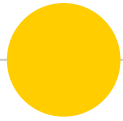


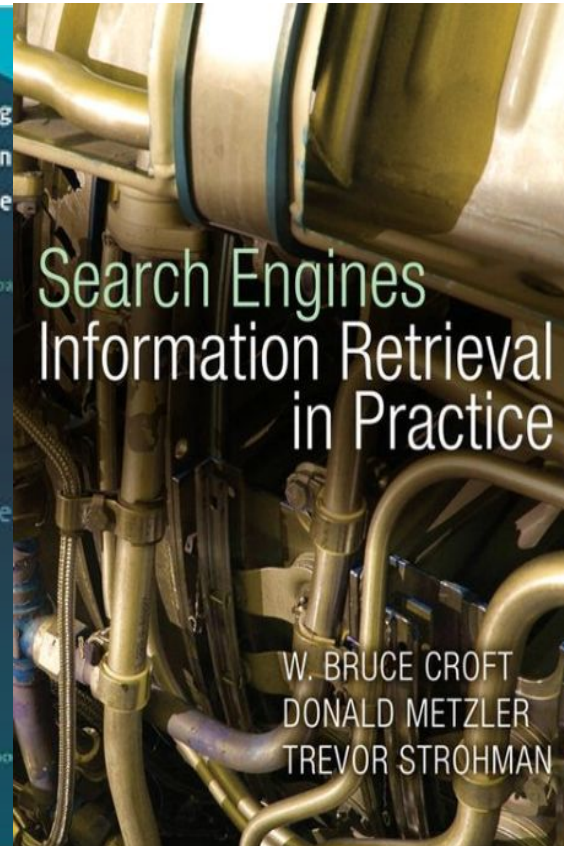
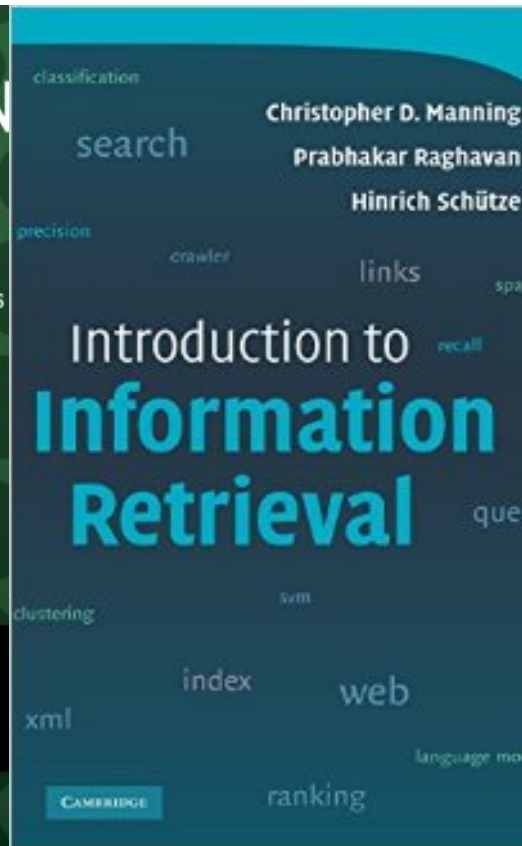
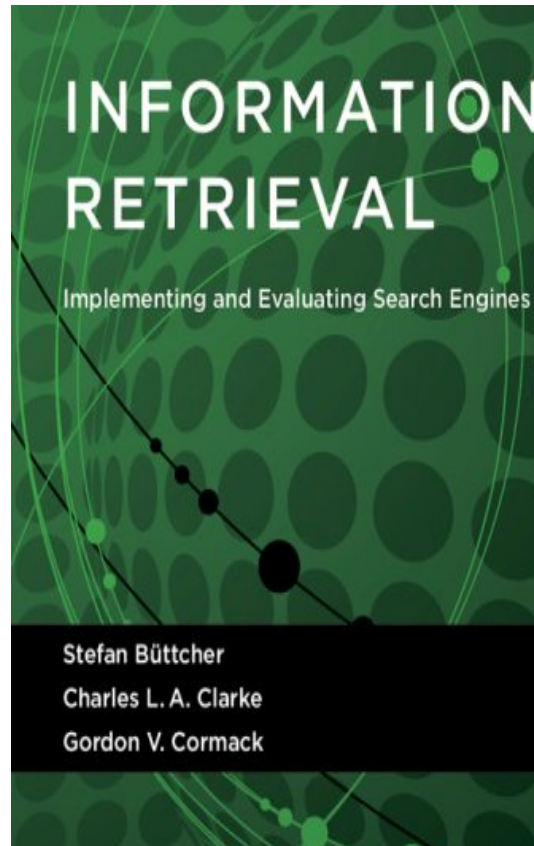
# Sistem Temu Kembali Informasi

