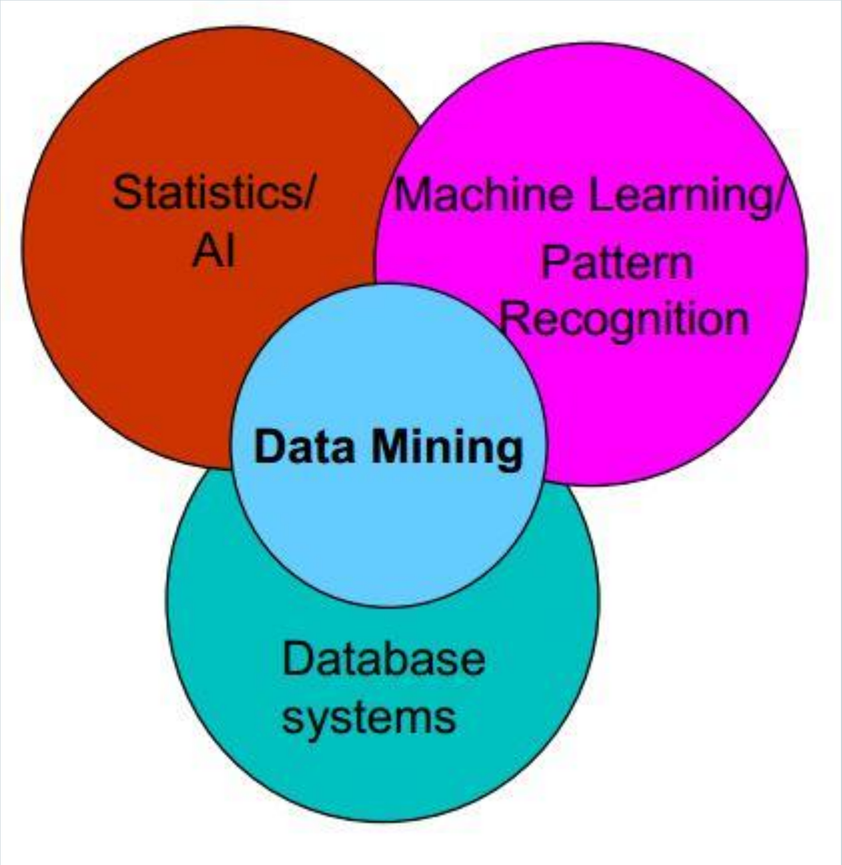
Clustering & Intro to Algorithms

SU 5050
LECTURE 4
JESSICA L. MCCARTY, PH.D.

Core Disciplines of Data Mining



Origins of Data Mining

- Draws ideas from machine learning/AI, pattern recognition, statistics, and database systems
- Traditional techniques may be unsuitable due to:

1. Enormity of data;
2. High dimensionality of data;
3. Heterogeneous, distributed nature of data

Data Mining Tasks

- Prediction Methods
  - Use some variables to predict unknown or future values of other variables.

- Description Methods
  - Find human-interpretable patterns that describe the data.

Data Mining Tasks

- Classification *[Predictive]*
- Clustering *[Predictive]*
- Association Rule Discovery *[Descriptive]*
- Sequential Pattern Discovery *[Descriptive]*
- Natural Language Processing *[Descriptive]*
- Regression *[Predictive]*
- Deviation Detection *[Predictive]*

Data Mining Tasks

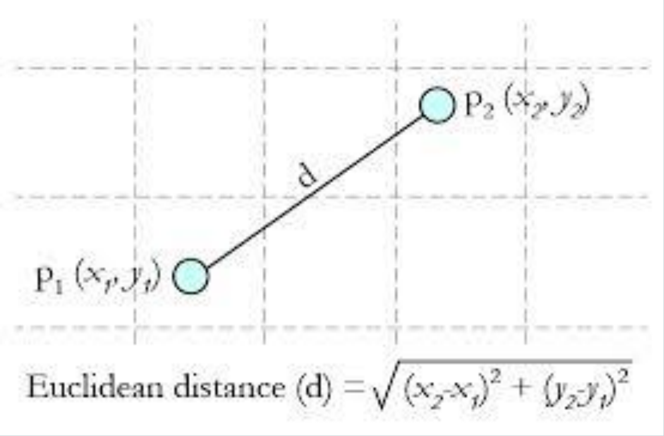
- Classification *[Predictive]*
- Clustering *[Predictive]*
- Association Rule Discovery *[Descriptive]*
- Sequential Pattern Discovery *[Descriptive]*
- Natural Language Processing *[Descriptive]*
- Regression *[Predictive]*
- Deviation Detection *[Predictive]*

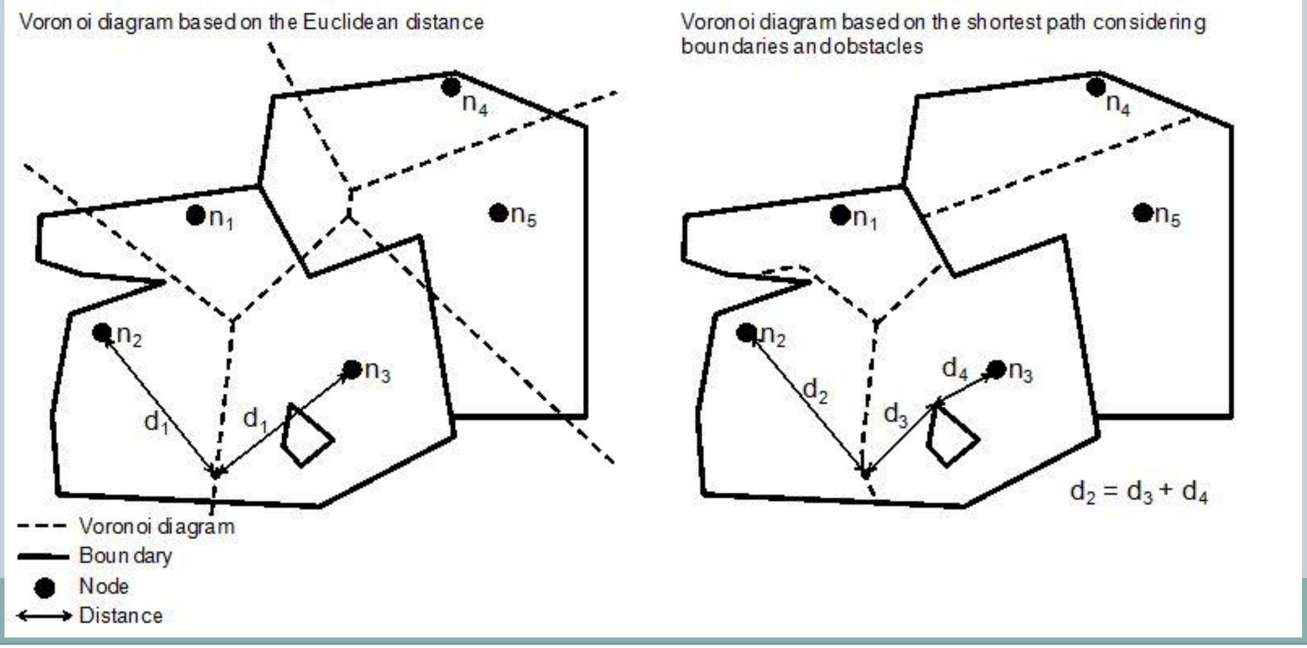
Clustering Definition

- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that
  1. Data points in one cluster are more similar to one another.
  2. Data points in separate clusters are less similar to one another.

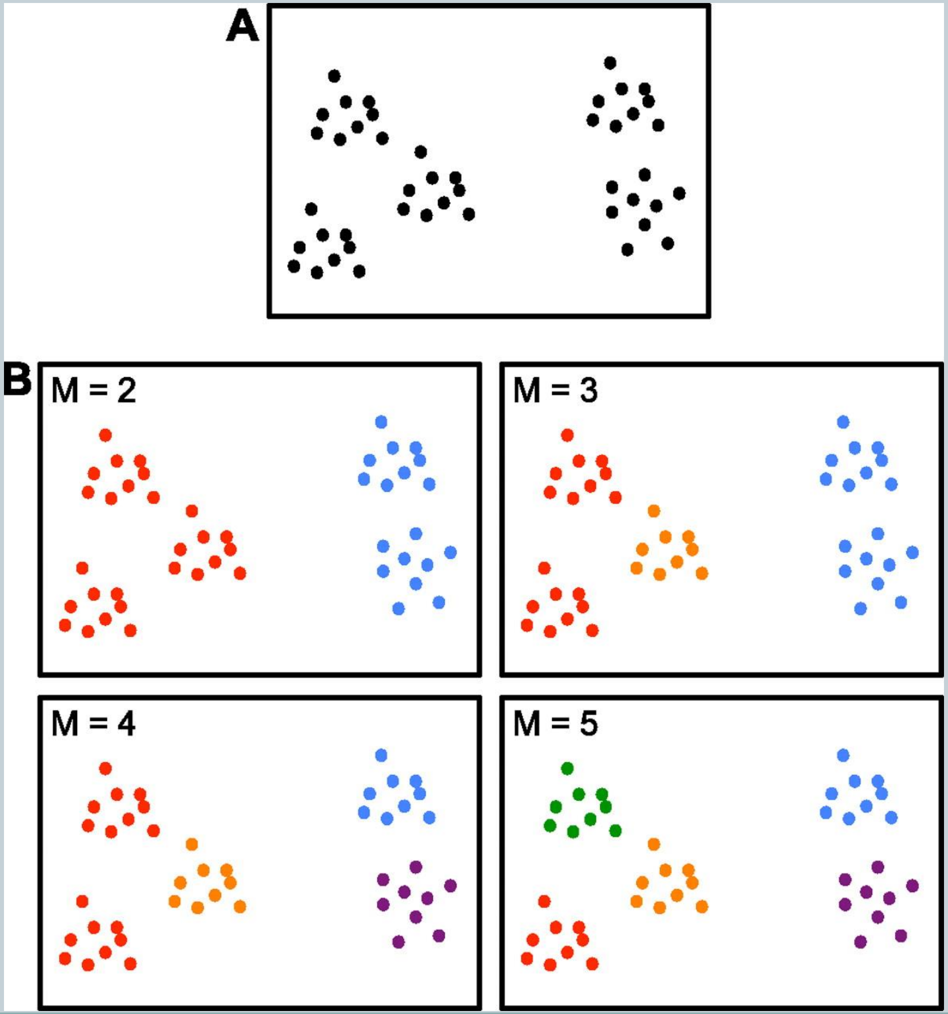
Clustering Definition

- Similarity Measures:
  1. Euclidean Distance if attributes are continuous.
  2. Other Problem-specific measures.

