Clustering: k-Means

Ali Akbar Septiandri

October 23, 2018

Universitas Al Azhar Indonesia

Daftar isi

Clustering

Clustering

- *Unsupervised learning*

Clustering

- *Unsupervised learning*
- Subpopulasi apa yang ada dalam data?

Clustering

- *Unsupervised learning*
- Subpopulasi apa yang ada dalam data?
- Apa kesamaan dari elemen di tiap subpopulasi?

Clustering

- *Unsupervised learning*
- Subpopulasi apa yang ada dalam data?
- Apa kesamaan dari elemen di tiap subpopulasi?
- Bisa digunakan untuk menemukan pencilan

Contoh Data

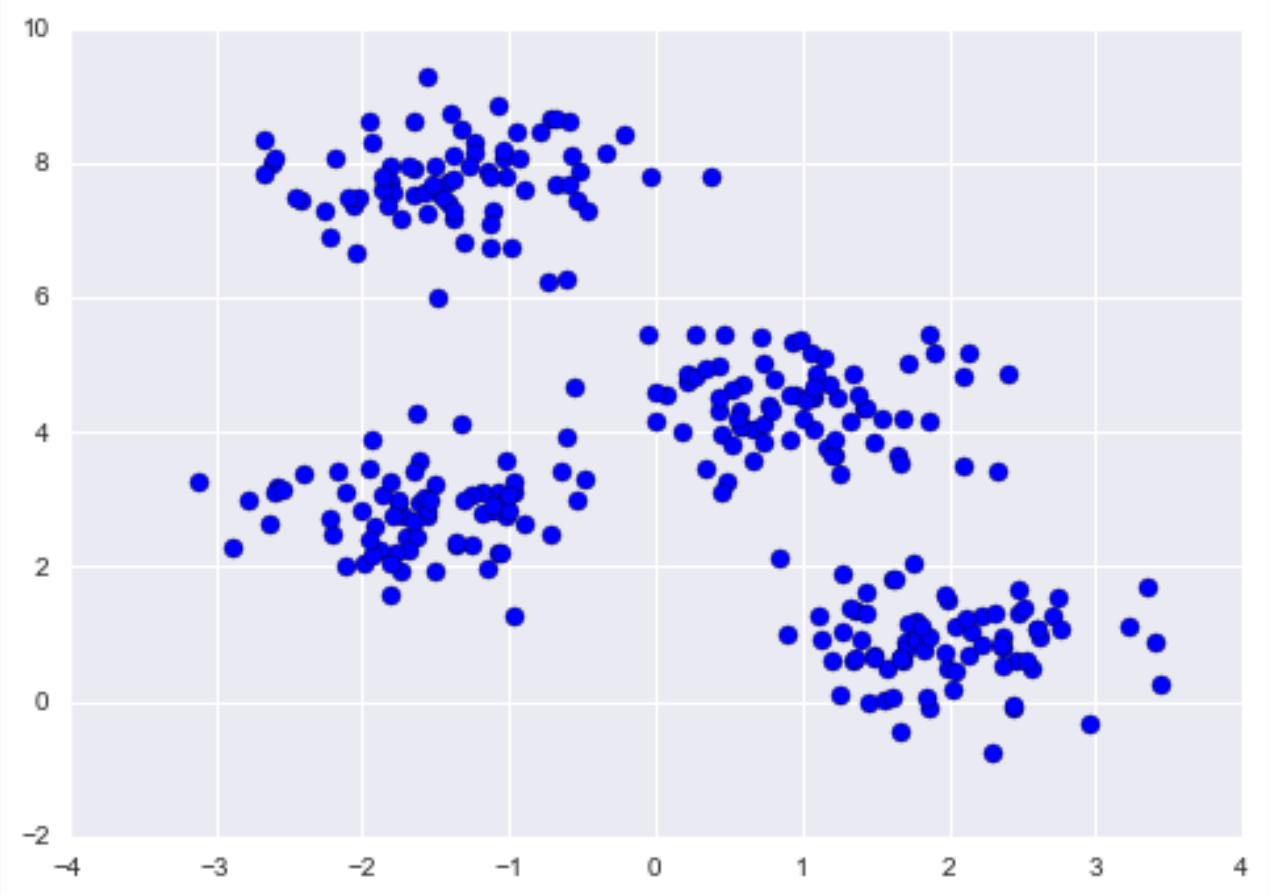


Figure 1: Contoh data dalam 2D [VanderPlas, 2016]