## “Evaluasi”

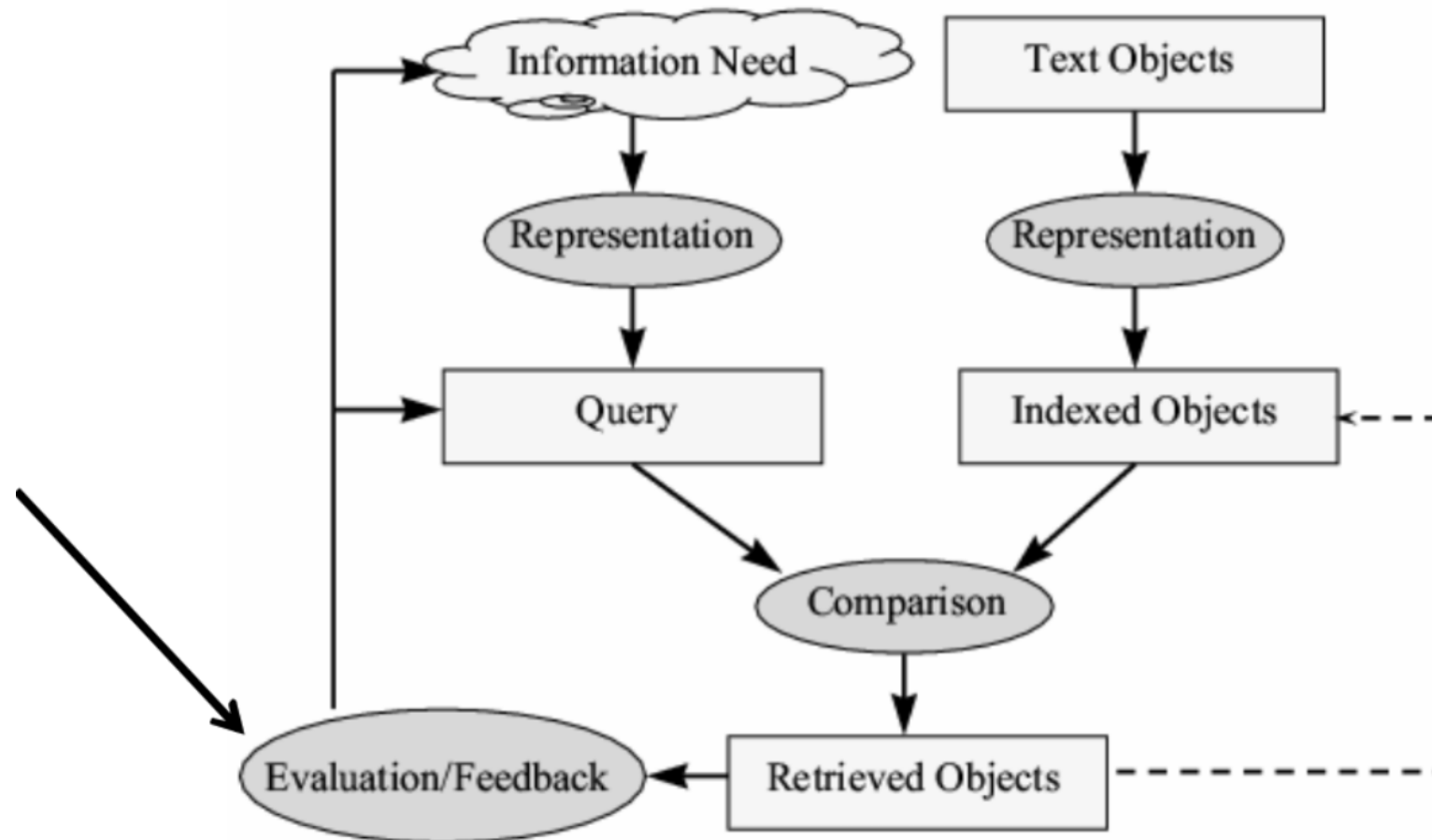


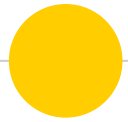


## Buku Penunjang & Literatur



# Evaluation «Sistem Temu Kembali Informasi»





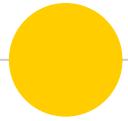
# Evaluation «Sistem Temu Kembali Informasi»