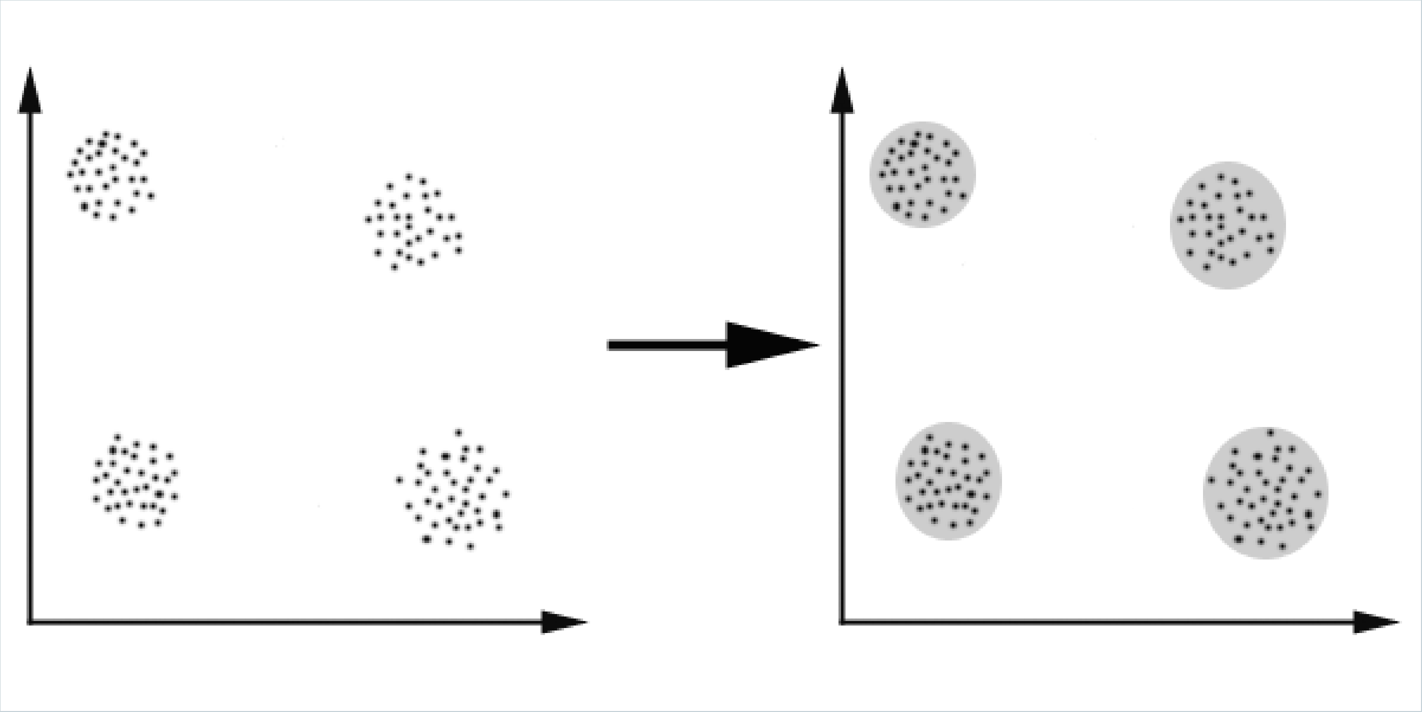




Illustrating Clustering

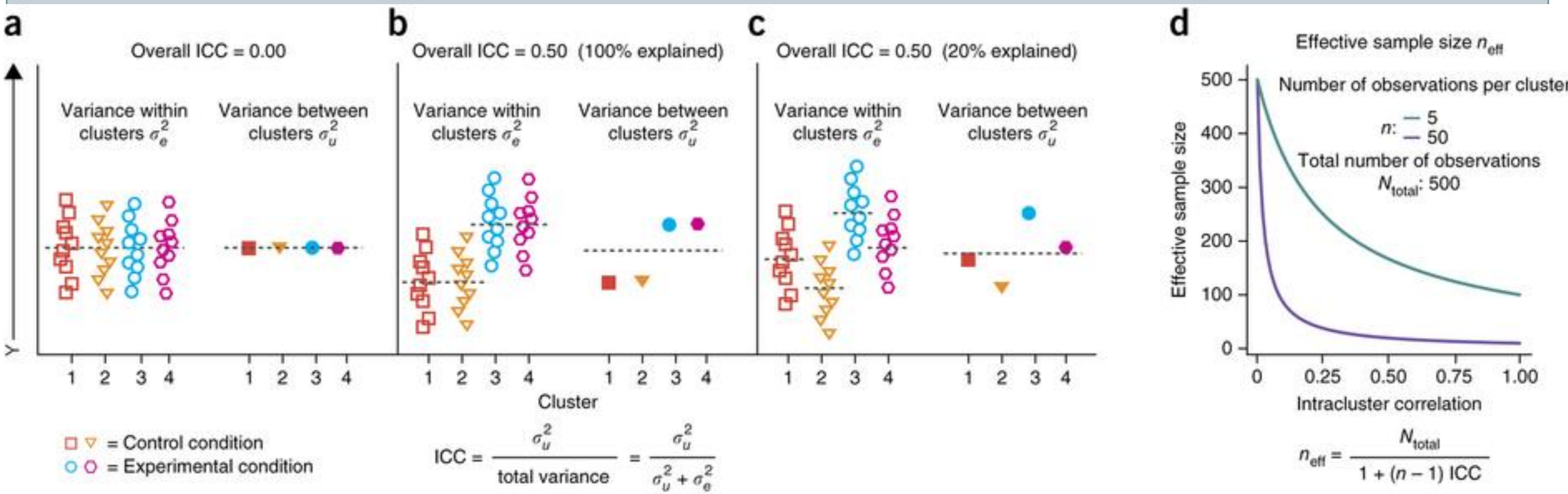


Illustrating Clustering



Illustrating Clustering

ICC = intracluster correlations



Emmeke Aarts, Matthijs Verhage, Jesse V Veenvliet, Conor V Dolan & Sophie van der Sluis. 2014. **A solution to dependency: using multilevel analysis to accommodate nested data.** Nature Neuroscience 17, 491–496 . doi:10.1038/nn.3648.

Clustering

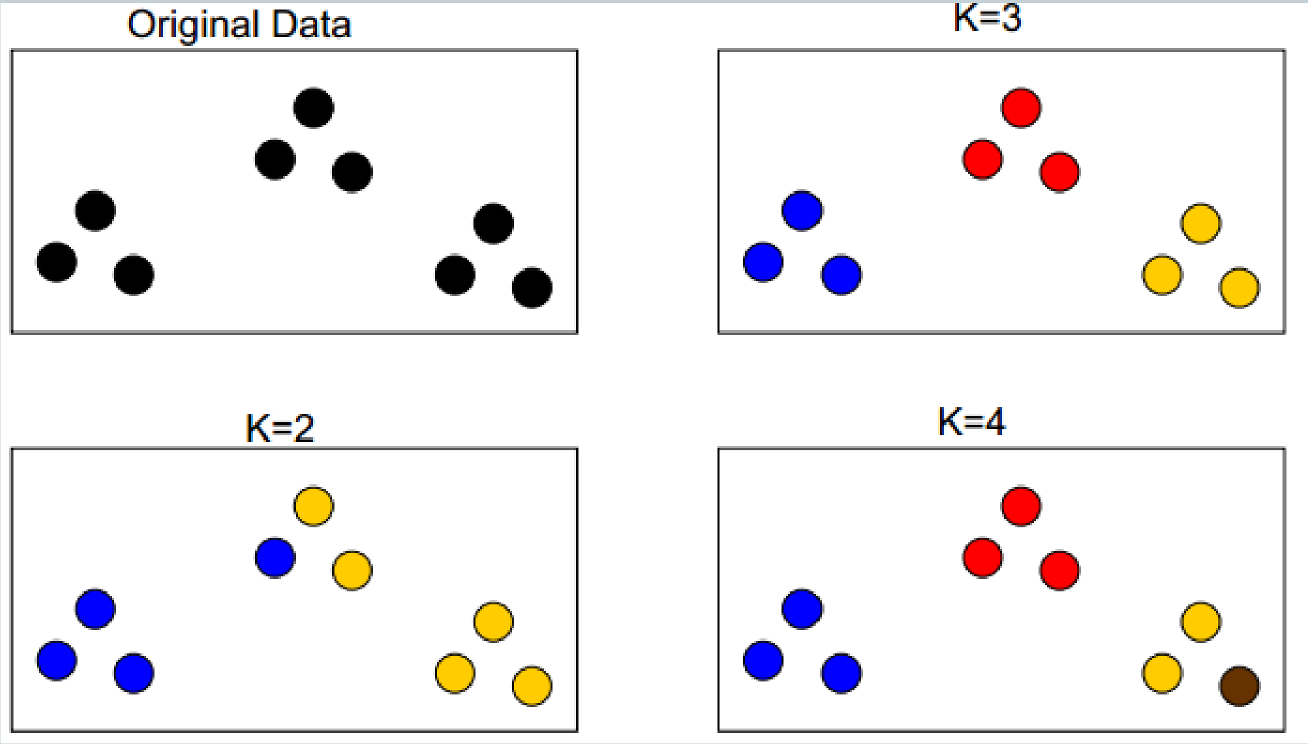
- Not **classification**, which identifies the group an item belongs to.

- **Example**
- *Classification* would determine the major of one student.

- *Clustering* puts the students into groups, but does not really know their true major.

Clustering is Ambiguous

- Divide a 2D dataset into K clusters



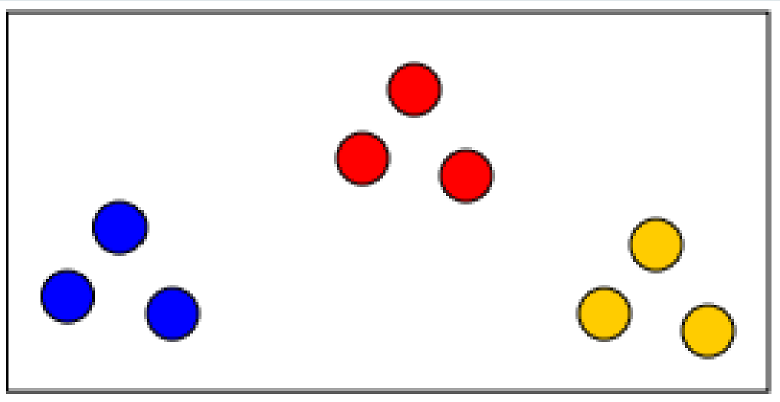
Distance Between Points

- Suppose each data point has length N.

$x=(x_1,x_2,\dots,x_N)$ $y=(y_1,y_2,\dots,y_N)$

- We need a measure of the distance between 2 points.
- The distance d(x,y) tells us how similar the objects x & y are.
  - d(x,y) small = x & y very similar
  - d(x,y) large = x & y not similar

- So goal of clustering is to put nearby points (small distance) into same group.



Euclidean Distance

- In space, the distance between points is given by the Euclidean or L2 distance.

$$d(x,y)=\sqrt{\sum_{i=1}^{N}(x_i-y_i)^2}=\sqrt{(x_1-y_1)^2+(x_2-y_2)^2+\cdots+(x_N-y_N)^2}$$

- <u>Ex</u> x=(1,2,3,4) y=(5,6,7,8)

$$d(x,y)=\sqrt{(1-5)^2+(2-6)^2+(3-7)^2+(4-8)^2}=\sqrt{4^2+4^2+4^2+4^2}=\sqrt{64}=8$$

- We often ignore the square root in our calculations.
- The Euclidean distance makes sense visually, but it is not always the best distance measure for general data.

Distance Matrix

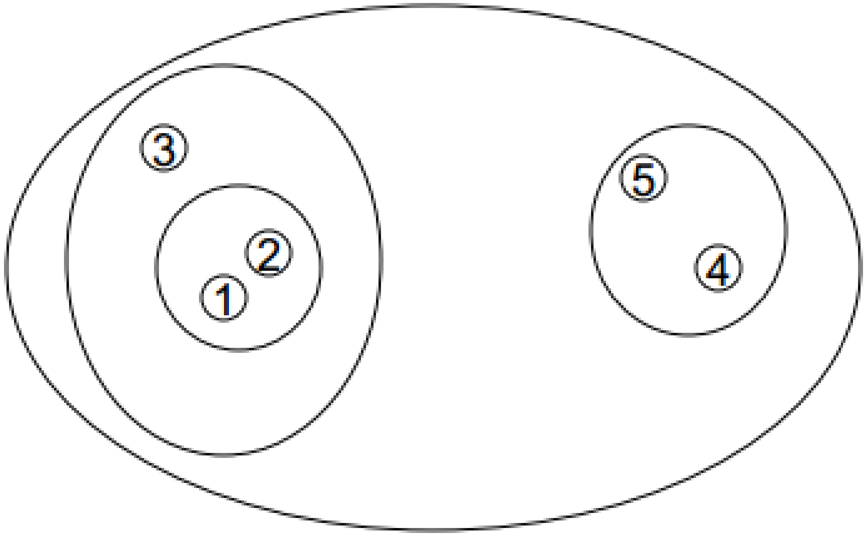
- After computing the pairwise distance between all points, it is often helpful to put the values in a 2D distance matrix.

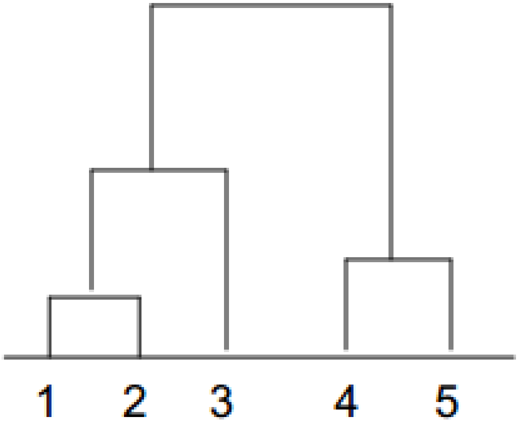
$$d = \begin{array}{cc} \begin{bmatrix} 0 & 8 & 14 \\ 8 & 0 & 5 \\ 14 & 5 & 0 \end{bmatrix} & \begin{array}{c} 1 \\ 2 \\ 3 \end{array} \\ \begin{array}{ccc} 1 & \quad 2 & \quad 3 \end{array} & \end{array}$$

- This is a symmetric matrix with zeros on the diagonal.