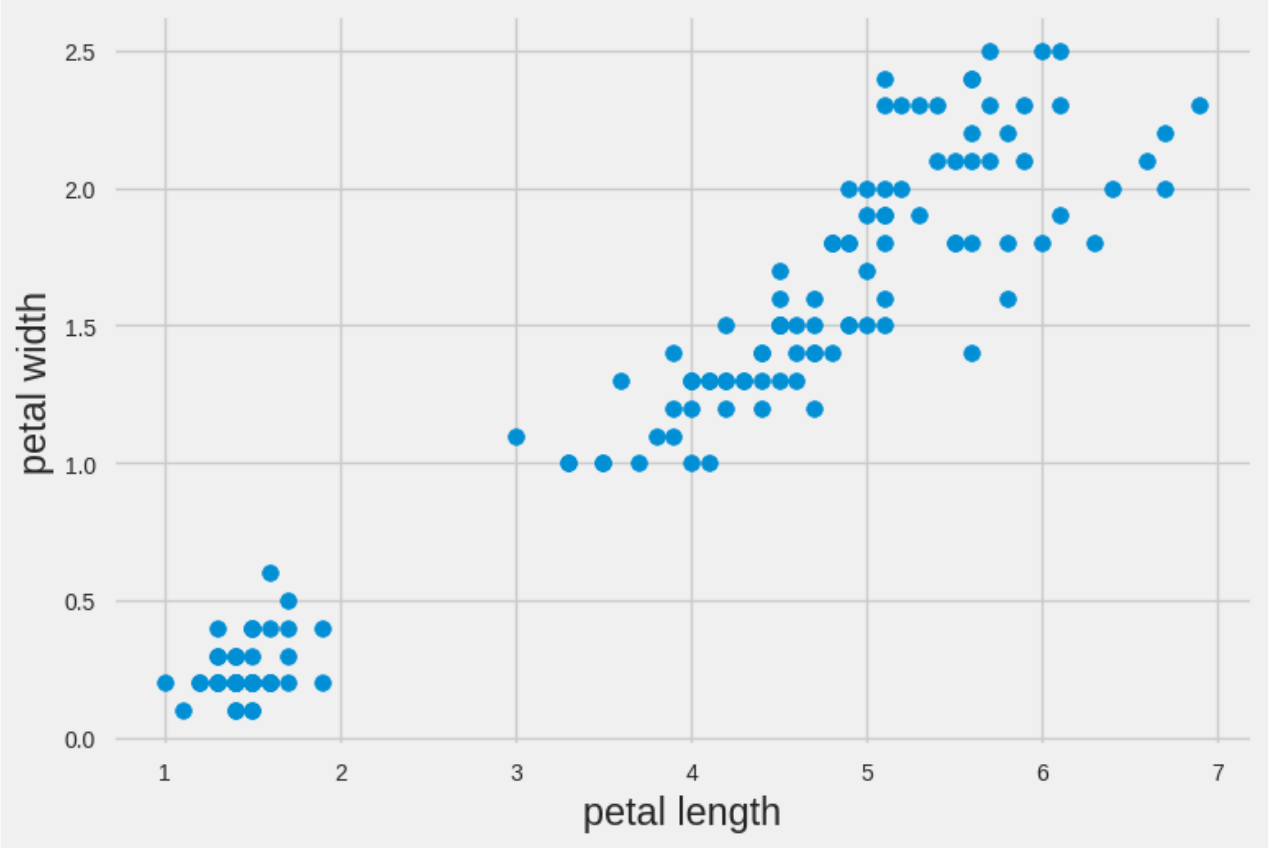
Subpopulasi

Figure 2: Subpopulasi dari algoritma k-Means [VanderPlas, 2016]

Tetangga dalam satu kompleks,
tanpa memedulikan kelasnya

Berapa jumlah subpopulasi (klaster)
yang ingin kita cari?

