k-Means Clustering

Ali Akbar Septiandri

Universitas Al-Azhar Indonesia

aliakbars@live.com

June 22, 2019

Selayang Pandang

1. Clustering
2. Evaluasi

Bahan Bacaan

1. VanderPlas, J. (2016). Python Data Science Handbook. (In Depth: k-Means Clustering) http://nbviewer.jupyter.org/github/jakevdp/PythonDataScienceHandbook/blob/master/notebooks/05.11-K-Means.ipynb
2. "Clustering dengan k-Means." *Cerita Tentang Data*. 14 November 2017. https://tentangdata.wordpress.com/2017/11/14/clustering-dengan-k-means/

Clustering

Clustering

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

Jenis-jenis Clustering

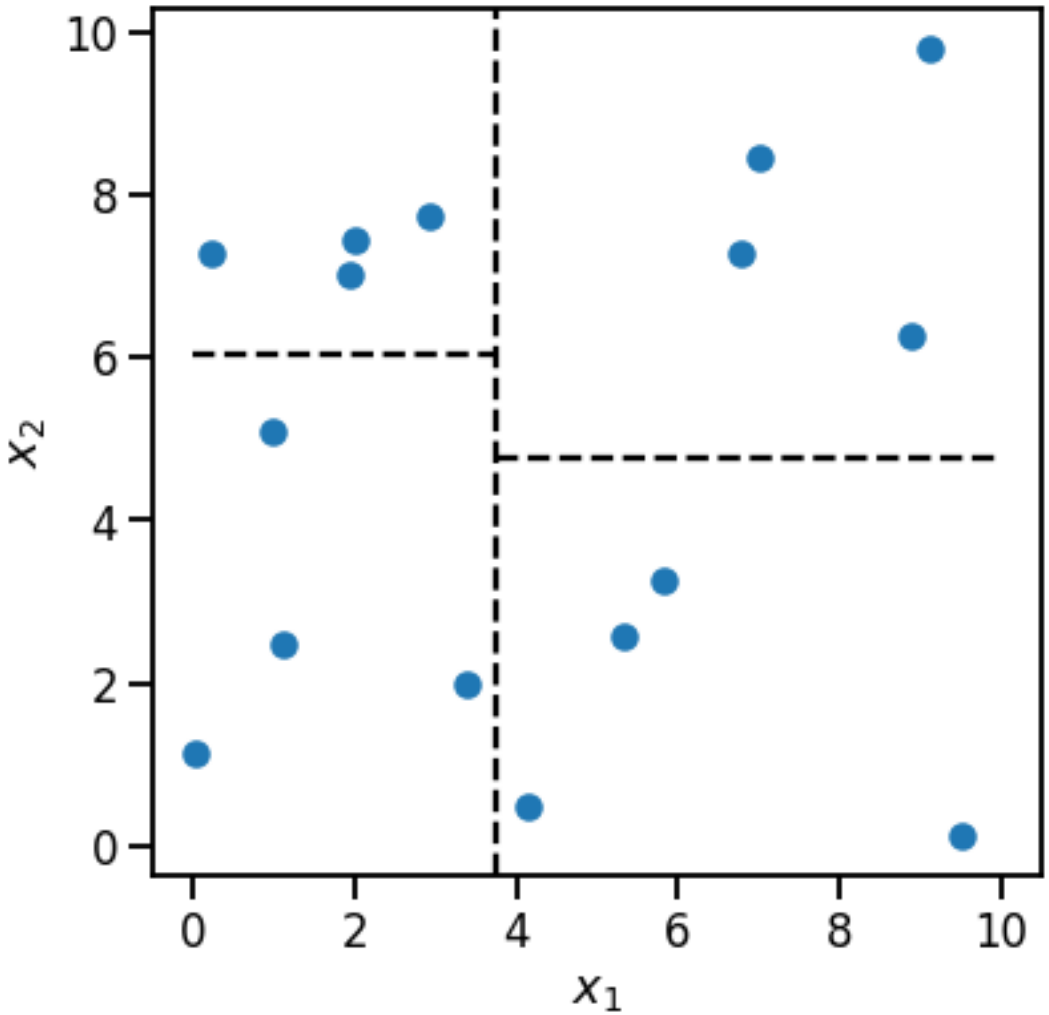
- Tujuan:
  - Monothetic: *common property*
  - Polythetic: kemiripan data dengan pengukuran jarak
- Irisan:
  - Hard clustering
  - Soft clustering
- Flat vs hierarchical

Metode Clustering

Metode *clustering* yang akan dibahas dalam kuliah ini:

- K-D Trees: *monothetic, hard boundaries, hierarchical*
- k-Means: *polythetic, hard boundaries, flat*
- Gaussian mixtures (EM algorithm): *polythetic, soft boundaries, flat*
- Agglomerative clustering: *polythetic, hard boundaries, hierarchical*

Clustering dengan K-D Trees

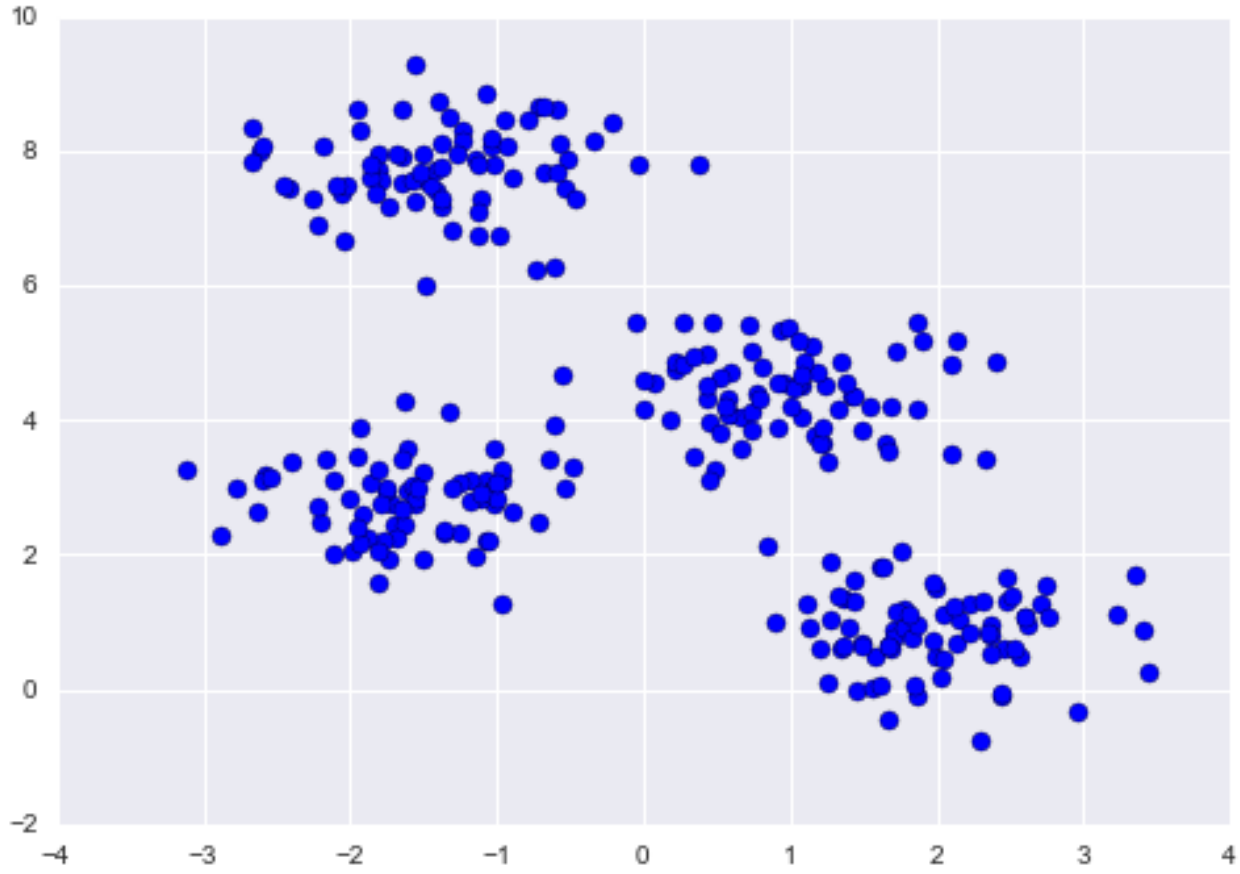


Gambar: Monothetic, hard boundaries, hierarchical

k-Means