Hierarchical Clustering

- *Hierarchical clustering* is a set of nested sets. We grow the partition by merging 2 clusters at a time.
- The *dendrogram* is a diagram that displays the partition. We grow the dendrogram upwards in order which clusters were merged. To make K clusters, we cut off the top of the dendrogram off to form K sets.

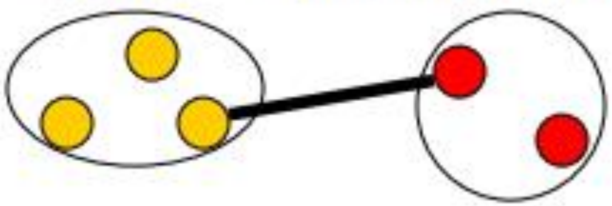




Distance Between Clusters

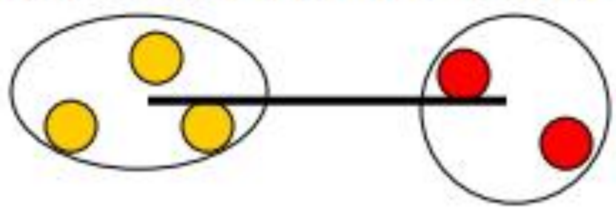
- To decide which 2 clusters to merge, we need a notion of distance between clusters.

<u>Min Link</u>: look at distance between 2 closest points



<u>Max Link</u>: look at distance between 2 farthest points

<u>Group Average</u>: look at distance between the cluster centroids



Examples

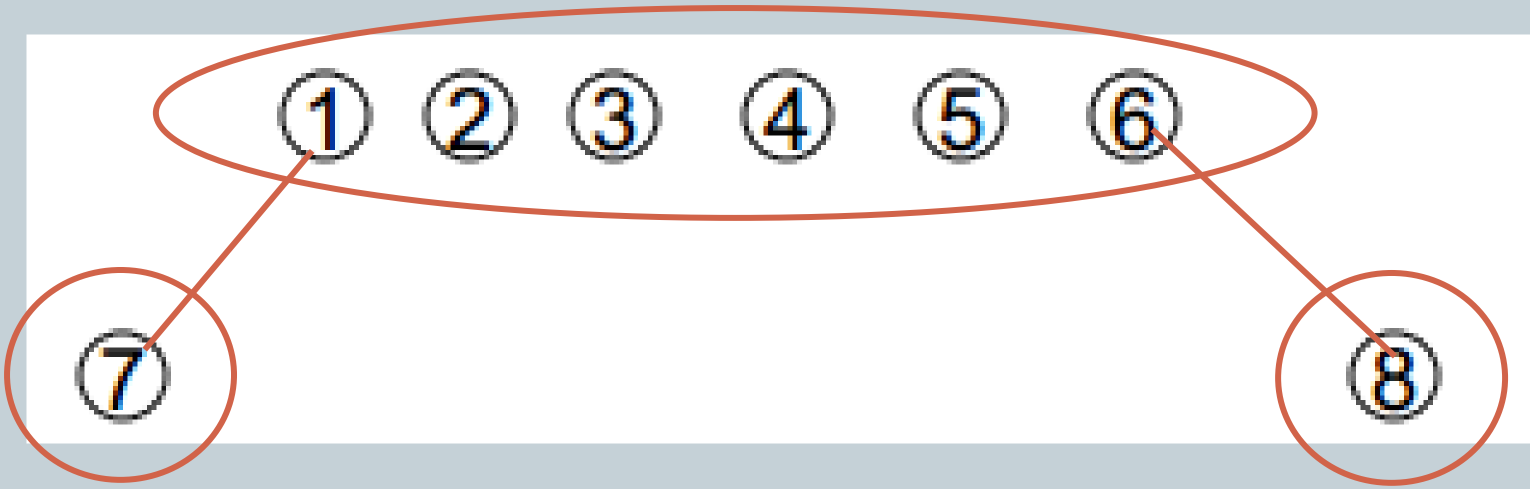
- How would the different methods cluster the data below into K=2 clusters?

① ② ③ ④ ⑤ ⑥

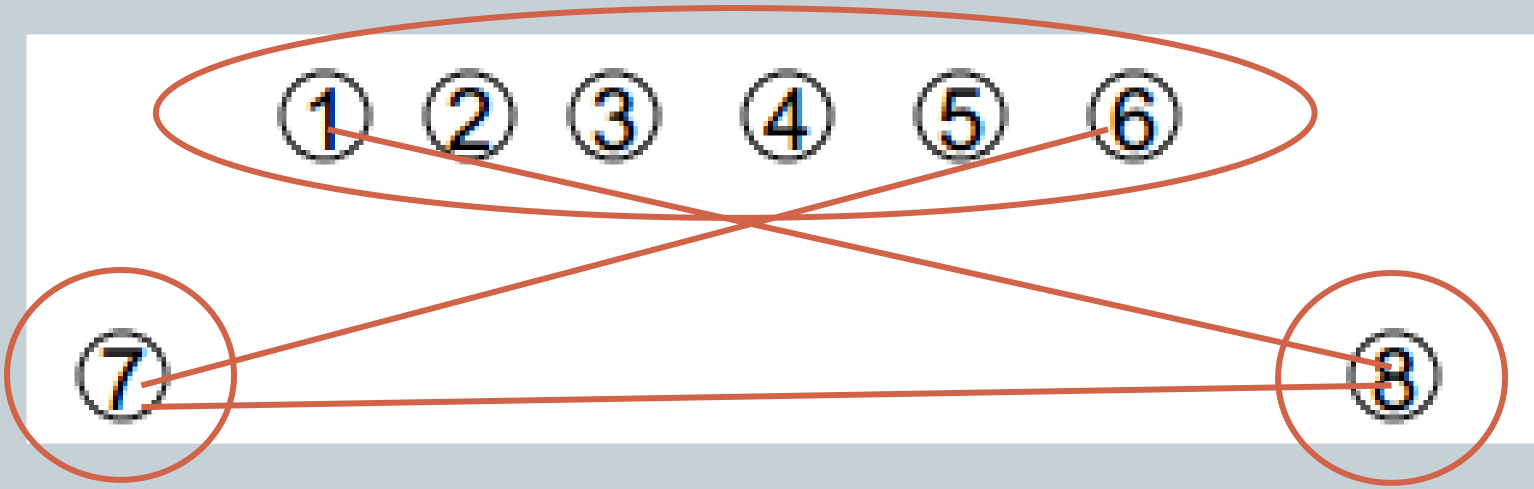
⑦ ⑧

- Min Link prefers contiguous clusters.
- Max Link and Group Average prefer globular clusters.

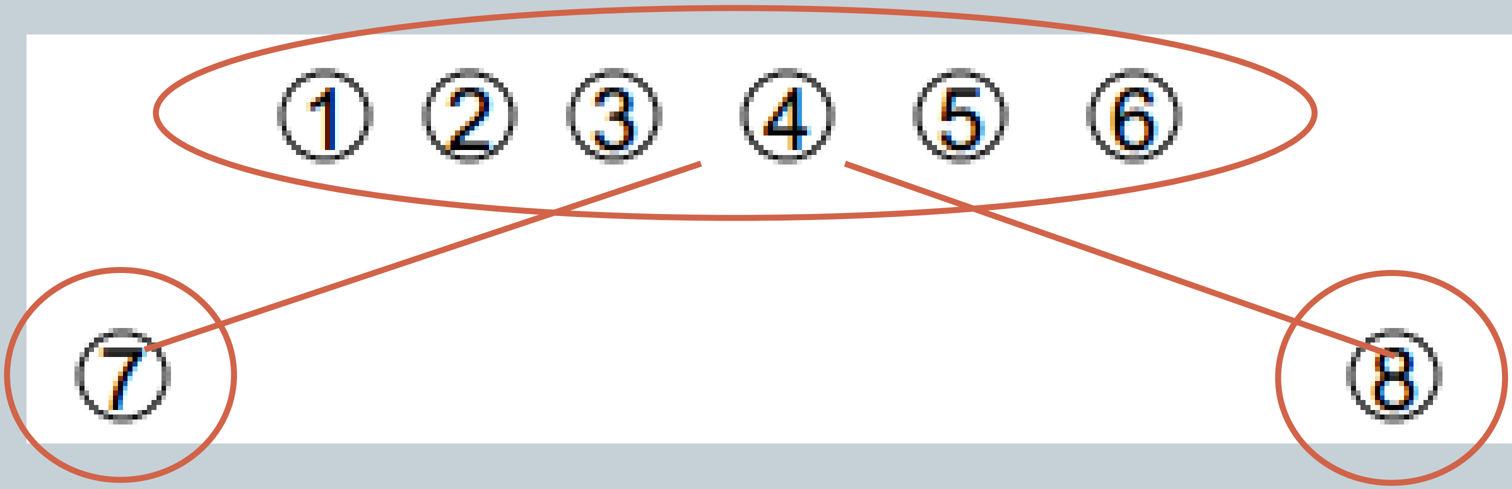
Cluster: Min Link



Cluster: Max Link



Group Average



K-Means Clustering

- Very popular partitional clustering algorithm.
- The algorithm starts from a random guess, so results may vary with each run.
- The idea is to calculate the centroid of each cluster and then put each point in a cluster with closest centroid.
- Then recalculate the centroids of the new clusters and continue.

K-Means Clustering Steps

1. Select K random data points as the initial centroids

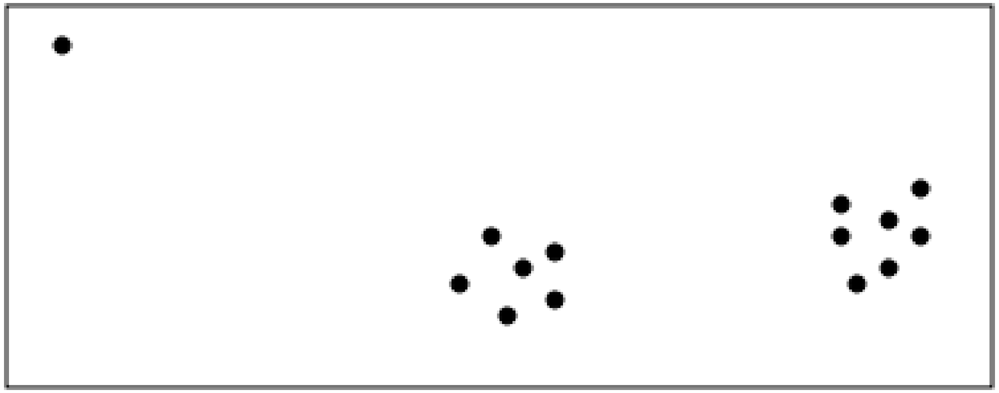
2. Assign each point with the nearest centroid

3. Recalculate the centroid of each cluster until centroids don't change.

Repeat

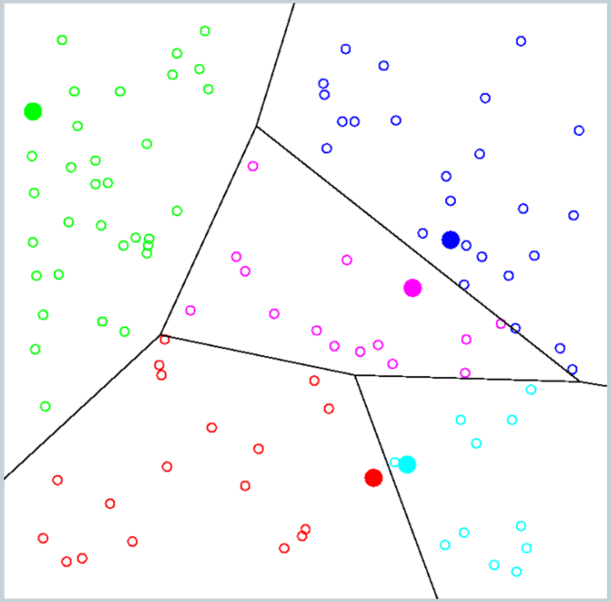
K-Means Example

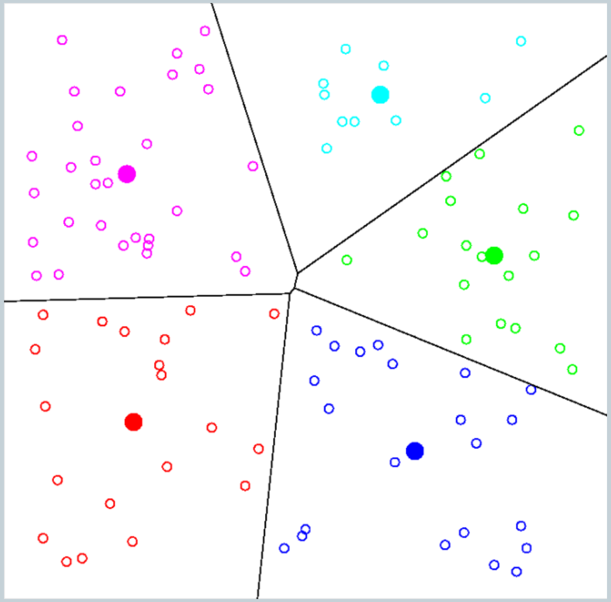
- How would K-means perform on the dataset below? (Answers may vary.)



