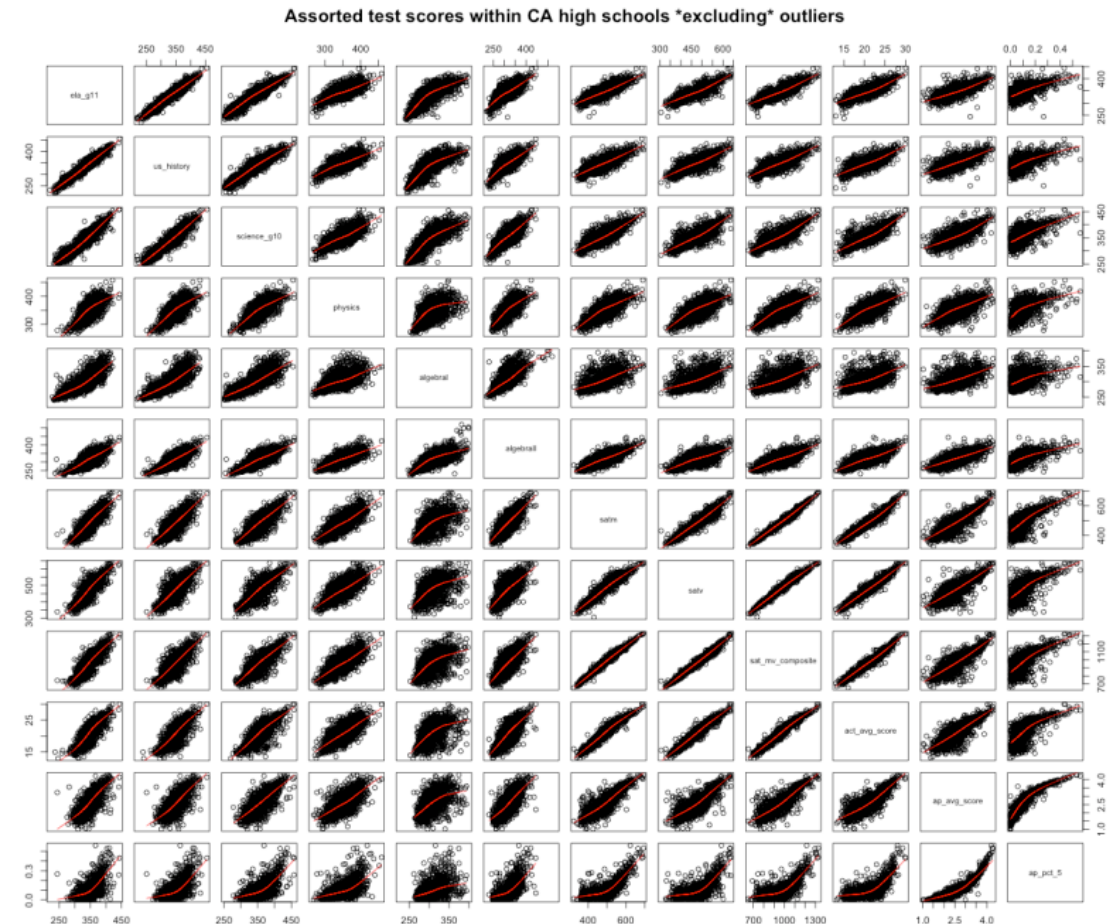
Graphical Outlier Detection

- 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.



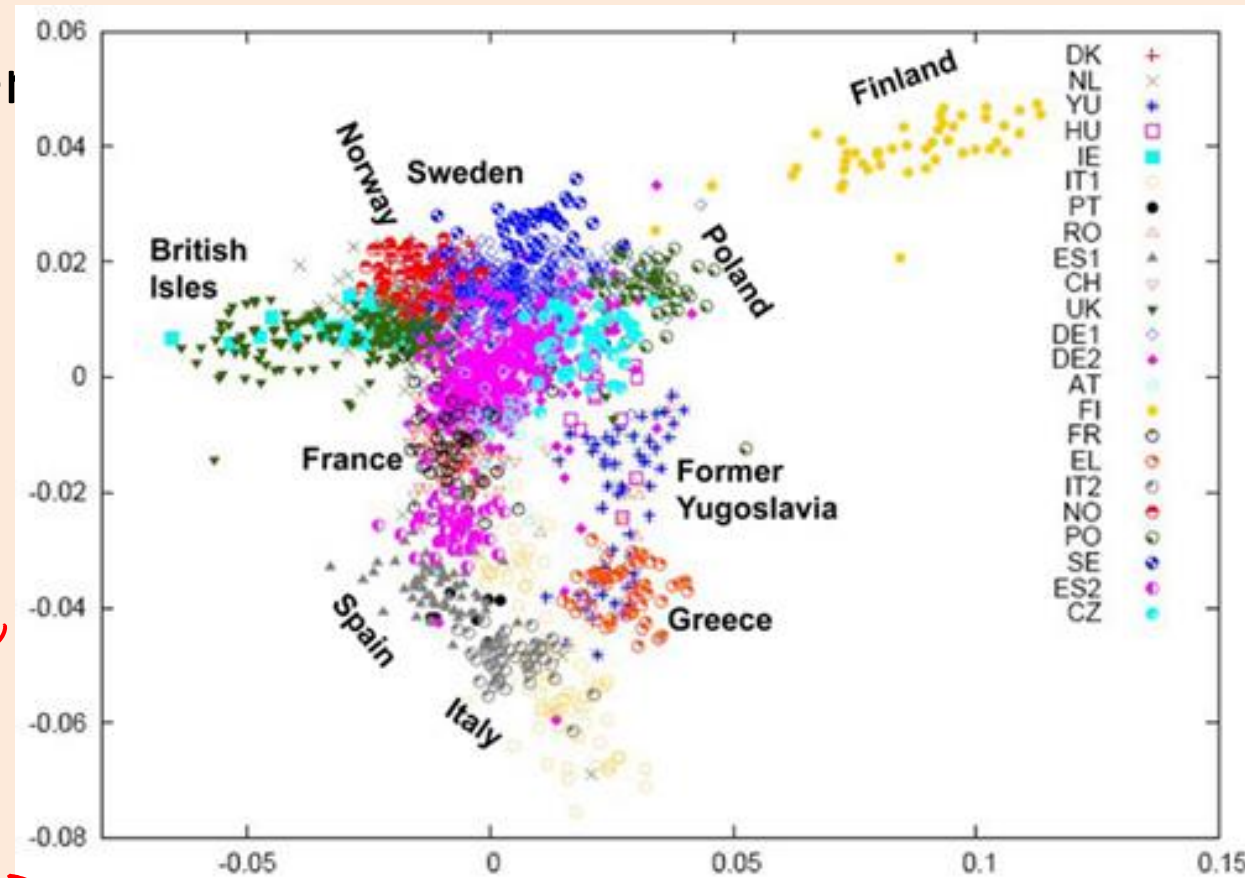
Graphical Outlier Detection

- Graphical approach to outlier detection:

1. Look at a plot of the data.
2. Human decides if data is an outlier

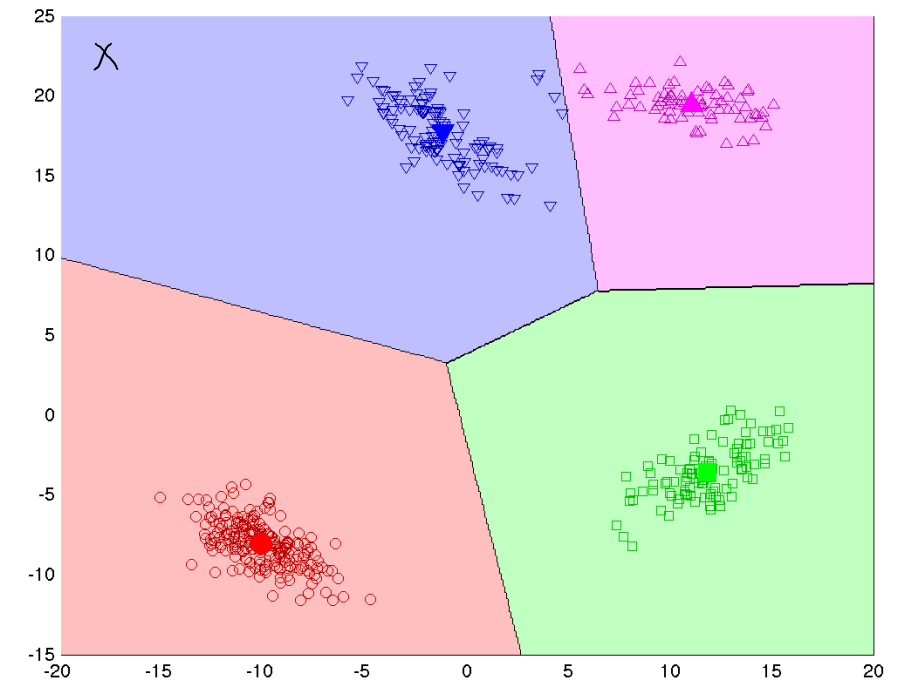
- Examples:

1. Box plot.
2. Scatterplot.
3. Scatterplot array.
4. Scatterplot of 2-dimensional PCA:
 - 'See' high-dimensional structure.
 - But loses information and sensitive to outliers.



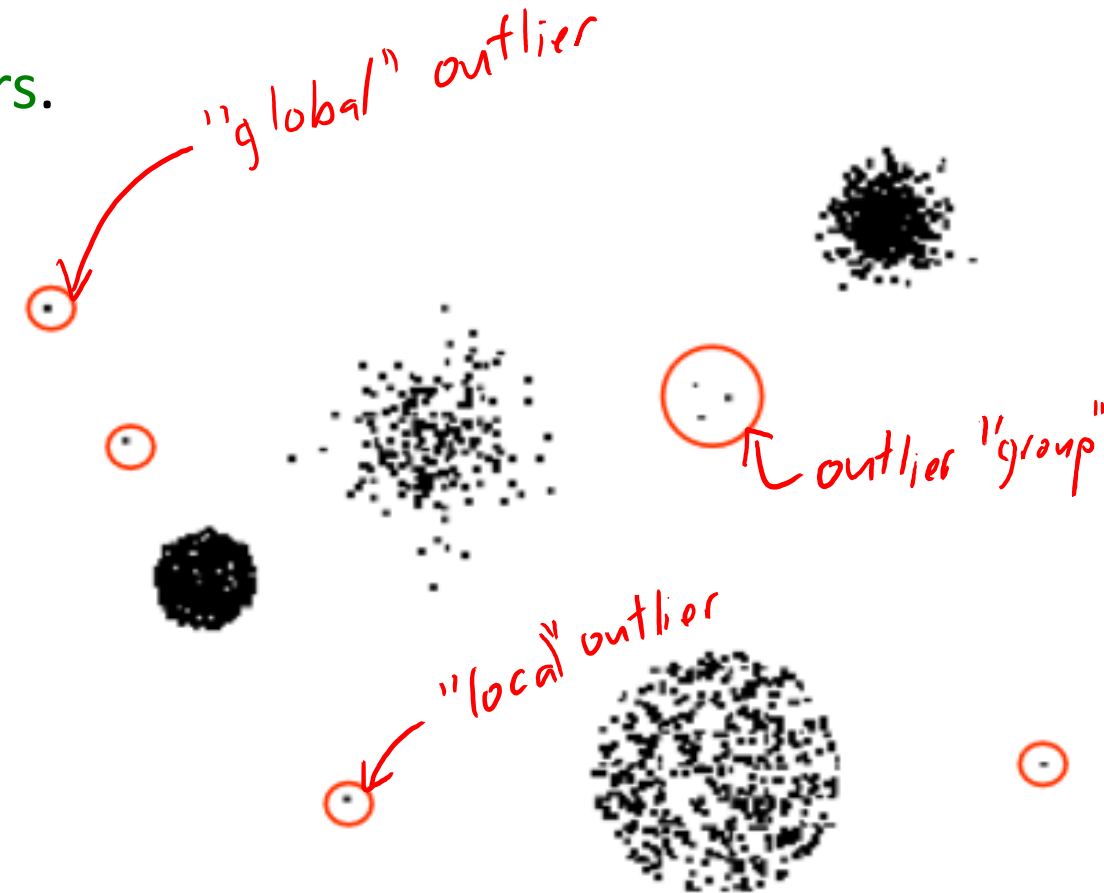
Cluster-Based Outlier Detection

- 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.



