

CPSC 340: Machine Learning and Data Mining

Hierarchical Clustering

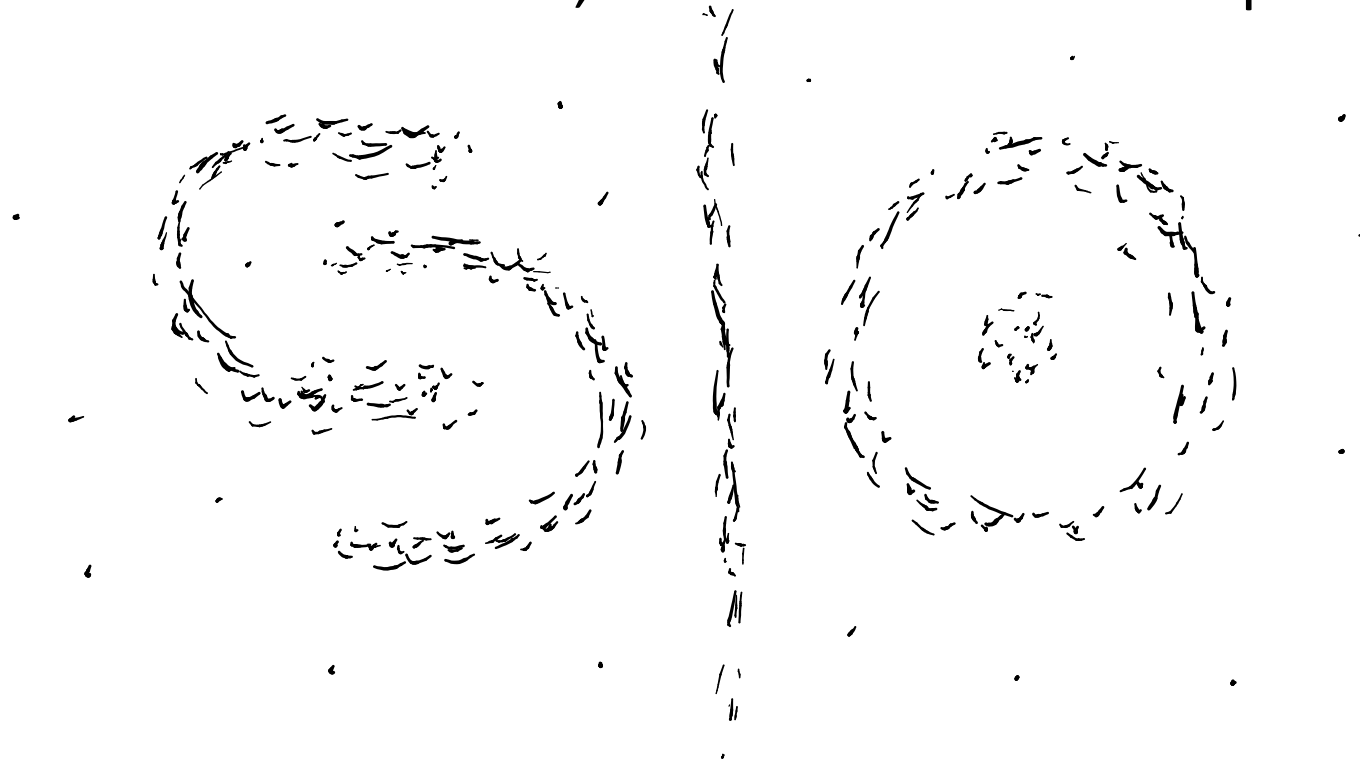
Fall 2017

Admin

- **Assignment 1** is due Friday.
 - Follow the assignment guidelines naming convention (a1.zip/a1.pdf).
- Assignment 0 grades posted on Connect.

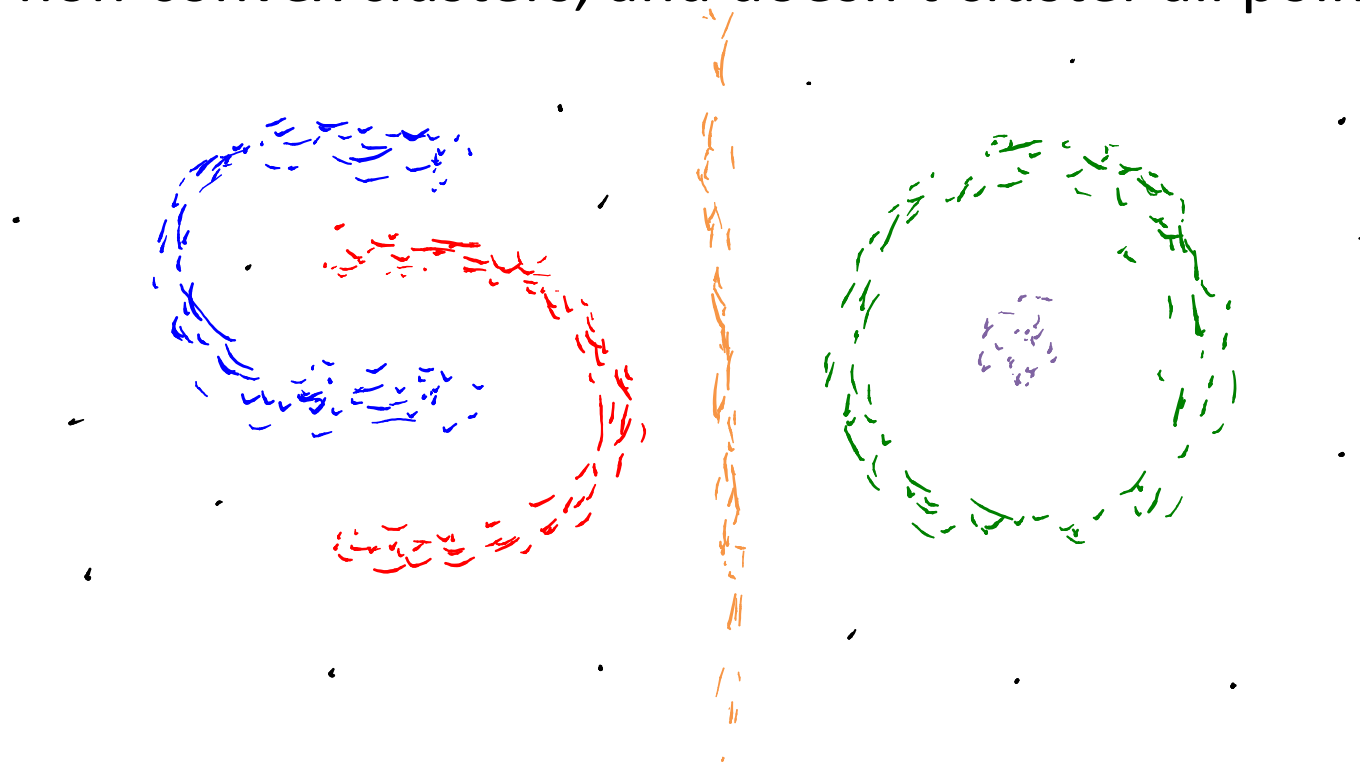
Last Time: Density-Based Clustering

- We discussed **density-based clustering**:
 - **Non-parametric** clustering method.
 - Based on finding **connected regions of dense** points.
 - Can find non-convex clusters, and doesn't cluster all points.