Data Iris



Klaster

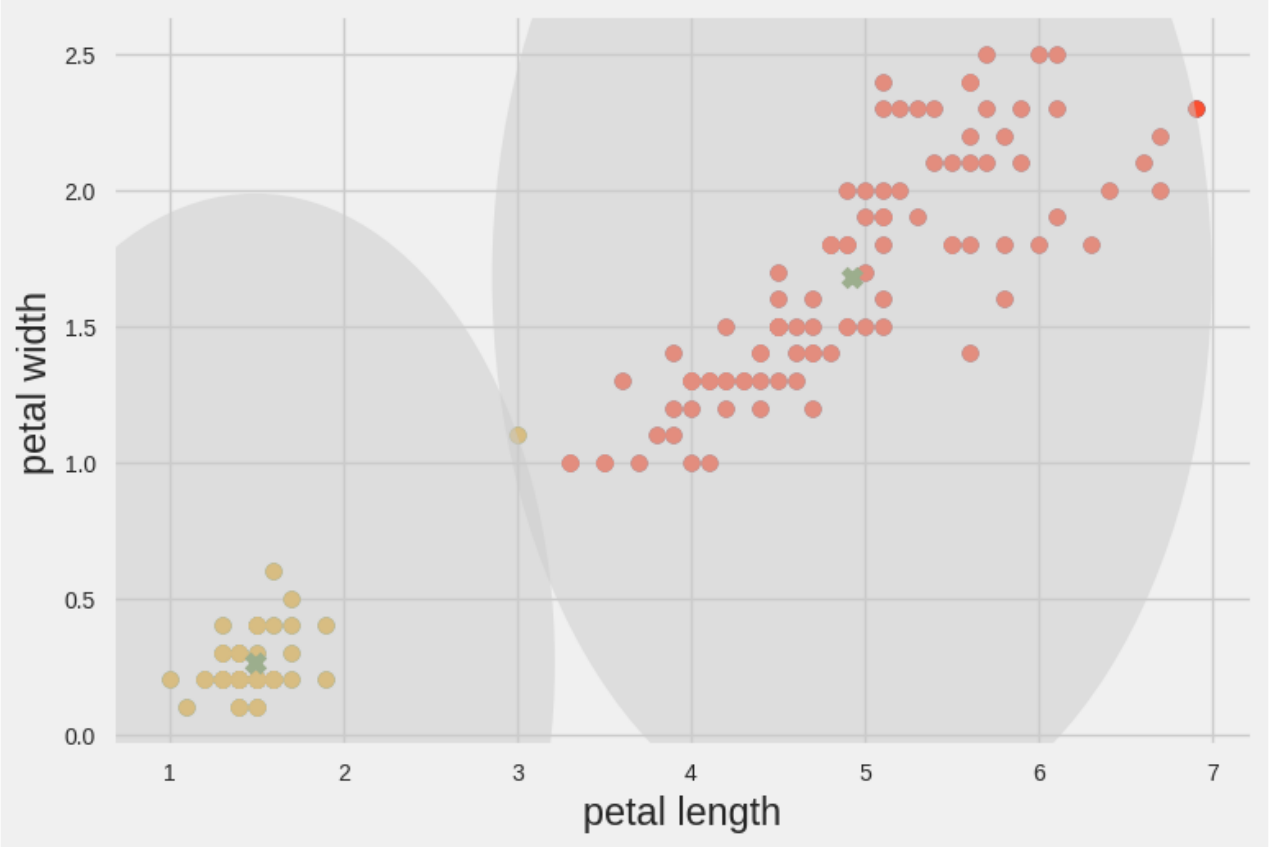


Figure 3: k-Means dengan $k = 2$

Klaster

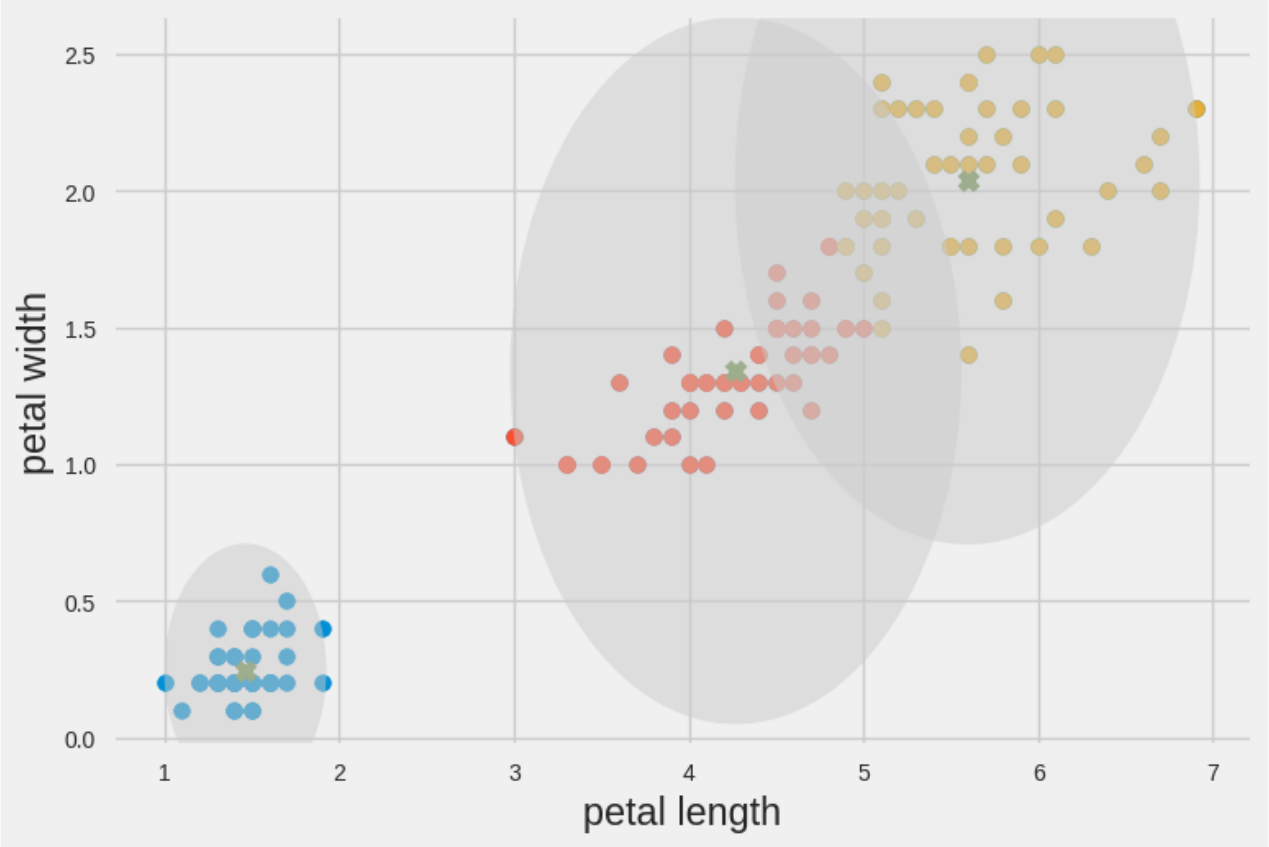


Figure 4: k-Means dengan $k = 3$

Klaster

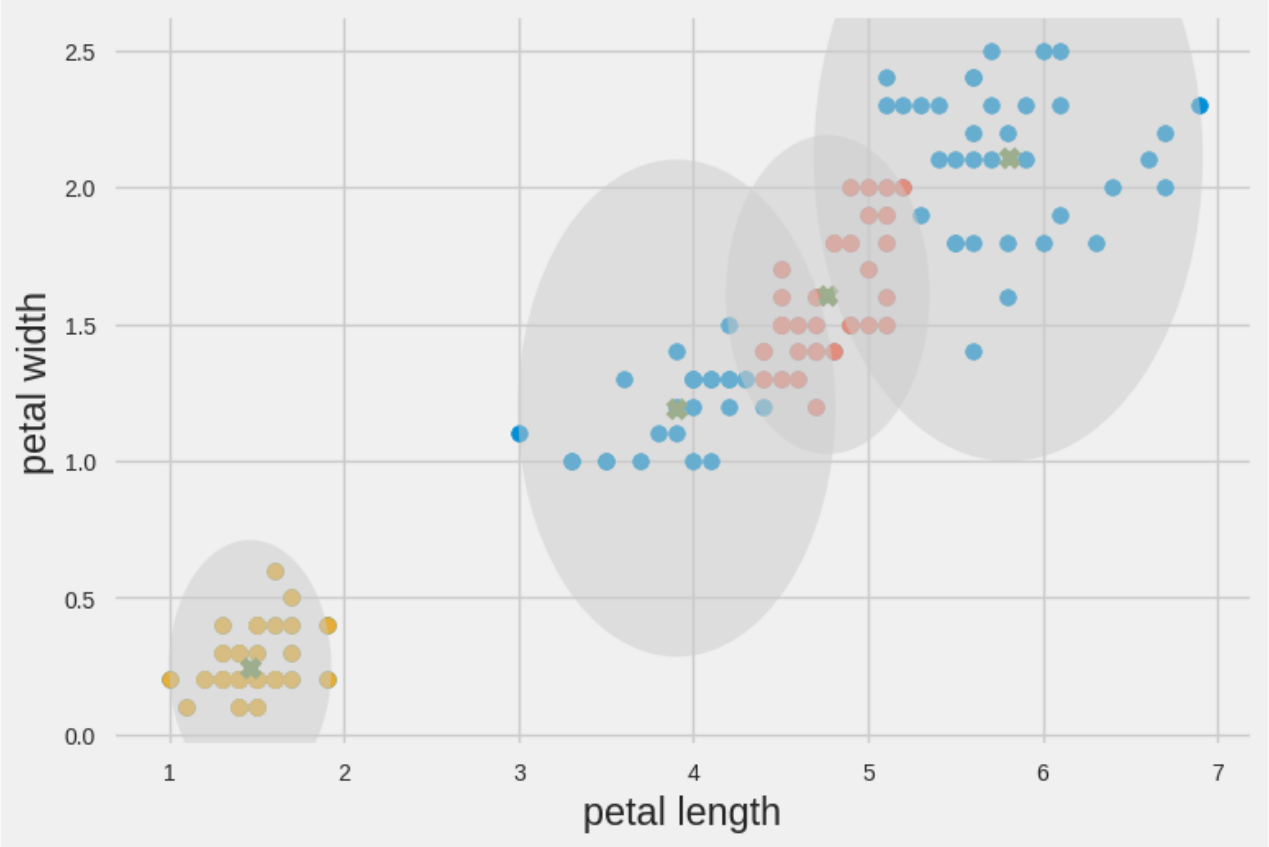


Figure 5: k-Means dengan $k = 4$

k-Means

- Jumlah k ditentukan dari awal
- Tidak memerlukan label