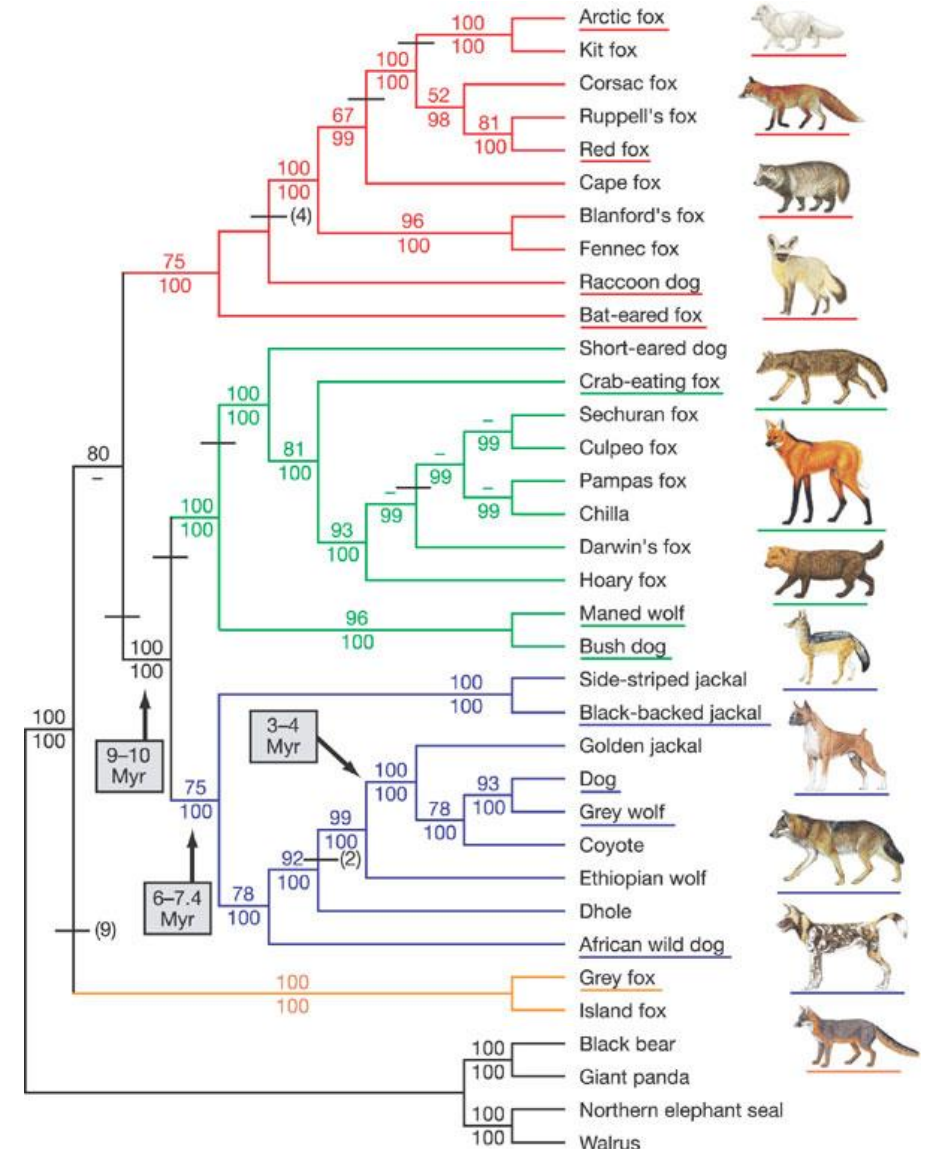
Cluster-Based Outlier Detection

- 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.



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.



Distance-Based Outlier Detection

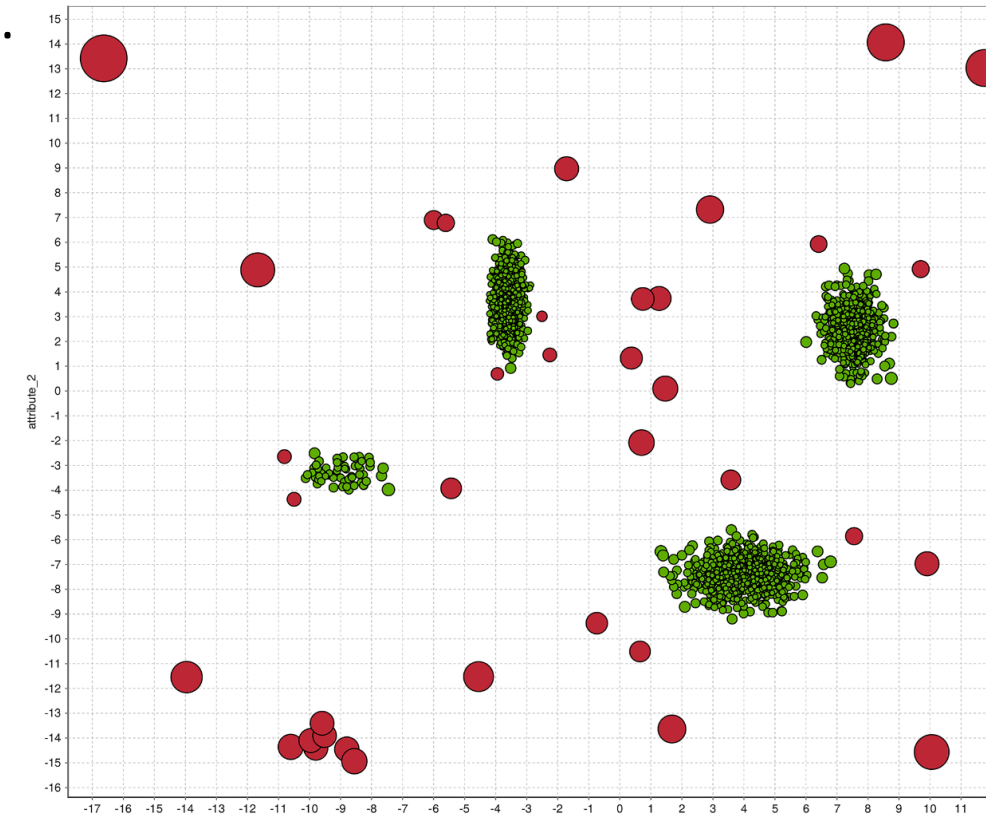
- Most outlier detection approaches are **based on distances**.
- Can we skip the model/plot/clustering and **just measure distances**?
 - How many points lie in a radius 'epsilon'?
 - What is distance to k^{th} nearest neighbour?
- UBC connection (first paper on this topic):

