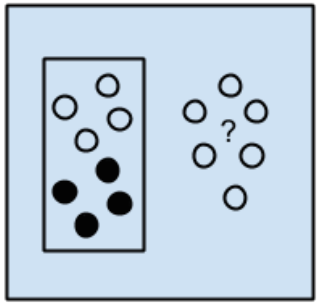
Pertemuan 3

Unsupervised Learning
Clustering

Supervised vs Unsupervised

Supervised

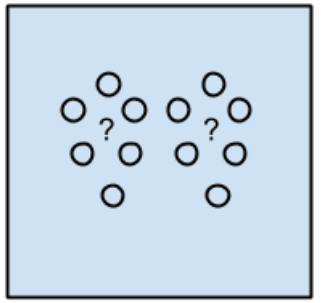
- Memiliki target
- Temukan fungsi yang dapat memetakan data pada targetnya
- Menemukan pola yang menghubungkan atribut dengan targetnya



Supervised Learning Algorithms

Unsupervised

- Tidak memiliki target
- Menemukan struktur data yang mendasarinya
- Tidak memprediksikan secara spesifik, hanya mengelompokan saja



Unsupervised Learning Algorithms

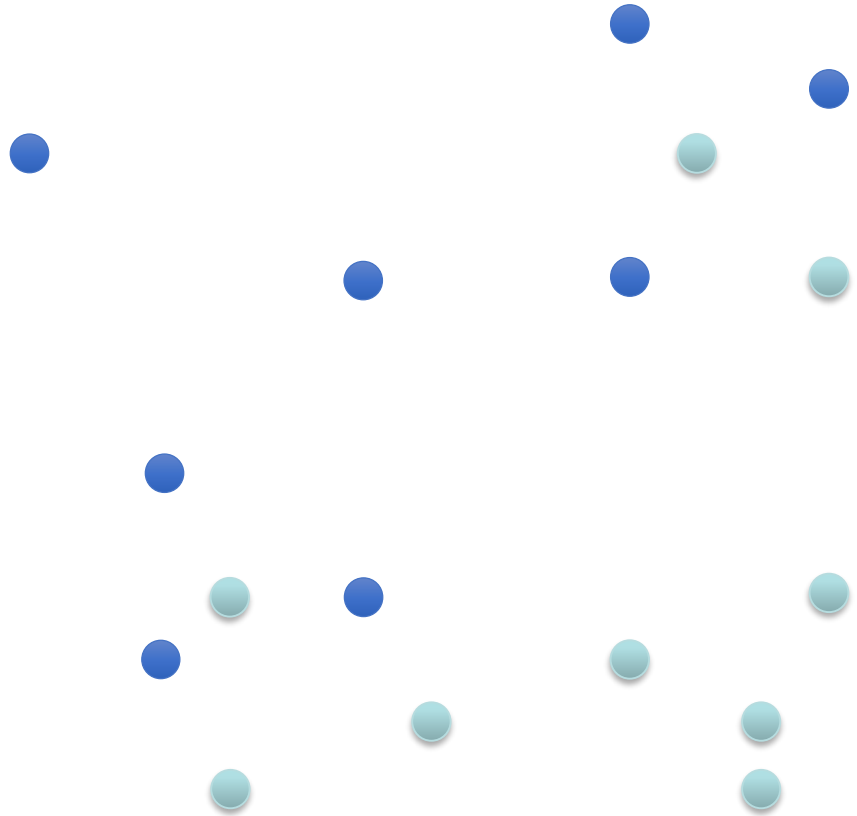
Clustering

- Clustering merupakan proses mengelompokan kumpulan objek menjadi beberapa kelas tertentu berdasarkan kemiripan antara objek-objek tersebut
  - Data pada sebuah kelas *(cluster)* harus berhubungan/mirip
  - Data antar kelas yang berbeda harus tidak saling berhubungan
- Cluster: Kelompok yang berisi data yang mirip
- Analisa Cluster: menemukan kemiripan antara data berdasarkan karakteristik lalu mengelompokkanya kedalam sebuah kelas

Masalah Clustering

- Berapa banyak kelas/*cluster* yang akan dihasilkan?
- Berapa jumlah data pada masing-masing cluster?
- Apakah data pada sebuah cluster memiliki kemiripan?
- Apakah data pada sebuah cluster tidak memiliki kemiripan dengan data pada cluster lain?