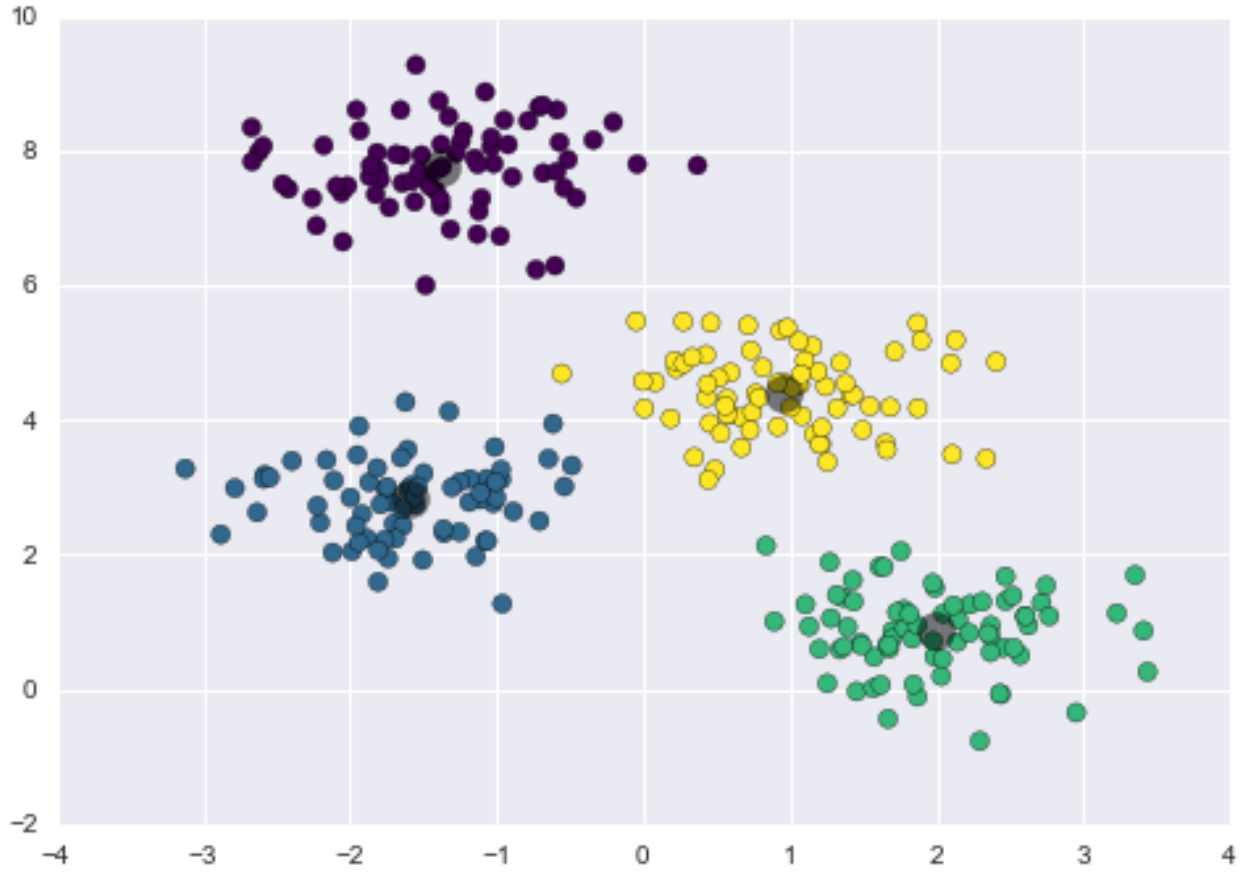
- Jumlah k ditentukan dari awal
- Tidak memerlukan label
- Menggunakan *centroid*, i.e. rata-rata nilai dari objek yang masuk dalam *cluster* tersebut
- Mencari *centroid* terdekat dari tiap objek

Contoh Data



Gambar: Contoh data dalam 2D [VanderPlas, 2016]

Hasil k-Means



Gambar: Setelah algoritma k-Means dijalankan [VanderPlas, 2016]