Algorithms for Mining Distance-Based Outliers in Large Datasets

Edwin M. Knorr and Raymond T. Ng
Department of Computer Science
University of British Columbia

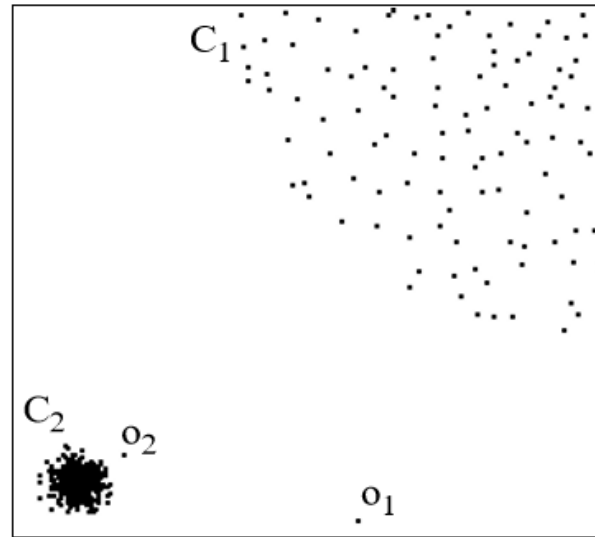
Global Distance-Based Outlier Detection: KNN

- KNN outlier detection:
 - For each point, compute the **average distance to its KNN**.
 - Choose points with biggest values (or values above a threshold) as outliers.
 - “Outliers” are points that are far from their KNNs.
- Goldstein and Uchida [2016]:
 - Compared 19 methods on 10 datasets.
 - **KNN best for finding “global” outliers.**
 - “Local” outliers best found with **local distance-based** methods...



Local Distance-Based Outlier Detection

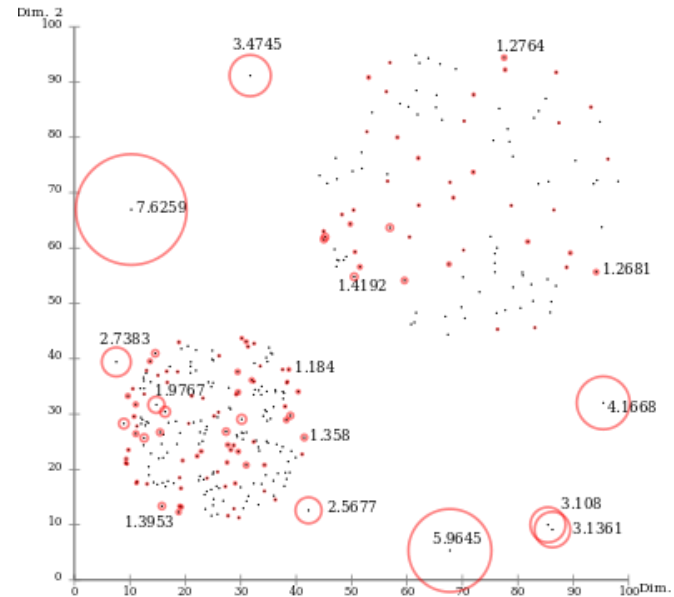
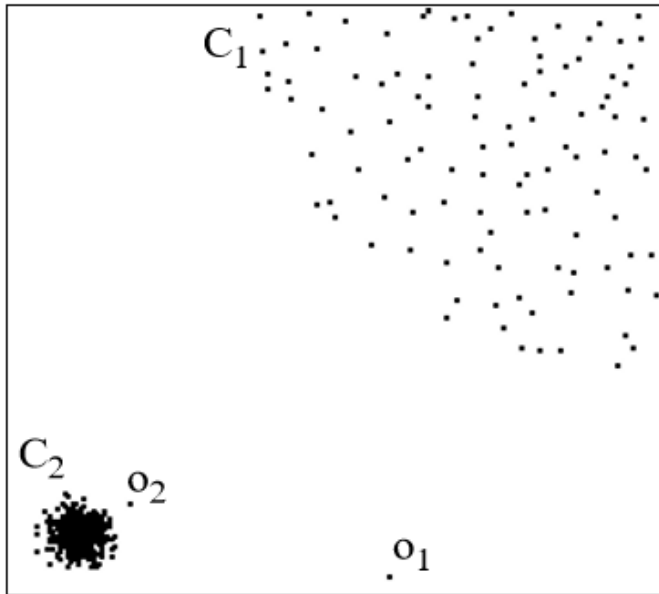
- As with density-based clustering, **problem with differing densities:**



- Outlier o_2 has similar density as elements of cluster C_1 .
- Basic idea behind **local distance-based** methods:
 - Outlier o_2 is “**relatively**” **far** compared to its neighbours.

Local Distance-Based Outlier Detection

- “Outlierness” ratio of example ‘i’:
$$\frac{\text{average distance of 'i' to its } KNN_5}{\text{average distance of neighbours of 'i' to their } KNN_5}$$
- If outlierness > 1 , x_i is further away from neighbours than expected.