- K-means also prefers globular clusters.
- K-means prefers to put outliers into their own cluster.
- K-means has problems with clusters of different sizes.

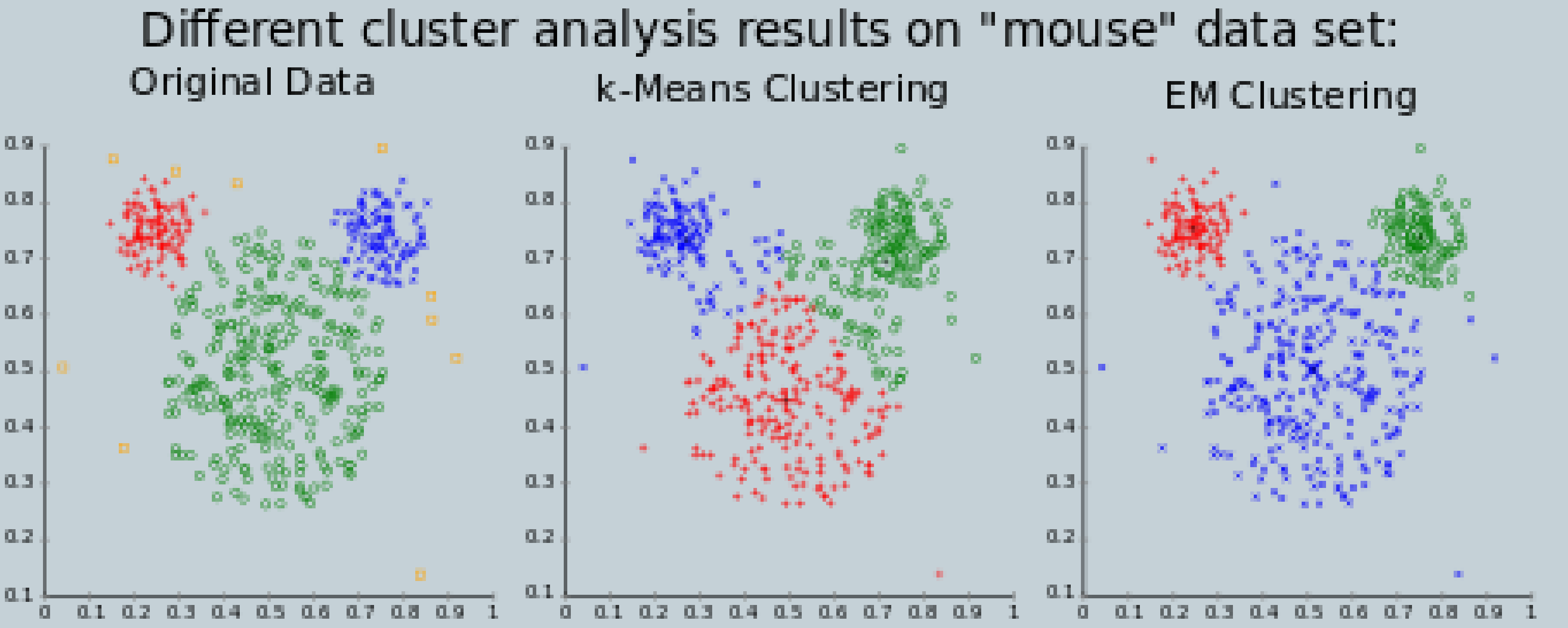
K-Means Example





K-Means Example

EM= Expectation-Maximization

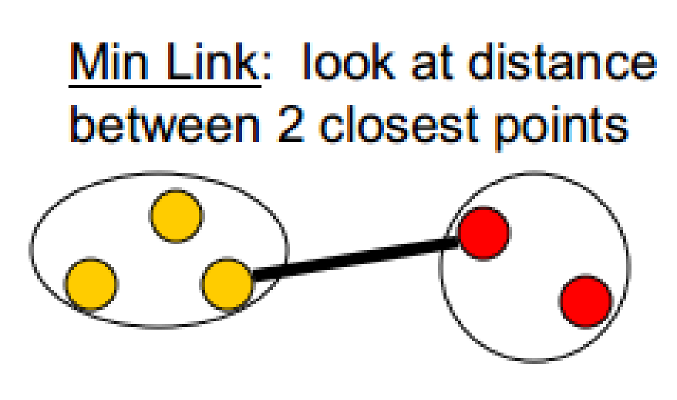


Clustering Validity

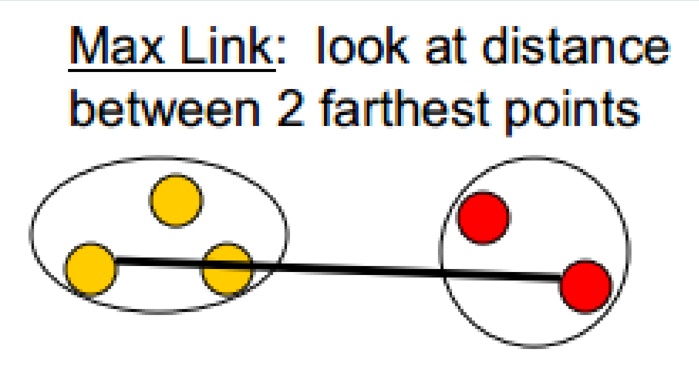
- How do we tell if our clustering method did a good job?
- We can visualize the distance matrix (*black=low, white=high*).
- If we re-order the matrix based on the clusters, ideally we will see black squares on the diagonal.

Comparing Methods

- Min Link
  - Prefers contiguous clusters
  - Can handle non-elliptical shapes
  - Sensitive to noise and outliers

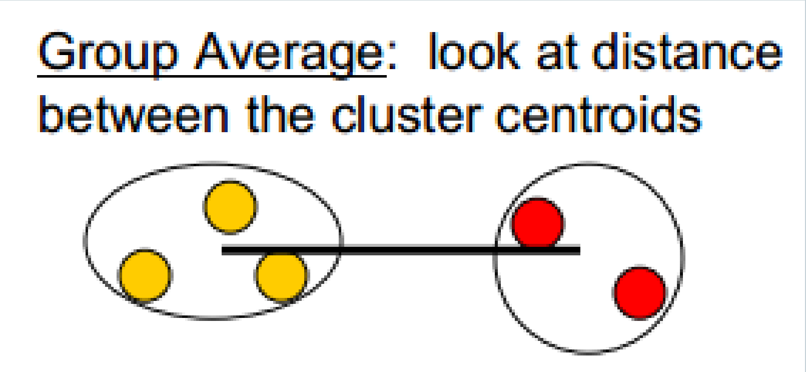


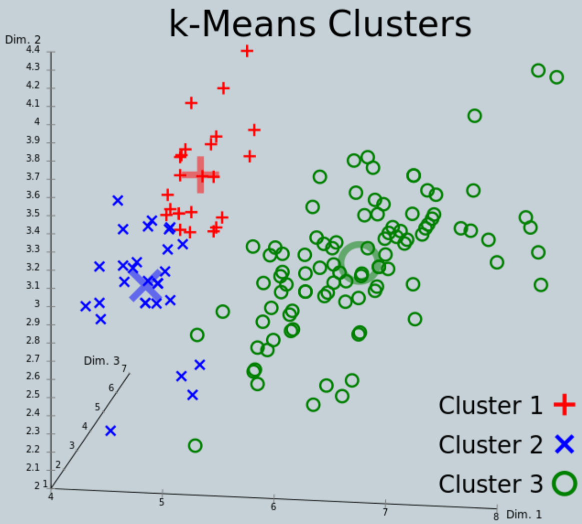
- Max Link
  - Prefers globular clusters
  - Less sensitive to noise and outliers
  - Tends to break up large clusters



Comparing Methods

- Group Average
  - Somewhere in between Min and Max Link
- K-means
  - Prefers globular clusters, equal sized clusters
  - Based on random initialization, so can give different answers each time





Applications of Clustering

- Document analysis
  - Group similar website together to make suggestions
- Biology
  - Group similar gene sequences
- Market research
- Crime Analysis
- Sociology
- Image processing

Clustering: Application 1

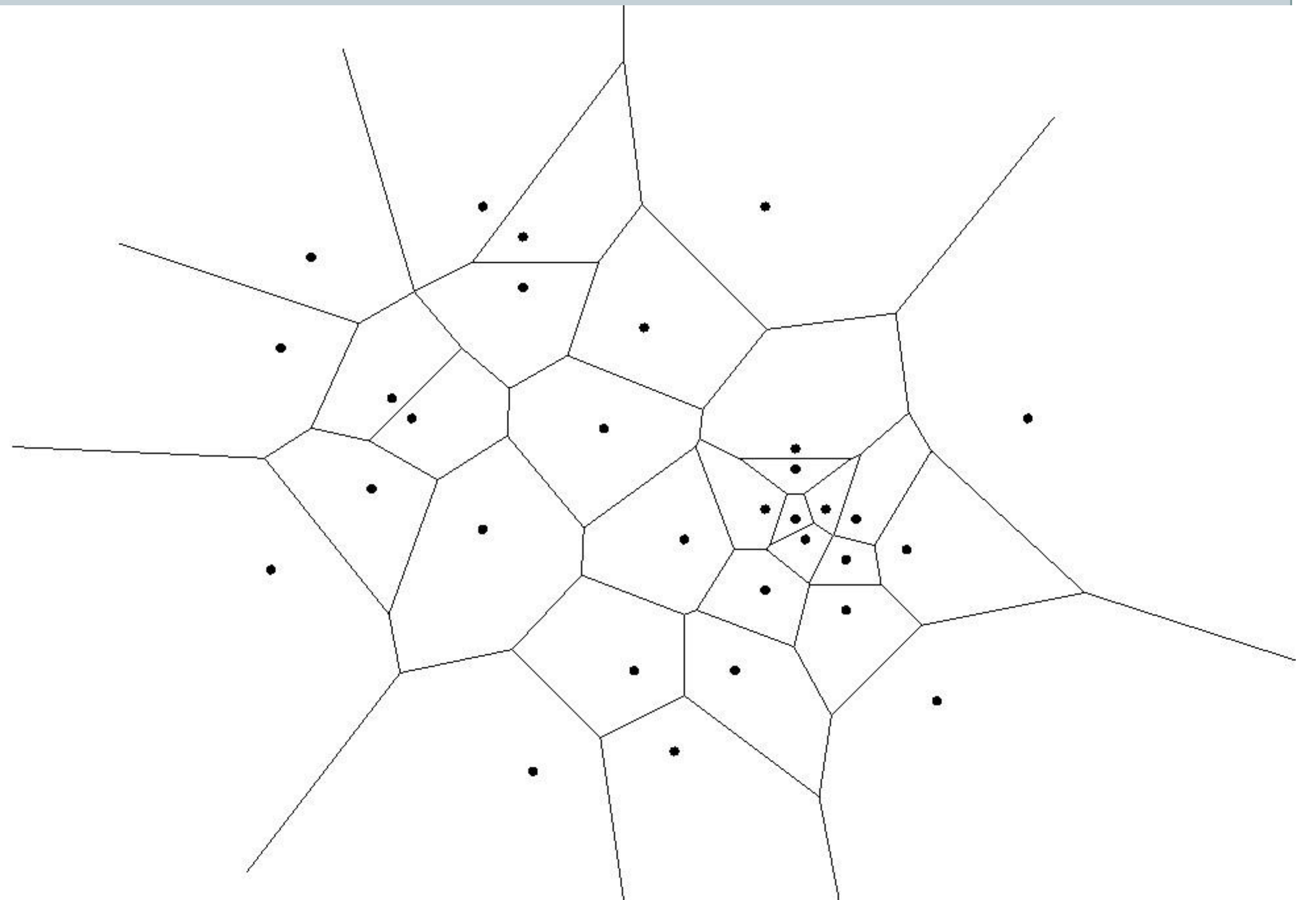
- Market Segmentation:
  - Goal: subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix.
  - Approach:
    - Collect different attributes of customers based on their geographical and lifestyle related information.
    - Find clusters of similar customers.
    - Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters.