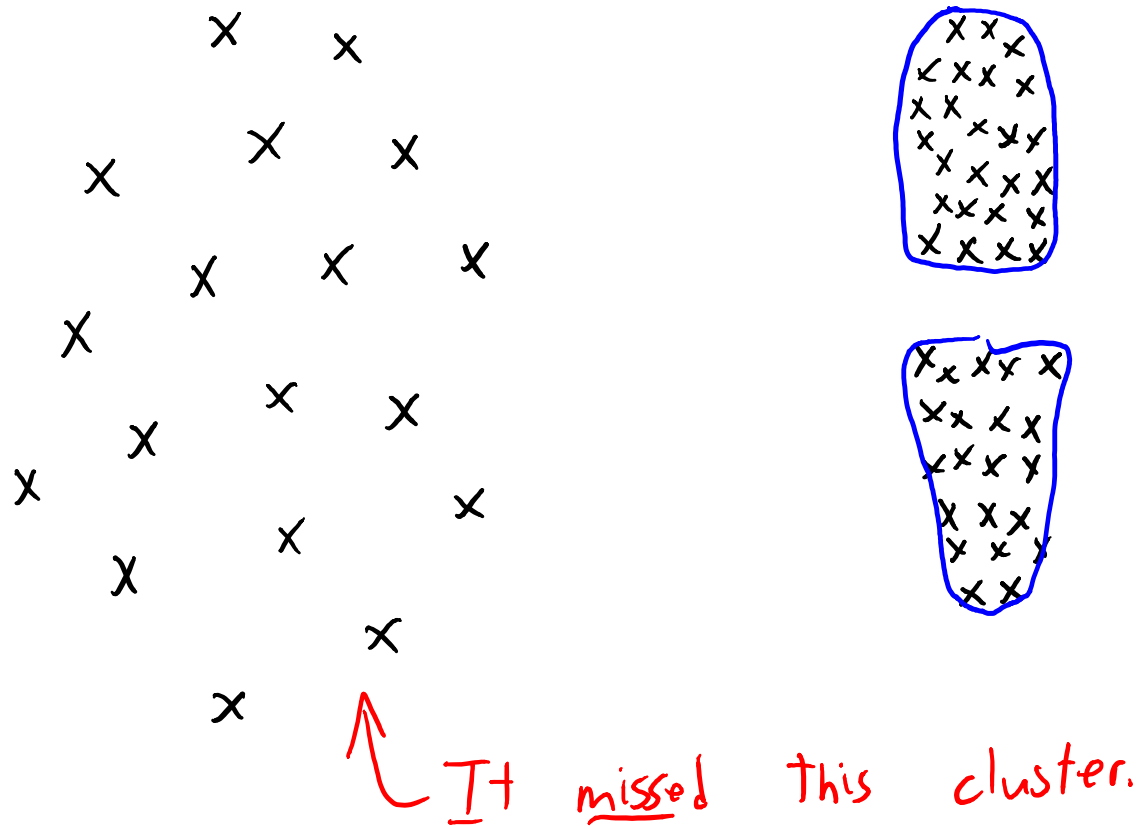
Last Time: Density-Based Clustering

- We discussed **density-based clustering**:
 - **Non-parametric** clustering method.
 - Based on finding **connected regions of dense** points.
 - Can find non-convex clusters, and doesn't cluster all points.



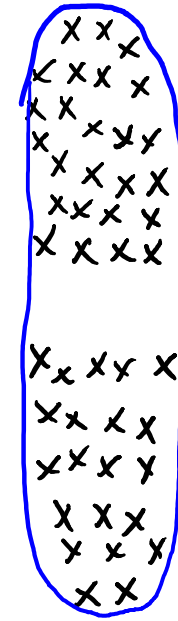
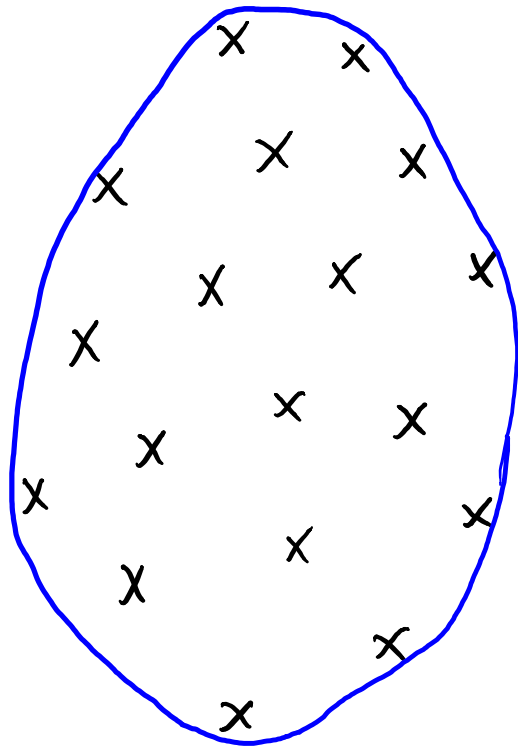
Differing Densities

- Consider density-based clustering on this data:



Differing Densities

- Increase the radius and run it again:

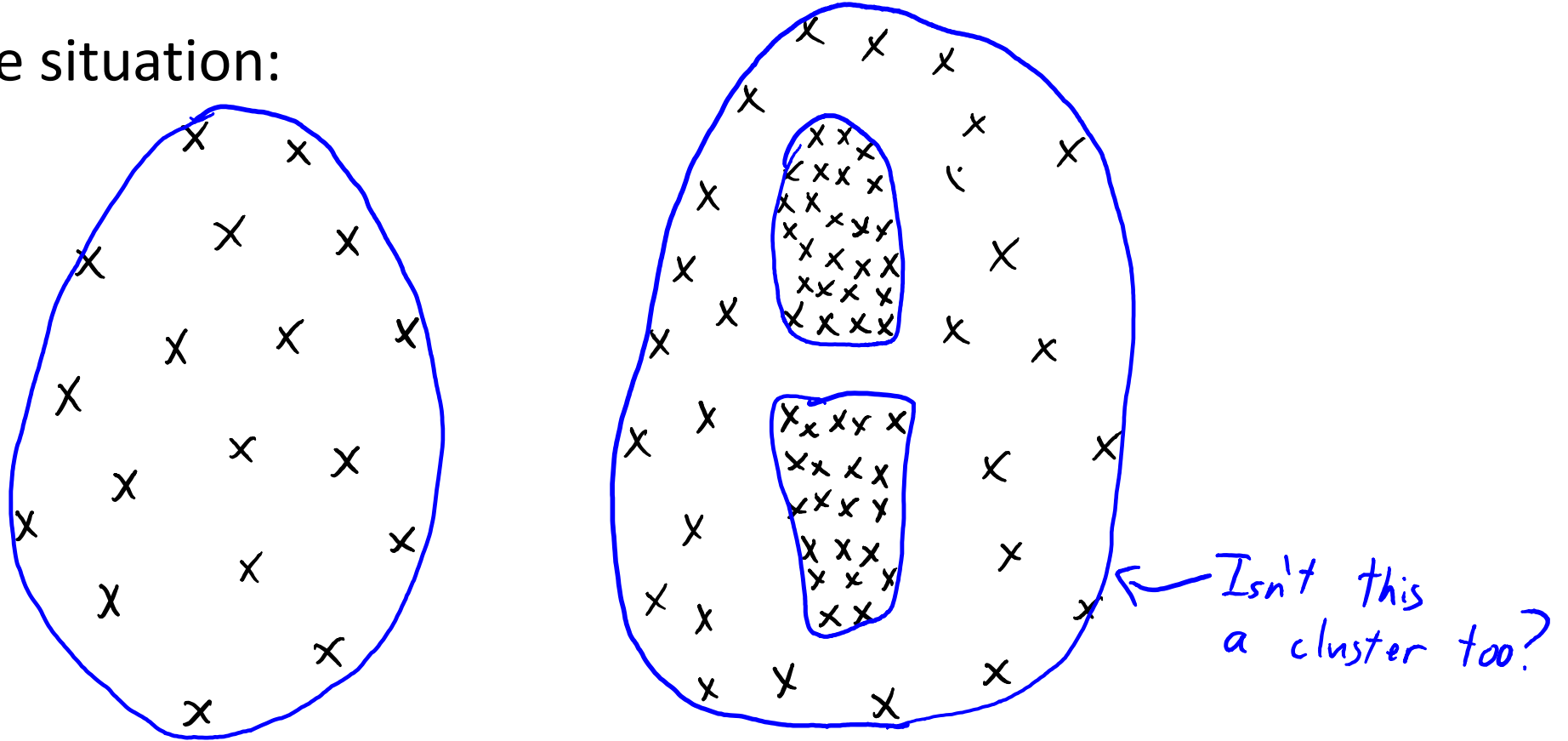


These 2 clusters
are now "close."

- There may no density-level that gives you 3 clusters.

Differing Densities

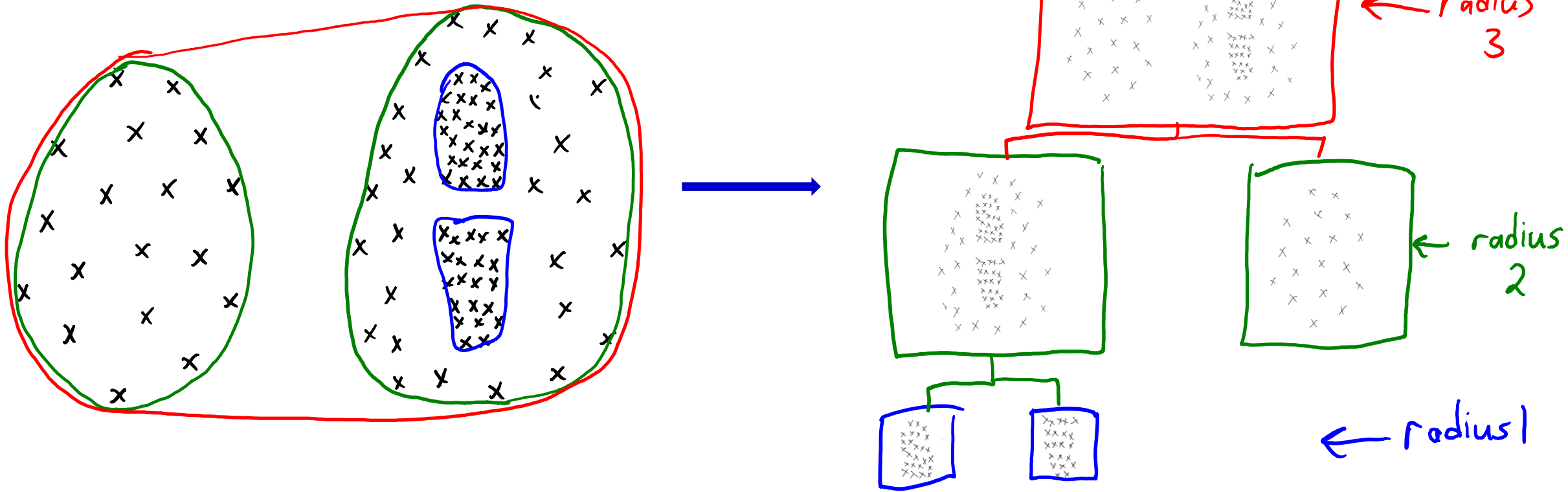
- Here is a worse situation:



- Now you need to choose between coarse/fine clusters.
- Instead of fixed clustering, we often want [hierarchical clustering](#).