Masalah Clustering



Dapat menjadi berapa cluster?

Apakah cluster yang dihasilkan berkualitas?

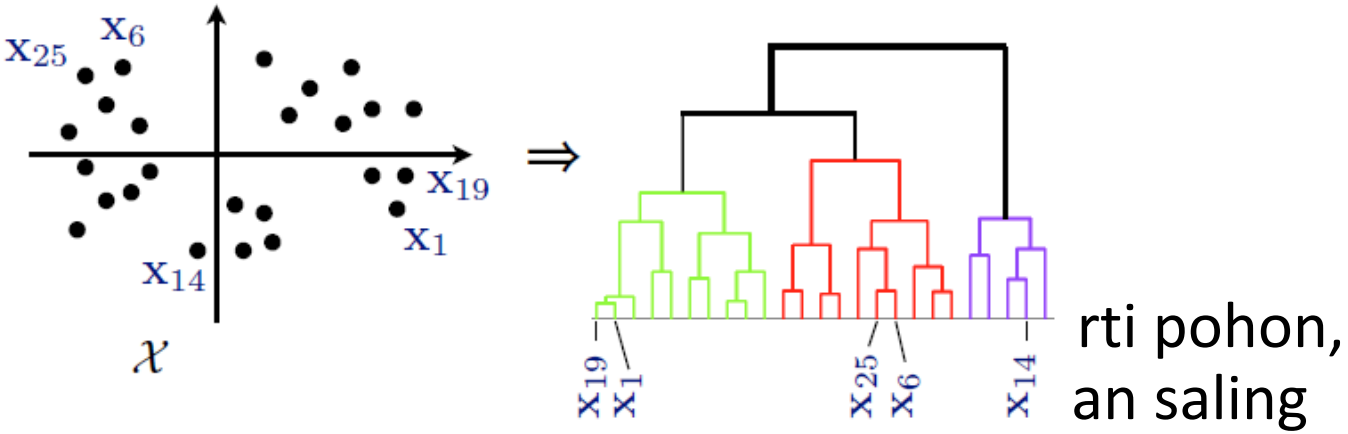
Bagaimana cara mengelompokan data tesebut?

Jenis Clustering

- Hirarkikal (Hierarchical)
  - Objek menjadi lebih terkait dengan objek di dekatnya daripada objek yang lebih jauh
- Partisional (Partitional)
  - Setiap cluster diwakili oleh centroid
  - Ditentukan oleh pengukuran kedekatan objek dengan centroid pada cluster tertentu

Hirarchical Clustering

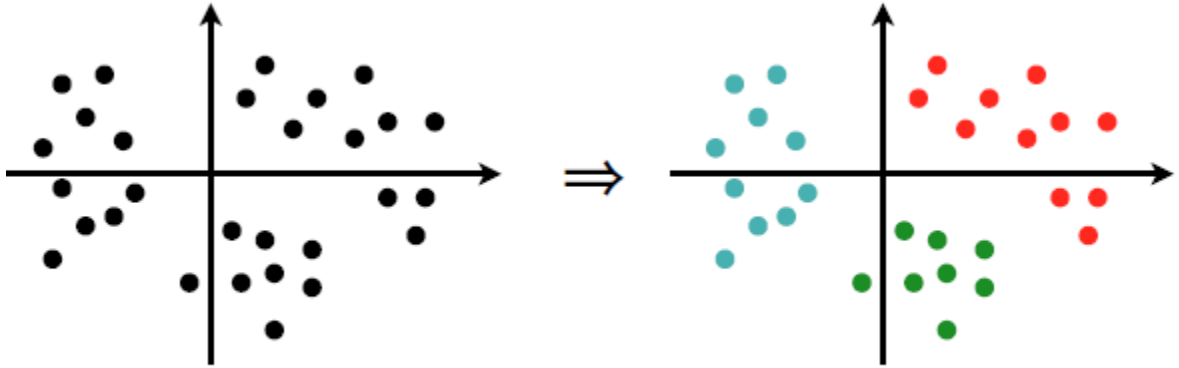
Objek menjadi lebih terkait dengan objek di dekatnya daripada objek yang lebih jauh



Pada pe rti pohon, dimana an saling berdekatan satu sama lain.