Clustering: Application 2

- Document Clustering:
  - Goal: To find groups of documents that are similar to each other based on the important terms appearing in them.
  - Approach: To identify frequently occurring terms in each document. Form a similarity measure based on the frequencies of different terms. Use it to cluster.
  - Gain: Information Retrieval can utilize the clusters to relate a new document or search term to clustered documents.

Voronoi diagrams and clustering

- Stores proximity among points in a set



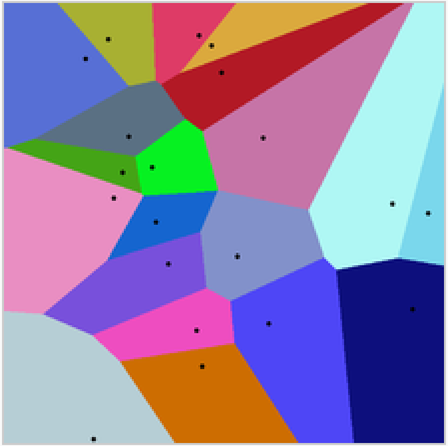
Voronoi diagrams and clustering

- Single-link clustering attempts to maximize the distance between any two points in different sets

Voronoi diagram

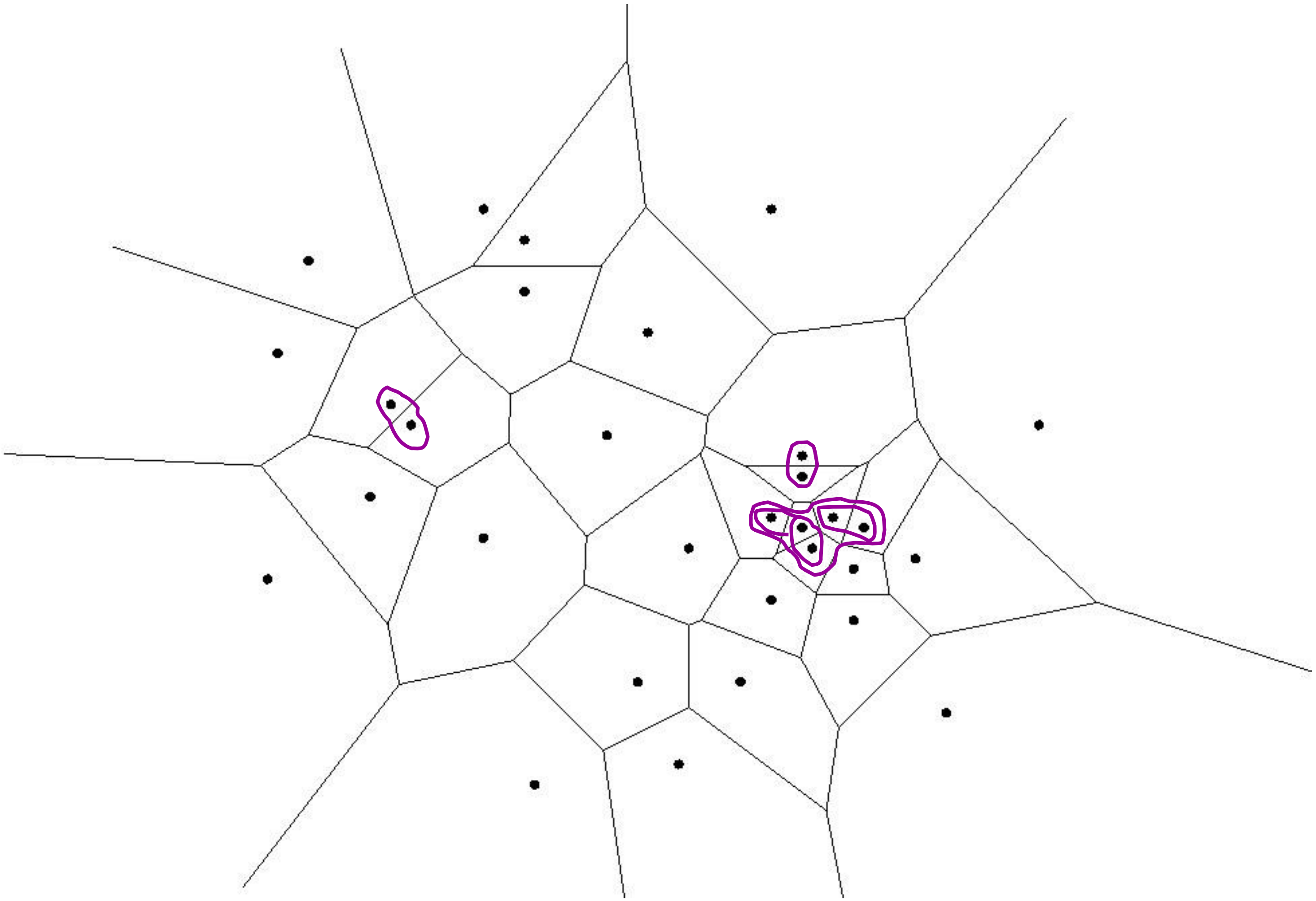
From Wikipedia, the free encyclopedia

In mathematics, a **Voronoi diagram** is a way of dividing space into a number of regions. A set of points (called seeds, sites, or generators) is specified beforehand and for each seed there will be a corresponding region consisting of all points closer to that seed than to any other. The regions are called Voronoi cells. It is dual to the Delaunay triangulation.

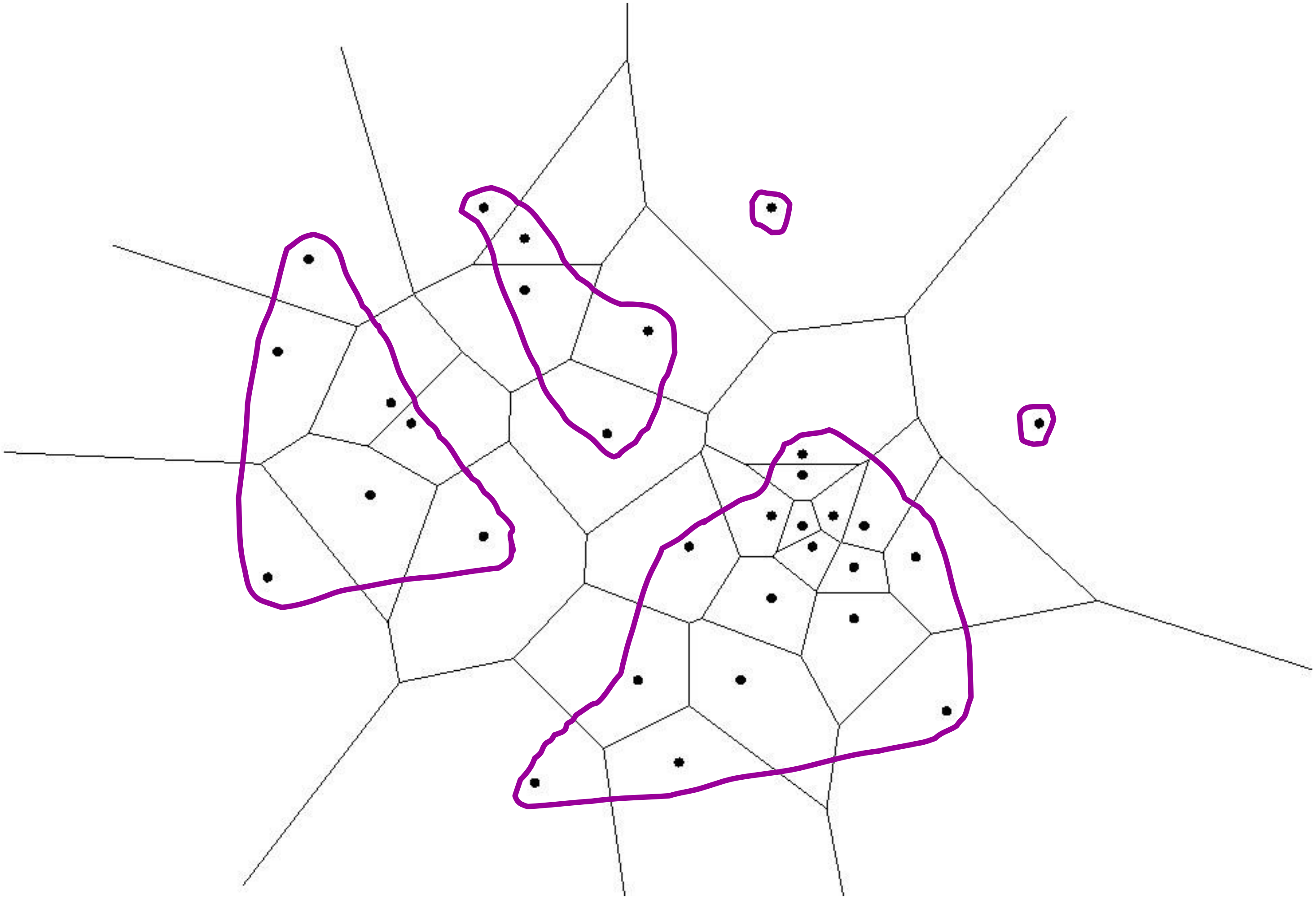

20 points and their Voronoi cells (larger version below).

It is named after Georgy Voronoy, and is also called a **Voronoi tessellation**, a **Voronoi decomposition**, a **Voronoi partition**, or a **Dirichlet tessellation** (after Peter Gustav Lejeune Dirichlet). Voronoi diagrams can be found in a large number of fields in science and technology, even in art, and they have found numerous practical and theoretical applications.[1][2]

Voronoi diagrams and clustering



Voronoi diagrams and clustering



Voronoi diagrams and clustering

- Algorithm (point set P; desired: k clusters):
  - Compute Voronoi diagram of P
  - Take all $O(n)$ neighbors and sort by distance
  - While #clusters > k do
    - Take nearest neighbor pair p and q
    - If they are in different clusters, then merge them and decrement #clusters (else, do nothing)

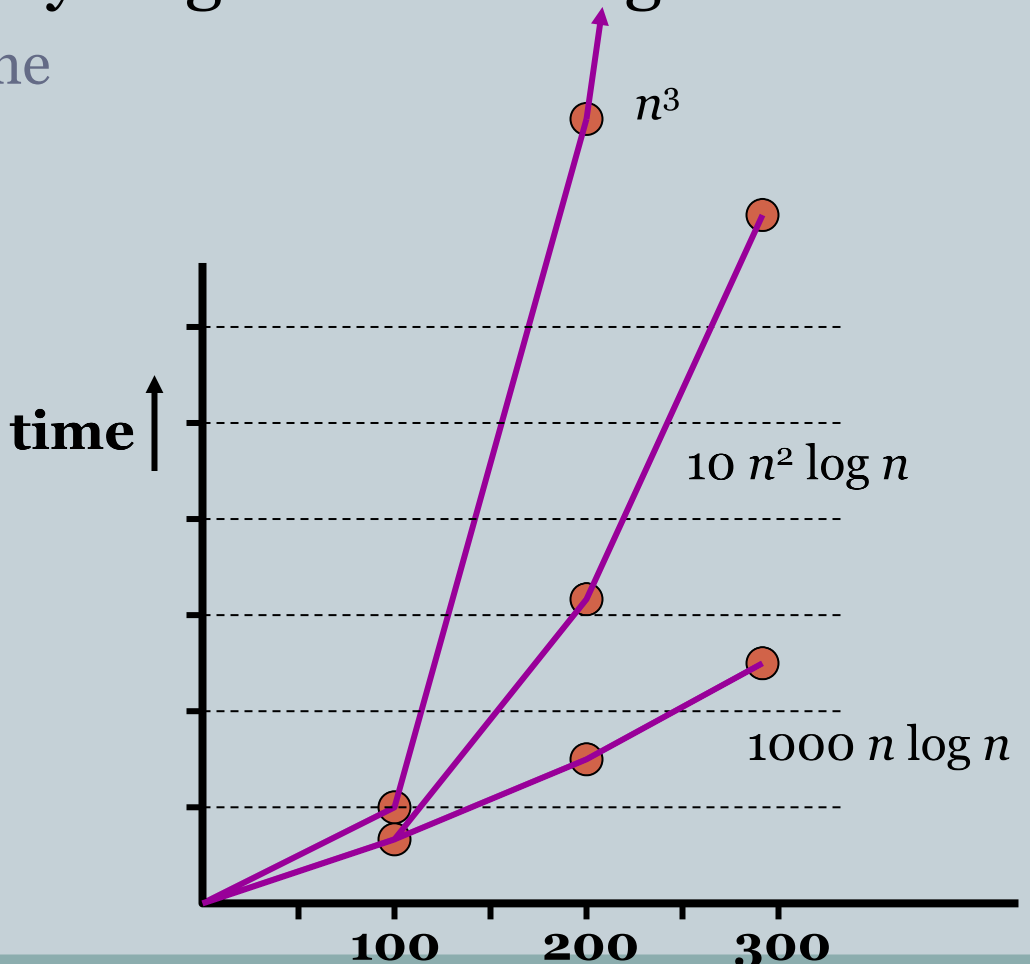
Voronoi diagrams and clustering

- Analysis; n points in P:
  - Compute Voronoi diagram: $O(n \log n)$ time
  - Sort by distance: $O(n \log n)$ time
  - While loop that merges clusters: $O(n \log n)$ time (using union-find structure)