Algoritma: Expectation-Maximization

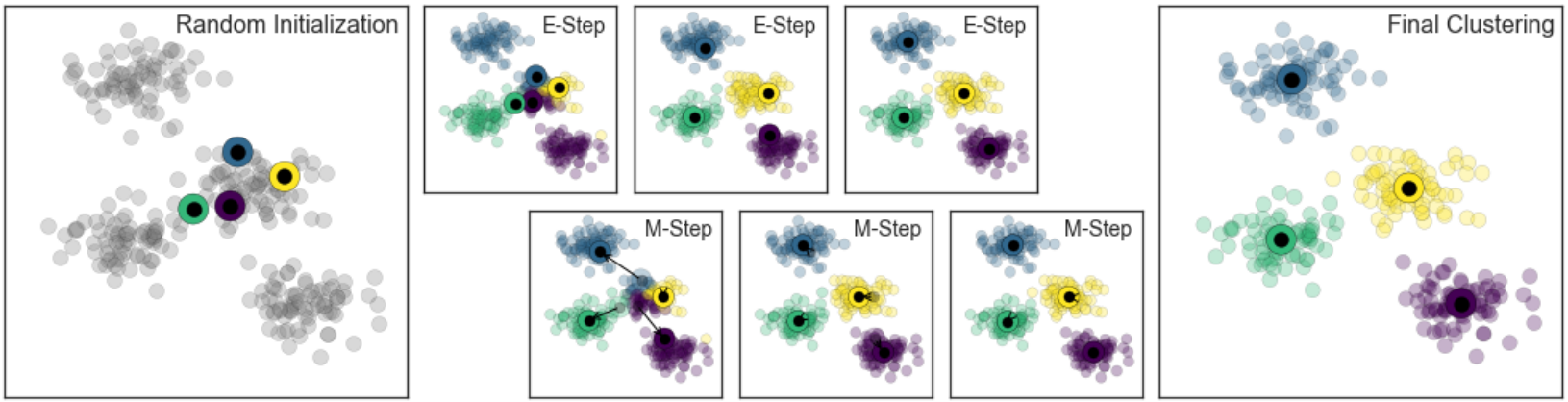
1. Inisialisasi k *centroid* secara acak
2. Ulangi hingga konvergen
   A. E-step: Masukkan tiap titik/objek ke *centroid* terdekat

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

   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



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

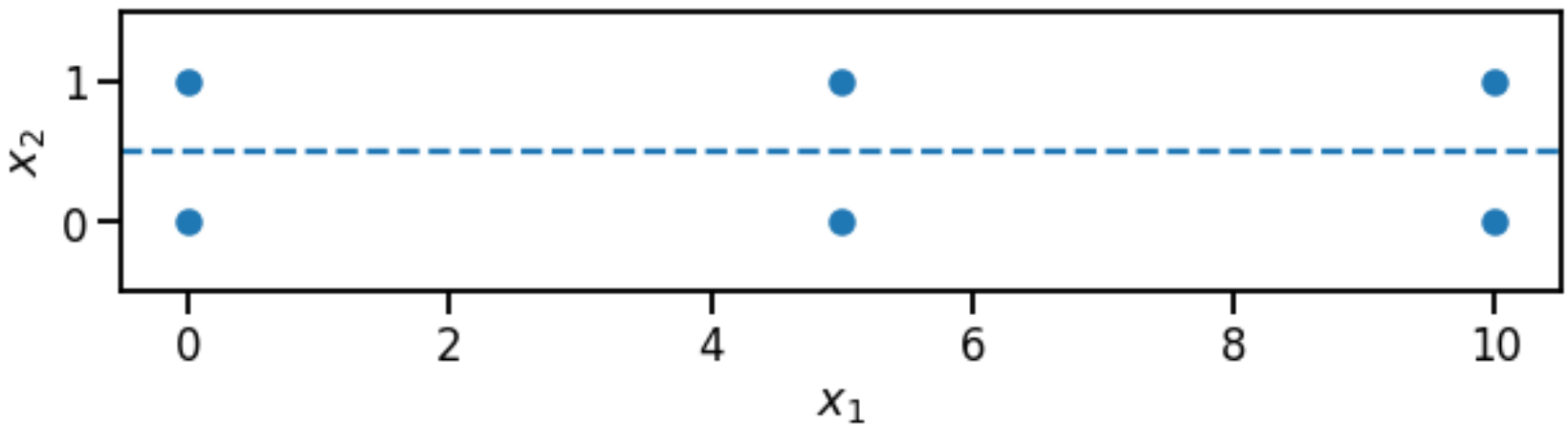
Perhatikan bahwa algoritma ini sangat bergantung pada inisialisasi *centroid*!

Properti dari k-Means

- Meminimalkan jarak agregat intra-klaster

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

- Konvergensi ke **minimum lokal**
- Poin yang berdekatan mungkin masuk ke klaster yang berbeda



Berapa nilai k yang optimal?