● Bagaimana mengetahui bahwa hasil yang diperoleh adalah relevan?

- Mengevaluasi suatu “*Search Engine*”
  - Benchmark (Patokan).
  - Presisi dan Recall.
  - Akurasi.
  - Ketidaksepakatan antar hakim.
  - Normalisasi potongan untung kumulatif.
  - Pengujian A/B.

● Rangkuman hasil :

- Membuat hasil yang bermanfaat bagi pengguna.

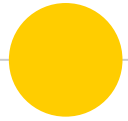


## Ukuran bagi Search Engine (SE)

- ⦿ Seberapa cepat **membangun indeks**
  - Jumlah dokumen/jam.
  - (Rerata ukuran dokumen).
- ⦿ Seberapa cepat **melakukan pencarian**
  - Latency sebagai fungsi dari ukuran indeks.
- ⦿ Ekspresi dari **bahasa query**
  - Kemampuan mengekspresikan kebutuhan informasi yang kompleks
  - Kecepatan pemrosesan query kompleks.
- ⦿ **UI (*User Interface*)**: Tertata dan mudah digunakan?
- ⦿ Gratis atau membayar?

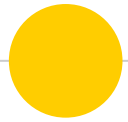
## ● Ukuran bagi Search Engine (SE) [2]

- Semua kriteria tersebut **measurable**: dapat dihitung kecepatan/ukurannya
  - Dapat diekspresikan dengan tepat.
- Tetapi ukuran kunci: **happiness** (kebahagian pengguna):
  - Apa itu?
  - Kecepatan respon/ukuran dari indeks adalah faktor penting.
  - Tetapi tidak asal cepat, jawaban yang tak berguna membuat pengguna kecewa.
- Perlu cara menghitung kepuasan pengguna.



# Ukuran Kebahagiaan Pengguna

- **Masalah:** Siapa pengguna yang akan dibuat bahagia?
  - Tergantung pada seting.
- **Web Engine:**
  - Pengguna mencari apa yang diinginkan dan Kembali ke Engine
    - Dapat diukur **angka** pengguna yang kembali.
  - Pengguna melengkapi tugasnya: **pencarian sebagai alat (sarana), bukan akhir.**
- **Situs eCommerce:** pengguna mencari apa yang diinginkan dan beli
  - Kepuasan bagi end-user atau situs eCommerce?
  - Waktu belanja atau % pencarian yang menjadi pembelian?
- **Recommender System:** pengguna mencari rekomendasi yang berguna atau sistem mampu memprediksi rating pengguna?

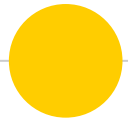


# Mengukur Kebahagiaan Pengguna

☉ **Enterprise** (perusahaan/pemerintah/kampus)  
harus *peduli* dengan “***produktifitas pengguna***”

- Berapa waktu yang dihemat oleh pengguna ketika mencari informasi?
- Banyak kriteria lain yang harus diperhatikan, terutama yang berkaitan dengan keleluasaan, kemudahan dan keamanan akses.