- Total: $O(n \log n) + O(n \log n) + O(n \log n) = O(n \log n)$ time

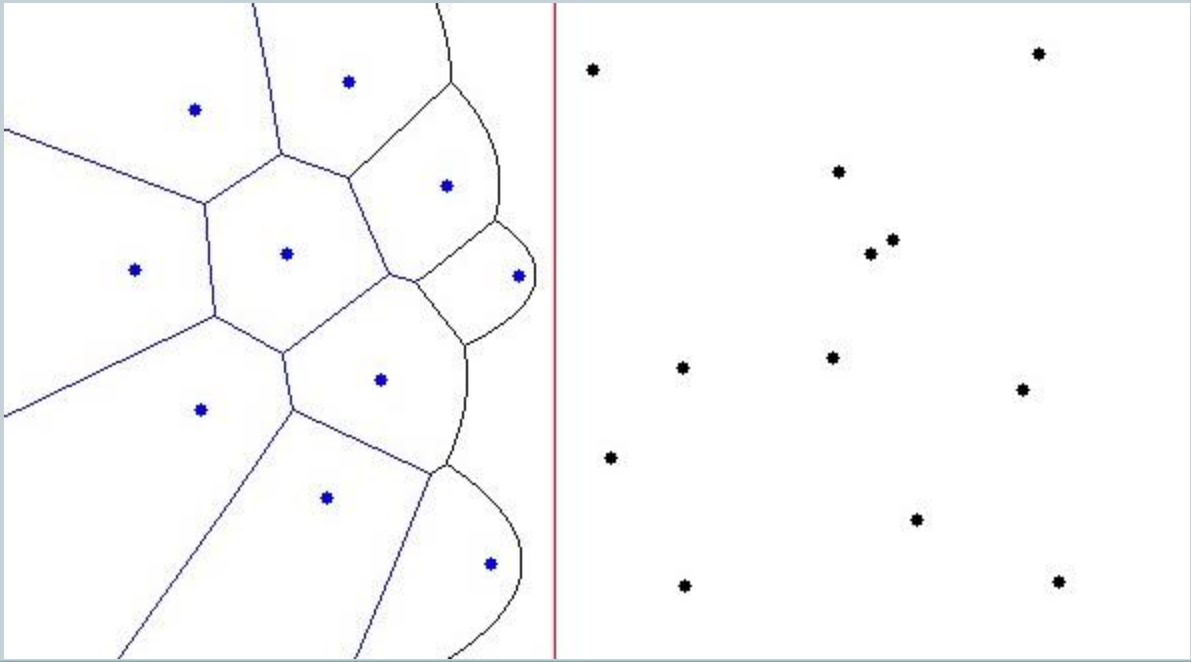
Voronoi diagrams and clustering

- What would an "easy" algorithm have given?
  - really easy: $O(n^3)$ time
  - slightly less easy: $O(n^2 \log n)$ time



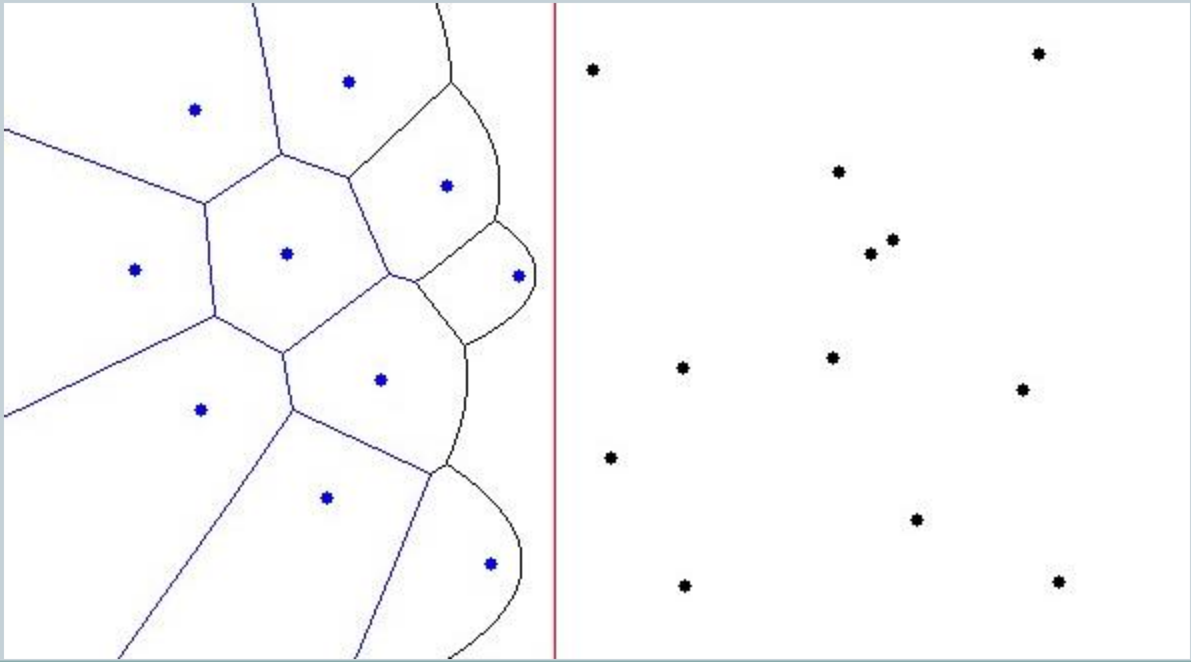
Computing Voronoi diagrams

- Fortune's sweep line algorithm (1987)
  - An imaginary line moves from left to right
  - The Voronoi diagram is computed while the known space expands (left of the line)



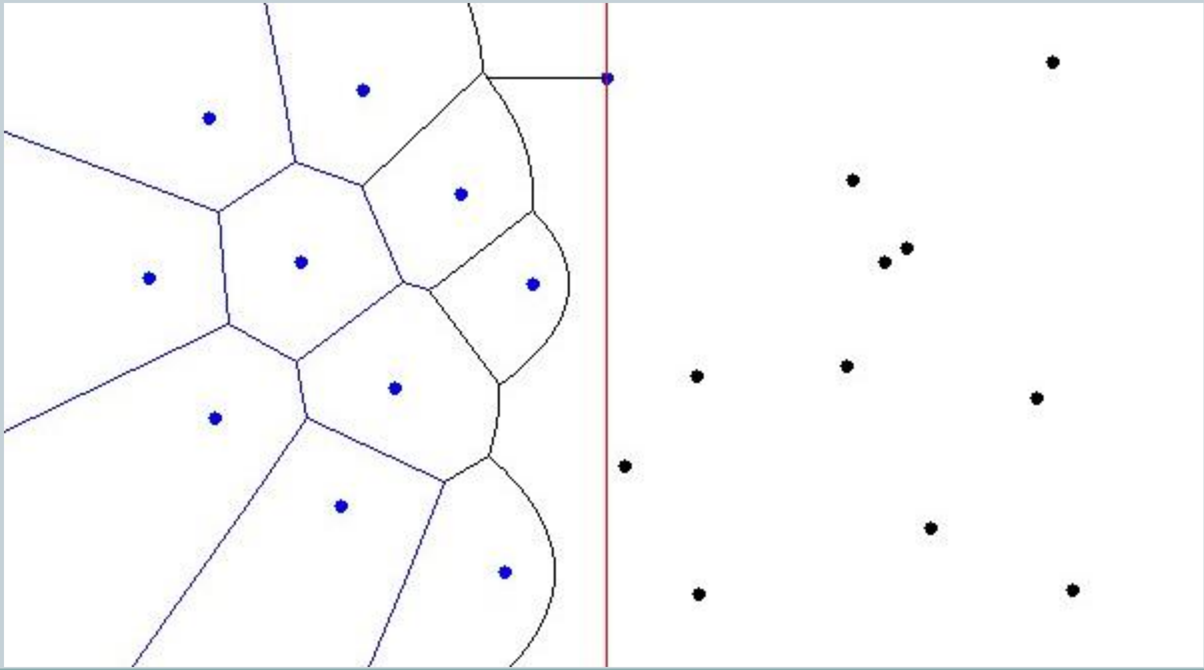
Computing Voronoi diagrams

- Beach line: boundary between known and unknown → sequence of parabolic arcs
  - Geometric property: beach line is y-monotone → it can be stored in a balanced binary tree



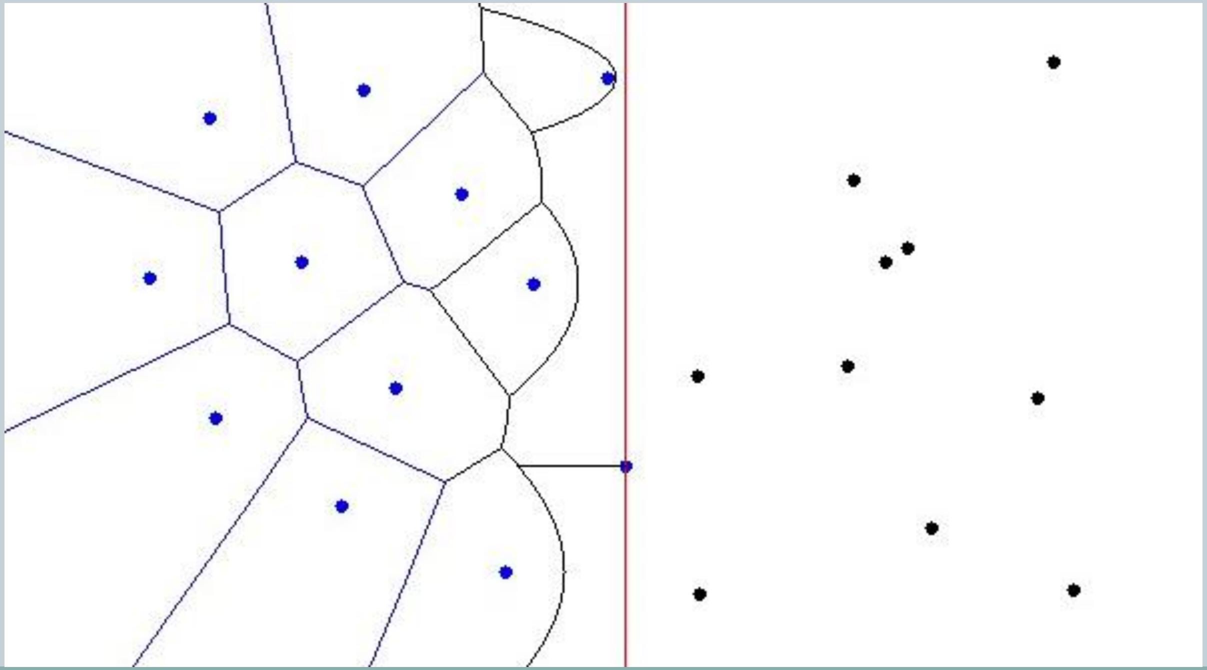
Computing Voronoi diagrams

- Events: changes to the beach line = discovery of Voronoi diagram features
  - Point events