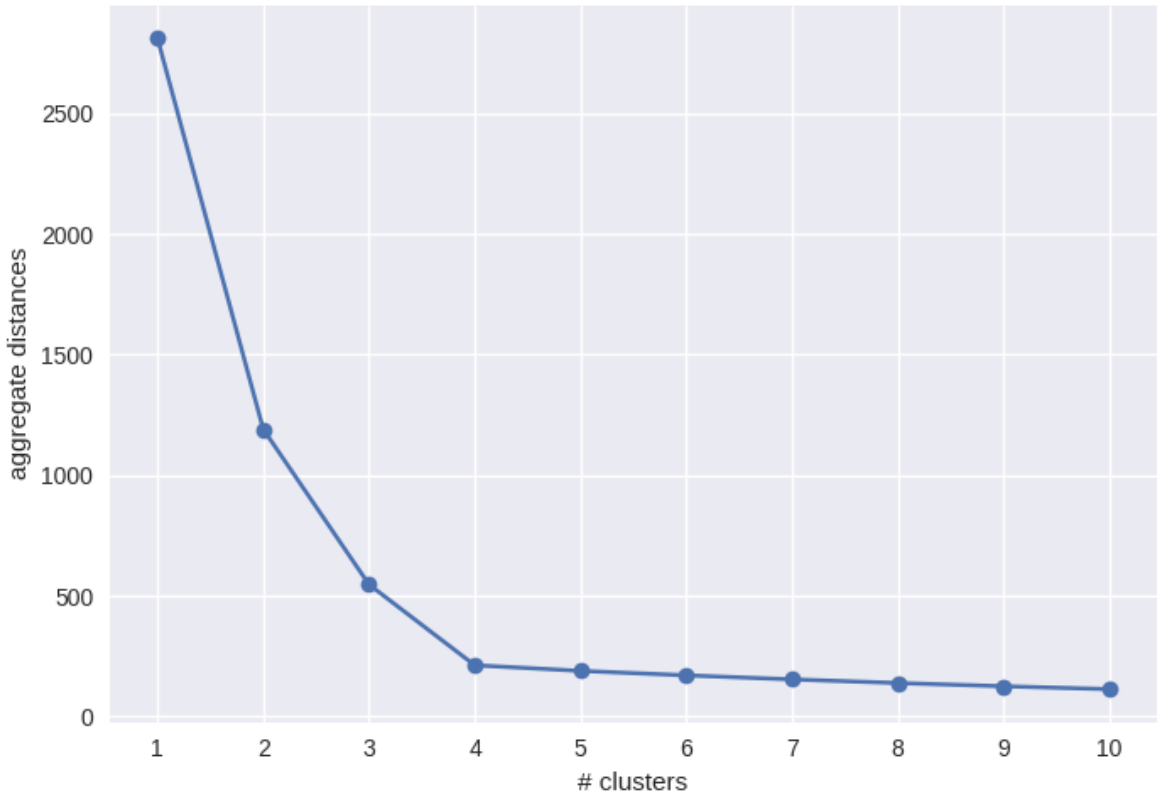
Menentukan Nilai k

- 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



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

Evaluasi

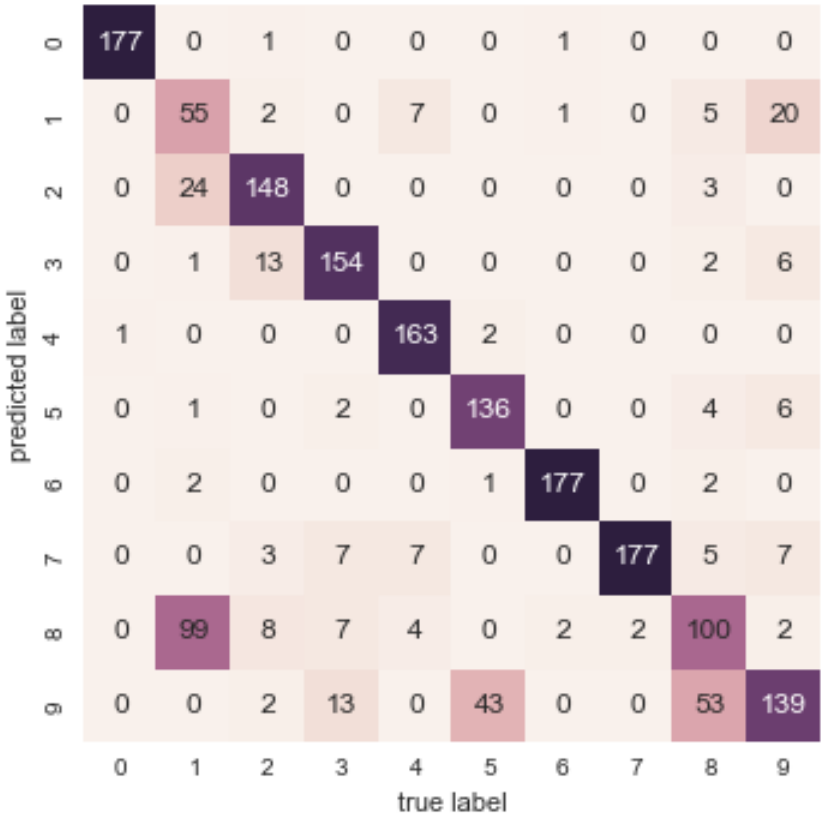
Evaluasi Klaster

- Ekstrinsik: untuk mengerjakan tugas lain
  - Representasi gambar dengan fitur berupa klaster
  - Menemukan pencilan
- Intrinsik: untuk diri sendiri
  - Memahami data – deskriptif
  - Klaster $\sim$ kelas, e.g. MNIST $\rightarrow$ 10 klaster
  - Perbandingan pasangan data dari klaster oleh manusia