Partitional Clustering

Setiap cluster diwakili oleh centroid dan diukur berdasarkan pengukuran jarak



Biasany[illegible] ak ditentukan sebelumnya/

Algoritma Clustering

1. K-Means
2. Fuzzy C Means
3. Agglomerative
4. K-D Trees
5. EM Clustering
6. Quality Threshold

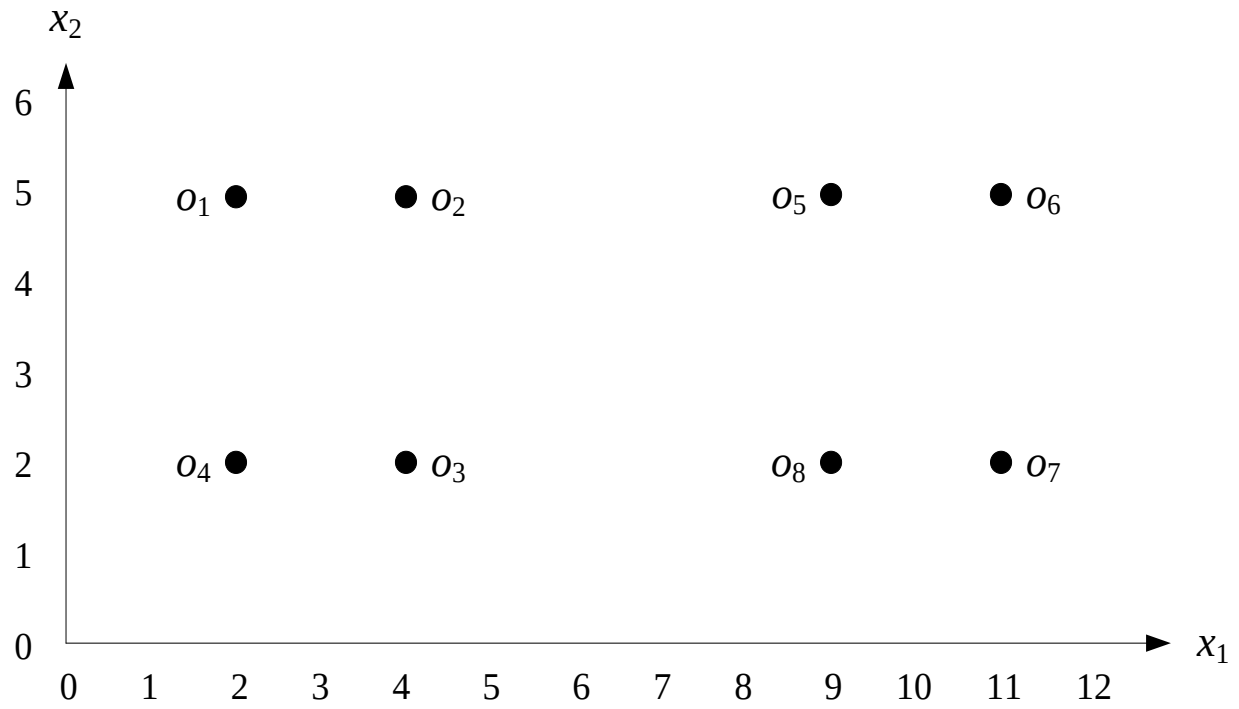
Kegunaan Clustering

1. Menemukan kelas pada dataset yang tidak memiliki target
2. Dimensionality reduction
3. Color-based image segmentation
4. Analisa jejaring media social
5. Segmentasi pasar

Algoritma K-Means

- Algoritma paling sederhana dan paling sering digunakan untuk kasus clustering
- Data dipartisi/dikelompokan menjadi *k* cluster *(k* merupakan jumlah cluster yang diinginkan*)*
- Setiap data pada sebuah cluster mirip dengan centroidnya