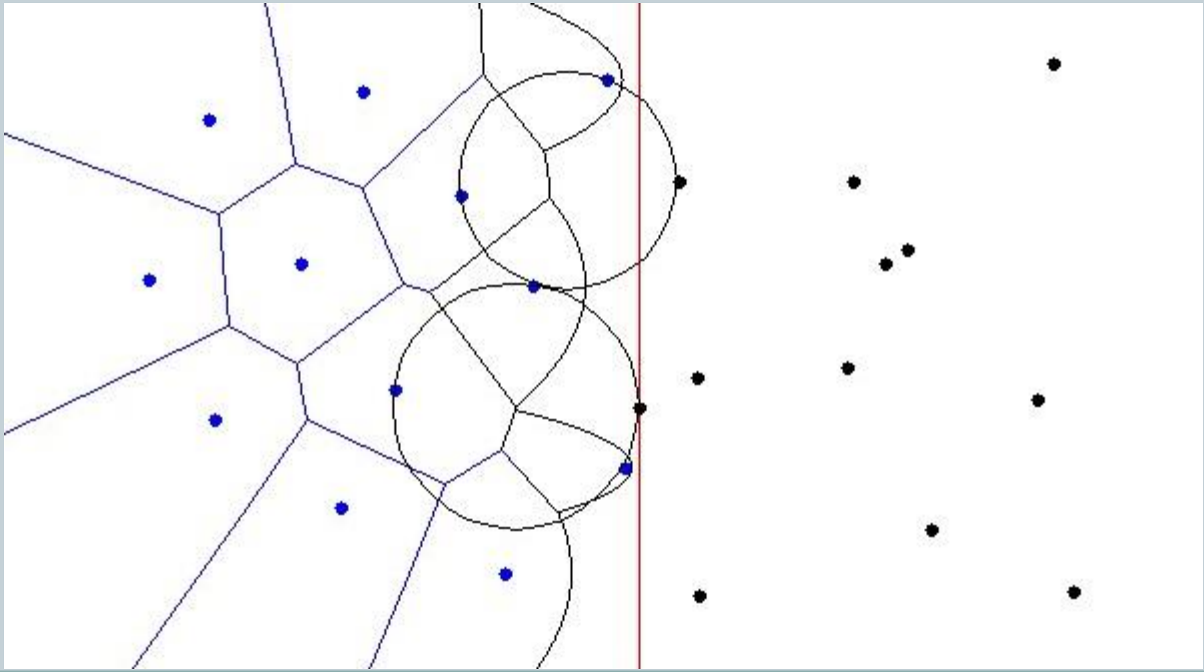
Computing Voronoi diagrams

- Events: changes to the beach line = discovery of Voronoi diagram features
  - Point events



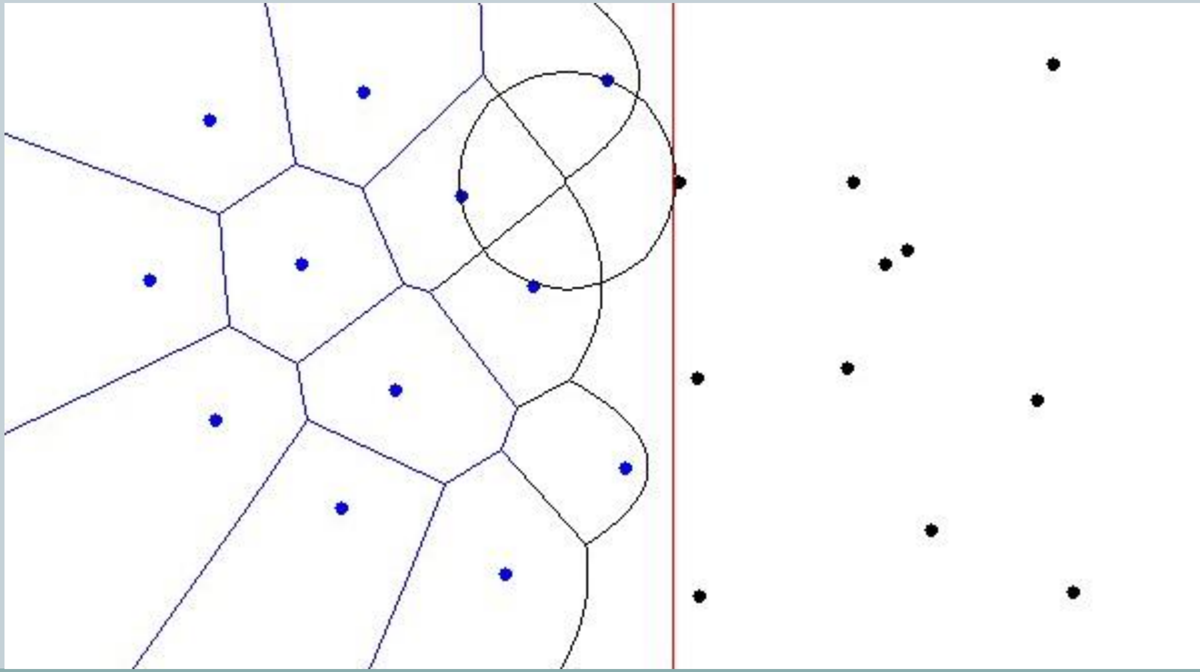
Computing Voronoi diagrams

- Events: changes to the beach line = discovery of Voronoi diagram features
  - Circle events



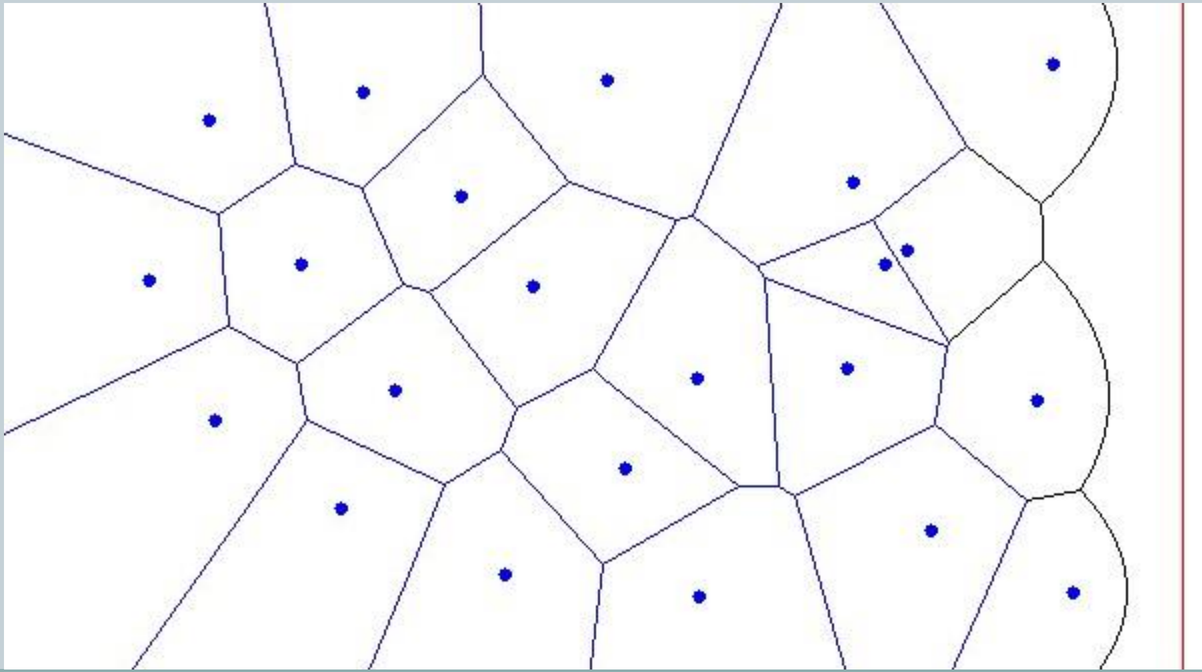
Computing Voronoi diagrams

- Events: changes to the beach line = discovery of Voronoi diagram features
  - Circle events



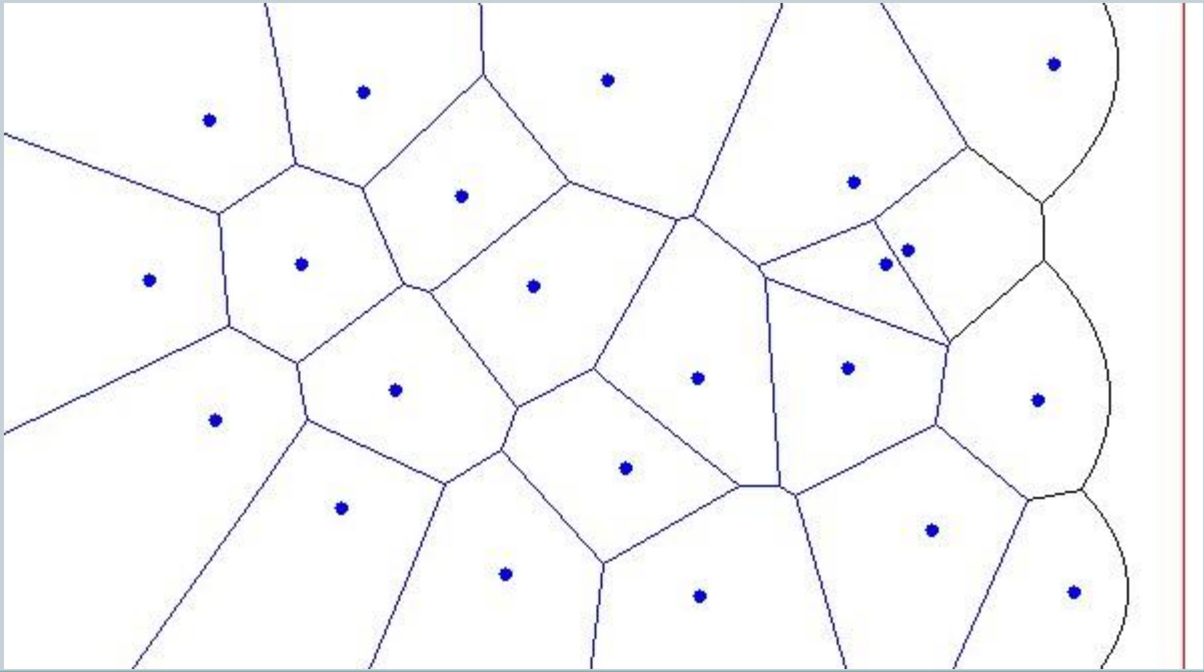
Computing Voronoi diagrams

- Events: changes to the beach line = discovery of Voronoi diagram features
  - Only point events and circle events exist