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 akurasi atau F1
  - Bagaimana jika $N \neq K$?
  - Ada banyak cara, paling mudah dengan pendekatan *greedy*

Contoh Evaluasi Intrinsik



Gambar: Confusion matrix dari MNIST clustering [VanderPlas, 2016]

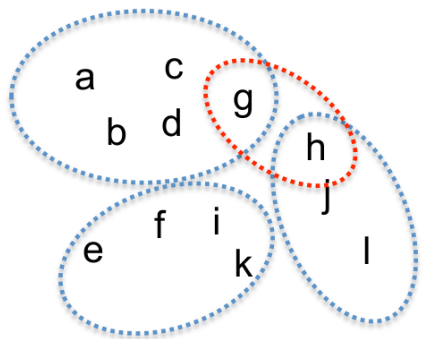
Contoh Evaluasi Intrinsik

	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

Gambar: Klaster karakter dalam Julius Caesar

Evaluasi Intrinsik: Perbandingan Antarpasangan

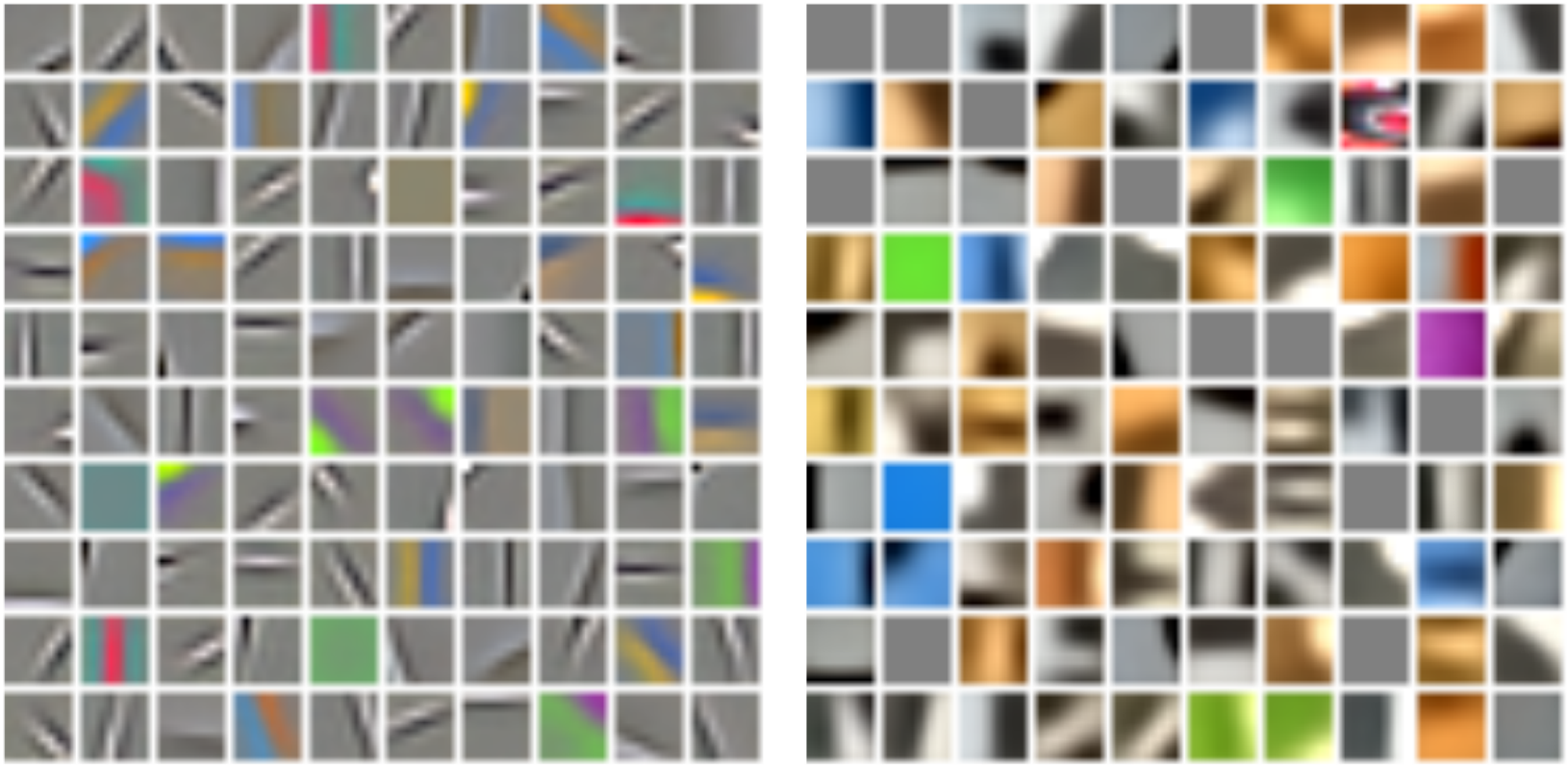
- Pasangan x_i, x_j apakah seharusnya berada dalam klaster yang sama?
- Hitung error, akurasi, F1
  - FN: pasangan x_i, x_j yang harusnya cocok, tapi berada dalam klaster yang lain (e,h)
  - FP: pasangan x_i, x_j yang harus tidak cocok, tapi berada dalam klaster yang sama (c,d)