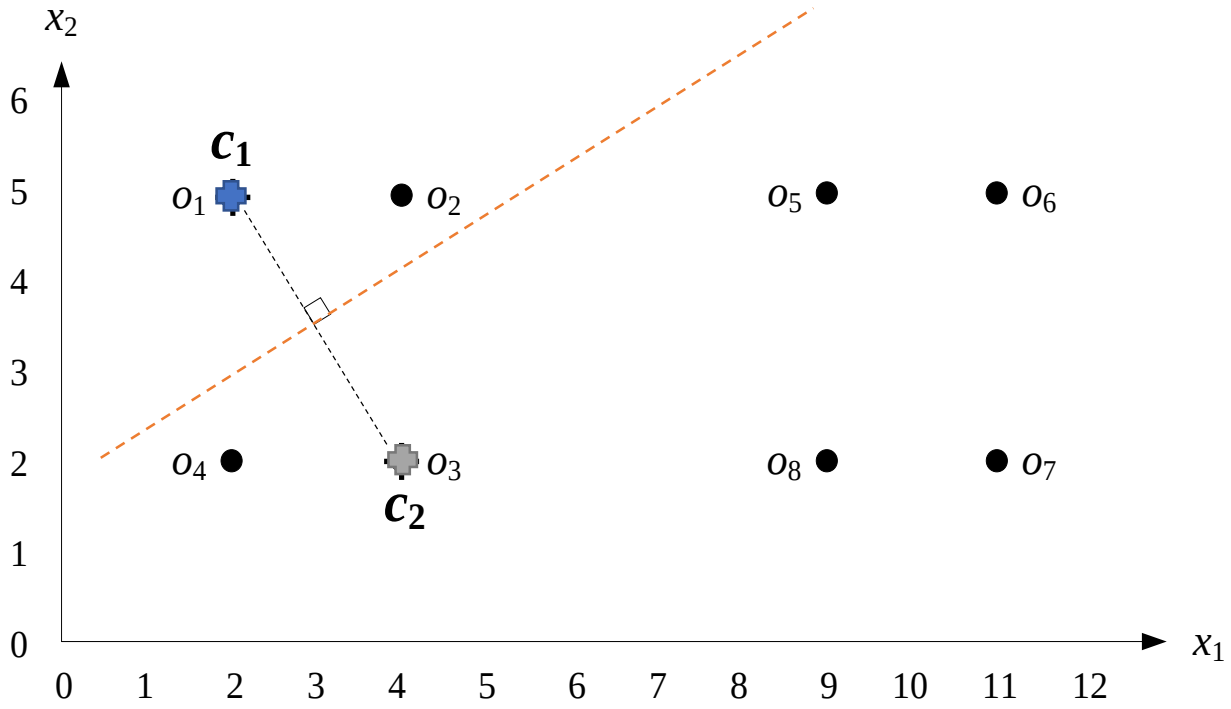
Contoh Kasus K-Means



Jumlah Cluster *(k) = 2*

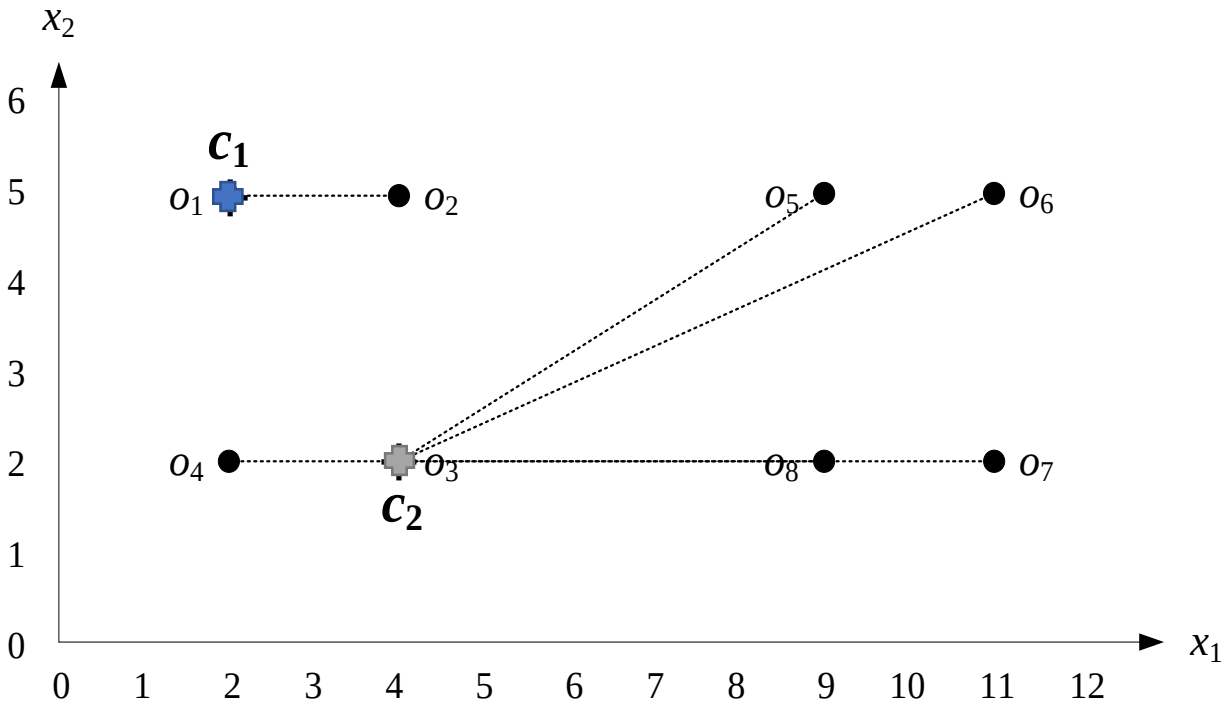
Inisialisasi 2 buah centroid secara acak