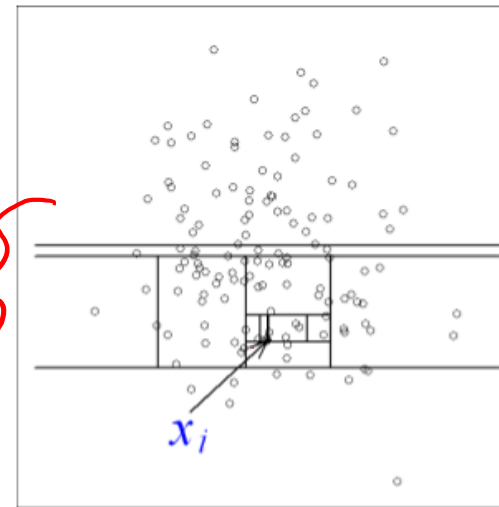


Isolation Forests

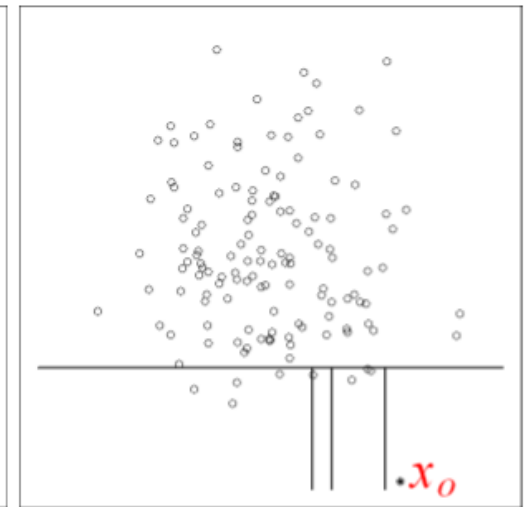
- Recent method based on random trees is **isolation forests**.
 - Grow a tree where **each stump uses a random feature and random split**.
 - Stop when each example is “isolated” (each leaf has one example).
 - The “**isolation score**” is the depth before example gets isolated.
 - Outliers should be isolated quickly, inliers should need lots of rules to isolate.

- Repeat for different random trees, take average score.

Depth 12:
- needed 12
rules to isolate
so may be inlier.



(a) Isolating x_i

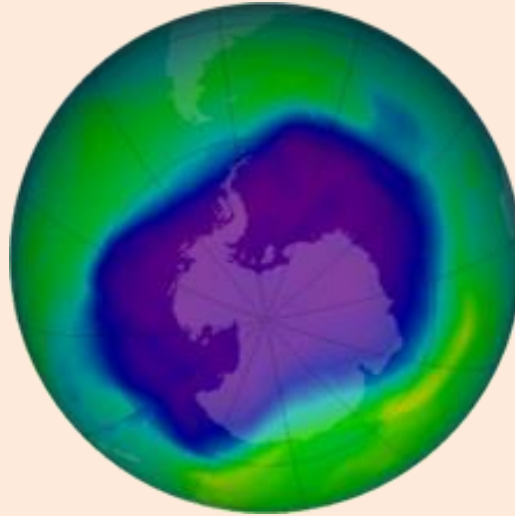


(b) Isolating x_o

depth 4
so more
likely to
be outlier

Problem with Unsupervised Outlier Detection

- Why wasn't the hole in the ozone layer discovered for 9 years?



- Can be **hard to decide when to report** an outlier:
 - If **you report too many non-outliers, users will turn you off.**
 - Most antivirus programs do not use ML methods (see ["base-rate fallacy"](https://en.wikipedia.org/wiki/Base_rate_fallacy))

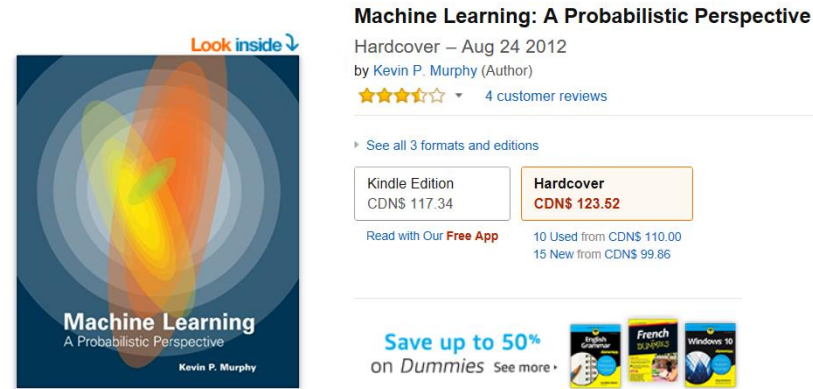
Supervised Outlier Detection

- Final approach to outlier detection is to use supervised learning:
 - $y_i = 1$ if x_i is an outlier.
 - $y_i = 0$ if x_i is a regular point.
- We can use our methods for supervised learning:
 - We can find very complicated outlier patterns.
 - Classic credit card fraud detection methods used decision trees.
- But it needs supervision:
 - We need to know what outliers look like.
 - We may not detect new “types” of outliers.

(pause)

Motivation: Product Recommendation

- A customer comes to your website looking to buy at item:



- You want to **find similar items** that they might also buy:

Customers Who Bought This Item Also Bought

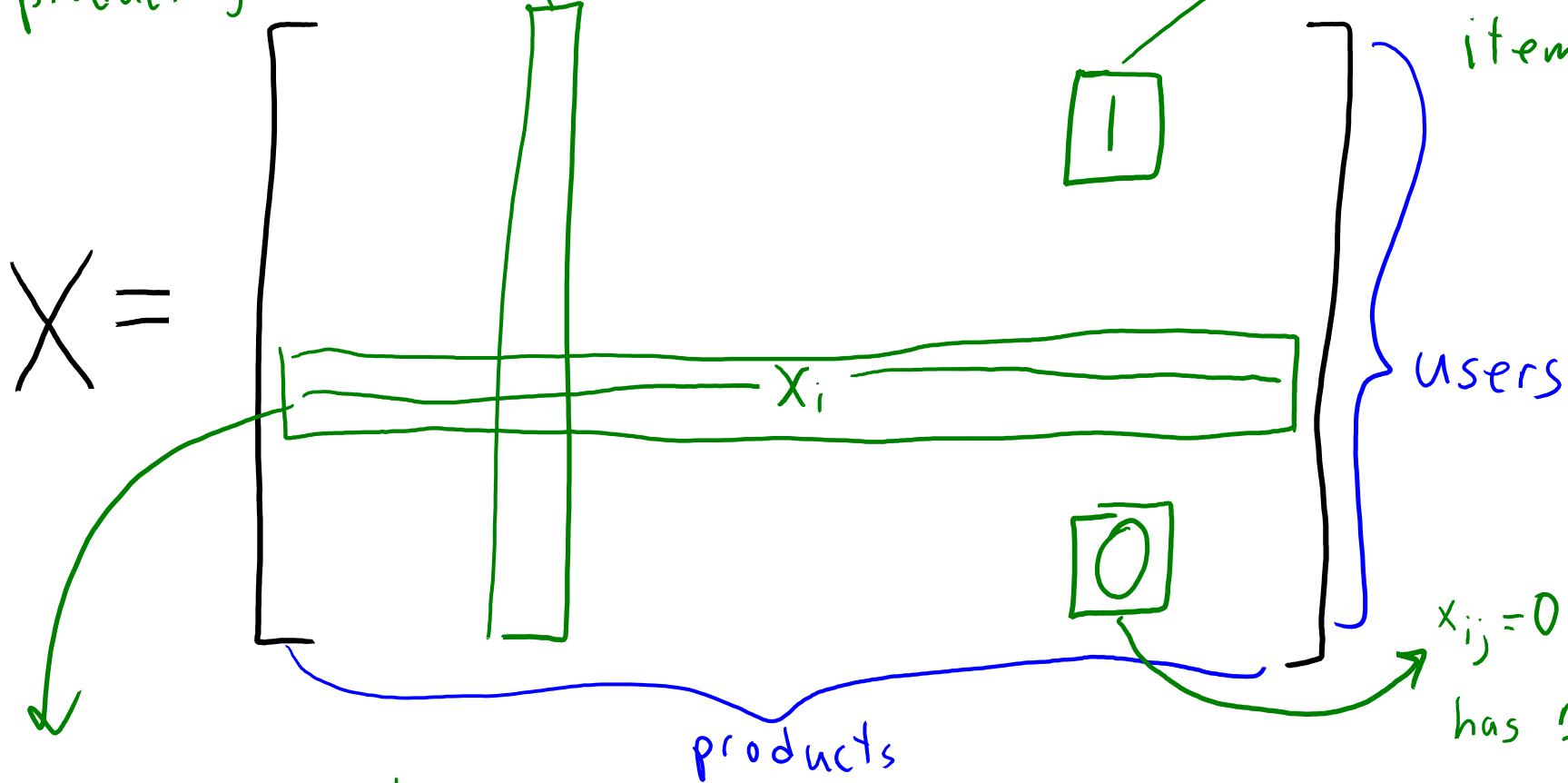
Page 1 of 20



User-Product Matrix

Column x^j gives
all users that
bought product ' j '

$x_{ij} = 1$ means
user ' i ' bought
item ' j '!



$x_{ij} = 0$ means user ' i '
has not buy item ' j '

Row x_i gives all items bought by user ' i '.

Amazon Product Recommendation

- Amazon product recommendation method:

A hand-drawn diagram illustrating a matrix X . The matrix is represented by two vertical rectangles enclosed in large square brackets. A green arrow labeled 'vs.' points from the top of the first rectangle to the top of the second rectangle. A green arrow labeled 'product' points from below the second rectangle. A black arrow labeled 'user' points from the right side of the brackets.

$$X = \left[\begin{array}{c} \text{vs.} \\ \text{product} \end{array} \right] \leftarrow \text{user}$$

- Return the **KNNs across columns**.
 - Find 'j' values minimizing $||x^i - x^j||$.
 - **Products that were bought by similar sets of users.**
- But first **divide each column by its norm**, $x^i / ||x^i||$.
 - This is called **normalization**.
 - Reflects whether product is bought by many people or few people.

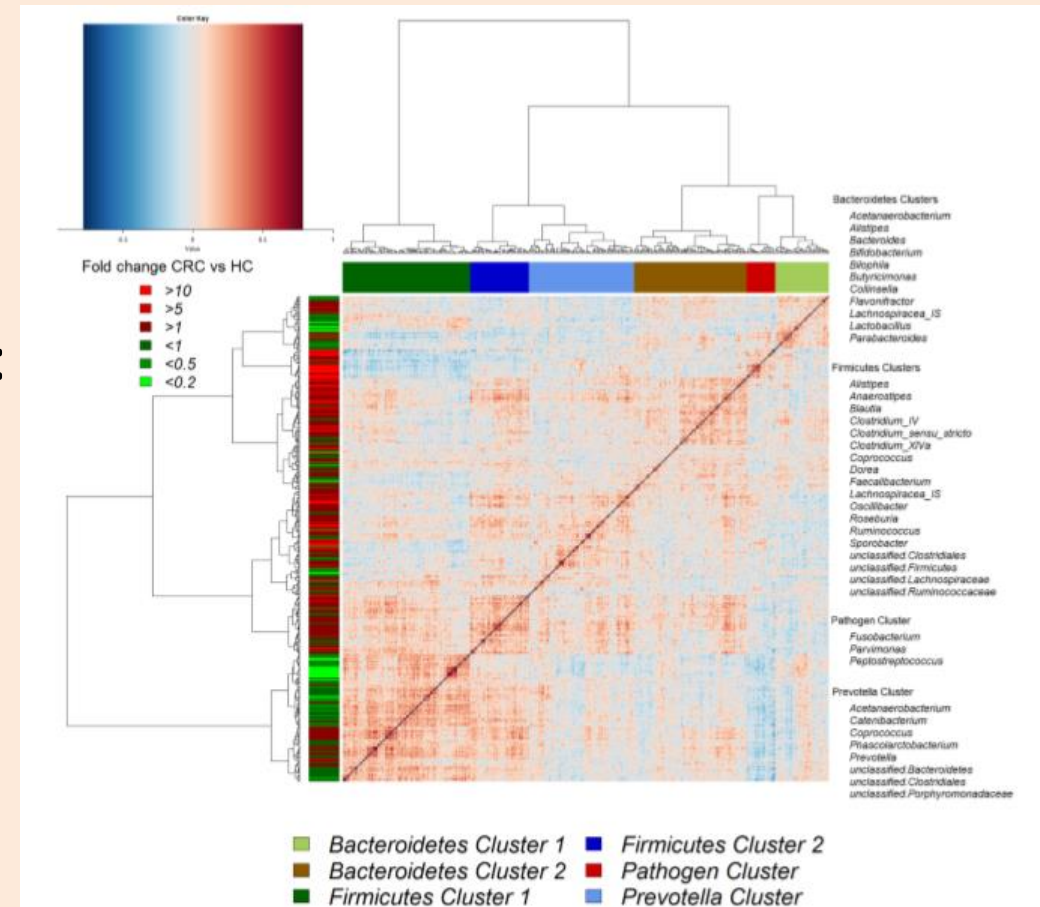
Summary

- **Biclustering**: clustering of the examples *and* the features.
- **Outlier detection** is task of finding unusually different example.
 - A concept that is very difficult to define.
 - **Model-based** find unlikely examples given a model of the data.
 - **Graphical** methods plot data and use human to find outliers.
 - **Cluster-based** methods check whether examples belong to clusters.
 - **Distance-based outlier detection**: measure (relative) distance to neighbours.
 - **Supervised-learning for outlier detection**: turns task into supervised learning.
- **Amazon product recommendation**:
 - Find similar items using (normalized) nearest neighbour search.
- Next time: detecting genes, viruses, plagiarism, and fingerprints.