Aplikasi Clustering

- Pemelajaran fitur [Coates, 2012]
- Kompresi gambar [VanderPlas, 2016]
- Sistem rekomendasi

Aplikasi: Pemelajaran fitur

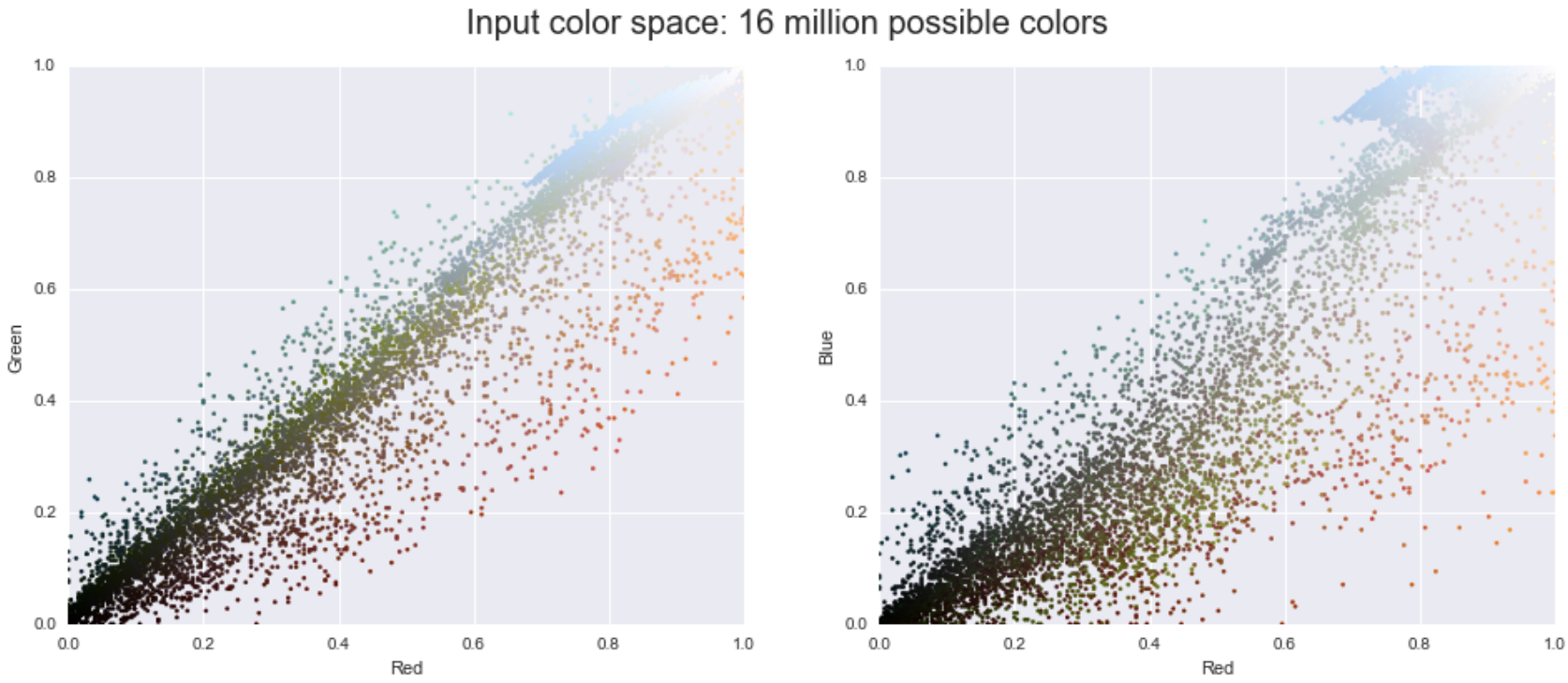


Gambar: Centroids dari CIFAR-10 dengan dan tanpa pemutihan [Coates, 2012]