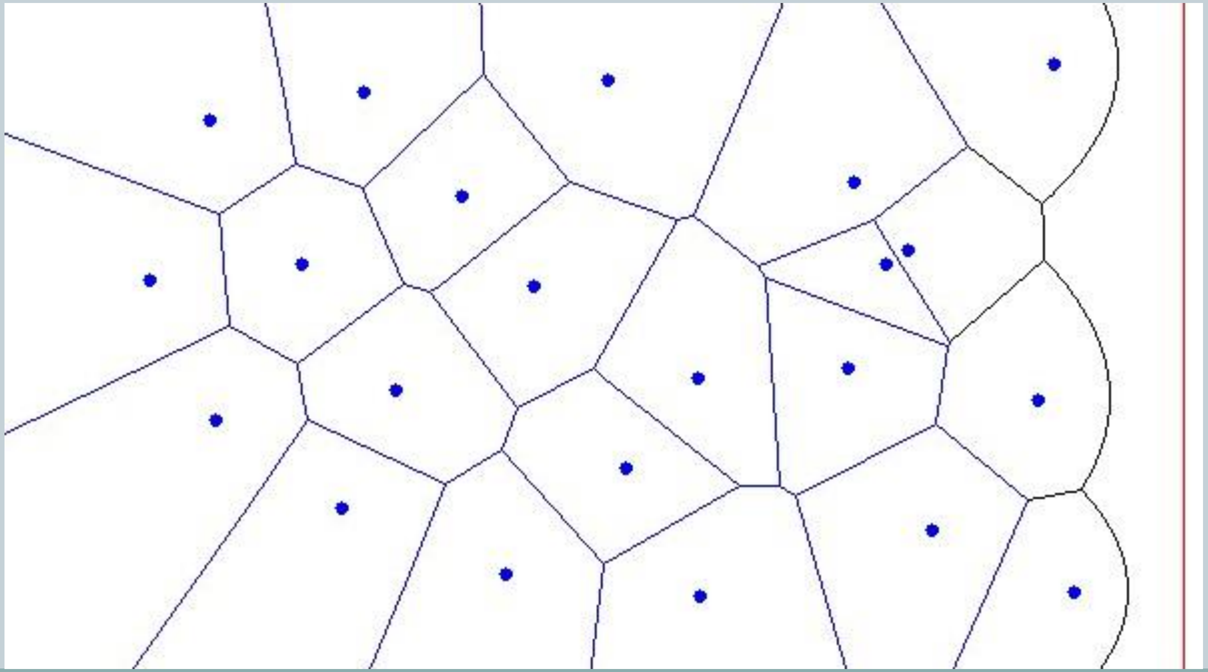
Computing Voronoi diagrams

- For n points, there are
  - n point events
  - at most $2n$ circle events



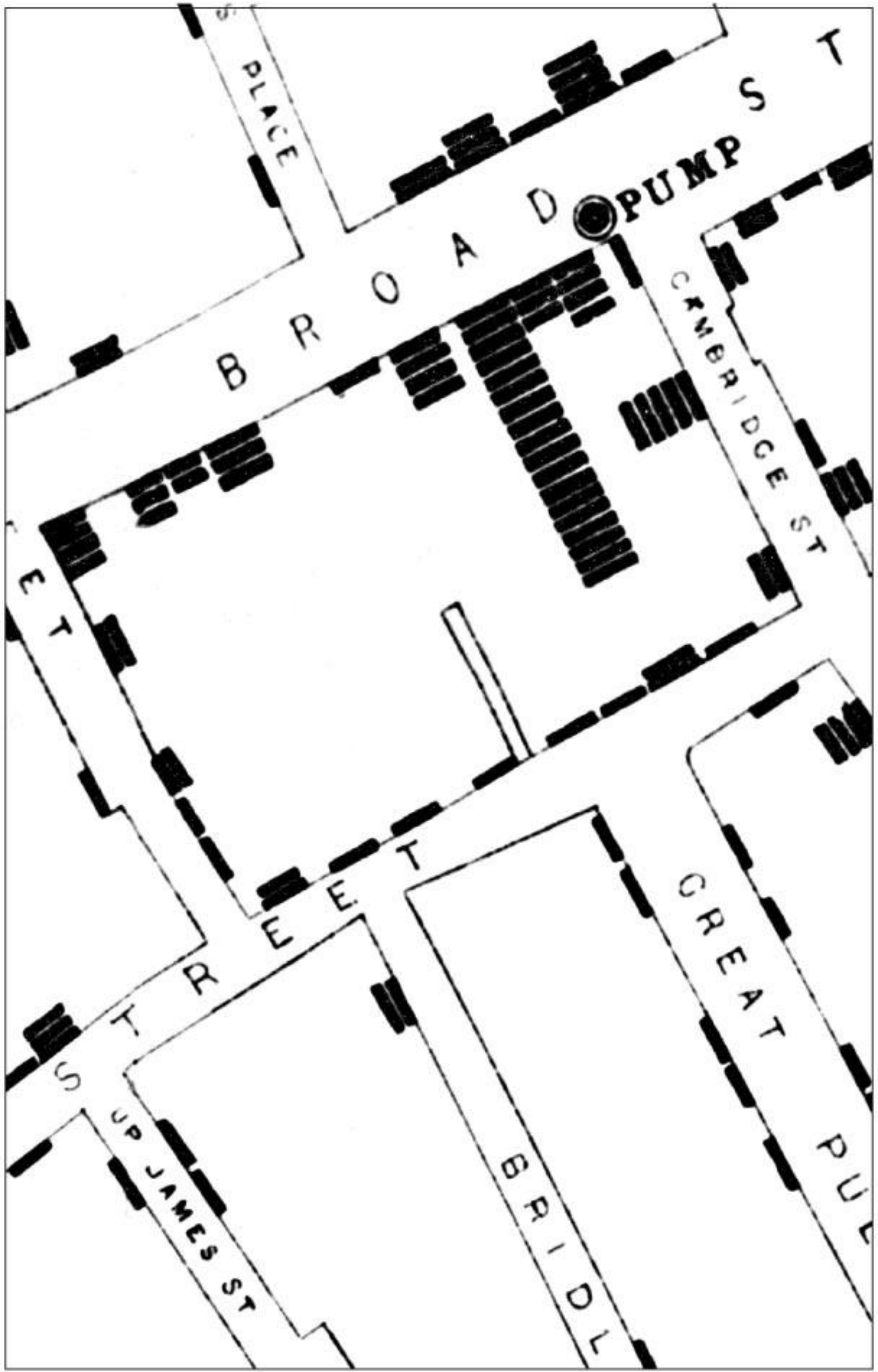
Computing Voronoi diagrams

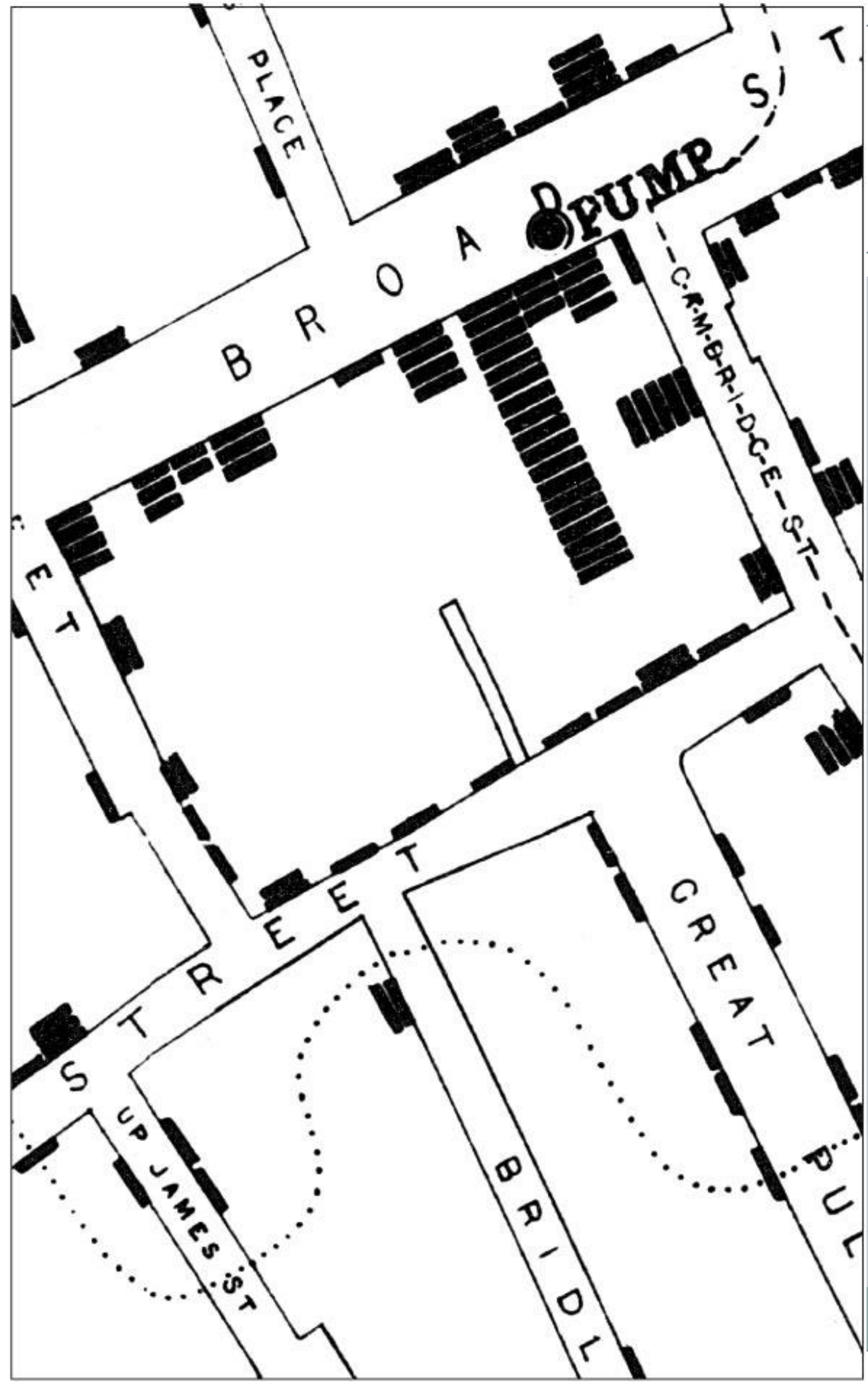
- Handling an event takes $O(\log n)$ time due to the balanced binary tree that stores the beach line → in total $O(n \log n)$ time



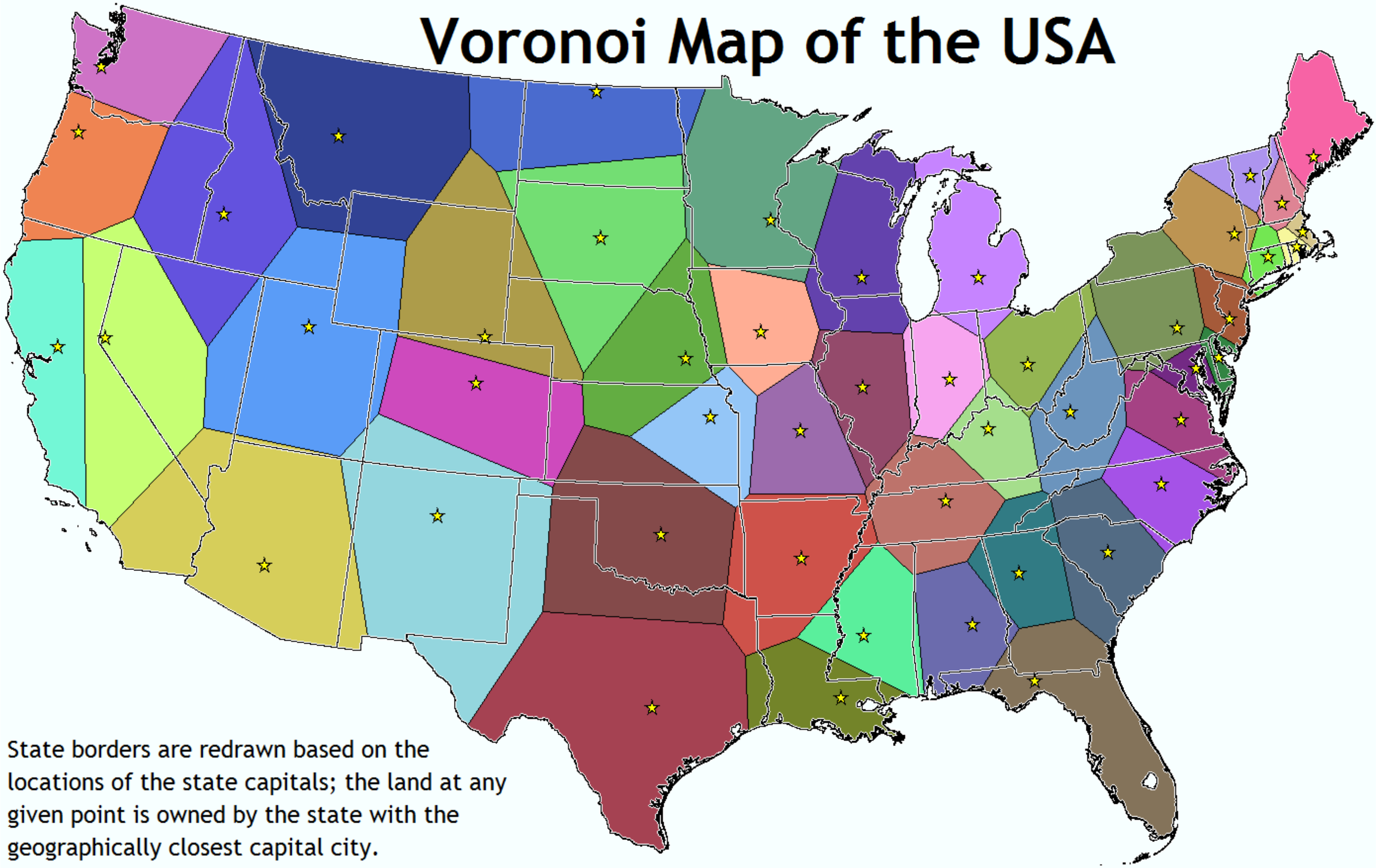
John Snow & The Cholera Outbreak in London

- Discovered that the pump on Broad Street was the source using Voronoi polygons


PLACE
BROAD ST
PUMP
CAMBRIDGE ST
STREET
GREAT PUL
BRIDL
UP JAMES ST


PLACE
BROAD ST
PUMP
CAMBRIDGE ST
STREET
GREAT PUL
BRIDL
UP JAMES ST

Voronoi Map of the USA



State borders are redrawn based on the locations of the state capitals; the land at any given point is owned by the state with the geographically closest capital city.

Voronoi and Corporate Data Mining

- http://www.markbaincreative.com/Unilever-Visual-Identity-2012

Data Mining Tasks

- Classification *[Predictive]*
- Clustering *[Predictive]*
- Association Rule Discovery *[Descriptive]*
- Sequential Pattern Discovery *[Descriptive]*
- Natural Language Processing *[Descriptive]*
- Regression *[Predictive]*
- Deviation Detection *[Predictive]*