- Menggunakan *centroid*, i.e. rata-rata nilai dari objek yang masuk dalam *cluster* tersebut
- Mencari *centroid* terdekat dari tiap objek

Algoritma: Expectation-Maximization

1. Inisialisasi k *centroid* secara acak
2. Ulangi hingga konvergen
   A. E-step: Masukkan tiap titik/objek ke *centroid* terdekat

$$\arg\min_j D(x_i, c_j)$$

   B. M-step: Ubah nilai *centroid* menjadi rata-rata dari tiap titik/objek

$$c_j(a) = \frac{1}{n_j} \sum_{x_i \to c_j} x_i(a), \text{for } a = 1..d$$

Visualisasi EM

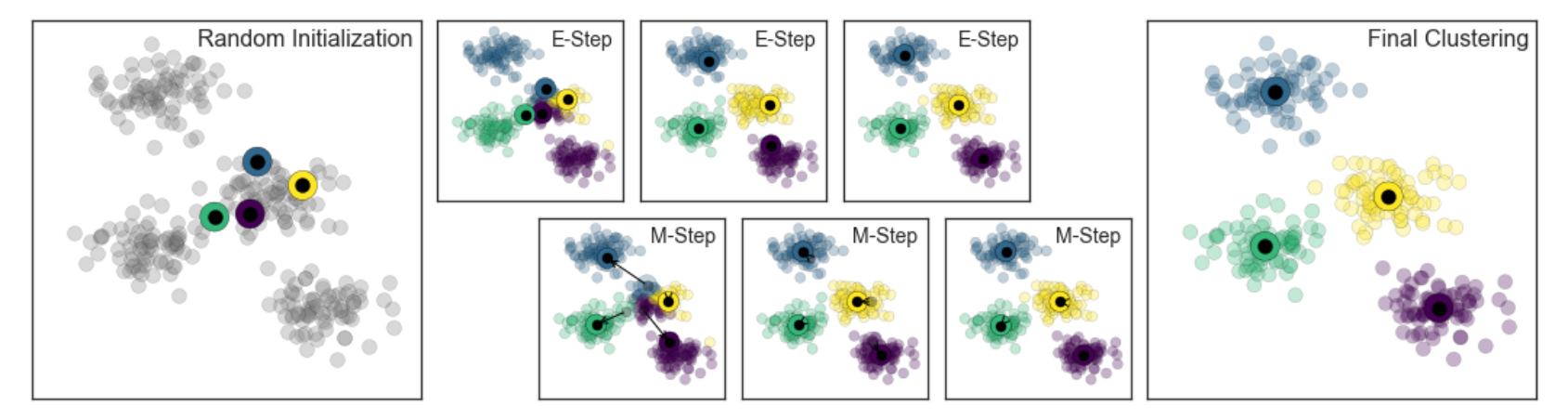


Figure 6: Konvergensi klaster tercapai hanya dalam tiga iterasi [VanderPlas, 2016]

Algoritma ini sangat bergantung pada inisialisasi *centroid*!

Berapa nilai k yang optimal?