Aplikasi: Kompresi gambar

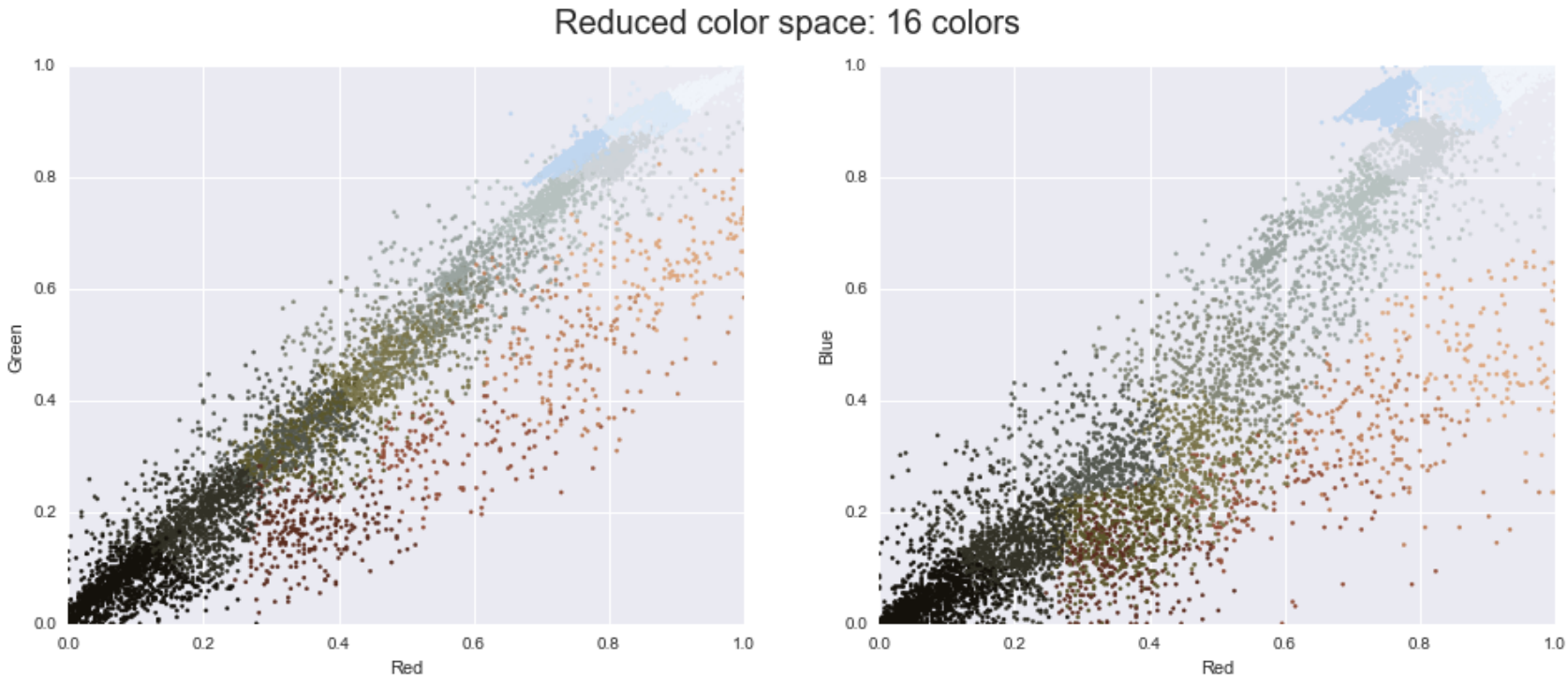
Gambar: Gambar yang akan dikompresi dengan *clustering* [VanderPlas, 2016]

Klaster warna



Gambar: *Clustering* warna dengan kompresi [VanderPlas, 2016]

Klaster warna



Gambar: *Clustering* warna dengan kompresi [VanderPlas, 2016]

Hasil kompresi gambar

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

Salindia ini dibuat dengan
sangat dipengaruhi oleh Lavrenko (2014)

Referensi

Jake VanderPlas (2016)

In Depth: k-Means Clustering

http://nbviewer.jupyter.org/github/jakevdp/PythonDataScienceHandbook/blob/master/notebooks/05.11-K-Means.ipynb

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