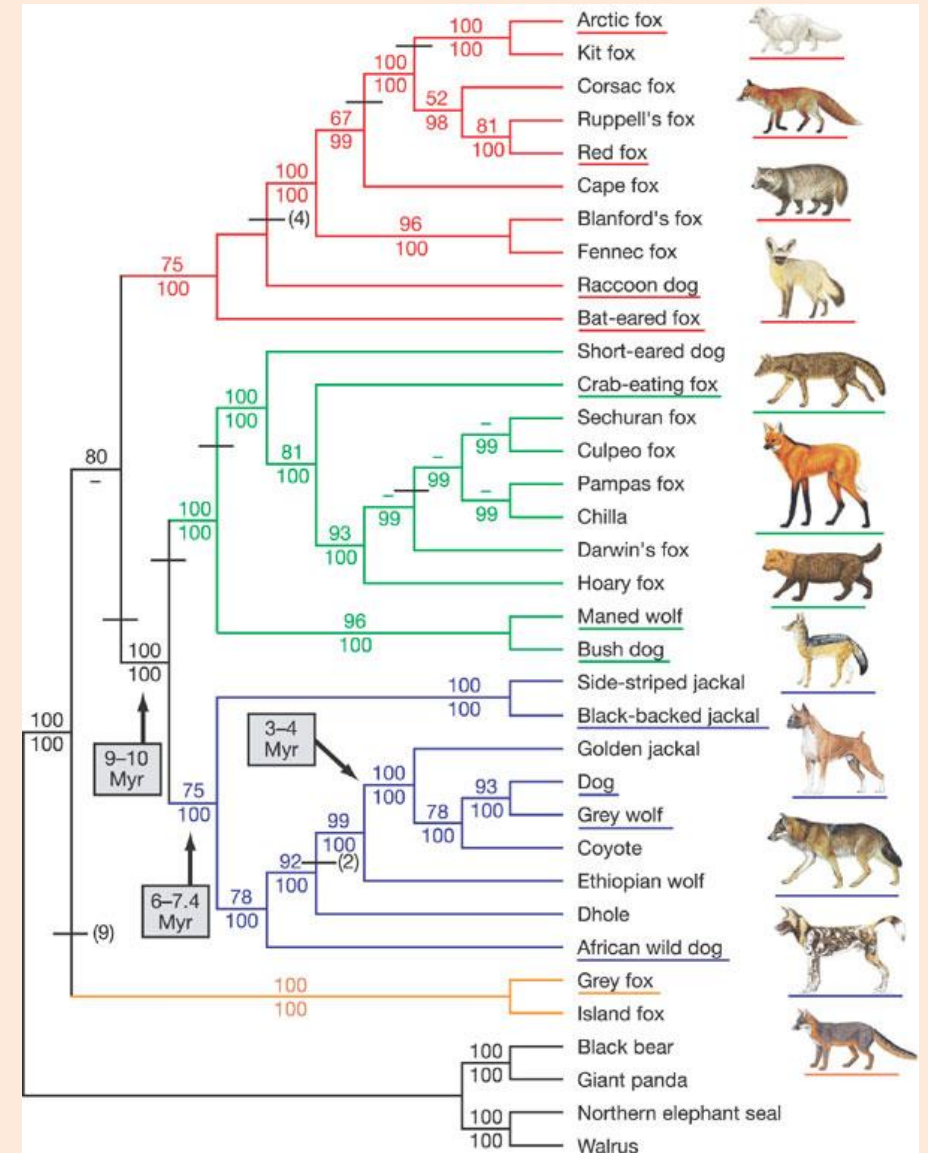
Density-Based Hierarchical Clustering

- A simple way to make a **hierarchical DBSCAN**:
 - Fix minPoints, **record clusters as you vary the radius**.
 - Much more information than using a fixed radius.



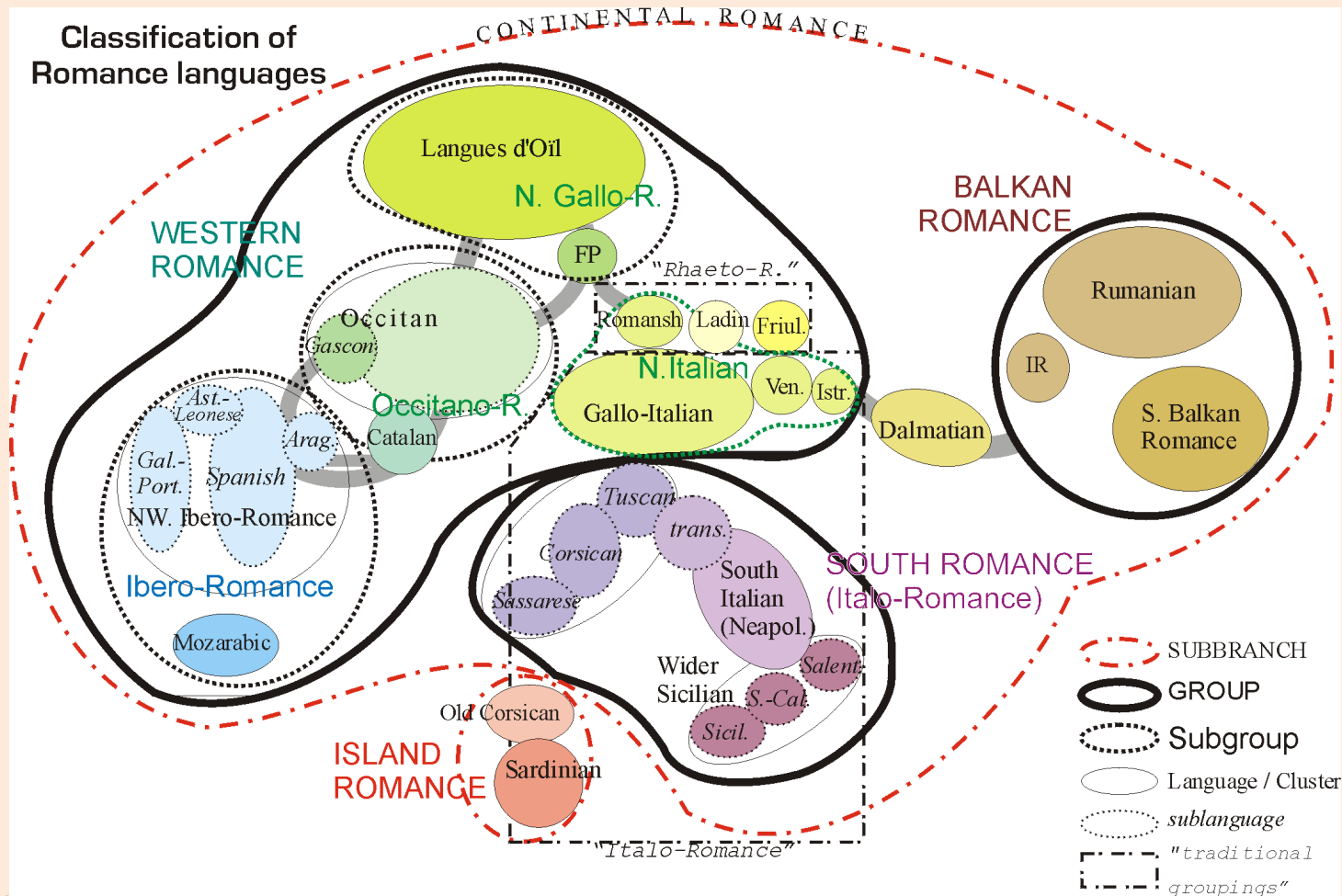
Application: Phylogenetics

- We sequence genomes of a set of organisms.
- Can we construct the “tree of life”?
- Comments on this application:
 - On the right are individuals.
 - As you go left, clusters merge.
 - Merges are ‘common ancestors’.
- More useful information in the plot:
 - Line lengths: chose here to approximate time.
 - Numbers: #clusterings across bootstrap samples.
 - ‘Outgroups’ (walrus, panda) are a sanity check.