Menentukan Nilai k

- Gunakan label kelas, e.g. 10 untuk MNIST
- Gunakan V untuk menggambarkan *scree plot*

$$V = \sum_{j} \sum_{x_i \to c_j} D(c_j, x_i)^2$$

lalu gunakan *elbow method*, i.e. nilainya dapat dicari dengan menggunakan nilai optimal turunan kedua

Scree Plot

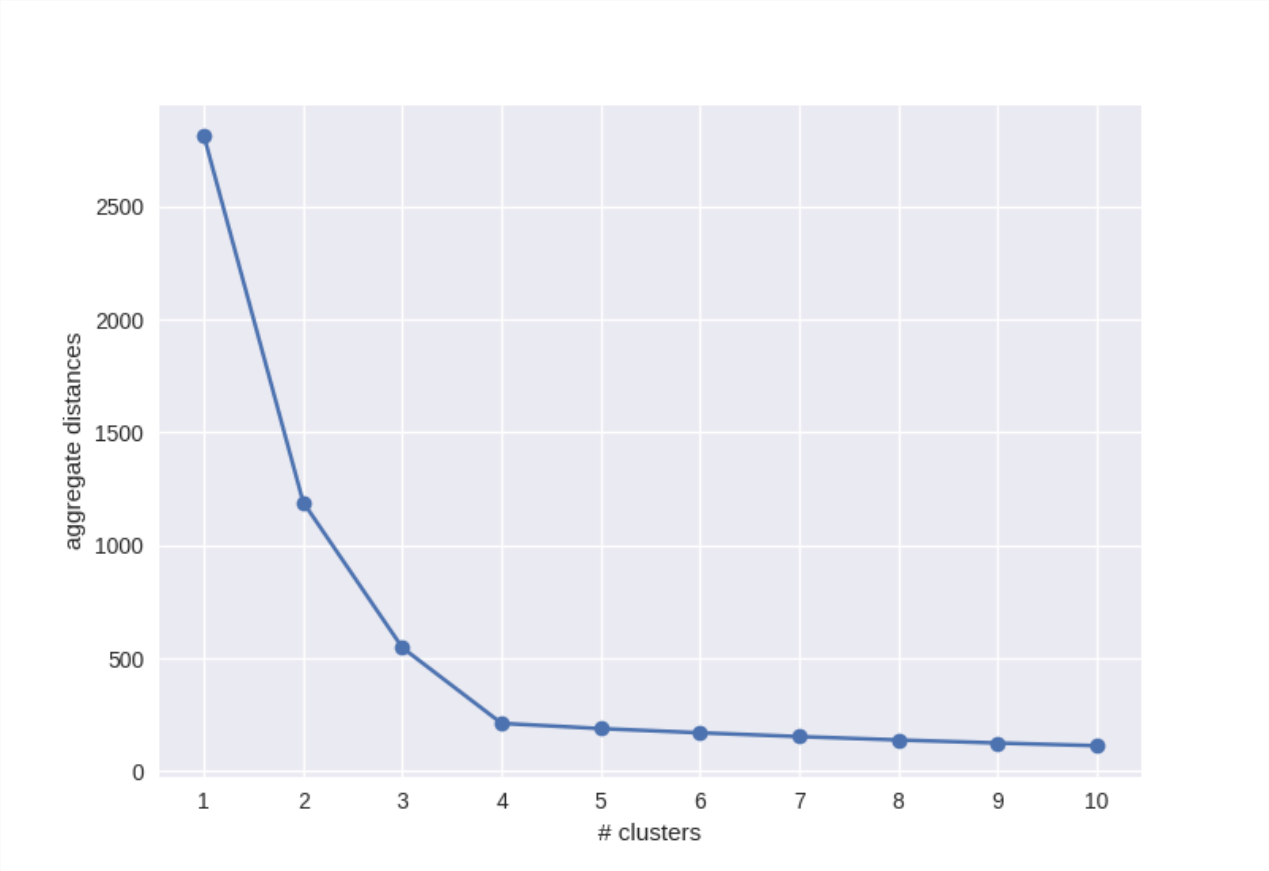


Figure 7: Secara visual, scree plot menunjukkan nilai optimal $k = 4$

Mengevaluasi Clustering

Evaluasi Intrinsik: Klaster $\sim$ Kelas

- Klaster $c_1, c_2, ..., c_K$
- Kelas $R_1, R_2, ..., R_N$
- Cocokkan R_i dengan c_j, hitung akurasinya

