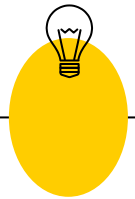
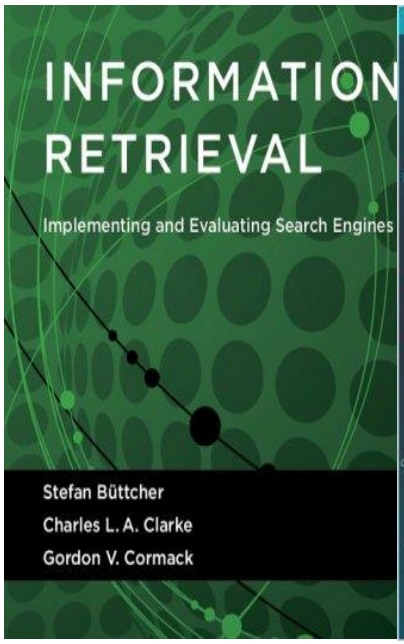
# Sistem Temu Kembali Informasi

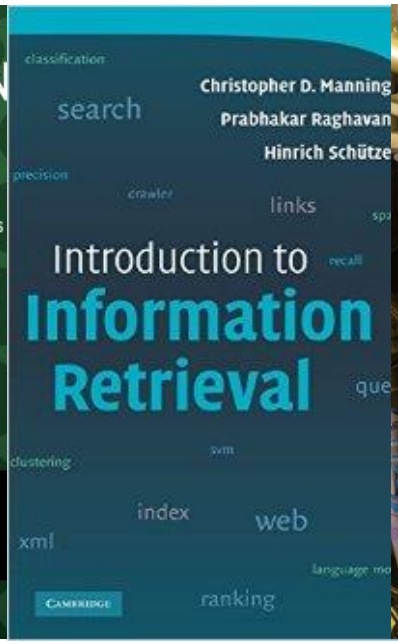
"Klasifikasi Dokumen dengan K-Nearest Neighbor (K-NN)"