Application: Medical data

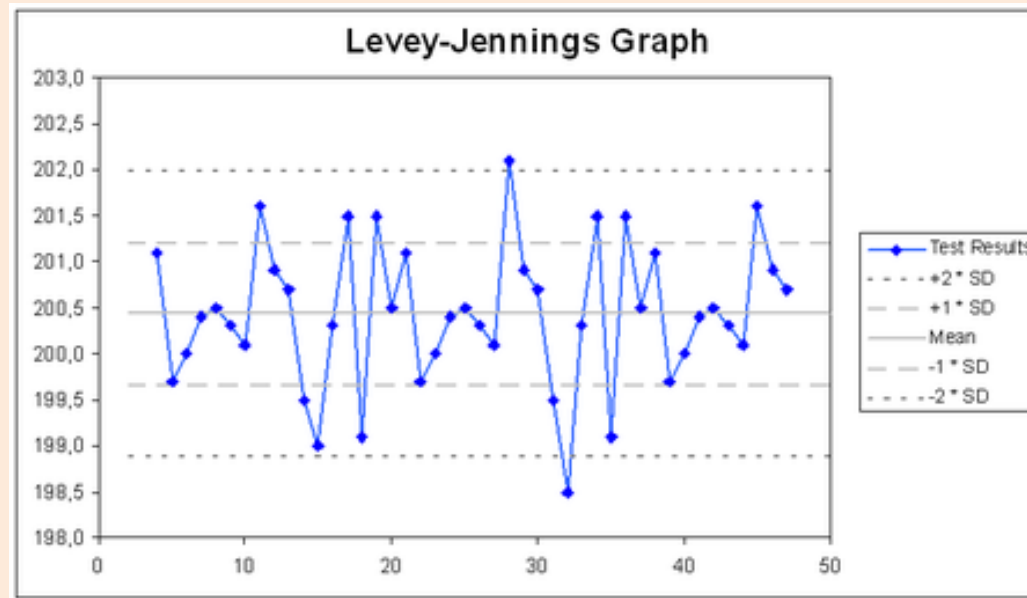
- Hierarchical clustering is very common in **medical data analysis**.
 - Clustering different samples of colorectal cancer:

- This plot is different, it's not a biclustering:
 - The matrix is 'n' by 'n'.
 - **Each matrix element gives correlation.**
 - Clusters should look like “blocks” on diagonal.
 - **Order of examples is reversed in columns.**
 - This is why diagonal goes from bottom-to-top.
 - Please don't do this reversal, it's confusing to me.



“Quality Control”: Outlier Detection in Time-Series

- A field primarily focusing on outlier detection is **quality control**.
- One of the main tools is plotting z-score thresholds over time:



- Usually don't do tests like " $|z_i| > 3$ ", since this happens normally.
- Instead, identify problems with tests like " $|z_i| > 2$ twice in a row".

Outlierness (Symbol Definition)

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

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

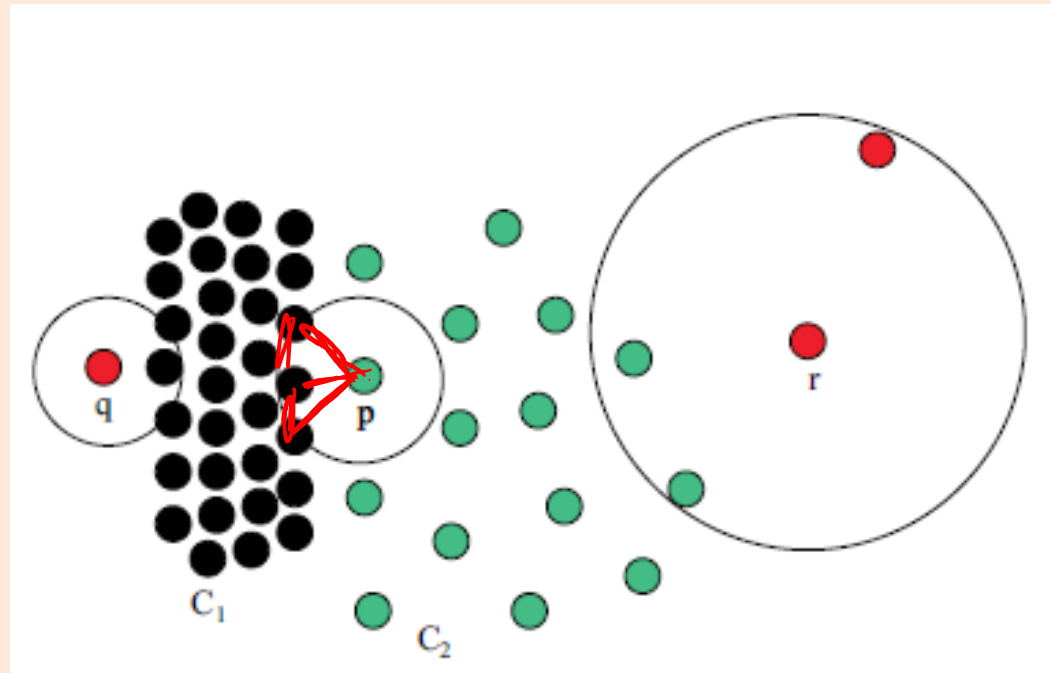
- **Outlierness** is ratio of $D_k(x_i)$ to average $D_k(x_j)$ for its neighbours 'j':