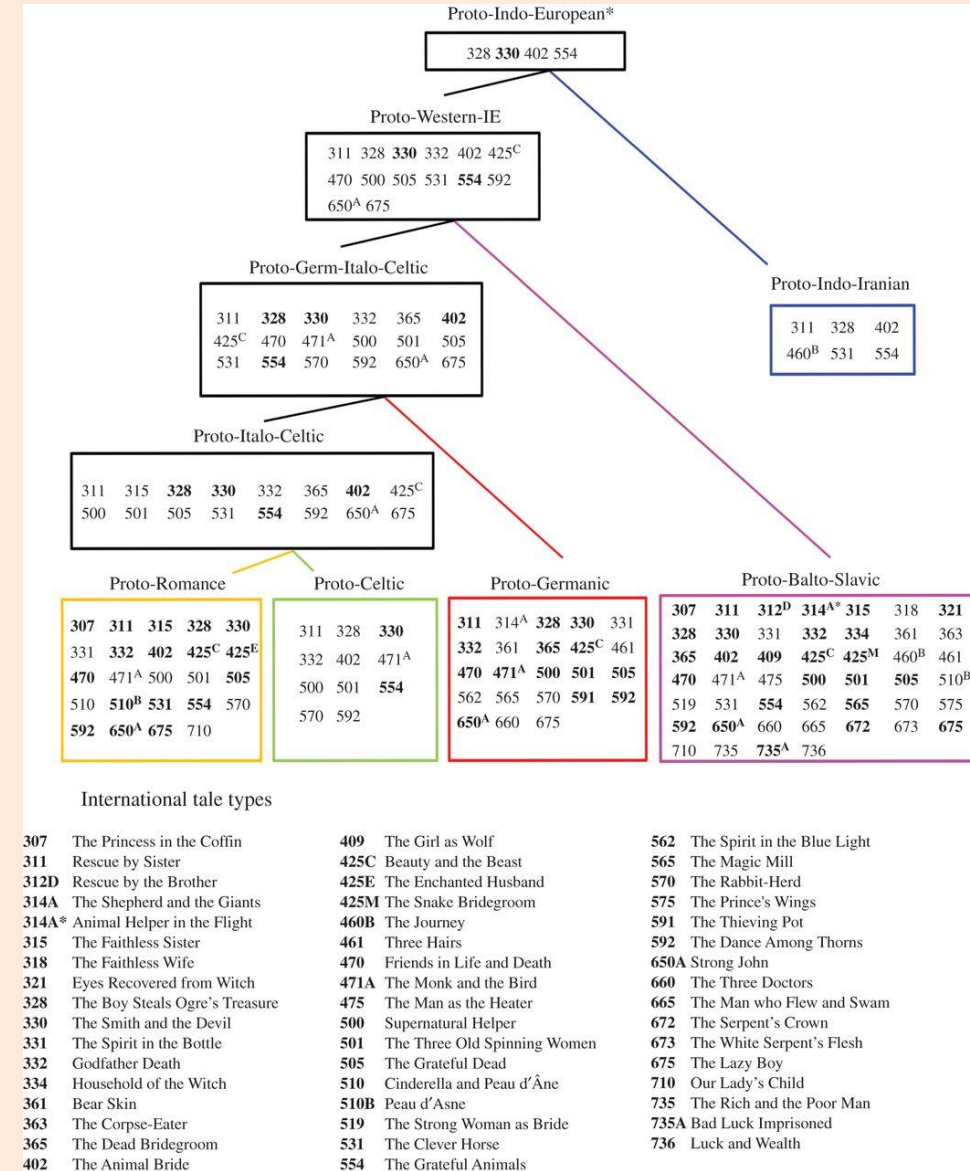
Application: Phylogenetics

- Comparative method in linguistics studies evolution of languages:



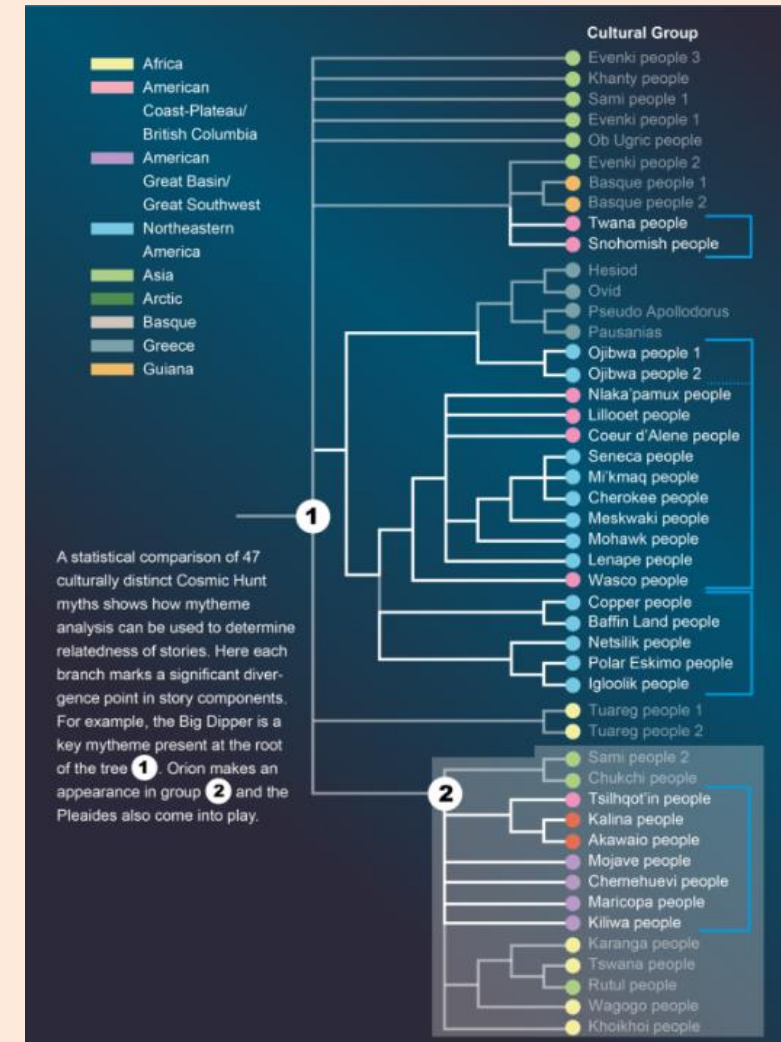
Application: Phylogenetics

- January 2016: evolution of fairy tales.
 - Evidence that “Devil and the Smith” goes back to bronze age.
 - “Beauty and the Beast” published in 1740, but might be 2500-6000 years old.



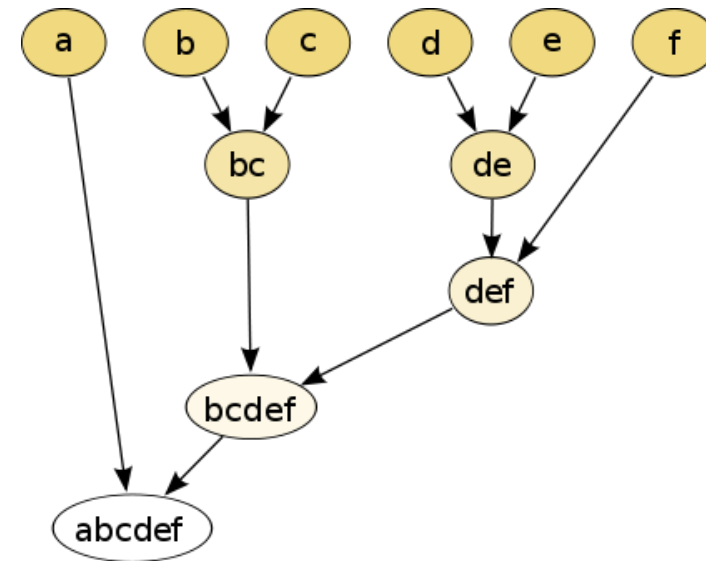
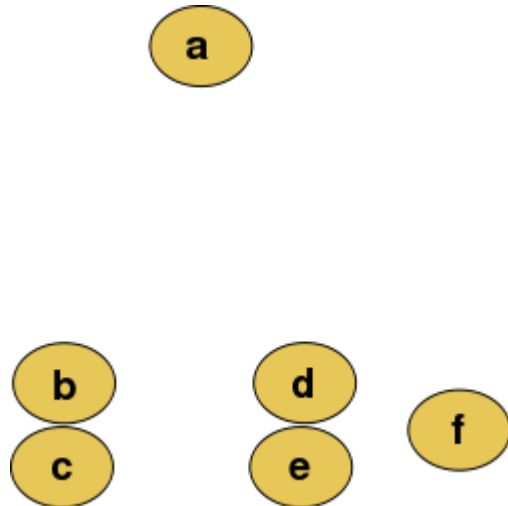
Application: Phylogenetics

- January 2016: evolution of fairy tales.
 - Evidence that “Devil and the Smith” goes back to bronze age.
 - “Beauty and the Beast” published in 1740, but might be 2500-6000 years old.
- September 2016: evolution of myths.
 - “Comic hunt” story:
 - Person hunts animal that becomes constellation.
 - Previously known to be at least 15,000 years old.
 - May go back to paleolithic period.