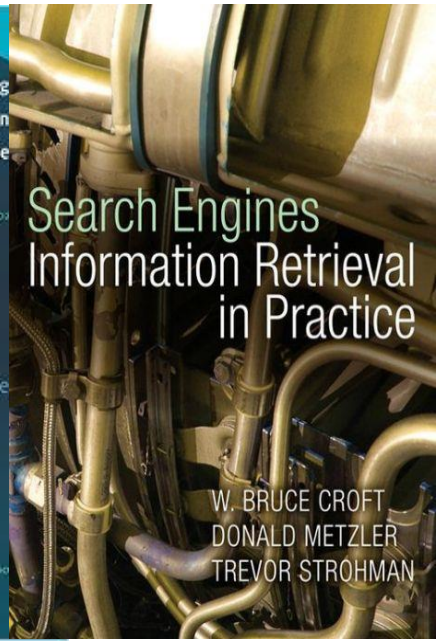


***Tim pengampu Dosen STKI***

# Buku Penunjang & Literatur



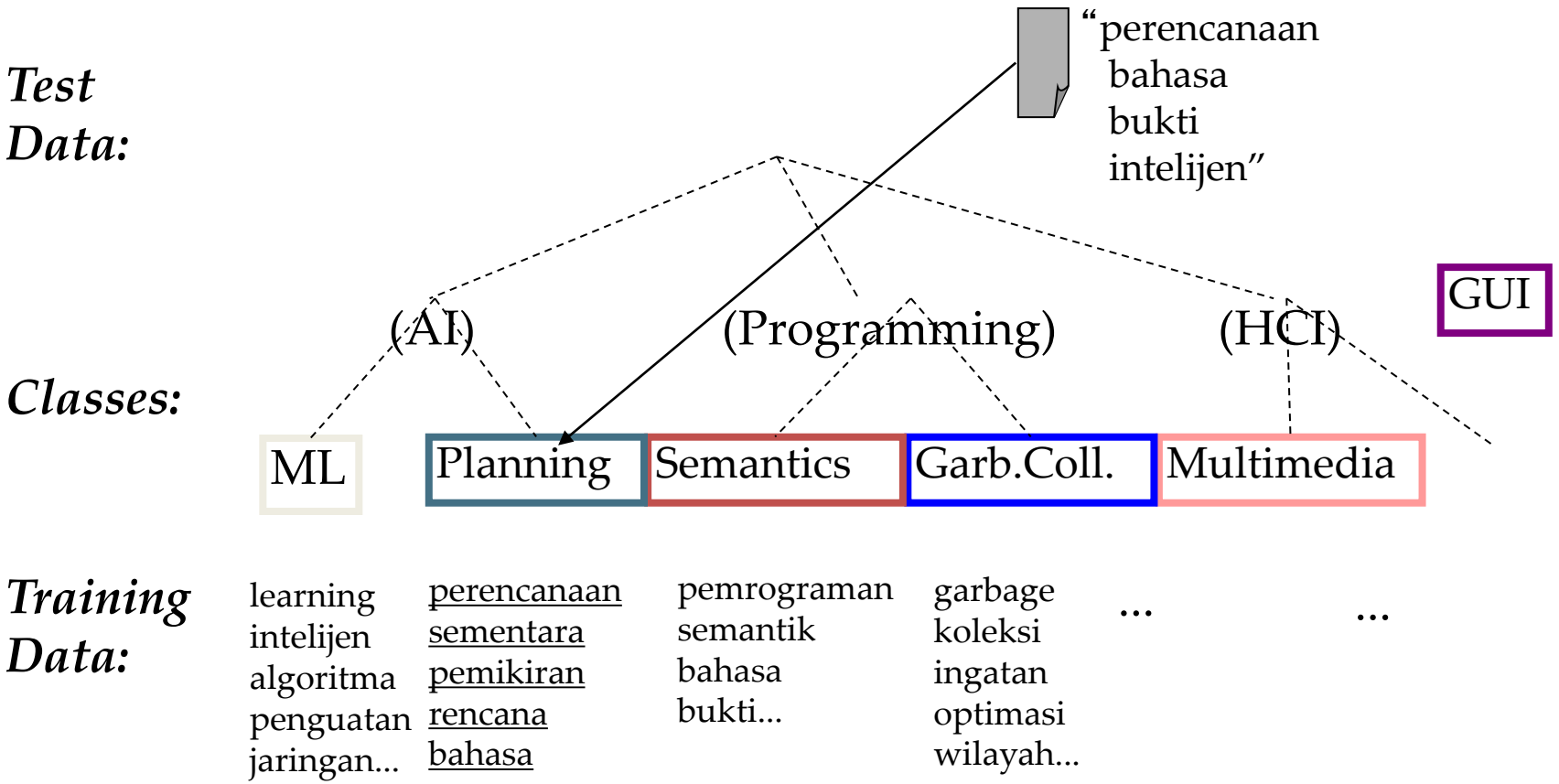



# Intro

**Algoritma K-Nearest Neighbor (K-NN)** adalah sebuah metode klasifikasi terhadap sekumpulan data maupun dokumen berdasarkan pembelajaran data yang sudah terklasifikasikan sebelumnya.

Termasuk dalam **supervised learning**, dimana hasil query instance yang baru diklasifikasikan berdasarkan mayoritas kedekatan jarak dari kategori yang ada dalam **K-NN**.

# Klasifikasi Dokumen



# Karakteristik K-Nearest Neighbor (K-NN)

- Klasifikasi K-NN umumnya didasarkan pada jarak *Euclidean* antara data uji dan data training yang ditentukan.

- K-NN menentukan kelas suatu objek data baru (data testing) dengan cara mencari pada kelompok k objek dalam data training yang paling dekat (mirip).

# Penerapan K-Nearest Neighbor (K-NN)

- K-NN digunakan dalam banyak aplikasi data mining, statistical pattern recognition, image processing, dll.
- Beberapa aplikasinya meliputi :
  - Pengenalan tulisan tangan
  - Satellite image
  - ECG pattern (menghasilkan pola yang mencerminkan aktivitas elektrik jantung)

# Kelebihan K-Nearest Neighbor (K-NN)

- **Sangat Nonlinear :** K-NN merupakan salah satu algoritma (model) pembelajaran mesin yang bersifat non-parametrik (model yang tidak mengasumsikan apa-apa mengenai distribusi instance di dalam data maupun dokumen).
- Mudah **dipahami dan diimplementasikan**