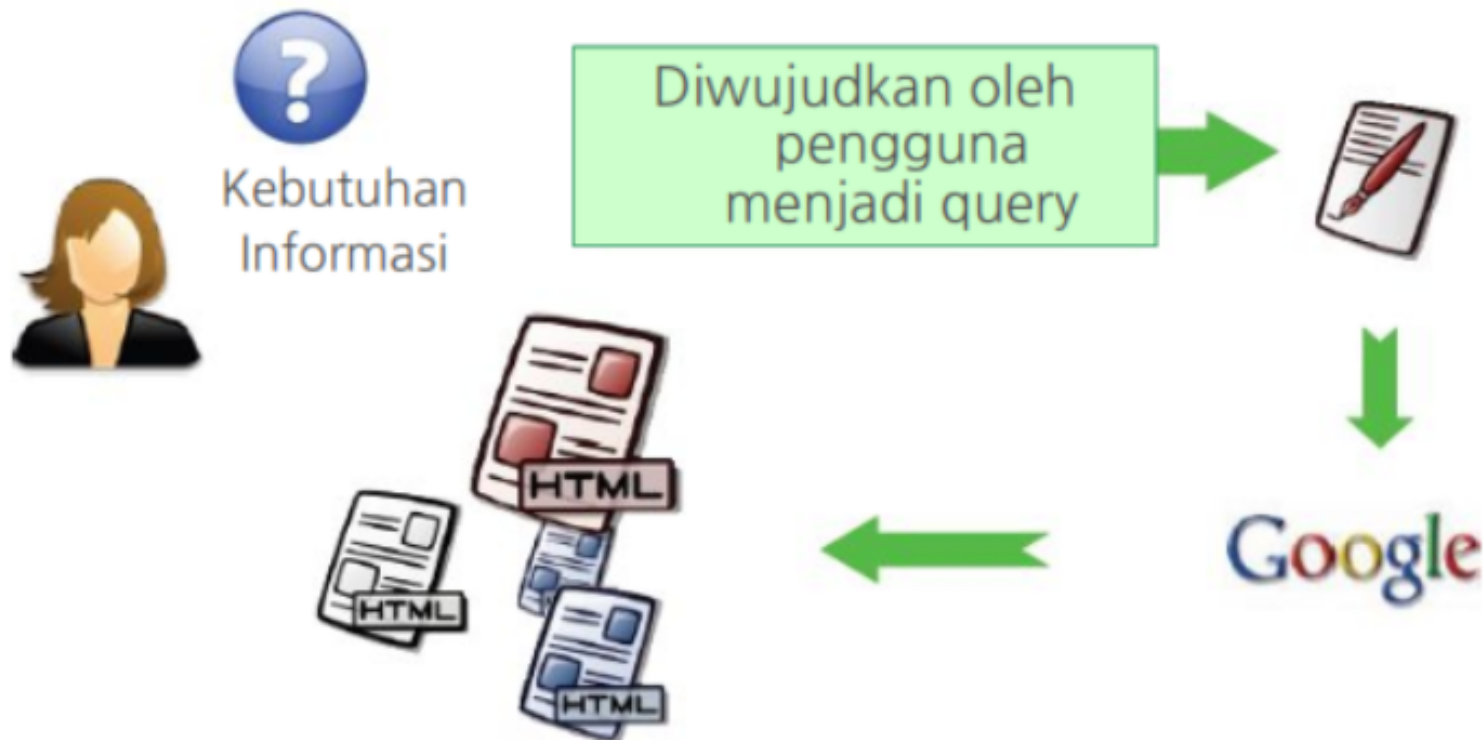




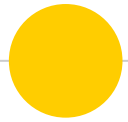
## Kebahagiaan: Sukar diukur

- ☉ Proxy paling umum: **Relevansi hasil pencarian.**
- ☉ Tetapi bagaimana mengukur relevansi?
- ☉ Ada metodologi => ada persoalan yang muncul.
- ☉ Ukuran relevansi memerlukan 3 elemen:
  1. Koleksi dokumen **benchmark.**
  2. Paket query **benchmark.**
  3. Biasanya taksiran biner: **Relevan** atau **Tak-Relevan** untuk setiap **query** dan setiap **dokumen.**
    - Ada yang tak biner, tapi tak standard.

# ● Kebutuhan => Query

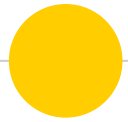


Kebutuhan informasi => Query => Search Engine => Hasil =>  
Browse atau Query => ...



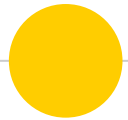
## Mengevaluasi Sistem STKI

- ☉ Kebutuhan informasi diterjemahkan ke dalam Query.
- ☉ Relevansi ditaksir relatif terhadap kebutuhan informasi, bukan **Query**.
- ☉ Misal, kebutuhan informasi: “I’m looking for information on whether using olive oil is effective at reducing your risk of heart attacks.”
- ☉ Query: *olive oil heart attack effective*
- ☉ **Evaluasi**: apakah dokumen menjawab kebutuhan informasi, atau hanya menyesuaikan kata-kata yang terkandung dalam query.