Contoh Kasus K-Means



Setiap data point *o*, temukan centroid *c* yang paling dekat

Contoh Kasus K-Means

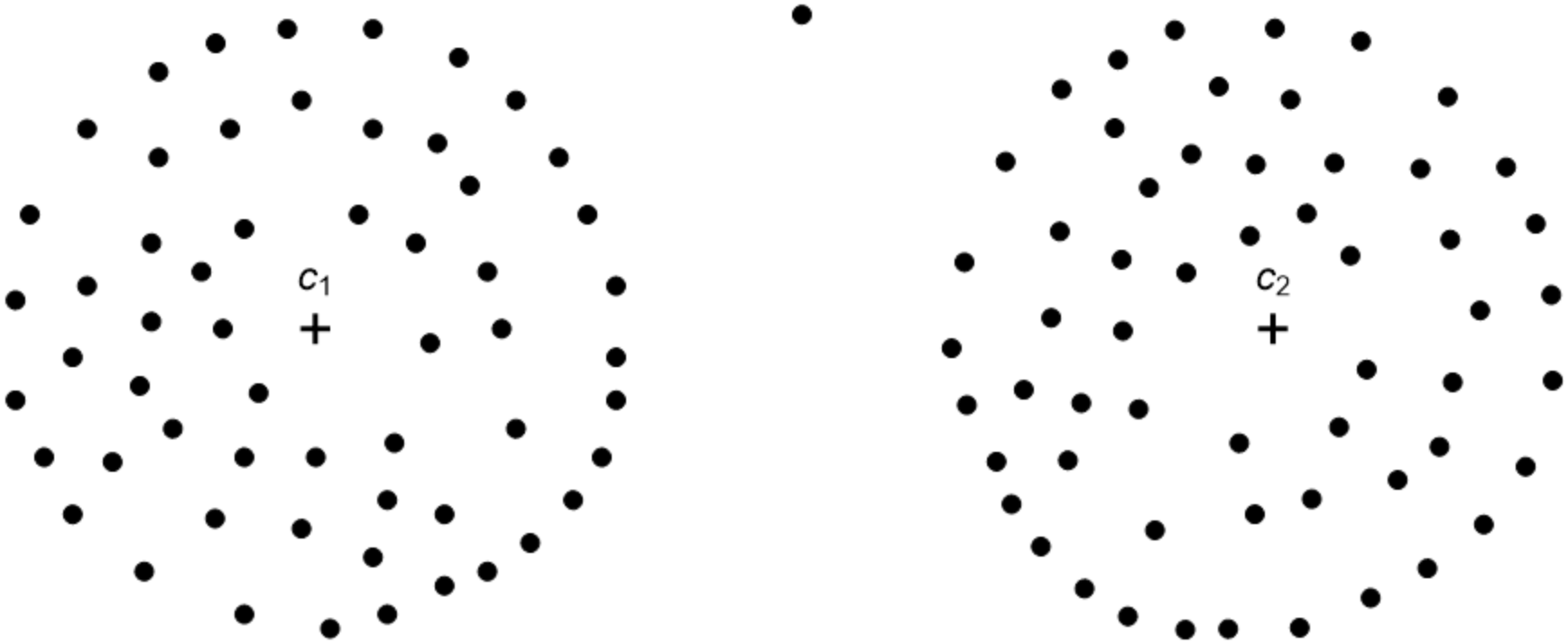


Kalkulasikan rata-rata data dari setiap cluster

Lalu update centroid nya

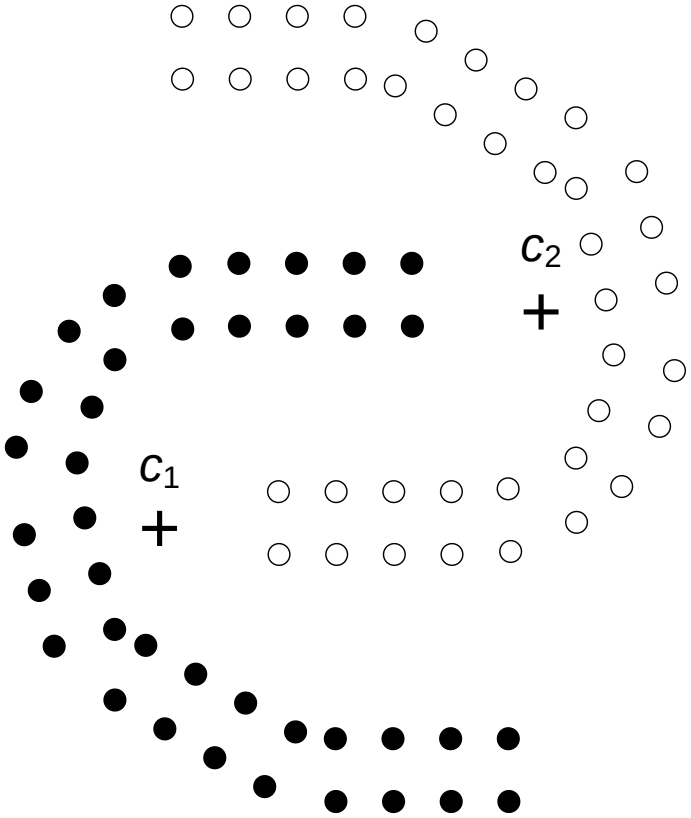
Masalah k-Means

Data point ini termasuk pada cluster yang mana?



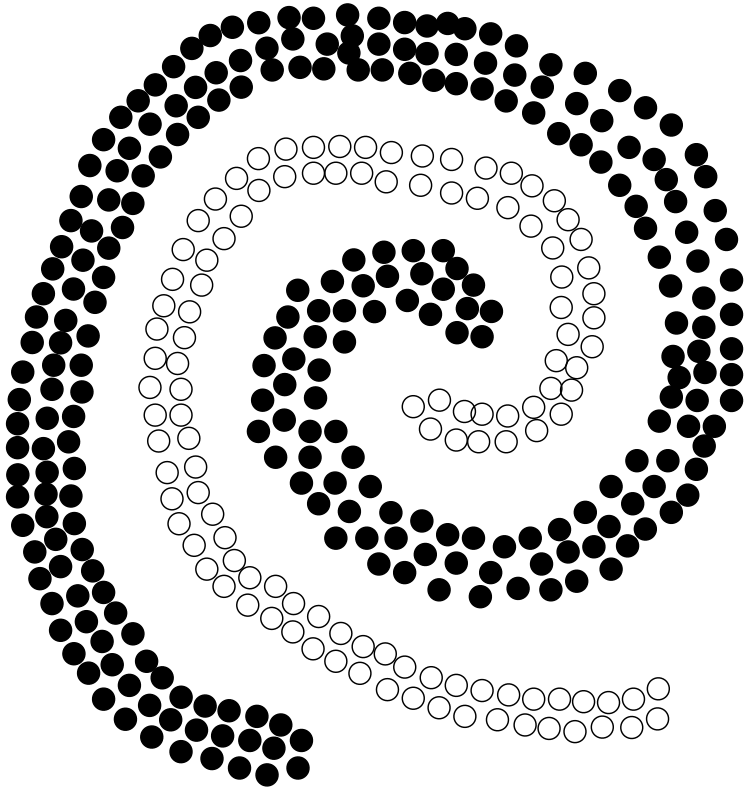
Masalah k-Means

Apakah hasil clustering ini berkualitas?
Apakah centroid ini tepat?
Apakah akurat?



Masalah k-Means

Dimana harus meletakan centroidnya?



Kelebihan dan Kekurangan K-Means

- Kelebihan:
  - Relatif sederhana untuk diimplementasikan
  - Lebih efisien dari segi waktu dan biaya komputasi
  - Cocok untuk data yang rapih dan terstruktur
- Kekurangan:
  - Sulit menentukan *k*
  - Sensitif terhadap penentuan centroid awal
  - Tidak cocok untuk data yang sebarannya terlalu bervariasi