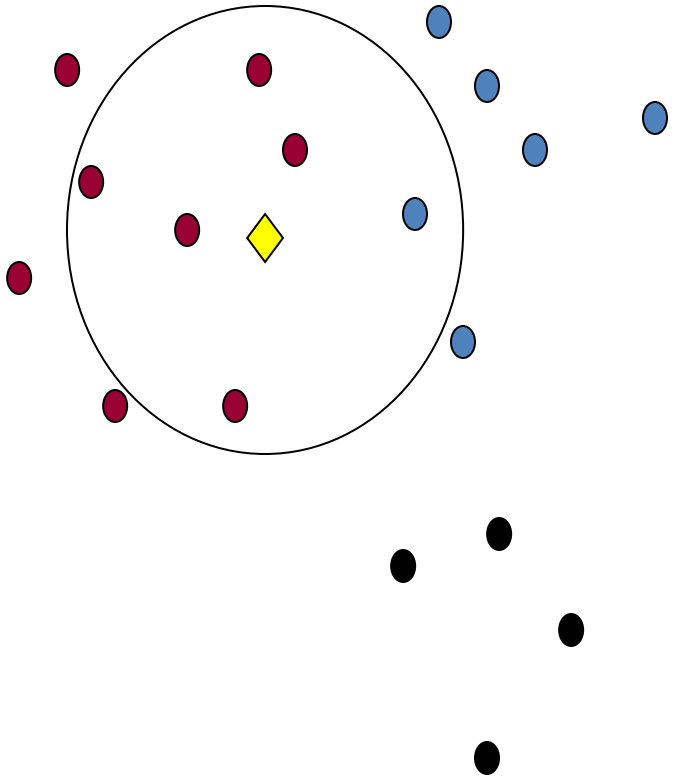
# Kekurangan K-Nearest Neighbor (K-NN)

- Perlu menunjukkan parameter K (jumlah tetangga terdekat).
- Tidak menangani nilai hilang (missing value) secara implisit.
- Sensitif terhadap data pencilan (outlier) terlebih yang terdapat ditengah-tengah class.
- Rentan terhadap variabel yang non-informative.
- Rentan terhadap dimensionalitas (banyaknya variabel) yang tinggi karena semakin banyak dimensi, ruang yang bisa ditempati instance semakin besar, sehingga semakin besar pula kemungkinan bahwa nearest neighbour dari suatu instance sebetulnya sama sekali tidak “near“.
- Rentan terhadap perbedaan rentang variable.
- Nilai komputasi yang tinggi.

# Tahapan Klasifikasi dengan K-NN

1. Lakukan **Pre-Processing** terhadap semua dokumen.
2. Hitung **kemiripan** vektor **dokumen yang dicari** dengan setiap dokumen yang telah terklasifikasi.
3. Urutkan hasil perhitungan kemiripan.
4. Ambil sebanyak k yang paling tinggi tingkat kemiripannya dengan dokumen yang dicari dan tentukan kelas dari dokumen tersebut.

# Example: k=6 (6NN)



P(science| ◆ )?

- Government
- Science
- Arts

# Studi Kasus
# Klasifikasi Dokumen dengan K-NN

| DOKUMEN | TERM YANG MEWAKILI DOKUMEN | CLASS |
|---|---|---|
| D1 | Partai Golkar dan Demokrat pada bertanding dalam kampanye 2009. | POLITIK |
| D2 | Pertandingan pertama antara Persema dan Persebaya diadakan di Malang. | OLAHRAGA |
| D3 | Sebagian besar wasit pada beberapa pertandingan sepakbola sering tidak adil. | OLAHRAGA |
| D4 | Partai demokrat memenangkan pemilu 2009 karena figur SBY | POLITIK |
| D5 | Pertandingan sepakbola Persebaya pada masa kampanye Pemilu 2009 akan ditunda. | ? |

- Terdapat 4 Dokumen Training (D1,D2,D3,D4) dan 1 Dokumen Testing (D5).
- Dokumen-dokumen teks tersebut dikategorikan (*classification*) menjadi 2, yaitu :
  - Class 1 => Politik
  - Class 2 => Olahraga

# Langkah 1. Pre-processing Dokumen

- **Langkah 1a.** Lakukan tokenisasi, stop word removal, dan stemming.

- Hasilnya :

| DOKUMEN | TERM YANG MEWAKILI DOKUMEN | CLASS |
|---|---|---|
| D1 | partai golkar demokrat tanding kampanye 2009 | POLITIK |
| D2 | tanding pertama persema persebaya malang | OLAHRAGA |
| D3 | besar wasit tanding sepakbola adil | OLAHRAGA |
| D4 | partai demokrat menang pemilu 2009 figur sby | POLITIK |
| D5 | tanding sepakbola persebaya kampanye pemilu 2009 tunda | ? |