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

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

Outlierness with Close Clusters

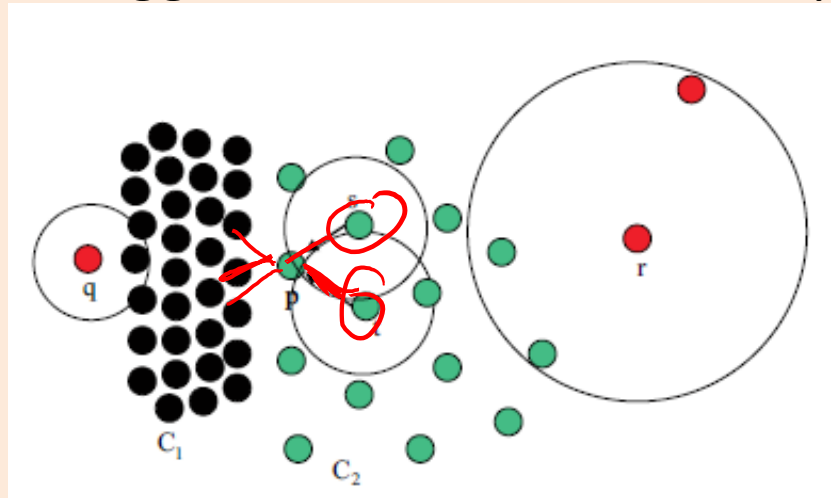
- If clusters are close, outlierness gives unintuitive results:



- 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'.

Outlierness with Close Clusters

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



- But still has problems:
 - Dealing with hierarchical clusters.
 - Yields many false positives if you have “global” outliers.
 - Goldstein and Uchida [2016] recommend just using KNN.

Training/Validation/Testing (Supervised)

- A typical supervised learning setup:
 - Train parameters on dataset D_1 .
 - Validate hyper-parameters on dataset D_2 .
 - Test error evaluated on dataset D_3 .
- What should we choose for D_1 , D_2 , and D_3 ?
- Usual answer: should all be IID samples from data distribution D_s .

Training/Validation/Testing (Outlier Detection)

- A typical outlier detection setup:
 - Train parameters on dataset D_1 (there may be no “training” to do).
 - For example, find z-scores.
 - Validate hyper-parameters on dataset D_2 (for outlier detection).
 - For example, see which z-score threshold separates D_1 and D_2 .
 - Test error evaluated on dataset D_3 (for outlier detection).
 - For example, check whether z-score recognizes D_3 as outliers.
- D_1 will still be samples from D_s (data distribution).
- D_2 could use IID samples from another distribution D_m .
 - D_m represents the “none” or “outlier” class.
 - Tune parameters so that D_m samples are outliers and D_s samples aren't.
 - Could just fit a binary classifier here.

Training/Validation/Testing (Outlier Detection)

- A typical outlier detection setup:
 - Train parameters on dataset D_1 (there may be no “training” to do).
 - For example, find z-scores.
 - Validate hyper-parameters on dataset D_2 (for outlier detection).
 - For example, see which z-score threshold separates D_1 and D_2 .
 - Test error evaluated on dataset D_3 (for outlier detection).
 - For example, check whether z-score recognizes D_3 as outliers.
- D_1 will still be samples from D_s (data distribution).
- D_2 could use IID samples from another distribution D_m .
- D_3 could use IID samples from D_m .
 - How well do you do at recognizing “data” samples from “none” samples?

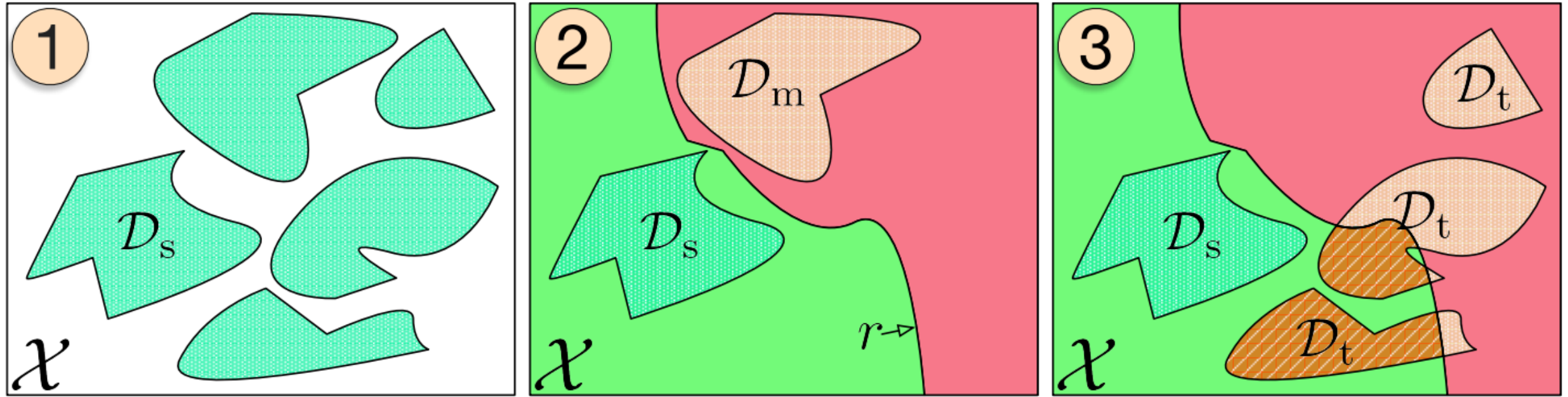
Training/Validation/Testing (Outlier Detection)

- Seems like a reasonable setup:
 - D_1 will still be **samples from D_s** (data distribution).
 - D_2 could use **IID samples from another distribution D_m** .
 - D_3 could use **IID samples from D_m** .
- What can go wrong?
- You **needed to pick a distribution D_m** to represent “none”.
 - But in the wild, your **outliers might follow another “none” distribution**.
 - This procedure can overfit to your D_m .
 - You can **overestimate your ability to detect outliers**.

OD-Test: a better way to evaluate outlier detections

- A reasonable setup:
 - D_1 will still be **samples from D_s** (data distribution).
 - D_2 could use **IID samples from another distribution D_m** .
 - ~~– D_3 could use **IID samples from D_m** .~~
 - D_3 could use **IID samples from yet-another distribution D_t** .
- “How do you perform at detecting different types of outliers?”
 - Seems like a harder problem, but arguably closer to reality.

OD-Test: a better way to evaluate outlier detections



- “How do you perform at detecting different types of outliers?”