Agglomerative (Bottom-Up) Clustering

- More common hierarchical method: **agglomerative clustering**.
 1. Starts with each point in its own cluster.
 2. Each step merges the two “closest” clusters.
 3. Stop when everything is in one big cluster.

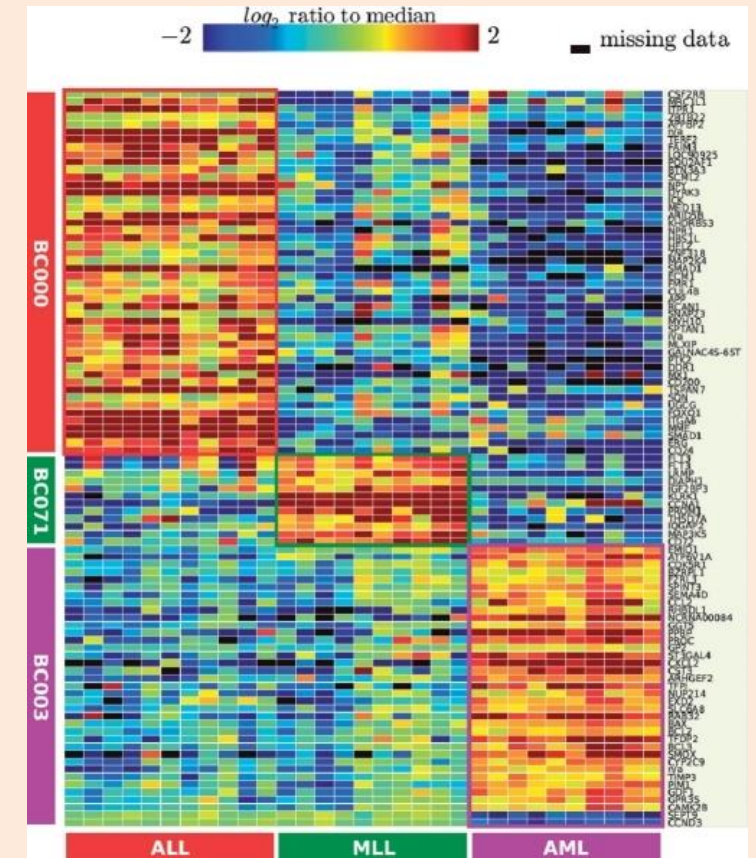


Agglomerative (Bottom-Up) Clustering

- Reinvented by different fields under different names (“UPGMA”).
- Needs a “distance” between two clusters.
- A standard choice: distance between means of the clusters.
 - Not necessarily the best, many choices exist (bonus slide).
- Cost is $O(n^3d)$, so this only makes sense for medium-size datasets.

Other Clustering Methods

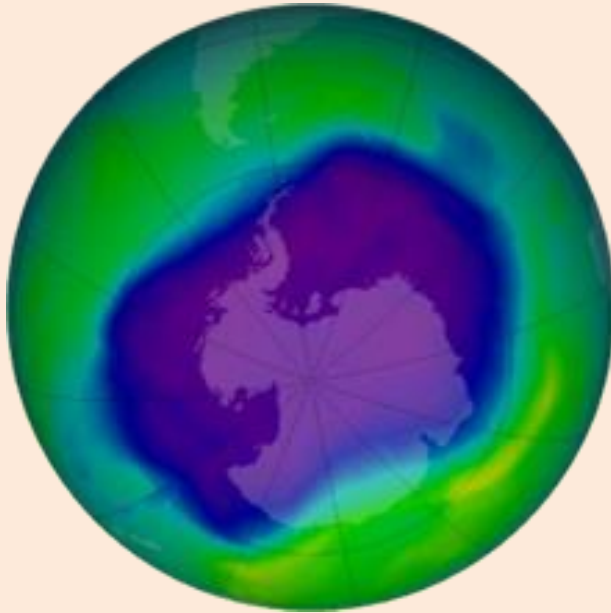
- Mixture models:
 - Probabilistic clustering.
- Mean-shift clustering:
 - Finds local “modes” in density of points.
- Bayesian clustering:
 - A variant on ensemble methods.
 - Averages over models/clustering, weighted by “prior” belief in the model/clustering.
- Biclustering:
 - Simultaneously cluster objects and features.
- Spectral clustering and graph-based clustering:
 - Clustering of data described by graphs.



(pause)

Motivating Example: Finding Holes in Ozone Layer

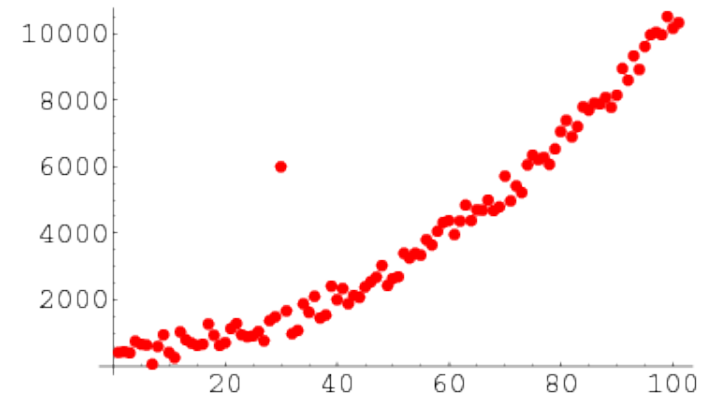
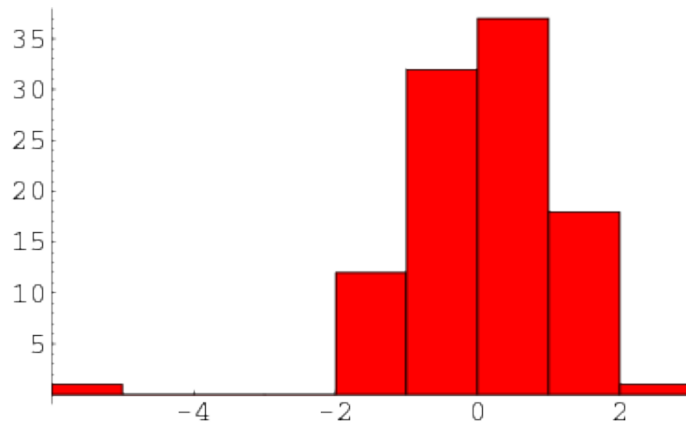
- The huge Antarctic ozone hole was “discovered” in 1985.



- It had been in satellite data since 1976:
 - But it was flagged and filtered out by 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 (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.
 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	...	Sick?
0	0.7	0	0.3	0	0		1
0.3	0.7	0	0.6	0	-1		1
0	0	0	"sick"	0	0		0
0.3	0.7	1.2	0	0.10	0.01		-1
900	0	1.2	0.3	0.10	0.01		1

- We should check basic things like:
 - 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**?
- These simple errors are VERY common in real data.

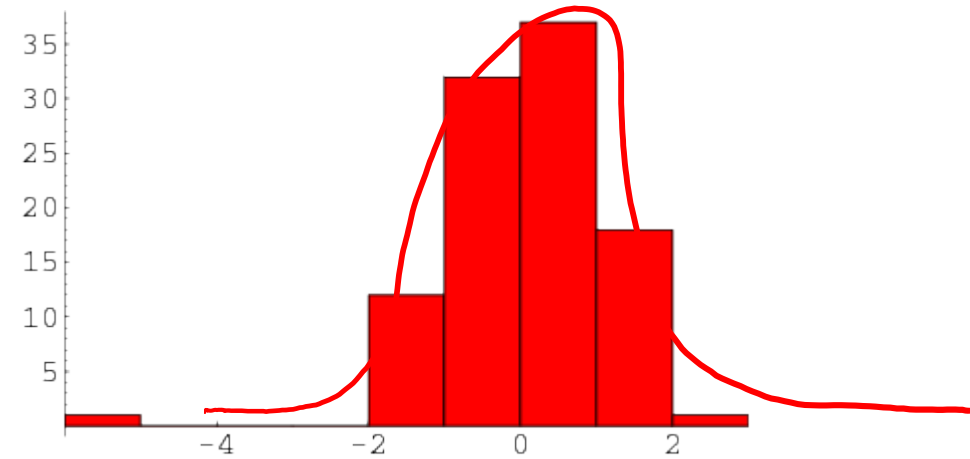
Model-Based Outlier Detection

- 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:

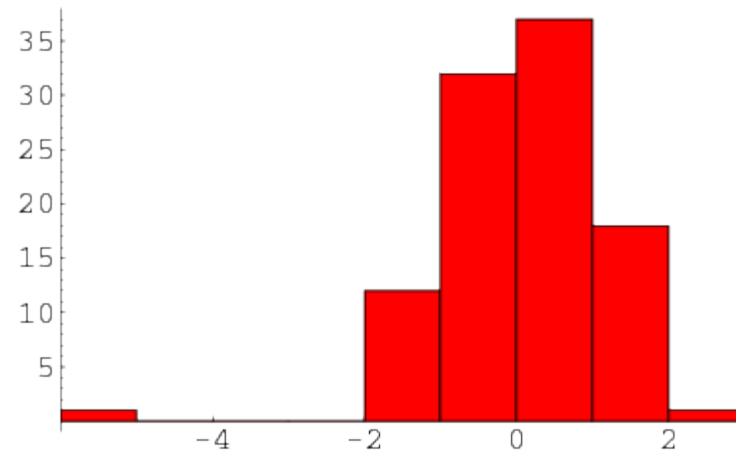
$$Z_i = \frac{x_i - \mu}{\sigma} \quad \text{where } \mu = \frac{1}{n} \sum_{i=1}^n x_i \quad \text{and} \quad \sigma = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2}$$