# Langkah 1. Pre-processing Dokumen

- **Langkah 1b.** Tentukan bobot untuk setiap term dari 5 dokumen yang terlibat menggunakan Term Weighting TF-IDF.
- Hasilnya :

| | tf | | | | | | idf | Wdt=tf.idf | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Term | D1 | D2 | D3 | D4 | D5 | df | log(n/df) | D1 | D2 | D3 | D4 | D5 |
| partai | 1 | | | 1 | | 2 | 0,39794 | 0,39794 | 0 | 0 | 0,39794 | 0 |
| golkar | 1 | | | | | 1 | 0,69897 | 0,69897 | 0 | 0 | 0 | 0 |
| demokrat | 1 | | | 1 | | 2 | 0,39794 | 0,39794 | 0 | 0 | 0,39794 | 0 |
| tanding | 1 | 1 | 1 | | 1 | 4 | 0,09691 | 0,09691 | 0,09691 | 0,09691 | 0 | 0,09691 |
| kampanye | 1 | | | | 1 | 2 | 0,39794 | 0,39794 | 0 | 0 | 0 | 0,39794 |
| 2009 | 1 | | | 1 | 1 | 3 | 0,22185 | 0,221849 | 0 | 0 | 0,221849 | 0,221849 |
| pertama | | 1 | | | | 1 | 0,69897 | 0 | 0,69897 | 0 | 0 | 0 |
| persema | | 1 | | | | 1 | 0,69897 | 0 | 0,69897 | 0 | 0 | 0 |
| persebaya | | 1 | | | 1 | 2 | 0,39794 | 0 | 0,39794 | 0 | 0 | 0,39794 |
| malang | | 1 | | | | 1 | 0,69897 | 0 | 0,69897 | 0 | 0 | 0 |
| besar | | | 1 | | | 1 | 0,69897 | 0 | 0 | 0,69897 | 0 | 0 |
| wasit | | | 1 | | | 1 | 0,69897 | 0 | 0 | 0,69897 | 0 | 0 |
| sepakbola | | | 1 | | 1 | 2 | 0,39794 | 0 | 0 | 0,39794 | 0 | 0,39794 |
| adil | | | 1 | | | 1 | 0,69897 | 0 | 0 | 0,69897 | 0 | 0 |
| menang | | | | 1 | | 1 | 0,69897 | 0 | 0 | 0 | 0,69897 | 0 |
| pemilu | | | | 1 | 1 | 2 | 0,39794 | 0 | 0 | 0 | 0,39794 | 0,39794 |
| figur | | | | 1 | | 1 | 0,69897 | 0 | 0 | 0 | 0,69897 | 0 |
| sby | | | | 1 | | 1 | 0,69897 | 0 | 0 | 0 | 0,69897 | 0 |
| tunda | | | | | 1 | 1 | 0,69897 | 0 | 0 | 0 | 0 | 0,69897 |

# Langkah 2. Hitung Kemiripan Vektor Dokumen

## Langkah 2.

- Kemiripan antar dokumen dapat menggunakan cosine similarity. Rumusnya adalah sebagai berikut :

$$\cos(\Theta_{ij}) = \frac{\sum_k (d_{ik} d_{jk})}{\sqrt{\sum_k d_{ik}^2} \sqrt{\sum_k d_{jk}^2}}$$

# Langkah 2. Hitung Kemiripan Vektor Dokumen

## Langkah 2a.

- Hitung hasil perkalian skalar antara D5 dengan 4 dokumen (D1, D2, D3 & D4) yang telah terklasifikasi. Hasilnya perkalian dari setiap dokumen dengan D5 dijumlahkan (sesuai pembilang pada rumus sebelumnya).

| WD5*WDi | | | |
|---|---|---|---|
| D1 | D2 | D3 | D4 |
| 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 |
| 0,009392 | 0,009392 | 0,009392 | 0 |
| 0,158356 | 0 | 0 | 0 |
| 0,049217 | 0 | 0 | 0,049217 |
| 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 |
| 0 | 0,158356 | 0 | 0 |
| 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 |
| 0 | 0 | 0,158356 | 0 |
| 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0,158356 |
| 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 |
| 0,216965 | 0,167748 | 0,167748 | 0,207573 |

# Langkah 2. Hitung Kemiripan Vektor Dokumen

## Langkah 2b.

- Hitung panjang setiap dokumen, termasuk D5. Caranya, kuadratkan bobot setiap term dalam setiap dokumen, jumlahkan nilai kuadrat tersebut dan kemudian akarkan.