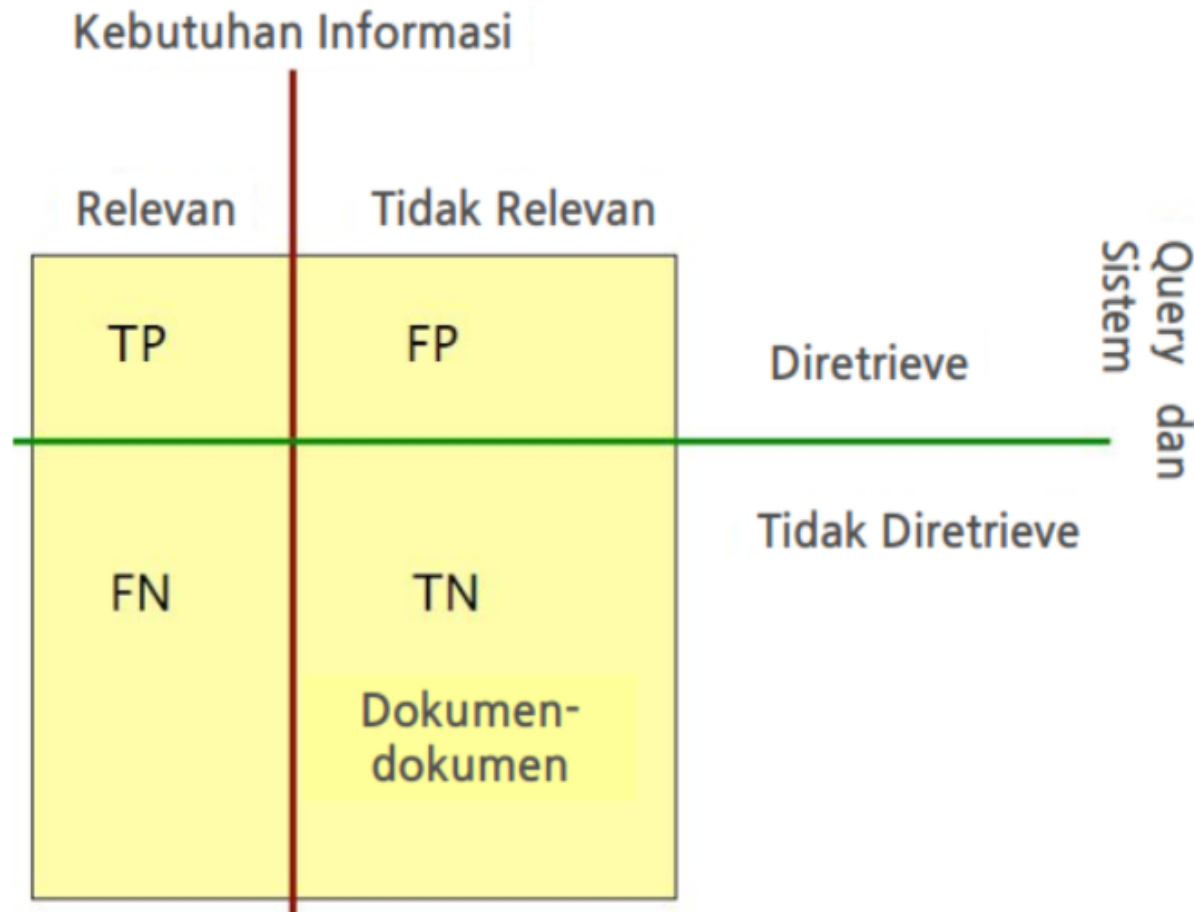


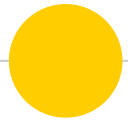
## Benchmark Relevansi Standard

- **TREC** – *National Institute of Standards and Technology (NIST)* telah menjalankan test bed TKI besar selama bertahun-tahun.
- Dokumen benchmark: **Reuters dan lainnya**.
- “Tugas-tugas Retrieval” ditetapkan:
  - Kadang kala sebagai query
- Pakar manusia menilai kedekatan setiap query dengan untuk setiap dokumen: **Relevan** atau **Tak-relevan**.
  - Atau setidaknya untuk subset dari dokumen yang dikembalikan oleh sistem untuk query tersebut.