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

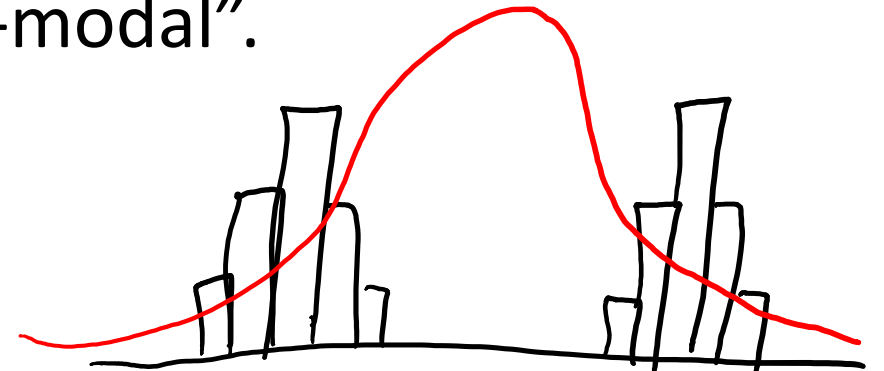


Problems with Z-Score

- 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.



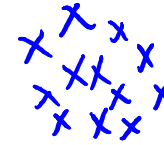
Global vs. Local Outliers

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



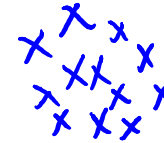
Global vs. Local Outliers

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



Global vs. Local Outliers

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

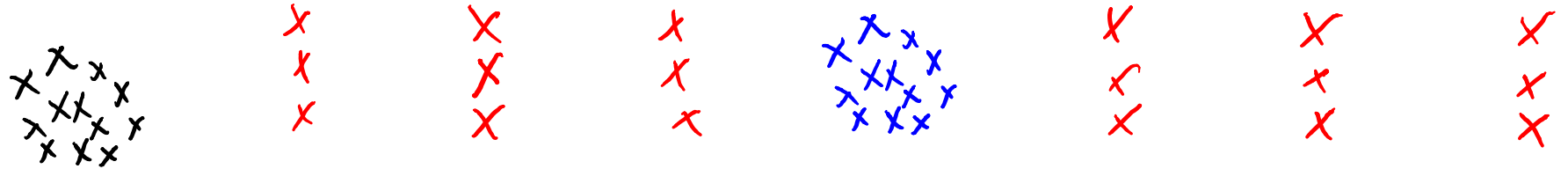
- Is the **red point** an outlier? What if add the **blue points**?



- 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 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

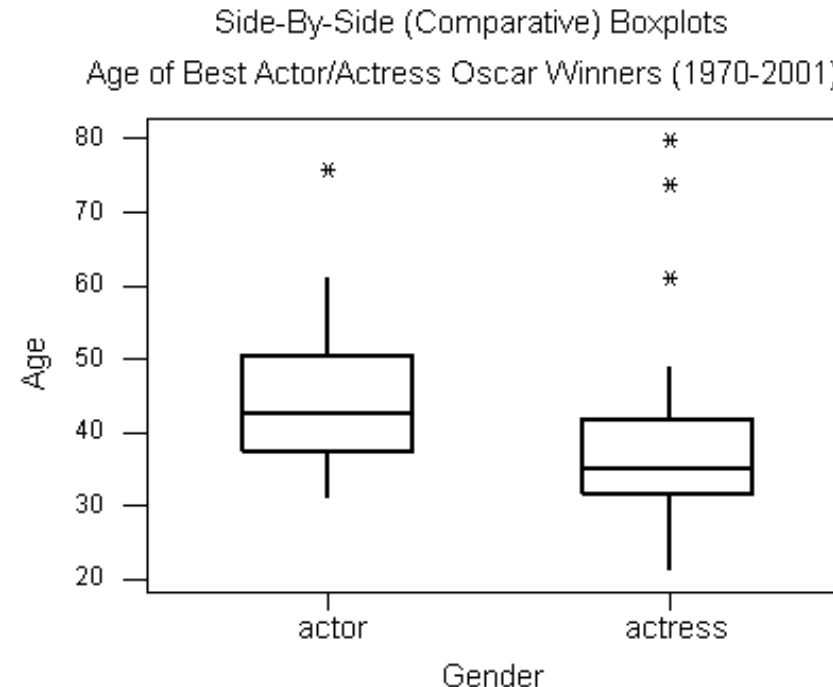
- Graphical approach to outlier detection:

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.



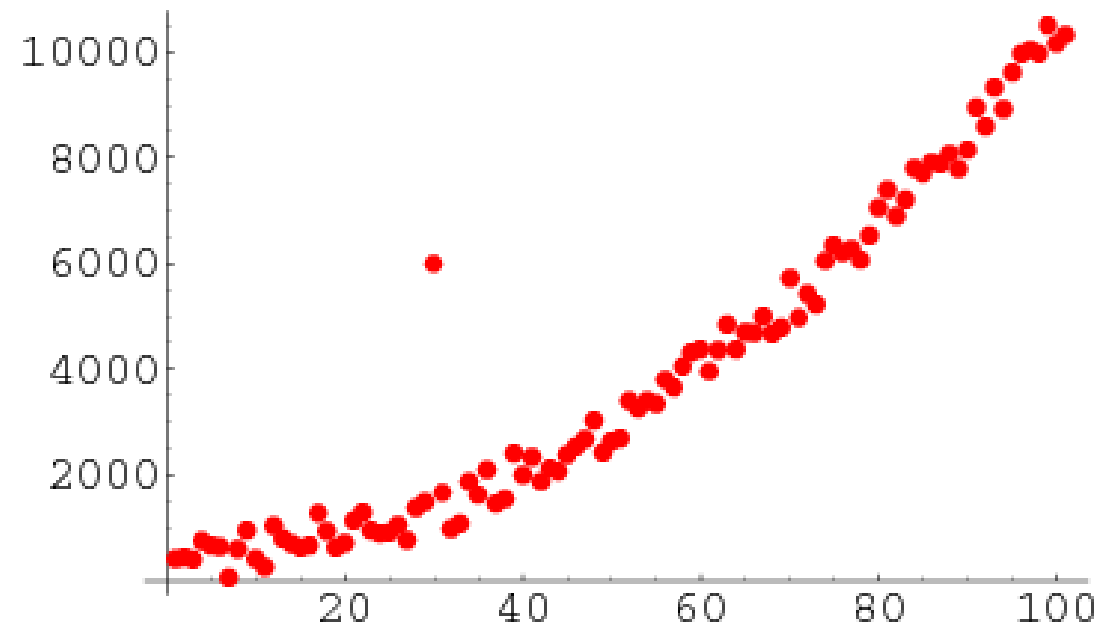
Graphical Outlier Detection

- 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.