Contoh Evaluasi Intrinsik

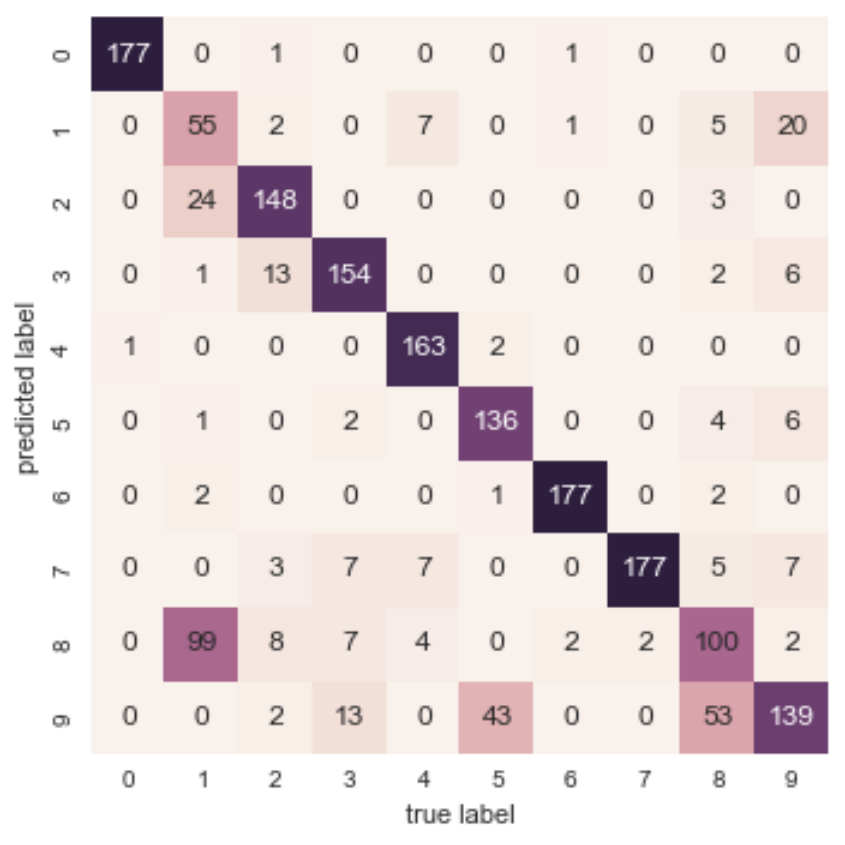


Figure 8: Confusion matrix dari MNIST clustering [VanderPlas, 2016]

Contoh Evaluasi Intrinsik

	G1	G2	G3	G4	G5	G6
C1	1	**7**	0	1	4	0
C2	0	0	0	0	2	**7**
C3	0	0	**2**	0	0	0
C4	**3**	1	0	0	1	0

Figure 9: Klaster karakter dalam Julius Caesar

Aplikasi

Aplikasi Clustering

- Representasi gambar: *bag of cluster id* atau fitur lain (lihat [Coates, 2012])
- Kompresi gambar
- Sistem rekomendasi

Aplikasi: Kompresi gambar

Figure 10: Gambar yang akan dikompresi dengan *clustering* [VanderPlas, 2016]