| Panjang vektor | | | | |
|---|---|---|---|---|
| D1 | D2 | D3 | D4 | D5 |
| 0,158356 | 0 | 0 | 0,158356 | 0 |
| 0,488559 | 0 | 0 | 0 | 0 |
| 0,158356 | 0 | 0 | 0,158356 | 0 |
| 0,009392 | 0,009392 | 0,009392 | 0 | 0,009392 |
| 0,158356 | 0 | 0 | 0 | 0,158356 |
| 0,049217 | 0 | 0 | 0,049217 | 0,049217 |
| 0 | 0,488559 | 0 | 0 | 0 |
| 0 | 0,488559 | 0 | 0 | 0 |
| 0 | 0,158356 | 0 | 0 | 0,158356 |
| 0 | 0,488559 | 0 | 0 | 0 |
| 0 | 0 | 0,488559 | 0 | 0 |
| 0 | 0 | 0,488559 | 0 | 0 |
| 0 | 0 | 0,158356 | 0 | 0,158356 |
| 0 | 0 | 0,488559 | 0 | 0 |
| 0 | 0 | 0 | 0,488559 | 0 |
| 0 | 0 | 0 | 0,158356 | 0,158356 |
| 0 | 0 | 0 | 0,488559 | 0 |
| 0 | 0 | 0 | 0,488559 | 0 |
| 0 | 0 | 0 | 0 | 0,488559 |
| 1,022236 | 1,633425 | 1,633425 | 1,989963 | 1,180592 |
| 1,011057 | 1,278055 | 1,278055 | 1,41066 | 1,086551 |

# Langkah 2. Hitung Kemiripan Vektor Dokumen

## Langkah 2c.

- Terapkan rumus cosine similarity. Hitung kemiripan D5 dengan D1, D2, D3 & D4 sebagai berikut:
  - Cos(D5,D1) = 0,21696/(1,08655*1,01106) = 0,1975
  - Cos(D5,D2) = 0,16775/(1,08655*1,27806) = 0,1208
  - Cos(D5,D3) = 0,16775/(1,08655*1,27806) = 0,1208
  - Cos(D5,D4) = 0,20757/(1,08655*1,41066) = 0,13542

- Hasil perhitungan tersebut diperlihatkan tabel berikut :

| D1 | D2 | D3 | D4 |
|---|---|---|---|
| 0,1975 | 0,1208 | 0,1208 | 0,13542 |

## Langkah 3. Urutkan Hasil Perhitungan Kemiripan

- Dari hasil sebelumnya untuk jarak setiap dokumen:

| D1 | D2 | D3 | D4 |
|---|---|---|---|
| 0,1975 | 0,1208 | 0,1208 | 0,13542 |

- Diurutkan berdasarkan jarak dengan nilai terbesar ke terkecil, sehingga menjadi:

| 1 | 2 | 3 | 4 |
|---|---|---|---|
| D1 | D4 | D2 | D3 |
| 0,1975 | 0,13542 | 0,1208 | 0,1208 |

# Langkah 4. Menentukan Kelas dari D5

- Ambil sebanyak k (k=3) yang paling tinggi tingkat kemiripannya dengan D5 dan tentukan kelas dari D5. Hasilnya :

| D1 | D4 | D2 |
|---|---|---|
| POLITIK | POLITIK | OLAHRAGA |

- Pilih kelas yang paling banyak kemunculannya. Untuk k=3 :
  - Kelas **POLITIK**, diwakili oleh 2 dokumen yaitu D1 dan D4.
  - Kelas **OLAHRAGA**, hanya diwakili oleh 1 dokumen, yaitu D2.

- Kesimpulan D5 terklasifikasi ke kelas ***POLITIK***.

# Kesimpulan & Review

- K-Nearest Neighbor merupakan salah satu metode machine learning yang melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut.
- Beberapa sifat pengklasifikasian dengan KNN yaitu :
  - Representasi teks biasanya sangat tinggi dimensi.
  - Algoritma bias tinggi yang mencegah overfitting umumnya bekerja paling baik dalam ruang berdimensi tinggi.
  - Untuk sebagian besar tugas pengkategorisasian teks, ada banyak fitur yang relevan dan banyak hal yang tidak relevan.

# Kuis (Latihan Soal)

| | | Teks | Class |
|---|---|---|---|
| **Training** | D1 | Sepakbola PSIS tahun ini semakin indah. | Olahraga |
| | D2 | Presiden Indonesia menaikan harga BBM. | Politik |
| | D3 | Partai politik Indonesia berburu suara. | Politik |
| | D4 | Timnas Indonesia gagal juara AFC | Olahraga |
| | | | |
| **Testing** | D5 | Presiden menaikan gaji Timnas Indonesia | ? |

- Jika dokumen-dokumen teks diatas diklasifikasikan ke dalam dua kelas yaitu Class 1 (Olahraga) = D1&D4 dan Class 2 (Politik) = D2&D3.
- **Pertanyaan:**
  - **Menggunakan k-NN, dengan nilai K=3, Tentukan kelas dari D5!**

# Thanks!

***Any questions ?***