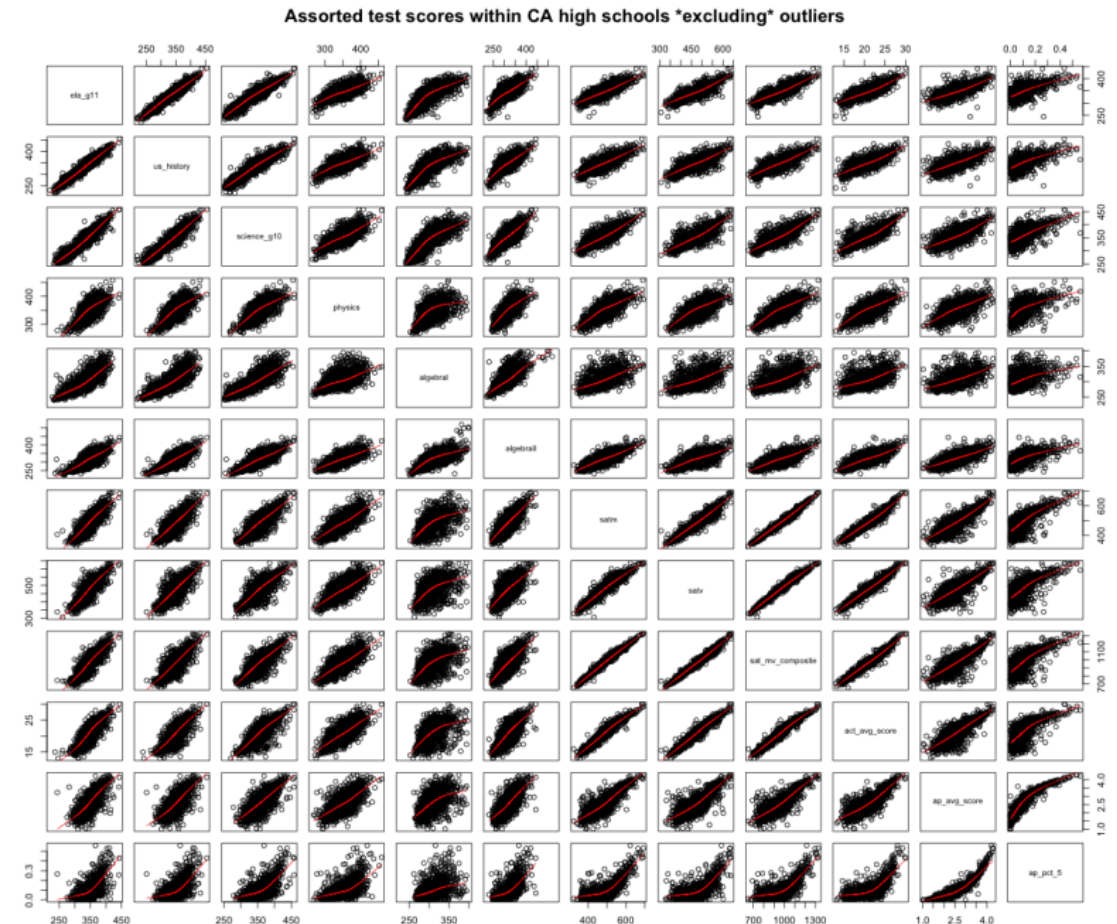
Graphical Outlier Detection

- 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.



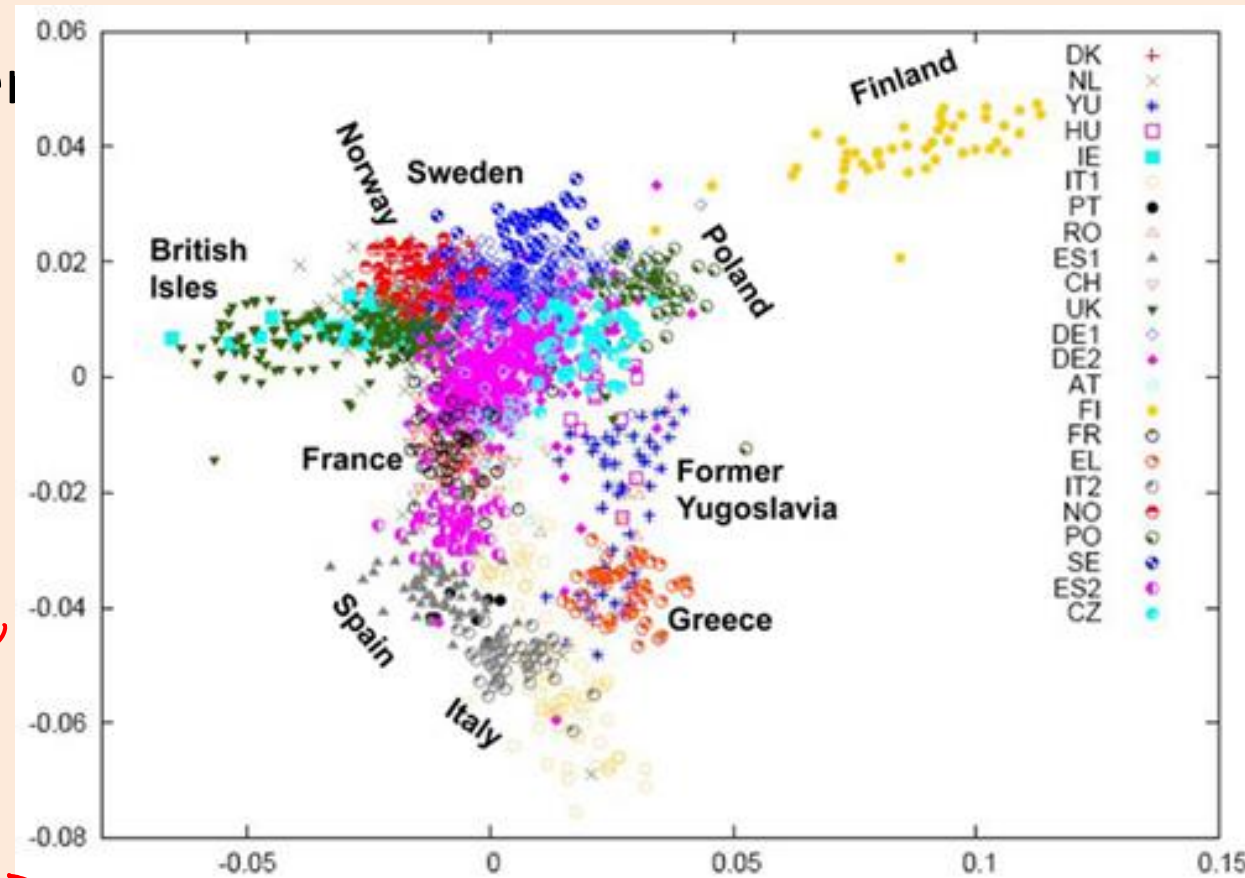
Graphical Outlier Detection

- Graphical approach to outlier detection:

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

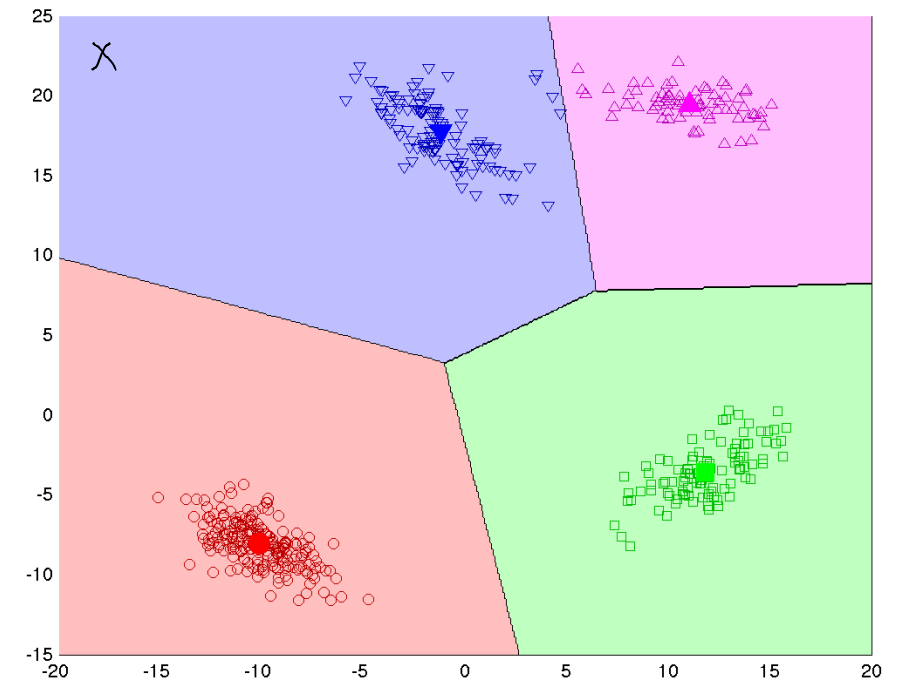
- Examples:

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.