Klaster warna

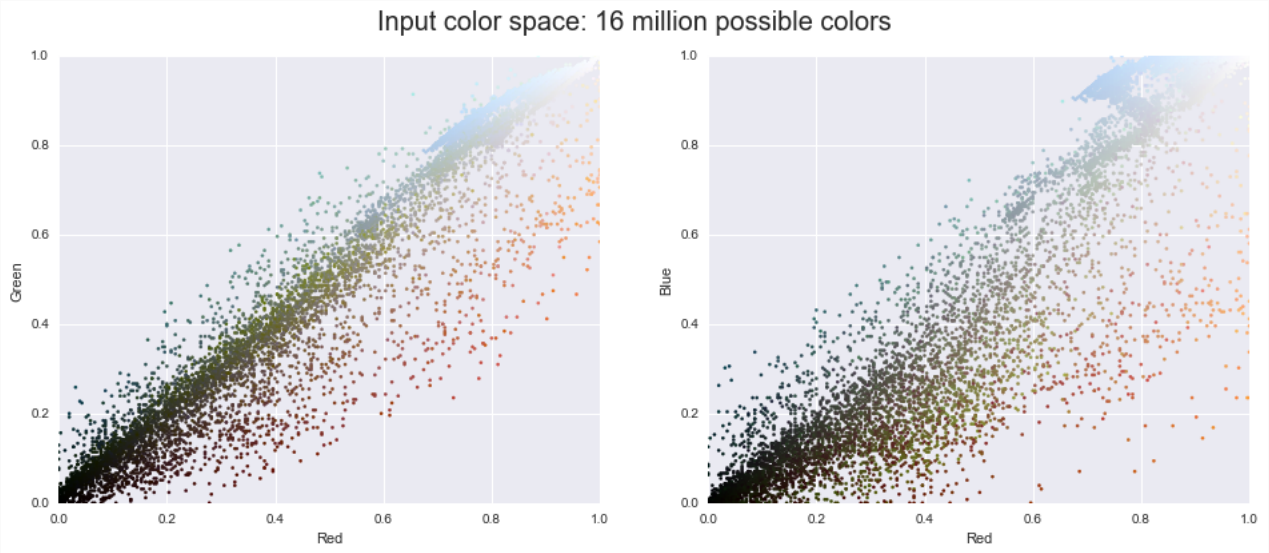


Figure 11: *Clustering* warna dengan kompresi [VanderPlas, 2016]

Klaster warna

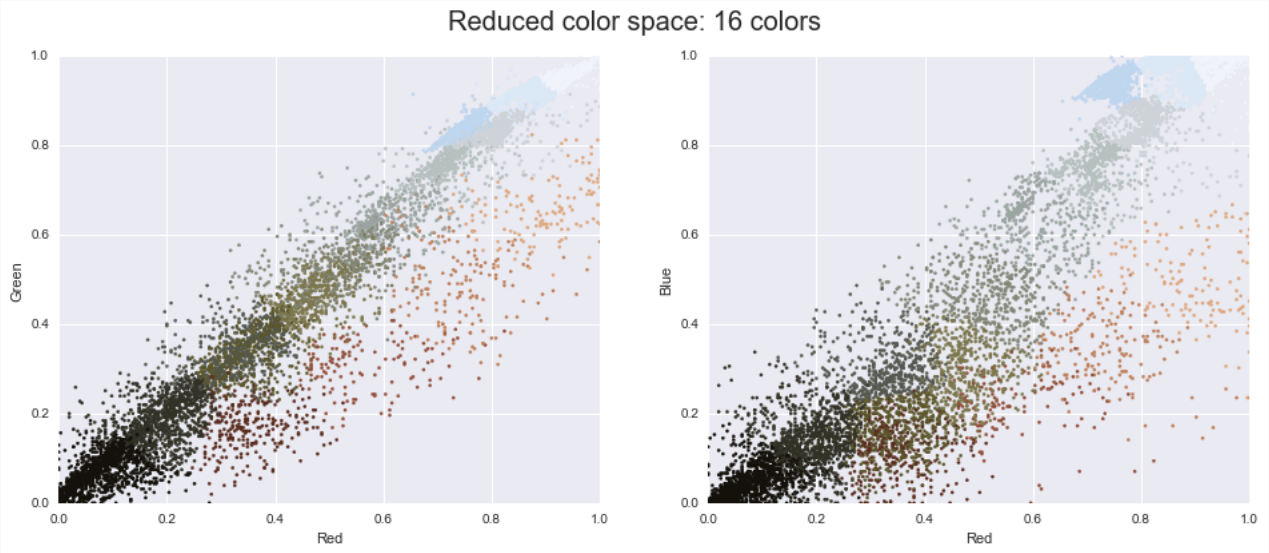


Figure 11: *Clustering* warna dengan kompresi [VanderPlas, 2016]

Hasil kompresi gambar

Figure 12: Kompresi dengan faktor hingga 1 juta dengan *clustering* [VanderPlas, 2016]

Ikhtisar

- *Clustering* merupakan salah satu tugas *unsupervised learning*, i.e. tidak memerlukan label
- Nilai k merupakan jumlah klaster dalam algoritma k-Means
- k-Means sangat bergantung pada inisialisasi centroid

Pertemuan berikutnya

- Kuliah tamu: Go-Jek
- Penggunaan AI di Go-Jek

Referensi

Jake VanderPlas (2016)

In Depth: k-Means Clustering

`https://jakevdp.github.io/PythonDataScienceHandbook/05.11-k-means.html`

Adam Coates & Andrew Y. Ng (2012)

Learning feature representations with k-means.

Neural networks: Tricks of the trade (pp. 561-580). Springer Berlin Heidelberg.

Terima kasih