# Relevansi & Dokumen Yang Ditemukan Kembali





## Evaluasi Retrieval Tak-Teranking: Presisi & Recall

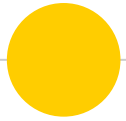
● **Presisi:** % dokumen yang diretrieve dan relevan =  $P(\text{relevan} \mid \text{diretrieve})$

● **Recall:** % dokumen relevan yang berhasil diretrieve =  $P(\text{diretrieve} \mid \text{relevan})$

	Relevant	Nonrelevant
Retrieved	tp	fp
Not Retrieved	fn	tn

● Presisi  $P = tp / (tp + fp) = tp / \text{diretrieve}$

● Recall  $R = tp / (tp + fn) = tp / \text{relevan}$



# Akurasi

- ⦿ Diberikan suatu query, suatu engine (***classifier***) mengelompokkan setiap dokumen sebagai “Relevan” atau “Tak-relevan”.
  - Apakah yang diretrieve terklasifikasi oleh engine sebagai “relevan” dan apakah yang tidak diretrieve diklasifikasikan sebagai “tak-relevan”.
- ⦿ Akurasi dari engine: % ketepatan dari klasifikasi
  - $(tp + tn) / (tp + fp + fn + tn)$
- ⦿ **Akurasi** adalah ukuran evaluasi yang umum digunakan dalam kerja klasifikasi ***machine learning***.
- ⦿ Mengapa ini bukan ukuran evaluasi yang sangat penting dalam STKI?

# ● Mengapa Tidak Hanya Akurasi?

- Bagaimana membangun Search Engine yang akurat 99.9999% dengan biaya rendah?

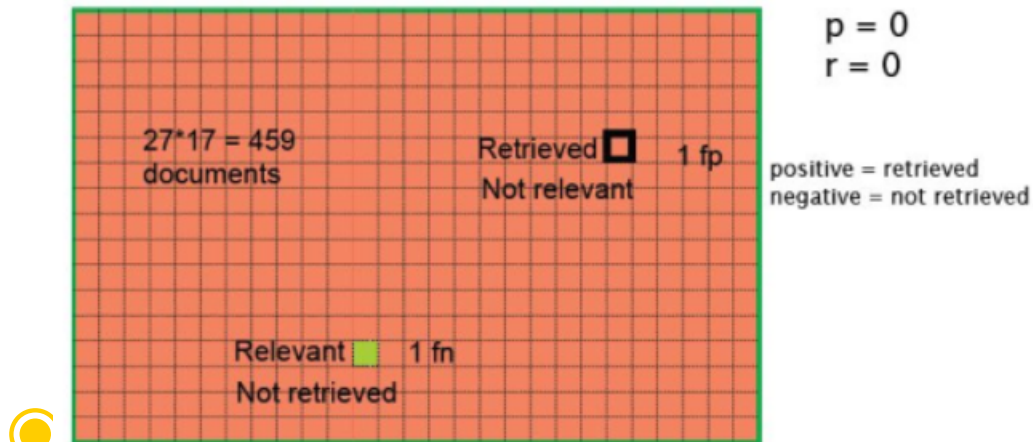


- **Sistem Temu Kembali Informasi** digunakan untuk mendapatkan sesuatu dan mempunyai toleransi tertentu terhadap sampah (*junk*).

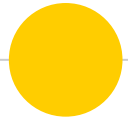


# Presisi, Recall & Akurasi

Presisi sangat rendah, Recall sangat rendah, Akurasi tinggi.



$$= (0 + (27*17 - 2)) / (0+1+1+(27*17 - 2))$$
$$= 0.996$$



## Kuis (Latihan Soal)

- ☉ Cari ***paper*** atau ***jurnal STKI*** tentang materi ***Evaluasi***, kemudian rangkumlah perbandingan antara beberapa metode dalam ***mengevaluasi*** konsep STKI tersebut kedalam bentuk artikel (***min. 500 kata***).



# Thanks!

*Any **questions** ?*