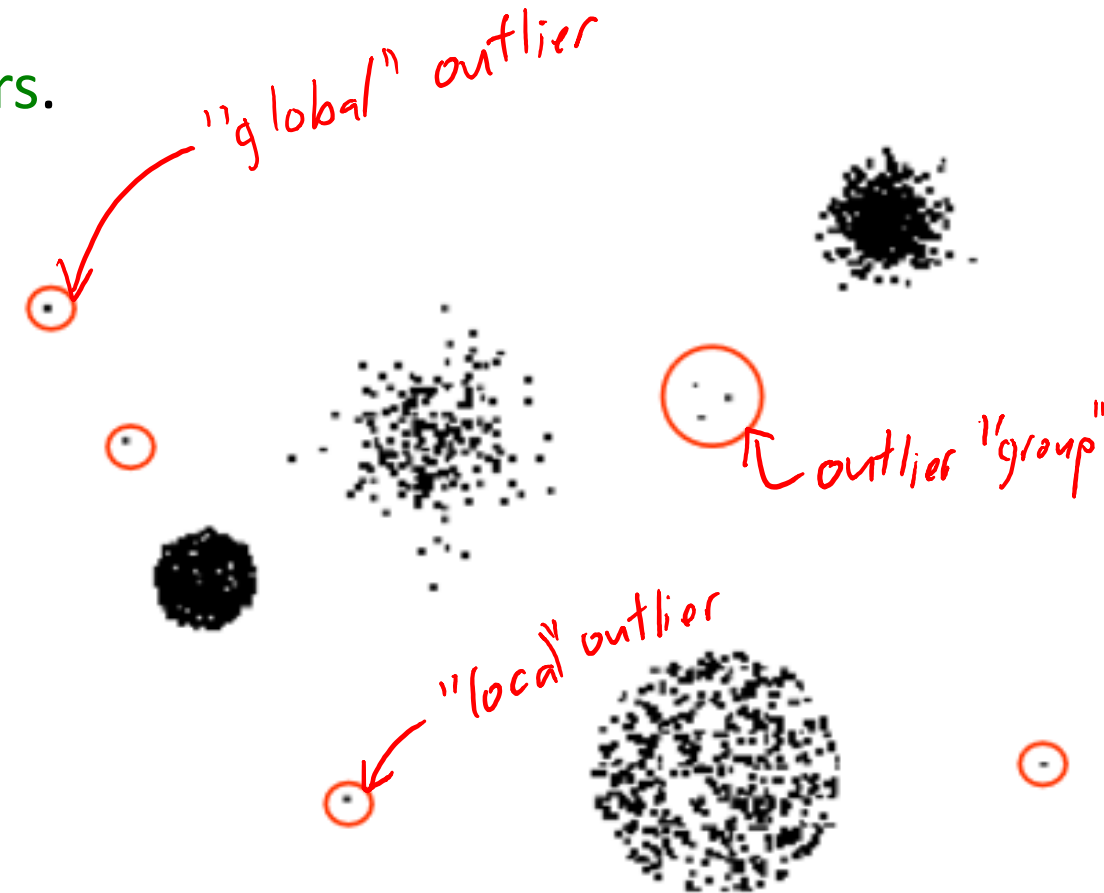
Cluster-Based Outlier Detection

- 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.



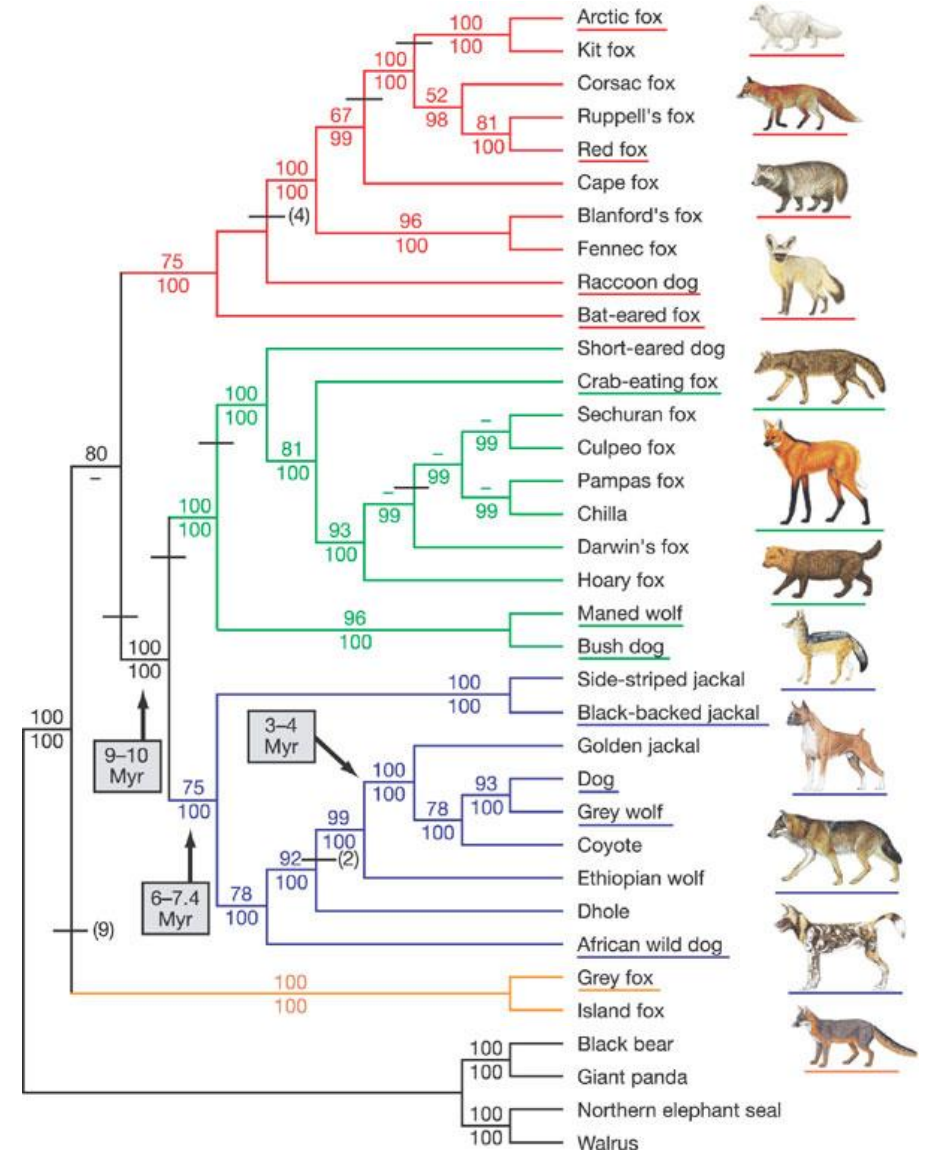
Cluster-Based Outlier Detection

- 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.

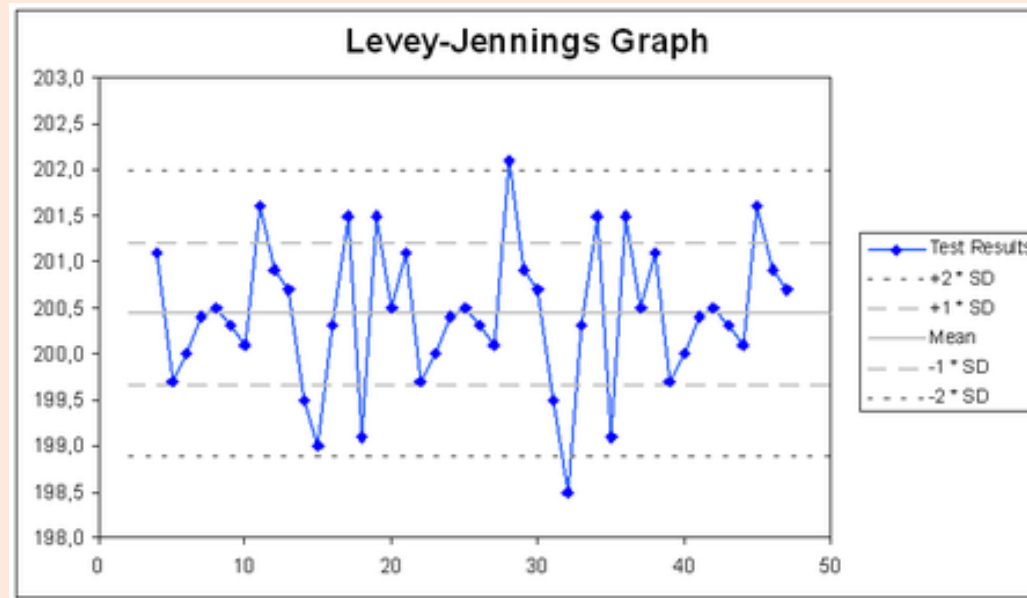


Summary

- Hierarchical clustering: more informative than fixed clustering.
- Agglomerative clustering: sequentially merges clusters.
- Outlier detection is task of finding unusually different object.
 - A concept that is very difficult to define.
 - Model-based find unlikely objects given a model of the data.
 - Graphical methods plot data and use human to find outliers.
 - Cluster-based methods check whether objects belong to clusters.
- Next time: “customers who bought this item also bought”.

“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".

Distances between Clusters

- Other choices of this distance between two clusters:
 - “Single-link”: minimum distance between points in clusters.
 - “Average-link”: average distance between points in clusters.
 - “Complete-link”: maximum distance between points in clusters.
 - Ward’s method: minimize within-cluster variance.
 - “Centroid-link”: distance between a representative point in the cluster.
 - Useful for distance measures on non-Euclidean spaces (like Jaccard similarity).
 - Centroid often defined as point in cluster minimizing average distance to other points.

Cost of Agglomerative Clustering

- One step of agglomerative clustering costs $O(n^2d)$:
 - We need to do $O(d)$ distance calculation between up to $O(n^2)$ points.
 - This is assuming the standard distance functions.
- We do at most $O(n)$ steps:
 - Starting with 'n' clusters and merging 2 clusters on each step, after $O(n)$ steps we'll only have 1 cluster left.
- This gives a total cost of $O(n^3d)$.
- This can be reduced to $O(n^2d \log n)$ with a priority queue:
 - Store distances in a sorted order, only update the distances that change.
- For single- and complete-linkage, you can get it down to $O(n^2d)$.
 - “SLINK” and “CLINK” algorithms.

Bonus Slide: Divisive (Top-Down) Clustering

- Start with all objects in one cluster, then start dividing.
- E.g., run k-means on a cluster, then run again on resulting clusters.
 - A clustering analogue of decision tree learning.

